Hydrol. Earth Syst. Sci., 15, 3327–3341, 2011
www.hydrol-earth-syst-sci.net/15/3327/2011/
doi:10.5194/hess-15-3327-2011

Hydrology and Earth System Sciences

Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 2: Generalization in time and space

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: 27 September 2011 – Accepted: 10 October 2011 – Published: 4 November 2011

Abstract. An uncertainty cascade model applied to stream flow forecasting seeks to evaluate the different sources of uncertainty of the complex rainfall-runoff process. The current trend focuses on the combination of Meteorological Ensemble Prediction Systems (MEPS) and hydrological model(s). However, the number of members of such a HEPS may rapidly increase to a level that may not be operationally sustainable. This paper evaluates the generalization ability of a simplification scheme of a 800-member HEPS formed by the combination of 16 lumped rainfall-runoff models with the 50 perturbed members from the European Centre for Medium-range Weather Forecasts (ECMWF) EPS. Tests are made at two levels. At the local level, the transferability of the 9th day hydrological member selection for the other 8 forecast horizons exhibits an 82 % success rate. The other evaluation is made at the regional or cluster level, the transferability from one catchment to another from within a cluster of watersheds also leads to a good performance (85 % success rate), especially for forecast time horizons above 3 days and when the basins that formed the cluster presented themselves a good performance on an individual basis. Diversity, defined as hydrological model complementarity addressing different aspects of a forecast, was identified as the critical factor for proper selection applications.

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

1 Introduction

The competency of probabilistic forecast to encompass the many sources of uncertainty in Hydrological Ensemble Prediction Systems (HEPS) has already been demonstrated (Roulin, 2007; Rousset et al., 2007; Velázquez et al., 2011). Yet the simultaneous consideration of the uncertainty associated with both the meteorological inputs and the structural and parametric configuration of the hydrological models can lead to systems consisting of too many members to be computationally and operationally implementable.

Nonetheless, reliability as a crucial feature in ensemble forecasting may be achieved through the uncertainty cascade model as proposed by Pappenberger et al. (2005). This approach states that the output uncertainty of a hydrological model is affected by several components: uncertainty from the meteorological data used to drive the model, initialization uncertainty (i.e. the initial state of the model), and the model uncertainty (from parameter identification to model conceptualization).

Combining information derived from the many Meteorological Ensemble Prevision Systems (MEPS) is an avenue that has been shown to improve early flood warning systems (He et al., 2009) – the THORPEX Interactive Grand Global Ensemble (TIGGE) (Bougeault et al., 2010) favours this new opportunity. Moreover, if the parametric uncertainty of hydrological models is assessed under the principle of equifinality (Beven and Binley, 1992) and if the structural uncertainty is tackled through a multi-model approach, the number of scenarios in the uncertainty cascade model may rapidly

turn out to be quite large. Simplification of such a HEPS thus becomes a mandatory step from an operational standpoint.

In such a context, the hydrological and meteorological community has focused their efforts on many lines of simplification. For instance, Pappenberger et al. (2005) evaluated 10-day ahead rainfall forecasts, consisting of one deterministic, one control, and 50 ensemble forecasts, into a rainfall-runoff model (LisFlood) for which parameter uncertainty was represented by six different parameter sets identified through a Generalized Likelihood Uncertainty Estimation (GLUE) analysis and functional hydrograph classification. Raftery et al. (2005) proposed the Bayesian Model Average methodology (BMA) as a means for the statistical post-processing of the forecast ensembles derived from numerical weather prediction models. The BMA predictive probability density function (PDF) is a weighted average of the PDFs centred on the bias-corrected forecasts from a set of different models. The weights assigned to each model reflect that model's contribution to the forecasting skill over a training period (Vrugt et al., 2006). In line with that, Vrugt et al. (2008) proposed evaluating BMA weights with the DiffeRential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) algorithm.

Other studies identified the meteorological forecasts as the most uncertain component of the cascade model (Todini, 2004; Pappenberger et al., 2005; Jaun et al., 2008), triggering interest in novel member selection techniques. For example, Marsigli et al. (2001); Molteni et al. (2001) and Jaun et al. (2008) select MEPS members based on lagging ensembles, and derived representative members through hierarchical clustering over the domain of interest. Ebert et al. (2007) analysed the relation between the atmospheric circulation patterns and extreme discharges to select representative members of MEPS. Finally, Xuan et al. (2009) establish, in a deterministic way ("best match" approach), the location of the forecast that is the most similar to the rainfall pattern of the catchment.

In the companion paper, Brochero et al. (2011) described in depth the hydrological member selection methodology adopted here: a Backward Greedy Selection combined with Cross Validation, hereafter BGS-CV, to retain the uncertainty properties of a 800-member HEPS derived from the fifty members of the European Center for Medium-range Weather Forecasts (ECWMF) propagated through sixteen simple lumped hydrological models.

Another aspect of particular interest in the evaluation of probabilistic forecast, and therefore in hydrological member selection, is the identification of a pertinent criteria set. In conventional forecasting, i.e. when confronting an observation against a single prediction, it is now generally accepted that the calibration of hydrological models should be approached as a multi-objective problem (Gupta et al., 1998, 1999; Yapo et al., 1998; Wagener et al., 2001; Confesor and Whittaker, 2007). Probabilistic forecasting is not different in that regard. In fact, the complexities of confronting an observation against an ensemble of predictions calls for a variety of criteria, here called scores, that specifically focus on one or more characteristics of the probabilistic sets. So, to assess these properties, several statistical measures should be considered concurrently (Wilks, 2005; Cloke and Pappenberger, 2009). Few studies have experimented hydrological member selection from a multi-score point of view.

Vrugt et al. (2006) posed the BMA inverse problem in a multi-objective framework, examining the Pareto set of solutions between the Continuous Ranked Probability Score (CRPS), the Mean Absolute Error (MAE), and the Ignorance Score with the AMALGAM method (Vrugt and Robinson, 2007). In continuity with that, the companion paper shows that a combined criterion which groups various characteristics of the probabilistic forecast is adequate to guide the selection of hydrological members with BGS-CV method. At this point, it is important to note that the BGS-CV method offers the possibility of combining results from different studies, which is highlighted as one of the aspects related to the improvement of HEPS (Cloke and Pappenberger, 2009).

In this paper we evaluate the generalization of a simplification scheme of the complex 800-member HEPS presented in Sect. 2. A brief description of the selection of hydrological members is given in Sect. 3. The generalization methodology, with local and regional orientation, is explained in Sect. 4. Thus, we test the hydrological members' selection obtained in sixteen catchments for the 9-day lead time, for the other 8 lead times. Additionally we evaluate the ability to extrapolate the selections to neighbouring catchments. Finally we present the integration of results from different catchments within a regional framework. Results and discussion are gathered in Sect. 5, while conclusions and a guideline for future work are given in Sect. 6.

2 HEPS configuration and catchment locations

As already mentioned, the 800-member HEPS at hand is the propagation of 50 perturbed members from the ECMWF EPS, that are a priori assumed to be equally likely (Gouweleeuw et al., 2005), through sixteen lumped hydrological models. Details of the HEPS conformation can be found in Brochero et al. (2011).

This HEPS was implemented over 28 French catchments, representing a large range of hydro-climatic conditions (Fig. 1), and evaluated over a 17-month period. The main characteristics of these catchments are summarized in Table 1. Henceforth each basin in Table 1 will be identified only with the first three characters.

