Hydrol. Earth Syst. Sci. Discuss., 8, C2221–C2255, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C2221/2011/

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.

To provide an idea of the complexity that can be achieved in HEPS, represented for example by the number of members to handle, it is worth mentioning the principal areas of uncertainty associated with hydrological process. It is also important to reference the resources available in the scientific community to determine the extent of possible combinations that can be achieved. Thus, the various sources of uncertainty can be summarized as follows:

- Uncertainty from the meteorological data : in this case, the MEPS are responsible for providing this information. Different centres around the world are currently working of this issue. The number of perturbed members varies between 10 and 50, for a total of 259 members in the THORPEX Interactive Grand Global Ensemble (TIGGE, Bougeault et al. (2010)). In relation with that, Bao et al. (2011) have shown that an HEPS comprised of meteorological members derived from multiple meteorological centres may actually perform better as compared to an ensemble derived from a single meteorological model (Bao et al., 2011).
- Uncertainty from the rainfall-runoff model: each hydrological model combines two important elements regarding the uncertainty associated with the hydrolog-

ical process: the initialization uncertainty (i.e. the initial state of the model) and the model uncertainty (from parameter identification to model conceptualization). In this regard, the methodology proposed by Beven and Binley (1992) provides the evaluation of parameter uncertainty from the point of view of equifinality. For example Pappenberger et al. (2005) have shown the advantages of GLUE incorporating both hydrological and hydraulic models in conjunction with the European Centre for Medium-range Weather Forecasts (ECMWF) EPS to flood inundation predictions, resulting in 3120 different inundation distributions.

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

Cloke and Pappenberger (2009) have already highlighted the computational demand of using MEPS for flood forecasting as one of the main points to overcome in the future, either by new technologies (stochastic chip technology) or by efficient use of computing clusters. Thus, the selection of hydrological members as part of a simplified model can be useful given the computational cost of running models and creating ensembles. Vrugt et al. (2008) have suggested the selection of hydrological models as an additional task that can be run based on the results of the post-processing using Bayesian Model Averaging (BMA) proposed by Raftery et al. (2005).

The members' selection with Backward Greedy Selection (BGS) proposed in this pa-

per, as well as the BMA technique, may thus be seen as a post-processing tool which assesses the contribution of each hydrological model which can be called a probabilistic subset of high performance. The qualities of BGS lie mainly in the fact that it does not assume any probability distribution in the ensembles. Likewise, the notion of interchangeability of the meteorological members (e.g. ECMWF EPS) does not condition the use of the selection technique, because the occurrence of participation of hydrological models in the subset of selected members is sufficient to guide the post-processing, as shown below in this discussion.

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

2 Interchangeability of the MEPS and HEPS members (Major observation)

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.

3 Applicability

After discussing the influence and the way in which it is necessary to interpret the results of BGS according to the interchangeability of the MEPS as a member of the HEPS base, we must admit, as the reviewer has suggested, that the paper lacks clarity on the scope and application of the proposed methodology. However, it is worth stating that the BGS analysis seeks to assess the weight that each model must represent within the response of a subset that offers the same or better performance than the reference set (800 members). This is the ultimate goal establishing a simplified model that would issue real-time forecasts in a relatively short computational time.

The enormous computational cost of coupling 50 ECMWF EPS perturbed members and 16 hydrological models, as well as further analysis of the selection of members with the BGS methodology, would be a step to execute each time that the centre forecasting manager considers it appropriate (e.g. the availability of more extreme events), since we assume in advance that a greater amount of information would improve the quality of the selection of members and therefore the forecast. Moreover, the availability of a greater number of extreme events is a desirable characteristic in the database, since one of the added values of the probabilistic forecast is mostly concentrated in the evaluation and impact of such events.

4 General Comments

4.1 If the ECMWF EPS is quite consistent during the 2005-2006 period, it evolves often, modifying its characteristics (probabilistic, spatial, etc.... . .). Moreover the hydrological model can also evolve. How reliable can thus the results of this study be after modification of this features?

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

4.2 One of my concerns is about the study period. This period is rather short, only 17 months, which is already a low limit for drawing conclusions about HEPS and their scores because of the small number of extreme (i.e. interesting) events during this period. The authors are aware of this problem but do not discuss it in the article. In the study, this period is split in training, validation and test periods, which reduces a lot their length. It seems to me important to extend the study period, even if I know that the CPU cost is enormous.

Although the verification period is relatively short, it is sufficient to cover the hydrological cycle in the basins studied, furthermore this period was also used by Velázquez et al. (2011) in the study used to compare the selection of members shown in this article. It is also important to note that the period under review (11 March 2005 to 31 July 2006) was used only for the evaluation of the forecasts. The forecast verification period is thus independent of the calibration/validation period of the hydrological models

employed.

Furthermore, the usefulness of ECMWF EPS coupled with hydrological models for flood forecasting has already been demonstrated by several researchers, notably in the project European Flood Alert System – EFAS (Ramos et al., 2007).

