Hydrol. Earth Syst. Sci. Discuss., 8, 2739–2782, 2011
www.hydrol-earth-syst-sci-discuss.net/8/2739/2011/
doi:10.5194/hessd-8-2739-2011

Hydrology and Earth System Sciences Discussions

Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria

D. Brochero[1,2], F. Anctil[1], and C. Gagné[2]

[1]Chaire de recherche EDS en prévisions et actions hydrologiques, Department of Civil and Water Engineering, Université Laval, Québec (Québec), G1V 0A6, Canada
[2]Computer Vision and Systems Laboratory (CVSL), Department of Electrical and Computer Engineering, Université Laval, Québec (Québec), G1V 0A6, Canada

Received: 21 February 2011 – Accepted: 1 March 2011 – Published: 11 March 2011

Correspondence to: D. Brochero (darwin.brochero.1@ulaval.ca)

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

Abstract

Hydrological Ensemble Prediction System (HEPS), obtained by forcing rainfall-runoff models with Meteorological Ensemble Prediction Systems (MEPS), have been recognized as useful approaches to quantify uncertainties of hydrological forecasting systems. This task is complex both in terms of the coupling of information and computational time, which may create an operational barrier. The main objective of the current work is to assess the degree of simplification (reduction of members) of a HEPS configured with 16 lumped hydrological models driven by the 50 weather ensemble forecasts from the European Center for Medium-range Weather Forecasts (ECMWF). Here, the selection of the most relevant members is proposed using a Backward greedy technique with *k*-fold cross-validation, allowing an optimal use of the information. The methodology draws from a multi-criterion score that represents the combination of resolution, reliability, consistency, and diversity. Results show that the degree of reduction of members can be established in terms of maximum number of members required (complexity of the HEPS) or the maximization of the relationship between the different scores (performance).

1 Introduction

In hydrology, as in many applications, it is accepted that there is no superior model for every application under all circumstances (Duan et al., 2007; Alpaydin, 2010). In general, the models show different capacities to retain certain aspects of the hydrological processes, so focus on the forecasting process in a single model often leads to predictions that represent either a certain type of phenomena at the expense of others (Duan et al., 2007). In that sense, a multi-model approach may circumvent structural errors or inductive biases of the models (Georgakakos et al., 2004; Alpaydin, 2010).

The integration of the responses is central to the multi-model approach. The most likely solution is often taken as the average one or the result of a robust method of

HESSD

8, 2739–2782, 2011

integration such as neural networks (Jeong and Kim, 2009). However, this type of integration ignores the uncertainty assessment of hydrological forecasting – it is deterministic.

HEPEX (Hydrological Ensemble Prediction Experiment) pointed mainly to the following sources of uncertainty: model parameters, initial hydrological conditions, and hydroclimatological forecasts (Schaake et al., 2007). In this respect, the interest of the hydrological and meteorological community is clearly represented by numerous Hydrological Ensemble Prediction Systems (HEPS) to couple Meteorological Ensemble Prediction Systems (MEPS) and hydrological forecasting systems (Cloke and Pappenberger, 2009). The advantage of HEPS in terms of both forecast accuracy and the relative economic value in the decision-making has been shown (Roulin, 2007). An additional gain has been reported when there is a meta-combination involving multiple hydrological models and a MEPS (Velázquez et al., 2010).

The complexity of such HEPS becomes an operational burden when one has to evaluate several hundreds of scenarios at each time step. This situation may be even more dramatic considering the current trend of also considering several MEPS (Bougeault et al., 2010). Although computer capabilities continuously improve, multiple model runs are time-consuming, whereas real-time forecasts are needed as much for day to day management as for emergency actions.

This study considers the selection of members as a step towards simplification of a HEPS setup. More specifically, the HEPS under study is formed of 16 lumped hydrological models forced by the 50 meteorological inputs of the European Center for Medium-range Weather Forecasts (ECWMF), leading to a grand-ensemble of 800 members (see Sect. 3). Cloke and Pappenberger (2009) has already highlighted the computational demand of such a system as one of the main points to overcome in the future, either by new technologies (stochastic chip technology) or by efficient use of computing clusters.

As a compromise researchers have attempted to cluster MEPS for flood predictions in various ways: by lagging ensembles and deriving representative members through

HESSD
8, 2739–2782, 2011

hierarchical clustering over the domain of interest, and thus to produce a reduced ensemble set at higher resolution (Marsigli et al., 2001), or by analyzing the relation between atmospheric circulation patterns and extreme discharges (Ebert et al., 2007) or in a deterministic way ("best match" approach) by determining the location of the forecast that most resembled the rainfall pattern in the catchment (Xuan et al., 2009).

In this study the selection of members is based directly on the analysis of the weather uncertainty propagation into hydrological domains. This approach was tested in 10 catchments located in France for a period of sixteen months (from March 2005 to July 2006). The simplification was focused on the conservation of resolution, reliability, consistency and diversity of the HEPS with a small number of members. The evaluation of these characteristics of HEPS is based on the scores defined in Sect. 2. The Backward greedy selection technique, well-known in Machine Learning, is presented in Sect. 4.

Another important feature of the HEPS at hand is the short duration of the series. This has been highlighted by several authors as a negative point in the evaluation of system performance in the case of extreme events (Renner et al., 2009; Cloke and Pappenberger, 2009). Similarly the technique displayed for the selection of members should be trained and validated on different data sets to avoid the known problem of overfitting (Alpaydin, 2010). This additional requirement highlights the need for schemes to benefit from information such as bootstrapping or cross-validation. Indeed, one variation of k-fold cross-validation with random choice of blocks of information is proposed in Sect. 5. In the same section, is presented the mean rank of members' elimination as a technique to combine the experiments results of the k-fold cross-validation to obtain a solution that shows the integration of the behavior of HEPS at different periods of time. In Sect. 6, the results and discussion are presented, while conclusions and a guideline for future work are drawn in Sect. 7.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

2 Verification statistics for ensemble forecasts

In ensemble forecasting, the focus is taken off finding the single best estimate of the streamflow (i.e. finding a "perfect" model), and aims at finding the best possible estimate of the forecasts' uncertainty. Therefore, instead of forecasting a single streamflow value for each lead time, an ensemble forecasting system produces *d* member forecasts. 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 (Boucher et al., 2009).

To assess the quality of the ensemble forecast, several statistical measures should be considered concurrently (Cloke and Pappenberger, 2009). In this paper several metrics are considered, including the Continuous ranked probability score (CRPS), the ignorance score (IGNS), the reliability diagram, the rank histogram, the median coefficients of variation, and the combination of all of these.

In some cases it is necessary to establish a priori a probabilistic distribution function that fits systematically the prevision ensembles for each time step. In the hydrological community is accepted that an adjustment of the gamma distribution makes more sense than a normal distribution given asymmetry in the distribution of precipitation and discharge (Vrugt et al., 2008), however, the gamma function evaluation involves more complex than the normal distribution which has explicit mathematical expressions. Székely (2003) proposes Monte Carlo techniques for the adjustment of any distribution to the ensembles. For this study, some simulations were performed to evaluate differences between normal and gamma distributions in the case of CRPS and IGNS, the results showed minor variations in contrast with a high computational cost (1.7 h vs. 47 h of calculation). However, it is important to note that this similarity is evaluated inside the ensembles with previsions varying between 30 and 800 members, as detailed below; in small samples it is expected that the results represent the expected asymmetry of information.

HESSD
8, 2739–2782, 2011

2.1 Continuous ranked probability score

Assuming that the forecast ensembles ($\boldsymbol{y}^t$) are normally distributed with mean μ_t and variance σ_t^2 at time t, the Continuous Ranked Probability Score (CRPS) at the time t is defined by Eq. (1) (Gneiting and Raftery, 2007):

$$\mathrm{CRPS}(F(\boldsymbol{y}^t), o^t) = \sigma_t \left[\frac{1}{\sqrt{\pi}} - 2\phi\left(\frac{o^t - \mu^t}{\sigma^t}\right) - \left(\frac{o^t - \mu^t}{\sigma^t}\right)\left(2\Phi\left(\frac{o^t - \mu^t}{\sigma^t}\right) - 1\right)\right], \quad (1)$$

where $F(\boldsymbol{y})^t$ represents the cumulative distribution function, o^t denotes the observation; ϕ and Φ denote the normalized variables for probability density function and cumulative distribution function, respectively. The goal is to minimize the mean CRPS. In the rest of this article when a CRPS is represented for a scalar value, this evokes the mean value of this score. This criterion combines two components: reliability and potential CRPS (Hersbach, 2000; Boucher et al., 2010), making it theoretically a multi-purpose criterion.

2.2 Ignorance score

The ignorance score (IGNS) is related to the concepts of information quantity and entropy developed in information theory (Williams, 1997). The IGNS, as a particular version of the relative entropy, is calculated from the evaluation of the logarithm of the ensemble probability density function at the point corresponding to the observation o^t (Roulston and Smith, 2002).

$$\mathrm{IGNS}(\boldsymbol{y}, o)^t = -\log_2\left[f(\boldsymbol{y}^t)_{o^t}\right] \quad (2)$$

Note that if $f(\boldsymbol{y}^t)_{o^t} = 0$ then, according to Eq. (2), an infinite number of bits[1] is assigned to this score, so forecasters should replace zero forecast probabilities with small probabilities based on the uncertainties in the $f(\boldsymbol{y})$ forecast (Roulston and Smith, 2002).

[1]When logs are taken to the base two, a "bit" is the unit of information (Williams, 1997).

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

Here, it was assumed that the ensemble forecasts are normally distributed and the individual infinite scores were replaced by the next worst non-infinite value. Usually the mean IGNS represents this score.

