Hydrol. Earth Syst. Sci., 15, 3307–3325, 2011
www.hydrol-earth-syst-sci.net/15/3307/2011/
doi:10.5194/hess-15-3307-2011

Hydrology and Earth System Sciences

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 Engineering and Water Engineering, Université Laval, Québec, G1V 0A6, Canada
[2]Computer Vision and Systems Laboratory (CVSL), Department of Electrical Engineering and Computer Engineering, Université Laval, Québec, G1V 0A6, Canada

Received: 21 February 2011 – Published in Hydrol. Earth Syst. Sci. Discuss.: 11 March 2011
Revised: 14 September 2011 – Accepted: 10 October 2011 – Published: 4 November 2011

Abstract. Hydrological Ensemble Prediction Systems (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 the number of hydrological members) that can be achieved with a HEPS configured using 16 lumped hydrological models driven by the 50 weather ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF). Here, Backward Greedy Selection (BGS) is proposed to assess the weight that each model must represent within a subset that offers similar or better performance than a reference set of 800 hydrological members. These hydrological models' weights represent the participation of each hydrological model within a simplified HEPS which would issue real-time forecasts in a relatively short computational time. The methodology uses a variation of the k-fold cross-validation, allowing an optimal use of the information, and employs a multi-criterion framework 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). Today, the availability of the Meteorological Ensemble Prediction Systems (MEPS) and its subsequent coupling with multiple hydrological models offers the possibility of building Hydrological Ensemble Prediction Systems (HEPS) relying on a large number of members. But the complexity of such HEPS becomes an operational burden when one has to evaluate several hundreds of scenarios at each time step.

To provide an idea of the complexity that can be achieved in HEPS, represented for example by the number of members to handle, it is worth mentioning the principal areas of uncertainty associated with the hydrological process (Schaake et al., 2007) as follows:

- Uncertainty from the meteorological data: in this case, the MEPS are responsible for providing this information. Different centres around the world are currently working on this issue, for example the TIGGE initiative consists of ensemble forecast data from ten global centres, for a total of 259 members (TIGGE, Bougeault et al., 2010). In relation to this, Bao et al. (2011) have shown that a HEPS comprised of meteorological members derived from multiple meteorological centres may actually perform better as compared to an ensemble derived from a single meteorological model.
- Uncertainty from the rainfall-runoff model: each hydrological model combines two important elements regarding the uncertainty associated with the hydrological

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

process: the initialization uncertainty (i.e. the initial state of the model) and the model uncertainty (from parameter identification to model conceptualization). In this regard, the methodology proposed by Beven and Binley (1992) provides the evaluation of parameter uncertainty from the point of view of equifinality. For example Pappenberger et al. (2005) have shown the advantages in HEPS to flood inundation predictions coupling MEPS with both hydrological and hydraulic models that have been evaluated at the same time with the GLUE methodology.

Another way of conceptualizing the uncertainty of the model focuses on a multi-model approach, making good use of the resources invested in the development of dozens of hydrological models. For instance, Velázquez et al. (2011) have shown, based on the database of the present paper, that the ensemble predictions produced by a combination of several hydrological model structures and meteorological ensembles have higher skill and reliability than ensemble predictions given either by a single hydrological model fed by weather ensemble predictions or by several hydrological models driven by a deterministic meteorological forecast.

Cloke and Pappenberger (2009) have already highlighted the computational demand of using MEPS for flood forecasting 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. Thus, the selection of hydrological members as part of a simplified model can be useful given the computational cost of running models and creating ensembles. Vrugt et al. (2008) have suggested the selection of hydrological models as an additional task that can be run based on the results of the post-processing using Bayesian Model Averaging (BMA) in a multi-criteria framework.

As a compromise, researchers have attempted to cluster MEPS for flood prediction in various ways: by lagging ensembles and deriving representative members through hierarchical clustering over the domain of interest, and thus to produce a reduced ensemble set at higher resolution (Marsigli et al., 2001); by analyzing the relation between atmospheric circulation patterns and extreme discharges (Ebert et al., 2007), or by establishing, in a deterministic way (“best match” approach), the location of the forecast that is the most similar to the rainfall pattern of the catchment (Xuan et al., 2009).

Here, we propose the selection of hydrological members directly in the HEPS with a technique called Backward Greedy Selection under the different scores presented in Sect. 2. In the case of MEPS with interchangeable members (the case presented here), the selection is oriented to evaluate the hydrological models participation inside a subset of a few members.

The HEPS under study is formed of 16 lumped hydrological models forced by the 50 meteorological inputs of the ECWMF EPS, leading to a grand-ensemble of 800 members. This approach was tested in 10 catchments located in France for a period of seventeen months (from March 2005 to July 2006). 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). This condition imposes the use of resampling and recombination techniques in the proposed methodology shown in Sect. 3.

Other studies that focused on periods of analysis very similar to the one used in this paper have also proven the usefulness of the ECMWF EPS. For example Rousset et al. (2007) evaluated hundreds of French catchments from 4 September 2004 to 31 July 2005 showing that the information given by the ensemble forecast is useful for flood warning and water management agencies. Similarly, Thirel et al. (2008), in a comparative analysis of short-range meteorological forecasts from the ECMWF EPS and PEARP EPS of Météo-France under the scheme of SIM coupling, analysed the competence jurisdiction of each of the two EPS from 11 March 2005 to 30 September 2006, showing that the ECMWF EPS seemed best suited for low flows and large basins while the PEARP EPS was best suited for floods and small basins.

We do emphasize that the results shown in this first phase focused primarily on the analysis of the scores in the process of selecting hydrological members. Furthermore, we evaluated the notion of interchangeability of the MEPS and HEPS members, concluding that the participation of the hydrological models in the subset of selected members is sufficient to guide the members’ selection, as shown below in Sect. 4. Finally, conclusions are drawn and a guideline for future work is given in Sect. 5.

2 Verification statistics for ensemble forecasts

Following the guidelines given by Cloke and Pappenberger (2009), we consider several metrics in the selection of hydrological members with BGS. We thus quote some of the features that are evaluated in probabilistic forecasting. The reader is referred to Murphy (1993) and Wilks (2005) for a detailed description of these features.

- Bias: correspondence between mean forecast and mean observation.
- Reliability: correspondence between conditional mean observation and conditioning forecast, averaged over all forecasts.
- Resolution: degree to which the forecasts sort the observed events into groups that are different from one another. It is related to reliability, in that both are concerned with the properties of the conditional distributions of the observations given the forecasts.

- Sharpness: variability of forecasts as described by their distribution.
- Consistency: degree to which the ensembles apparently include the observations being predicted as equiprobable members.

Additionally, we propose the use of the diversity concept studied in machine learning, i.e. the members should be as correct as possible, and when they make errors, these errors should be complementary (Kuncheva, 2004). Thus, the scores used in this research have been chosen because they quantify different aspects of the ensemble prediction's quality.

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, it is accepted that an adjustment of the gamma distribution makes more sense than a normal distribution given the asymmetry in the distribution of precipitation and discharge (Vrugt et al., 2008); however, the gamma function evaluation involves a distribution which is 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 the Continuous Ranked Probability Score (CRPS) and the Ignorance Score (IGNS). The results showed minor variations in contrast with a high computational cost. It is nonetheless important to note that this similarity is evaluated inside the ensembles with previsions varying between 30 and 800 hydrological members, as detailed below; in small samples it is expected that the results represent the expected asymmetry of information.

Note that the CRPS can be evaluated directly from the cumulative distribution of observed frequencies (Hersbach, 2000). However, considering the computational cost in evaluating this score thousands of times, a normal distribution was assumed.

The mathematical notation of each element in the scores, explained below, is drawn from Appendix A.

2.1 Continuous ranked probability score (CRPS)

The CRPS simultaneously evaluates reliability, resolution, and uncertainty (Hersbach, 2000; Gneiting and Raftery, 2007). Smaller values indicate better performance. Its minimal value of zero is only achieved in the case of a perfect deterministic forecast. Note that the CRPS has the dimension of the observation o^t. Its mean value is equivalent to the mean absolute error for a deterministic forecast (Hersbach, 2000). Assuming that the forecast ensembles ($\boldsymbol{y}^t$) are normally distributed, the CRPS at the time t is defined by Eq. (1) (Gneiting and Raftery, 2007):

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

2.2 Ignorance score (IGNS)

Proposed by Good (1952) as the logarithmic score, the IGNS is given by Eq. (2):

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

This score is described in detail by Roulston and Smith (2002). It is used to evaluate the sharpness or spread (Vrugt et al., 2006). It severely penalizes the bias, since positioning the observation in forecast regions of low probability lead to values that tend to infinity. It is defined simply as the logarithm of the ensemble probability density function ($f\left(\boldsymbol{y}^t\right)$) at the point corresponding to the observation (o^t). Smaller values indicate better performance.

