Hydrol. Earth Syst. Sci. Discuss., 8, C2226–C2248, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C2226/2011/

Hydrology and Earth System Sciences Discussions

Interactive comment on "

Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria" *by* D. Brochero et al.

D. Brochero et al.

darwin.brochero.1@ulaval.ca

Received and published: 15 June 2011

1 Why is the member's selection done ?

Today, the availability of the Meteorological Ensemble Prediction Systems (MEPS) and its subsequent coupling with multiple hydrological models offers the possibility of building Hydrological Ensemble Prediction Systems (HEPS) relying on a large number of members. However this task is complex both in terms of the coupling of information and computational time, which may create an operational barrier.

So, the selection of members within a HEPS may be viewed as a post-processing stage which seeks to maintain or improve the quality of probabilistic forecasts with a number of "x" members drawn from a super ensemble of "d" members ($x < d$), allowing the reduction of the computational time required to issue the probabilistic forecast.

2 General Comments

2.1 ...there is one point with which I don't agree. This is about the participation of the members of the ECMWF EPS in the selected 30-member ensemble (p.2757, l.18-24). I don't see the relevance of using the member's numbering as criteria since at each new forecast, the 50 initial states are assumed to be equally likely (as written p2749, l.11 & p.2757, 21-24!).

You are right. We agree that it is necessary to rewrite some paragraphs of this article, in particular the paragraph that you quote (p.2757, L.18 -24). Indeed, we are not selecting members of the ECMWF MEPS, but the product of their filtering by many different non-linear hydrological models.

This observation, as a key issue in HEPS conformation, leads us to introduce and manipulate the interchangeability of members as a variable in the selection process. So, we propose to insert explicitly this discussion and its implications in the selection

process through:

- A new section called “Interchangeability of the MEPS and HEPS members” (see below).
- The elimination of Fig. 4 of this article (the histogram in this figure leads to disorientation, moreover this space is necessary for the discussion above).
- Random combinations evaluations as a justification of the selection process (see Fig. 5 and 6).

Interchangeability of the MEPS and HEPS members
We are aware and agree with you regarding the interchangeability of the members of the ECMWF EPS. Nonetheless we stress the fact that the selection task performed here is not made on members of the ECMWF EPS but rather on the hydrological response of the 16 models used in the formation of the super ensemble of 800 members.

We also agree that some sections of the proposed article should be improved regarding that issue, namely that the selection focuses on the participation of the hydrological models. For instance, Fig. 4 in Brochero et al. (2011a) will be removed because it leads to some confusion.

From this point of view the final selection, result of the proposed methodology, should be directed more clearly to a method of selection and weighting of hydrological models based on participation of the hydrological models in the selected subset. So, if for example the final selection of 30 members shows a participation of the hydrological models 1, 4, 6, 13 and 14 to 10, 3, 5, 2 and 10 members respectively, then it immediately follows that hydrological models 1 and 14 have a strong influence as a direct interpretation of the selection made.

In order to illustrate the interchangeability of the members of the ECMWF EPS and equiprobability of this system, Fig. 6 shows both the performance of the subset found

with the Backward Greedy Selection methodology proposed (BGS) and the boxplot diagrams of 200 random experiments of 50 members with the guidance of the BGS solution (Fig. 6a), and without any guidance (Fig. 6b). The random selection on the solution oriented by BGS is based on the usage of the ratio of members per model given by the BGS methodology. For example, based on the example above, the random selections should retain 10 members for the hydrological model 1 and 14, and 4 members for the hydrological model 3, and so on.

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

The performance of selections based on the proportion of members found in the BGS solution is evident in Fig. 6a since the third quartile Q_3 (top line of boxes) is, in 90% of the cases, less than the normalized sum reference (except for the catchment B21 where $Q_3(NS) > 5$). So, it is demonstrated that the proportion of members for a hydrological model is a sufficient criterion to reduce the number of members while improving the balance of the scores represented by the normalized sum. For comparison, Fig. 6b illustrates the system response to random selections without any a priori guidance, showing that in all cases the normalized sum is greater than 5 and have recurring extremes greater than 7.

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

The generalization of the BGS method is discussed in detail in the companion paper, where the temporal and spatial extrapolation is executed for a nearby catchment. However, Fig. 6a shows that the catchments H36, K73 and U25 obtained combinations with a normalized sum lower than those obtained with the BGS method, which can be associated with the integration of experiments carried out in a subdivision database for

each catchment or the BGS algorithm structure – it is known that the classical BGS algorithm is unable to detect the collective influence of these variables.

