Hydrol. Earth Syst. Sci. Discuss., 7, 4023–4058, 2010
www.hydrol-earth-syst-sci-discuss.net/7/4023/2010/
doi:10.5194/hessd-7-4023-2010

Hydrology and Earth System Sciences Discussions

Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments

J. A. Velázquez[1], F. Anctil[1], and C. Perrin[2]

[1]Chaire de recherche EDS en prévisions et actions hydrologiques, Département de génie civil et de génie des eaux, 1065, avenue de la Médecine, Québec, Qc, G1V 0A6 Canada
[2]Cemagref, Hydrosystems and Bioprocesses Research Unit, Parc de Tourvoie, BP 44, 92163 Antony Cedex, France

Received: 26 May 2010 – Accepted: 15 June 2010 – Published: 29 June 2010

Correspondence to: J. A. Velázquez (juan-alberto.velazquez.1@ulaval.ca)

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

Abstract

This work investigates the added value of ensembles constructed from seventeen lumped hydrological models against their simple average counterparts. It is thus hypothesized that there is more information provided by all the outputs of these models than by their single aggregated predictors. For all available 1061 catchments, results showed that the mean continuous ranked probability score of the ensemble simulations were better than the mean average error of the aggregated simulations, confirming the added value of retaining all the components of the model outputs. Reliability of the simulation ensembles is also achieved for about 30% of the catchments, as assessed by rank histograms and reliability plots. Nonetheless this imperfection, the ensemble simulations were shown to have better skills than the deterministic simulations at discriminating between events and non-events, as confirmed by relative operating characteristic scores especially for larger streamflows. From 7 to 10 models are deemed sufficient to construct ensembles with improved performance, based on a genetic algorithm search optimizing the continuous ranked probability score. In fact, many model subsets were found improving the performance of the reference ensemble. This is thus not essential to implement as much as seventeen lumped hydrological models. The gain in performance of the optimized subsets is accompanied by some improvement of the ensemble reliability in most cases. Nonetheless, a calibration of the predictive distribution is still needed for many catchments.

1 Introduction

In hydrology, traditional approaches focus on a single model thought to be the best possible for a given application. In opposition, multimodel combination aims at extracting as much information as possible from a group of existing models. The idea is that each model of the group provides specific information that might be combined to produce a better overall simulation. This concept has been widely tested because no hydrological model could yet be identified as the “best” model in all circumstances (Oudin et al., 2006).

For example, Shamseldin et al. (1997) compared three combinational methods over five rainfall-runoff models and eleven catchments. The methods were the simple model average (SMA), the weighted average, and artificial neural networks. Results showed that the combined outputs were more accurate than the best single one. Later, Georgakakos et al. (2004) tested a multimodel approach over six catchments. Combined outputs were constructed with both calibrated and uncalibrated distributed model simulations, using the SMA. Results confirmed the better performance of the combined series over individual ones; furthermore, the authors claimed that multimodel simulations should be considered as an operational tool. Ajami et al. (2006) examined yet another method of combination, namely the multimodel superensemble of Krishnamurti et al. (1999), using outputs from seven distributed models. They found that more sophisticated combination techniques may further improve simulation accuracy, that at least four models are required to obtain consistent multimodel simulations, and that the multimodel accuracy is related to the accuracy of the individual member models (longer dataset and more models might then improve multimodel combination results). Viney et al. (2009) compared predictions for one catchment exploiting ten models of different model types, covering lumped, semi-distributed, and fully distributed models combined in many ways. Their results differ from Ajami et al. (2006) in that the best ensembles are not necessarily those containing the best individual models. For the same catchment and models as Viney et al. (2009), Bormann et al. (2007) suggested that a number of at least 6 models are required for a multimodel ensemble to ensure good model performance and that any number above six may not considerably improve the performance of the ensemble.

Another multimodel combinational method has been proposed by Oudin et al. (2006) who resorted to two different parameterizations of the same model.

An alternative idea, which is gaining ground, combines models through optimization. For example, Devineni et al. (2008) proposed an algorithm combining streamflow forecast from individual models based on their skill, as assessed from the rank probability score. The methodology assigns larger weights to models leading to better predictability under similar prediction conditions. This multimodel combination has been tested over a single catchment, combining two statistical models. Seven multimodel combinations techniques were tested and results showed that developing optimal model combinations contingent on the predictor lead to improve predictability.

Multimodel combination has also been applied in an operational context. Loumagne et al. (1995) combined model outputs using weights adapted to the state of the flood forecasting system. This procedure proved to be more effective than choosing the best model at each time step. Coulibaly et al. (2005) combined three structurally different hydrologic models to improve the accuracy of a daily reservoir inflow forecast based on the weighted average method. They found that model combination can offer an alternative to the daily operational updating of the models, providing a cost-effective solution to operational hydrology. Marshall et al. (2007) used a hierarchical mixture of experts (HME) allowing changes in the model structure, depending on the state of the catchment. The framework was tested on 10 Australian catchments, combining results from two parameterizations of a conceptual model. Results showed that the HME improves performance over the model taken alone.

The view shared by the above studies is the production of improved hydrological simulations through the aggregation of a group of outputs into a single predictor. The present study hypothesizes that there is more value exploiting all the outputs of this group than the single aggregated one, following the philosophy of meteorological ensemble prediction (Schaake et al., 2007). All the members of the ensemble are then used to fit a probability density function (the predictive distribution), useful evaluating confidence intervals for the outputs, probability of the streamflow being above a certain threshold value, and more. In other words, an ensemble allows appreciating the uncertainty of the simulation. He et al. (2010) used predictions from six meteorological agencies, for the Huai River catchment in China, to drive a hydrological model forecasting the July–September 2008 flood event. Their results established multimodel as

HESSD
7, 4023–4058, 2010
Performance and reliability of ensemble simulations
J. A. Velázquez et al.

a promising tool for 10-day-ahead discharge forecasts.

The present study aims assessing the added value of ensembles constructed from seventeen lumped hydrological models (the probabilistic simulations) against their simple average counterparts (the deterministic simulations). It resorts to 1061 French daily streamflow time series extending over a ten-year period, in order to generalize conclusions. The probabilistic performance based on all seventeen outputs is first compared to the deterministic one. Then the reliability of the ensembles is assessed as well as their operational value in terms of hit rate and false alarm rate. Further ensemble performance improvement is finally sought through model selection: subsets of the seventeen lumped hydrological model outputs are objectively constructed using a genetic search algorithm optimizing the Continuous Ranked Probability Score.

The methodology is described in the next section. Results are presented in Sect. 3, while conclusions are given in Sect. 4.

2 Methodology

Catchments and models are presented along scores and tools used to evaluate the performance and reliability of the ensembles. The genetic search algorithm is described last.

