Evolutionary Multiobjective Optimization for Selecting Members of an Ensemble Streamflow Forecasting Model

Darwin Brochero[1,2], Christian Gagné[2], and François Anctil[1]
[1] Chaire de recherche EDS en prévisions et actions hydrologiques, Dép. de génie civil et de génie des eaux
[2] Laboratoire de vision et systèmes numériques, Département de génie électrique et de génie informatique
Université Laval, Québec (Québec), Canada G1V 0A6
darwin.brochero.1@ulaval.ca, christian.gagne@gel.ulaval.ca, francois.anctil@gci.ulaval.ca

ABSTRACT

We are proposing to use the Nondominated Sorting Genetic Algorithm II (NSGA-II) for optimizing a hydrological forecasting model of 800 simultaneous streamflow predictors. The optimization is based on the selection of the best 48 predictors from the 800 that jointly define the "best" ensemble in terms of two probabilistic criteria. Results showed that the difficulties in simplifying the ensembles mainly originate from the preservation of the system reliability. We conclude that Pareto fronts generated with NSGA-II allow the development of a decision process based explicitly on the trade-off between different probabilistic properties. In other words, evolutionary multiobjective optimization offers more flexibility to the operational hydrologists than *a priori* methods that produce only one selection.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search—*Heuristic methods*; I.6.6 [**Simulation and Modelling**]: Simulation Output Analysis

Keywords

Evolutionary multiobjective optimization, Probabilistic forecasting, Streamflow forecasting, Hydrological ensemble prediction system, Uncertainty cascade model

General Terms

Experimentation

1. INTRODUCTION

The physics of hydrological processes inevitably leads us to different sources of uncertainty when forecasting their outcomes. First comes the uncertainty associated with meteorological variables. In this regard, in recent years, Meteorological Ensemble Prediction Systems (MEPS) have become increasingly popular. Second comes the uncertainty in the conceptualization of the hydrological processes such as evapotranspiration and the interactions among vegetation, land surface, and groundwater. Finally comes hydraulic routing conceptualization and parametric uncertainty in both the hydraulic and hydrological models. This uncertainty chain propagation model has been called *uncertainty cascade model* [24].

In recent years, the hydrometeorological community has focused on the development of Hydrological Ensemble Prediction Systems (HEPS) taking into account the uncertainty process chain evaluation as an important part of the decision making stage [7, 26, 28]. Consequently, hydrological response is seen as a pool of multiple probable scenarios, accepting the paradigm of complementarity (diversity) for forecasting purposes.

The classical evaluation of multiple forecast scenarios leads to the use of reductionist decision schemes based on combining functions such as average or more elaborated combination functions, ignoring the importance of uncertainty evaluation. In this context, the hydrometeorological community has developed probabilistic performance metrics, called scores, used not only to evaluate the most likely prediction, but also its uncertainty. Probabilistic response properties such as reliability, resolution, sharpness, and consistency have been highlighted as complex features in HEPS [3, 31].

However, the HEPS complexity may become an operational burden when one has to evaluate several hundreds of scenarios at each time step, a situation easily achieved with the increasing computational resources and the advancements of each component of the uncertainty chain presented above. As a result, simplification of such a HEPS becomes a mandatory step from an operational standpoint [5].

At this complexity level, represented by the number of scenarios and its probabilistic properties interaction, an "overproduce and select" mechanism appears as a natural procedure of simplification into the so-called *combiner response level* of a multiple classifier system [20]. So, in this work, the overproduce step is a consequence of evaluating multiple prediction scenarios from a pool of hydrological models, while the selection step is tackled with the NSGA-II algorithm. At this point, it is important to note the analogy between the classical problem of finding the inputs that give us the most information for training a given forecasting model (i.e. features selection) and selecting the predictors making the ensembles, which is investigated here. Henceforth we will refer to predictors selection as the effect of applying a features selection tool for selecting members at the response predictors level.

GECCO'13, July 6–10, 2013, Amsterdam, The Netherlands.

Features selection algorithms, as well as predictors selection, can roughly be grouped into two categories depending on their application model, that is *filter* methods and *wrapper* methods. Filter methods allow the selection to be made without involving the chosen learning/combining system, using instead some other measures, generally statistical ones. In contrast, wrapper methods use the performance of the chosen learning/combining system to guide the selection [1]. It is generally accepted that wrapper methods lead to higher performance [18] at the expense of high computational cost compared to filter approaches. It is frequently categorized as a "brute force" method, although it is not necessarily so [16]. Meta-heuristic procedures such as Genetic Algorithms (GA) have been proposed [9, 30, 36] as a wrapper method for features selection, with the advantage of reducing the computational cost, but also showing a capacity to find better solutions given its global search capabilities.

Another remarkable aspect in predictors selection based on wrapper methods is the central role played by the performance measures guiding the optimization process. Moreover, depending on the application, one might want to use several measures of different nature available for selecting the predictors. In that context, a common approach consists in aggregating these performance measures into a scalar global criterion, which has been called by some a global criterion method [21]. That approach has its drawback, foremostly requiring to articulate the trade-off between the performance measure beforehand, and being inefficient at handling an aggregation of objectives that have non linear (non convex) relations. Thus, a multiobjective framework emerges as a more natural and practical model of predictors selection with several performance measures. Indeed, in such a context, it may be more efficient to resort to an *a posteriori* multiobjective technique such as the Nondominated Sorting Genetic Algorithm II (NSGA-II) [8], where the optimization is generating as output a group of solutions that represents various trade-offs between the objectives, corresponding to the so-called Pareto front. An example of such an application, but in a deterministic context and with two UCI datasets, is presented in Waqas et al. [33].