2.2 The authors (p.2756, l.7-10) compare the results with the ensemble of 16 hydrological models driven by the deterministic forecasts. Maybe analyzing the results without using the "mean rank of elimination" but taking advantage of the resampling procedure (Section 5) could explain the apparent difference. This major criticism might be withdrawn if multi-model meteorological ensembles could be used or – possible to achieve with the same material – if a reduced number of ECMWF EPS members were drawn randomly.

In this regard, the comparison between the two schemes studied by Velázquez et al. (2011) does not use the mean rank of elimination. This measure, proposed here, is used for the integration of cross-validation results. In this case your observation takes into account three key issues in this study:

- Previous results on the number of members and the HEPS conformation: Velázquez et al. (2011) have shown, based on the database of the present paper, that the ensemble predictions produced by a combination of several hydrological model structures and meteorological ensembles (800-member set) have higher skill and reliability than ensemble predictions given either by a single hydrological model fed by weather ensemble predictions (50-member set) or by several hydrological models driven by a deterministic meteorological forecast (16-member set). So, our goal was focused on at least replicating the good quality of the 800-member set with fewer members.
- Implications of the BGS method and the length of the series: In some algorithms, such as the BGS, the overfitting[1] is highlighted as a structural problem. So, one

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

method for improving generalization which is called early stopping (Hudson et al, 2011), well-know in neural network community, is used in the methodology proposed here.

In this technique the available data is divided into three subsets. The first subset is the training set, which is used in BGS for sequentially removing the members. The second subset is the validation set. The error on the validation set is monitored during the training process. The validation error normally decreases during the initial phase of training, as does the training set error. However, when the selection begins to overfit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of members the training is stopped. The test set error is not used during training, but it is used to compare different models.

This dataset subdivision, combined with the short length of the series, imposes the use of resampling techniques such as cross-validation, which maximizes the utilization of the available information. Moreover, methodologically the combination of experiments with the so-called mean rank of elimination is shown as a mechanism avoiding overfitting in the solution found with BGS.

- Random selection of members as the performance criterion of BGS: As has already been introduced, this paper will be complemented by this analysis to show the usefulness of the BGS.

2.3 The choice of the scores should be presented more carefully... These aspects should be clearly defined in Section 2 (with appropriate references) and the choice of the scores should be shown to cover all of them.

To define or choose the error function(s) used in the selection of members methodology with BGS, we quote some of the features that are evaluated in probabilistic forecasting.

The reader is referred to Murphy (1993) and Wilks (2005) for a detailed description of these features.

- Bias: correspondence between mean forecast and mean observation.
- Reliability: correspondence between conditional mean observation and conditioning forecast, averaged over all forecasts.
- Resolution: degree to which the forecasts sort the observed events into groups that are different from each other. It is related to reliability, in that both are concerned with the properties of the conditional distributions of the observations given the forecasts.
- Sharpness: variability of forecast as described by distribution of forecast.
- Consistency: degree to which the ensembles apparently include the observations being predicted as equiprobable members.

Additionally, we propose the use of the diversity, concept studied in machine learning in some multiple classifiers systems. So the following paragraph extracted from Kuncheva (2004) summarizes this concept: "If we have a perfect classifier[2] that makes no errors, then we do not need an ensemble. If, however, the classifier does make errors, then we seek to complement it with another classifier, which makes errors on different objects. The diversity of the classifier outputs is therefore a vital requirement for the success of the ensemble."

Thus, the scores used in this research have been chosen because they quantify different aspects of ensemble prediction's quality. So, CRPS simultaneously evaluates reliability, resolution and uncertainty (Hersbach, 2000; Gneiting and Raftery, 2007).

[2]In the supervised category (called also supervised learning), each object in the data set comes with a preassigned class label. Our task is to train a classifier to do the labelling "sensibly" (Kuncheva, 2004).

The logarithmic or ignorance score, described in detail by Roulston and Smith (2002), is called to evaluate the sharpness or spread (Vrugt et al., 2006) and strongly the bias, since positioning the observation in forecast regions of low probability lead to values that tend to infinity. Reliability is directly evaluated by the RD_{mse} and, the consistency and the bias of the ensemble is assessed by the delta ratio. Finally, the maximization of the MDCV function (or minimization of the relationship $z_2 -$ MDCV) seeks to increase the diversity of the ensemble, which is equivalent to increasing its spread.