2.1 Catchments and models

Deterministic and probabilistic streamflow simulations from seventeen hydrological models are analyzed on 1061 French catchments. The dataset was built by Le Moine (2008) and used by Le Moine et al. (2007). Catchments are spread over the French territory (Fig. 1) in order to representing a large variety of physical conditions in terms of size, topography, geology, soil, land use, and climate, which ranges from oceanic to Mediterranean to continental (Table 1). Catchments with important snow accumulation are not included, avoiding the need for a snowmelt module. Temperature, precipita-

HESSD
7, 4023–4058, 2010

tion and flow data were available at a daily time step over a 10-year period extending from 1996 to 2005. Daily streamflows come from the French database Banque Hydro. Daily precipitation and temperature values over a 8-km grid originate from the meteorological analysis system SAFRAN of Météo-France (Durand et al., 1993). Potential evapotranspiration is estimated from air temperature, using the radiation-based formulation proposed by Oudin et al. (2005).

The first half of the time series is used for calibration, while the second half is used for validation. All results provided herein concern the validation sub-dataset.

All seventeen hydrological models are of low to moderate complexity: the number of parameters ranging from 4 to 13. Table 2 lists the tested model structures along with the number of optimized parameters and stores for their tested version. Most of these models were used by Perrin et al. (2001) and Mathevet (2005). All model structures were applied in a lumped mode. These models correspond to various conceptualizations of the rainfall-runoff transformation at the catchment scale. They all include a soil moisture accounting procedure but with various formulations (linear or non linear, possibly with several soil layers). The routing module includes from 1 to 5 linear or non linear stores, and unit hydrographs or pure time delays. Some of the models include a non conservative function (correction factors of inputs or groundwater exchange functions) used to adjust the water balance. All the models were applied in the same conditions, i.e. ran at a daily time step using the same rainfall and potential evapotranspiration inputs and calibrated with the same procedure. This single application framework provides more comparable results between model structures. This is one of the reasons why the original model structures were modified as they sometimes had specificities that did not match this framework. Note that the objective here was not to test the original structures but to have a variety of conceptualizations. To avoid confusion with the original model from which they are derived, only 4 letter acronyms are used in Table 2 and identification numbers will be used in the text and figures. Model's structure description is available from authors.

Calibration was performed using a local search procedure, as described by Edijatno

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

HESSD

7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

et al. (1999), applied in combination with a pre-screening of the parameter space as proposed by Mathevet (2005). This pre-screening provides a likely starting point for the search algorithm and limits the risks to be trapped in local optima. Mathevet (2005) showed that this approach is competitive for this type of models, in terms of efficiency and effectiveness, when compared with more sophisticated global search procedures.

2.2 Performance and reliability

Deterministic simulations were aggregated using the simple average method (SMA). This is the simplest procedure for combining outputs from an ensemble of individual models (Shamseldin et al., 1997). Ensembles were constructed in different forms. First, a simple pooling of all seventeen model outputs was considered. Then, subsets of the seventeen lumped hydrological model outputs were identified objectively using the genetic search algorithm described in Sect. 2.3 and the Continuous Ranked Probability Score as the objective function. Finally, subsets of eight models, selected according to their deterministic performance, were tested for comparison.

2.2.1 The absolute error criteria

The evaluation of the performance of the deterministic simulations is based on the absolute error (AE), a linear scoring rule that describes the average magnitude of the errors without considering their direction. The main advantage of the AE over alternative deterministic scores is that it can be directly compared to the Continuous Ranked Probability Score – described next – of the probabilistic simulations (Gneiting and Raftery, 2007). It thus provides a way to compare the performance of ensemble simulations against the performance of deterministic simulations, for each individual catchment.

2.2.2 The continuous ranked probability score

Performance evaluation of the probabilistic simulations implies the verification of a probability distribution. Therefore the simulation error cannot be estimated from a routine comparison between the model output and a verifying value. The performance depends of the correspondence between the predicted probability and the actual frequency of occurrence (Atger, 1999). The selected score is the Continuous Ranked Probability Score (CRPS) (Matheson and Winkler, 1976), which is a proper score widely used in atmospheric and hydrologic sciences (e.g., Gneiting et al., 2005; Candille and Talagrand, 2005; Weber et al., 2006; Boucher et al., 2009). The CRPS is defined as:

$$\mathrm{CRPS}(F_t, x_t) = \int_{-\infty}^{\infty} (F_t(x) - H\{x \geq x_t\})^2 dx \tag{1}$$

where F_t is the cumulative predictive distribution function for the time t, x is the predicted variable (here streamflow) and x_t is the corresponding observed value. The function $H\{x{\geq}x_t\}$ is the Heaviside function which equals 1 for simulated values larger than the observed value and 0 for simulated values lower than the observation. The CRPS is positive and a zero value indicates a perfect simulation. An analytical solution of Eq. (1) exists only for normal predictive distributions (Gneiting and Raftery, 2007). However, because the normality of the predictive distribution is not always true in the present study, a Montecarlo approximation to Eq. (1) has been used instead (Székely et al., 2003; Gneiting et al., 2007):

$$\mathrm{CRPS} = E|X - x_t| - 0.5E|X - X'| \tag{2}$$

where X and X' are independent copies of a random variable in a vector with distribution function F_t.

As already mentioned, an interesting property of the CRPS is that it reduces to the AE score in the case of a deterministic simulation (Gneiting and Raftery, 2007). However, because the score obtained by a particular ensemble simulation for a certain time has no meaning, we rather consider the average of all individual scores as a measure of the quality of the simulation system, thus comparing mean AE (MAE) and mean CRPS ($\overline{\mathrm{CRPS}}$), which values are directly proportional to the magnitude of the observations.

HESSD
7, 4023–4058, 2010

We also aim to evaluate the performance gain in terms of $\overline{\text{CRPS}}$ that may bring the optimization procedure. Based on the skill score (e.g., Wilks, 1995), the percentage of improvement over the reference is given by:

$$\text{gain}(\%) = \left(1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{ref}}}\right) \times 100 \tag{3}$$

2.2.3 Reliability

Reliability refers to the statistical consistency between simulations and observations. For instance, a reliable 90% confidence interval calculated using the predictive distribution function should contain the observed value in 9 cases out of 10 on average. On the other hand, the potential CRPS corresponds to the best possible CRPS value that could be obtained with the database and the particular simulation system that is used, if the latter was made to be perfectly reliable. Because of the complex nature of the CRPS, other means of assessing the reliability is often used in parallel, such as the rank histogram and the reliability diagram. Unreliable simulations can be misleading and should be used with caution, if at all. Statistical procedures exist to calibrate unreliable probabilistic simulations (e.g., Raftery et al., 2005; Fortin et al., 2006; Stensrud and Yussouf, 2007).

The reliability of the predictive distribution can be visually assessed using the rank histogram (Talagrand et al., 1999; Hamill, 2001). To construct it, the observed value x_t is added to the ensemble simulation. That is, if the simulation has n members, the new set consists of $n+1$ values. Then, the rank associated with the observed value is determined. This operation is repeated for all simulations and corresponding observations in the archive. The rank histogram is obtained by constructing the histogram of the resulting N ranks. The interpretation of the rank histogram is based on the assumptions that all the members of the ensemble simulation along with the observations are independent and identically distributed; under these hypotheses, if the predictive distribution is well calibrated, then the rank histogram should be close to uniformity (equally

HESSD
7, 4023–4058, 2010

distributed). An asymmetrical histogram is usually an indication of a bias in the mean of the simulations. If the rank histogram is symmetric and "U" shaped, it may indicate that the predictive distribution is under-dispersed. If it has an arch form, the predictive distribution may be over dispersed.

