Hydrol. Earth Syst. Sci. Discuss., 8, 2783–2820, 2011
www.hydrol-earth-syst-sci-discuss.net/8/2783/2011/
doi:10.5194/hessd-8-2783-2011

Hydrology and Earth System Sciences Discussions

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

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

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

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

Abstract

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 article evaluates a 94% simplification of an initial 800-member HEPS, forcing 16 lumped rainfall-runoff models with the European Center for Medium-range Weather Forecasts (ECMWF MEPS). More specifically, it tests the time (local) and space (regional) generalization ability of the simplified 50-member HEPS obtained using a methodology that combines 4 main aspects: (i) optimizing information of the short-length series using *k*-folds cross-validation, (ii) implementing a backward greedy selection technique, (iii) guiding the selection with a linear combination of diversified scores, and (iv) formulating combination case studies at the cross-validation stage. At the local level, the transferability of the 9th day member selection was proven for the other 8 forecast horizons at an 82% success rate. At the regional level, a good performance was also achieved when the 50-member HEPS was applied to a neighbouring catchment within the same cluster. Diversity, defined as hydrological model complementarities addressing different aspects of a forecast, was identified as the critical factor for proper selection applications.

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., 2010). 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 may only be achieved through the uncertainty cascade model proposed by Pappenberger et al. (2005), which 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 another 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) favors 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 huge. 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, resorting to 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 forecast ensembles derived from numerical weather prediction models. The BMA predictive probability density function (PDF) is a weighted average of the PDF's centered 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

HESSD
8, 2783–2820, 2011

(Vrugt et al., 2006). In this same line, 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 uncertainty cascade model (Todini, 2004; Jaun et al., 2008; Pappenberger et al., 2005), triggering interest in novel member selection techniques. For example, Marsigli et al. (2001); Jaun et al. (2008); Molteni et al. (2001) select MEPS members based on lagging ensembles and deriving representative members through hierarchical clustering over the domain of interest. Ebert et al. (2007) analyzed the relation between the atmospheric circulation patterns and extreme discharges to select representative members of MEPS or, in a deterministic way ("best match" approach), to determine the location of the forecast that most resembled the rainfall pattern over the catchment (Xuan et al., 2009). In a companion paper, Brochero et al. (2011) presented the HEPS member selection methodology adopted here: a Backward greedy selection algorithm (Alpaydin, 2010) retaining the uncertainty properties of a 800-member ensemble derived from the 50 members of the European Center for Medium-range Weather Forecasts (ECWMF) and 16 simple lumped hydrological models (see Sect. 2).

Another aspect of particular interest in the evaluation of probabilistic forecast, and therefore in 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 member selection with a multi-score focus. Vrugt et al. (2006) posed the BMA inverse

HESSD
8, 2783–2820, 2011

problem in a multiobjective framework, examining the Pareto set of solutions between the Continuous Rank Probability Score (CRPS), the Mean Absolute Error (MAE), and the Ignorance Score. This is achieved using the A Multi-ALgorithm Genetically Adaptive Multiobjective (AMALGAM) method, which combines two new concepts: a simultaneous multi-method search and self-adaptive offspring creation (Vrugt and Robinson, 2007). The present paper explores member selection from a linear combination of different scores (see Sect. 3). Selection based on individual criterion has already been presented in the companion paper (Brochero et al., 2011).

Finally, it has been shown that the enhancement of HEPS is related to the quality of available information, particularly the number of extreme events, and the possibility of combining results of different studies (Cloke and Pappenberger, 2009). Although more case studies are needed, the improvement of HEPS should be aligned under reforecasting studies (Hamill et al., 2004). Here, the optimal use of the information is addressed through k-folds cross-validation and subsequent combination of the k experiments (see Sect. 4), local evaluation of selection in different forecasting horizons and regional integration of the selection based on a k-means clustering algorithm (Sect. 5.2). Results and discussion are gathered in Sect. 6, while conclusions and a guideline for future work are drawn in Sect. 7.

2 HEPS configuration

The basic configuration of the HEPS was established by Velázquez et al. (2010) in their work on the comparison of the qualities of different types of HEPS, combining uncertainties from the meteorological input and from the hydrological model structure. The uncertainty of the inputs is represented by the 50 members from the ECMWF that are a priori assumed to be equally likely (Gouweleeuw et al., 2005) – a detailed description of the ECWMF model can be found in Molteni et al. (1996) or Buizza (2005). As for the uncertainty of the hydrological process, 16 independent structures are explored. These hydrological models are lumped reservoir-type models proposing

HESSD
8, 2783–2820, 2011

various conceptualizations of the rainfall-runoff transformation at the catchment scale (Table 1). It is important to note that some models such as the HM02 were specifically devised for the catchment scale, whereas others such the HM06 or HM08, inspired from distributed models, have suffered a series of substantial changes bringing them to a lumped state.

On average, 29 years of observations were available for the hydrological model calibration, based on the RMSE objective function. It is beyond the scope of this article to present these models and their respective different settings. A detailed explanation of each model and the various adjustments made can be found in Perrin (2000) with the exception of models GR4J (Perrin et al., 2003), MORDOR (Garçon, 1999), MOHYSE (Fortin and Turcotte, 2006; Roy et al., 2010), SIMHYD (Chiew et al., 2002) and Probability-distributed store (PDS) (Yadav et al., 2007).

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 2, where maximum values stress the problem variability: mean rainfall is equal to 2.6 mm with a standard deviation of 0.5 mm and a mean flow equal to 1.0 mm with a standard deviation of 0.5 mm. Codification and flow data are from the Banque Hydro database (http://www.hydro.eaufrance.fr) while rainfall observations originate from Météo-France SAFRAN. As already mentioned, the rainfall probabilistic forecasts are from the ECMWF.

3 Scores: verification metrics

There are various attributes to ensemble forecast quality such as bias, sharpness, reliability and consistency that are considered important for accurate probabilistic weather forecasting (Murphy, 1993; Wilks, 2005). Each of them is represented to a greater or lesser degree by a different metric (score).

HESSD
8, 2783–2820, 2011

In some cases, it is necessary to establish an a priori probabilistic distribution function that systematically fits the prediction ensembles at each time step. In hydrology, it is generally accepted that the adjustment of the gamma distribution is more appropriate than that of the normal distribution, given asymmetry in the distribution of precipitation and discharge (Vrugt et al., 2008); however, the gamma function evaluation involves more complex computations than for the normal distribution, for which an explicit mathematical expression exists.

Here, after some tests assessing disparities between normal and gamma distributions in the case of the CRPS and the IGNS, results showed small differences in contrast to the high computational cost imposed by the gamma distribution (1.7 h vs. 47 h of calculations). However, it is important to note that the disparities may increase for ensembles of less than 30 members, which is not an issue here.