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

The present paper explicitly demonstrates that the cross-validation, as a vital part of the proposed methodology systematically dealing with the issue of the short length of the series (see the existing reference to this dilemma in the following lines: p.2740 line 11, p.2742 lines:18-21, p.2764 lines: 6-7).

In series or datasets that are large enough, analysed with models for which overfitting[1] does not stand out as a structural problem, it would be easy to define two subsets for model evaluation, a training and a test set. However, BGS sometimes overfits the solution. It is thus prudent to define an additional subset called validation to guide the selection in terms of the ability of generalization. Thus, the need to define three subsets to run the BGS and the short length of the series impose the use of resampling techniques such as cross-validation, which maximizes the utilization of the available information. Moreover, methodologically the combination of experiments with the so-

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

called mean rank of elimination is shown as a mechanism avoiding overfitting in the solution found with BGS.

With regard to the need to extend the evaluation period to give more validity to the results, it is certainly a feasible option and always advantageous but not absolutely necessary because the methodology was conceptualized from the start for small datasets. Furthermore, the companion paper (Brochero et al., 2011b) demonstrates the generalizability of the findings.

4.3 The 10 catchments used in this study and presented in Tab. 1 are quite small. Could you discuss how an 8x8km meteorological analysis (SAFRAN) and the even coarser ECMWF EPS are used (any kind of interpolation / disaggregation?) and can represent the processes at this small scale?

The spatial downscaling method, proposed by Rousset et al. (2007) and based on that used by the SAFRAN analysis system (Quintana-Seguí et al, 2008), was set up in order to adapt the 1.5° ensemble forecasts from ECMWF (temperature and precipitations) to the catchments scale, is described below :

- Data from ECMWF is interpolated horizontally onto the SAFRAN zones (615 non regular zones, about 20x20 km each) using distance-dependent weights ($1/r^2$ interpolation). For the forecast precipitation the value of the gradient is close to the one generally used by SAFRAN (0.7 mm year^{-1} m^{-1}). The gradient was calibrated over approximately one year (from 4 September 2004 to 31 July 2005).
- Subsequently, for each catchment, areal mean rainfall forecasts were computed by averaging the rainfall amounts of each grid above the catchment, weighted by the percentage of the catchment area inside the grid (Velázquez et al., 2011).

As regards the representation of processes at the scale of the basins of this study, the results presented by Velázquez et al. (2011) show that the combination between the downscalling of the ECMWF EPS and the 16 lumped hydrological models evaluated capture with good precision the uncertainty of the different events.

4.4 The main characteristics of these 10 catchments are described and the authors state that it represents “a large range of hydro-climatic conditions”. Thus it would be interesting to try to link the results to these hydro-climatic conditions. Are the improvements higher for smaller/larger basins? For rainy/dry basins? Etc. . .

This observation is very interesting in the light of the known effects of scale that meteorological models can have (see Thirel et al. (2008)), however the results show that the selection of members provide equal or better results than those obtained with the whole reference set (800 members) regardless the catchment in evaluation. It should be noted that work done by Velázquez et al. (2011) is shown explicitly as the combination of perturbed members of the ECMWF EPS and 16 lumped hydrological models capture adequately the uncertainty of the hydrological process, compared to the other two schemes : 16-member ensemble (all 16 models are driven by the deterministic forecast), and 50-member ensemble (each individual hydrological model is driven by the 50-member ECMWF EPS forecast).

4.5 The backward greedy selection technique is applied through removing one by one the members that decrease the error the most. The justification of this choice is missing and has to be provided in this paper before publication. Furthermore, it should be explained what is "the error" in this case (i.e. that it is one of the given statistical scores, or the CC)

Member selection is justified by the computational cost to issue a hydrological forecast based on the combination of meteorological models and hydrological models. In this line, the selection of members without sacrificing the quality of a forecast stands out as an operational option. Thus, in a general context of selection, this problem is well-known in machine learning as Subset Selection or Feature Selection and numerous methods have been developed. Thus, there are greedy selection methods (Sequential backward or forward selection) but also methods such as integer programming and evolutionary algorithms.

Thus, the approach presented in this article is based on the simplest method to execute the selection. Precisely the work that is currently being developed focuses on the application of evolutionary algorithms in multi-score framework for the optimal selection of members.

Again, we must agree that the methodology needs a paragraph stating that to study the interaction of the different scores the BGS runs by varying the error function "E" (that it is one of the given statistical scores, or the MDCV function or the CC) to obtain the results reflected in Figure 5 and Tables 5 and 6.

4.6 I also think that a discussion is needed on the impact of selecting some members of some ensembles and using them together. All the members of a given ensemble are equally likely. Can we consider when we use 5 members from 1 ensemble, plus 26 members from another one and finally 19 from a last one (for example), that all the members of the newly created ensemble are still equally likely? If not, can we still use the rank diagram that relies on this specific assumption?