Because it is not practical to present all 1061 rank histograms, results will be synthesised using the ratio δ metric proposed by Candille and Talagrand (2005): a numerical indicator reflecting the squared deviation from flatness in individual rank histograms. It is given by

$$\delta = \frac{\Delta}{\Delta_0} \tag{4}$$

where:

$$\Delta = \sum_{k=1}^{n+1} \left(s_k - \frac{N}{n+1} \right)^2 \tag{5}$$

and s_k is the number of elements in the kth interval of the rank histogram. For a reliable system, s_k has an expectation of $N/(n+1)$. Then, Δ_0 is the ratio that would be obtained by a perfectly reliable system:

$$\Delta_0 = \frac{Nn}{n+1} \tag{6}$$

leading to a target value of δ=1. Of course, a perfectly reliable system is a theoretical concept. In practice, a system is declared unreliable whenever its δ value is quite larger than 1 (Candille et al., 2005). However, the exact δ threshold, above which a system may be declared unreliable, has to be established for each investigation, notably because the δ metric is proportional to the length of the time series (the threshold value adopted here will be discussed later on). Some applications of the δ metric include evaluating the degree of reliability of meteorological ensembles by comparing δ values according to their series lengths (e.g., Jiang et al., 2009).

HESSD

7, 4023–4058, 2010

The reliability diagram is another approach used to graphically represent the performance of probability simulations of dichotomous events. A reliability diagram consists of the plot of observed relative frequency as a function of simulation probability and the 1:1 diagonal perfect reliability line (Wilks, 1995). In the present study, ten confidence intervals have been calculated with nominal confidence level of 5% to 95%, with an increment of 5% for each emitted simulation. Then, for each simulation and for each confidence interval, it was established whether or not each confidence intervals covered the observation. This is repeated for all simulation-observation pairs and its mean is then plotted (Boucher et al., 2009).

2.2.4 Hit over threshold criteria

The relative operating characteristic (ROC) curve (Peterson et al., 1954; Mason, 1982) plots the probability of detection (POD) versus the probability of false detection (POFD), which are given by:

$$\mathrm{POD} = \frac{\mathrm{hits}}{\mathrm{hits} + \mathrm{misses}} \tag{7}$$

$$\mathrm{POFD} = \frac{\mathrm{false\_alarms}}{\mathrm{correct\_negatives} + \mathrm{false\_alarms}} \tag{8}$$

The four combinations of simulations (yes or no) and observations (yes or no), called the *joint distribution*, are: hit (the event simulation to occur and did occur), miss (the event simulation not to occur, but did occur), false alarm (event simulation to occur, but did not occur) and correct negative (event simulation not to occur and did not occur) (e.g., Wilks, 1995). The area under the ROC curve characterizes the quality of a simulation system's ability to correctly anticipate the occurrence or non occurrence of the events. In constructing a ROC curve, simulations are expressed in binary as "warnings" or "not warnings" indicating whether or not the defined event is expected to occur. The ROC area ranges from 0 to 1, 0.5 indicating no skill and 1 being the perfect

HESSD
7, 4023–4058, 2010

score. ROC measures the ability of the simulation to discriminate between two alternative outcomes, thus measuring resolution. It is not sensitive to bias in the simulation, so says nothing about reliability. A biased simulation may still have good resolution and produce a good ROC curve, which means that it may be possible to improve the simulation through calibration. The ROC is thus a basic decision-making criterion that can be considered as a measure of potential usefulness (WMO, 2002).

2.3 Genetic algorithm

Genetic algorithm is a technique for optimization of problems or systems. It is inspired from biology, more specifically by genetic codes, where solutions are typically translated into binary code string. The search of optimal solution is regulated by rules based on Darwin's theory on the survival of the fittest, by which the strings are allowed to survive from one generation (i.e. iteration) to another and to trade part of their genetic material with other strings depending of their robustness as defined by the objective function (e.g., Anctil et al., 2006).

The present work uses genetic algorithm to identify model subsets optimizing the Continuous Ranked Probability Score. The rules of reproduction, crossover and mutation employed here are well described in Goldberg (1989).

The coded string consists of seventeen elements or positions, each one representing a specific model: 0 values identify models that are not used, while 1 values identify models that are retained. A total of 131 054 combinations of at least two models can be generated from a pool of seventeen candidates. The processes of reproduction, crossover and mutations regulate the search in the domain of all these possible combinations, where the objective function is the inverse squared CRPS. At each generation, 50 combinations are thus investigated. From the initial generation, 20 others are created, leading to the consideration of 1000 model subsets. This search is repeated over all 1061 catchments.

As already mentioned, the first half of the time series is used for optimization, while the second half is used for validation. All results provided herein concern strictly the

HESSD
7, 4023–4058, 2010

validation sub-dataset.

3 Results

3.1 Individual model performance

MAE values are used to compare individual model performance, based on their frequency of occurrence in the top 5 ranking for each catchment (Fig. 2). There are clear differences between models. Some of them are more frequently in the top five, such as models 1, 2, and 3, while others are rarely present, such as model 17 and 16 – note that Fig. 2 justify the model ordering in Table 2. The selected seventeen models thus offer a wide range of individual performance.

3.2 Comparison of deterministic and probabilistic simulations

The main scope of the present study is to answer the following question: is there more valuable information in the ensemble simulations than in the deterministic ones? This question is first tackled by comparing the $\overline{\text{CRPS}}$ and the MAE values for the C0 reference ensemble formed by all seventeen models. In Fig. 3, all 1061 catchments lead to a $\overline{\text{CRPS}}$ value lower than the MAE ones, confirming the added value of retaining all the components of the ensembles over their average deterministic values. Note that simulations for each catchment have been standardized by their corresponding mean streamflow observation to facilitate comparison between them.

However, it remains possible that some individual models surpass in performance the C0 reference ensemble. Indeed such situations occur quite frequently when relying on deterministic simulations, which provides the lowest MAE for only 38% of the catchments; while for example model 1 surpasses the performance of all the other models including the deterministic simulation in 21% of the catchments (Fig. 4a). The performance gain following the usage of the SMA aggregating multiple model outputs is thus not as universal as proposed by Shamseldin et al. (1997) or Georgakakos et al. (2004). However, the situation gets considerably better when using the probabilistic ensemble simulations (i.e. keeping all individuals model outputs), which improve on the performance of all individual models in 96% of the catchments (Fig. 4b). These striking results confirm the superiority of the probabilistic approach over the deterministic one.

The next question concerns the reliability of the ensemble simulations, as assessed by the rank histograms and the reliability plots. Figure 5 presents some examples of rank histograms in order to interpret their corresponding ratio δ values. As mentioned earlier, a threshold δ value has to be established for each experimental set-up because this metric is proportional to the length of the time series. From Fig. 5, it is assessed that, for the simulation system and series length at hand, a ratio δ value of about 20 may be used as a practical upper limit of reliability (Fig. 5c), while value of about 100 is without a doubt under-dispersed (Fig. 5f) – as confirmed by the corresponding reliability diagrams drawn in Fig. 6. Now turning to the entire database, the cumulative frequency of the ratio δ in Fig. 7 shows that reliability is achieved for about one third of the catchments (δ values below 20) and that the system is clearly unreliable for at least 20% of the catchments (δ values larger than 100), the other cases being debatable. An operating simulation system based on the C0 reference ensembles would thus need to include the calibration of the predictive distribution for an important number of catchments, in order to improve their reliability.