In our work, based on the concepts outlined above, we apply NSGA-II in a multi-*score* framework for the selecting of the best predictors (better complementarity) in a real world application, that is a pool of 800 streamflow forecasting models. The remainder of the paper is organized as follows. Sec. 2 provides a presentation of the HEPS of reference and the cases studied. Sec. 3 summarizes the basic concepts underlying the scores to evaluate the performances of the HEPS. Follows in Sec. 4 a presentation of the working hypothesis used for our experiments. Sec. 5 details the experimental setup, while results and the corresponding analysis are presented in Sec. 6. Finally, some conclusions and a guideline for future work are given in Sec. 7.

2. HEPS SETUP AND LOCATIONS

The HEPS under study is formed of 16 hydrological models combined with 50 meteorological inputs given by the ECWMF EPS, leading to a grand ensemble of 800 members, usable for 1 to 9 days ahead forecasting. The precipitation inputs are *a priori* assumed to be equally likely [15]. Another important aspect of the HEPS at hand is the short duration of the series, from March 2005 to July 2006. This HEPS was implemented over 28 French catchments with an average response time of 3.2 days, representing a large range

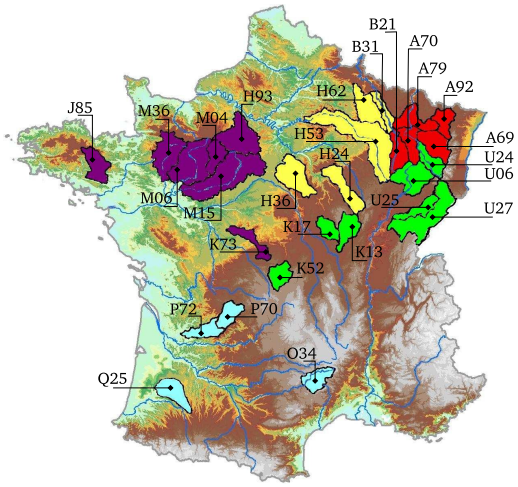

Figure 1: Catchments location.

of hydro-climatic conditions (Fig. 1). The results presented here are based on a selection of four basins that show different behaviours according to simplification. Basins H24, H93, M06, and P70 are used for training, whereas corresponding neighbouring basins H36, M04, M15, and P72 are used for testing.

The high performance of the 800-member HEPS for the 9-th day Forecast Time Horizon (FTH), and a lesser extent in the other FTHs, has been demonstrated in [31]. However, preliminary analysis has shown the high redundancy of the models response, opening the way to the hydrological model selection and precipitation clustering as an efficient simplification idea.

3. PERFORMANCE EVALUATION

In the machine learning community, the evolution of the bias-variance dilemma [12] to the accuracy-diversity breakdown [19], and the bias-variance-covariance decomposition [29], naturally supports the multiple classifier systems approach and, more generally speaking, the evaluation of multiple scenarios. However, this is frequently reduced to a single response to use in a scalar objective function. But in a probabilistic context, it is imperative to use the distribution-oriented approach (scores) that reveals more details about the response uncertainty.

The distribution-oriented approach reveals that forecast quality is inherently multifaceted in nature [7, 23]. In the following, we quote two of the properties commonly evaluated in probabilistic forecasting, that is the bias and the reliability. The reader is referred to Wilks [35] for a more detailed description of these and other features. Note that we use Gaussian distributions in the following examples for the sole purpose of facilitating the explanations.

The **bias**, also called unconditional bias or systematic bias, measures the correspondence between the mean forecast and the mean observation. In Fig. 2 we illustrate two cases regarding *pdf* A. The observation o^t first occurs near the mean forecast[1] (case 1), i.e. a system with low bias. Second, the observation o^t is located further from the central value (case 2), i.e. a highly biased system.

[1]Note that mean, median, and mode coincide with the peak of the Gaussian distribution.

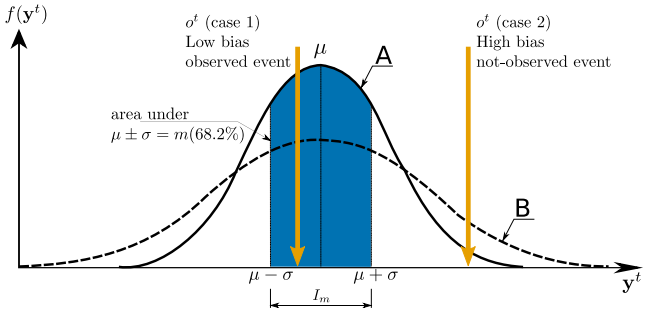

Figure 2: Probabilistic forecasting evaluation.