It is important to note that this study focuses on evaluating the probabilistic hydrological forecasting from a cooperative point of view seeking diversity in the final hydrological members' selection, i.e. that each member acts as a complement to the others. This clarification is relevant in order to avoid misinterpretation of competitiveness in the different

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)	P (mm)	ET (mm)	Q (mm)	Catchment codes	Area (km^2)	P (mm)	ET (mm)	Q (mm)
A6921010	2780	3.04	1.79	1.18	M0680610	7380	2.04	1.93	0.56
A7930610	9387	2.78	1.80	1.21	O3401010	2170	3.19	1.80	1.90
A9221010	1760	2.49	1.83	0.91	Q2593310	2500	2.52	2.24	0.75
B2130010	2290	2.57	1.80	0.87	U2542010	4970	3.63	1.75	1.88
B3150020	3904	2.58	1.80	1.09	**A7010610**	6830	2.99	1.78	1.46
H2482010	2982	2.31	1.89	0.84	**H6221010**	2940	2.50	1.83	0.92
H3621010	3900	1.98	1.95	0.45	**M3600910**	3910	2.31	1.88	0.80
H5321010	8818	2.41	1.85	0.93	**K1341810**	2277	2.65	1.89	1.02
J8502310	2465	2.36	1.89	0.81	**M1531610**	7920	1.85	1.95	0.36
K1773010	1465	2.65	1.94	1.07	**P7001510**	1863	2.88	2.08	1.19
K7312610	1712	2.13	2.01	0.68	**P7261510**	3752	2.65	2.14	0.87
M0421510	1890	2.04	1.89	0.62	**U2722010**	7290	3.63	1.79	2.07

P: precipitation, ET: potential evapotranspiration, Q: flow. For the distinction of the basins used in training and testing, the latter are highlighted in bold.

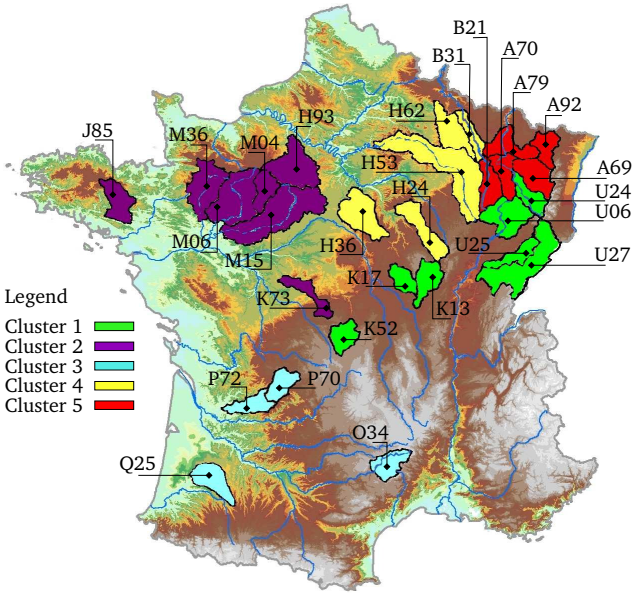

Fig. 1. Location of the catchments grouped by clusters. Some of them have been used in the BGS-CV process, while the others have been used for extrapolation. The colours identify the five regions evaluated in this paper.

conceptualizations of the sixteen hydrological models used. It should be clear that the comparison would not be fair because some models such as the GR4J were specifically devised for the catchment scale, whereas others have suffered a series of substantial changes bringing them to a lumped state.

3 Hydrological members' selection

The hydrological members' selection is described in detail in the companion paper (Brochero et al., 2011). It is executed basically in three steps:

Step 1: Resampling with a variation of the k-fold cross-validation. Because the series are short-length (500 forecast-observation pairs), a rigorous application of the selection requires evaluating different types of events in the training, validation, and test sets. Thus, the process of selecting data follows a k-fold cross-validation technique.

Step 2: Backward greedy selection. Optimization for a preselected number of hydrological members (nmim) relies on the Combined Criterion (CC), which brings together the Continuous Ranked Probability Score (CRPS), the IGNorance Score (IGNS), the Mean Squared Error (MSE) evaluated in the Reliability Diagram (RD), the δ ratio evaluated in the rank histogram and the MeDian of Coefficients of Variation (MDCV):

$$\mathrm{CC} = w_1 \frac{\overline{\mathrm{CRPS}}_{\mathrm{se}}}{\overline{\mathrm{CRPS}}_{\mathrm{ie}}} + w_2 \frac{z_1 - \overline{\mathrm{IGNS}}_{\mathrm{se}}}{z_1 - \overline{\mathrm{IGNS}}_{\mathrm{ie}}} + w_3 \frac{\mathrm{RD}_{\mathrm{MSE}_{\mathrm{se}}}}{\mathrm{RD}_{\mathrm{MSE}_{\mathrm{ie}}}} + w_4 \frac{\delta_{\mathrm{se}}}{\delta_{\mathrm{ie}}} + w_5 \frac{z_2 - \mathrm{MDCV}_{\mathrm{se}}}{z_2 - \mathrm{MDCV}_{\mathrm{ie}}}, \quad (1)$$

where the result of each criterion in the selection ensemble (se subscript) is divided by the criterion calculated on the initial 800-member HEPS (ie subscript). z_{m} represents some thresholds to orient a direct minimization; w_{cp} are the weights assigned to each component. Here, the weight assigned to the reliability (the critical factor) is twice that of the other factors, which have a unit weight.

The mechanism of member elimination begins with all members, removing at each step the hydrological member

that, when it is removed, has the greater impact on the training set error (i.e. minimises training error the most).

Step 3: Combination of results. It is highly likely that variability in the five experiments configured in the first step will lead to different solutions. An integration mechanism is thus needed for a global solution for each catchment. The importance of each hydrological member 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.

Attention is given to the interpretation of results of the final hydrological members' selection, because if the HEPS is 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, in the simplest case, 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.

Note that the CC could be used to compare the performance of the hydrological 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 some feature of the 800-member set. Hereafter, this particular condition of unit weights in the CC will be called the normalized sum (NS). This distinction is important to display the priority that can be defined a priori to any feature in the hydrological members' selection training with BGS-CV.

It is important to note that the normalized sum may hide some deterioration compensated by one or more other metrics. It is thus necessary to accompany this measure with the results of each of its components, for a collective analysis. In this sense, the analysis is facilitated if each component is associated with an index that reflects the gain or loss of the selected subset over the initial 800-member set:

$$\text{Gain}(\%) = 100 \times \frac{\text{Score}_{\text{ie}} - \text{Score}_{\text{se}}}{|\text{Score}_{\text{ie}}|}. \quad (2)$$

Note that the absolute value is used in the denominator for accounting for possible negative values of the IGNS. The MDCV function further requires the inversion of the numerator, because the purpose of this metric is to maximize the dispersion of the selected subset of hydrological members.

4 Generalization test methodology

The generalization ability of a hypothesis, namely, the quality of its inductive bias, can be measured if there is access to data outside of the training process. The methodology proposed in the companion paper simulates this by dividing the training set into two parts. One part is used for training (i.e. to find a hypothesis) and the remaining part (validation set) is used to test the generalization ability. Nevertheless, if it is necessary to report the error to approximate the expected selection error, it is compulsory to make use of a third set, a test set, sometimes also called the publication set, containing examples not used in training or validation (Alpaydin, 2010; Hudson and Demuth, 2011).

Thus, the method of combining results, based on the mean rank of elimination, is derived on the use of all series as a means of optimizing the use of information in a short-length series (seen from the point of view of the periodicity of the hydrological cycle). However, results of this procedure can be conceived as indicators of a relative performance or otherwise as an optimistic estimate of the hydrological members' selection process (Diamantidis et al., 2000).

Figure 2 shows the generalization or test methodology of the hydrological members' selection at two levels: the local focuses on the extrapolation of results to different FTH within the same catchment and another named regional, while the regional level tests the temporal and spatial performance in nearby catchments, or under a broader perspective on the integration of regional results.

4.1 Extrapolation to different forecast time horizons

The hydrological members' selection is performed on the results of sixteen hydrological models fed with the 9th day FTH of the ECMWF MEPS. Thus, the application of this selection of members for the other eight FTHs (1 to 8 days) is a first level test. It has to be stressed that the idea of simplifying the HEPS is only valuable if the hydrological member selection is invariant in regard to the FTH. However, one may always argue that the assumption of statistical independence between the test and training data, principally for FTHs next to the ninth, may be somewhat questionable.

4.2 Extrapolation to a different catchment

Transferring selected members to a neighbouring catchment, and even further to a different FTHs, constitutes a rigorous test of the generalization ability of results at both the temporal and spatial scales. The choice of the second catchment could first be viewed as a simple nearest neighbour problem. However, we explored the possibility of regionalizing the selection of hydrological members from the grouping of catchments by k-means clustering and subsequent integration of results to select the most representative hydrological members.