The logarithmic score involves a harsh penalty for low probability events and therefore is highly sensitive to extreme cases (Gneiting and Raftery, 2007). To rule out the possibility that the results solely reflect the effect of a few outliers, we analysed trimmed means of the IGNS series excluding the highest and lowest 2% data values, following Weigend and Shi (2000). Infinite values were replaced by the next worst non-infinite value, following Boucher et al. (2010).

2.3 Reliability diagram – mean square error ($\mathbf{RD_{MSE}}$)

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):

$$\bar{o}_m = \frac{1}{N}\sum_{t=1}^{N} r^t \quad \text{where} \quad r^t = \begin{cases} 1 & \text{if } o^t \in I_m \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

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. In this study, as discussed later in Sect. 3.3, it is necessary to establish a single target value, so the Mean Square Error between the probability forecast and the observed frequency in the Reliability Diagram ($\mathrm{RD_{MSE}}$), as suggested by Wilks (2005), is evaluated by Eq. (4):

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

These distances are all small for well-calibrated forecasts.

2.4 Normalized deviation of the rank histogram from flatness (δ ratio)

The reliability, consistency and bias of the ensemble are evaluated in this score. That is, the rank histogram is used to evaluate whether the ensembles apparently include the observations being predicted as equiprobable members. The rank histogram is a graphical approach that was devised independently by Anderson (1996); Hamill and Colucci (1997) and Talagrand et al. (1997). The rank of the observations within each ensemble is evaluated and then plotted in the form of a histogram. In the case of equality of observation with one or more of the ensemble members, the rank is chosen randomly. For a reliable system, over all $d+1$ members, the number of elements in each interval of the rank histogram (S_c) has an expected value $N/(d+1)$, while the deviation (Δ) of the histogram from flatness is measured by Eq. (5) (Talagrand et al., 1997):

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

A reliable system has an expectation of $\Delta_0 = \frac{dN}{d+1}$. The δ ratio ($\delta = \Delta/\Delta_0$), proposed by Talagrand et al. (1997) is used 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.

Given the difficulty of assessing the probabilistic nature of the studied HEPS, the use of the rank histogram is totally dependent upon eventually relaxing the ensemble members distribution, such as has been proposed by some authors (see Sect. 2c in Anderson, 1996 and Sect. 3a in Hamill and Colucci, 1997).

2.5 Median of coefficients of variation (MDCV)

Velázquez et al. (2011) showed that the reliability of the studied HEPS improved in two ways: first with the combination of all perturbed members from ECMWF EPS and the 16 hydrological models studied, and second, by increasing the lead time. A common feature is that the higher the observed dispersion, the greater the HEPS reliability.

The standard deviation is a classical measure of dispersion; however, it preserves the magnitude of the observed variable, complicating the joint interpretability of the results of the 10 basins in evaluation. So, the coefficient of variation (CV) as a dimensionless measure is useful in comparing different data sets with respect to central location and dispersion (Kottegoda and Rosso, 2009).

In this research, the analysis of the HEPS dispersion, through CV (results are omitted in this article), showed an increase proportional to the lead time, so the first lead time has a mean CV of 0.05 while longer lead times (e.g. 9 days), reached a mean value of 0.6. Note that CV is calculated for each time step. However, the mean CV is not a good measure of location in the skewed CV series evaluated for each basin. The MeDian of the Coefficients of Variation (MDCV), given by Eq. (6), turns out to be a much better measure:

$$\mathrm{MDCV}(\mathbf{Y}) = \operatorname*{med}_{t=1}^{N} \mathrm{CV}\left(\boldsymbol{y}^{\mathrm{t}}\right). \tag{6}$$

The hypothesis under the maximization of the MDCV is that a gain in dispersion should increase the reliability of the HEPS.

2.6 Combined criterion (CC)

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. So, we propose the following guidelines to define the CC:

$$\mathrm{CC} = w_1 \frac{\overline{\mathrm{CRPS}}_{\mathrm{se}}}{\overline{\mathrm{CRPS}}_{\mathrm{ie}}} + w_2 \frac{z_1 - \overline{\mathrm{IGNS}}_{\mathrm{se}}}{z_1 - \overline{\mathrm{IGNS}}_{\mathrm{ie}}} + w_3 \frac{\mathrm{RD_{MSE_{se}}}}{\mathrm{RD_{MSE_{ie}}}} \tag{7}$$
$$+ w_4 \frac{\delta_{\mathrm{se}}}{\delta_{\mathrm{ie}}} + w_5 \frac{z_2 - \mathrm{MDCV_{se}}}{z_2 - \mathrm{MDCV_{ie}}},$$

- The combination should assign weights to each of the scores as a direct measure prioritizing some of the characteristics of the HEPS in evaluation. Additionally, these weights, in a general framework, offer the possibility of constructing a trade-off among different objectives. In our case, weights were used only to give priority to the reliability in the selection, because Velázquez et al. (2011) showed that this was the most influential aspect in the evaluation of the HEPS studied here. For this reason the weight assigned to the reliability corresponds to twice that of the other factors, which have a unit weight.

- Each score in the selected ensemble of hydrological members (se subscript) is normalized from the division by the corresponding score in the initial 800-member ensemble (ie subscript), placing each component on the same scale.

- All scores except the MDCV function are oriented for minimization. However, the IGNS has the peculiarity

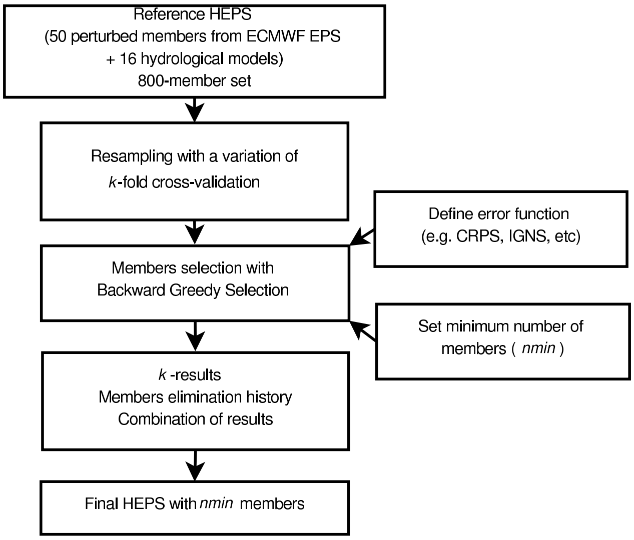
Reference HEPS
(50 perturbed members from ECMWF EPS
+ 16 hydrological models)
800-member set
Resampling with a variation of
k-fold cross-validation
Define error function
(e.g. CRPS, IGNS, etc)
Members selection with
Backward Greedy Selection
Set minimum number of
members (*nmin*)
k -results
Members elimination history
Combination of results
Final HEPS with *nmin* members

Fig. 1. Hydrological members' selection methodology.

of having negative values, making it necessary to establish a threshold (z_1) in the normalization so as to manipulate the duality of having a positive (or negative) score in the selection and a negative (or positive) score in the 800-member set. Thus, we establish $z_1 = -2$, since the preliminary analysis of selection under different scenarios (different catchments and number of members to be selected) showed minimum values for this score of about −1.5. With regard to the MDCV function, a threshold of $z_2 = 1$ is used to change the orientation since the objective is to maximize dispersion, as testing different scenarios showed maximum values of about 0.8.