The following scores are considered here: the mean Continuous Ranked Probability Score ($\overline{\text{CRPS}}$) that principally measure the ensemble sharpness or spread (Gneiting and Raftery, 2007; Hersbach, 2000; Boucher et al., 2009), the mean ignorance score ($\overline{\text{IGNS}}$) that strongly reacts to bias (Roulston and Smith, 2002), the rank histograms uniformity evaluated from the δ ratio that determines the consistency condition (Candille and Talagrand, 2005), the reliability diagram error (RD_{MSE}) that evaluates the reliability properties in a single scalar value, and finally a simple measure of spread proposed by Brochero et al. (2011): the median of the coefficients of variation evaluated for each ensemble (Eq. 5). A detailed version of each metric can be found in the companion paper (Brochero et al., 2011). These five metrics can be computed as follows:

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

HESSD
8, 2783–2820, 2011
Simplifying a hydrological ensemble prediction system – Part 2
D. Brochero et al.

$$\overline{\text{IGNS}}(\mathbf{Y},\boldsymbol{o}) = -\frac{1}{N}\sum_{t=1}^{N}\log_2\left[f(\boldsymbol{y}^t)_{o^t}\right] \tag{2}$$

$$\text{RD}_{\text{MSE}}(\mathbf{Y},\boldsymbol{o}) = \frac{1}{M}\sum_{m=1}^{M}(\bar{o}_\text{m} - l_\text{m})^2 \tag{3}$$

$$\text{where } \bar{o}_\text{m} = \frac{1}{N}\sum_{t=1}^{N} r^t, \quad r^t = \begin{cases} 1 \text{ if } o^t \in l_\text{m} \\ 0 \text{ otherwise} \end{cases}$$

$$\delta(\mathbf{Y},\boldsymbol{o}) = \frac{\sum_{c=1}^{d+1}(S_c - h_\text{ref})^2}{\Delta_0} \tag{4}$$

$$\text{where } h_\text{ref} = \frac{N}{d+1} \quad \text{and} \quad \Delta_0 = \frac{dN}{N+1}$$

$$\text{MDCV}(\mathbf{Y}) = \operatorname*{med}_{t=1}^{N} CV(\boldsymbol{y}^t), \tag{5}$$

where in the Eq. (1) $F(\boldsymbol{y})^t$ represents the cumulative distribution function fitted to the ensemble $\boldsymbol{y}^t$ with mean μ_t and variance σ_t^2 at the time t, o^t denotes the observation; ϕ and Φ denote the normalized variables for the probability density function and cumulative distribution function, respectively, and N is the total number of observations. The ignorance score (IGNS) is calculated from the evaluation of the logarithm of the $f(\boldsymbol{y}^t)$ at the point corresponding to the observation o^t (Eq. 2). Note that the IGNS score equals the average of the natural logarithms evaluated at the observations, and the CRPS is equivalent to the mean absolute error for a deterministic forecast (Hersbach, 2000). Smaller values of the diagnostic measures are preferred and indicate a better performance.

RD_MSE represents the mean squared differences in the reliability corresponding to squared vertical distances between the conditional event relative frequency $\bar{o}_\text{m} = P(o^t|l_\text{m})$ and the M forecast probabilities l_m (Eq. 3). These distances are all small for

HESSD
8, 2783–2820, 2011

well-calibrated forecasts.

The rank histogram (S_c occurrences inside each rank or class c) is used to evaluate whether the ensembles apparently include the observations being predicted as equiprobable members. Candille and Talagrand (2005) proposed the δ ratio as a measure of the reliability of an ensemble prediction system for a scalar variable, where d indicates the total number of members in each forecast ensemble. A value of δ that is significantly larger than 1 is a proof of unreliability.

Here, instead of using a single measure or characteristic to select the representative members of the HEPS, a linear multi-score function proposed by Brochero et al. (2011) is considered for calibration of the forecast ensemble (Eq. 6). This metric reflects different but complementary scalar measures.

$$\mathrm{CC} = w_1 \frac{\overline{\mathrm{CRPS}_{\mathrm{se}}}}{\overline{\mathrm{CRPS}_{\mathrm{ie}}}} + w_2 \frac{z_1 - \overline{\mathrm{IGNS}_{\mathrm{se}}}}{z_1 - \overline{\mathrm{IGNS}_{\mathrm{ie}}}} + w_3 \frac{\mathrm{RD}_{\mathrm{MSEse}}}{\mathrm{RD}_{\mathrm{MSEie}}} + w_4 \frac{\delta_{\mathrm{se}}}{\delta_{\mathrm{ie}}} + w_5 \frac{z_2 - \mathrm{MDCV}_{\mathrm{se}}}{z_2 - \mathrm{MDCV}_{\mathrm{ie}}} \tag{6}$$

As part of the member selection framework, in order to normalize each of the components of the combined criterion (CC), the result of each criterion in the selection ensemble (se subscript) is divided by the criterion calculated on the 800-member initial ensemble (ie subscript). Special formulation requires the normalization of the IGNS with the threshold z_1 to take into account the scale of positive and negative values that can take this score. It is also necessary to set a threshold z_2 for the MDCV function to change the maximization orientation in terms of the minimization of all the factors together. Preliminary analysis showed that $z_1 = -2$ and $z_2 = 1$ covers all scenarios of the series at hand. Finally the last part of the combined criterion consists of the weights assigned to each component (w_{cp}). Here, the weight assigned to the reliability (the critical factor) is twice that of the other factors, which have a unit weight.

HESSD
8, 2783–2820, 2011

4 Member selection methodology

More details on the member selection methodology can be found in the companion paper (Brochero et al., 2011). Figure 2 summarizes the selection procedure applied to the 800-member HEPS configured from the 50 9th day ECMWF forecasts and the 16 lumped hydrological models. The selection is executed in 3 steps, which are described below:

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 (χ_t), validation (χ_v) and test sets (χ_p). Thus, the process of selecting data follows a k-fold cross-validation technique. The dataset (χ) is divided into 5 equal-sized parts in order to create 5 experiments. In each experiment, a part is set aside for testing, while the remaining four parts are grouped into n blocks of 10 consecutive pairs of observations-ensemble forecast, randomly choosing 75% of the blocks for the training set and the remaining 25% for the validation set.

Step 2: Backward greedy member selection. Optimization relies on the combined criterion (Eq. 6) for a preselected number of members (nm). The mechanism of member elimination begins with all members (d), removing at each step the one that decreases the error the most (or increases it the least). A pseudocode is given (Algorithm 1).