2.3 Reliability diagram error

Given that m denotes the different M thresholds of probability to assess, the reliability of the system can be directly measured from the comparison of these M thresholds with the conditional probability of observation as a function of the forecast (o_m). Since observation of the event is dichotomous ($r^t = 1$ if the event occurred and $r^t = 0$ otherwise) such conditional probability or relative frequency observed $\bar{o}_m$ is given by Eq. (3):

$$\bar{o}_m = \frac{1}{N}\sum_{t=1}^{N} r^t, \qquad r^t = \begin{cases} 1 \text{ if } o^t \in I_m \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

where N is the number of forecast-observation pairs used in verification. The goal is to have well-calibrated forecast systems where the relative frequency is essentially equal to the probability of the forecast, i.e. $\bar{o}_m \approx I_m$ (Wilks, 2005). The plot of the conditional probability versus the probability of the forecast (I_m) is called the reliability diagram (RD). In this study, as discussed later in Sect. 4, it is necessary to establish a single target value, so the reliability is evaluated from the calculation of the mean square error (MSE) between the probability forecast and the observed frequency (Eq. 4), as has been suggested by Wilks (2005). These distances are all small for well-calibrated forecasts.

$$\mathrm{RD}_{\mathrm{MSE}}(\mathbf{Y}, \boldsymbol{o}) = \frac{1}{M}\sum_{m=1}^{M} (\bar{o}_m - I_m)^2 \tag{4}$$

HESSD
8, 2739–2782, 2011

2.4 Rank histogram (δ ratio)

Consider the evaluation of N ensemble forecasts, each of which consists of d ensemble members, in relation to the N corresponding observed values for the predictand. For each of these N sets, if the d members and the single observation all have been drawn from the same distribution, then the rank of the observation within these $d+1$ values is equally likely to take on any of the values $c = \{1, 2, 3, \ldots, d+1\}$. For example, if the observation is smaller than all d ensemble members, then its rank or class is $c = 1$. If it is larger than all the ensemble members, then its rank is $c = d+1$ (in the case of equality of observation with one or more of the ensemble members, the rank is chosen randomly). For each of the N forecasting occasions, the rank of the observation (S_c) within this $d+1$ member distribution is tabulated. Collectively these N verification ranks are plotted in the form of a histogram to produce the rank histogram (Wilks, 2005). The flatness of the corresponding rank histogram is therefore a measure of the reliability of the prediction system. Because of the finiteness of the number N of realizations of the prediction process over which the validation is carried out, the rank histogram cannot be expected to be exactly flat. For a reliable system, S_c has expectation $N/(d+1)$, while the deviation of the histogram from flatness (Δ) is measured by Eq. (5) (Talagrand et al., 1997).

$$\Delta = \sum_{c=1}^{d+1} (S_c - h_{\text{ref}})^2 \quad \text{where} \quad h_{\text{ref}} = \frac{N}{(d+1)}, \tag{5}$$

A reliable system has expectation of $\Delta_0 = dN/(N+1)$. The δ ratio, proposed by Talagrand et al. (1997) is used ($\delta = \Delta/\Delta_0$) as a measure of the reliability of an ensemble prediction system for a scalar variable. A value of δ that is considerably larger than 1 is a proof of unreliability. A value that would be considerably less than 1 would indicate that the successive realizations of the prediction process are not independent, and that the verifying observation tends to fall preferentially in intervals in which it has fallen less frequently in the previous realizations. Such a situation is of course very unlikely (Wilks, 2005; Talagrand et al., 1997).

2.5 Median of coefficients of variation

The coefficient of variation (cv) is defined as the ratio of the standard deviation to the mean. Although this measure is not established in the hydrological community as a probabilistic forecast evaluation criterion, its assessment leads to a measure of the reliability of the central value (Kottegoda and Rosso, 2009): the ensemble dispersion. Maximizing this metric thus favors dispersion. However, it is important to understand that the successful implementation of this ratio depends largely on the degree of individual adjustment of the hydrological models, because the cv is unable to detect a significant bias of the central value. A scalar value is obtained from the calculation of median cv (Eq. 6).

$$\mathrm{MDCV}(\mathbf{Y}) = \mathrm{med}_{t=1}^{N}\ cv(\boldsymbol{y}^t) \tag{6}$$

2.6 Combined criterion

Selecting only one criterion may give a partial view of the forecast performance and even be misleading. The combination of several metrics into one diagram has already been evaluated (Taylor, 2001), but is inappropriate for this study because a scalar objective value is required for the selection procedure. Thus, the integration of scores is addressed in obtaining a single index, whose interpretation can sometimes be difficult because good performance should be defined from a good performance in all respects and not great performance on some aspects and low performance on others. Another element to consider is the weight associated with each component of the combined criterion.

A criterion combining the CRPS, the IGNS, the reliability diagram, the rank histogram and the coefficient of variation is proposed. The combined score (Eq. 7) exploits values

HESSD
8, 2739–2782, 2011

of the initial 800-member set (ie) as a normalization factor for the scores of the subset of members (se).

$$CC = w_1 \frac{\overline{\text{CRPS}}_{\text{se}}}{\overline{\text{CRPS}}_{\text{ie}}} + w_2 \frac{z_1 - \overline{\text{IGNS}}_{\text{se}}}{z_1 - \overline{\text{IGNS}}_{\text{ie}}} + w_3 \frac{\text{RD}_{\text{MSEse}}}{\text{RD}_{\text{MSEie}}} + w_4 \frac{\delta_{\text{se}}}{\delta_{\text{ie}}} + w_5 \frac{z_2 - \text{MDCV}_{\text{se}}}{z_2 - \text{MDCV}_{\text{ie}}} \tag{7}$$

It is also necessary to set thresholds to define the behavior of MDCV and IGNS in terms of the minimization of all the factors together. Preliminary analysis showed that $z_1 = -2$ and $z_2 = 1$ covers all scenarios of the series at hand. The weight assigned to the reliability (as a critical factor) corresponds to twice that of the other factors, which have a unit weight.

3 HEPS configuration

The basic configuration of HEPS was established by Velázquez et al. (2010) in their work on the comparison of the performance of different types of HEPS, using uncertainties from the meteorological input and from the hydrological model structure. The scenarios are built on the basis of 16 different lumped rainfall-runoff model structures and 9-day ECMWF ensemble and deterministic forecasts. The systems were implemented over 28 French catchments, representing a large range of hydro-climatic conditions, and evaluated over a period of 17 months. The present study relies on a random selection of 10 catchments (Fig. 1), whose characteristics are presented in Table 1. The codification follows the Banque Hydro database (http://www.hydro.eaufrance.fr). The rainfall observations originate from the SAFRAN Météo-France and the flow data are for the Banque Hydro database. As already mentioned, the rainfall probabilistic forecast are from the ECMWF.

HESSD
8, 2739–2782, 2011

3.1 ECMWF – global atmospheric model

The current ECWMF deterministic forecast provides a horizon of up to 240 h. Here, the rainfall amounts were accumulated at 24 h time steps, starting at 0 h to match with observed daily data, which resulted in nine daily lead times.

A detailed description of the ECWMF model can be found in Molteni et al. (1996) or Buizza (2005). The 50 perturbed forecasts of the ensemble system result from a perturbation technique, which is based on a mathematical method called singular vector analysis (Buizza and Palmer, 1995). It attempts to locate the dynamically most unstable regions of the atmosphere by calculating where small initial uncertainties would affect a 48-h forecast most rapidly. From these, 50 alternative forecasts are produced. All the different initial states are a priori assumed to be equally likely, i.e. the unperturbed forecast is not necessarily the most probable (Gouweleeuw et al., 2005).

3.2 Hydrological models

The sixteen hydrological models applied in this study are lumped reservoir-type models and correspond to various conceptualizations of the rainfall-runoff transformation at the catchment scale (Table 2).

They are of low to moderate complexity: the number of parameters to calibrate against observed data ranges from 3 to 13. They all include a soil moisture accounting procedure in their representation of the hydrological production function, but with various formulations (linear or non linear, possibly with several soil layers, etc.).

The routing module includes from 1 to 5 linear or non linear stores, as well as unit hydrographs or pure time delays. Some of the models include a non-conservative function to adjust the water balance (correction factors of inputs or groundwater exchange functions). All the models were applied in the same conditions by Velázquez et al. (2010); some original model structures were modified to match this framework. They were run at a daily time step, using the same rainfall and potential evapotranspiration inputs, and were calibrated with the same optimization procedure using a local search procedure (Edijatno et al., 1999), applied in combination with a pre-screening of the parameter space (Mathevet, 2005).

The hydrological models were calibrated with 29 years as mean length. The objective function was the RMSE. It is beyond the scope of this article to present these models and the different settings in each one of them had to place. A detailed explanation of each model and the various adjustments can be found in (Perrin, 2000) with the exception of models GR4J (Perrin et al., 2003), MORDOR (Garçon, 1999), MOHYSE (Fortin and Turcotte, 2006; Roy et al., 2010), SIMHYD (Chiew et al., 2002) and Probability-distributed store (PDS) (Yadav et al., 2007).

3.3 HEPS results without selection of members

Velázquez et al. (2010) highlighted mainly two types of HEPS: 1) The 16-member ensemble: obtained by running all 16 hydrological models with the deterministic meteorological forecast as input (control meteorological condition), and 2) The 800-member ensemble: all 16 models are driven by the 50 forecast members from the MEPS and all the outputs are considered as a single ensemble (perturbed meteorological conditions). Figure 2 provides an illustration of both HEPS, for the 10 selected catchments.

Scores for both HEPS are grouped in Table 3. It is quite obvious that the 800-member HEPS provides a better performance than the 16-member HEPS; however, the implementation of such a system is computationally demanding. Thus, the work developed in this article seeks to evaluate the most representative members of the 800 member HEPS without sacrificing the performance and reliability of the forecast. With this goal, the selection of members is evaluated with a simple machine learning technique called backward greedy selection, which is explained next.

HESSD
8, 2739–2782, 2011

4 Backward greedy selection technique

The evaluation of multiple models for simulation or prediction of an event, and to further select those which together enhance or simplify a condition for adjustment, is known as an overproduce and select precess. Machine Learning tasks similar to this (feature selection or subset selection) can be performed with simple methods such as forward or backward greedy selection or with more sophisticated methods like genetic algorithms. In feature selection, the simplest approach is based on a deterministic selection that can run in both directions: from zero members and gradually increasing the number of members (forward selection) or starting with all members and reducing the number of members (backward selection). In the context of probabilistic forecast, the forward selection technique loses applicability since the consistency of the scores depends on a minimum number (nmin) of members to evaluate.