4.2.1 k-means clustering

The k-means clustering algorithm is used to define 5 regions based on the combination of different characteristics of the catchments, such as geographic location of the basin outlet, minimum, mean, and maximum precipitation,

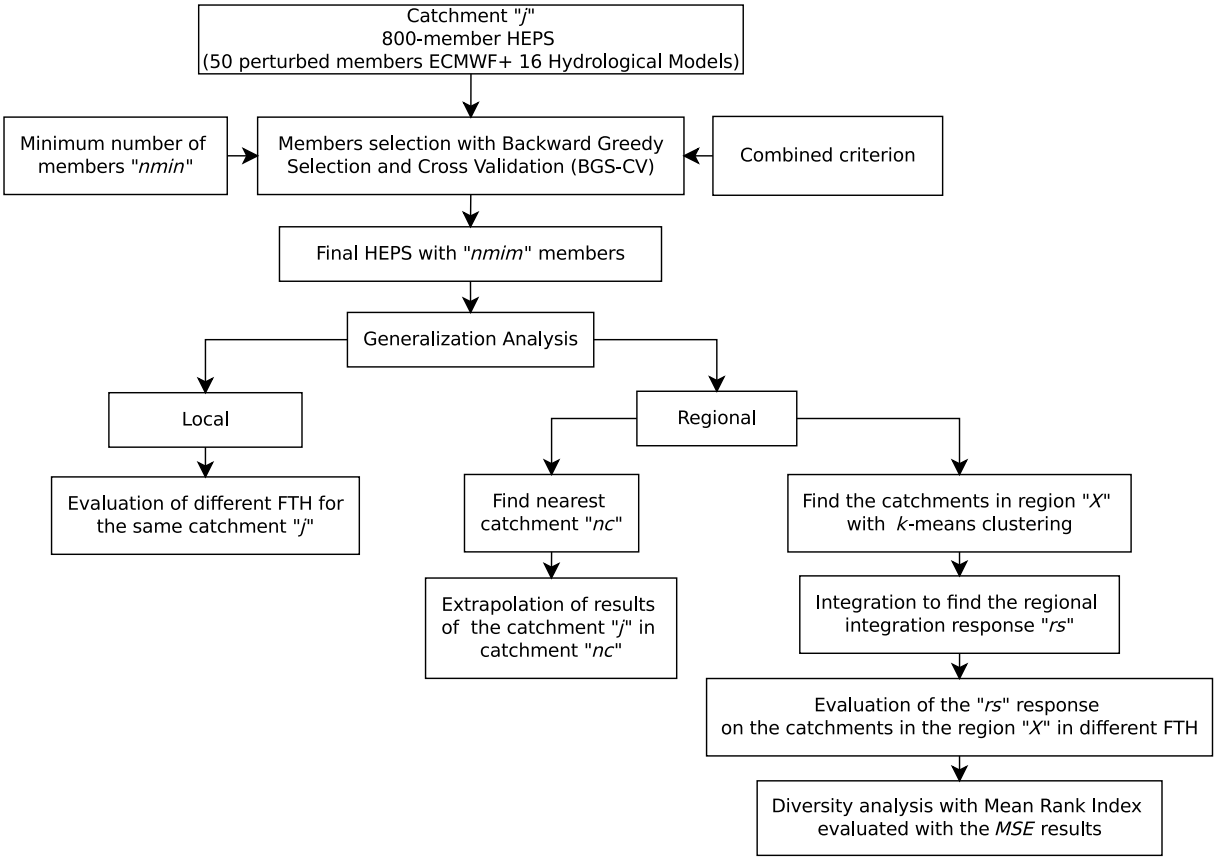

Fig. 2. Generalization test methodology for the hydrological members' selection found with BGS-CV.

evapotranspiration and flow (see Table 1). Of course, every possible combination of features will yield a different distribution of catchments that will be evaluated through the integration mechanism that will be presented in Sect. 4.2.2.

It is convenient at this point to define some notation to describe the assignment of catchments to a region or cluster. The property set $\boldsymbol{x}^l$ for each catchment is introduced into a corresponding set of binary indicator variables $b_k^l \in \{0,\ 1\}$, where $k = 1, ..., K$ describe which of the K clusters the catchment l or its property set $\boldsymbol{x}^l$ is assigned to, so that if $\boldsymbol{x}^n$ is assigned to cluster k then $b_k^n = 1$, and $b_j^n = 0$ for $j \neq k$. Then an objective function is given by:

$$J = \sum_{l=1}^{L} \sum_{k=1}^{K} b_k^l \|\boldsymbol{x}^l - \boldsymbol{m}_k\|^2, \tag{3}$$

which represents the sum of the squares of the distances of each catchment to its assigned vector $\boldsymbol{m}_k$. The goal is to find values for the b_k^l and the $\boldsymbol{m}_k$ so as to minimise J. Then the iterative application of Eq. (3) leads to the following procedure for finding the $\boldsymbol{m}_k$ centres:

Algorithm 1 k-means pseudo-code

1. Define the number of clusters (K), (here $K = 5$)
2. Initialize randomly centres $\boldsymbol{m}_k \quad (k = 1, \cdots, K)$

repeat
  for all $\boldsymbol{x}^l, \quad l = 1, \ldots, L$ **do**

$$b_k^l = \begin{cases} 1 & l = \text{argmin}_k \|\boldsymbol{x}^l - \boldsymbol{m}_k\| \\ 0 & \text{otherwise} \end{cases}$$

  end for
  for all $\boldsymbol{m}_k$ **do**

$$\boldsymbol{m}_k = \frac{\sum_{l=1}^{L} b_k^l \boldsymbol{x}^l}{b_k^l}$$

  end for
until $\boldsymbol{m}_k$ converges

Details of the k-means clustering algorithm are given by Bishop (2006). Figure 1 shows an example of k-means clustering results based only on the geographic location of the basin outlets.

4.2.2 Regional integration mechanism

The hydrological members' selection integration for region X, consisting of C catchments, is defined from matrix

S, which has C columns with $nmin$ rows representing the most $nmin$ important hydrological members as assessed by the mean rank of elimination ($\overline{R}$) for each catchment. Then, the process of forming a regional solution **rs** with q members is based on taking the most important members of each catchment without replacement until the number of members in **rs** is equal to the desired q, i.e. each member cannot be selected again later. Algorithm 2 details this procedure:

Algorithm 2 Regional integration mechanism pseudo-code

1. Determine the C catchments in the X region (clustering process).
2. Define the matrix $\mathbf{S} = \{\mathbf{s_1}, \mathbf{s_2}, \cdots, \mathbf{s_C}\}$
3. Establish the number of hydrological members q in the regional solution **rs**
4. Initialize **rs** = {}, $h = 0$ and $i = 1$

repeat
  for $j = 1, \ldots, C$ **do**
    if $S_{i,j} \notin$ **rs then**
      rs = **rs** + $S_{i,j}$
      $h = h + 1$
    end if
  end for
  $i = i + 1$
until $h > q$

4.2.3 Diversity evaluation

The participation of hydrological models in the regional selection stresses the importance of the integration of models with different characteristics. To view this in a deterministic framework, an index based on the performance rank assigned to each model in each catchment is proposed. Its calculation is summarized as follows:

- MSE for catchment i and hydrological model j is first calculated ($\mathrm{MSE}_{i,j}$).
- Performances are next ranked for each catchment, leading to $\mathrm{PR}_{i,j}$, for which the model with the lowest MSE is assigned the rank PR = 16 and the highest MSE is assigned the rank PR = 1.
- Finally, the mean rank of performance or rank index RI_j for each model is estimated based on the results of all 28 basins:

$$\mathrm{RI}_j = \frac{1}{28} \sum_{i=1}^{28} \mathrm{PR}_{i,j}. \quad (4)$$

5 Results and discussion

In the companion paper we have shown the high performance of the 800-member HEPS for the 9th day FTH. However, as one of the objectives of this paper is to show the transferability of the hydrological members selections to other FTHs, it is necessary to show the performance of the 800-member HEPS in such scenarios to clearly establish our point of reference concerning the quality of the hydrological members' selection. In the companion paper we also stressed that on the δ ratio and the $\mathrm{RD_{MSE}}$ scores rest the main advantages of the 800-member HEPS.