There are two approaches to answer this question: the first considers that the use of the rank histogram (RH) eventually relaxes the hypotheses regarding the distribution of ensemble members, the second is a question that is probably more complex than the object itself of this paper: how to assess the propagation of equiprobability of meteorological models under the multi-model hydrological scheme, or even more under the trend of using multiple meteorological models (e.g. the TIGGE initiative) ?

The applicability of the rank histogram (RH) to evaluate whether the ensembles apparently include the observations being predicted as equiprobable members can be discussed from the same formulations of this tool, which according to Wilks (2005) was devised independently by Anderson (1996) as Binned Probability Ensemble (BPE) forecast, by Hamill and Colucci (1997) as verification rank and, finally as Talagrand diagram by Talagrand et al. (1997).

In this respect, Anderson (1996) develops his theoretical model assuming each member to be for a unique initial condition probability distribution; however, later in this same paper, section 2c eventually relaxes the perfect model context in order to demonstrate use of the BPE method for validating ensemble forecasts. It is emphasized that as a non-parametric method, the BPE does not depend on any of the details of the probability distribution of the forecasts or the initial conditions, so a large set of ensemble forecasts can be grouped for validation without difficulty.

Similarly, Hamill and Colucci (1997), established under a different hypothesis that each forecast should have independent and identically distributed (iid) errors. However, it is

recognized that these “are unrealistically ideal assumptions because any systematic error in the forecast model can result in forecasts with non-iid errors. Similarly, if the initial conditions are not equally plausible, but some are less likely than others, then the subsequent forecasts cannot be expected to exhibit equal accuracy”.

On the other hand, Talagrand et al. (1997) in his definition of the rank histogram make no explicit assumption of the probability distribution of the members as a condition, however for testing as a case study used the ECMWF EPS, whose members are equiprobables as a feature of this meteorological model.

Finally Wilks (2005) summarizes the HR distribution hypothesis as the following: the members and the single observation have all been drawn from the same distribution.

Now, to examine the probability distribution of the response obtained with BGS, it is important to show that there are several links to connect, and thus the task becomes so complex that the use of HR is totally dependent upon eventually relaxing of the ensemble members distribution, such as has been proposed by the authors cited above. The complexity of the distribution of members in the HEPS must be viewed based on:

- The probabilistic assessment of the propagation of the uncertainty associated with the ECMWF EPS in 16 hydrological models which act as non-linear filters, in addition to the supposed independence between models.
- The selection of members of HEPS, in this paper reflects the importance of each hydrological model, and it is constant in the time in each series evaluated, however seasonal effects could further refine the selection of members and add an area of complexity to the final probability distribution of the selected set.

4.7 Did you compute scores for a selection of n (with n=30, 50, 100,. . .) members randomly chosen from the 800? Are these scores really worse than those coming from the backward greedy technique? I think this would be a good justification of the fact that the improvement comes from the selection method and not from any statistical artifact based on the number of members for example.

To reflect the power of BGS in the selection of members, we added in Fig. 5 (Fig. 6 in Brochero et al. (2011a)) the evaluation of the normalized sum (NS) with 200 random selections of 30, 50, 100, 200 and 400 members in terms of gain index. It is clear that BGS selection with positive gains are always obtained, i.e. improving the balance of the scores. Otherwise in random experiments the percentiles 10, 25, 50, 75 and 90 are shown generally in the range of negative gain index (i.e. a detriment to the balance of the criteria). This tendency is obviously stronger in random selections of 100 or fewer members where the probability of taking the most representative hydrological responses is lower. It is important to note how even in the random selection of 200 and 400 members (25% and 50% of the 800 hydrological members) the NS in 75% of the evaluations shows negative gain index.

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, Table 1 shows standardized scores with respect to the 800-member set in the case of the basin H36 for the median of 200 random selections. It is important to note that the scores evaluated in the 800-member set are excellent reference points (see Table 3 in Brochero et al. (2011a)).

Table 1 shows several aspects that highlight the importance of the systematic selection of members. This analysis is used to evaluate the sensitivity of the scores with respect to the selection of members in the database under study. So, it is possible to point out the following:

- The NS shows the direct relationship between the number of members and the conservation of the initial forecast performance (i.e. the scores of the 800-member set). It is the greatest challenge selecting a small set of members, for example 30 or 50.
- CRPS indifference to the selection of members, and to a lesser extent, both the low variability of the IGNS and the MDCV function.
- The members' selection presents its greatest challenges in maintaining or improving reliability and the consistency of the ensemble represented by the delta ratio. Therefore, to define the combined criteria, such as an error term in BGS, the reliability term (RD_{mse}) has more weight to guide the optimization in this way. At this point it should be noted that consistency has a direct relationship with reliability, although ensemble consistency does not necessarily imply that probability forecasts constructed from the ensemble are reliable in the sense of conditional outcome relative frequencies being equal to the forecast probabilities yielding a 45° calibration function on a reliability diagram, unless either the ensemble size is relatively large or the forecasts are reasonably skillful, or both (Wilks, 2011).