At some point, the Backward Greedy Selection may lose its ability to generalize the results, a problem known as overfitting. A procedure circumventing this problem divides the available information into three subsets: training (χ_t), validation (χ_v) and test or publication set (χ_p). The training subset is employed so as to look for a solution (hypothesis evaluation), the generalization capability is evaluated on the validation subset, and the test subset is used for calculating the expected error in examples never presented to the training-validation process (Alpaydin, 2010).

The members' elimination mechanism begins with all members (d) and removes them one by one, at each step removing the one that decreases the error the most (or increases it the least). The removal mechanism is as follows:

1. It begins with a subdivision of the dataset (χ) in training (χ_t), validation (χ_v) and test set (χ_p).

2. The reference set $\mathbf{G}^d$, containing all the original d members, is presented.

$$\mathbf{G}^d = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_d\}$$

HESSD

8, 2739–2782, 2011

3. For iter $= d-1, d-2, \ldots,$ nmin
   The member $\boldsymbol{y}_j$ that causes the least error in the training set χ_t is chosen.

   $$\boldsymbol{y_j} = \mathrm{argmin}_{\boldsymbol{y}_j \in \mathbf{G}^{\mathrm{iter}+1}} E\left(\mathbf{G}^{\mathrm{iter}+1} \backslash \{\boldsymbol{y}_j\} | \chi_\mathrm{t}\right)$$

   The reference set is then updated by removing $\boldsymbol{y}_j$ member in **G**.

   $$\mathbf{G}^{\mathrm{iter}} = \mathbf{G}^{\mathrm{iter}+1} \backslash \boldsymbol{y}_j$$

4. At this point, the error E in the validation set χ_v, excluding the $\boldsymbol{y}_j$ member, is evaluated.

   $$E_v^t = E\left(\mathbf{G}^t | \chi_\mathrm{v}\right)$$

5. The subset $\mathbf{G}^{\mathrm{nmin}}$ of the selected members is achieved, then the whole selection process is analyzed on the training and validation results.

In the general framework for the selection of members, the minimum number of members (nmin) must be such as to ensure the consistency of the calculations of individual scores. Here the minimum is taken as 30 members.

Backward Greedy Selection is a local search procedure that does not guarantee finding the optimal subset. For example, $\boldsymbol{y}_x$ and $\boldsymbol{y}_p$ by themselves may not be pertinent but together they may decrease the error substantially. But, because the algorithm is greedy and adds attributes one by one, it may not be able to detect this. Here, the Backward Greedy Selection is executed with a resampling technique based on k-fold cross-validation. However, some considerations are made to take into account the influence of temporal correlation in the proper formation and interpretation of the subsets used in the evaluation of the selection of members.

5 Resampling and results combination method

Given the high degree of linear correlation exhibited in the first lags of the correlogram of the series at hand (Table 4), the choice of the training and validation data should be directed in order to temporarily avoiding near data to form the two subsets. For example, suppose that the linear correlation between o^t and o^{t+1} is equal to 0.8 and that the selection of members has been trained in o^t and validated in o^{t+1}. The validation could consequently be highly contaminated by the effect of the correlation between data. Correlation contamination is avoided by forming training and validation subsets from groups of 10 consecutive data (blocks) rather than from individual data. It is important to note that contrarily to standard hydrology applications, the order of the events is of no importance.

The proposed process of selecting data for training, validation, and test follows k-fold cross-validation. Here, the dataset is divided into 5 equal-sized parts in order to create 5 experiments. In each experiment, a part is kept out for testing, while the remaining four parts are randomly combined to form training and validation subsets. The detailed process develops in two steps:

Step 1: Data and test set configuration. The test set is set up from simple cut-offs to "guarantee" statistical independence with the training-validation process. To build the test set, the series is subdivided into five folds, each of which corresponds to the test set of each experiment. For example, if N denotes the length of the series, the test set of the first experiment corresponds to the first fold ($i = 1$ to $\lfloor N/5 \rfloor$), similarly the test set of the fifth experiment will be the last fold ($i = \lceil 4N/5 \rceil$ to N). Thus, strong linear correlation between training-validation and the test data is limited only to the values situated near the cut-off line.

Step 2: Blocks' selection of the training and validation sets. The remaining 4 parts are grouped into K blocks of consecutive pairs of observations-ensemble forecast, then randomly choosing 75% of the blocks for the training set and the remaining 25% sets for the validation set.

HESSD
8, 2739–2782, 2011

However, the variability of each experiment (xp), given by the cross-validation technique, increases the probability of reaching different member selections. So, it is necessary to determine an integration mechanism for a global solution for each catchment. Here, the importance of each member $\boldsymbol{y}_i$ within the ensemble is then assumed as being directly proportional to the iteration number at which it was eliminated during the selection process in each experiment ($\mathrm{iter}_{\mathrm{xp}}^{\boldsymbol{y}_i}$). The combined ranking is thus the mean rank of elimination as defined in Eq. (8), For example, if the rank of elimination of member $\mathbf{y}_i$ is 50, 60, 200, 10 and 150 in the five experiments, then the mean rank of elimination is equal to 94.

$$\overline{R}(\boldsymbol{y}_i) = \frac{1}{5}\sum_{\mathrm{xp}=1}^{5} \mathrm{iter}_{\mathrm{xp}}^{\boldsymbol{y}_i} \tag{8}$$

Finally, the final selection (s) of the nm best members corresponds to the members which have the highest mean rank of elimination (Eq. 9).

$$s = \{\overline{R}_p, y_p\}_{p=1}^{nm}, \overline{R}_i \geq \overline{R}_j \text{ where } 1 \leq i \leq j \leq d \tag{9}$$

It should be noted that another possibility to integrating the results could have been based on the frequency of selection of each member of the ensemble, and later to elect the members with the highest frequency, but as this integration leads to low performance, these results are omitted from this article.

Finally, the mean rank selection was evaluated on the complete series, as a performance relative indicator. The word relative stresses the fact that the process of combining results involves a high degree of integration of the entire series in the training-validation step in the selection of members. This type of performance evaluation in Machine Learning is known as an optimistic estimate since the algorithm, in this case the selection of members, is tested with examples that could be part of the training and validation sets. In this regard the rigorous evaluation of an algorithm or a model must be based on instances not used in training or validation. A detailed discussion on the degree of optimism of an estimator (fairness of the solution) and the creation

HESSD
8, 2739–2782, 2011

of training and validation sets statistically or structurally homogeneous is described in detail by Diamantidis et al. (2000).

It is important also to highlight that in order to compare the performance of different scores as objective functions in the selection of members, individual experimentation and subsequent combination of results creates a fair basis for comparison between different scores. In the companion paper (Brochero et al., 2011) more rigorous tests of the degree of efficiency of selection of members are produced, including test evaluations with data from other catchments never used in the learning process of the selection algorithm.

6 Results and analysis

Analysis of the median coefficient of variation (MDCV), as a measure of the diversity of the HEPS, revealed the following characteristics:

- The variability is low at least for the first three days of predictions (MDCV < 0.12), many models showing no variability (i.e. the same response for all members). As shown by Velázquez et al. (2010), part of this difficulty may be inherited from the meteorological ensembles, which are not reliable prior to about a 3-day lead time. More importantly, it is believed that not including uncertainties associated with the hydrological initial conditions at the onset of the forecasts also takes its toll on reliability, at least for the first few time steps of the hydrological predictions, i.e. until the mean characteristic response time scale of the studied catchments (3.2 days) is reached.

- As for the incremental variability, it depends on the forecast horizon. MDCV for 4- to 9-day predictions reached between 0.2 and 0.6, respectively.

Consequently, the results presented in this paper are strictly based on the 9-day forecast horizon. This decision is justified on the variability within the ensemble forecasts as well as on the fact that the selection of members as a method of simplifying HEPS

HESSD
8, 2739–2782, 2011

should be unique regardless of the forecast horizon. The companion paper (Brochero et al., 2011) assesses the transferability of the 9-day member selection to other forecast horizons.

An example of the results obtained are shown in Fig. 3, which compares the 30-member and the 800-member results for the M0421510 catchment, after an optimization based on the δ criterion. In general the 30-member scores are better or as good as the reference set. The selected members make use of 13 of the 16 available lumped models, which is an interesting combination of models and meteorological members, especially taking into account the much poorer performance of the 16-member multi-model approach driven by the deterministic prediction (Table 3).

Specifically, Fig. 3a shows that the 30-member CRPS equals the reference value. Also, taking into account that the CRPS generalizes the mean absolute error (MAE) for a point forecast (Gneiting and Raftery, 2007), it is important to stress that the CRPS values are always lower than the MAE values, when the deterministic counterpart was taken as the mean of each daily ensemble, in agreement with results obtained by other authors (Boucher et al., 2009). Another remarkable feature of CRPS is its direct relationship with the flow magnitude; the shapes of the CRPS and of the hydrograph are similar. A direct strategy of optimization could then focus on the maximum values from the cost function. The selection not only preserves the mean CRPS (0.16) but also the structure of the CRPS series.

Figure 3b shows that the 30-member 5% trimmed mean ignorance score (−1.01) has also improved over the initial value (−0.99). Regarding the time structure of the IGNS, it is observed that both the 30-member and 800-member values surpass 6 for certain peak flows (as $6 = \log_2 0.015$).

With regard to the reliability diagram, Fig. 3c shows a considerable agreement improvement (1.09e-3) over the initial value (1.74e-3). This gain in reliability may be traced back to the optimization criterion used: the δ ratio, which is entirely based on the integration of the whole range in terms of corresponding verifications (observations). Similarly, Fig. 3d reveals that the rank histograms have a nearly uniform distribution,

HESSD
8, 2739–2782, 2011

even if the first rank reflects a slight bias. Those imperfections demonstrate the difficulty inherent in minimizing the δ ratio.