Figure 3 shows the HEPS' behaviour with different set-up and different FTH. Results focus on the reliability ($\mathrm{RD_{MSE}}$) and the ensemble consistency (δ ratio) for two schemes formed from sixteen hydrological models, one led by the deterministic ECMWF forecast and the other by the 50 perturbed members from ECMWF EPS. The results in Fig. 3, expressed in terms of interquartile range (iqr) and median, are due to the grouping of the scores obtained in the 28 basins evaluated here. Note that the δ ratio and $\mathrm{RD_{MSE}}$ scores are directly comparable since their scale is independent of the measured variable.

Figure 3 illustrates that the 800-member HEPS advantages becomes apparent after the 4th day FTH. According to 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. Furthermore the spread in the results of both the $\mathrm{RD_{MSE}}$ and the δ ratio, viewed from the interquartile range, shows two features: first, the 16-member HEPS has greater dispersion than the 800-member HEPS, and second, the 800-member HEPS spread diminishes with increasing lead time.

5.1 Selection process

The optimal number of hydrological members simplifying the HEPS was identified in the companion paper to be between 50 and 100, depending on the catchment. In most cases a significant gain with respect to the balance of the different criteria evaluated from the initial 800-member HEPS was then achieved. Results presented in this section are based on a selection of 50 hydrological members.

Table 2 presents the results of the 50-member selection based on the combined criterion, for 16 catchments uniformly distributed over France (see Fig. 1). The overall performance is the normalized sum given by Eq. (1) with unit weights definition, values lower than 5 indicate a selection of higher performance than the base set of 800 hydrological members, and values greater than 5 indicate the detriment of any feature of the 800-member set.

To facilitate the visualization of results, Table 2 shows the performance of one selection oriented with the hydrological members' proportion found in the BGS-CV process. However, Fig. 4 and 6 present an analysis that shows the performance of multiple selections oriented by the BGS-CV solution and a random choice of the meteorological members from ECMWF.

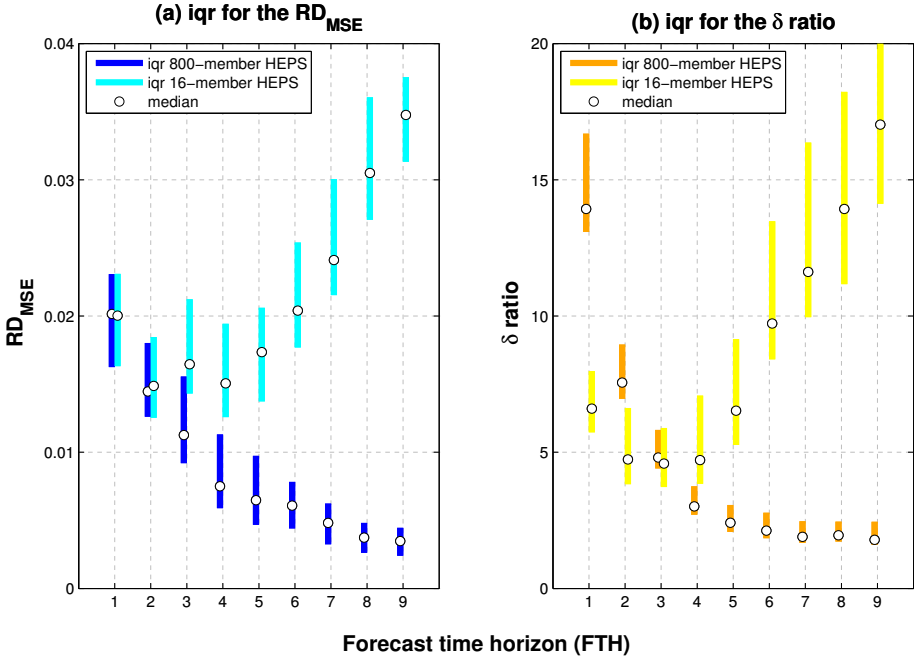

Fig. 3. Interquartile range (iqr) of RD_{MSE} and δ ratio assessed in the 28 catchments under two HEPS schemes: 16-member HEPS (16 hydrological models are driven by the deterministic forecast from ECMWF) and the 800-member HEPS (16 hydrological models are driven by the 50-perturbed member forecast from ECMWF).

Table 2 shows that in all cases the normalized sum (NS) is always lower than 5, indicating the superiority of the 50-member HEPS, even after a size reduction equivalent to a 94% compression of the initial 800-member HEPS (i.e. 750 members are removed).

Based on the gain score formulation (Eq. 2), it is noted that for the 50-member selection, the CRPS and the MDCV show low variability with mean gain indexes around 2 % and 5 %, respectively.

RD_{MSE} shows a minimum gain of 49 % (catchment B21) and a maximum gain of 87 % (catchment K17), reflecting the emphasis given to this property in the formulation of the combined criterion used in the selection process. With respect to the IGNS, index gains between −5 % and 27 % (excluding the catchment B21) reflect an acceptable behaviour.

Finally, the δ ratio is the score more difficult to minimise or preserve; a positive index gain was obtained for only 25% of the cases (4/16), while the spread ranged from −39 % for catchment H53 to 27 % for catchment B31. Note that the δ ratio has an inverse relationship with the number of members of the selection, so it directly follows the complexity in maintaining the value of the initial 800-member HEPS in the selection process. Nonetheless, it was shown in the companion paper that the δ ratio is the best individual metric for the hydrological members' selection.

5.2 Generalization test

5.2.1 Local analysis

For operational convenience, it is fundamental that the 50 hydrological members selected for the 9th day FTH are also appropriate for the 8 previous time horizons. A lack of transferability of the selected members would considerably reduce the actual level of achieved simplification.

Here, temporal transferability is first evaluated comparing the normalized sum of the performance of the 50-member selection to the 800-member performance, whose normalized sum equals 5 in all cases. It is then compared to the performance of 200 random combinations with 50 hydrological members, in order to evaluate if any good performance may only be attributable to chance. Results for the 8 first FTHs and sixteen basins are gathered in box-plot diagrams (Fig. 4), where the performance of the solution is based on random experiments that are set-up following these guidelines:

- Experiments considering the participation of hydrological models found with BGS-CV: taking into account the participation of hydrological models to assign to each model a number of members chosen randomly from ECMWF EPS.

- Without considering any "a priori" participation of hydrological models: hydrological members are picked randomly from the initial 800-member HEPS.

Table 2. Selection of 50 hydrological members based on combined criterion and the BGS-CV process on the 9-day FTH. Beside each score is presented the gain index evaluated by Eq. (2). NS represents the normalized sum (Eq. 1 with unit weights). NHM indicates the number of hydrological models participating in the selection. RD_{MSE} values are expressed on a 10^{-3} basis.