2.7 Elements to compare the performance of members' selection (NS, G_{NS}, G_{SC})

Note that the CC could be used to compare the performance of the members' selection with respect to the 800-member set. So, in a general framework, if all features of the ensemble forecast have the same importance, one members' selection with equal performance to the 800-member set will lead to a CC equal to 5, values lower than 5 indicate a selection of higher performance than the base set of 800 members, and values greater than 5 indicate the detriment of any feature of the 800-member set. Hereafter this particular condition of unit weights in the CC will be called Normalized Sum (NS). This distinction is important to display the priority that can be defined a priori to any feature in the members' selection training with BGS. In this way, it is possible to define a gain index for the scores balance with respect to 5 (Eq. 8):

$$G_{\mathrm{NS}}(\%) = 100 \times \left(\frac{5}{\mathrm{NS}} - 1\right). \tag{8}$$

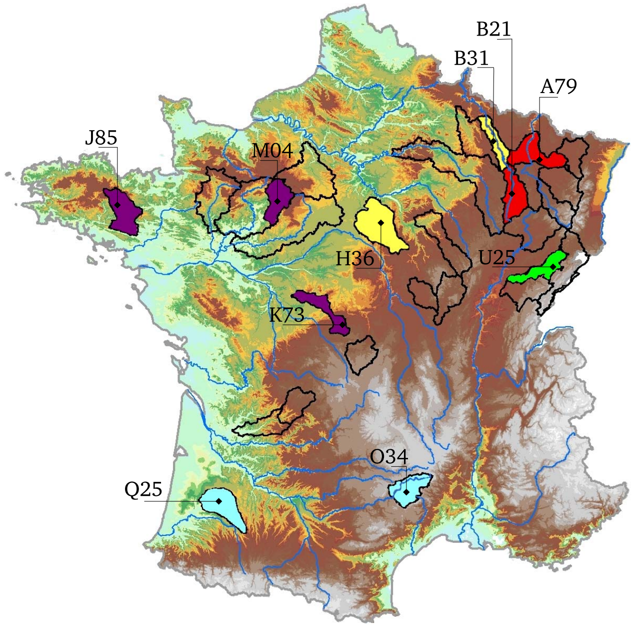

Fig. 2. Selected catchments are identified with the first three digits of each code used in Table 1. The other delimited basins are part of the study of results' generalization shown in Brochero et al. (2011).

It is possible that the NS evaluated in the selected sets with BGS hides undesirable effects on the balance of the scores, for example to substantially improve one score with respect to the other score(s). To check this condition, a gain index for each score is also proposed:

$$G_{\mathrm{SC}}(\%) = 100 \times \frac{\mathrm{Score}_{\mathrm{ie}} - \mathrm{Score}_{\mathrm{se}}}{|\mathrm{Score}_{\mathrm{ie}}|}. \tag{9}$$

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 have positive and negative values.

3 Experimental set-up

Figure 1 shows the selection procedure applied to the 800-member HEPS. The main elements of the methodology are described below.

3.1 Database: 800-member HEPS

Database details can be found in Velázquez et al. (2011). The study is conducted over 10 French catchments with a typical response time of 3 days. These catchments represent a large variety of hydro-climatic conditions (Fig. 2 and Table 1), and were evaluated over a period of 17 months (from March 2005 to July 2006).

Temperature, rainfall, and flow data are available at a daily time step over the period extending from 1970 to 2005, and

Table 1. Main characteristics of the studied basins (mean annual values) based on a 36 year length of the series (1970–2006).

Catchment codes	Area (km^2)	Altitude (m a.m.s.l.)	P (mm)	ET (mm)	Q (mm)
A7930610	9387	155	2.78	1.80	1.21
B2130010	2290	227	2.57	1.80	0.87
B3150020	3904	162	2.58	1.80	1.09
H3621010	3900	48	1.98	1.95	0.45
J8502310	2465	4	2.36	1.89	0.81
K7312610	1712	85	2.13	2.01	0.68
M0421510	1890	56	2.04	1.89	0.62
O3401010	2170	349	3.19	1.80	1.90
Q2593310	2500	17	2.52	2.24	0.75
U2542010	4970	201	3.63	1.75	1.88

P: precipitation, ET: potential evapotranspiration, Q: flow.

were used for the calibration and validation of the hydrological models. Observed data for the period 11 March 2005 to 31 July 2006 was used only for the evaluation of the forecasts. The forecast verification period is thus independent of the calibration/validation period. Rainfall data come from the meteorological analysis system SAFRAN of Météo-France (see Quintana-Seguí et al., 2008 for details). They consist of rainfall accumulated at a daily time step and available for the entire country of France at an 8 × 8-km grid resolution. Daily streamflow data come from the French database Banque Hydro (http://www.hydro.eaufrance.fr/). The length of available observed streamflow time series varies according to the catchment, with, on average, 29 years of available daily data for the catchment dataset used here.

The 50 perturbed forecasts from ECMWF was provided at a 0.5° × 0.5° lat/lon grid resolution. A detailed description of the ECWMF EPS model can be found in Molteni et al. (1996) or Buizza (2005). Forecasts are issued at 12:00 UTC and extend over 240 h. 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. No bias removal or disaggregation was performed. For each catchment, areal mean rainfall forecasts were computed by averaging the rainfall amounts of each grid above the catchment, weighted by the percentage of the catchment area inside the grid.

The sixteen hydrological models are lumped models and correspond to various conceptualizations of the rainfall-runoff transformation at the catchment scale. Some original model structures were modified. Thus, to avoid unfair comparisons of models, they will be referred to hereafter as HM## (Table 2). It is beyond the scope of this article to present these models. References with a detailed explanation of each model structure can be found in Velázquez et al. (2011).

On the other hand, analysis of the median coefficient of variation (MDCV), as a measure of the diversity of the HEPS, revealed the following characteristics:

Table 2. Hydrological models.

Hydrological models	Base model and parameters		Hydrological models	Base model and parameters	
HM01	CEQU	9	HM09	CREC	8
HM02	GR3J	3	HM10	GR4J	4
HM03	HBV0	9	HM11	SIMH	8
HM04	IHAC	6	HM12	MOHY	7
HM05	MORD	6	HM13	PDM0	8
HM06	SACR	13	HM14	HYM0	5
HM07	SMAR	9	HM15	TANK	10
HM08	TOPM	8	HM16	WAGE	8

- 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. (2011), 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 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 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 hydrological members as a method of simplifying HEPS should be unique regardless of the forecast horizon. The companion paper (Brochero et al., 2011) assesses the transferability of the 9-day members' selection to other forecast horizons.

3.2 Resampling technique

In some algorithms, such as the BGS, the overfitting[1] is highlighted as a structural problem. So, one method for improving generalization which is called early stopping (Hudson and Demuth, 2011; Alpaydin, 2010), well-known in the neural network community, is used in the methodology proposed here.

In this technique, the available data is divided into three subsets. The first subset is the training set, which is used in BGS for sequentially removing the members. The second subset is the validation set. The error on the validation set

[1] When the error on the training set is driven to small values, but the error of the model is large on new data.

is monitored during the training process. The validation error normally decreases during the initial phase of training, as does the training set error. However, when the selection begins to overfit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of members, the training is stopped. The test set error is not used during training, but it is used to compare different models.

The need to define three subsets to run the BGS and the short length of the series impose the use of resampling techniques such as k-fold cross-validation, which maximizes the utilization of the available information.

Moreover, one notes the high degree of linear correlation exhibited in the first lags of the correlogram of the flow series at hand (e.g. in the 80 % of the catchments evaluated, the correlation using a lag of three days was greater than 0.82). So, the choice of the training and validation data should be directed in order to temporarily avoid 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 not important in the BGS process.

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, a priori divided in blocks, 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 dataset 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.

3.3 Backward Greedy Selection (BGS)

In Machine Learning, 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. In a general context of selection, numerous methods have been developed. There are greedy selection methods (Backward or Forward Selection) but also methods such as integer programming and evolutionary algorithms.

Here, BGS and the idea of subdividing the data into three subsets to improve the generalization are applied. For its implementation it is necessary to define the error function "E" (that it is one of the given statistical scores shown in Sect. 2) and the minimum number of members. With regard to the minimum number of members, which was arbitrarily defined as 30 here, the choice is mainly due to the high availability of initial members (800), for example with 30 hydrological members a level of compression of information equivalent to 96.25 % is reached. It is certain that if the selection task had started with a pool of 50 members, then the minimum number of members could have been defined as 10, for example. Moreover, the minimum number of members is just a stopping criterion of selection with BGS because the number of members to define as optimal should focus on specific analysis in each basin.

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) into training (χ_t), validation (χ_v), and test set (χ_p).