Figure 3e draws the occurrence of each lumped model from the 30-member ensemble. A wide selection of models alone could justify the multi-model approach advocated here. Results show that 13 models out of 16 were selected in this case, and that no models were selected more than 7 times. Knowing that these models are not of equal quality with regards to RMSE performance, for instance, this suggests that the selection favored a diversity of errors.

At the end of the selection process, the median coefficients of variation (MDCV) has slightly decreased, from 0.37 to 0.35. This confirms that optimization with the δ criterion seeks diversity of the ensemble forecasts in the correct way, not necessarily maximizing the MDCV. This aspect can also be seen on Fig. 4, which shows that the minimum and maximum predicted discharge are not necessarily part of the group members selected (see Fig. 2g for the reference of the whole spectrum of responses of the 800 members). Note that the lower and upper limits on the spectrum of response selection does not correspond to a particular member, but for each time step such limits are evaluated within the chosen set.

Regarding the participation of members of the ECMWF MEPS in the 30-member HEPS, the histogram in Fig. 4 shows that for this particular case 73% of the selected members were drawn from the top half of the MEPS (entries 25 to 50); members 25, 35, 44 and 46 were picked more than once. One should not interpret these findings as a lack of equiprobability, notably because the selection is influenced by various hydrological structures that may distort their probabilistic response. However, such behavior tends to disappear for a larger number of members.

Results presented up to this point concerned δ ratio optimization for a specific catchment. Table 5 summarized results for more catchments and optimization criteria. The 30-member comparison is based on a normalized sum (NS) with unit weights (Eq. 7) so that the normalized sum corresponding to the 800-member HEPS is equal to 5. In this way, NS lower than 5 indicates an overall improved performance. Performance for all criteria are also given in Table 5 for completeness, and the best optimization criterion for each catchment is identified in bold letters.

Overall, the combined criterion (CC) offers an effective and direct rule, finding balance between features offered by each of the criteria. However, it is important to point out the two cases for which the δ criterion provides a slightly better optimum. This reflects the limitations of the Backward Greedy Selection technique[2], because if the objective function (CC) is equal to the criterion used to compare results obtained with different objectives, the CC criterion should obviously always find the best solution within the vision of a global optimization tool.

The δ ratio criterion, based on a rank histogram which is the most common approach for evaluating whether a collection of ensemble forecasts for a scalar predictand satisfies the consistency condition (Wilks, 2005), comes to a close second. It led to the best performance for two catchments and to the second best performance for seven other catchments. This is particularly interesting considering the simplicity of this approach with respect to the combined approach. In addition, the δ criterion favored the highest average participation of hydrological models.

The CRPS and IGNS led to a poorer selection, to the point that they were not considered further after experimenting with the first four catchments allowing an economy in computational time[3]. The CRPS showed low variability, so it is not very sensitive to changes in the selection of members. The IGNS demonstrated a negative relationship with reliability, leading to poor performance in terms of reliability diagram (RD) and δ ratio. Both criteria led to the selection of a lower number of different hydrological models. They are also correlated, optimizing one criterion often favoring the improvement of the other one.

[2] A greedy search procedure is an heuristic based on the results of local explorations, with no guarantee of finding a global optimal value.

[3] The Backward selection method may be costly because to decrease the dimensions from d to k, it is necessary to train and test the system $d+(d-1)+(d-2)+\cdots+(d-k)$ times, which has a complexity of $\mathcal{O}(d^2)$.

Specifically the behavior of the optimization of each score could be also described from the following relationships observed in Table 5:

- Optimization based on CRPS is detrimental to the reliability, especially in terms of the RD and the δ ratio. For example, it had the effect of increasing RD by a factor of 10, for catchment Q2593310. The CRPS also decreases diversity of the members (MDCV), except for catchment B31150020 where that value remained stable.

- The combined criterion (CC) leads to stable CRPS values. The most remarkable gains come in terms of RD, as provided in the weights definition of the Eq. (7). With reference to δ ratio, evaluations reveal the difficulty in maintaining the stability of this criterion, but the differences between the selection and the reference set are not pronounced. As for the MDCV, the diversity is in most cases maintained or improved. The IGNS performance is often slightly decreased. In conclusion, the CC promotes overall good performance, increasing the reliability of the system (decrease of the RD score) and ensuring the stability or slight declines in the other scores.

- Selection based on the RD score is detrimental to the CRPS. As for reliability, there are some cases for which the error increases. This condition is surprising given that the combined criterion always achieved reductions of this error, but it could not last under the assumption of a greater weight of this score in the combination because the relationship is completely constant, which highlights the interaction between the scores as a mechanism implicit in the reduction of error reliability (RD). The δ ratio is never improved, while diversity (MDCV) is lost except in three cases (B3150020, Q2593310 and U2542010) where interestingly the MDCV increased (theoretically consistent effect). Finally, the IGNS shows a negative trend to the minimization of the RD.

- By definition, the δ ratio focuses exclusively on the reliability and the consistency of the ensemble. In fact, it leads to better reliability performance in terms of RD,

HESSD
8, 2739–2782, 2011

than when the selection is optimized with RD itself. The δ ratio also preserves the resolution of the forecast, as shown by the CRPS and IGNS results. All of this is accompanied by a slight loss in performance in terms of δ ratio, which can be explained by the direct relationship of this score with the number of members. However, this dependency rather than becoming an obstacle in the selection stands as a logical consequence of the system, since statistically a better performance is expected from a system that combines a larger number of models (Alpaydin, 2010). Finally, with respect to MDCV, it is shown once again that diversity, hypothetically represented by MDCV, fluctuates between values that indicate the extent to which such diversity needs to be maintained in the ensemble.

- When the selection process focuses on the maximization of MDCV, the relationship with CRPS, the IGNS and δ ratio is always negative. However, there are four cases in which the reliability is improved by increasing the diversity index, from which it follows that while reliability improves the resolution drops.

In summary, the interaction of different scores, as seen from the 30-member selection, shows that the optimization focused on scores that define mainly the resolution of the ensemble (CRPS, IGNS) has a negative impact on the reliability, consistency and ensemble diversity. It also reveals that if the selection is based only on a reliability view, the ensemble loses resolution and consistency. Maximization of the MDCV is in general detrimental to the other criteria, but sometimes improves reliability, a condition that can easily be understood from a theoretical point of view. The δ ratio improves reliability while maintaining resolution. The combined approach stands out as the most balanced criterion.

The above analysis focused exclusively on 30-member selections. However, a global vision demands the analysis of the evolution of the scores as the number of members is reduced. Such an analysis is specific to each catchment, so as an example, Fig. 5 shows evolution diagrams of the various scores as a function of the number of members, for catchment A7930610.

HESSD
8, 2739–2782, 2011

A gain index was used (Eq. 10) to evaluate the scores obtained in the selected ensemble (se) in reference to the initial ensemble forecasts of 800 members (ie). A positive index indicates superior performance of the selected set. The absolute value in the denominator is needed to assess the performance of IGNS, which can take positive and negative values. The normalized sum of scores with unit weights (Eq. 7) is again used for assessing the joint evolution of all scores.

$$\text{Gain}(\%) = 100 \times \frac{\text{Score}_{\text{ie}} - \text{Score}_{\text{se}}}{|\text{Score}_{\text{ie}}|} \tag{10}$$

Figure 5a,e clearly shows that an optimization based on resolution of the system (CRPS or IGNS) is detrimental to the reliability. However, this effect is less pronounced in the IGNS optimization when the number of members is kept above 100. Figure 5 also highlights the correspondence of CRPS and IGNS throughout the selection process, when the optimization is focused on one or the other.

RD optimization (Fig. 5b) is surprisingly unfavorable to the δ ratio (negative gain index), which is related to the indifference of the RD with respect to the location of the observation within the ensemble, while this location analysis creates a solid indicator of the system consistency. Likewise, it is remarkable that the normalized sum for RD is equal to 4.96 when the number of members is equal to 100. This is strictly because loss in consistency (negative gain index in the δ ratio of 50%) and resolution (IGNS equivalent to losses of 10%) is balanced by a positive gain of about 50% in RD.

The δ ratio (Fig. 5c) displays a gradual overall improvement of individual scores to a selection of about 70 members, when the various scores show a tendency to decrease in performance. At this point it is important to note that the normalized sum (NS) reached 4.53.

Figure 5d shows that criteria focusing on the resolution and the consistency have a negative relationship with the maximization of the diversity (MDCV), overall gains (NS) are achieved only when the number of members is greater than 400.

The combined criterion (Fig. 5f) improves collective performance of all scores in the selection, with an optimal number of members of again 70 for this catchment, coinciding

HESSD
8, 2739–2782, 2011

with the interaction shown in the minimization of the δ ratio (Fig. 5c). Scores tend to lose quality afterward.

Taking into account the detailed analysis for the 30-member selections and the global analysis performed for each of the catchments, the combined criterion leads to the best Backward Greedy selection. However, the issue of the optimal number of members remains somehow blurred. So, Fig. 6 revisits that question in terms of a gain index defined in Eq. (11). The gain index is constructed around the 800-member NS value (5), thus a negative value indicates that the selection reduces the quality of the reference score combination.

$$\text{Gain}(\%) = 100 \times \left(\frac{5}{\text{NS}} - 1\right) \tag{11}$$

Figure 6 emphasizes that the 30-member selection always displays a positive gain index, however 100 members would be a more optimal choice for most catchments. However, one should keep in mind that the optimal number of members should be based on an individual analysis of the different scores.

Table 6 groups the 100-member scores following optimization with the combined score and the δ ratio, the two best one. These values confirm the superiority of the combined score, leading to the smallest NS for all catchments, mainly because of the great influence on minimizing reliability. This also maximizes MDCV to such an extent that it allows a proper balance between reliability, resolution, and consistency. It is also remarkable that for 8 catchments out of 10, the δ ratio is minimized even more than when the optimization is focused on the δ ratio itself. Optimization based on the δ ratio also improved scores over the initial 800-member values (NS < 5) for 9 catchments out of 10. This single criterion is thus also very appealing, especially because it makes use of all 16 models in its selection.