Catchment Codes	Scores				MDCV function	NS	NHM
	CRPS	RD_{MSE}	δ	IGNS			
A69	0.284 (+0 %)	1.3 (+81 %)	1.5 (+18 %)	0.67 (+14 %)	0.39 (+5 %)	4.0	9
800 members	0.284	7.0	1.8	0.78	0.37	5.0	16
A79	0.254 (+3 %)	1.5 (+69 %)	3.6 (−11 %)	0.34 (+23 %)	0.41 (−1 %)	4.4	11
800 members	0.263	5.1	3.3	0.44	0.41	5.0	16
A92	0.183 (+4 %)	0.3 (+86 %)	2.3 (−28 %)	−0.42 (+27 %)	0.57 (+0 %)	4.4	11
800 members	0.192	2.4	1.8	−0.33	0.57	5.0	16
B21	0.232 (−1 %)	1.2 (+49 %)	2.6 (−16%)	−0.18 (−38%)	0.63 (+9 %)	4.6	13
800 members	0.230	2.4	2.2	−0.29	0.57	5.0	16
B31	0.134 (+1 %)	1.3 (+72 %)	2.0 (+27 %)	−0.84 (−5 %)	0.24 (+7 %)	4.0	11
800 members	0.135	4.5	2.7	−0.88	0.22	5.0	16
H36	0.157 (+2 %)	0.7 (+80 %)	2.0 (−37%)	−1.02 (+2 %)	0.36 (−1 %)	4.5	14
800 members	0.161	3.5	1.5	−0.99	0.37	5.0	16
H53	0.165 (+3 %)	1.9 (+74 %)	4.3 (−39 %)	−0.76 (+8 %)	0.36 (+8 %)	4.6	11
800 members	0.171	7.4	3.1	−0.71	0.33	5.0	16
H24	0.180 (+2 %)	2.2 (+68 %)	3.8 (−32%)	−0.82 (+9 %)	0.37 (+6 %)	4.6	12
800 members	0.185	7.1	2.9	−0.76	0.35	5.0	16
K17	0.205 (+4 %)	0.5 (+87 %)	1.8 (−9 %)	−0.73 (+12 %)	0.38 (−2 %)	4.2	12
800 members	0.213	3.6	1.7	−0.65	0.39	5.0	16
U25	0.290 (+0 %)	0.9 (+74 %)	2.6 (−1 %)	−0.40 (+13 %)	0.38 (+7 %)	4.2	14
800 members	0.289	3.4	2.5	−0.36	0.35	5.0	16
J85	0.159 (+2 %)	0.4 (+80 %)	1.7 (−5 %)	−1.00 (+2 %)	0.40 (+8 %)	4.2	14
800 members	0.163	2.2	1.7	−0.98	0.37	5.0	16
K73	0.160 (+3 %)	0.9 (+70 %)	2.1 (−5 %)	−0.93 (+0 %)	0.38 (+9 %)	4.3	11
800 members	0.165	3.1	2.0	−0.93	0.35	5.0	16
M04	0.158 (+1 %)	0.6 (+68 %)	1.6 (−2 %)	−0.98 (−1 %)	0.37 (+2 %)	4.3	13
800 members	0.160	1.7	1.6	−0.99	0.37	5.0	16
M06	0.153 (+4 %)	0.3 (+79 %)	1.6 (−4 %)	−1.09 (+6 %)	0.39 (+1 %)	4.2	13
800 members	0.159	1.4	1.5	−1.03	0.38	5.0	16
O34	0.166 (+2 %)	1.0 (+71 %)	1.6 (+1 %)	−0.91 (+5 %)	0.37 (+3 %)	4.2	13
800 members	0.169	3.5	1.6	−0.86	0.36	5.0	16
Q25	0.159 (+3 %)	0.6 (+73 %)	1.1 (+22 %)	−0.94 (−5 %)	0.39 (+4 %)	4.0	12
800 members	0.163	2.1	1.4	−0.98	0.37	5.0	16

Figure 4 shows that the median of 200 evaluations of 50-member HEPS for the 9th day FTH is superior to the 800 reference members in 82 % of the evaluated cases. It is also noteworthy that in only 11 % of the cases (14/128) the 50 hydrological members selected oriented by the BGS-CV process lead to a worse performance than the 25 percentile of 200 random combinations test. Note that all these cases correspond to short lead times (1 to 3 days), remarkably in the 2-day FTH. Another aspect that draws attention is the low dispersion of the BGS-CV selections represented by the interquartile range, highlighting the importance of the hydrological models participation in the selection process.

Figure 4 also shows that the selection slowly loses efficiency as it moves away from the 9th day FTH. It also detects a systematic deficiency for catchment A69 and to a lesser extent for catchment B21. Nonetheless, these results are very encouraging.

5.2.2 Regional analysis

As described in Sect. 4.2, the regional analysis assesses the generalization ability of the hydrological member selection for a specific catchment with respect to another one. For example, Fig. 5 explores the transferability of the 50-member

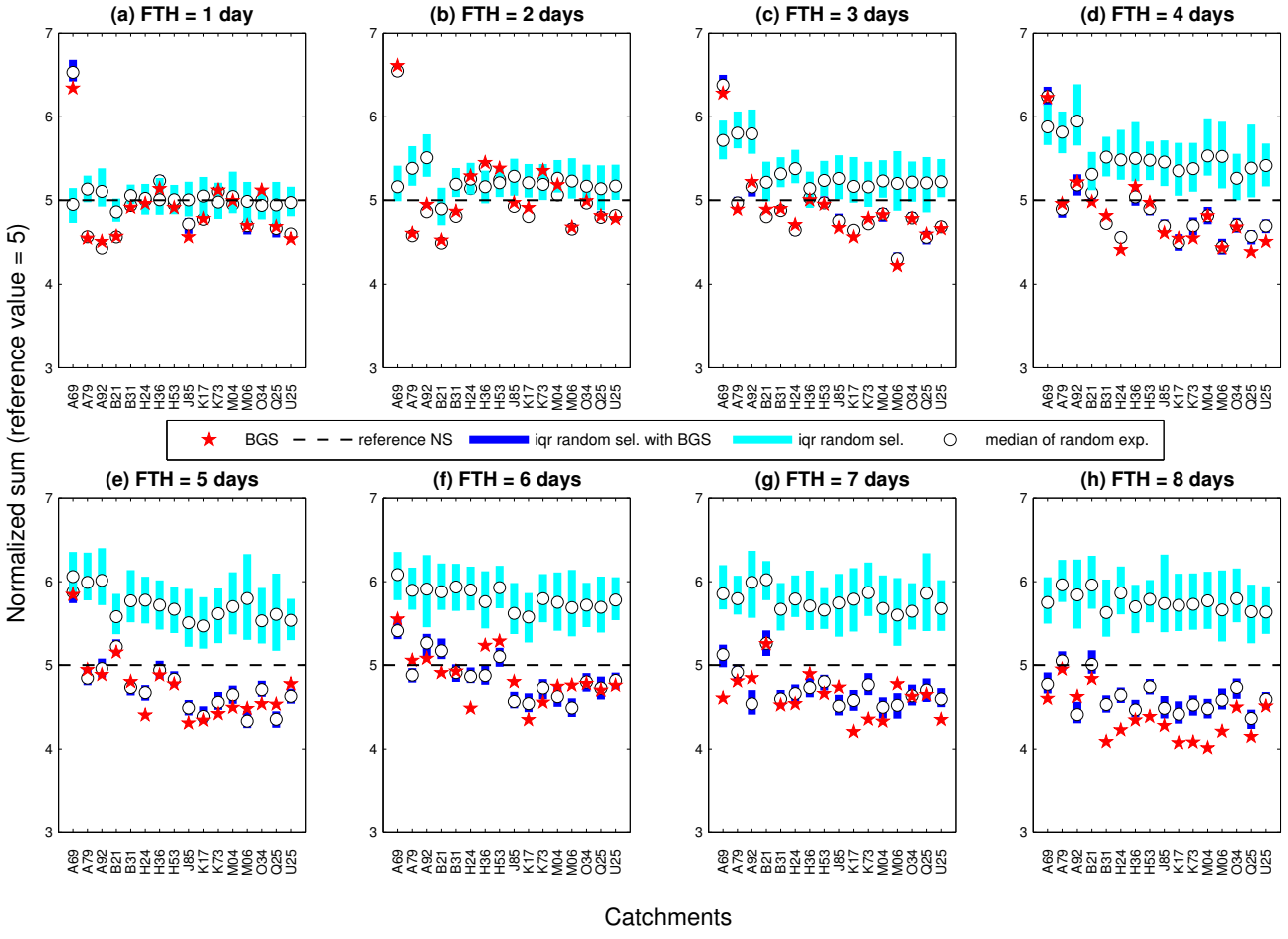

Fig. 4. Evolution of the normalized sum (NS) to evaluate the response sensibility with regard to the interquartile range (iqr) of 200 random experiments in different FTHs following these guidelines: (1) Considering the participation of hydrological models found with BGS-CV (vertical blue bars), and (2) Without regard to any "a priori" participation of hydrological models, i.e. completely random selection (vertical cyan bars).

selection obtained for catchment Q25 for a lead time of 9 days to catchment P72 for the 4-day lead time.