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

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

3. For iter = $d-1$, $d-2$, ..., nmin
The hydrological member "$\boldsymbol{y}_j$" corresponds to the one that, when it is removed, has the greater impact on the training set error E (i.e. minimise train error the most). It is important to note that E must be a scalar or single value.

$$\boldsymbol{y}_j = \underset{\boldsymbol{y}_i \in \mathbf{G}^{\mathrm{iter}+1}}{\mathrm{argmin}} \; E\left(\mathbf{G}^{\mathrm{iter}+1} \backslash \{\boldsymbol{y}_i\} | \chi_t\right)$$

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

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

4. At this point, the error E in the validation set $\boldsymbol{\chi}_\mathrm{v}$, 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 analysed on the training and validation results.

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 removes hydrological members one by one, it may not be able to detect this. Here, the BGS is executed with a resampling technique explained in Sect. 3.2.

3.4 Combination of results

The variability of each experiment set-up in the cross-validation step increases the probability of reaching different hydrological member' selections. So, it is necessary to determine an integration mechanism for a global solution for each catchment. Here, the importance of each hydrological 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. (10):

$$\overline{R}(\boldsymbol{y}_i) = \frac{1}{5}\sum_{\mathrm{xp}=1}^{5}\mathrm{iter}_{\mathrm{xp}}^{\boldsymbol{y}_i}. \qquad (10)$$

For example, if the rank of elimination of the hydrological 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. Finally, the final selection (s) of the nm[2] best members corresponds to the members which have the highest mean rank of elimination given by Eq. (11):

$$s = \left\{\overline{R}_p,\, y_p\right\}_{p=1}^{nm}, \quad \overline{R}_i \geq \overline{R}_j \ \text{ where } \ 1 \leq i \leq j \leq d. \qquad (11)$$

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

3.5 Interpretability of hydrological members' selection

In the case of MEPS in which the members are not perfectly interchangeable (e.g. Meteorological Service of Canada – MSC, TIGGE database), the selection of hydrological members with BGS focuses directly on the combinations of hydrological members that maintain or improve characteristics of the super ensemble of reference.

Table 3. Performance for the 16-member ensemble (16 hydrological models are driven by the deterministic forecast from ECMWF EPS) and the 800-member ensemble (16 hydrological models are driven by the 50 perturbed members forecast from ECMWF EPS) for a 9-day forecast time horizon. Hereafter, $\mathrm{RD}_{\mathrm{MSE}}$ values are expressed on a 10^{-3} basis.

Catchment codes	HEPS	Scores				MDCV function
		CRPS	IGNS	$\mathrm{RD}_{\mathrm{MSE}}$	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

But in the HEPS driven by a MEPS with interchangeable members (e.g. ECMWF EPS), the selection should be directed more clearly to a method of selection and weighting of hydrological models based on their participation in the final selected subset. Therefore, we can create a new simplified high-performance HEPS using the same proportion of the hydrological members associated with a random choice of the meteorological members.

For example if the final selection shows that the simplified HEPS should consist of ten members for the hydrological model "A" and thirty members for the hydrological model "B", then we should expect to achieve a high performance HEPS if we randomly pick ten meteorological members to evaluate the hydrological model "A" and thirty meteorological members, randomly chosen once again, to assess the hydrological model "B". Section 4.3 presents such an analysis.

4 Results and analysis

Table 3 presents a comparison of results for two HEPS schemes analysed by Velázquez et al. (2011). It should be stressed that the 800-member HEPS serves as a reference for results in the selection of members and that their scores show

[2] nm is not necessarily equal to $nmin$ because nm reflects the analysis of the error on the validation set regarding the number of selected members.

a desired behaviour. Note that the δ ratio and $\mathrm{RD}_{\mathrm{MSE}}$, scores on which rest the main advantages of the 800-member HEPS, are directly interpretable since the scale is independent of the measured variable.

With respect to the IGNS score, mean values are generally negative, which shows that on average the system has an acceptable bias. Finally, in terms of CRPS, Velázquez et al. (2011) show in detail the efficiency of CRPS in this 800-member HEPS.

Note that results discussed in this paper correspond to a "pseudo test dataset" for comparing the performance between different scores in the process of selecting hydrological members, since the data used to minimize all error functions are exactly the same.

It is a "pseudo test dataset" because there is a high probability that the data used in testing (the complete series) have been used in the BGS training process, becoming the indicator of an optimistic estimator of the selection (Diamantidis et al., 2000); however, we do emphasize that the first part of this research focuses on an analysis of scores in the BGS process with the subsequent integration of results, and the second phase presented in a companion paper (Brochero et al., 2011) shows a rigorous test of generalization in time and space.

Validation results were omitted mainly because they have a trend similar to the training ones, except for some experiments where the random distribution of the training and validation sets was not statistically homogeneous.

In order to illustrate the interchangeability of the members of the ECMWF EPS and equiprobability of this system, Sect. 4.3 shows both the performance of the subset found with the BGS and the boxplot diagrams of 200 random experiments of 50 members, with and without the guidance of the BGS solution.

4.1 Selection performance

An example of the results obtained is 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.

We stress the fact that the selection task focuses on the participation of the hydrological models. For instance, Fig. 3e shows that the selected hydrological members make use of 13 of the 16 available lumped models, however, the strong participation of the models 3, 7, 9, and 14 is displayed, which is an interesting combination of hydrological models, especially taking into account the much poorer performance of the 16-member multi-model approach driven by the deterministic prediction (Table 3) and knowing that these hydrological models are not of equal quality with regards to MSE performance. This suggests that the selection favoured a diversity of errors.

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; Velázquez et al., 2011). 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 removing the hydrological members that have a large impact in the daily extreme CRPS values. Note also that 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 4 % 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 have many extreme values which suggest low assessments of the predictive distribution of the ensembles, i.e. a bias problem in the forecasts (note that a value of 4.5 corresponds to an evaluation of the *pdf* near 0.0442).

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

At the end of the selection process, the (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.

Figure 3e illustrates 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.

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 BGS results. The next section presents this analysis. However, the issue of the optimal number of hydrological members remains somehow blurred. So, Fig. 4 revisits that question in terms of the gain index based on NS defined in Eq. (8). Figure 4 emphasizes that the 30-member selection always displays a positive gain index. However, one should keep in mind that the optimal number of members should be based on an individual analysis of the different scores