Additionally, the δ ratio can be highlighted as a simple optimization criterion, which for 100% of the catchments, makes use of the participation of all hydrological models in the formation of the solution, which is not the case for the combined score optimization.

HESSD
8, 2739–2782, 2011

7 Conclusions

Results presented here support the idea that selecting HEPS members is viable. It is in general even possible to expect a better balance of scores in the subset of selected members than in the original much larger ensemble, based on standard scores such as the CRPS, the IGNS, the reliability diagram, and the δ ratio. The diversity, sought in the multi-model approach with MEPS, may also be maintained in the final selection.

The simplification of the HEPS can be addressed from two points of view: as a function of the maximum simplification of the number of members or as a function of the maximization of the balance of the scores. Simplification of the number of members involves the definition of a limit ensuring statistical consistency of the scores assessed. A trade-off exists between the number of members and the level of improvement in scores. For example, in this study, the balance of scores is achieved with about 100 members which maximized the qualities of the system: reliability, consistency, resolution, and diversity. This corresponds to a 87.5% compression level. The ultimate level of compression is in fact a compromise between the gain index and the complexity of the system. The ultimate decision should be established according to the requirements and the operational capacity of the probabilistic forecast system.

The evaluation of five individual scores as criteria for optimizing the selection process revealed the complexity of the relationship between them. In many situations, improving one score is achieved at the expense of another score. Therefore, the design of a combined criterion (CC) led to an important methodological improvement that integrates many characteristics of each score. The δ ratio is the best single optimization criterion, not very distant to the achievements of the CC criterion.

The CRPS is often the primary score used for evaluating HEPS performance. However, results here indicate that it is not a good choice for member selection. In fact, it was often possible to preserve or minimize the CRPS using others objective criteria. Likewise, the centralization of the selection process in the IGNS heavily penalized the reliability and the consistency of the system. With respect to the MDCV, the

HESSD
8, 2739–2782, 2011

uncontrolled maximization of this parameter, which describes diversity, leads to a deterioration of the other sought qualities of the system. There exists a threshold beyond which the system abruptly loses reliability, resolution, and consistency. On the other hand, experiments showed that both the δ ratio and the CC criterion improve the balance of the scores.

The resampling methodology presented in this study is adapted to restrictions imposed by the short-length of the series (500 observations). However, it is believed that it is widely applicable to any length of condition series.

Finally, the encouraging results of this study will lead to an interest in testing other global search (non-greedy) tools such as Genetic Algorithms.

Appendix A

Notations

t	Time-step
N	Number of pairs observations-forecasts
d	Total number of members in the forecast ensembles
M	Total number of m intervals to analyze the reliability diagram
c	Identification of the rank or class to analyze the uniformity in the rank histogram
o^t	Observed flow at the time t
$\boldsymbol{y}^t$	Ensemble flow forecast at the time t
y_i^t	ith flow forecast member in $\boldsymbol{y}^t$
$\mathbf{Y}$	Ensemble flow forecast from $t = 1$ to N
$\boldsymbol{o}$	Observations vector from $t = 1$ to N
F	Cumulative distribution function
f	Probability density function

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

$\bar{o}_m$	Conditional probability of the event as a function of the interval I_m assigned to the forecast $m \to P(o^t \mid I_m)$
r^t	Binary indicator, 1 if the event occurs for the tth forecast-event pair, 0 if it does not
S_c	Absolute frequency or occurrences in each rank c in the rank histogram
$\mathrm{med}_{t=1}^{N}$	Median value evaluated from $t = 1$ to N
μ_t	Mean ensemble flow forecasts at the time t
σ_t^2	Variance ensemble flow forecasts at the time t
χ_t	Training set
χ_v	Validation set
χ_p	Test or publication set
$\{x^t\}_{t=1}^{N}$	Set of x with index t ranging from 1 to N
$\mathrm{argmin}_\theta g(x \mid \theta)$	The argument θ for which g has its minimum value
$E(\theta \mid \chi)$	Error function with parameters θ on the sample χ
w_cp	Weights of the components of the combined criterion (CC)
$\mathrm{iter}_\mathrm{xp}^{\boldsymbol{y}_i}$	Iteration number at which was eliminated the $\boldsymbol{y}_i$ member during the selection process in the xp experiment
$\overline{R}(\boldsymbol{y}_i)$	Mean rank of elimination of the $\boldsymbol{y}_i$ member
s	Final selection of the nm best members in the individual selection process

Acknowledgements. The authors acknowledge NSERC and the Institut EDS for financial support, CEMAGREF and ECWMF for making the database available, and the CVSL for the computational resources. They also thanks Annette Schwerdtfeger for proofreading this manuscript.

References

Alpaydin, E.: Introduction to Machine Learning. Adaptive Computation and Machine Learning, 2 edn., Cambridge, The MIT Press, 2010. 2740, 2742, 2751, 2760

Boucher, M.-A., Perreault, L., and Anctil, F.: Tools for the assessment of hydrological ensemble forecasts obtained by neural networks, J. Hydroinform., 11, 297–307, doi:10.2166/hydro.2009.037, 2009. 2743, 2756

Boucher, M.-A., Laliberté, J.-P., and Anctil, F.: An experiment on the evolution of an ensemble of neural networks for streamflow forecasting, Hydrol. Earth Syst. Sci., 14, 603–612, doi:10.5194/hess-14-603-2010, 2010. 2744

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., and Worley, S.: The THORPEX interactive grand global ensemble, B. Am. Meteorol. Soc., 91, 1059–1072, doi:10.1175/2010BAMS2853.1, 2010. 2741

Brochero, D., Anctil, F., and Gagné, C.: Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 2: Generalization in time and space Hydrol. Earth Syst. Sci. Discuss., 8, 2783–2820, doi:10.5194/hessd-8-2783-2011, 2011. 2755, 2756

Buizza, R.: EPS skill improvements between 1994 and 2005, ECMWF Newsl., 104, 10–14, 2005. 2749

Buizza, R. and Palmer, T. N.: The singular-vector structure of the atmospheric global circulation, J. Atmos. Sci., 52, 1434–1456, 1995. 2749

Chiew, F. H. S., Peel, M. C., and Western, A. W.: Application and testing of the simple rainfall-runoff model SIMHYD, in: Mathematical models of small watershed hydrology and applications, edited by: Singh, V. P., Frevert, D., Water Resources Publications, Littleton, CO, 335–367, 2002. 2750

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009. 2741, 2742, 2743

Diamantidis, N., Karlis, D., and Giakoumakis, E.: Unsupervised stratification of cross-validation for accuracy estimation, Artif. Intell., 116, 1–16, doi:10.1016/S0004-3702(99)00094-6, 2000. 2755

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Adv. Water Resour., 30, 1371–1386, doi:10.1016/j.advwatres.2006.11.014, 2007. 2740

Ebert, C., Bárdossy, A., and Bliefernicht, J.: Selecting members of an EPS for flood forecasting systems by using atmospheric circulation patterns, Geophysical Research Abstracts, European Geosciences Union, Vienna, Austria, 9, 2007. 2742

Edijatno, De Oliveira Nascimiento, N., Yang, X., Makklouf, Z., and Michel, C.: GR3J: a daily watershed model with three free parameters/GR3J: un modèle pluie-débit journalier à trois paramètres testé en France, Hydrolog. Sci. J., 44, 263–277, 1999. 2750

Fortin, V. and Turcotte, R.: Le modèle hydrologique MOHYSE: notes de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Université du Québec à Montréal, 2006. 2750

Garçon, R.: Modèle global pluie-débit pour la prévision et la prédétermination des crues, La Houille Blanche, 0, 88–95, doi:10.1051/lhb/1999088, 1999. 2750

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, The Distributed Model Intercomparison Project (DMIP), J. Hydrol., 298, 222–241, doi:10.1016/j.jhydrol.2004.03.037, 2004. 2740

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, doi:10.1198/016214506000001437, 2007. 2744, 2756

Gouweleeuw, B. T., Thielen, J., Franchello, G., De Roo, A. P. J., and Buizza, R.: Flood forecasting using medium-range probabilistic weather prediction, Hydrol. Earth Syst. Sci., 9, 365–380, doi:10.5194/hess-9-365-2005, 2005. 2749

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, 2000. 2744

Jeong, D. I. and Kim, Y.-O.: Combining single-value streamflow forecasts – a review and guidelines for selecting techniques, J. Hydrol., 377, 284–299, doi:10.1016/j.jhydrol.2009.08.028, 2009. 2741

Kottegoda, N. T. and Rosso, R.: Applied Statistics for Civil and Environmental Engineers, electronic version, Wiley, Chichester, 2009. 2747

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F., and Buizza, R.: A strategy for high-resolution ensemble prediction, II: Limited-area experiments in four Alpine flood events, Q. J. Roy. Meteor. Soc., 127, 2095–2115, doi:10.1002/qj.49712757613, 2001. 2742

HESSD
8, 2739–2782, 2011

Mathevet, T.: Quels modèles pluie-débit globaux pour le pas de temps horaire? Développement empirique et comparaison de modèles sur un large échantillon de bassins versants, Ph.D. thesis, ENGREF (Paris), Cemagref (Antony), France, 463 pp., 2005. 2750

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Q. J. Roy. Meteor. Soc., 122, 73–119, doi:10.1002/qj.49712252905, 1996. 2749

Perrin, C.: Vers une amélioration d'un modèle global pluie-débit au travers d'une approche comparative, Ph.D. thesis, Institut National Polytechnique de Grenoble (INPG), France, 2000. 2750

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003. 2750

Renner, M., Werner, M., Rademacher, S., and Sprokkereef, E.: Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376, 463–475, doi:10.1016/j.jhydrol.2009.07.059, 2009. 2742

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725–737, doi:10.5194/hess-11-725-2007, 2007. 2741

Roulston, M. S. and Smith, L. A.: Evaluating probabilistic forecasts using information theory, Mon. Weather Rev., 130, 1653–1660, 2002. 2744

Roy, A., Royer, A., and Turcotte, R.: Improvement of springtime streamflow simulations in a boreal environment by incorporating snow-covered area derived from remote sensing data, J. Hydrol., 390, 35–44, doi:10.1016/j.jhydrol.2010.06.027, 2010. 2750

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: HEPEX: the hydrological ensemble prediction experiment, B. Am. Meteorol. Soc., 88, 1541–1547, 2007. 2741

Székely, G. J.: E-Statistics: The Energy of Statistical Samples, Tech. Rep. 2003-16, Department of Mathematics and Statistics, Bowling Green State University, Ohio, 2003. 2743

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Workshop on predictability, edited by: for Medium-Range Weather Forecasts, E. C., 1–25, Shinfield Park, Reading, Berkshire RG2 9AX, UK, 1997. 2746, 2747

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, 2001. 2747

Velázquez, J. A., Anctil, F., Ramos, M.-H., and Perrin, C.: Comparison of the performance and reliability of 18 lumped hydrological models driven by ECMWF rainfall ensemble forecasts: a case study on 29 French catchments, EGU General Assembly, Austria, 12, 4929, 2010.