3 Specific Comments

3.1 p.2744, l.1 Add the reference to Appendix A.

You are right. Prior to presenting each score the reader will be directed to Appendix A to understand the respective formulation. Thus, the presentation of each score will focus on:

- The main feature related to the score.
- Measuring scale.
- Calculation assumptions.

3.2 p.2744, l.2 Herbasch (2000) shows how to the compute CRPS of an ensemble without the need to assume normality.

In lines 21-27 p.2743, we put in evidence the little difference in results when using the gamma or normal distribution for the studied database, despite the enormous compu-

tational cost, estimated at 1.7 h for the evaluation assuming a normal distribution and 47 h in the case of assuming a gamma distribution.

3.3 p.2748, l.19 Explain SAFRAN. l.20 from.

Given the complexity of the model SAFRAN, it will be referenced to Quintana-Seguí et al (2008) for readers who wish to know in depth one of the key elements in the HEPS conformation here studied.

3.4 p.2750, l.2 Daily data are probably observed around 6 UTC and if the 0 UTC forecasts are used, rainfall predictions are accumulated from 6 to 30, etc. Please clarify.

Forecasts are issued at 12:00 UTC and extend over 240 h. Rainfall amounts were accumulated at 24 h time steps, starting at 0 h to match with observed daily data, which resulted in nine daily lead times. No bias removal or disaggregation was performed (Velázquez et al., 2011).

3.5 p.2751, l.4 I don't understand the end of the first sentence of this paragraph.

To clarify this paragraph we propose the following: In Machine Learning the evaluation of multiple models for simulation or prediction of an event, and to further select those which together enhance or simplify a condition for adjustment, is known as "overproduce and select" (Kuncheva, 2004).

3.6 p.2752, l.12 Explain "consistency" and tell why a minimum of 30 members has been chosen.

A short definition of consistency can be found in Sect. 2.3 of this report. A more detailed definition and interpretation, drawn from Wilks (2005) is:

"A necessary condition for ensemble consistency is an appropriate degree of ensemble dispersion. If the ensemble dispersion is consistently too small, then the observation will often be an outlier in the distribution of ensemble members, implying that ensemble relative frequency will be a poor approximation to probability. If the ensemble dispersion is consistently too large, then the observation may too often be in the middle of the ensemble distribution. The result will again be that ensemble relative frequency will be a poor approximation to probability. If the ensemble distribution is appropriate, then the observation may have an equal chance of occurring at any quantile of the distribution that is estimated by the ensemble".

With regard to the minimum number of members, which was arbitrarily defined as 30 here, his choice is mainly due to the high availability of initial members (800), for example with 30 members is reached a level of compression of information equivalent to 96.25%. It is certain that if the selection task had started with a pool of 50 members, then the minimum number of members could had been defined as 10, for example. Moreover, the minimum number of members is just a stopping criterion of selection with BGS because the number of members to define as optimal should focus on specific analysis in each basin. Fig. 5 presents an example of such an analysis based on the number of members.

3.7 p.2751, l.21 & p.2752, l.2-4. Isn't it a contradiction?

No, line 2 p.2752 possibly could have caused such confusion, which will be rewritten as follows:

The member "$\vec{y_j}$" corresponds to the one that has the greater impact on the training set error (i.e. minimise train error the most).

It is important to note that the notation used for the exclusion of "$\vec{y_j}$" member of the ensemble $\mathbf{G}^{\text{iter}+1}$ is $\mathbf{G}^{\text{iter}+1} \backslash \vec{y}_j$.

3.8 p.2753, l.12-15 An easier link with the subdivision of the dataset in three subsets explained in Section 4 should be provided.

As it was discussed in Sect. 2.2, Hudson et al (2011) clearly presents the reasons and the idea of subdividing the data into three subsets to improve the generalization, in that case of artificial neural networks. This concept was also applied here in the case of the BGS.

3.9 p.2754, l.1-2 & l.14-17. Could be also enhanced by the independence between the EPS members.

You are right. So, the new paragraph will be: However, the variability of each experiment, given by the cross-validation technique and possibly by the independence between the MEPS members (input in the HEPS studied), increases the probability of reaching different member selections.