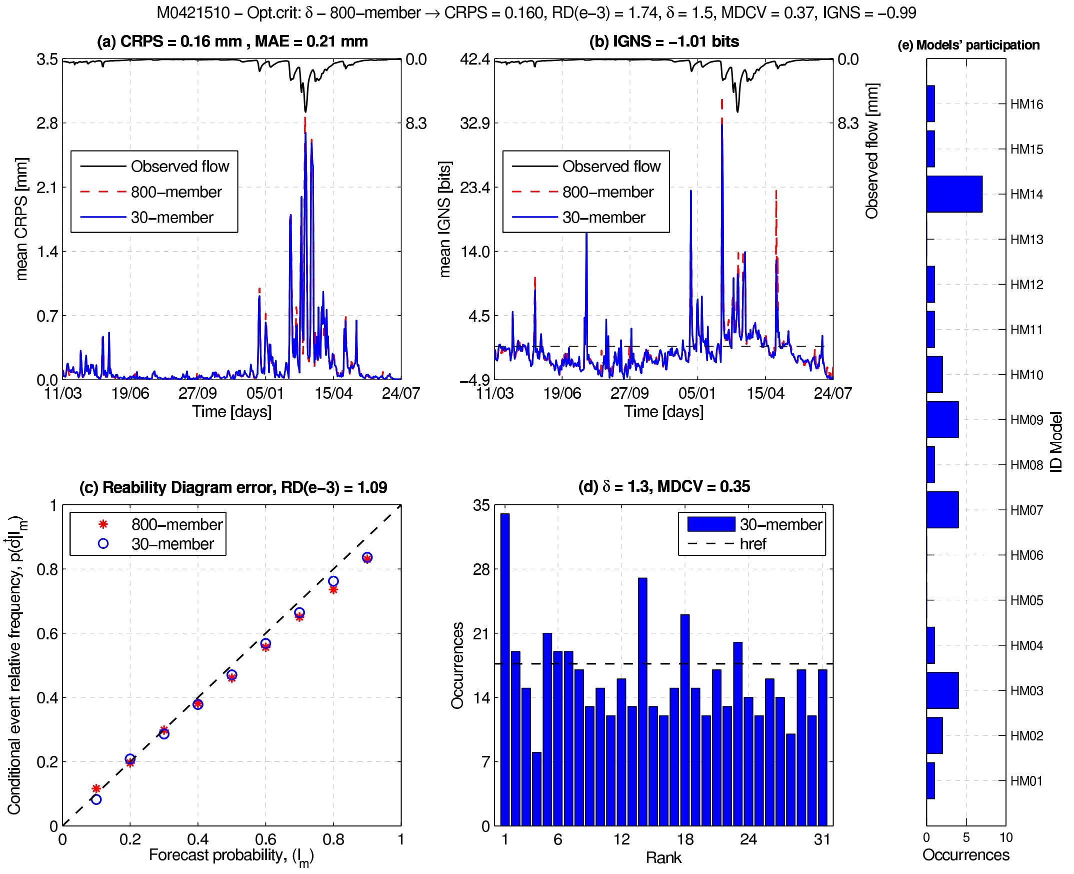

Fig. 3. Comparison between the initial ensemble (800 members) and the ensemble selected (30 members) for the lead time 9 in the catchment M0421510. **(a)** Figure above: observed flow; figure below: CRPS, x-axis indicates day/month. Note the correspondence between higher observed flows and higher mean CRPS. **(b)** Figure above: observed flow; figure below: IGNS. Note that there is no full correspondence between the higher IGNS and higher observed flow, x-axis indicates day/month. **(c)** Reliability diagram error ($\mathrm{RD}_{\mathrm{MSE}}$ based on vertical distances between the points). **(d)** Rank histogram for the 30 selected hydrological members. The horizontal dashed line indicates the frequency $(N/d+1)$ attained by a uniform distribution. **(e)** Occurrences of the employed models in the final solution of 30 hydrological members.

balance, i.e. evaluating that the normalized sum does not mask the detriment in a score(s) with gains made in other.

On the other hand, to reflect the BGS performance in the selection, Fig. 4 also presents the NS evaluation with 200 random selections of 30, 50, 100, 200, and 400 members in terms of gain index defined in Eq. (8). It is clear that BGS selection with positive gains are always obtained – improving the balance of the scores. Otherwise in random experiments the percentiles 10, 25, 50, 75, and 90 are shown generally in the range of a negative gain index (i.e. a detriment to the balance of the criteria). This tendency is obviously stronger in random selections of 100 or fewer hydrological members where the probability of taking the most representative hydrological responses is lower. It is important to note how even in the random selection of 200 and 400 members (25 % and 50 % of the 800 hydrological members) the NS in 75 % of the evaluations shows a negative gain index.

To check each score individually, Table 4 shows the median of 200 random selections for basin H36 optimized with the combined criterion. The random selections pick 50 hydrological members to evaluate each score in a standardized fashion, that is, dividing the score obtained in the selection subset by the reference score of all 800 members of base (see each component in Eq. 7 without weight parameter).

Table 4 shows an analysis to evaluate the sensitivity of the scores with respect to the selection of hydrological members in the database under study. So, it is possible to point out the following:

- In the hydrological members' selection, the greatest challenge is selecting a small set of members, for example 30 or 50.

- CRPS indifference to the selection of members, and to a lesser extent, both the low variability of the IGNS and the MDCV function.

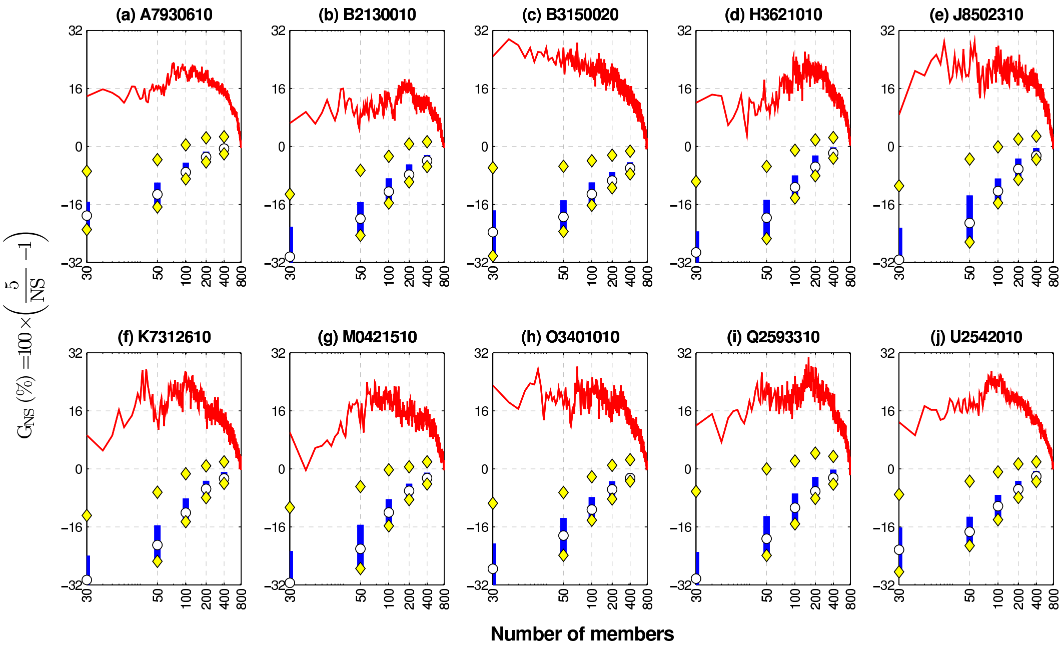

Fig. 4. Evolution of the normalized sum (NS) in terms of gain index for the lead time 9, optimization with the combined criterion. Logarithmic scale on the x-axis. The normalized sum equal to 5 represents the performance of the initial 800-member ensemble. The thin red line represents the normalized sum under different numbers of members found with BGS. Symbols for the 200 random selection experiments: the blue vertical line identifies the interquartile range, white circles represent the median and yellow diamonds corresponding to the percentiles 10 and 90.

Table 4. Median of 200 random selections in catchment H36 for the lead time 9. The scores are presented in a standardized fashion and oriented at the minimization coinciding with the formulation of each component of the combined criterion (Eq. 7).

Members	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS
30	1.01	1.50	1.80	1.05	1.11	6.47
50	1.01	1.25	1.53	1.03	1.06	5.88
100	1.00	1.09	1.28	1.02	1.02	5.41
200	1.00	1.02	1.11	1.01	1.01	5.15
400	1.00	0.98	1.03	1.00	1.00	5.01

Table 5. Results of BGS in catchment H36 for the lead time 9 with the combined criterion as error function. The scores are presented in a standardized fashion and oriented at the minimization coinciding with the formulation of the Eq. 7.

Members	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS
30	1.00	1.00	0.96	1.00	1.00	4.96
50	1.00	0.92	0.99	1.00	1.00	4.91
100	1.00	0.80	1.01	1.00	1.00	4.81
200	1.00	0.58	0.97	0.99	1.00	4.54
400	0.99	0.45	0.88	0.98	1.00	4.30

– The hydrological members' selection presents its greatest challenges in maintaining or improving reliability and the consistency of the ensemble represented by the δ ratio, as shown in Table 3. Therefore, to define the combined criterion, such as an error term in BGS, the reliability term (RD_{MSE}) has more weight to guide the optimization in this way. At this point it should be noted that consistency has a direct relationship with reliability, although ensemble consistency does not necessarily imply that probability forecasts constructed from the ensemble are reliable in the sense of conditional outcome relative frequencies being equal to the forecast probabilities yielding a 45° calibration function on a reliability diagram, unless either the ensemble size is relatively large or the forecasts are reasonably skillful, or both (Wilks, 2011).

Finally, Table 5 shows detailed results for each score in the selection process with BGS for the basin H36. It shows that in the BGS methodology, with the combined criterion as error function, is not detrimental to any of the scores. Instead, gains in the balance scores (normalized sum) are mainly due to optimization of system reliability while preserving the quality of the other scores.