Nonetheless the reliability imperfection of our simulation system, its ability to discriminating between events and non-events is next confronted to the same ability of the deterministic simulations. For that purpose, ROC scores were calculated for threshold values respectively corresponding to quantiles 10, 25, 50, 75 and 90 of the observation time series. Results are gathered in Fig. 8, where it can be noted that the probabilistic ROC scores are in almost all instances superior to the deterministic ones, proving again the superiority of the ensemble philosophy over the aggregation philosophy, at least for better event detections, even if the produced ensemble could in many cases be further improved by the application of a calibration procedure. It is also noteworthy

HESSD
7, 4023–4058, 2010

that the predictive distributions are skilled for the large majority of the catchments (ROC values superior to 0.5) and that the system is better at detecting larger events such as quantiles 50 or higher, than low flow events such as quantile 10. For the latter case, the probabilistic simulations largely improve over the deterministic ones that prove to be unskilled for many catchments.

3.3 Looking for optimized model ensembles

Could the system performance be further improved through model selection? A genetic search algorithm is used to answer that question, objectively optimizing the $\overline{\text{CRPS}}$ value for each catchment. Such analysis will also help answer some other subsidiary questions like: Are seventeen models enough or too many to produce an operational ensemble? Are all models equally useful to the ensemble subsets or only the ones that performs better individually? Does any gain in performance through optimization come with the cost of a loss of reliability?

The optimization procedure described in Sect. 2.3 was applied to all catchments. Many model subsets showed improved performance over the C0 reference ensemble. More specifically, improvements were found for 1057 of the 1061 catchments, which represent 99.6% of the database. The gain in terms of $\overline{\text{CRPS}}$ resulting from the performed optimization is shown in Fig. 9 (see Eq. 3 where the reference value is C0). The gain varies from 0.3% to 93% with a median value of 5.5%. There is also a gain in the quality of the ensemble's reliability as seen in Fig. 10 that draws the initial ratio δ values against the ones of the optimized subsets: an improvement was obtained in 86% of the cases. However, those gains are not large enough to solve the under dispersion issue of the produced ensembles. A calibration procedure is thus still needed for most catchments.

Figure 11a shows the relative frequency of selection of the models in the best subset ensemble of each catchment. When compared to Fig. 2, which showed the frequency of occurrence in the top five ranking, it may be deduced that all models are

HESSD
7, 4023–4058, 2010

useful contributors to subset ensembles that outperform the C0 reference ensemble. Nonetheless, the optimization procedure does somehow favour models that lead to the best individual performance, namely 1, 2, 3, 4 and 5, then 7, 8 and 9. Furthermore, no links could be established between the level of complexity of the models (number of optimized parameters and storages) and their usefulness to optimized subset.

Figure 11b presents the relative frequency of the number of models in these subsets over all catchments. From 7 to 10 models are deemed sufficient to construct ensembles with improved performance.

Figure 12 provides yet another view of the optimized subsets, where they are categorized by number of models, which varies from 2 to 16. Boxplots were produced in order to illustrate the variability of the 1061 $\overline{\mathrm{CRPS}}$ values (standardized with their corresponding mean streamflow observation as in Fig. 3). In general, results show that there exist many subset sizes that improve on the C0 reference performance obtained by pooling all seventeen lumped model outputs (the median for the best optimized combination is 0.1850 and the median for C0 is 0.1976). Furthermore, these subsets are superior to the ones constructed with the best eight individual models (C1 with a median of 0.1965) and with the worst 8 individual models (C2 with a median value of 0.2240). This latter result supports the finding of Viney et al. (2009) that the best ensembles are not necessarily those containing the best individual models, but it seems that the inclusion of some good models is essential.

4 Conclusions

The main scope of this work was to compare the added value of ensembles constructed from seventeen lumped hydrological models against their simple average counterparts. Ensembles are probabilistic simulations that allow appreciating the uncertainty according to the spread of their predictive distribution at each time step. For example, they may be used to evaluate confidence intervals for the outputs or probabilities of the streamflow being above a certain threshold value. Conversely, the simple average of

HESSD
7, 4023–4058, 2010

the seventeen lumped outputs leads to a single aggregated predictor, which provides no specific information about its uncertainty.

For all 1061 catchments, results showed that the $\overline{\text{CRPS}}$ of the ensembles were lower than the MAE of the aggregated simulations, confirming the added value of retaining all the components of the ensembles over their aggregated deterministic values. Furthermore, the probabilistic simulations surpass the performance of all individual models in 96% of the catchments, while the same occur for only 38% of the catchments in the case of the aggregated deterministic simulations.

Reliability of the simulation ensembles is achieved for about 30% of the catchments. An operating simulation system would thus need to include a calibration of the predictive distributions in order to improve their reliability. In spite this imperfection, the ensembles were shown to be skilled at discriminating between events and non-events, based on the ROC scores, especially for larger streamflows. Again, the comparison between probabilistic and deterministic skills was favorable to the probabilistic approach.

Genetic algorithm was next used to identify model subsets optimizing the CRPS. Many model subsets were found improving the performance of the reference ensemble. In most cases, from 7 to 10 models selected among the 17 available models were deemed sufficient to construct ensembles with improved performance. However, even if an important disparity was noticed between the individual performances of the available models, all of them appeared in the many optimized subsets. Furthermore, the optimized subsets were found superior to the ones constructed with the best eight individual models, which means that the best ensembles are not necessarily those containing the best individual models. The gain in performance of the optimized subsets is accompanied by an improvement of the ensemble reliability in 86% of the cases. Nonetheless, a calibration procedure is still needed for many catchments.

More sophisticated aggregation methods may also have been tested, as discussed in the introduction. They may have improved the performance (MAE) of our deterministic simulations, as suggested by the results of previous studies. However, it is also true that the calibration of the predictive distribution should also improve the performance

HESSD
7, 4023–4058, 2010

($\overline{\mathrm{CRPS}}$) of the probabilistic simulation.

All in all, this work advocates the increased usage of multiple hydrological models for performance improvement and for uncertainty assessment. However, more work is needed concerning model selection and the sought after diversity that brings the essence of model ensembles: reliability. Future work should also investigate multiple model probabilistic forecasting in an operational context.

Acknowledgements. The authors acknowledge Institut EDS, MITACS, and CONACYT-Mexico financial contributions to the project, and Météo-France for making the meteorological data available for this study.

References

Ajami, N. K., Duan, Q., Gao, X., and Sorooshian, S.: Multi-model combination techniques for hydrological simulations: application to distributed model intercomparison project results, J. Hydrometeorol., 7, 755–768, 2006.