Step 3: Combination of results. It is highly likely that variability in the 5 experiments configured in step 1 leads to different solutions. An integration mechanism is thus needed for a global solution for each catchment. The importance of each member $\boldsymbol{y_i}$ within the ensemble is then assumed as being directly proportional to the iteration number at which it was eliminated during the selection process in each experiment ($\mathrm{iter}_{xp}^{\boldsymbol{y_i}}$). The combined ranking is thus the mean rank of elimination as defined in Eq. (7).

$$\overline{R}(\boldsymbol{y_i}) = \frac{1}{5}\sum_{xp=1}^{5} \mathrm{iter}_{xp}^{\boldsymbol{y_i}} \tag{7}$$

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

HESSD

8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

Algorithm 1 Backward greedy selection pseudocode

1. Subdivision of the dataset χ in train, validation and test sets.
$\chi \rightarrow \chi_t, \chi_v, \chi_p$
2. Define Combined criterion as a error function (CC) and the number of members to select (nm).
3. Initialize ensemble with all d members
$G^d = \{y_1, y_2, y_3, \ldots, y_d\}$
4. Remove the worst member
for iter $= d-1, d-2, \ldots,$ nm **do**
– Update selection set by removing currently worst member
$y_j = \mathrm{argmin}_{y_j \in G^{\mathrm{iter}+1}} \mathrm{CC}(G^{\mathrm{iter}+1} \backslash \{y_j\} | \chi_t)$
$G^{\mathrm{iter}} = G^{\mathrm{iter}+1} \backslash y_j$
– CC evaluation in the validation set
$\mathrm{CC}_v^{\mathrm{iter}} = \mathrm{CC}(G^{\mathrm{iter}} | \chi_v)$
end for

Finally, the final selection (**s**) of the nm best members corresponds to the members which have the highest mean rank of elimination (Eq. 8).

$$\mathbf{s} = \{\overline{R}_p, y_p\}_{p=1}^{\mathrm{nm}}, \; \overline{R}_i \geq \overline{R}_j \text{ where } 1 \leq i \leq j \leq d \tag{8}$$

5 Experimental set-up

The generalization ability of a hypothesis, namely, the quality of its inductive bias, can be measured if there is access to data outside 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 resort

HESSD
8, 2783–2820, 2011

to a third set, a test set, sometimes also called the publication set, containing examples not used in training or validation (Alpaydin, 2010).

Thus, the method of combining results based on the mean rank of elimination is founded 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 hydrologic cycle). However, results of this procedure can be conceived as indicators of relative performance or otherwise as an optimistic estimate of the selection process (Diamantidis et al., 2000).

This study rigorously tests the outcome of the selection methodology at two levels: one local that focuses on the extrapolation of results to different forecast time horizons in the same catchment and another named regional, testing the temporal and spatial performance in nearby catchments, or under a broader perspective on the integration of regional results.

5.1 Extrapolation to a different forecast time horizon

As described in Sect. 2, member selection is performed on the results of 16 hydrologic models fed with the 9th day forecast time horizon of the ECMWF MEPS. Thus, the application of this selection of members for the other eight forecast time horizons (1 to 8) is a first level test. It has to be stressed that the idea of simplifying the HEPS is only valuable if the member selection is invariant in regard to the forecast time horizon. However, one may always argue that the assumption of statistical independence between the test and training data, principally for forecast time horizons next to the ninth, may be somewhat questionable.

5.2 Extrapolation to a different catchment

Transferring selected members to a neighbouring catchment, and even further to a different forecast time horizons, constitutes a rigorous test of the generalization ability of results at both the temporal and spatial scales. The choice of the second catchment

HESSD
8, 2783–2820, 2011

could first be viewed as a simple nearest neighbour problem, but in a context of greater significance, called regional, this problem requires the definition of regions (clusters) of similar behaviour and a regional integration mechanism to select the final members.

5.2.1 *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, minimum, average and maximum precipitation, evapotranspiration and flow (see Table 2). 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. 5.2.2.

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 $k = 1,\ldots,K$ 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 (Eq. 9).

$$J = \sum_{l=1}^{L} \sum_{k=1}^{K} b_k^l \|\boldsymbol{x}^l - \boldsymbol{m}_k\|^2 \tag{9}$$

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 minimize J. Then the iterative application of Eq. (9) leads to the following procedure for finding the $\boldsymbol{m}_k$ centers (Algorithm 2). Details of the k-means clustering algorithm are given by Bishop (2007).

Figure 1 shows an example of k-means clustering based only on the geographic location of each catchment, the various symbols in the Fig. 1 identify the 5 regions.

HESSD
8, 2783–2820, 2011

Algorithm 2 $k-means$ pseudocode

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

repeat

  for all $\boldsymbol{x}^l, l = 1, \ldots, L$ **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 $\boldsymbol{m}_k, k = 1, \cdots, K$ **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

Algorithm 3 Regional integration mechanism pseudocode

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

repeat

  for $j = 1, \ldots, C$ **do**

    if $S_{i,j} \notin \mathbf{rs}$ **then**

      $\mathbf{rs} = \mathbf{rs} + S_{i,j}$

      $h = h + 1$

    end if

  end for

  $i = i + 1$

until $h < q$

HESSD

8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

5.2.2 Regional integration mechanism

The member selection integration for region k, consisting of C catchments, is defined from matrix **S**, which has C columns with nm rows representing the most nm important 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, i.e. each member cannot be selected again later, until the number of members in **rs** is equal to the desired q. Algorithm 3 details this procedure.

6 Results and discussion

The optimal number of 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 members, simplifying the assessment of the participation of members of the ECMWF MEPS (whose number is also 50) in the final selection.

Table 3 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. (6) with unit weights definition. A value of 5 thus indicates an identical performance for the 50-member selection as for the 800-member reference set. Values lower than 5 reflect a better performance of the 50-member selection.

Table 3 shows that in all cases the normalized sum (NS) is always lower than 5, indicating the superiority of the selected 50-member HEPS, even after a size reduction equivalent to a 94% compression of initial 800-member HEPS. 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

HESSD
8, 2783–2820, 2011

of its components, for a collective analysis. In this sense, the analysis is facilitated if each component is associated with an index (Eq. 10) that reflects the gain or loss of the selection (se subscript) over the initial 800-member set (ie subscript). Note that the absolute value is used in the denominator for accounting for negative values that can take 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 initial HEPS.

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

Based on this formulation, it is noted that for the 50-member selection, the CRPS and the MDCV show low variability with gain indexes around 2% and 5%, respectively. The $\mathrm{RD}_{\mathrm{MSE}}$ shows minimum gains of 50% (catchment B21) and 87% (catchment K17), reflecting the emphasis given to this property in the formulation of the combined criterion used for optimizing where this component was given a weight twice that of the others. With respect to the IGNS, index gains between −5% and 29% (excluding the catchment B21) reflect an acceptable behaviour.

Finally, the delta ratio is the most difficult to retain; a positive index gain was obtained for only 25% of the cases (4/16), while the spread ranged from −41% for catchment H36 to 31% for catchment B31. Note that the delta ratio has an inverse relationship with the number of members of the selection (Eq. 4), 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 delta ratio is the best individual metric for member selection, second only to the proposed combined criterion.

6.1 Local analysis

For operational convenience, it is fundamental that the 50 selected members for the 9th day forecast time horizon 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

HESSD
8, 2783–2820, 2011

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 50-member random combinations, in order to evaluate if any good performance may only be attributable to chance. Results for the 8 first time horizons and 16 watersheds are gathered in box-plot diagrams (Fig. 3).