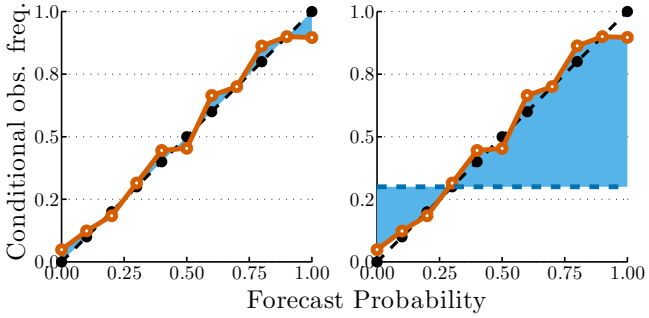

Figure 3: Reliability diagram.

The **reliability** relates to the occurrence of event o^t given a probability threshold m, averaged over all N observation - forecast pairs. In Fig. 2, the reliability evaluation of a threshold probability equal to 68.2% in *pdf* A. So, at each time step, it must be determined whether the event falls or not into the I_m region bounded by this probability value. Subsequently, the conditional observed frequency $\bar{o}_m$ is evaluated for the N observation - forecast pairs by Eq. 1:

$$\bar{o}_m = \frac{1}{N}\sum_{t=1}^{N} r^t, \text{ where } r^t = \begin{cases} 1 & \text{if } o^t \in I_m \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

Finally, for this probability threshold, the system is perfectly reliable if $\bar{o}_m$ is equal to m. For instance, in a perfectly reliable system with a probability threshold of 68.2%, we can imagine that for 1000 cases the event fell 682 times within the intervals evaluated. We say that the system is overforecasting if $\bar{o}_m$ is less than m, and underforecasting otherwise.

Given that m denotes the different M thresholds of probability to assess, the reliability of the system can be directly measured from the comparison of these thresholds with the M observed conditional probabilities. The goal is to have well-calibrated forecast systems for which the relative frequency is essentially equal to the probability of the forecast, i.e. $\bar{o}_m = m$.

In the following, we present two scores that represent bias and reliability, that is the ignorance score and the reliability diagram. Note that this antagonistic behaviour can be easily seen as another dimension of the bias-variance dilemma.

3.1 Ignorance score

Proposed by Good [14] as the logarithmic score, the IGNorance Score (IGNS), given in Eq. 2, is defined as the logarithm of the ensemble probability density function $f(\mathbf{y}^t)$ at the point corresponding to the observation o^t, all evaluated at time t:

$$\text{IGNS}\left(\mathbf{y}^t,\ o^t\right) = -\log_2 \left[f\left(\mathbf{y}^t\right)_{o^t}\right]. \quad (2)$$

Note that this score can take negative values because the *pdf* may be larger than one[2]. Smaller values indicate better performance. The IGNS is a local measure that severely penalizes the bias because positioning the observation in forecast regions of low probability leads to values that tend to infinity. The logarithmic score involves a harsh penalty for low probability events and therefore is highly sensitive to extreme cases [13]. To rule out the possibility that the results solely reflect the effect of a few outliers, we analyzed trimmed means of the IGNS series excluding the highest and lowest 2% data values, following Weigend and Shi [34]. Infinite values were replaced by the next worst non-infinite value, following Boucher et al. [2].

[2] *pdf* values denotes an intensity of probability or a probability rate.

3.2 Reliability diagram

The reliability diagram or attributes diagram is a graphical representation of the joint distribution of the forecasts and observations. For its construction, we define M probability thresholds, often deciles, then we compute the conditional observed frequency for each of these M thresholds. Finally, we illustrate the relationship between forecast probabilities and conditional observed frequency. In a perfectly reliable system, $\bar{o}_m$ will be equal to m, i.e. the distance or area between the 1:1 line and computed pairs $(m, \bar{o}_m)$, will be very small (left panel of Fig. 3). Consequently we can evaluate the reliability of the system from a distance measure such as the Mean Square Error (MSE):

$$\text{RD}_{\text{MSE}}(\mathbf{Y},\ \mathbf{o}) = \frac{1}{M}\sum_{i=1}^{M}(\bar{o}_{m_i} - m_i)^2, \quad (3)$$

where $m_i \in [0, 1]$.

Note that Eq. 3 corresponds to the reliability as defined in the Brier score decomposition [35]. It is clear that to maximize the reliability, one seeks at minimizing Eq. 3. The reliability diagram proposes a direct assessment of reliability and resolution of a probability forecast. Regarding the resolution (ability of the forecast to distinguish situations with different frequencies of occurrence), its measure is given by the difference between each of conditional observed probabilities $\bar{o}_m$ and the overall unconditional mean observation $\bar{o}$ (see No-resolution line in right panel of Fig. 3).

Finally, a reliability diagram diagnosis leads to determining overforecasting or underforecasting. For example if the curve is below the 1:1 line, that indicates that the average forecast is larger than the average observation (overforecasting). But, if the curve is above the 1:1 line (underforecasting), the average forecast is smaller than the average observation.

