Modeling Earth Systems and Environment (2023) 9:1103–1118
https://doi.org/10.1007/s40808-022-01540-1

ORIGINAL ARTICLE

Improving the predictive skills of hydrological models using a combinatorial optimization algorithm and artificial neural networks

Juan F. Farfán[1] · Luis Cea[1]

Received: 1 July 2022 / Accepted: 18 September 2022 / Published online: 7 October 2022

Abstract
Ensemble modelling is a numerical technique used to combine the results of a number of different individual models in order to obtain more robust, better-fitting predictions. The main drawback of ensemble modeling is the identification of the individual models that can be efficiently combined. The present study proposes a strategy based on the Random-Restart Hill-Climbing algorithm to efficiently build ANN-based hydrological ensemble models. The proposed technique is applied in a case study, using three different criteria for identifying the model combinations, different number of individual models to build the ensemble, and two different ANN training algorithms. The results show that model combinations based on the Pearson coefficient produce the best ensembles, outperforming the best individual model in 100% of the cases, and reaching NSE values up to 0.91 in the validation period. Furthermore, the Levenberg-Marquardt training algorithm showed a much lower computational cost than the Bayesian regularisation algorithm, with no significant differences in terms of accuracy.

Highlights

- Random-Restart Hill-Climbing algorithm could be adapted to identify optimal ensemble models.
- Ensemble models enhance model results in terms of linear correlation, bias, and variability.
- At least 7 individual models are necessary to ensure a good fitting ensemble model.
- Individual models with a high Pearson or high NSE can yield robust ensembles.
- Levenberg-Marquardt shows similar results to Bayesian Regularization with much lower computationalcost.

Keywords Ensemble model · Artificial neural networks · Hydrological model · Water resources

Introduction

Watershed modelling is an essential tool in hydrological studies that can be used for different purposes as water resources management or the analysis of extreme events such as floods and droughts, among others. However, since hydrological models are rough representations of real watersheds, the accuracy of discharge predictions generated with the models can directly affect the management of the water resources (Wang et al. 2017; Farfán et al. 2020).

A technique that has been applied in different areas of environmental modelling to overcome this limitation and obtain more robust and accurate results is the use of ensemble models, which consists of combining different individual models to compensate the deficiencies of each other (Schäfer Rodrigues Silva et al. 2022; Najafi and Moradkhani 2016; Farfán et al. 2020; Wang et al. 2009; Viney et al. 2009; Velázquez et al. 2011; Duan et al. 2007).

In the context of hydrological systems modeling, the application of ensemble models covers a very wide range of studies, such as streamflow modeling, groundwater

✉ Juan F. Farfán
j.farfan@udc.es

Luis Cea
luis.cea@udc.es

[1] Center for Technological Innovation in Construction and Civil Engineering (CITEEC), University of A Coruña, A Coruña, Spain

modeling, global hydrological simulations, droughts, floods, reservoir modeling, water resources management, agricultural irrigation, land use change, climate change, among others (Shen et al. 2022; Jafarzadeh et al. 2022; Qi et al. 2021; Wang et al. 2009, 2017; Liu et al. 2022; Nourani et al. 2022; Viney et al. 2009; Bermúdez et al. 2021; Zhang and Yang 2018; Farfán et al. 2020). To this end, ensemble modelling combines simulations obtained with different parameter sets, initial conditions, or model structures (Liu et al. 2022; DeChant and Moradkhani 2014; Najafi et al. 2012; Najafi and Moradkhani 2016), which allows the modeller to extract a large amount of information from a set of existing models and combine them in an optimal way (Li and Sankarasubramanian 2012; Viney et al. 2009). Ensemble models can be categorized into 2 different types: (1) single-model ensembles (SME) and (2) multi-model ensembles (MME) (Li et al. 2022; Viney et al. 2009; Najafi and Moradkhani 2016). In SME the individual models are obtained from the same hydrological model using different parameter sets. Since only one hydrological model is used. Its main strength is the reduction of the uncertainty related to the model parameters, which tends to be dominant for modelling processes at finer time scales (Li et al. 2016). On the other hand, in MME the individual models are obtained from different hydrological models and thus, they could help to reduce the uncertainty associated with the model structure, which tends to be dominant at coarser time scales (i.e. monthly) (Viney et al. 2009; Najafi et al. 2012; DeChant and Moradkhani 2014; Najafi and Moradkhani 2016; Li et al. 2016; Zhang and Yang 2018).

The techniques used in ensemble modelling include approaches such as calculating the average or weighted average of the individual models (Madadgar and Moradkhani 2014; Arsenault et al. 2015; Duan et al. 2007; Dong et al. 2013; Zhang and Yang 2018; Tyralis et al. 2021), multiple linear regression techniques such as constrained and unconstrained least squares (Kumar et al. 2015; Najafi and Moradkhani 2016), and more complex methods involving the application of artificial intelligence techniques inspired in the structure of the brain such as the the well known Artificial Neural Networks (ANN) (Li et al. 2022; Farfán et al. 2020; Li et al. 2018). Comparisons between different techniques can be found in the works of Andraos and Najem (2020), Tyralis et al. (2021), Shamseldin et al. (2007).