The 50 selected members for the 9th day forecast time horizon 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 selected members lead to a worse performance than the 25 percentile of 200 random combinations test. Figure 3 also shows that the selection slowly loses efficiency as it moves away from the 9th day forecast horizon. It also detects a systematic deficiency for catchment A69 and to a lesser extent for catchment B21. Nonetheless, these results are very encouraging.

Figure 4 shows the histogram of the participation of members of the ECMWF MEPS in the final selection for each of the 16 catchments. The results hint at some degree of uniformity in the histograms – there are only three occurrences for which a meteorological member is selected 5 times out of 50 (see catchments A69, A79 and B21). Such uniformity and the already mentioned high selection level of all hydrological models (last column of Table 3) support the multi-model trend as an effective mechanism to shelter the uncertainty related to hydrological processes in the probabilistic forecasting of stream flows.

6.2 Regional analysis

As described in Sect. 5.2, the regional analysis assesses the generalization ability of the member selection for a specific catchment with respect to another one at a different forecast time horizon. For example, Fig. 5 explores the transferability of the 50-member selection obtained for catchment Q25 (FTH = 9) to catchment P72 at FTH = 4.

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

HESSD
8, 2783–2820, 2011

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 (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., 2010).

Another remarkable feature of CRPS is its direct relationship with the flow magnitude; the shapes of the CRPS and of the hydrograph are similar. A direct strategy of optimization could then focus on the maximum values from the cost function. The selection thus not only preserves the mean CRPS (0.13) but also the structure of the CRPS series.

Figure 5b shows that the 50-member IGNS (-1.65) is also 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, considering that a value of 10 corresponds to the local assessment of 9.7×10^{-4} as a probability density function on-site point observation.

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 set. Similarly, Fig. 5d reveals that the rank histograms have a nearly uniform distribution, even if the first and the last rank reflects a slight bias. Those imperfections demonstrate the difficulty inherent in minimizing the δ ratio.

Figure 5e illustrates the occurrence of each lumped model within the 50-member 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

HESSD
8, 2783–2820, 2011
Simplifying a hydrological ensemble prediction system – Part 2
D. Brochero et al.

selection favored a diversity of errors. At the end of the selection process, the median of the coefficients of variation series (MDCV) has slightly increased, from 0.15 to 0.16.

The regional analysis, which seeks to identify features that facilitate the combination of results from different basins, 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 4, after application of the *k*-means algorithm and of the regional integration procedure already described in Sect. 5.2.

In Table 4, the normalized sum (NS) for FTH = 9 is always lower than 5 for catchments subjected to the regional integration. Furthermore, in 38% of such assessments (catchments H24, K17, U25, K73, M04 and H36) the regional integration gave better results than the local performance relative indicators showed in Table 3.

Although the regional integration in clusters 1, 2 and 3 shows that 93% of the normalized sums are lower than 5, it is less efficient for clusters 4 and 5, whose normalized sums are higher to 5 in 57% of the cases. Such expense may be assessed by the participation of members of the ECMWF MEPS in the selection of each catchment (Fig. 4) or the participation of the hydrological models in the regional selection (Fig. 6). However, MEPS members selected catchments for clusters 4 and 5 (Fig. 4a–d) do not show a different behaviour with respect to the other catchments, so the difference may be attributable to the selection hydrological models.

In this regard, Fig. 6 shows that 70% of the cluster 4 members originate from only 3 hydrological models (HM03, HM06 and HM14), which is a quite different behaviour than for clusters 1, 2 and 3 where the portion of the three most selected models reaches 58%, 56% and 44%, respectively. In that regard, cluster 5 is similar to cluster 3; however, it is important to note that cluster 5 integrates 4 catchments while cluster 3 integrates only 2.

The participation of hydrological models in the regional selection stresses the importance of the integration of models with different characteristics. To view this in

HESSD
8, 2783–2820, 2011

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:

- The mean square error $\mathrm{MSE}_{i,j}$ for catchment i and hydrological model j is first calculated.
- The 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 RI_j for each model is estimated based on the results of all 28 basins (Eq. 11).

$$\mathrm{RI}_j = \frac{1}{28}\sum_{i=1}^{28}\mathrm{PR}_{i,j} \tag{11}$$

Mean ranks RI_j are illustrated in Fig. 6f. It is noteworthy that the most selected models (HM01, HM03, HM06, HM09, and HM14) occupy quite different ranks. For instance, HM03 and HM09 are high performance while HM01, HM06 and HM14 are of lower performance. This feature exemplifies the notion of diversity discussed in different stages of the scientific community concerning ensemble methods.

Alpaydin (2010) shows statistically that if an ensemble of d models, 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, 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 modeling. 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. 6 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 rank in the ensemble.

7 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 forecasts 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 9th day time horizon forecasts of the 800-member ensemble. When applied to the other 8 time horizons, the 50 selected members again improved performance over the initial 800-member HEPS in 82% of the situations. It was also quite 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 structure. Indeed, it has been shown that most 50-member HEPS relied on a broad selection of meteorological members and hydrological models.

Comparing scores obtained for the 50 representative member ensembles to the ones of the initial 800-member ensembles showed 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. 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. 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 was proposed by Vrugt et al. (2006) that consists in the optimization of Bayesian Model Averaging

HESSD
8, 2783–2820, 2011

weights and variance using the A Multi-ALgorithm Genetically Adaptive Multiobjective (AMALGAM) method.

Consistency of the HEPS, evaluated from the rank histogram or its scalar variable named the delta ratio, turned out to be important. However, the delta ratio is inversely related to the number of members of the ensemble, which becomes an unfavorable factor regarding the direct comparison between two sets of different numbers of members. An alternative to the evaluation of the uniformity of the rank histograms should be considered. For example, it could be based on a classic fit measure such as the MSE between the relative frequency of the rank histogram and the value that represents an equiprobable histogram behaviour.

Appendix A

Notations

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

HESSD
8, 2783–2820, 2011

$[\bar{o}_\mathrm{m}]$	Conditional probability of the event as a function of the interval I_m assigned to the forecast $m \rightarrow P(o^t \vert I_\mathrm{m})$
$[r^t]$	Binary indicator, 1 if the event occurs for the tth forecast-event pair, 0 if it does not
$[S_c]$	Absolute frequency or occurrences in each rank c in the rank histogram
$[\mathrm{med}_{t=1}^{N}]$	Median value evaluated from $t=1$ to N
$[\mu_t]$	Mean ensemble flow forecasts at the time t
$[\sigma_t^2]$	Variance ensemble flow forecasts at the time t
$[\chi_\mathrm{t}]$	Training set
$[\chi_\mathrm{v}]$	Validation set
$[\chi_\mathrm{p}]$	Test or publication set
$[\{x^t\}_{t=1}^{N}]$	Set of x with index t ranging from 1 to N
$[\mathrm{argmin}_\theta g(x\vert\theta)]$	The argument θ for which g has its minimum value
$[E(\theta\vert\chi)]$	Error function with parameters θ on the sample χ
$[w_\mathrm{cp}]$	Weights of the components of the combined criterion (CC)
$[\mathrm{iter}_{xp}^{\boldsymbol{y}_i}]$	Iteration number at which was eliminated the $\boldsymbol{y}_i$ member during the selection process in the xp experiment
$[\overline{R}(\boldsymbol{y}_i)]$	Mean rank of elimination of the $\boldsymbol{y}_i$ member
$[\mathbf{s}]$	Final selection of the nm best members in the individual selection process
$[\boldsymbol{x}^l]$	Property set of the l catchment in the clustering process
$[k]$	Cluster indicator
$[m_k]$	Center of the kth cluster
$[b_k^l]$	Membership binary indicator of the l catchment to the kth cluster
$[\mathbf{S}]$	Matrix which includes local solutions $\mathbf{s}$ in each cluster
$[\mathbf{rs}]$	Regional ensemble of the best q members

HESSD
8, 2783–2820, 2011

$[\mathrm{PR}_{i,j}]$	Performance rank based on the MSE for catchment i and hydrological model j
$[RI_j]$	Mean rank RI_j of performance for each hydrological model j

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. They also thank Annette Schwerdtfeger for proofreading this manuscript.