Anctil, F., Lauzon, N., Andréassian, V., Oudin, L., and Perrin, C.: Improvement of rainfall-runoff forecasts through mean areal rainfall optimization, J. Hydrol., 328, 717–725, 2006.

Atger, F.: Verification of intense precipitation forecasts from single models and ensemble prediction systems, Nonlinear Proc. Geoph., 8, 401–417, 2001.

Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall-runoff model, Nord. Hydrol., 4, 147–170, 1973.

Bormann, H., Breuer, L., Croke, B., et al.: Reduction of predictive uncertainty by ensemble hydrological modeling of catchments processes and land use change effects, in: Procedings of the 11th Conference of the Euromediterranean Network of Experimental and Representative Basins, Luxembourg, September 2006, IHP-VI/Technical documents in hydrology, 81133–81139, 2007.

Boucher, M. A., Perreault, L., and Anctil, F.: Tools for the assessment of hydrological ensemble forecasts obtained by neural networks, J. Hydroinform., 11, 297–307, 2009.

Burnash, R. J. C., Ferral, R. L., and Mc Guire, R. A.: A generalized streamflow simulation system – conceptual modeling for digital computers, US Department of Commerce, National Weather Service and State of California, Department of Water Resources, 1973.

Candille, G. and Talagrand, O.: Evaluation of probabilistic prediction systems for a scalar variable, Q. J. Roy. Meteor. Soc., 131, 2131–2150, 2005.
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. and Frevret, D. K., Water Resources Publications, Highlands Ranch, 335–367, 2002.
Cormary, Y. and Guilbot, A.: Étude des relations pluie-débit sur trois bassins versants d'investigation, IAHS-AISH P., 108, 265–279, 1973.
Coulibaly, P., Hache, M., Fortin, V., and Bobee, B.: Improving daily reservoir inflow forecasts with model combination, J. Hydrol. Eng., 9(2), 91–99, 2005.
Devineni, N., Sankarasubramanian, A., and Ghosh, S.: Multimodel ensembles of streamflow forecasts: role of predictor state in developing optimal combinations, Water Resour. Res., 44, W09404, doi:10.1029/2006WR005855, 2008.
Durand, Y., Brun, E., Mérindol, L., Guyomarc'h, G., Lesassre, B., and Martin, E.: A meteorological estimation of relevant parameters for snow models, Ann. Glaciol., 18, 65–71, 1993.
Edijatno: Nascimento, N. O., Yang, X., Makhlouf, Z., and Michel, C.: GR3J: a daily watershed model with three free parameters, Hydrolog. Sci. J., 44(2), 263–277, 1999.
Fortin, V., Favre, A. C., and Said, M.: Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member, Q. J. Roy. Meteor. Soc., B132, 1349–1369, 2006.
Fortin, V. and Turcotte, R.: Le modèle hydrologique MOHYSE, Note de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Université du Québec à Montréal, 2006.
Garçon, R.: Modèle global Pluie-Débit pour la prévision et la prédétermination des crues, La Houille blanche, 7 (8), 88–95, 1999.
Georgakakos, K. P., Seo, D-J., Gupta, H., Schaake, J., and Butts, M. B.: Characterizing streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298, 222–241, 2004.
Girard, G., Morin G., and Charbonneau, R.: Modèle précipitations-débits à discrétisation spatiale, Cahiers ORSTOM Série Hydrologie, IX 4, 35–52, 1972.
Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133 (5), 1098–1118, 2005.
Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am.

HESSD
7, 4023–4058, 2010

Stat. Assoc., 102(477), 359–378, 2007.
Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA, Addison-Wesley, 412 pp., 1989
Hamill, T.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001.
He, Y., Wetterhall, F., Bao, H., Cloke, H., Li, Z., Pappenberger, F., Hu, Y., Manful, D., and Huang, Y.: Ensemble forecasting using TIGGE for the July–September 2008 floods in the Upper Huai catchment: a case study, Atmos. Sci. Lett., 11(2), 132–138, 2010.
Jakeman, A. J., Littlewood, I. G., and Whitehead, P. G.: Computation of the instantaneous unit hydrograph and identifiable component flows with applications to two small upland catchments, J. Hydrol., 117, 275–300, 1990.
Jiang, Z. and Mu, M.: A comparision study of the methods of conditional nonlinear optimal perturbations and singular vectors in ensemble predictions, Adv. Atmos. Sci., 26(3), 465–470, 2009.
Krishnamurti, T. N.: Improved weather and seasonal climate forecasts from multimodel superensemble, Science, 285(1548), 1548–1550, 1999.
Le Moine, N.: Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modèles pluie-débit?, Ph.D. thesis, Pierre et Marie Curie University, Paris, France, 324 pp., 2008.
Le Moine, N., Andréassian, V., Perrin, C., and Michel, C.: How can rainfall-runoff model handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, Water Resour. Res., 43, W06428, doi:10.1029/2006WR005608, 2007.
Loumagne, C., Vidal, J. J., Torterotot, J. P., and Roche, P. A.: Procédure de décision multimodèle pour une prévision des crues en temps réel application au bassin supérieur de la Garonne, Revue des sciences de l'eau, 8, 539–561, 1995.
Marshall, L., Nott, D., and Sharma, A.: Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework, Hydrol. Process., 21(7), 847–861, 2007.
Mason, S. J.: A model for assessment of weather forecast, Aust. Meteorol. Mag., 30, 291–303, 1982.
Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Manage. Sci., 22, 1087–1096, 1976.
Mathevet, T.: Quels modèles pluie-débit globaux pour le pas de temps horaire? Développement empirique et comparaison de modèles sur un large échantillon de bassins versants, Ph.D.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

thesis, ENGREF (Paris), Cemagref (Antony), France, 463 pp., 2005.
Michel, C., Perrin, C., and Andréassian, V.: The exponential store: a correct formulation for rainfall–runoff modeling, Hydrolog. Sci. J., 48(1), 109–124, 2003.
Moore, R. J. and Clarke, R. T.: A distribution function approach to rainfall-runoff modeling, Water Resour. Res., 17(5), 1367–1382, 1981.
Nielsen, S. A. and Hansen, E.: Numerical simulation of the rainfall-runoff process on a daily basis, Nord. Hydrol., 4, 171–190, 1973.
O'Connell, P. E. and Clarke R. T.: Adaptive hydrological forecasting. A review, Hydrolog. Sci. Bulletin, 26(2), 179–205, 1981.
Oudin, L., Andréassian, V., Mathevet, T., and Perrin, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resour. Res., 42, W07410, doi:10.1029/2005WR004636, 2006.
Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a rainfall-runoff model? Part 2 – Towards a simple and efficient PE model for rainfall-runoff modeling, J. Hydrol., 303(1–4), 290–306, 2005.
Perrin, C., Michel, C., and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 242(3–4), 275–301, 2001.
Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, 2003.
Peterson, W. W., Birdsall, T. G., and Fox, W. C.: The theory of signal detectability, Trans. IRE Prof. Group. Inf. Theory, PGIT, 2–4, 171–212, 1954.
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.
Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: The hydrological ensemble prediction experiment, B. Am. Meteorol. Soc., 88, 1541–1547, 2007.
Shamseldin, A. Y., O'Connor, K. M., and Liang G. C.: Methods for combining the outputs of different rainfall–runoff models, J. Hydrol., 197, 203–229, 1997.
Stensrud, D. J. and Yussouf, N.: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system, Weather Forecast., 22(1), 3–17, 2007.
Sugawara, M: Automatic calibration of the tank model, Hydrolog. Sci. J., 24(3), 375–388, 1979.
Székely, G. J. and Rizzo, M. L.: A new test for multivariate normality, J. Multivariate Anal., 93(1), 58–80, 2005.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of the probabilistic prediction systems, in: Proceedings, ECMWF Workshop on Predictability, Shinfield Park, Reading, Berkshire, ECMWF, 1–25, 1999.