4.2 Scores interaction in the selection

Table 6 summarizes results for more catchments and optimization criteria. The 30-member comparison is based on a normalized sum (Sect. 2.7). In this way, a value of NS lower than 5 indicates an overall improved performance. Performance for all criteria are also given in Table 6 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 BGS technique or the effects of the combination of results, because if the objective function (CC) is equal to the criterion used to compare results obtained with different objectives, the CC 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 five other catchments. This is particularly interesting considering the simplicity of this approach with respect to the combined approach. In addition, the δ criterion favoured 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. The CRPS showed low variability, so it is not very sensitive to changes in the selection of hydrological members, as was shown in Tables 4 and 5. The IGNS demonstrated a negative relationship with reliability, leading to poor performance in terms of the reliability diagram (RD) and δ ratio. They are also correlated, optimizing one criterion often favouring the improvement of the other one.

Specifically the behaviour of the optimization of each score could also be described from the following relationships observed in Table 6:

- Optimization based on CRPS is detrimental to the reliability. For example, it had the effect of increasing RD by a factor of 10, for catchment Q25. The CRPS also decreases diversity of the members (MDCV), except for catchment B31 where it 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 $\mathrm{RD}_{\mathrm{MSE}}$ score) and ensuring the stability 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 constant, which highlights the interaction between the scores as a mechanism implicit in the reduction of error reliability ($\mathrm{RD}_{\mathrm{MSE}}$). The δ ratio is never improved, while diversity (MDCV) is lost except in three cases (B31, Q25, and U25) 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 on the reliability and the consistency of the ensemble. In fact, it leads to better reliability performance in terms of RD, 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 members (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 mainly define 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

Table 6. Selection of 30 hydrological members based on different scores in all the basins or the lead time 9. NHM indicates the number of hydrological models participating in the solution.

Basin	Optimization criterion	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS	NHM
A79	CRPS	0.24	7.0	4.3	0.34	0.41	5.7	8
	CC	**0.26**	**1.8**	**3.4**	**0.40**	**0.38**	**4.4**	**10**
	RD	0.26	2.8	4.6	0.40	0.49	5.0	7
	δ	0.27	5.1	3.7	0.40	0.48	5.2	13
	MDCV	0.28	11.1	5.0	0.46	0.65	6.8	7
	IGNS	0.24	9.6	4.8	0.31	0.38	6.5	6
	800 members	0.26	5.0	3.3	0.41	0.44	5.0	16
B21	CRPS	0.21	4.0	4.5	0.49	−0.48	6.7	8
	CC	**0.23**	**1.3**	**2.6**	**0.63**	**−0.16**	**4.7**	**13**
	RD	0.23	2.5	3.9	0.53	−0.33	5.9	8
	δ	0.23	2.1	3.0	0.56	−0.27	5.2	14
	MDCV	0.24	5.2	3.7	0.61	−0.26	6.8	8
	IGNS	0.22	23.2	8.0	0.39	−0.33	16.7	7
	800 members	0.23	2.4	2.2	0.57	−0.29	5.0	16
B31	CRPS	0.18	5.9	4.6	0.22	−0.97	5.9	7
	CC	**0.13**	**0.9**	**2.0**	**0.23**	**−0.85**	**4.0**	**10**
	RD	0.15	3.5	5.2	0.24	−0.62	6.2	8
	δ	0.13	2.9	3.3	0.23	−0.86	4.9	12
	MDCV	0.14	12.1	7.3	0.24	−0.70	8.7	7
	IGNS	0.12	17.4	7.1	0.17	−0.97	9.5	8
	800 members	0.14	4.5	2.7	0.22	−0.88	5.0	16
Q25	CRPS	0.14	21.9	5.9	0.25	−0.96	17.2	6
	CC	0.16	0.7	1.8	0.37	−0.97	4.5	9
	RD	0.17	1.7	3.1	0.38	−0.84	6.0	5
	δ	**0.16**	**0.6**	**1.6**	**0.37**	**−0.98**	**4.4**	**13**
	MDCV	0.18	3.9	3.5	0.45	−0.74	7.3	5
	IGNS	0.15	32.0	12.5	0.18	−0.41	26.9	6
	800 members	0.16	2.2	1.5	0.37	−0.98	5.0	16
H36	**CC**	**0.16**	**1.1**	**1.7**	**0.36**	**−0.97**	**4.5**	**11**
	RD	0.16	2.9	2.5	0.34	−1.00	5.5	7
	δ	0.16	2.4	1.9	0.36	−1.02	4.9	13
	MDCV	0.17	2.5	3.8	0.44	−0.79	6.4	6
	800 members	0.16	3.5	1.5	0.37	−0.99	5.0	16
J85	**CC**	**0.16**	**0.5**	**2.3**	**0.39**	**−0.98**	**4.6**	**12**
	RD	0.17	2.3	3.1	0.35	−0.91	6.2	7
	δ	0.16	1.3	1.6	0.36	−0.99	4.6	13
	MDCV	0.18	1.6	2.5	0.44	−0.74	5.5	6
	800 members	0.16	2.2	1.6	0.37	−0.98	5.0	16
K73	**CC**	**0.16**	**1.3**	**2.4**	**0.36**	**−0.96**	**4.6**	**9**
	RD	0.17	3.4	3.7	0.35	−0.89	6.1	7
	δ	0.16	2.1	3.3	0.33	−0.95	5.5	13
	MDCV	0.17	2.5	4.2	0.43	−0.68	6.2	6
	800 members	0.17	3.1	1.9	0.35	−0.93	5.0	16
M04	CC	0.16	0.7	1.9	0.36	−0.99	4.5	12
	RD	0.16	2.1	2.9	0.36	−0.92	6.3	6
	δ	**0.16**	**1.1**	**1.3**	**0.35**	**−1.01**	**4.4**	**13**
	MDCV	0.17	2.6	3.3	0.44	−0.75	7.0	5
	800 members	0.16	1.7	1.5	0.37	−0.99	5.0	16
O34	**CC**	**0.17**	**0.9**	**1.3**	**0.36**	**−0.87**	**4.1**	**13**
	RD	0.17	2.5	4.2	0.36	−0.67	6.6	5
	δ	0.17	1.9	1.8	0.37	−0.85	4.7	12
	MDCV	0.19	5.7	4.9	0.44	−0.51	7.9	4
	800 members	0.17	3.5	1.5	0.36	−0.86	5.0	16
U25	**CC**	**0.29**	**1.2**	**2.9**	**0.37**	**−0.34**	**4.4**	**12**
	RD	0.30	2.9	5.0	0.37	−0.25	5.8	6
	δ	0.29	1.8	2.9	0.34	−0.32	4.7	15
	MDCV	0.30	3.0	3.7	0.43	−0.10	5.4	5
	800 members	0.29	3.4	2.6	0.35	−0.36	5.0	16