References

Alpaydin, E.: Introduction to Machine Learning. Adaptive Computation and Machine Learning, The MIT Press, 2 edn., Cambridge, 2010. 2786, 2794, 2802

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. 2785

Bishop, C. M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006. 2795

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. 2789, 2800

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. 2785

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. Discuss., 8, 2739–2782, doi:10.5194/hessd-8-2739-2011, 2011. 2786, 2787, 2789, 2791, 2792, 2803

Buizza, R.: EPS skill improvements between 1994 and 2005, ECMWF Newslett., 104, 10–14, 2005. 2787

Candille, G. and Talagrand, O.: Evaluation of probabilistic prediction systems for a scalar variable, Q. J. R. Meteorol. Soc., 131, 2131–2150, doi:10.1256/qj.04.71, 2005. 2789, 2791

Chiew, F. H. S., Peel, M. C., and Western, A. W.: Application and testing of the simple rainfall-runoff model SIMHYD, in: Mathematical models of small watershed hydrology and applications, edited by: Singh, V. P., Frevert, D., Water Resources Publications, Littleton, CO, 335–367, 2002. 2788

Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009. 2786, 2787

Confesor, R. B. and Whittaker, G. W.: Automatic calibration of hydrologic models with multi-objective evolutionary algorithm and pareto optimization, J. Am. Water Resourc. Assoc., 43, 981–989, doi:10.1111/j.1752-1688.2007.00080.x, 2007. 2786

Diamantidis, N., Karlis, D., and Giakoumakis, E.: Unsupervised stratification of cross-validation for accuracy estimation, Artif. Intel., 116, 1–16, doi:10.1016/S0004-3702(99)00094-6, 2000. 2794

Ebert, C., Bárdossy, A., and Bliefernicht, J.: Selecting members of an EPS for flood forecasting systems by using atmospheric circulation patterns, Geophys. Res. Abs., Eur. Geosci. Union, 9, 08177, Vienna, Austria, 2007. 2786

Fortin, V. and Turcotte, R.: Le modèle hydrologique MOHYSE: Notes de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Université du Québec à Montréal, Montréal, Canada, 2006. 2788

Garçon, R.: Modèle global pluie-débit pour la prévision et la prédétermination des crues, La Houille Blanche, 0, 88–95, doi:10.1051/lhb/1999088, 1999. 2788

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. 2789, 2800

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. 2787

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. 2786

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. 2786

Hamill, T. M., Whitaker, J. S., and Wei, X.: Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts, Mon. Weather Rev., 132, 1434–1447, 2004.

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. 2787

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. 2785

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. Hazard. Earth Sys., 8, 281–291, 2008. 2789, 2790

Kuncheva, L. I.: Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, New York, 2004. 2786

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. R. Meteorol. Soc., 127, 2095–2115, doi:10.1002/qj.49712757613, 2001. 2803

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Q. J. R. Meteorol. Soc., 122, 73–119, doi:10.1002/qj.49712252905, 1996. 2786

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. R. Meteorol. Soc., 127, 2069–2094, doi:10.1002/qj.49712757612, 2001. 2787

Murphy, A.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather Forecast., 8, 281–293, 1993. 2786

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. 2788

Perrin, C.: Vers une amélioration d'un modèle global pluie-débit au travers d'une approche comparative, Ph.D. thesis, Institut National Polytechnique de Grenoble (INPG), France, 2000. 2785, 2786

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for stream-

HESSD

8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

flow simulation, J. Hydrol., 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003. 2788

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. 2788

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. 2785

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. 2784

Rousset, F., Habets, F., Martin, E., and Noilhan, J.: Ensemble streamflow forecasts over France, ECMWF Newslett., 111, 21–27, 2007. 2789

Roy, A., Royer, A., and Turcotte, R.: Improvement of springtime streamflow simulations in a boreal environment by incorporating snow-covered area derived from remote sensing data, J. Hydrol., 390, 35–44, doi:10.1016/j.jhydrol.2010.06.027, 2010. 2784

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrol. Process., 18, 2743–2746, doi:10.1002/hyp.5687, 2004. 2788

Velázquez, J. A., Anctil, F., Ramos, M.-H., and Perrin, C.: Comparison of the performance and reliability of 18 lumped hydrological models driven by ECMWF rainfall ensemble forecasts: a case study on 29 French catchments, EGU General Assembly, Austria, 12, 4929, 2010. 2786

Vrugt, J. A. and Robinson, B. A.: Improved evolutionary optimization from genetically adaptive multimethod search, P. Natl. Acad. Sci. USA, 104(3), 708–711, doi:10.1073/pnas.0610471104, 2007. 2784, 2787, 2800

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. 2787

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. 2786, 2803

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. 2786, 2789, 2802

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, vol. 91, Academic Press, 2 edn., Burlington, MA, London, 2005. 2786

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. 2786, 2788

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resourc., 30, 1756–1774, doi:10.1016/j.advwatres.2007.01.005, 2007. 2786

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. 2788

2786

HESSD

8, 2783–2820, 2011

Table 1. Hydrological models. The number of parameters used in each model are shown in brackets.

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

HESSD
8, 2783–2820, 2011

Table 2. Main characteristics of the studied catchments based on a 36 year length of the series (1970–2006). P: diary precipitation, ET: diary potential evapotranspiration, Q: diary observed flow.