Finally, Table 2 shows detailed results for each score in the selection process with BGS for the basin H36. It shows that in the BGS methodology, with the combined criterion as error function, is not detrimental to any of the scores. Instead, gains in the balance scores (normalized sum) are mainly due to optimization of system reliability while preserving the quality of the other scores. At this point it is important to call attention to the results in Table 5 and Figure 5 reported by Brochero et al. (2011a), which shows the duality between reliability (RD_{mse}) and consistency (δ ratio) when the optimization is focused exclusively on the reliability as error criterion in the BGS.

5 Specific Comments

5.1 Page 2741 lines 22-23: please add "EPS" in "of the European Center for Medium-range Weather Forecasts (ECMWF) EPS". Moreover the correct spelling is "Centre", not "Center".

Done.

5.2 Section 2: could you please discuss why you chose these 5 scores and not other ones?

To define or choose the error function(s) used in the selection of members methodology with BGS, then 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 multi-assembly methods. 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). 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.

5.3 CRPS description (section 2.1): the authors should add what is the best value for this score, and what is a bad value. This could help the readers not knowing this score very well.

The following sentences will be added to the description of CRPS: Its minimal value of zero is only achieved in the case of a perfect deterministic forecast. Note that the CRPS has the dimension of the observation o^t.

[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 labeling "sensibly" (Kuncheva, 2004).

5.4 The IGNS (section 2.2) is not a classical score for hydrologists, thus I think that more efforts have to be put in order to introduce it in this paper. Like it is now, it is not clear what this score shows, what a good score is, what it brings more than the other scores. Please try to improve this part.

You are right, the new description would be:

The logarithmic score, or ignorance score (Eq. 1), proposed by Good (1952), is simply the logarithm of the ensemble probability density function ($f(\vec{y}^t)$) at the point corresponding to the observation (o^t). Roulston and Smith (2002) gave an information theoretic perspective and an interpretation in terms of gambling returns.

$$\text{IGNS}(\vec{y}, o)^t = -\log_2 \left[f(\vec{y}^t)_{o^t} \right] \tag{1}$$

The logarithmic score is strictly proper but involves a harsh penalty for low probability events and therefore is highly sensitive to extreme cases (Gneiting and Raftery, 2007). Weigend and Shi (2000) noted similar concerns and considered the use of trimmed means when computing the logarithmic score, an alternative that was also adopted in this research. The CRPS is less sensitive to extreme cases or outliers (Gneiting and Raftery, 2007). We use the logarithmic score in the negative orientation for reasons of minimization. In addition, when the observation falls outside of the predictive distribution, the corresponding probability density is zero. This produces an infinite value, which affects the calculation of the mean score. Here, we chose to replace those individual infinite scores by the next worst non-infinite value, following Boucher et al. (2010).

5.5 Section 2.4 line 10: it seems to me that Sc is not the rank of the observation, but the population of each interval.

You are right, S_c is the number of elements of the cth interval of the histogram ($c = 1, \cdots, d+1$).

5.6 Section 2.4 line 16-17: Delta is the deviation, not the flatness, please put the delta symbol earlier in this sentence.

Done.

5.7 The MDCV (section 2.5); please discuss what you consider to be the best MDCV value (from eq. (7) it seems to be 1) and why.

In an effort to justify the MDCV function, and thus clarify its use, the following paragraph will be added:

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

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

In this research, the analysis of the HEPS dispersion, through CV (results are omitted in

this article), showed an increase proportional to the lead time, so the first lead time has an average CV of 0.05 while longer lead times (e.g. 9 days), reached an average value of 0.5. Note that CV is calculated for each time step. Given the observed asymmetry in the CV series the use of the median measure is proposed. Having defined the median of the coefficients of variation (MDCV) as a function that determines the HEPS dispersion, the hypothesis under this function is that a gain in dispersion increases the reliability of HEPS, as it is shown again in Table 3 of this research, in which two HEPS combination between the ECMWF EPS and 16 hydrological models are compared.

5.8 Could you explain the sentence "Preliminary analysis showed. . ." (lines 5-6 section 2.6)? What do you mean by "covering all scenarios"? Isn't this combined criterion supposed to give the best forecast when CC is minimal?