In general, Fig. 5 shows that results for the different scores are very similar for the 800-member and 50-member sets, except for the RD_{MSE} where the gain index reaches 51 %. In particular, Fig. 5a shows that the 50-member CRPS equals the reference value. Taking into account that the CRPS generalizes the mean absolute error (CRPS) 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 on 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 5b shows that the trimmed mean IGNS for the 50-member HEPS (-1.65) also presents an improvement over the initial value (-1.59). Regarding the time structure of the IGNS, it is observed that both the 50-member and 800-member series have high values for extreme events, showing a systemic problem in terms of ensemble bias.

With regard to the reliability diagram, Fig. 5c shows a considerable agreement improvement (4.21×10^{-3}) over the initial value (8.67×10^{-3}). This gain in reliability may be traced back to the optimization criterion used: the combined criterion (CC) that focuses primarily on system reliability as defined by its weights. Similarly, Fig. 5d reveals that the rank histograms have a nearly uniform distribution, even if the first and the last rank reflect a slight bias. Those imperfections demonstrate the difficulty inherent in minimizing the δ ratio.

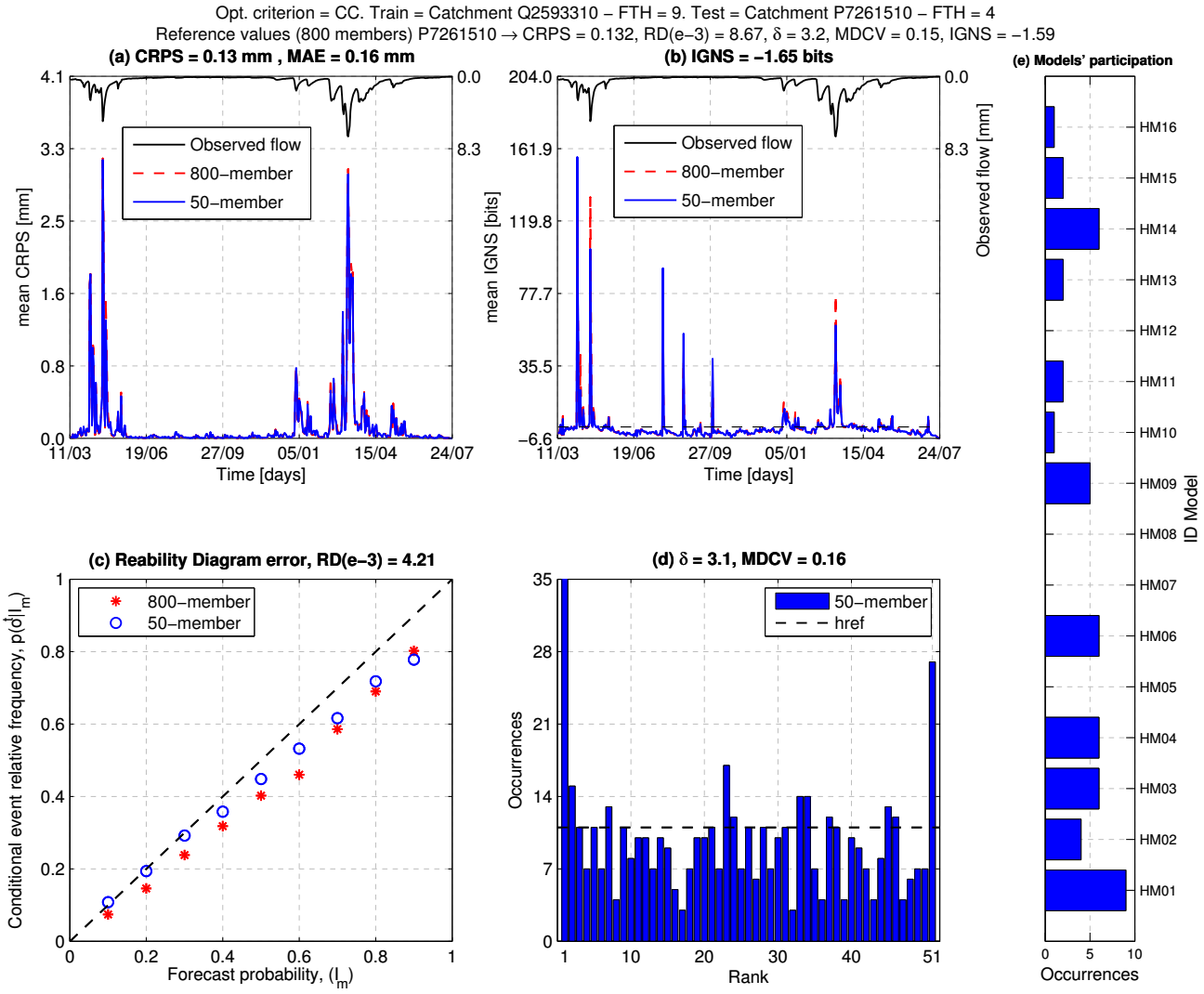

Fig. 5. Comparison between the initial ensemble (800 members) and the ensemble selected (50 members) for a lead time of 9 days. **(a)** Figure above: observed flow; figure below: CRPS (x-axis formatted as: day/month). Note the correspondence between higher observed flows and higher CRPS. **(b)** Figure above: observed flow; figure below: IGNS (x-axis formatted as: day/month). **(c)** Reliability diagram error (MSE based on vertical distances between the points). **(d)** Rank histogram for the 50 hydrological members selected. 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 50 hydrological members.

Figure 5e illustrates the occurrence of each lumped model within the 50-member hydrological ensemble. A wide selection of models alone could justify the multi-model approach advocated here. Results show that 12 models out of 16 were selected in this case, and that no models were selected more than 9 times. Knowing that these models are not of equal quality with regards to MSE performance, for instance, this suggests that the selection favoured a diversity of errors. At the end of the selection process, the MDCV has slightly increased, from 0.15 to 0.16.

To display an overview of the extrapolation of results to the nearest basin, Fig. 6 shows such an assessment under the same selection schemes analysed in Fig. 4, i.e. analyzing various combinations considering or ignoring the solution found with BGS-CV. Although in general the solution found with BGS-CV (red stars in Fig. 5) exhibits the highest performance, given the interchangeability of MEPS members as input of hydrological models, solutions focus on comparing the median of the evaluations that follow the participation of hydrological models found with BGS-CV.

Additionally, it is clear that the dispersion of the BGS-CV selections, evaluated from the interquartile range, is less than that assessed in completely random selections. Likewise, the median of the BGS-CV selections is usually better than the reference set of 800 hydrological members, which corresponds to a normalized sum equal to 5.