2741, 2748, 2749, 2750, 2755
Vrugt, J., Diks, C., and Clark, M.: Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling, Environ. Fluid Mech., 8, 579–595, doi:10.1007/s10652-008-9106-3, 2008. 2743
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, vol. 91, 2 edn., Academic Press, Burlington, MA, London, 2005. 2745, 2746, 2747, 2758
Williams, G. P.: Chaos Theory Tamed, 1st edn., National Academies Press, Washington D.C., 1997. 2744
Xuan, Y., Cluckie, I. D., and Wang, Y.: Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction, Hydrol. Earth Syst. Sci., 13, 293–303, doi:10.5194/hess-13-293-2009, 2009. 2742
Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774, doi:10.1016/j.advwatres.2007.01.005, 2007. 2750

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

HESSD

8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

Table 1. Main characteristics of the studied catchments based on a 36 year length of the series (1970–2006). P: diary precipitation, ET: diary potential evapotranspiration, Q: diary observed flow.

Catchment codes	Area (km^2)	Altitude (m)	Pmean	Pmax	ETmean (mm)	ETmax	Qmean	Qmax
A7930610	9837	155	2.77	56.99	1.80	4.07	1.20	18.22
B2130010	2290	227	2.57	55.11	1.80	4.05	0.89	17.02
B3150020	3904	162	2.57	56.63	1.80	4.04	1.08	12.79
H3621010	3900	48	1.97	51.29	1.96	4.26	0.45	6.56
J8502310	2465	4	2.35	49.34	1.90	3.94	0.81	15.18
K7312610	1712	85	2.13	45.00	2.01	4.30	0.67	15.14
M0421510	1890	56	2.04	39.42	1.90	4.11	0.61	7.13
O3401010	2170	349	3.18	182.83	1.81	4.01	1.88	81.62
Q2593310	2500	17	2.52	53.63	2.25	4.52	0.73	12.72
U2542010	4970	201	3.63	59.04	1.76	3.97	1.88	22.08

Table 2. Hydrological models. The number of parameters used in each model are shown in brackets.

Hydrological models	Base model (parameters)	Origin	Main objective
HM01	CEQUEAU (9)	France	Flood forecasting
HM02	GR3J (3)	France	Application in ungauged basins
HM03	HBV (9)	Sweden	Flood forecasting, Nordic countries
HM04	IHACRES (6)	Australia	Regionalization, water quality
HM05	MORDOR (6)	France	Monitoring of water resources
HM06	SAC-SMA (13)	USA	Flood forecasting
HM07	SMAR (9)	Ireland	Flood forecasting, regionalization
HM08	TOPMODEL (8)	UK	Many applications, SIG coupling
HM09	CREC (8)	France	Flood forecasting
HM10	GR4J (4)	France	Application in ungauged basins
HM11	SIMHYD (8)	Australia	Flood forecasting
HM12	MOHYSE (7)	Canada	Identification of the components of the process
HM13	PDM (8)	UK, Brazil	Flood forecasting
HM14	PDS (5)	USA, UK	Ensemble predictions in ungauged basins
HM15	TANK (10)	Japan	Flood forecasting
HM16	WAGENINGEN (8)	Netherlands	Identification of the components of the process

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

Table 3. Performance for the deterministic (16 members) and probabilistic (800 members) HEPS for a 9-day forecast time horizon.

Catchment codes	HEPS	Scores				MDCV function
		CRPS	IGNS	RD(e-3)	δ	
A7930610	16	0.338	4.51	93.95	42.5	0.18
	800	0.263	0.44	5.06	3.3	0.41
B2130010	16	0.282	1.05	39.29	23.3	0.32
	800	0.230	−0.29	2.43	2.2	0.57
B3150020	16	0.164	0.77	39.21	21.3	0.13
	800	0.135	−0.88	4.51	2.7	0.22
H3621010	16	0.181	0.84	34.89	17.4	0.19
	800	0.161	−0.99	3.50	1.5	0.37
J8502310	16	0.184	0.69	34.49	15.8	0.20
	800	0.163	−0.98	2.16	1.6	0.37
K7312610	16	0.184	0.53	33.98	15.8	0.19
	800	0.165	−0.93	3.09	1.9	0.35
M0421510	16	0.177	0.49	27.24	13.7	0.19
	800	0.160	−0.99	1.74	1.5	0.37
O3401010	16	0.198	0.77	36.39	16.8	0.19
	800	0.169	−0.86	3.46	1.5	0.36
Q2593310	16	0.186	0.66	32.89	14.9	0.21
	800	0.163	−0.98	2.15	1.5	0.37
U2542010	16	0.390	3.29	39.73	21.0	0.19
	800	0.289	−0.36	3.39	2.6	0.35

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

Table 4. First 10 lags of the correlogram (linear autocorrelation diagram).

Lag	Catchment codes									
	A7930610	B2130010	B3150020	H3621010	J8502310	K7312610	M0421510	O3401010	Q2593310	U2542010
1	0.922	0.929	0.975	0.974	0.974	0.974	0.973	0.974	0.973	0.956
2	0.787	0.810	0.927	0.918	0.917	0.917	0.916	0.917	0.917	0.882
3	0.675	0.698	0.873	0.856	0.856	0.856	0.854	0.855	0.855	0.820
4	0.604	0.623	0.824	0.802	0.803	0.802	0.801	0.801	0.801	0.774
5	0.563	0.579	0.783	0.759	0.759	0.759	0.757	0.758	0.758	0.740
6	0.523	0.539	0.746	0.716	0.717	0.716	0.714	0.714	0.715	0.715
7	0.475	0.491	0.707	0.670	0.671	0.671	0.669	0.668	0.669	0.698
8	0.432	0.437	0.670	0.624	0.624	0.624	0.622	0.622	0.623	0.684
9	0.402	0.396	0.638	0.582	0.582	0.582	0.581	0.579	0.581	0.662
10	0.386	0.372	0.619	0.554	0.554	0.554	0.553	0.550	0.552	0.625

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

Table 5. Selection of 30 members based on different scores. NS represents the normalized sum with unit weights (Eq. 7). NHM indicates the number of hydrological models participating in the solution. Bold indicates the best results according to the NS measure.

Catchment codes	Optimization criterion	CRPS	RD(e-3)	δ	MDCV	IGNS	NS	NHM
	CRPS	0.239	6.96	4.26	0.34	0.41	5.70	8
	CC	**0.257**	**1.82**	**3.44**	**0.40**	**0.38**	**4.39**	**10**
	RD	0.263	2.80	4.60	0.40	0.49	5.02	7
A7930610	δ	0.269	5.17	3.68	0.40	0.48	5.21	13
	MDCV	0.282	11.14	5.05	0.46	0.65	6.83	7
	IGNS	0.244	9.61	4.83	0.31	0.38	6.46	6
	Ref. values (800 members)	0.263	5.06	3.26	0.41	0.44	5.00	16
	CRPS	0.208	4.00	4.54	0.49	−0.48	6.67	8
	CC	**0.234**	**1.33**	**2.60**	**0.63**	**−0.16**	**4.70**	**13**
	RD	0.230	2.46	3.92	0.53	−0.33	5.86	8
B2130010	δ	0.229	2.06	3.02	0.56	−0.27	5.24	14
	MDCV	0.243	5.24	3.69	0.61	−0.26	6.83	8
	IGNS	0.224	23.25	8.03	0.39	−0.33	16.72	7
	Ref. values (800 members)	0.230	2.43	2.24	0.57	−0.29	5.00	16
	CRPS	0.117	5.93	4.57	0.22	−0.97	5.91	7
	CC	**0.133**	**0.92**	**2.02**	**0.23**	**−0.85**	**4.01**	**10**
	RD	0.152	3.53	5.19	0.24	−0.62	6.16	8
B3150020	δ	0.130	2.95	3.28	0.23	−0.86	4.92	12
	MDCV	0.139	12.12	7.28	0.24	−0.70	8.73	7
	IGNS	0.122	17.44	7.07	0.17	−0.97	9.45	8
	Ref. values (800 members)	0.135	4.51	2.66	0.22	−0.88	5.00	16
	CRPS	0.142	21.88	5.93	0.25	−0.96	17.17	6
	CC	0.158	0.67	1.78	0.37	−0.97	4.47	9
	RD	0.171	1.68	3.06	0.38	−0.84	5.97	5
Q2593310	**δ**	**0.161**	**0.57**	**1.59**	**0.37**	**−0.98**	**4.36**	**13**
	MDCV	0.175	3.93	3.45	0.45	−0.74	7.34	5
	IGNS	0.152	32.01	12.47	0.18	−0.41	26.93	6
	Ref. values (800 members)	0.163	2.15	1.46	0.37	−0.98	5.00	16
	CC	**0.156**	**1.10**	**1.69**	**0.36**	**−0.97**	**4.46**	**11**
	RD	0.163	2.89	2.46	0.34	−1.00	5.49	7
H3621010	δ	0.160	2.44	1.87	0.36	−1.02	4.94	13
	MDCV	0.170	2.51	3.83	0.44	−0.79	6.36	6
	Ref. values (800 members)	0.161	3.50	1.49	0.37	−0.99	5.00	16

Table 5. Continued.