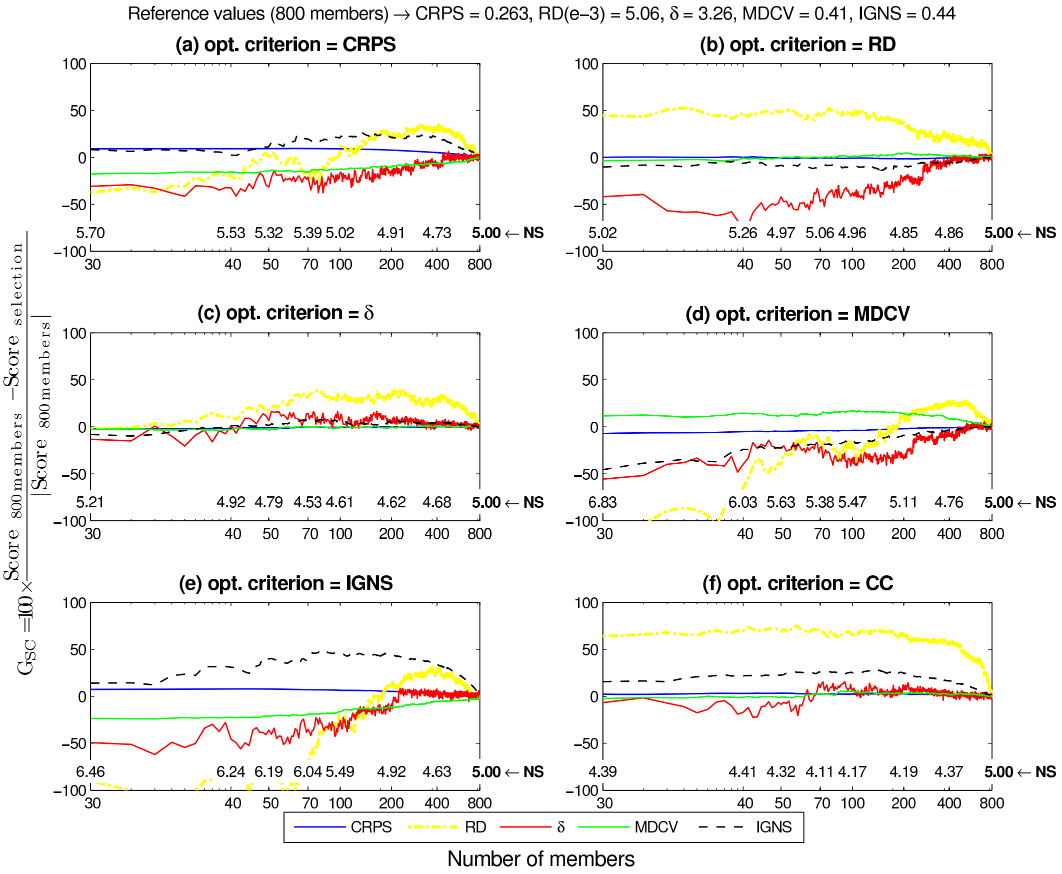

Fig. 5. Evolution of the gain index for each score under different optimization schemes in the basin A79 for the lead time 9. A logarithmic scale is used on the x-axis. The chosen optimization criterion in the selection is shown at the top of each subfigure. The lower part of each subfigure indicates the values of the normalized sum (NS) for the number of hydrological members shown on the x-axis.

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 requires the analysis of the evolution of the scores as the number of hydrological 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 the catchment A79.

In order to assess the joint evolution of all scores the gain index defined by Eq. (9) was used. Figure 5a and 5e clearly show that an optimization based on resolution of the system (CRPS or IGNS) is detrimental to the reliability. 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 unfavourable 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 hydrological members is equal to 100. This is strictly because loss in consistency (negative gain index in the δ ratio of 40 %) 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 in a selection of about 70 hydrological 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 are achieved only when the number of hydrological members is greater than 400.

The combined criterion (Fig. 5f) improves collective performance of all scores in the selection, with an optimal number of hydrological members of 70 for this catchment, coinciding with the interaction shown in the minimization of the δ ratio (Fig. 5c). Scores tend to lose quality afterwards.

Table 7. Selection of 100 hydrological members based on the combined (CC) and δ criteria. NHM indicates the number of hydrological models participating in the solution.

Basin	Optimization criterion	CRPS	$\mathrm{RD}_{\mathrm{MSE}}$	δ	MDCV	IGNS	NS	NHM
A79	CC	0.26	1.8	3.0	0.43	0.33	4.2	13
	δ	0.27	3.5	3.0	0.41	0.43	4.6	16
	800 members	0.26	5.1	3.3	0.41	0.44	5.0	16
B21	CC	0.23	1.0	2.3	0.63	−0.19	4.4	14
	δ	0.28	1.2	2.4	0.59	−0.28	4.5	16
	800 members	0.23	2.4	2.2	0.57	−0.29	5.0	16
B31	CC	0.13	1.0	2.4	0.25	−0.83	4.2	14
	δ	0.14	2.3	2.5	0.23	−0.85	4.5	16
	800 members	0.14	4.5	2.7	0.22	−0.88	5.0	16
Q25	CC	0.16	0.4	1.3	0.40	−0.98	4.0	16
	δ	0.16	0.6	1.4	0.36	−1.05	4.2	16
	800 members	0.16	2.2	1.5	0.37	−0.98	5.0	16
H36	CC	0.16	0.6	1.6	0.38	−1.03	4.2	14
	δ	0.16	2.5	1.8	0.36	−1.04	4.8	16
	800 members	0.16	3.5	1.5	0.37	−0.99	5.0	16
J85	CC	0.16	0.4	1.5	0.39	−0.98	4.0	15
	δ	0.16	1.3	1.7	0.38	−1.00	4.6	16
	800 members	0.16	2.2	1.6	0.37	−0.98	5.0	16
K73	CC	0.16	0.6	1.7	0.39	−0.91	4.0	14
	δ	0.16	2.6	2.2	0.34	−0.95	5.0	16
	800 members	0.17	3.1	1.9	0.35	−0.93	5.0	16
M04	CC	0.16	0.3	1.7	0.37	−1.00	4.2	15
	δ	0.16	0.8	1.3	0.36	−1.03	4.2	16
	800 members	0.16	1.7	1.5	0.37	−0.99	5.0	16
O34	CC	0.17	0.7	1.4	0.38	−0.87	4.1	16
	δ	0.17	2.2	2.1	0.37	−0.89	4.9	16
	800 members	0.17	3.5	1.5	0.36	−0.86	5.0	16
U25	CC	0.29	0.9	2.2	0.39	−0.38	4.1	14
	δ	0.29	1.4	2.5	0.36	−0.42	4.3	16
	800 members	0.29	3.4	2.6	0.35	−0.36	5.0	16

Table 7 groups the 100-member scores following optimization with the combined score and the δ ratio, the two best ones. 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 ($\mathrm{NS} < 5$) 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 optimization with the CC.

4.3 Interchangeability of MEPS members as input of hydrological models

In order to illustrate the interchangeability of the members of the ECMWF EPS and equiprobability of this system, Fig. 6a shows that a random selection oriented only with the hydrological models' participation in the BGS has a chance to have even better performance than the 800-member HEPS upper 90 % (top of the box diagram). These box plots are constructed by retaining the participation of hydrological models in the response but with a random selection of members of the MEPS. On the other hand, Fig. 6b shows the same kind of results under different random selections but without considering the participation of hydrological models found with BGS.

Figure 6 highlights three main aspects: high-performance solutions based on the proportion given by the BGS, low variability, and high performance of the BGS solutions.

The performance of selections based on the proportion of members found in the BGS solution is evident in Fig. 6a. So, it is demonstrated that the proportion of members for a hydrological model is generally a sufficient criterion to reduce the number of members while improving the balance of the scores represented by the normalized sum. For comparison, Fig. 6b illustrates the system response to random selections without any a priori guidance, showing that in all cases the normalized sum is greater than 5 and have recurring extremes greater than 7.

Regarding the variability of the normalized sum evaluated in random selections guided by the BGS solution, it can be seen that the interquartile range ($Q_3 - Q_1$) is at worst equal to 0.3 (catchment H36), which is a much lower value than for the purely random selection, as shown in Fig. 6b where the latter interquartile range is equal to 0.6.

The generalization of the BGS method is discussed in detail in the companion paper, where the temporal and spatial generalization is evaluated for a nearby catchment. However, Fig. 6a shows that the catchments H36 and J85 obtained combinations with a normalized sum lower than those obtained with the BGS method (see only cross points at the bottom in Fig. 6a), which can be associated with the integration of experiments carried out in a subdivision database for each catchment or the BGS algorithm structure – it is known that the classical BGS algorithm is unable to detect the collective influence of the variables.

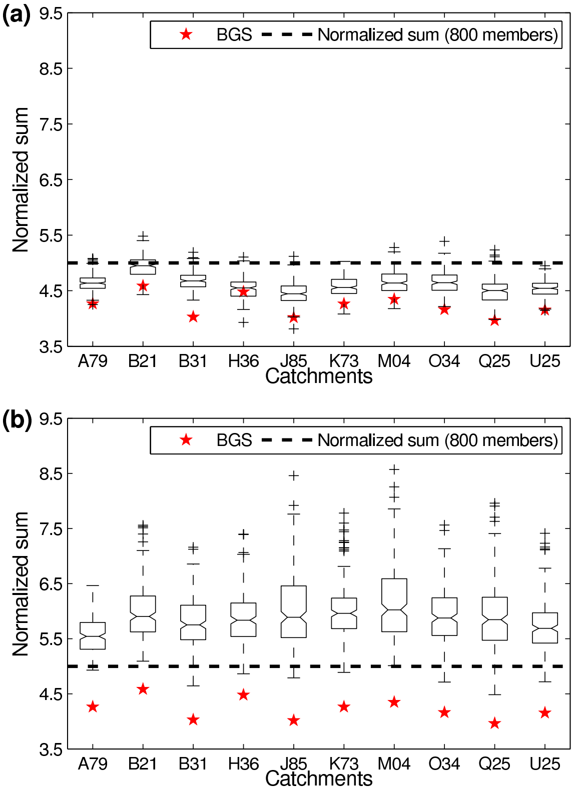

Fig. 6. Backward Greedy Selection (BGS) and Box-plots in 200 random experiments of 50 hydrological members for the lead time 9. On each box, the central mark is the median, the edges of the box are the 25th (Q_1) and 75th percentiles (Q_3), the whiskers or limits to consider the outliers extend from $Q_1 - 1.5 \times (Q_3 - Q_1)$ to $Q_3 + 1.5 \times (Q_3 - Q_1)$ points (but not necessarily correspond to observed data), and the outliers are plotted individually as cross points. **(a)** Random selection oriented with the frequency observed in the BGS to check the interchangeability in the 800 member-set, **(b)** Random selection without any guidance to check the BGS performance.