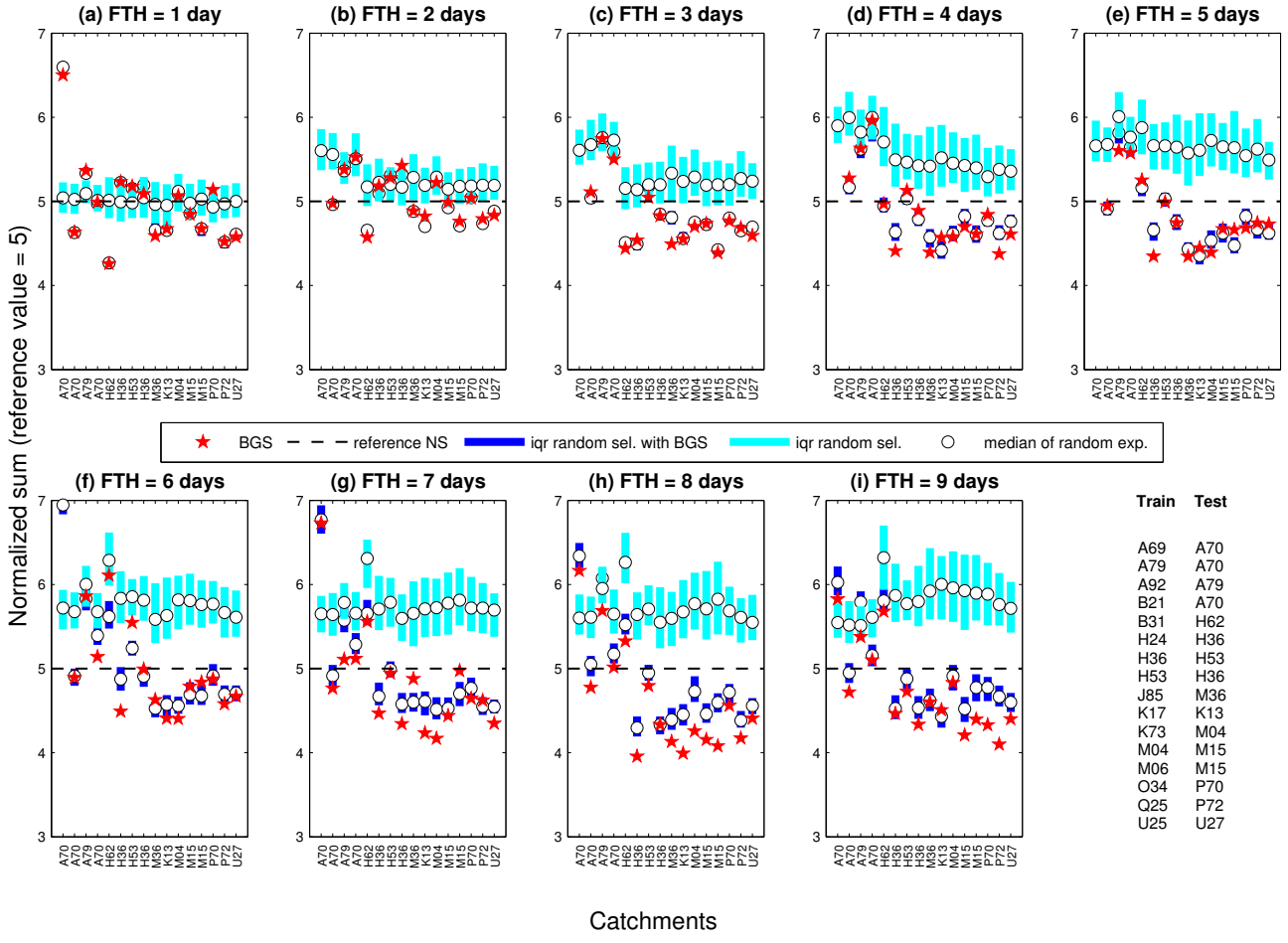

Fig. 6. Evolution of the normalized sum (NS) to evaluate the response sensibility of the extrapolation of results in the nearest catchments. Each vertical bar represents the interquartile range (iqr) of 200 combinations of 50 hydrological members under the following guidelines: the combination is oriented with the same proportion of hydrological models found with BGS-CV (blue vertical bars), the selection is completely random (cyan vertical bars). Note the deficiency of the selections' extrapolation in basin A69 to basin A79, notably for early lead times (2 to 5 days); these results do not appear in the figure because they are above 7.

Another aspect that stands out in the extrapolation is the recurrent deficiency of selection in the basins A69, A92, B21 and B31, i.e. 25 % of the basins tested. Initially, the deficiency in these basins at different FTHs shows the temporal consistency of HEPS, as if the deficiency of a given selection disappears at certain lead times would reflect inconsistency of the selection task.

Likewise, it is noteworthy that extrapolation of the results of selection in the basins A69, A79 and B21 are tested in the basin A70; however, only the results of the hydrological members' selection in the basin A79 show considerable efficiency in most of the FTHs evaluated. It follows that while the geographic location of the basin outlet is an acceptable feature to run the extrapolation of results, it is not sufficient in some cases, which requires a more detailed analysis of other factors such as hydrometeorological and physiographic characterization of the basins.

The regional analysis that integrates several basins, which seeks to identify features that facilitate the combination of results, revealed that geographical location is the most important feature, followed by evapotranspiration, precipitation and flow, when the normalized sum is used to evaluate the gain. However, consideration of the geographic location was found to be sufficient. Such results are presented in Table 3, after application of the k-means algorithm and the regional integration procedure already described in Sect. 4.2.2.

Note that the results in Table 3 are due to the evaluation of one combination of MEPS members randomly chosen, but respecting the participation of hydrological models found with BGS-CV. Additionally, for purposes of extrapolation of results, in the evaluation of the normalized sum, a threshold z_1 equal to -4 was used, because in the first lead times (1 to 4 days) some values lower than -2 were obtained for the trimmed mean IGNS.

Table 3. Test based on the normalized sum in new catchments and different FTHs of regional integration given by the analysis of clusters by geographical location of the basin outlets. Values lower than 5 determined that the scores of selection are better than the reference set. See clusters' distribution in Fig. 1. In each cluster, the catchments highlighted in bold represent the series that are not used by the hydrological members' selection training methodology.

FTH	Cluster 1								Cluster 2						
	H24	K17	U25	**K13**	**K52**	**U06**	**U24**	**U27**	J85	K73	M04	M06	**H93**	**M15**	**M36**
1	5.08	5.25	5.06	5.19	5.36	5.20	5.15	5.12	4.96	5.19	5.09	5.07	5.06	5.09	4.96
2	5.17	5.18	5.12	5.07	5.24	5.02	5.36	5.04	5.03	4.97	4.97	4.89	4.85	4.90	5.00
3	4.89	4.85	4.87	4.71	5.01	4.60	4.86	4.78	4.66	4.63	4.67	4.73	4.71	4.70	4.67
4	4.50	4.56	4.69	4.26	4.76	4.53	4.68	4.59	4.67	4.57	4.72	4.71	4.70	4.71	4.60
5	4.82	4.56	4.56	4.31	4.85	4.54	4.76	4.68	4.70	4.33	4.51	4.54	4.40	4.43	4.29
6	4.99	4.74	4.86	4.59	4.87	4.59	4.76	4.79	4.41	4.47	4.53	4.29	4.49	4.53	4.34
7	4.50	4.52	4.42	4.58	4.74	4.50	4.52	4.50	5.01	5.04	5.00	4.81	4.77	4.80	4.80
8	4.38	4.25	4.27	4.16	4.71	4.22	4.33	4.33	4.43	4.61	4.78	4.62	4.47	4.84	4.41
9	4.50	3.97	4.09	4.04	4.36	4.07	4.32	4.17	4.09	4.32	4.59	4.39	4.31	4.39	4.22

FTH	Cluster 3				Cluster 4				Cluster 5				
	O34	Q25	**P70**	**P72**	B31	H36	H53	**H62**	A69	A79	A92	B21	**A70**
1	4.88	4.68	4.74	4.78	5.69	5.21	4.92	5.09	4.20	4.78	4.42	4.98	4.94
2	4.83	4.61	4.73	4.81	5.85	5.11	4.64	5.15	4.40	4.98	4.78	4.52	5.22
3	4.16	4.36	5.98	4.74	5.83	4.69	7.24	4.65	5.03	5.42	5.02	4.96	5.45
4	4.77	3.43	4.47	4.28	5.97	4.49	5.23	7.01	5.19	5.57	5.58	5.11	6.22
5	4.80	4.53	4.69	4.68	5.71	5.29	5.24	5.60	5.10	5.80	4.74	5.50	5.60
6	4.68	4.47	4.59	4.55	5.78	4.96	5.41	5.45	4.78	5.62	5.32	5.31	5.45
7	4.62	4.74	4.45	4.32	5.24	4.60	4.81	5.16	5.12	5.11	4.35	5.53	5.57
8	4.70	4.34	4.39	4.28	4.58	4.57	4.91	5.46	4.97	5.22	4.25	5.50	5.08
9	4.36	4.15	4.28	4.12	4.26	4.08	4.50	4.74	4.87	4.66	4.45	4.92	5.38

In Table 3, the normalized sum (NS) for the 9-day FTH is generally lower than 5 for catchments subjected to the regional integration (except basin A70). Furthermore, in 44 % of such assessments (catchments H24, K17, U25, J85, K73, H36, and H53) the regional integration presents better results than the local performance relative indicators shown in Table 2.

Although the regional integration in clusters 1, 2 and 3 shows that the 85 % of the normalized sums are lower than 5 and the remaining 15 % corresponds principally to the first lead times (1 to 3 days), the clustering and posterior regional integration is less efficient for the groups 4 and 5, whose normalized sums are higher to 5 in 65 % of the cases.

The behaviour in cluster 5 is inherited from the low extrapolation efficiency highlighted in basins A69, A92, and B21 (Fig. 6). As such, the proposed regional integration mechanism is shown as a consistent task since its efficiency is a function of performance of its components.