3.10 p.2755, l.4 This should be already announced in Section 3,3 (Fig. 2).

Section 3.3 presents the results of individual scores under two schemes analysed by Velázquez et al. (2011). These are:

- 16-member ensemble (16 hydrological models are driven by the deterministic forecast from ECMWF).

- 800-member ensemble (16 hydrological models are driven by the 50-member forecast from ECMWF).

Instead, the paragraph in question shows the conceptual treatment of the database to compare the scores studied in this paper.

3.11 p.2757, l.27 NS should be defined in Section 2 and justified regarding CC.

You are right, to facilitate understanding of the paper, it is also important to introduce in section 2 all the elements used in the comparison of results. Thus, using the new description of the CC from line 24 p.2747 (presented below), the elements of comparison of members' selection with respect to the 800-member set will be introduced in a new section 2.7.

2.6 Combined criterion ...
To define an approach that combines the joint evaluation of the features of the probabilistic forecasting (Sect. 2.3 in this report) the following guidelines define the conceptualization of Eq. 1 proposed in Brochero et al. (2011a):

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

- The combination should assign weights to each of the scores as a direct measure to prioritize some of the characteristics of HEPS in evaluation. Additionally, these weights, in a general framework, offer the possibility of constructing trade-off among different objectives known as Pareto fronts (Marler and Arora, 2004). In our case, weights were used only to give priority to the reliability in the selection, because Velázquez et al. (2011) showed that this was the most influential

aspect in the evaluation of the HEPS studied here. For this reason the weight assigned to the reliability corresponds to twice that of the other factors, which have a unit weight.

- To establish the main goal of the selection of members under the conservation of the different scores calculated in the initial ensemble of 800 members (ie subscript), and also to put each component on the same scale, we define the normalization of each score in the selected ensemble of members (se subscript) from the division by the corresponding score in the 800-member set.

- All scores except the MDCV function are oriented to direct minimization. However the IGNS has the peculiarity of having negative values, making necessary to establish in the normalization a threshold (z_1) to manipulate the duality of having a positive (or negative) score in the selection and a negative (or positive) score in the 800-member set. Thus, we establish $z_1 = -2$, since the preliminary analysis of selection under different scenarios (different catchments and number of members to be selected) showed minimum values for this score of about -1.5. With regard to the MDCV function, the threshold $z_2 = 1$ simply changes the orientation since the objective is to maximize dispersion, again different scenarios showed maximum values of about 0.8.

2.7 Elements to compare the performance of members' selection

Note that the CC could be used to compare the performance of the members' selection with respect to the 800-member set. So, in a general framework, if all features of the ensemble forecast have the same importance, one members' selection with equal performance to the 800-member set will lead to a CC equal to 5. Values lower than 5 indicate a selection of higher performance than the base set of 800 members, and values greater than 5 indicate the detriment of any feature of the 800-member set. Hereafter this particular condition of unit weights in the CC will be called normalized sum (NS), this distinction is important to display the priority that can be defined a priori

to any feature in members' selection training with BGS. In this way, it is possible to define a gain index for the scores balance with respect to 5 (Eq. 2).

$$\mathsf{G_{NS}}(\%) = 100 \times \left(\frac{5}{\mathsf{NS}} - 1 \right) \tag{2}$$

It is possible that the NS evaluated in the selected sets with BGS hides undesirable effects on the balance of the scores, for example to substantially improve a score over the other(s) score(s). To check this condition, a gain index for each score is also proposed (Eq. 3). A positive index indicates superior performance of the selected set. The absolute value in the denominator is needed to assess the performance of IGNS, which can take positive and negative values.

$$\mathsf{G_{sc}}(\%) = 100 \times \frac{\mathsf{Score_{ie}} - \mathsf{Score_{se}}}{|\mathsf{Score_{ie}}|} \tag{3}$$

3.12 p.2761, Eq.10 & p.2762, Eq. 11 should be defined earlier

See Sect. 3.11 of this report.

3.13 Table 3. The distinction between "deterministic" and "probabilistic" HEPS seems inappropriate since both are probabilistic.

You are right. The new title for this table will be: Table 3. Performance for the 16-member ensemble (16 hydrological models are driven by the deterministic forecast from ECMWF) and the 800-member ensemble (16 hydrological models are driven by the 50-member forecast from ECMWF).