Thiery, D.: Utilisation d'un modèle global pour identifier sur un niveau pièzometrique des influences multiples dues à diverses activités humaines, IAHS-AISH P., 136, 71–77, 1982.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Gräff, T., et al.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32, 147–158, 2009.

Warmerdam, P. M., Khole, J., and Chormansky, J.: Modelling rainfall-runoff process in the hupsele breek research bassin, ecohydrological process in small basins, in: Proceedings of the Strasbourg Conference (24–26 September 1996), IHP-V/Technical documents in Hydrology, 14, 155–160, 1997.

Weber, F., Perreault, L., Fortin, V., and Gaudet, J.: Performance measures for probabilistic hydrologic forecasts used at BC-Hydro and Hydro-Québec, in: EGU Conference, April 2006.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 1995.

WMO, Exchange of long range forecasts verification scores: a verification system for long range forecasts, in: Commission for basic systems meeting of expert team to develop a verification system for long range forecasts, Montreal, Canada, 22–26 April 2002, 2002.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30(8), 1756–1774, 2007.

HESSD

7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

Table 1. Characteristics of the 1061 catchment dataset.

	Area (km^2)	Mean annual rainfall (mm)	Mean annual potential evapotranspiration (mm)	Mean annual discharge (mm)
Minimum	10	662	339	31
Median	163	980	657	352
Maximum	32 400	2182	870	3493

Table 2. Models identification and characteristics.

ID	Model	Number of optimized parameters	Number of storages	Derived from
1	GR4J	4	2	Perrin et al. (2003)
2	PDM0	8	4	Moore et al. (1981)
3	MORD	6	4	Garçon (1999)
4	TOPM	8	3	Michel et al. (2003)
5	SACR	13	6	Burnash et al. (1973)
6	SMAR	9	3	O'Connell et al. (1981)
7	NAM0	10	7	Nielsen et al. (1973)
8	TANK	10	5	Sugawara (1979)
9	HBV0	9	3	Bergström et al. (1973)
10	CREC	8	3	Cormary et al. (1973)
11	WAGE	8	4	Warmerdam et al. (1997)
12	IHAC	6	3	Jakeman et al. (1990)
13	GARD	7	3	Thiery (1982)
14	SIMH	8	3	Chiew et al. (2002)
15	MOHY	7	2	Fortin et al. (2006)
16	CEQU	9	3	Girard et al. (1972)
17	HYM0	6	5	Yadav et al. (2007)

HESSD
7, 4023–4058, 2010

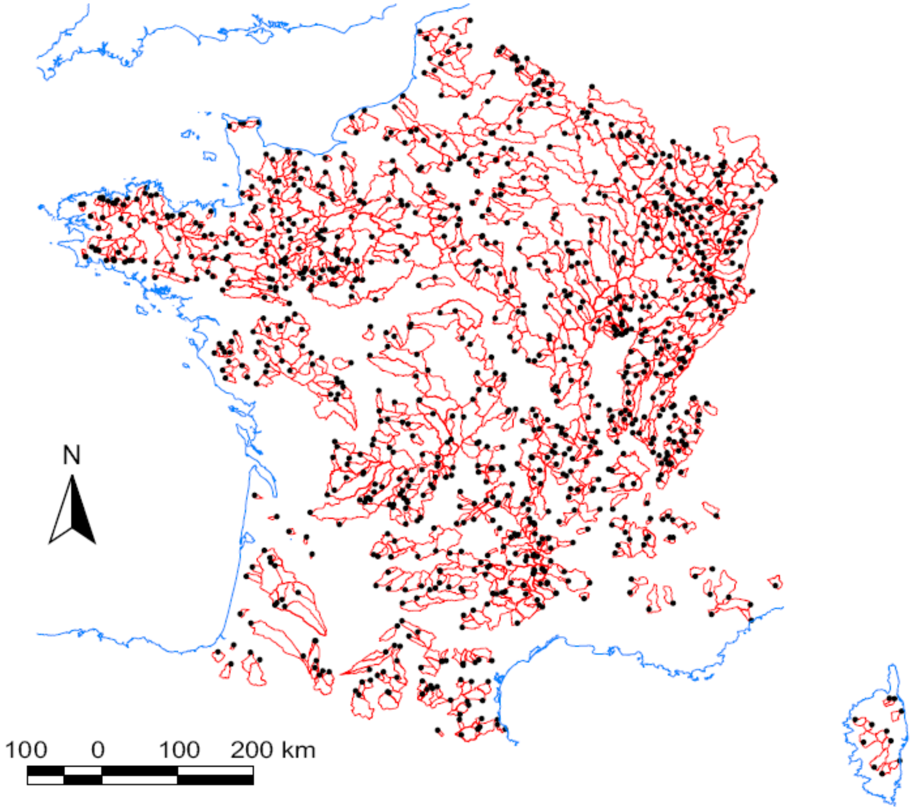

Fig. 1. Location of the 1061 gauging stations and corresponding catchment boundaries (Le Moine, 2008).

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

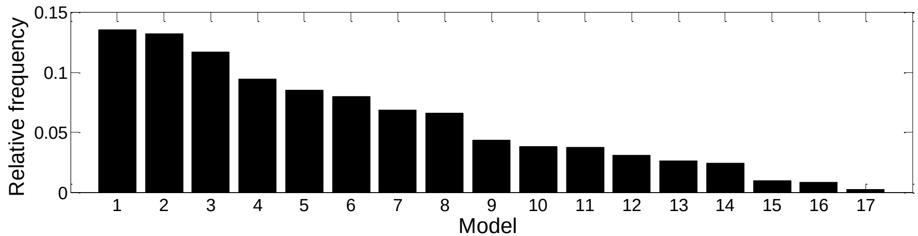

Fig. 2. Relative frequency of occurrence in the top 5 ranking, based on individual MAE values for all catchments.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

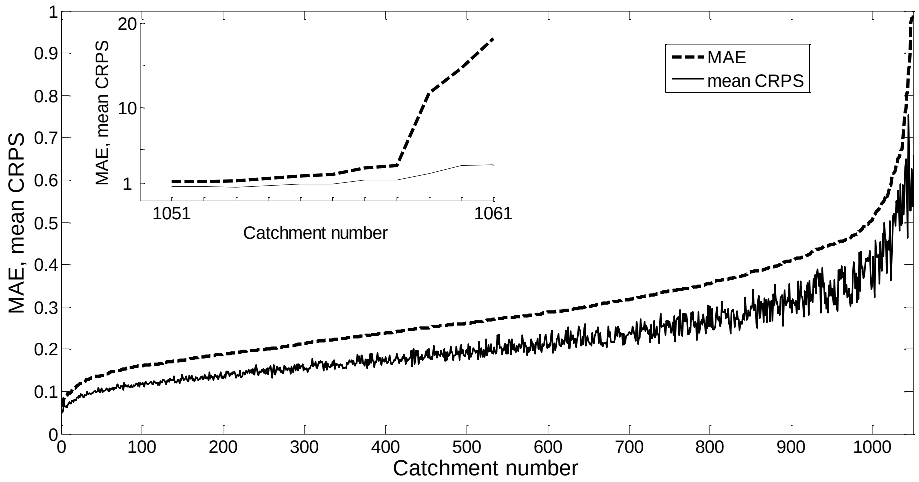

Fig. 3. Mean probabilistic and deterministic scores comparison. Catchments are ordered according to their MAE value.