With regard to cluster 4, the regional solution shows a lower diversity of hydrological models. This factor is evident in Fig. 7 which illustrates that for this cluster 70 % of the hydrological members originate from only three hydrological models (HM03, HM06, and HM14), which is quite a different behaviour than for clusters 1, 2 and 3 where the portion of the three most selected models reaches 58 %, 56 %, and 44 %, respectively.

Thus it seems that diversity as characteristic of the final selection of hydrological members appears to be a factor with a significant impact on the performance of the selection. In other words, the participation of hydrological models in the regional selection stresses the importance of the integration of models with different characteristics. To view this in a deterministic framework, the index based on the performance rank assigned to each model in each catchment (Sect. 4.2.3) shows that the most selected models (HM01, HM03, HM06, HM09, and HM14) occupy quite different ranks (Fig. 7). For instance, HM03 and HM09 present a high performance while HM01, HM06 and HM14 are of lower performance. This feature exemplifies the notion of the diversity discussed in different stages of the scientific community concerning ensemble methods.

Alpaydin (2010) statistically showed that if an ensemble of d models with outputs that are independent and identically distributed, has a negative correlation between their error, the error variance of the average ensemble decreases proportionally with d^2. For hydrological model combination,

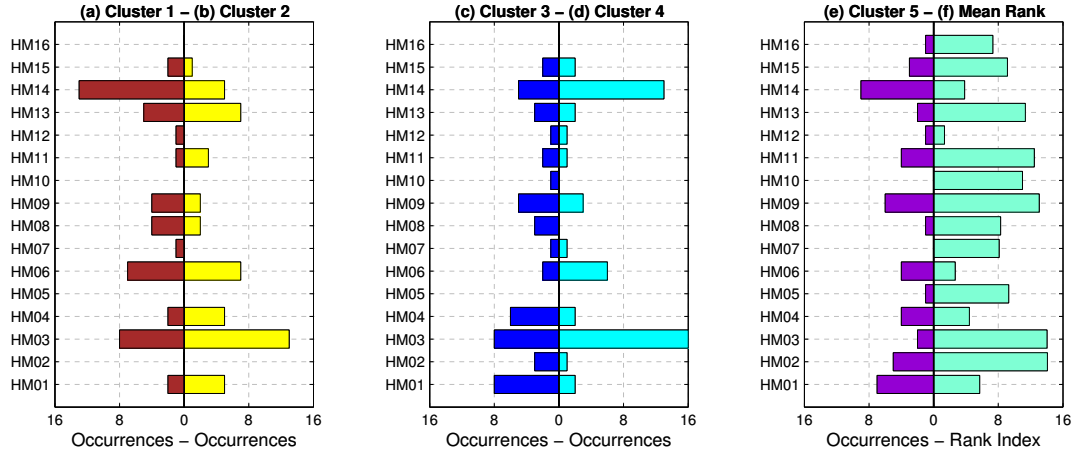

Fig. 7. Hydrological Models participation. Distribution in the five regions (clusters) are presented in **(a)**–**(e)**. Model performance evaluated as the mean rank index is shown in **(f)**.

Vrugt et al. (2008) proposed positive correlation (lack of diversity) as an efficient mechanism for removal of members of an ensemble.

Diversity can be defined as the search for models that complement their skills, so that each model focuses on different objects. Diversity in the ensemble is thus a vital requirement for successful modelling. In practice, it appeared to be difficult to define a single measure of diversity and even more difficult to relate that measure to the ensemble performance in a neat and expressive dependency (Kuncheva, 2004). Nevertheless, the regional clusters in Fig. 7 make use of most of the 16 available models, whatever their performance rank. For example, the most frequently selected models in cluster 2 are HM03 and HM06 despite the fact that HM02 exhibits the same rank of performance as HM03 and that HM06 presents one of the lowest ranks in the ensemble.

6 Conclusions

A companion paper has already demonstrated the success of the backward greedy member selection technique for simplifying a 800-member HEPS combining the 50 perturbed members from the ECMWF MEPS with 16 lumped hydrological models (Brochero et al., 2011). The present paper has focused on the generalization quality in time and space of a 50-member HEPS selected from the 800-member ensemble correspondent to the 9-day FTH. When applied to the other 8 time horizons, the 50 selected members also improved performance over the initial 800-member HEPS in 82 % of the situations. It was particularly successful when applied to a nearby catchment of the same cluster. Member diversity seems to be the key to this simplified HEPS that makes use of only 6.25 % of the initial structures (50 members/800 members). Indeed, it has been shown that most 50-member HEPS relied on a broad selection of hydrological models, which gives further support to the multi-model hydrological approach.

Comparing scores obtained for the 50 representative hydrological members to the ones of the initial 800-member ensemble indicated that the proposed selection methodology, which is based on cross-validation and the combination of scores into a single function, generally leads to good performance in terms of gains of individual scores. However, these gains were not entirely transferable under the scheme of extrapolation evaluated here. This drawback may in part be attributable to the simple selection methodology used here along a linear integration of scores that has no real control over balance, or the need to evaluate more features to enhance such transferability in the clustering approach.

A more sophisticated approach would optimize all performance diagnostics simultaneously or find a Pareto set of solutions identifying trade-offs among the various performance metrics. Such a framework, but in a context of combination rather than selection of hydrological members, was proposed by Vrugt et al. (2006). It consists in the optimization of Bayesian Model Averaging weights and variance using the A Multi-ALgorithm Genetically Adaptive Multiobjective (AMALGAM) method.

Finally, it would be interesting, in the case of a HEPS driven by interchangeable meteorological members, to combine the participation of hydrological models found with BGS-CV with the meteorological members chosen by a technique such as that proposed by Molteni et al. (2001) instead of testing them randomly.

Acknowledgements. The authors acknowledge NSERC and the Institute EDS for financial support, CEMAGREF and ECWMF for making the database available, and CVSL for the computational resources. We thank G. Thirel and an M. Zappa 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, 2nd Edn., The MIT Press, Cambridge, 2010.

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.

Bishop, C. M.: Pattern Recognition and Machine Learning (Information Science and Statistics), ISBN0387310738, Springer-Verlag New York, Inc., Secaucus, NJ, USA,2006.

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.

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 1: Optimization criteria, Hydrol. Earth Syst. Sci., Hydrol. Earth Syst. Sci., 15, 3307–3325, doi:10.5194/hess-15-3307-2011, 2011.

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

Confesor, R. B. and Whittaker, G. W.: Automatic calibration of hydrologic models with multi-objective evolutionary algorithm and pareto optimization, J. Am. Water Resour. Assoc., 43, 981–989, doi:10.1111/j.1752-1688.2007.00080.x, 2007.

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.

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.

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.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, doi:10.1029/97WR03495, 1998.

Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Parameter estimation of a land surface scheme using multicriteria methods, J. Geophys. Res., 104, 19491–19503, doi:10.1029/1999JD900154, 1999.

He, Y., Wetterhall, F., Cloke, H. L., Pappenberger, F., Wilson, M., Freer, J., and McGregor, G.: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, Meteorol. Appl., 16, 91–101, doi:10.1002/met.132, 2009.

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.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schär, C.: A probabilistic view on the August 2005 floods in the upper Rhine catchment, Nat. Hazards Earth Syst. Sci., 8, 281–291, doi:10.5194/nhess-8-281-2008, 2008.

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., Marsigli, C., Montani, A., Nerozzi, F., and Paccagnella, T.: A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments, Q. J. Roy. Meteorol. Soc., 127, 2069–2094, doi:10.1002/qj.49712757612, 2001.

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.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, doi:10.1175/MWR2906.1, 2005.

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.

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

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrol. Process., 18, 2743–2746, doi:10.1002/hyp.5687, 2004.

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. and Robinson, B. A.: Improved evolutionary optimization from genetically adaptive multimethod search, P. Natl. Acad. Sci. USA, 104, 708–711, doi:10.1073/pnas.0610471104, 2007.

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.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13–26, doi:10.5194/hess-5-13-2001, 2001.

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

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.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, J. Hydrol., 204, 83–97, doi:10.1016/S0022-1694(97)00107-8, 1998.