To clarify the meaning of this sentence, we propose the following changes from line 24 in Section 2.6:

To define an approach that combines the joint evaluation of the features of the probabilistic forecasting (Sect. 5.2) the following guidelines define the conceptualization of Eq. 2 proposed in Brochero et al. (2011a):

$$CC = w_1 \frac{\overline{\mathsf{CRPS}}_{\mathsf{se}}}{\overline{\mathsf{CRPS}}_{\mathsf{ie}}} + w_2 \frac{z_1 - \overline{\mathsf{IGNS}}_{\mathsf{se}}}{z_1 - \overline{\mathsf{IGNS}}_{\mathsf{ie}}} + w_3 \frac{\mathsf{RD}_{\mathsf{MSEse}}}{\mathsf{RD}_{\mathsf{MSEie}}} + w_4 \frac{\delta_{\mathsf{se}}}{\delta_{\mathsf{ie}}} + w_5 \frac{z_2 - \mathsf{MDCV}_{\mathsf{se}}}{z_2 - \mathsf{MDCV}_{\mathsf{ie}}} \qquad (2)$$

- 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.

5.9 The Velázquez et al. (2010) reference used in part 3 to refer to the description of the HEPS is not enough. Consider replacing this EGU abstract reference with the Velázquez et al. (2011) paper published in Advances in Geosciences, which is much more complete.

Done.

5.10 Section 3: please add SAFRAN reference. If the 50 year reanalysis has been used, please add: Vidal et al. (2010)

Reference for SAFRAN was added. The 50 year reanalysis was not used.

5.11 Section 3.1: from 10-day MEPS forecasts you obtain 9-day HEPS. Are you using the EPS forecasts issued at 12:00? If yes, this information is missing.

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).

5.12 Section 3.3: “The hydrological models were calibrated with 29 years as mean length”. Could you explain where this difference comes from, since you use the same input?

The length of available observed streamflow time series varies according to the catchment, with, on average, 29 years of available daily data for the catchment dataset used here (Velázquez et al., 2011).

5.13 Section 4: it is not clear for me how the training and validation subsets are used. If I understand right, the training subset is used to find the best members with the backward greedy technique, the test subset is used for verification (scores discussed in Part 6), but what is the validation subset used for? Could you explain it better?

Partitioning into training and validation data sets guides the selection task in order to avoiding overfitting. Cross-validation resampling and the subsequent integration of the 5 experiments with mean rank of elimination further reinforce the generalization of BGS. Thus, for the 5 experiments performed in each basin we could show training and validation results. Detailed results were omitted for two reasons:

- Validation results showed in general a trend similar to the training ones, except for some experiments where the random distribution of the training and validation sets was not statistically homogeneous (see Sect. 5 in Brochero et al. (2011a)). However, as expected, in the training process the tendency was always in minimizing the target score, while in some cases validation revealed some overfitting of the selection.
- In reaction, the methodological integration of experiments was designed in a way to avoid overfitting of the selection, which could be verified in the companion paper (Brochero et al., 2011b).

Results discussed in Sect. 6 in Brochero et al. (2011a) thus correspond to a "pseudo test dataset" for comparing the performance between different scores in the process of selecting members, since the data used to minimize all error functions are exactly the same.

It is a "pseudo test dataset" because there is a high probability that the data used in testing have been used in the BGS training process, becoming the indicator of an opti-

mistic estimator of the selection (Diamantidis et al., 2000); however, we do emphasize that the first part of this research focuses on an analysis of scores in the BGS process with the subsequent integration of result, and the second phase presented in a companion paper (Brochero et al., 2011b) shows a rigorous test of generalization in time and space.

5.14 Page 2757, line 27: why unit weights are used here? The RD weight was 2, why did you modify it?

Please see Sect. 5.8 above.

5.15 Conclusions page 2763 lines 12-14: the 100 members as optimal number is not proved in this paper. This sentence should thus be less categorical.

You are right, to clarify these lines will be rewritten as follows: For example, in this study, the best balance of scores is achieved with a number of members fluctuating between 30 and 100, maximizing the qualities of the system: reliability, consistency, resolution, and diversity. So in the worst case this corresponds to a 87.5% (700 members/800 members) compression level.

5.16 References: please indicate the link for downloading Talagrand et al. (1997), because it is not so easy to find.

Done. http://www.ecmwf.int/publications/newsletters/chronological_list.html

5.17 Fig. 1: please draw the area of the 10 basins.

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

5.18 Fig. 3: it is difficult to see anything on the time series plots of CRPS and IGNS, please consider improving them.

Idem.

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

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, Journal of Climate, 9, 1518–1530, 1996.

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

Beven, K. and Binley, A.: The future of distributed models: model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.

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, 2010.