Catchment codes	Area (km^2)	Altitude (m)	Pmax (mm)	ETmax (mm)	Qmax (mm)	Catchment codes	Area (km^2)	Altitude (m)	Pmax (mm)	ETmax (mm)	Qmax (mm)
A6921010	2780	200	58.28	4.06	12.68	K5220910	1836	193	57.08	4.08	12.61
A7010610	6830	184	57.44	4.04	20.49	K7312610	1712	85	45.00	4.30	15.14
A7930610	9837	155	56.99	4.07	18.22	M0421510	1890	56	39.42	4.11	7.13
A9221010	1760	195	68.94	4.15	22.58	M0680610	7380	22	35.47	4.17	7.80
B2130010	2290	227	55.11	4.05	17.02	M1531610	7920	21	38.38	4.22	4.90
B3150020	3904	162	56.63	4.04	12.79	M3600910	3910	26	39.79	4.02	16.04
H2482010	2982	85	42.16	4.20	9.77	O3401010	2170	349	182.83	4.01	81.62
H3621010	3900	48	51.29	4.26	6.56	P7001510	1863	87	70.88	4.35	18.83
H5321010	8818	110	43.95	4.11	4.64	P7261510	3752	6	64.50	4.43	9.88
H6221010	2940	76	45.21	4.09	8.73	Q2593310	2500	17	53.63	4.52	12.72
H9331010	4598	21	43.48	4.08	2.25	U0610010	3740	195	57.39	4.14	19.94
J8502310	2465	4	49.34	3.94	15.18	U2402010	3420	305	63.70	3.86	20.67
K1341810	2277	237	49.73	4.19	14.84	U2542010	4970	201	59.04	3.97	22.08
K1773010	1465	196	51.31	4.25	17.75	U2722010	7290	181	59.78	4.03	20.39

HESSD

8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

Table 3. Selection of 50 members based on combined criterion and the combination of k-fold cross-validation results on the forecast time horizon 9. Beside each score is presented the gain index evaluated by Eq. (10). NS represents the normalized sum (Eq. 6 with unit weights). NHM indicates the number of hydrological models participating in the solution.

Catchment codes	Scores				MDCV	NS	NHM
	CRPS	RD[$e-3$]	δ	IGNS	function		
A6921010	0.284 (0%)	1.321 (81%)	1.550 (13%)	0.667 (14%)	0.392 (5%)	3.99	9
Ref. Values (800 members)	0.284	6.953	1.778	0.780	0.374	5.00	16
A7930610	0.254 (3%)	1.548 (69%)	3.525 (−8%)	0.344 (23%)	0.407 (−1%)	4.32	11
Ref. Values (800 members)	0.263	5.057	3.264	0.445	0.410	5.00	16
A9221010	0.183 (4%)	0.328 (86%)	2.399 (−33%)	−0.417 (27%)	0.565 (0%)	4.36	11
Ref. Values (800 members)	0.192	2.367	1.802	−0.328	0.567	5.00	16
B2130010	0.232 (−1%)	1.231 (49%)	2.432 (−9%)	−0.180 (−38%)	0.627 (9%)	4.57	13
Ref. Values (800 members)	0.230	2.427	2.240	−0.291	0.575	5.00	16
B3150020	0.134 (1%)	1.252 (72%)	1.828 (31%)	−0.840 (−5%)	0.240 (7%)	3.93	11
Ref. Values (800 members)	0.135	4.507	2.660	−0.882	0.224	5.00	16
H3621010	0.157 (2%)	0.708 (80%)	2.101 (−41%)	−1.018 (2%)	0.363 (−1%)	4.55	14
Ref. Values (800 members)	0.161	3.499	1.490	−0.995	0.365	5.00	16
H5321010	0.165 (3%)	1.924 (74%)	4.129 (−31%)	−0.763 (8%)	0.357 (8%)	4.43	11
Ref. Values (800 members)	0.171	7.396	3.160	−0.706	0.332	5.00	16
H2482010	0.180 (2%)	2.237 (68%)	4.007 (−36%)	−0.821 (9%)	0.368 (6%)	4.59	12
Ref. Values (800 members)	0.185	7.084	2.943	−0.756	0.349	5.00	16
K1773010	0.205 (4%)	0.458 (87%)	1.901 (−9%)	−0.729 (12%)	0.385 (−2%)	4.15	12
Ref. Values (800 members)	0.213	3.560	1.746	−0.650	0.393	5.00	16
U2542010	0.290 (0%)	0.889 (74%)	2.522 (4%)	−0.404 (13%)	0.376 (7%)	4.18	14
Ref. Values (800 members)	0.289	3.391	2.620	−0.356	0.350	5.00	16
J8502310	0.159 (2%)	0.438 (80%)	1.632 (0%)	−1.002 (2%)	0.396 (8%)	4.11	14
Ref. Values (800 members)	0.163	2.164	1.630	−0.982	0.368	5.00	16
K7312610	0.160 (3%)	0.936 (70%)	2.171 (−17%)	−0.932 (0%)	0.381 (9%)	4.41	11
Ref. Values (800 members)	0.165	3.087	1.860	−0.930	0.348	5.00	16
M0421510	0.158 (1%)	0.552 (68%)	1.726 (−14%)	−0.981 (−1%)	0.374 (2%)	4.45	13
Ref. Values (800 members)	0.160	1.737	1.510	−0.987	0.366	5.00	16
M0680610	0.153 (4%)	0.292 (79%)	1.567 (−6%)	−1.093 (6%)	0.389 (1%)	4.11	13
Ref. Values (800 members)	0.159	1.418	1.478	−1.028	0.384	5.00	16
O3401010	0.166 (2%)	1.003 (71%)	1.628 (−6%)	−0.906 (5%)	0.372 (3%)	4.26	13
Ref. Values (800 members)	0.169	3.459	1.540	−0.861	0.360	5.00	16
Q2593310	0.159 (3%)	0.587 (73%)	1.147 (21%)	−0.940 (−5%)	0.390 (4%)	4.03	12
Ref. Values (800 members)	0.163	2.148	1.460	−0.984	0.374	5.00	16

HESSD
8, 2783–2820, 2011

Table 4. Test based on the normalized sum with unit weights in new catchments and different forecast time horizons (FTH) of regional integration given by the analysis of clusters by location. Values lower than 5 determined that the scores of selection are better than the reference set. See clusters' distribution at the Fig. 1. In each cluster, the catchments highlighted in bold represent the series that are not used by the members' selection methodology.

FTH	Cluster 1								Cluster 2						
	H24	K17	U25	**K13**	**K52**	**U06**	**U24**	**U27**	J85	K73	M04	M06	**H93**	**M15**	**M36**
1	4.86	4.90	4.77	4.94	6.08	4.94	4.78	4.71	4.83	4.87	4.89	4.88	4.93	4.81	4.80
2	5.03	4.97	4.77	4.92	5.40	4.72	4.33	5.51	4.91	4.87	4.83	4.70	4.76	4.91	4.83
3	5.61	7.74	4.90	4.21	4.93	5.29	5.23	4.86	4.41	4.56	4.43	4.44	4.60	4.37	4.35
4	4.34	4.55	4.52	4.62	4.79	4.52	4.63	4.67	2.69	4.99	1.20	4.20	4.50	2.56	−1.81
5	4.92	4.66	4.59	4.68	4.87	4.69	4.80	4.65	4.80	4.44	4.71	4.90	4.91	4.69	4.48
6	4.95	4.79	4.79	4.68	4.92	4.68	4.92	4.78	4.60	4.51	4.54	4.35	4.41	4.68	4.42
7	4.46	4.46	4.42	4.44	4.66	4.42	4.55	4.39	5.08	5.05	4.92	5.01	4.79	4.78	4.91
8	4.34	4.25	4.22	4.22	4.68	4.24	4.36	4.22	4.46	4.68	4.77	4.56	4.44	4.83	4.55
9	4.48	3.95	4.05	4.02	4.32	4.12	4.26	4.19	4.19	4.35	4.32	4.35	4.25	4.43	4.27