4. BASIC ASSUMPTIONS

A subtle but key aspect in the simplification scheme is the equiprobability condition of the precipitation inputs. Note that while meteorological members are interchangeable, the occurrence of each hydrological model within HEPS stays invariable. For all time steps, the first 50 hydrological members correspond to the combination of 50 precipitation members and the hydrological model #1. Similarly the last

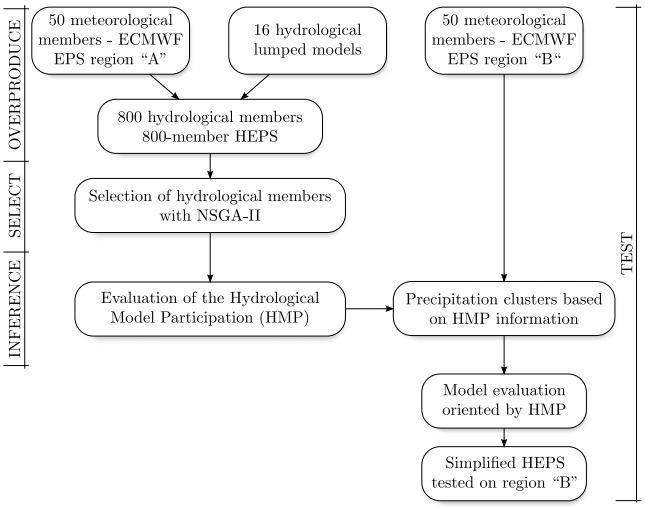

Figure 4: HEPS simplification based on clustering and different HMP.

50 hydrological members (751-800) correspond to the same combination with the hydrological model #16. It is clear that hydrological models act as non-linear filters in which one of the variables is precipitation. We assume that the hydrological models that form the HEPS of reference reflect different conceptualizations of the hydrological process. In this way, the removal of a member only represents a loss of model weight in the proposed simplified scheme that is based on the Hydrological Models' Participation (HMP) as a key concept of HEPS simplification.

In conclusion, we hypothesize that considering each hydrological member as a variable is not in conflict with the proposed methodology, because the selection of members, for subsequent interpretation in terms of HMP, is not made on members of the ECMWF EPS but rather on the 800-member hydrological response. The empirical validity of these assumptions will be evaluated particularly in Sec. 6 from the evaluation of multiple random predictors selections.

Beside that, Brochero et al. [3] and Velázquez et al. [31] showed that the 800-member HEPS has a high performance on the 9-th day FTH. Consequently, we apply the simplification process on the database corresponding to this lead time. This decision about the FTH is justified since the hydrological model participation as a method of simplifying HEPS should be unique regardless of the FTH.

5. EXPERIMENTAL SETUP

In Fig. 4 we illustrate the scheme of simplification and test. This procedure consists of four stages: overproduce, select, inference, and test. In this figure we outline how to train and test systematic selection results. For training, an 800-member HEPS database of a given catchment is analyzed for the 9-th day lead time. For testing, we use a neighbouring basin and other forecast time horizons.

5.1 Overproduce

The HEPS under study is formed of 16 hydrological models combined to one of the 50 meteorological inputs of the ECWMF EPS, leading to a grand-ensemble of 800 members. The MEPS members are *a priori* assumed to be equally likely [15]. Another important feature of the HEPS

Representation	Truncated permutations
Recombination	Partially mapped crossover (PMX)
Recombination probability	90%
Mutation	Swap
Mutation probability	2% for each allele
Parent selection	Best 2 out of random 4
Survival selection	Pareto-front rank and crowding distance
Population size	100
Initialization	Random
Termination condition	300 generations

Table 1: NSGA-II setup.

at hand is the short duration of the series, from March 2005 to July 2006. This has been highlighted by several authors as a negative point in the evaluation of system performance in the case of extreme events [7, 25]. However, other studies that focused on periods of analysis very similar to the one used here have also proven the usefulness of the ECMWF EPS. For example Rousset et al. [27] 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.

5.2 Predictors Selection

5.2.1 *Schemes to analyze*

Following [3, 4], who demonstrated in this same database that the best balance of scores is achieved with a number of selected members fluctuating between 30 and 100, here we use NSGA-II to select the best 48 predictors based on the scores presented in Sec. 3. This allows a fair comparison to be made with two reference models: i) a naive model that consists in a uniform participation of three forecasts for each of the sixteen hydrological models, resulting in a 48-member HEPS called 48UP, and ii) the Median of 200 Random member selections, scheme called 48MR.

5.2.2 *NSGA-II setup*

This technique uses a permutation representation GA following the NSGA-II algorithm. It has received the most attention because of its simplicity and demonstrated superiority over other multiobjective optimization methods [32]. A detailed description of the NSGA-II algorithm appears in [8]. Like any evolutionary algorithm, its configuration is given by way of representing potential solutions (genotype), evaluation function (or fitness function), population, initialization, parent selection mechanism, variation operators, survivor selection, and termination condition. Table 1 presents a summary of the configuration evaluated here.

Truncated permutation as representation is defined by the following guidelines: i) the individual should represent 48 *unique* members in order to compare solutions with other methods, and ii) in terms of selection of variables, the permutation of the same group of members does not represent a new solution. Thus, the representation of each individual (candidate solution) is a permutation of a set of 800 integers. However, only the first 48 alleles (string positions) represent the solutions to be tested. The other 752 alleles are reserved for the application of the variation operators presented in Table 1.