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.
Boucher, M.-A., Laliberté, J.-P., and Anctil, F.: An experiment on the evolution of an ensemble of neural networks for streamflow forecasting, Hydrol. Earth Syst. Sci., 14, 603–612, 2010.
Cloke, H. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., 375, 613–626, 2009.
Diamantidis, N., Karlis, D., and Giakoumakis, E.: Unsupervised stratification of cross-validation for accuracy estimation, Artif. Intell., 116, 1–16, 2000.
Good, I. J.: Rational Decisions, Journal of the Royal Statistical Society, 14, 107–114, 1952.
Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, 2007.
Hamill, T. M. and Colucci, S. J.: Verification of Eta–RSM Short-Range Ensemble Forecasts, Monthly Weather Review, 125, 1312–1327, 1997.
Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, 2000.
Kottegoda, N. T. and Rosso, R.: Applied Statistics for Civil and Environmental Engineers, electronic version, Wiley, Chichester, 2009.
Kuncheva, L. I.: Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, 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.
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, 2005.
Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, Journal of Applied Meteorology and Climatology, 47,

92–107, 2008.
Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.
Ramos, M. -H., Bartholmes, J., and Thielen-del Pozo, J.: Development of decision support products based on ensemble forecasts in the European flood alert system, Atmospheric Science Letters, 8, 113–119, 2007.
Roulston, M. S. and Smith, L. A.: Evaluating probabilistic forecasts using information theory, Mon. Weather Rev., 130, 1653–1660, 2002.
Rousset, F., Habets, F., Martin, E. and Noilhan, J.: Ensemble streamflow forecasts over France, ECMWF Newsletter, 111, 21–27, 2007.
Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Workshop on predictability, edited by: for Medium-Range Weather Forecasts, E. C., 1–25, Shinfield Park, Reading, Berkshire RG2 9AX, UK, http://www.ecmwf.int/publications/library/do/references/list/16233, 1997.
Thirel, G., Rousset-Regimbeau, F., Martin, E. and Habets, F.: On the Impact of short-range meteorological forecasts for ensemble streamflow predictions, Journal of Hydrometeorology, 9, 1301–1317, 2008.
Velázquez, J. A., Anctil, F., Ramos, M. H. and Perrin C.: Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures, 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.
Vrugt, J., Diks, C., and Clark, M.: Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling, Environ. Fluid Mech.,8, 579–595, 2008.
Weigend, A. S. and Shi, S.: Predicting daily probability distributions of S&P500 returns, Journal of Forecasting, 19, 375–392, 2000.
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, vol. 91, 2 edn., Academic Press, Burlington, MA, London, 2005.
Wilks, D. S.: On the Reliability of the Rank Histogram, Monthly Weather Review, 139, 311–316, 2011.

Table 1. Median of 200 random selections in catchment H36 for the lead time 9. Initial 800-member set (ie) as a normalization factor for the scores of the subset of members (se).

Selected Members	$\frac{\overline{\text{CRPS}}_{se}}{\overline{\text{CRPS}}_{ie}}$	$\frac{\text{RD}_{\text{MSE}_{se}}}{\text{RD}_{\text{MSE}_{ie}}}$	$\frac{\delta_{se}}{\delta_{ie}}$	$\frac{z_2-\text{MDCV}_{se}}{z_2-\text{MDCV}_{ie}}$	$\frac{z_1-\overline{\text{IGNS}}_{se}}{z_1-\overline{\text{IGNS}}_{ie}}$	NS
30	1.0111	1.5040	1.7997	1.0465	1.1061	6.4674
50	1.0056	1.2518	1.5331	1.0299	1.0568	5.8773
100	1.0038	1.0960	1.2833	1.0157	1.0207	5.4196
200	1.0016	1.0217	1.1113	1.0097	1.0087	5.1531
400	1.0006	0.9771	1.0333	1.0044	1.0020	5.0174

Table 2. Results of BGS in catchment H36 for the lead time 9 with the combined criterion as error function. Initial 800-member set (ie) as a normalization factor for the scores of the subset of members (se).

Selected Members	$\frac{\overline{\text{CRPS}}_{se}}{\overline{\text{CRPS}}_{ie}}$	$\frac{\text{RD}_{\text{MSE}_{se}}}{\text{RD}_{\text{MSE}_{ie}}}$	$\frac{\delta_{se}}{\delta_{ie}}$	$\frac{z_2-\text{MDCV}_{se}}{z_2-\text{MDCV}_{ie}}$	$\frac{z_1-\overline{\text{IGNS}}_{se}}{z_1-\overline{\text{IGNS}}_{ie}}$	NS
30	1.0000	1.0000	0.9610	1.0000	1.0000	4.9610
50	1.0004	0.9181	0.9896	0.9962	1.0022	4.9064
100	0.9986	0.8026	1.0139	0.9918	1.0003	4.8072
200	0.9976	0.5821	0.9707	0.9859	1.0055	4.5418
400	0.9954	0.4488	0.8829	0.9762	0.9987	4.3020