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

HESSD
8, 2783–2820, 2011

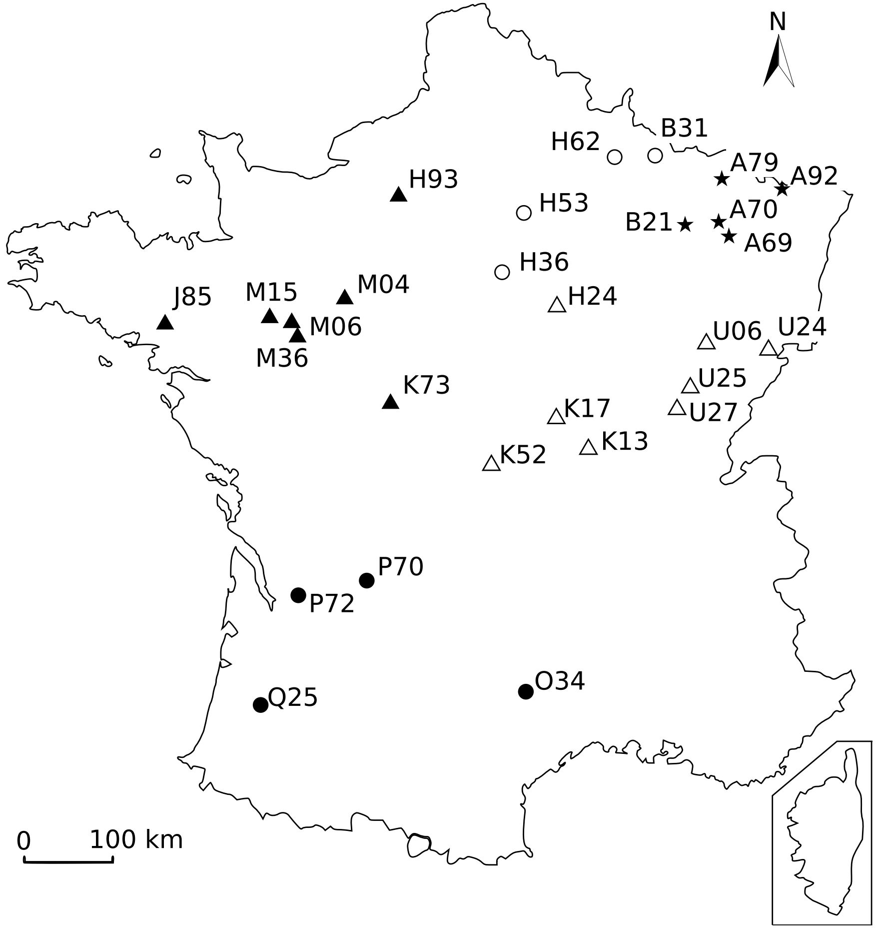

Fig. 1. Selected catchments. Each catchment is identified with the first three digits of each code used in Table 2.

HESSD

8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

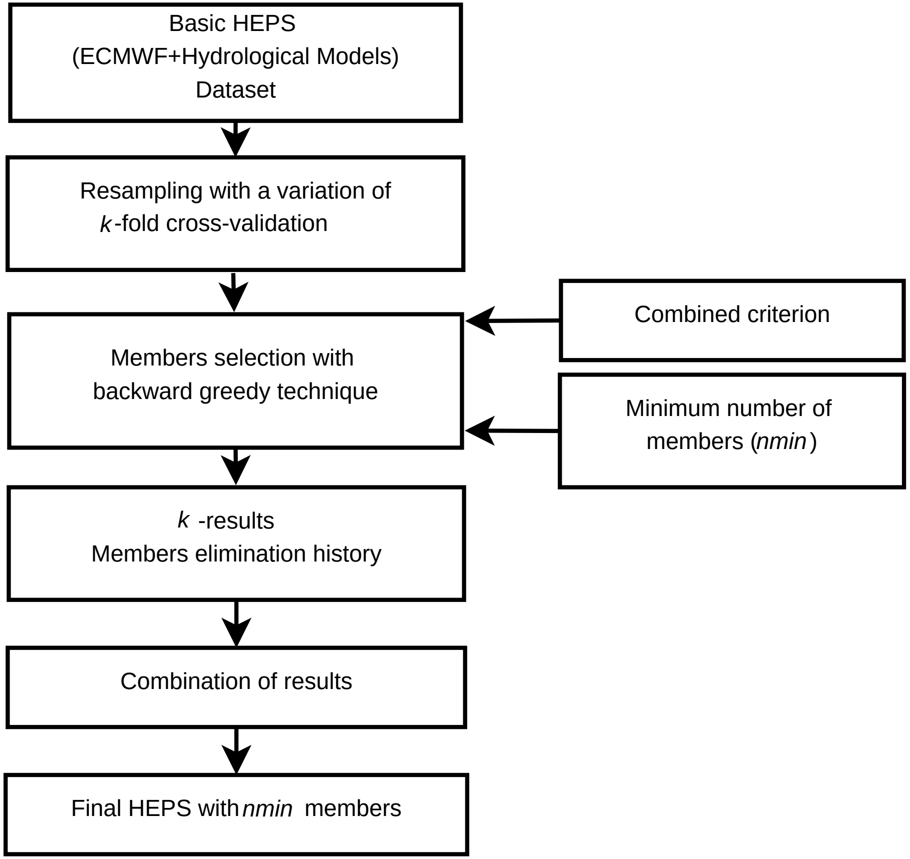

Fig. 2. Members selection methodology.