3.14 Fig. 2. The hydrograms are difficult to read. Select one. Refer to which forecast day these graphs correspond.

New figures (designed in colour for easy viewing) are given at the end of this document.

Note: Figure 4 in Brochero et al. (2011a) will be eliminated.

In addition to the interchangeability of members discussion. Figure 6, presented here, will be added.

The complete captions for the figures below are:

Fig. 1. Selected catchments for the first phase. Each catchment is identified with the first three digits of each code used in Table 1 in Brochero et al. (2011a).

Fig. 2. HEPS results in the catchment U25 for the lead time 9. Q25, Q50 and Q75 represent the first, second and third quantile. Note that 800-member HEPS scheme covers the two largest peaks, contrary to 16-member HEPS scheme. Note also the low dispersion of this second scheme.

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

Fig. 4. Evolution of the gain index for each score under different optimization schemes in the basin A7930610 for the lead time 9. A logarithmic scale is used on the x-axis.

The chosen optimization criterion in the selection is shown at the top of each subfigure. The lower part of each subfigure indicates the values of the normalized sum (NS) of all scores with unit weights (Eq. 7) for the number of members shown on the x-axis.

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

Fig. 6. Backward Greedy Selection (BGS) and Box-plots in 200 random experiments of 50 members for the lead time 9. (a) Random selection oriented with the frequency observed in the BGS to check the interchangeability in the 800 member-set, (b) Random selection without any guidance to check the BGS performance.

References

Brochero, D., Anctil, F., and Gagné, C.: Simplifying hydrological ensemble prediction system with a backward greedy selection of members, Part I: Optimization criteria, Hydrol. Earth Syst. Sci. Discuss., 8, 2739–2782, 2011a.

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

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 2007.

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

Hudson, M., Hagan, M., and Demuth, H.: Neural Network Toolbox - User's Guide, 9 edn., The MathWorks, Inc., Natick, MA, USA, http://www.mathworks.com/help/pdf_doc/allpdf.html, 2011.

Kuncheva, L. I.: Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, New York, 2004.
Marler, R. T. and Arora, J. S.: Survey of multi-objective optimization methods for engineering, Structural and Multidisciplinary Optimization, 26, 369–395, 2004.
Murphy, A.H.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather and forecasting, 8, 281–293, 1993.
Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, Journal of Applied Meteorology and Climatology, 47, 92–107, 2008.
Roulston, M. S. and Smith, L. A.: Evaluating probabilistic forecasts using information theory, Mon. Weather Rev., 130, 1653–1660, 2002.
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, Advances in Geosciences, 29, 33–42, 2011.
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, 2006.
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, vol. 91, 2 edn., Academic Press, Burlington, MA, London, 2005.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 8, 2739, 2011.

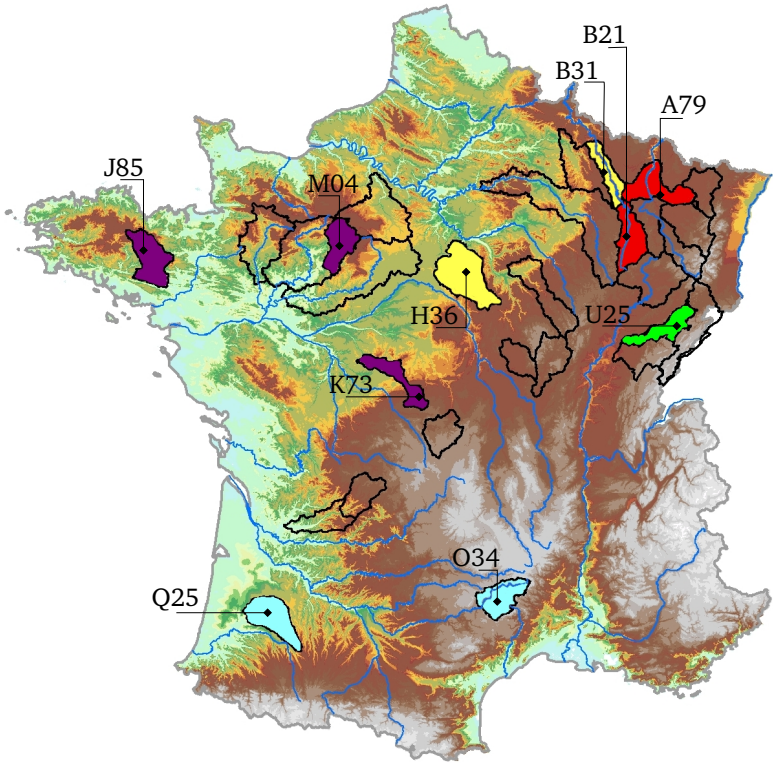

Fig. 1. Selected catchments for the first phase.