Moreover, with the goal of comparing NSGA-II with the other two selections schemes, a representative solution of the Pareto-front is necessary, in which case we orient this

30-member HEPS subset	800-member HEPS		HMP	HM weight(%)
	HM $\#i$	Members interval $[50i-49, 50i]$		
10, 12, 23, 25, 34,	1	1 – 50	7	23.3
42, 45, 55, 63, 70,	2	51 – 100	3	10.0
245, 247, 345, 350,	5	201 – 250	2	6.7
654, 680, 690, 700,	7	301 – 350	2	6.7
701, 710, 751, 753,	14	651 – 700	4	13.3
755, 757, 759, 760,	15	701 – 750	2	6.7
778, 780, 785, 800	16	751 – 800	10	33.3

Table 2: Hypothetical example to show the HMP.

selection with the post-Pareto front analysis proposed by Chaudhari et al. [6]. Therefore, the procedure consists in the following steps:

- Obtain a subset of solutions that represent the Pareto-optimal front.
- Apply k-means clustering on that Pareto front. For that purpose a normalization space is needed to avoid problems arising from the scale of the scores. Here we normalize each variable so that they will have zero mean and unit standard deviation. To find the "optimal" number of clusters, we evaluate the number of clusters that maximizes the mean silhouette value. The silhouette is a measure of how close each point forming a cluster is to others in the neighbouring clusters.
- For each cluster, select a representative solution. To do this, the solution that is closest to its respective cluster centroid is chosen as a good representative solution.
- Analyze the "knee" cluster or the k representative solutions. In this study the "knee" cluster is considered as the most interesting solution because we do not propose an operational scenario in which one could decide which of the two functions to minimize is more relevant at a given time.

5.3 Inference

Following the assumptions made in Sec. 4, the simplification of the 800-member HEPS is based on the direct systematic selection of certain hydrological members, which indirectly leads us to determine the hydrological models participation (HMP) as the base criterion of the simplification task.

Consider the example given in Table 2 for a simplified HEPS of 30 members presented in the first column. Assuming that this simplified scheme provides at least equal performance than the 800-member HEPS, Table 2 shows that the evaluation and posterior combination of Hydrological Models (HM) can be reduced substantially (7 instead of 16). Note that the last two columns show the apparent higher relevance of models 1 and 16; however, other models of less weight can be important to describe the uncertainty of the process with opposing views to those of the most relevant models.

5.4 Test

To assess hydrological models with representative precipitation members (RM), as proposed by several authors [17, 22] and following the trend of recent clustering products produced directly at ECMWF EPS [10], the HMP directly orients the evaluation of representative precipitation members at each time step to subsequently propagate them into its respective hydrological model. For this, we rely on the k-means technique configured with the Euclidean distance as the similarity measure, and the HMP to define the number of clusters, so the EPS member closest to its cluster centroid will define the representative member.

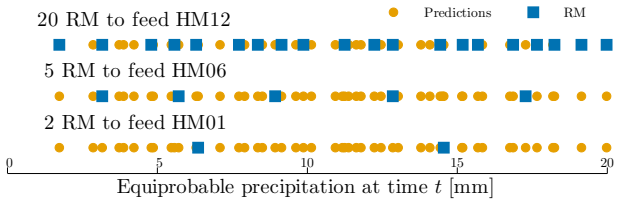

Figure 5: Representative members for simplification testing.

Consider the following example to clarify the previous concepts: we have a HMP, of 16 elements that represent the 16 hydrological models, after applying a particular selection technique and evaluating the final number of members corresponding to each HM.

$$\text{HMP} = \{\overbrace{2}^{\text{HM01}}, 0, 0, 0, 0, \overbrace{5}^{\text{HM06}}, 0, 0, \overbrace{16}^{\text{HM09}}, 0, 0, \overbrace{20}^{\text{HM12}}, 0, \overbrace{5}^{\text{HM14}}, 0, 0\}$$

So, for test purposes, at each time step we evaluate the two precipitation members closest to two cluster centroids and we propagate them into the hydrological model HM01. We continue so on to evaluate the five representative precipitation members into the HM14. Fig. 5 presents a visual exploration of Representative precipitation Members (RM).

Finally, we compare the HMP results based on the NSGA-II, the 48MR and the 48UP schemes. For comparison purposes, each score in the selected ensemble of hydrological members (*se* subscript) is normalized by dividing it with the corresponding score in the initial 800-member ensemble (*ie* subscript), therefore using a similar scale for each component. Also we estimated a normalization threshold for the IGNS equal to -3 to manipulate the negative values of this score:

$$\text{normalized } \overline{\text{IGNS}} = \frac{-3 - \overline{\text{IGNS}}_{\text{se}}}{-3 - \overline{\text{IGNS}}_{\text{ie}}}, \quad (4)$$

$$\text{normalized } \text{RD}_{\text{MSE}} = \frac{\text{RD}_{\text{MSE}_{\text{se}}}}{\text{RD}_{\text{MSE}_{\text{ie}}}}. \quad (5)$$

6. RESULTS AND ANALYSIS

Table 3 compares results of the initial 800-member HEPS and the two reference schemes: a naive model of a Uniform HMP (48UP) and the Median of 200 Random member selections (48MR). Hereafter, RD_{MSE} values are expressed on a 10^{-3} basis.

With reference to 48-member schemes, it is noteworthy that in all cases the IGNS of reference is improved, at the expense of reliability. Specifically the uniform HMP scheme is the most efficient in minimizing IGNS, while the median of random evaluations is a little less inefficient in terms of reliability (RD_{MSE}), but not yet matching the performance of reference (800 members). Consequently, simplification stands out as an optimization task involving hydrological models in a weight assignment problem.