Catchment codes	Optimization criterion	CRPS	RD(e-3)	δ	MDCV	IGNS	NS	NHM
	CC	**0.160**	**0.49**	**2.32**	**0.39**	**−0.98**	**4.60**	**12**
	RD	0.166	2.33	3.11	0.35	−0.91	6.20	7
J8502310	δ	0.163	1.34	1.63	0.36	−0.99	4.60	13
	MDCV	0.175	1.64	2.48	0.44	−0.74	5.47	6
	Ref. values (800 members)	0.163	2.16	1.63	0.37	−0.98	5.00	16
	CC	**0.158**	**1.26**	**2.44**	**0.36**	**−0.96**	**4.57**	**9**
	RD	0.167	3.36	3.72	0.35	−0.89	6.10	7
K7312610	δ	0.163	2.10	3.34	0.33	−0.95	5.46	13
	MDCV	0.173	2.51	4.22	0.43	−0.68	6.19	6
	Ref. values (800 members)	0.165	3.09	1.86	0.35	−0.93	5.00	16
	CC	0.159	0.71	1.85	0.36	−0.99	4.54	12
	RD	0.164	2.10	2.93	0.36	−0.92	6.25	6
M0421510	$\boldsymbol{\delta}$	**0.158**	**1.09**	**1.26**	**0.35**	**−1.01**	**4.41**	**13**
	MDCV	0.170	2.63	3.33	0.44	−0.75	6.96	5
	Ref. values (800 members)	0.160	1.74	1.51	0.37	−0.99	5.00	16
	CC	**0.166**	**0.94**	**1.31**	**0.36**	**−0.87**	**4.06**	**13**
	RD	0.172	2.48	4.22	0.36	−0.67	6.57	5
O3401010	δ	0.168	1.94	1.80	0.37	−0.85	4.69	12
	MDCV	0.189	5.70	4.85	0.44	−0.51	7.93	4
	Ref. values (800 members)	0.169	3.46	1.54	0.36	−0.86	5.00	16
	CC	**0.292**	**1.16**	**2.86**	**0.37**	**−0.34**	**4.43**	**12**
	RD	0.297	2.88	4.98	0.37	−0.25	5.83	6
U2542010	δ	0.291	1.79	2.86	0.34	−0.32	4.68	15
	MDCV	0.301	3.03	3.72	0.43	−0.10	5.41	5
	Ref. values (800 members)	0.289	3.39	2.62	0.35	−0.36	5.00	16

HESSD
8, 2739–2782, 2011

Table 6. Selection of 100 members based on the combined (CC) and δ criteria. NS represents the normalized sum (Eq. 7) with unit weights. NHM indicates the number of hydrological models participating in the solution.

Catchment codes	Optimization criterion	CRPS	RD	δ	MDCV	IGNS	NS	NHM
A7930610	CC	0.257	1.76	3.01	0.43	0.33	4.17	13
	δ	0.265	3.54	2.96	0.41	0.43	4.61	16
	Ref. values (800 members)	0.263	5.06	3.26	0.41	0.44	5.00	16
B2130010	CC	0.232	1.03	2.29	0.63	−0.19	4.37	14
	δ	0.227	1.16	2.40	0.59	−0.28	4.50	16
	Ref. values (800 members)	0.230	2.43	2.24	0.57	−0.29	5.00	16
B3150020	CC	0.134	0.99	2.44	0.25	−0.83	4.16	14
	δ	0.135	2.31	2.48	0.23	−0.85	4.48	16
	Ref. values (800 members)	0.135	4.51	2.66	0.22	−0.88	5.00	16
Q2593310	CC	0.160	0.36	1.30	0.40	−0.98	3.98	16
	δ	0.159	0.63	1.43	0.36	−1.05	4.18	16
	Ref. values (800 members)	0.163	2.15	1.46	0.37	−0.98	5.00	16
H3621010	CC	0.158	0.58	1.63	0.38	−1.03	4.18	14
	δ	0.158	2.45	1.80	0.36	−1.04	4.83	16
	Ref. values (800 members)	0.161	3.50	1.49	0.37	−0.99	5.00	16
J8502310	CC	0.161	0.38	1.49	0.39	−0.98	4.05	15
	δ	0.161	1.31	1.66	0.38	−1.00	4.63	16
	Ref. values (800 members)	0.163	2.16	1.63	0.37	−0.98	5.00	16
K7312610	CC	0.162	0.59	1.65	0.39	−0.91	4.04	14
	δ	0.164	2.60	2.22	0.34	−0.95	5.01	16
	Ref. values (800 members)	0.165	3.09	1.86	0.35	−0.93	5.00	16
M0421510	CC	0.157	0.29	1.65	0.37	−1.00	4.23	15
	δ	0.158	0.79	1.25	0.36	−1.03	4.24	16
	Ref. values (800 members)	0.160	1.74	1.51	0.37	−0.99	5.00	16
O3401010	CC	0.167	0.74	1.40	0.38	−0.87	4.07	16
	δ	0.166	2.18	2.10	0.37	−0.89	4.89	16
	Ref. values (800 members)	0.169	3.46	1.54	0.36	−0.86	5.00	16
U2542010	CC	0.289	0.89	2.22	0.39	−0.38	4.07	14
	δ	0.287	1.44	2.45	0.36	−0.42	4.34	16
	Ref. values (800 members)	0.289	3.39	2.62	0.35	−0.36	5.00	16

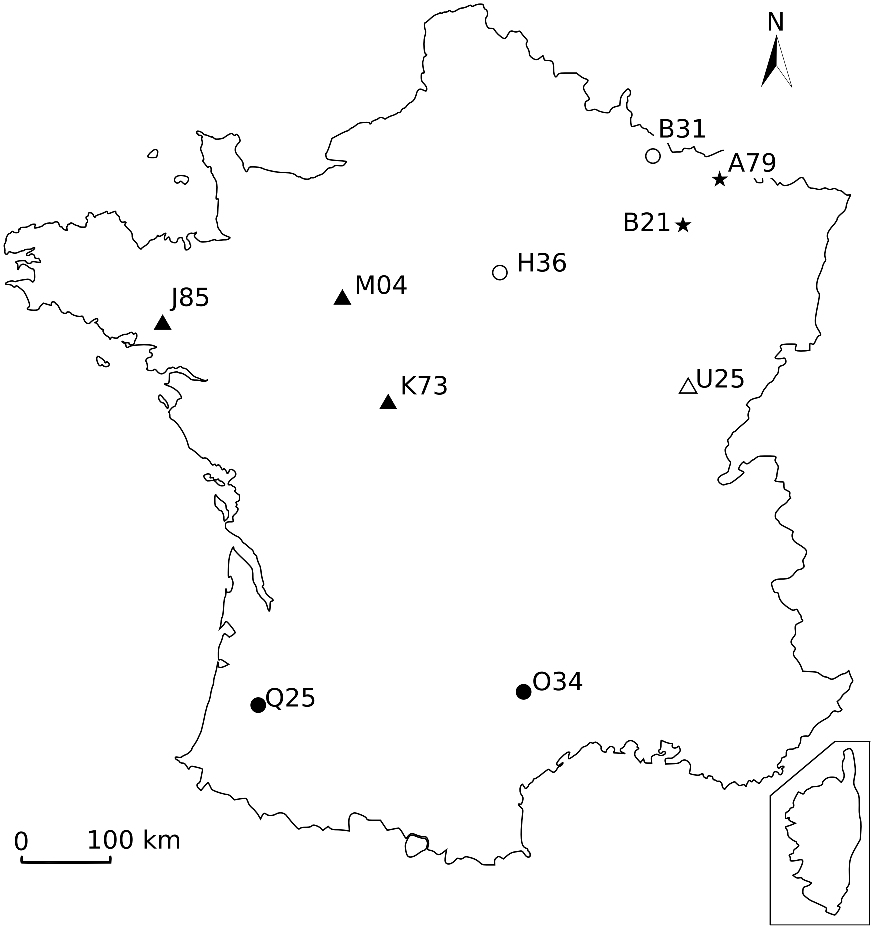

Fig. 1. Selected catchments for the first phase. Each catchment is identified with the first three digits of each code used in Table 1.

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.

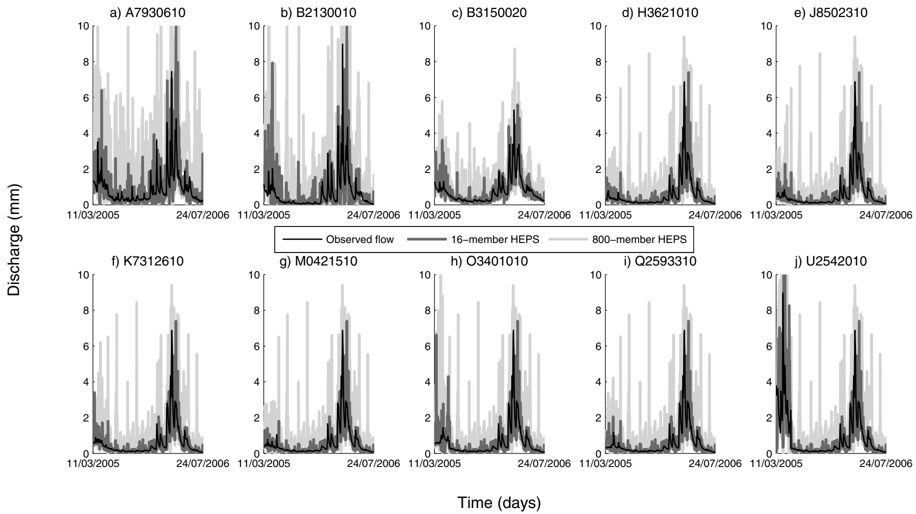

Fig. 2. HEPS results. The 16 member HEPS corresponds to the deterministic meteorological condition and the 800 member HEPS corresponds to the probabilistic meteorological condition.