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

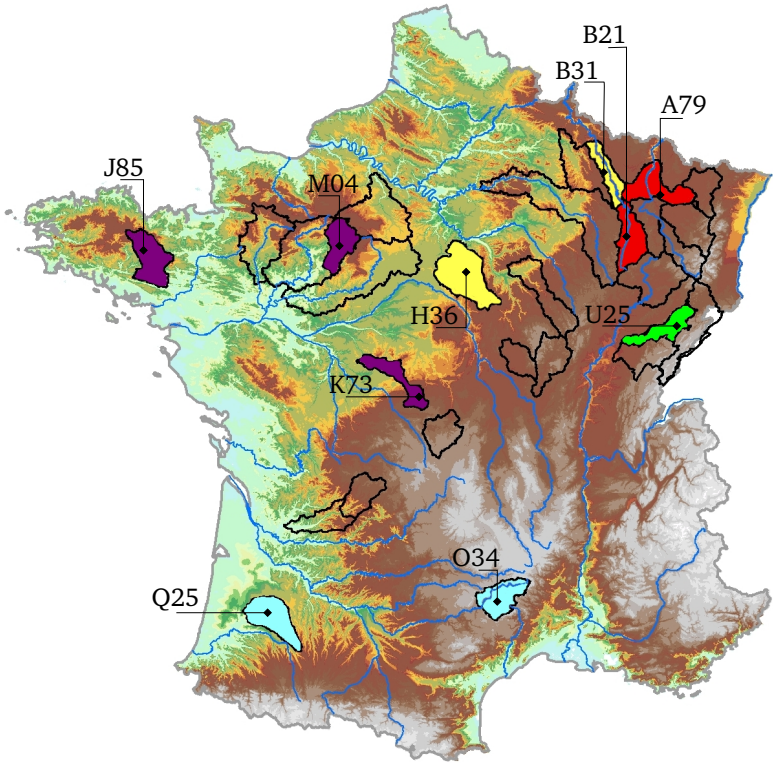

Fig. 1. Selected catchments for the first phase.

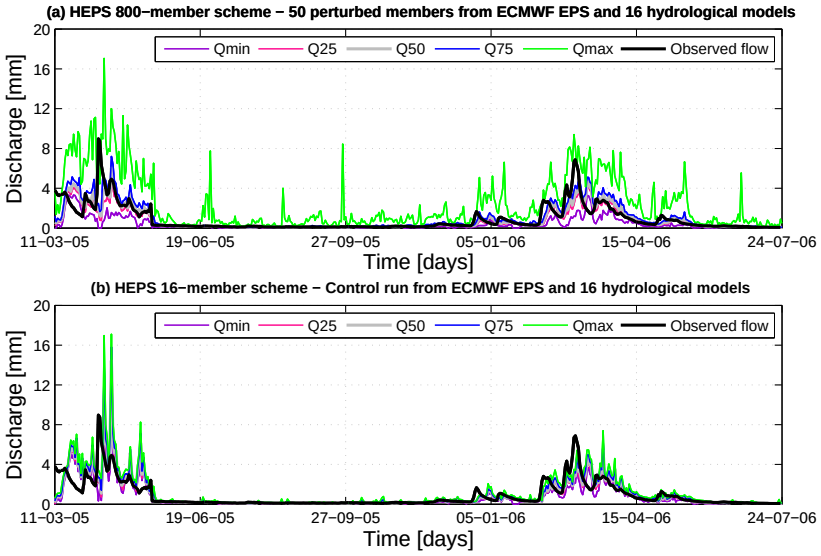

Fig. 2. HEPS results in the catchment U25 for the lead time 9.

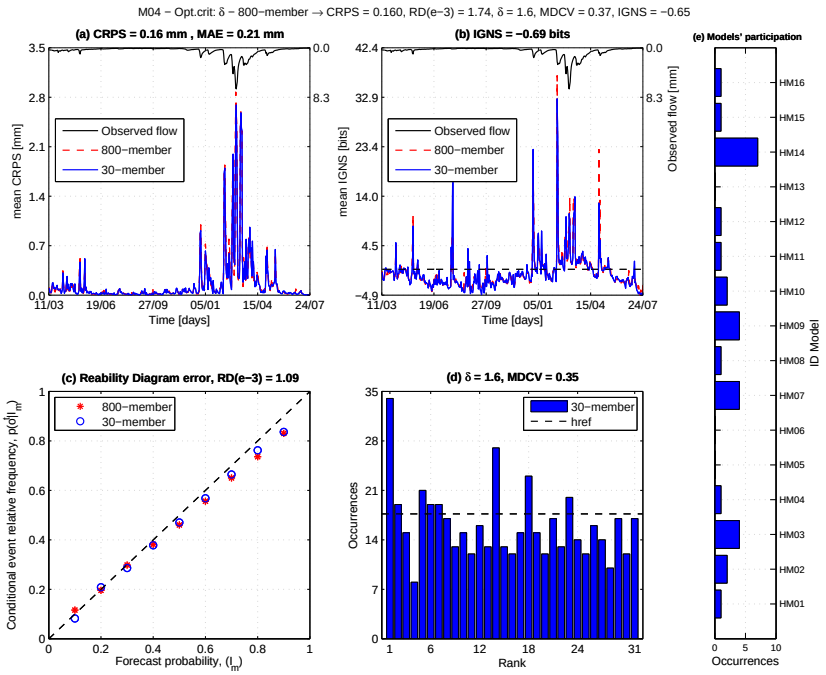

Fig. 3. Comparison between the initial ensemble (800 members) and the ensemble selected (30 members) for the lead time 9.