HESSD
8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

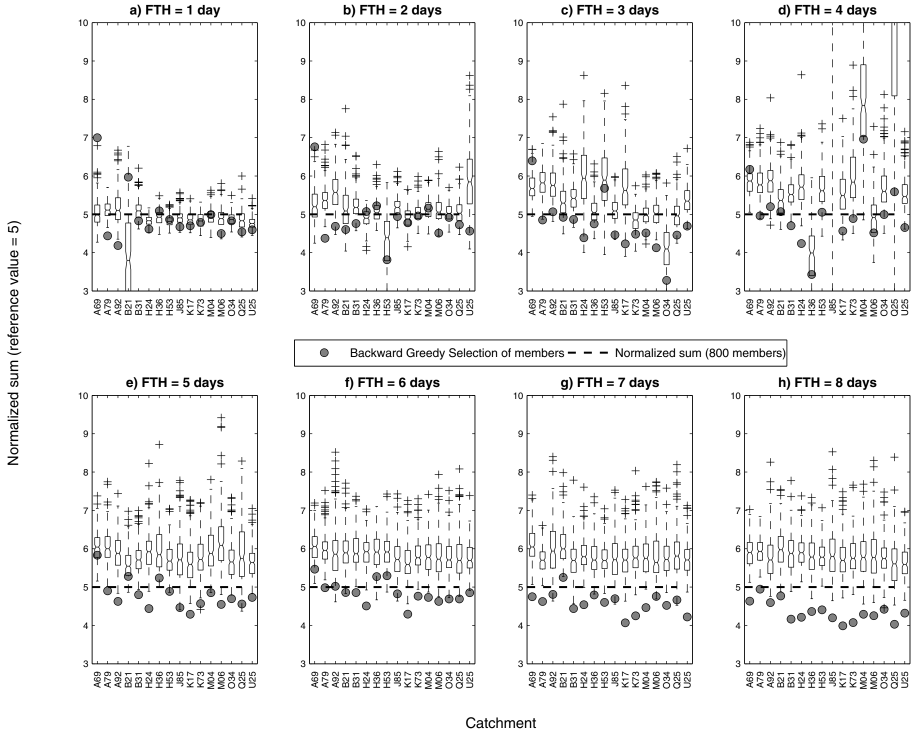

Fig. 3. Box-plot diagrams to evaluate the response sensibility with regard to 200 random experiments in different forecast time horizons (FTH). Each catchment is identified with the first three digits of each code used in Table 2. The crosses represent the outliers detected by the assumption of normality in the box plot diagram. In an effort to not lose sight of the behavior around the NS equal to 5 and the extent of the boxes, random results greater than 10 have been truncated.

HESSD
8, 2783–2820, 2011

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

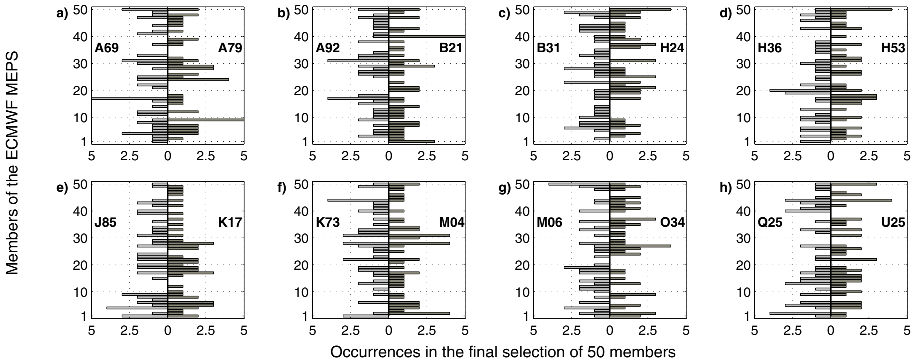

Fig. 4. Evaluation of the ECMWF members participation in the final selection of 50 members in the 16 catchments evaluated.

HESSD
8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

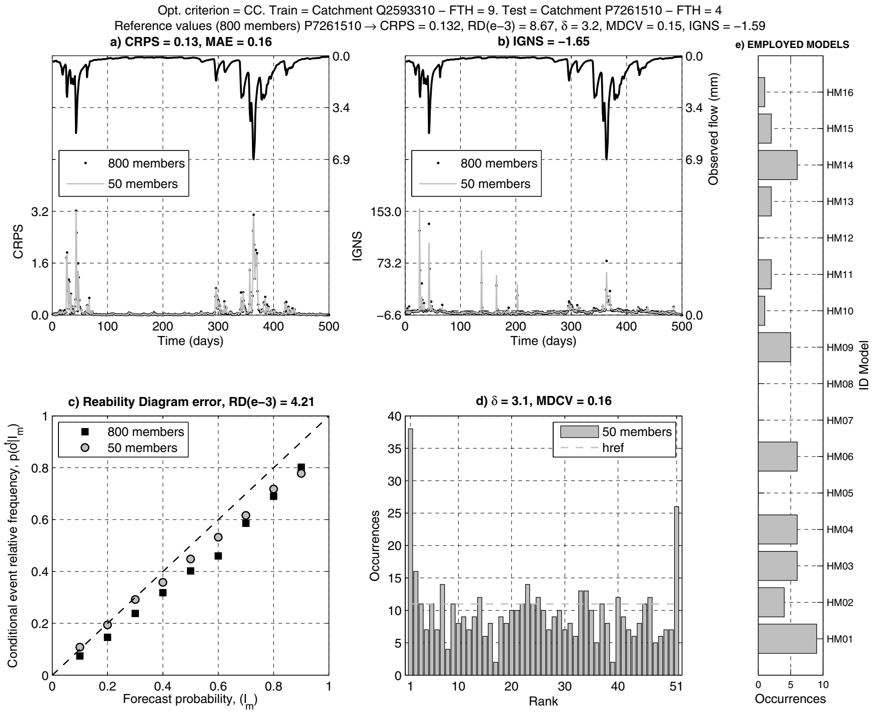

Fig. 5. Comparison between the initial ensemble in catchment P7261510 (800 members), FTH = 4 days, and the 50 selected members in catchment Q2593310 with FTH = 9 days. **(a)** Figure above: observed flow; figure below: CRPS. Note the correspondence between higher observed flows and higher CRPS. **(b)** Figure above: observed flow; figure below: IGNS. Note that there is no full correspondence between the higher IGNS and higher observed flow. **(c)** Reliability Diagram error (MSE based on vertical distances between the points). **(d)** Rank histogram for the 50 selected members. The horizontal dashed gray lines indicate the frequency ($N/d+1$) attained by a uniform distribution, **(e)** occurrences of the employed models in the final solution of 50 members.

HESSD
8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.

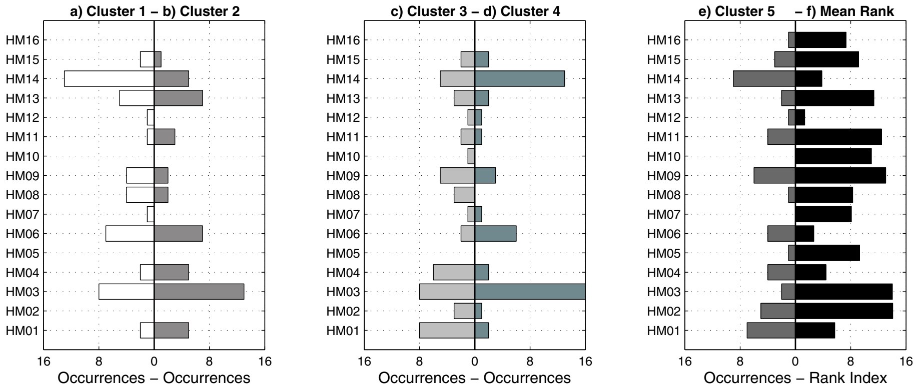

Fig. 6. Hydrological Models participation. **(a)** Distribution in the five regions (clusters) presented at the Fig. 4. **(b)** Model performance evaluated as the mean rank.

HESSD
8, 2783–2820, 2011

Simplifying a hydrological ensemble prediction system – Part 2

D. Brochero et al.