In an effort to visualize explicitly the trade-off between the scores, Fig. 6 presents such a Pareto-type analysis of results

Training catchments	HEPS members	Scores		Testing catchments	HEPS members	Scores	
		RD_{MSE}	IGNS			RD_{MSE}	IGNS
H24	800	7.08	0.41	H36	800	3.50	−0.09
	48UP	8.20	−0.49		48UP	5.00	−0.75
	48MR	7.82	−0.46		48MR	4.67	−0.74
H93	800	2.59	−0.27	M04	800	1.74	−0.03
	48UP	4.47	−0.98		48UP	4.31	−0.72
	48MR	3.84	−0.95		48MR	3.50	−0.70
M06	800	1.42	−0.14	M15	800	1.61	0.28
	48UP	4.25	−0.77		48UP	3.37	−0.64
	48MR	3.19	−0.74		48MR	2.62	−0.61
P70	800	4.14	3.29	P72	800	4.39	0.89
	48UP	5.22	−0.06		48UP	5.28	−0.39
	48MR	4.51	0.06		48MR	4.51	−0.33

Table 3: Performance for a 9-day FTH in different schemes.

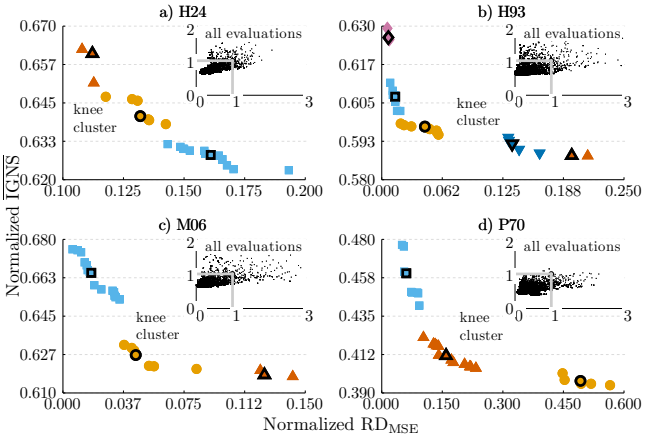

Figure 6: Evaluations of different selections with NSGA-II.

obtained with NSGA-II. Each panel illustrates the behaviour of different selections for each basin. The inset figure in the upper right corner of each panel shows the 30 000 tested selections in the optimization process. In these figures, we can see that the reliability of the system is the principal component of the variability of the results. Similarly, from the density of points outside the bounding rectangle, limited by unit normalized scores, we have evidence that the optimization process converges rapidly in the initial performance of the reference HEPS. Thus, the simplification process is shown primarily as a process of scores optimization, hence it may be referred to as a post-processing step.

The Pareto front obtained in the optimization is drawn in each panel, along with their respective clusters and centroids identification (symbols in bold). The scale of normalized scores of each panel is similar, except for basin P70 where the IGNS and the RD_{MSE} exhibit larger Pareto ranges, between 0.39 and 0.48, and between 0 and 0.6, respectively. This difference could be related to two factors specific to this basin: its lower drainage area and its higher relative flow. However, it is necessary to perform more experiments to identify the relationship of these factors.

Importantly, the Pareto front or simplification based on the centroid of each cluster offers a descriptive version of the optimization process, which allows the development of a decision process based on the characteristics of each score and the properties we want to prioritize in a particular case. In other words, NSGA-II offers more flexibility to the operational hydrologists than a single subset of predictors obtained with *a priori* multiobjective optimization methods.

However, a rigorous test requires evaluating the simplification model against new information. Thus, the above results are really optimistic indicators. Accordingly, Fig. 7 shows the results of different selection techniques in a framework of extrapolation in both time and space – executed on a nearby catchment at different forecast time horizons. bp-rand corresponds to box-plots of the 200 random selections executed.

As for the reliability of the system (left panels in Fig. 7), these results demonstrate again that the main difficulty lies in the preservation of this property. Furthermore, random selections show that the situation is more dramatic in FTH over 6 days, except in the basin P72 where the median of the random selections is around one. This behaviour is especially important if one considers that the benefits of the 800-member HEPS is focused primarily on these lead times. Also note that the interquartile range (length of the box-plots) exhibits an important dispersion in relation to the results observed with respect to IGNS (right panels in Fig. 7). Concerning the 48UP scheme, it is important to note that the tendency is similar than for the random selections. However, it is remarkable that in the initial FTHs (1 to 6 days) uniform selection is generally better than the first quartile of the 200 random selections tested. The high performances achieved with NSGA-II are obvious.

7. CONCLUSION

Given the singularity of HEPS in evaluation, where its 800 hydrological members come from the propagation of 50 interchangeable meteorological members, simplification scheme and/or scores optimization of the system based on the Hydrological Models Participation (HMP) has proven to be highly efficient. Clearly, the methodology presented here combines the HMP and the meteorological clustering stage as additional filters that facilitate the interpretation of the hydrological member selection. However, in the case of a HEPS conceived from non-interchangeable meteorological members (in Canada, for example), the selection task would