5 Conclusions

Previous results on the number of hydrological members and the HEPS conformation (Velázquez et al., 2011) have shown, based on the database of the present paper, that the ensemble predictions produced by a combination of several hydrological model structures and meteorological ensembles (800-member set) have higher skill and reliability than ensemble predictions given either by a single hydrological model fed by weather ensemble predictions (50-member set) or by several hydrological models driven by a deterministic meteorological forecast (16-member set). So, our goal was focused on at least replicating the good quality of the 800-member set with fewer hydrological members.

Hydrological member selection is justified by the computational cost to issue a hydrological forecast based on the combination of meteorological models and hydrological models. In this line, the selection of hydrological members without sacrificing the quality of a forecast stands out as an operational option.

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 hydrological 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 hydrological members or as a function of the maximization of the balance of the scores. Simplification of the number of hydrological members involves the definition of a limit ensuring statistical consistency of the scores assessed. A trade-off exists between the number of hydrological members and the level of improvement in scores. For example, in this study, the best balance of scores is achieved with a number of members fluctuating between 30 and 100, maximizing the qualities of the system: reliability, consistency, resolution, and diversity. So in the worst case this corresponds to a 87.5 % (700 members/800 members) 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 hydrological 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 combined criterion (CC).

The CRPS is often the primary score used for evaluating HEPS performances. However, results here indicate that it is not a good choice for hydrological members' selection in this case of study. In fact, it was often possible to preserve or minimize the CRPS using other 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 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 improve the balance of the scores.

The proposed methodology is part of the so-called data-driven models, so the design is independent of the database, in this case the evolution of MEPS or hydrological models. Precisely this point stands out as one of the advantages of the proposed methodology, since the selection of hydrological members could be implemented in any desired combination between any MEPS (e.g. ECMWF EPS, MSC, US National Centers for Environmental Prediction – NCEP) and hydrological models.

The cross-validation, a vital part of the proposed methodology, systematically deals with the issue of the short length of the series. However, 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 evolutionary algorithms.

Appendix A

Notations

t	Time-step
N	Number of pairs observations-forecasts
d	Total number of hydrological members in the forecast ensembles
M	Total number of m intervals to analyse the reliability diagram
c	Identification of the rank or class to analyse 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	i-th 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
ϕ	Normalized variables for probability density function
Φ	Normalized variables for cumulative distribution function
$\bar{o}_m$	Conditional probability of the event as a function of the interval I_m assigned to the forecast $m \rightarrow P(o^t \mid I_m)$
r^t	Binary indicator, 1 if the event occurs for the tth forecast-event pair, 0 if it does not
S_c	Number of elements of the c-th interval of the rank histogram ($c=1, ..., d+1$)
$\operatorname*{med}\limits_{t=1}^{N}$	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
$\boldsymbol{\chi}_{\mathrm{t}}$	Training set
$\boldsymbol{\chi}_{\mathrm{v}}$	Validation set
$\boldsymbol{\chi}_{\mathrm{p}}$	Test or publication set
$\{x^t\}_{t=1}^{N}$	Set of x with index t ranging from 1 to N
$\operatorname*{argmin}\limits_{\theta} \quad g(x \mid \theta)$	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 $\boldsymbol{y}_i$ hydrological member during the selection process in the xp experiment

$\overline{R}(\boldsymbol{y}_i)$	Mean rank of elimination of the $\boldsymbol{y}_i$ hydrological member
s	Final selection of the nm best hydrological members in the selection process

Acknowledgements. The authors acknowledge NSERC and the Institute EDS for financial support, CEMAGREF and ECWMF for making the database available, and the CVSL for the computational resources. We thank G. Thirel, and an anonymous reviewer for their careful revision, which helped improve this manuscript. We also thank Annette Schwerdtfeger for proofreading this manuscript.

Edited by: F. Pappenberger

References

Alpaydin, E.: Introduction to Machine Learning. Adaptive Computation and Machine Learning, The MIT Press, Cambridge, 2nd Edn., 2010.

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, 1996.

Bao, H.-J., Zhao, L.-N., He, Y., Li, Z.-J., Wetterhall, F., Cloke, H. L., Pappenberger, F., and Manful, D.: Coupling ensemble weather predictions based on TIGGE database with Grid-Xinanjiang model for flood forecast, Adv. Geosci., 29, 61–67, doi:10.5194/adgeo-29-61-2011, 2011.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, doi:10.1002/hyp.3360060305, 1992.

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

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.

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.

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., 15, 3327–3341, doi:10.5194/hess-15-3327-2011, 2011.

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

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

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.

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

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.

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.

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

Hamill, T. M. and Colucci, S. J.: Verification of Eta RSM Short-Range Ensemble Forecasts, Mon. Weather Rev., 125, 1312–1327, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2, 1997.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems, Weather Forecast., 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hudson, M. H. M. and Demuth, H.: Neural Network Toolbox – User's Guide, The MathWorks, http://www.mathworks.com/help/pdf_doc/allpdf.html, last access: October 2011.

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

Kuncheva, L. I.: Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, 2004.

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. Meteorol. Soc., 127, 2095–2115, doi:10.1002/qj.49712757613, 2001.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF Ensemble Prediction System: Methodology and validation, Q. J. Roy. Meteorol. Soc., 122, 73–119, doi:10.1002/qj.49712252905, 1996.

Murphy, A.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather Forecast., 8, 281–293, 1993.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and de Roo, A. P. J.: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), Hydrol. Earth Syst. Sci., 9, 381–393, doi:10.5194/hess-9-381-2005, 2005.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, J. Appl. Meteorol. Clim., 47, 92–107, doi:10.1175/2007JAMC1636.1, 2008.

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.

Roulston, M. S. and Smith, L. A.: Evaluating Probabilistic Forecasts Using Information Theory, Mon. Weather Rev., 130, 1653–1660, doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2, 2002.

Rousset, F., Habets, F., Martin, E., and Noilhan, J.: Ensemble streamflow forecasts over France, ECMWF Newsletter, 111, 21–27, 2007.

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.

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.

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., Shinfield Park, Reading, Berkshire RG2 9AX, UK, http://www.ecmwf.int/publications/library/do/references/list/16233, 1–25, 1997.

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

Thirel, G., Rousset-Regimbeau, F., Martin, E., and Habets, F.: On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Predictions, J. Hydrometeorol., 9, 1301–1317, doi:10.1175/2008JHM959.1, 2008.

Velázquez, J. A., Anctil, F., Ramos, M. H., and Perrin, C.: Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures, Adv. Geosci., 29, 33–42, doi:10.5194/adgeo-29-33-2011, 2011.

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.

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., and Robinson, B. A.: Multi-objective calibration of forecast ensembles using Bayesian model averaging, Geophys. Res. Lett., 33, L19817, doi:10.1029/2006GL027126, 2006.

Weigend, A. S. and Shi, S.: Predicting daily probability distributions of S&P500 returns, J. Forecast., 19, 375–392, doi:10.1002/1099-131X(200007)19:4<375::AID-FOR779>3.0.CO;2-U, 2000.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, vol. 91, 2nd Edn., Burlington, MA, London, Academic Press, 2005.

Wilks, D. S.: On the Reliability of the Rank Histogram, Mon. Weather Rev., 139, 311–316, doi:10.1175/2010MWR3446.1, 2011.

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.