HESSD
7, 4023–4058, 2010

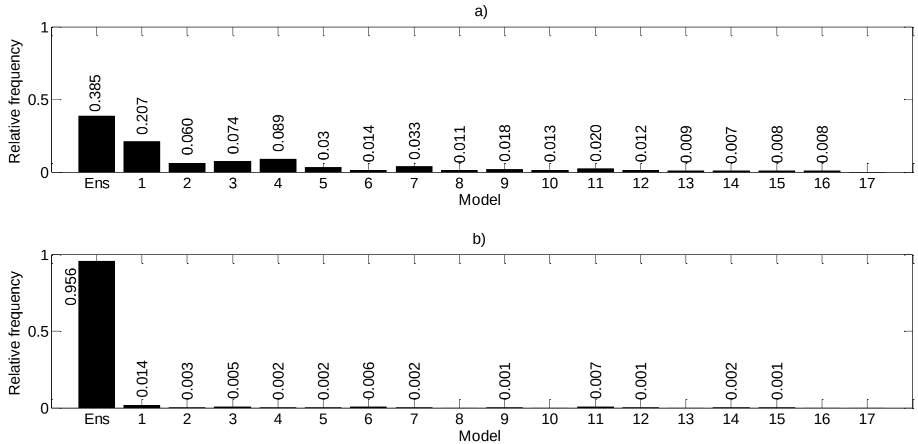

Fig. 4. Relative frequency of occurrence as the best model or ensemble: **(a)** deterministic ($\overline{\text{MAE}}$) and **(b)** probabilistic ($\overline{\text{CRPS}}$).

HESSD

7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

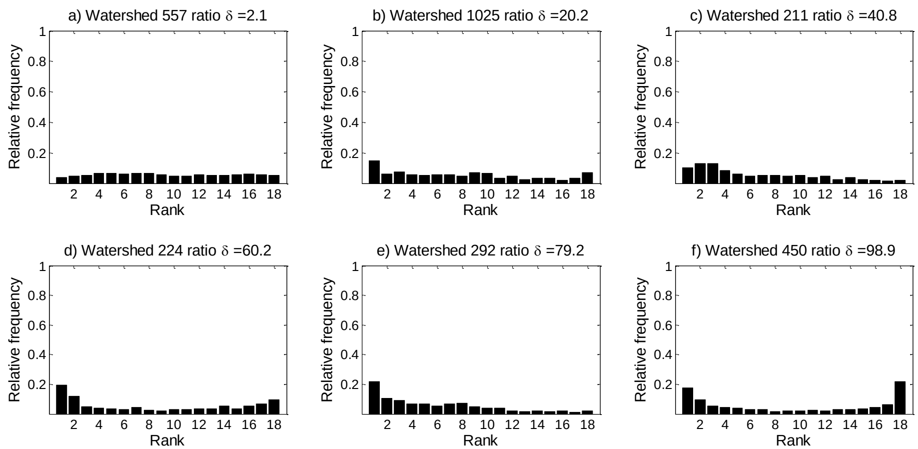

Fig. 5. Six examples of rank histograms with their ratio δ values.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

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

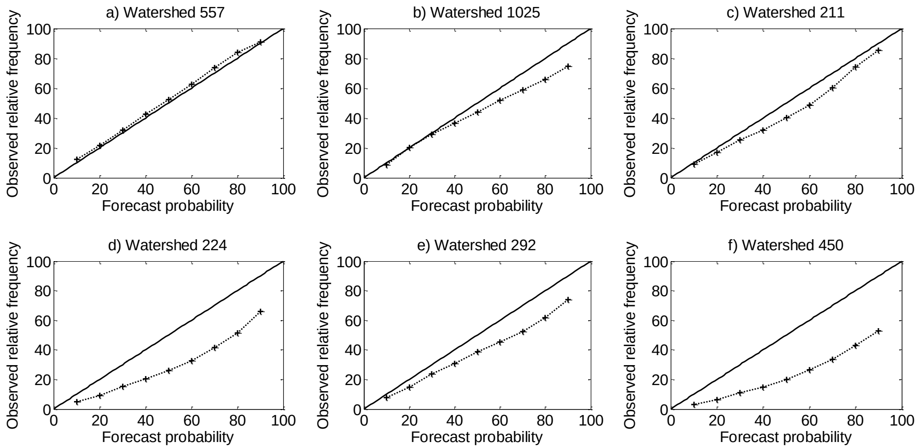

Fig. 6. Reliability plots for the same catchments as in Fig. 5.

HESSD

7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

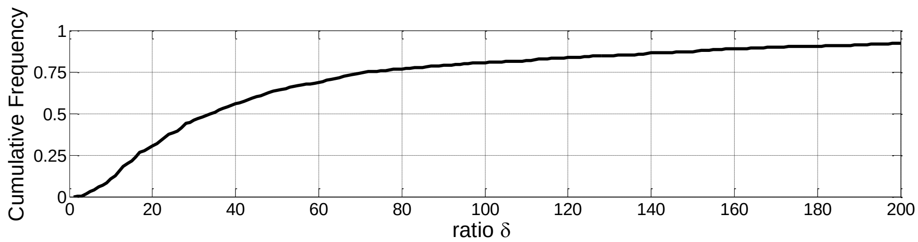

Fig. 7. Cumulative frequency of ratio δ for the C0 reference ensembles.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

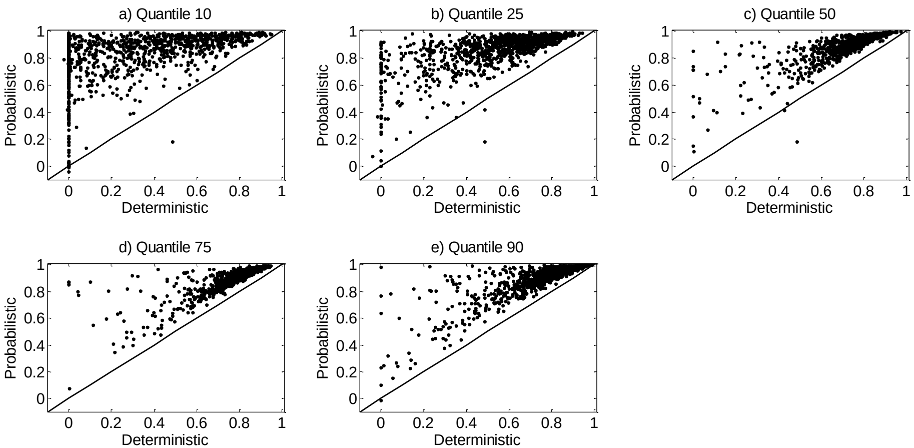

Fig. 8. Probabilistic and deterministic ROC scores for quantiles 10, 25, 50, 75 and 90.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