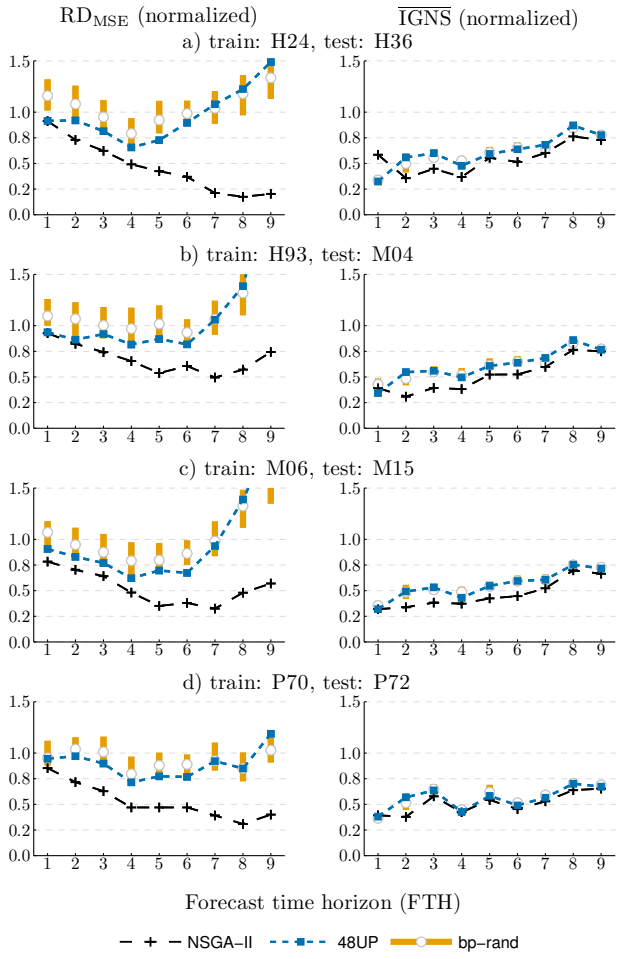

Figure 7: Test of different HEPS simplification schemes of 48 members.

then directly identify the importance of certain members in the propagation of uncertainty in flow prediction.

With respect to the reliability of the system, this property is notoriously the most difficult to maintain in the simplification scheme. The results of a uniform HMP and random selections show such difficulty. In this work, without loss of reference bias, it is important to note the high efficiency of the NSGA-II, which we consider to be the best choice among the techniques evaluated. At this point, operational hydrologists can highlight the following aspects:

- **Computational complexity** Although the Backward selection suggested in [3] is more intuitive, its complexity is the order of $O\left(d^2\right) = O\left(800^2\right)$ while the NSGA-II has a computational complexity of $O\left(s \cdot p^2\right)$, where s is the number of scores and p is the population size. Moreover, there are several open source user-friendly frameworks that include an efficient implementation of NSGA-II [11].
- **Trade-off between bias and reliability** Local optimization methods with a global criterion, as suggested in [3], do not allow a direct visualization of the relationship between multiple objectives. Although in the literature has proposed the manipulation of objectives weight to build a Pareto Front, the NSGA-II shows directly the different simplification schemes that highlight the trade-off among the evaluated objectives.
- **Optimization of the number of members to retain** Because the objective of our methodology is focused on a comparison of techniques with the same number of members, solution representation (genotype) with NSGA-II was established by permutations. However, a binary encoding would be more intuitive in order to optimize at the same time the number of members to hold out.
- **Search procedure** Finding the optimal subset is not guaranteed with a local search procedure [21]. For example, y_i and y_j by themselves may not be good, but taken together, they may decrease significantly the error. But because these algorithms are local or greedy and remove members one by one, they may not be able to detect this. In contrast, the NSGA-II has the capability to find the global optimum in its evolutionary search process in such a deceptive setting, although there is no absolute guarantee that it will find it.

Finally, we propose various directions for future research:

- Explore more probabilistic features through other scores as the δ ratio or the CRPS.
- Further research with longer databases is needed in order to identify the HEPS value in several type of events, e.g. peak events.
- Explore using an even larger number of basins in order to evaluate the relation between the selection performance and physiographic and hydro-climatological properties.
- Furthermore the diversity, evaluated from deterministic performance of each model, should be considered as an approximation to the true structural diversity of hydrological models, in this sense an explicit analysis of the relationship between the structural diversity of a group of hydrological models and their relevance in a probabilistic scheme should be studied in more detail.

Acknowledgements

The authors acknowledge NSERC (Canada) and the Institute EDS for financial support, CEMAGREF and ECWMF for making the databases available. We also thank J. Vrugt for providing a Matlab version of NSGA-II and A. Schwerdtfeger for proofreading this manuscript.

References

[1] E. Alpaydin. *Introduction to Machine Learning*. 2nd ed. The MIT Press, 2010.

[2] M.-A. Boucher, J.-P. Laliberté, and F. Anctil. "An experiment on the evolution of an ensemble of neural networks for streamflow forecasting". In: *Hydrol. Earth Syst. Sci.* 14.3 (2010), pp. 603–612.

[3] D. Brochero, F. Anctil, and C. Gagné. "Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria". In: *Hydrol. Earth Syst. Sci.* 15.11 (2011), pp. 3307–3325.

[4] D. Brochero, F. Anctil, and C. Gagné. "Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 2: Generalization in time and space". In: *Hydrol. Earth Syst. Sci.* 15.11 (2011), pp. 3327–3341.

[5] D. Brochero, F. Anctil, and C. Gagné. *An experience on the selection of members for simplifying a multimodel hydrological ensemble prediction system*. In: *CSHS Workshop: Operational River Flow and Water Supply Forecasting*. 2011.