HESSD
8, 2739–2782, 2011
Simplifying a hydrological ensemble prediction system – Part 1
D. Brochero et al.

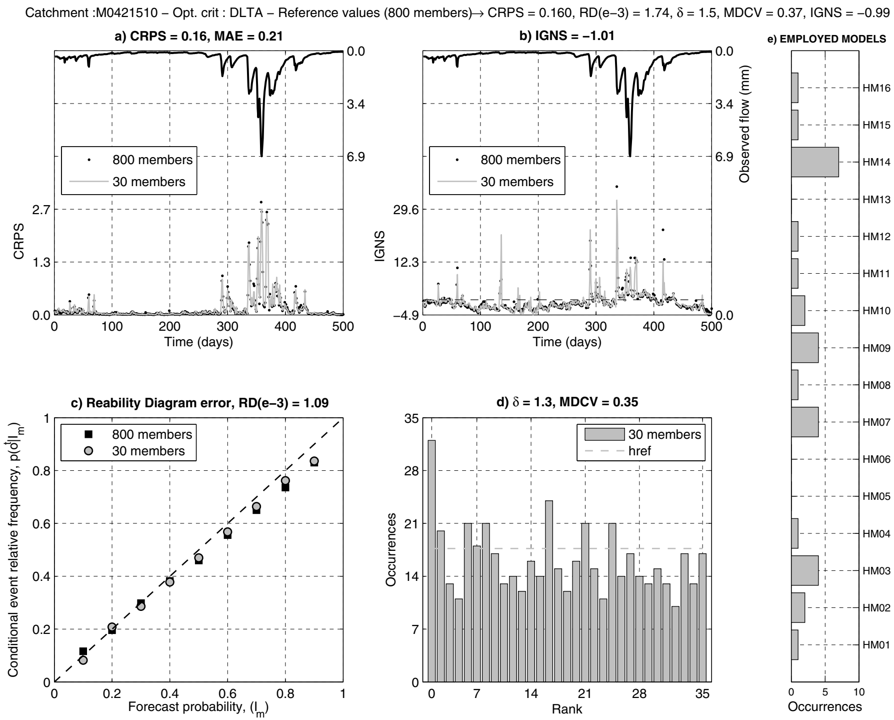

Fig. 3. Comparison between the initial ensemble (800 members) and the ensemble selected (30 members). **(a)** Figure above: observed flow; figure below: CRPS. Note the correspondence between higher observed flows and higher CRPS. **(b)** Figure above: observed flow; figure below: IGNS. Note that there is no full correspondence between the higher IGNS and higher observed flow. **(c)** Reliability diagram error (MSE based on vertical distances between the points). **(d)** Rank histogram for the 30 selected members. The horizontal dashed gray lines indicate the frequency ($N/d+1$) attained by a uniform distribution. **(e)** Occurrences of the employed models in the final solution of 30 members.

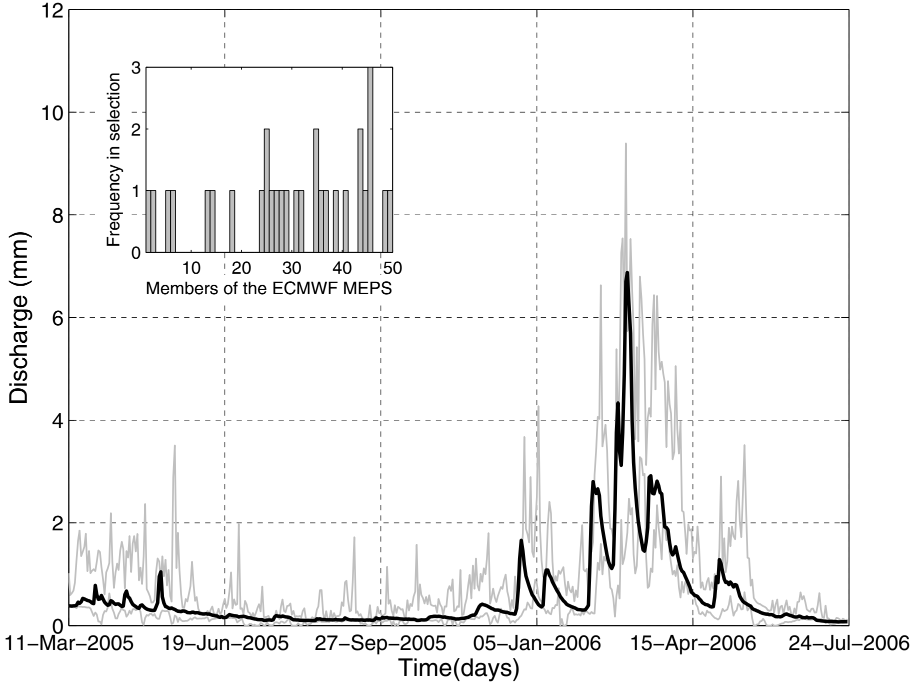

Fig. 4. The 30 members response spectrum in the selection in the catchment M0421510. The Black line indicates the observed flow; gray lines represent the minimum and maximum limits in the prediction of the selection. The histogram (top) indicates the occurrence of members of the ECMWF in the selection.

HESSD

8, 2739–2782, 2011

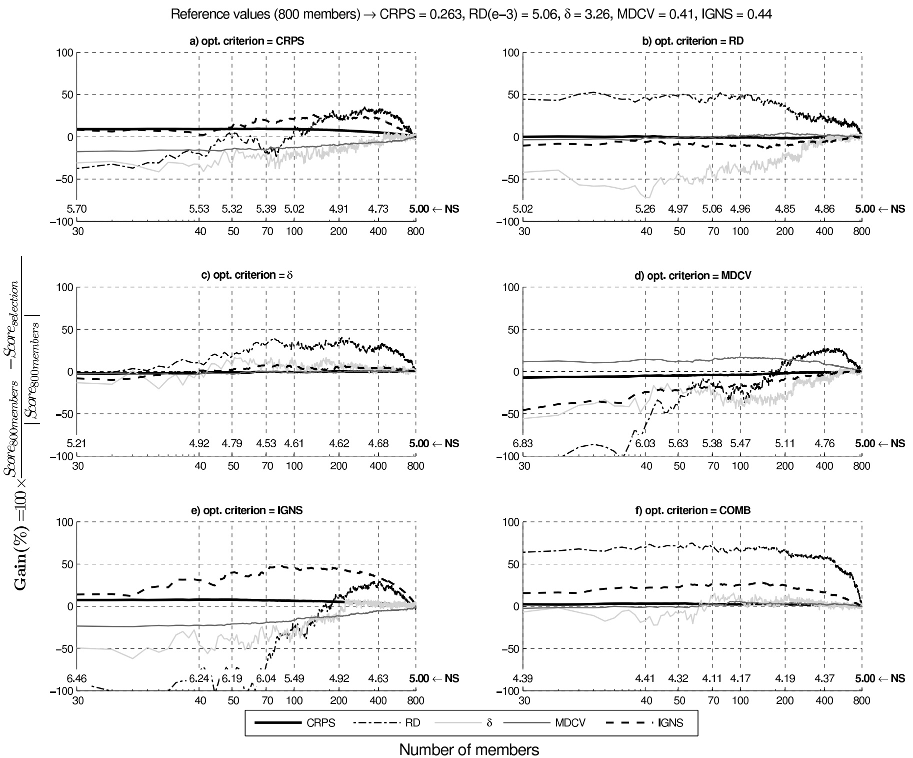

Fig. 5. Evolution of the gain index for each score under different optimization schemes in the basin A7930610. A logarithmic scale is used on the x-axis. The chosen optimization criterion in the selection is shown at the top of each subfigure. The lower part of each subfigure indicates the values of the normalized sum (NS) of all scores with unit weights (Eq. 7) for the number of members shown on the x-axis.

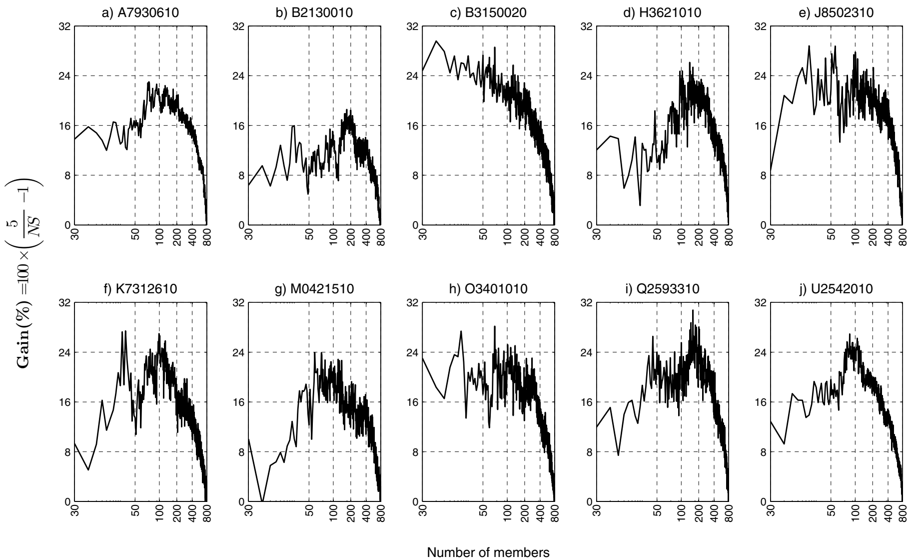

Fig. 6. Evolution of the normalized sum (NS) in terms of gain index. Logarithmic scale on the x-axis. Normalized sum equal to 5 represents the performance of the initial 800 members ensemble.

HESSD
8, 2739–2782, 2011

Simplifying a hydrological ensemble prediction system – Part 1

D. Brochero et al.