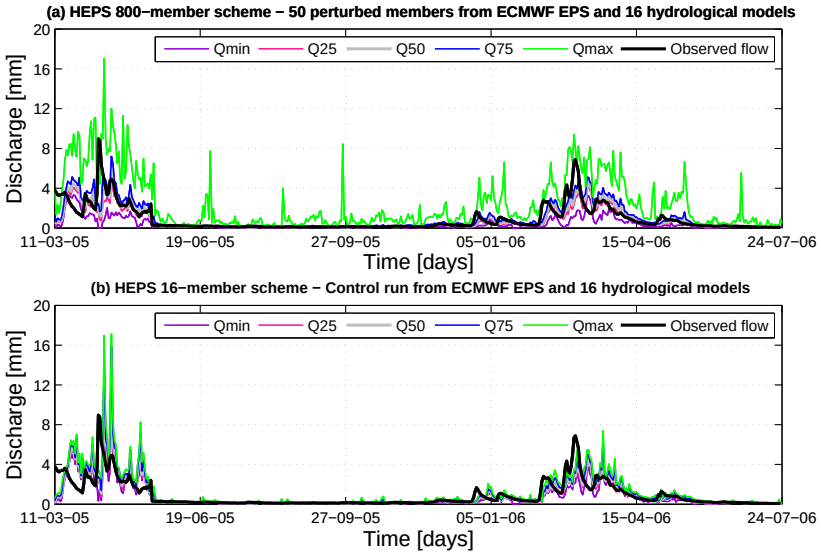

Fig. 2. HEPS results in the catchment U25 for the lead time 9.

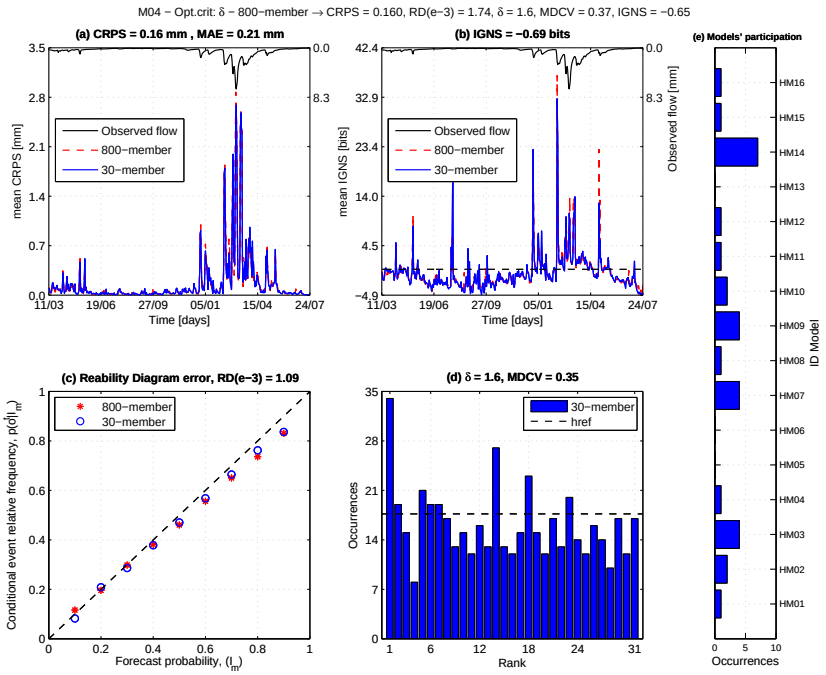

Fig. 3. Comparison between the initial ensemble (800 members) and the ensemble selected (30 members) for the lead time 9.