[6] P. Chaudhari, R. Dharaskar, and V. M. Thakare. "Computing the Most Significant Solution from Pareto Front obtained in Multi-objective Evolutionary". In: *IJACSA* 1.4 (2010), pp. 63–68.

[7] H. L. Cloke and F. Pappenberger. "Ensemble flood forecasting: A review". In: *J. Hydrol.* 375.3-4 (2009), pp. 613–626.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Trans. Evol. Comp.* 6.2 (2002), pp. 182–197.

[9] A. Eiben and J. Smith. *Introduction to evolutionary computing*. 1, Corr. 2nd printing. Springer, 2003.

[10] L. Ferranti and S. Corti. "New clustering products". In: *ECMWF Newsl.* 127 (2011), pp. 6–12.

[11] F.-A. Fortin, F.-M. D. Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné. "DEAP: Evolutionary Algorithms Made Easy". In: *Journal of Machine Learning Research* 13 (2012), pp. 2171–2175.

[12] S. Geman, E. Bienenstock, and R. Doursat. "Neural networks and the bias variance dilemma". In: *Neural Comput.* 4.1 (1992), pp. 1–58.

[13] T. Gneiting and A. E. Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *J. Am. Statist. Assoc.* 102.477 (2007), pp. 359–378.

[14] I. J. Good. "Rational Decisions". In: *J. R. Stat. Soc.* 14.1 (1952), pp. 107–114.

[15] B. T. Gouweleeuw, J. Thielen, G. Franchello, A. P. J. De Roo, and R. Buizza. "Flood forecasting using medium-range probabilistic weather prediction". In: *Hydrol. Earth Syst. Sci.* 9.4 (2005), pp. 365–380.

[16] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[17] S. Jaun, B. Ahrens, A. Walser, T. Ewen, and C. Schär. "A probabilistic view on the August 2005 floods in the upper Rhine catchment". In: *Nat. Hazards Earth Syst. Sci.* 8 (2008), pp. 281–291.

[18] R. Kohavi and G. H. John. "Wrappers for feature subset selection". In: *Artif. Intell.* 97.1–2 (1997), pp. 273–324.

[19] A. Krogh and J. Vedelsby. *Neural Network Ensembles, Cross Validation, and Active Learning*. In: *Advances in Neural Information Processing Systems 8*. MIT Press, 1995, pp. 231–238.

[20] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[21] R. Marler and J. Arora. "Survey of multi-objective optimization methods for engineering". In: *Struct. Multidiscip. O.* 26 (6 2004), pp. 369–395.

[22] F. Molteni, R. Buizza, C. Marsigli, A. Montani, F. Nerozzi, and T. Paccagnella. "A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments". In: *Q. J. R. Meteorolog. Soc.* 127 (2001), pp. 2069–2094.

[23] A. H. Murphy. "What is a good forecast? An essay on the nature of goodness in weather forecasting". In: *Wea. Forecasting* 8 (1993), pp. 281–293.

[24] F. Pappenberger, K. J. Beven, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, J. Thielen, and A. P. J. de Roo. "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)". In: *Hydrol. Earth Syst. Sci.* 9.4 (2005), pp. 381–393.

[25] M. Renner, M. Werner, S. Rademacher, and E. Sprokkereef. "Verification of ensemble flow forecasts for the River Rhine". In: *J. Hydrol.* 376.3-4 (2009), pp. 463–475.

[26] E. Roulin. "Skill and relative economic value of medium-range hydrological ensemble predictions". In: *Hydrol. Earth Syst. Sci.* 11.2 (2007), pp. 725–737.

[27] F. Rousset, F. Habets, E. Martin, and J. Noilhan. "Ensemble streamflow forecasts over France". In: *ECMWF Newsl.* 111 (2007), pp. 21–27.

[28] J. C. Schaake, T. M. Hamill, R. Buizza, and M. Clark. "HEPEX: The Hydrological Ensemble Prediction Experiment". In: *Bull. Am. Meteorol. Soc.* 88.10 (2007), pp. 1541–1547.

[29] N. Ueda and R. Nakano. *Generalization error of ensemble estimators*. In: *IEEE International Conference on Neural Networks*. Vol. 1. 1996, pp. 90–95.

[30] H. Vafaie and K. De Jong. *Genetic algorithms as a tool for feature selection in machine learning*. In: *Fourth International Conference on Tools with Artificial Intelligence*. 1992, pp. 200–203.

[31] J. A. Velázquez, F. Anctil, M. H. Ramos, and C. Perrin. "Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures". In: *Adv. Geosci.* 29 (2011), pp. 33–42.

[32] J. Vrugt and B. Robinson. *Improved evolutionary optimization from genetically adaptive multimethod search*. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 104. 3. 2007, pp. 708–711.

[33] K. Waqas, R. Baig, and S. Ali. *Feature subset selection using multi-objective genetic algorithms*. In: *IEEE International Multitopic Conference (INMIC)*. 2009, pp. 1–6.

[34] A. S. Weigend and S. Shi. "Predicting daily probability distributions of S&P500 returns". In: *J. Forecast.* 19.4 (2000), pp. 375–392.

[35] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Vol. 100. Academic Press, 2011.

[36] J. Yang and V. Honavar. "Feature subset selection using a genetic algorithm". In: *IEEE Intell. Syst.* 13.2 (1998), pp. 44–49.