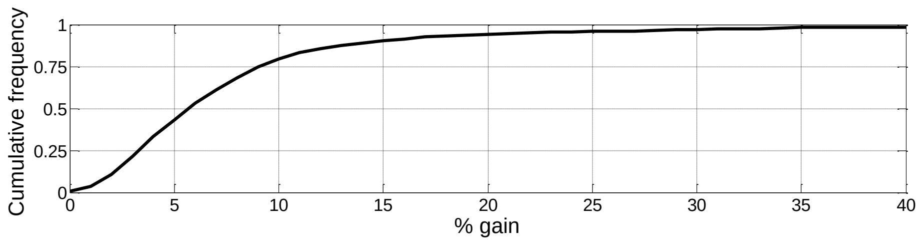

Fig. 9. Cumulative frequency of the $\overline{\text{CRPS}}$ gain after optimization.

HESSD
7, 4023–4058, 2010

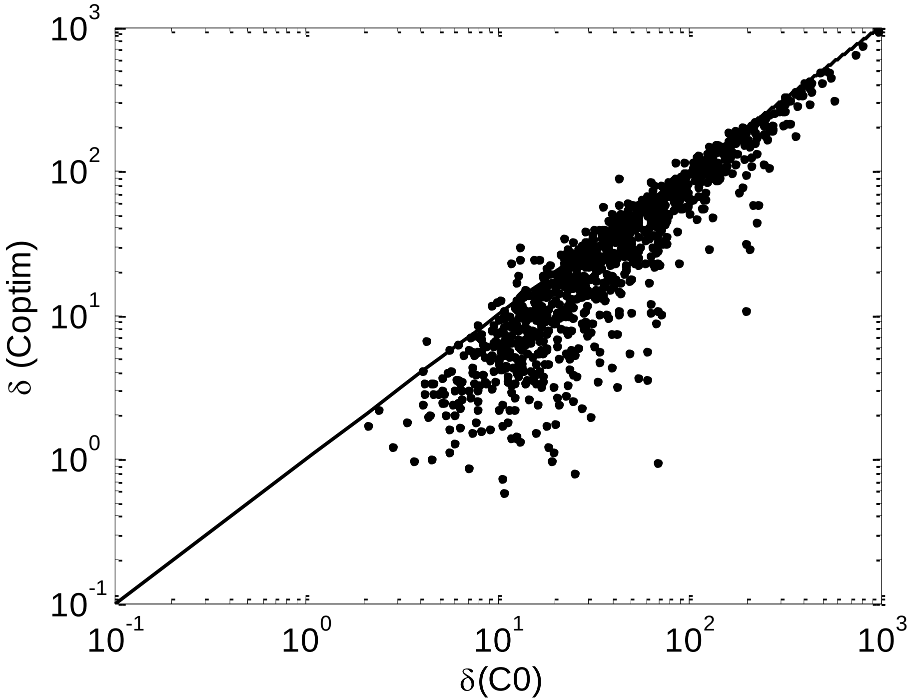

Fig. 10. Scatter plot ratio δ values without (C0) and with optimization.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

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

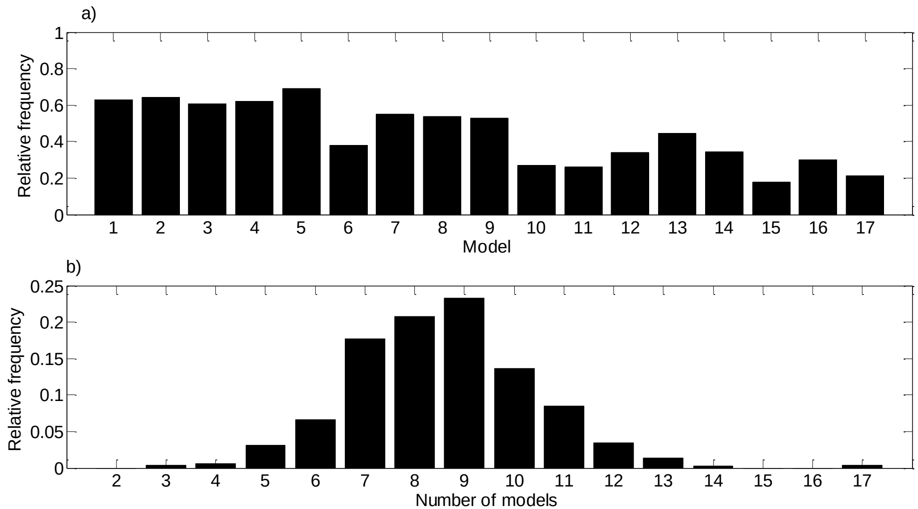

Fig. 11. Relative frequency of **(a)** the presence of each model in the optimized subset, and **(b)** the number of models in these subsets, for all 1061 catchments.

HESSD
7, 4023–4058, 2010

Performance and reliability of ensemble simulations

J. A. Velázquez et al.

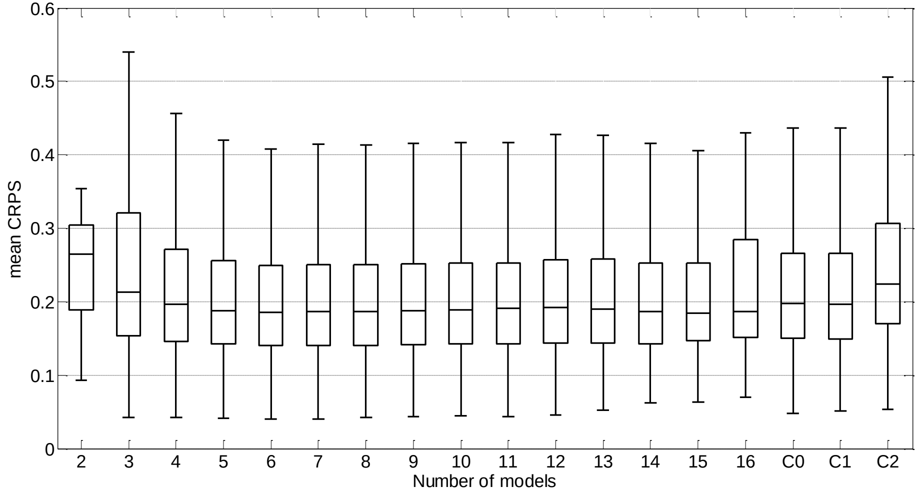

Fig. 12. Box plot of the $\overline{\text{CRPS}}$ over the 1061 catchments, as a function of the number of models per optimized subsets.

HESSD
7, 4023–4058, 2010