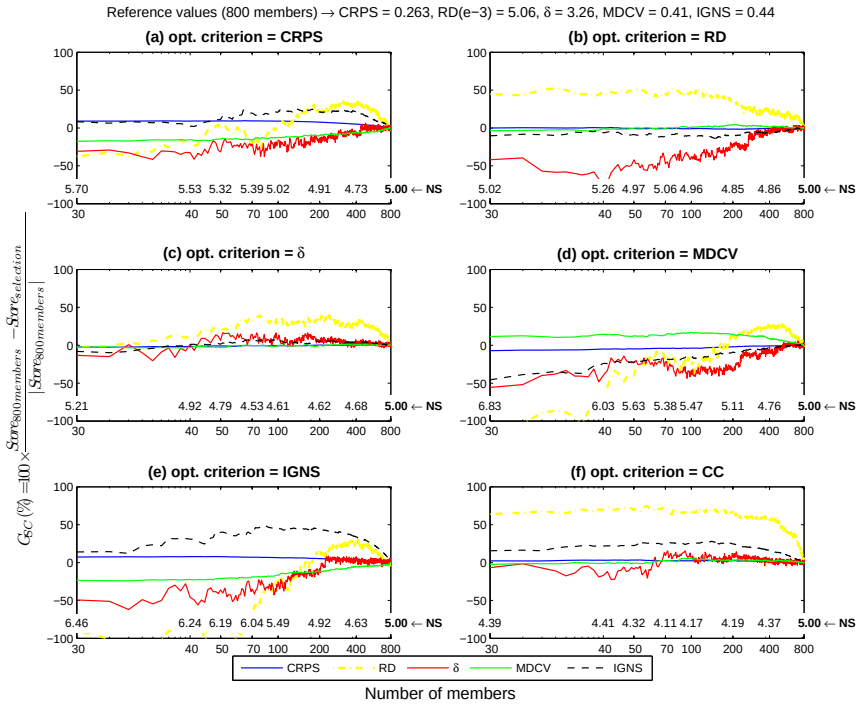

Fig. 4. Evolution of the gain index for each score under different optimization schemes in the basin A7930610 for the lead time 9.

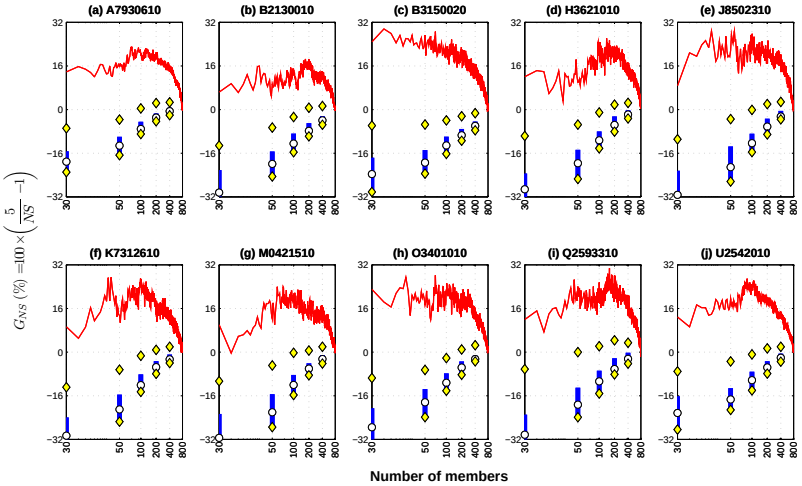

Fig. 5. Evolution of the normalized sum (NS) in terms of gain index for the lead time 9.

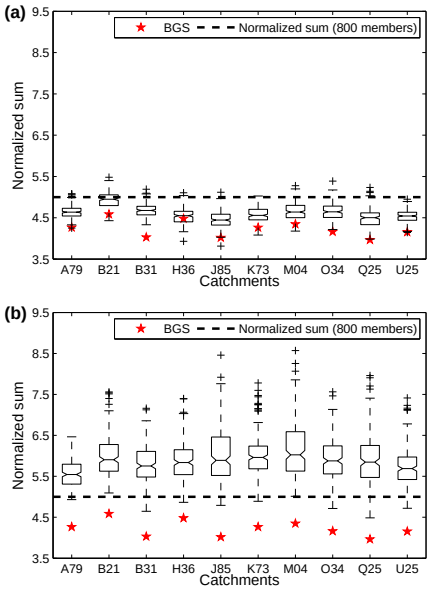

Fig. 6. Backward Greedy Selection (BGS) and Box-plots in 200 random experiments of 50 members for the lead time 9.