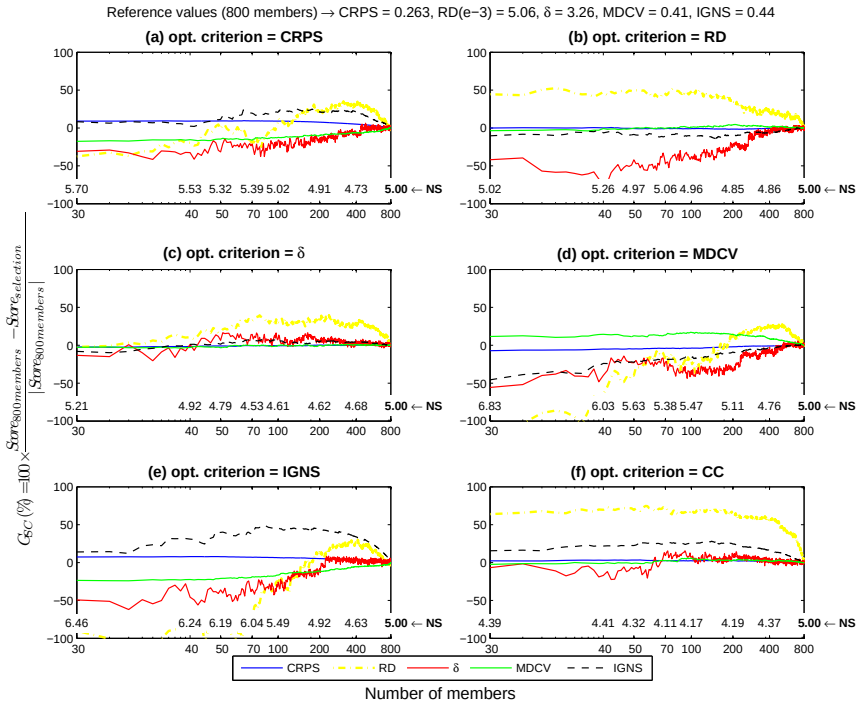

Fig. 4. Evolution of the gain index for each score under different optimization schemes in the basin A7930610 for the lead time 9.

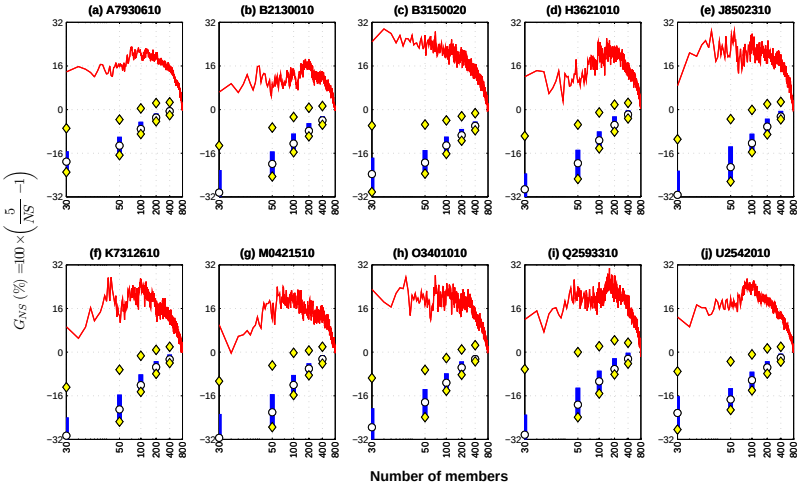

Fig. 5. Evolution of the normalized sum (NS) in terms of gain index for the lead time 9.

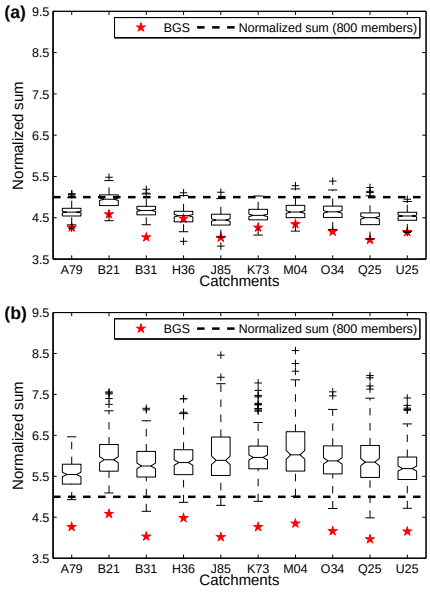

Fig. 6. Backward Greedy Selection (BGS) and Box-plots in 200 random experiments of 50 members for the lead time 9.