In addition to the technique for combining model outputs, the modeller must decide which individual models to include in the ensemble (Kumar et al. 2015; Li and Sankarasubramanian 2012; Viney et al. 2009). In this respect, several studies have reported that using the models with the highest individual performance does not necessarily produce the best ensemble. On the contrary, individual models with relatively lower fits may produce a better error compensation in the ensemble (Viney et al. 2009; Kumar et al. 2015).

In the particular case of ANN-based ensembles, additional decisions must be taken, as the number of neurons in the hidden layer and the ANN training algorithm. Several works have studied the number of neurons that must be used to build the ensemble models to avoid the adverse phenomena of overfitting (Farfán et al. 2020; Shamseldin et al. 2007). Other studies have analyzed the optimal number of individual models that should be used to build the ensemble (Londhe and Shah 2019; Phukoetphim et al. 2014; Arsenault et al. 2015).

Despite this important insight, the approaches to identify an optimal combination of individual models are scarce studied. In this regard, the referenced works, despite reporting good results after implementation, do not provide criteria for the selection of models to be used in the ensemble. This lack of criteria may lead to the training of ANN-based ensembles using individual models that are not the most appropriate for this purpose. This implies that the ensemble model results may fail to outperform the best individual model, discarding ANN as an option for building ensemble models since the selection of ensemble members is not a straightforward task. (Yaseen et al. 2015).

A possible method to identify adequate ANN-based (or others) ensemble models is to generate a large number of individual models and to evaluate the performance obtained using different combinations of them. Then, finding optimal combinations of individual models can be addressed as a combinatorial optimisation problem, which consists of searching for an object inside a finite collection of a large number of objects, when evaluating all of them one by one is not a feasible option (Schrijver 2003). The Hill-Climbing algorithm is a technique for solving combinatorial optimisation problems. The algorithm is computationally efficient since it compares only a current object against a randomly generated one in each iteration. Due to its easy applicability, this algorithm has been used in different works focused on branches of engineering and game theory (Lim et al. 2006; Ceylan 2006; Al-Betar et al. 2017; Alsukni et al. 2019).

Considering the potential advantage of the application of ensemble modeling to hydrological systems, and the lack of criteria for the selection of ensemble members, in the present work, we provide a strategy for the construction of ANN-based ensembles by means of a combinatorial optimisation approach. We use a continuous lumped hydrological model that is run with a number of parameter sets randomly sampled from its feasible space, to obtain a collection of individual hydrological models. Then, we propose to adapt the Random-Restart Hill-Climbing Algorithm (RRHC) to identify optimal ANN-based ensemble models that enhance the predictions of the individual hydrological models. The results are evaluated in terms of: (1) The effectiveness of the combinatorial optimisation approach to identify reliable ANN-based ensembles, (2) the goodness of fit of the identified ensembles, (3) The computational cost of the proposed technique.

The present paper uses as a case study a watershed located in Galicia, Northwest of Spain and is organised as follows: Sect. 2 describes the study area and available data. Then, in Sect. 3, we describe the methodology concerning the applied hydrological model, the ANN-based ensemble technique, the combinatorial optimisation algorithm (HR) and the criteria for the goodness of fit evaluation. Section 4 analyses and discusses the results obtained and finally, in Sect. 5 the conclusions of the work are provided

Study area and data

The Anllóns river basin is used as a study case to develop and test the methodology presented in this paper. The basin is located in the region of Galicia (NW Spain) and is affected by low pressures from the Arctic Polar front coming from the North Atlantic, which reaches the European coastline influenced by westerly winds during the months of October to March, bringing with it gusts of wind and intense rainfall (Cabalar Fuentes 2005). Historical rainfall and temperature data with a time resolution of one hour are publicly available at several observation stations from the regional meteorological agency (MeteoGalicia). From the MeteoGalicia meteorological network, 16 stations were used to interpolate the precipitation and temperature over the Anllóns catchment (Fig. 1). The average annual precipitation in the catchment is 1400 mm while the average annual maximum daily precipitation is 56 mm.

There is one stream gauge station in the basin, managed by the regional water administration (Augas de Galicia), with publicly available 10 min discharge data, although for this study the data were aggregated with a time step of 1 h to have the same temporal resolution as the rainfall and temperature data. The location of the stream gauge stations is shown in Fig. 1. The Anllóns stream gauge station is located about 53 km upstream of the river mouth, and its drainage area is 438 km^2. The basin is not affected by reservoir regulation, so the discharges measured are representative of natural conditions.

Calibration and validation periods

The available observed time series of precipitation, temperature and discharge were split into two periods that were used for the calibration and validation of the proposed methodology. There are 7 years of discharge data available, that have been split in two periods of 5 years for calibration and 2 years for validation. The calibration period spans from 2008 to 2013, while the validation periods starts in 2013 and ends in 2015. Table 1 provides some discharge statistics for the calibration and validation periods.

Methods

As mentioned in the introduction section, single model ensembles are adequate for the reduction of parameter-related uncertainty, which is dominant at fine time scales (Li et al. 2016). Therefore, since in the present study we

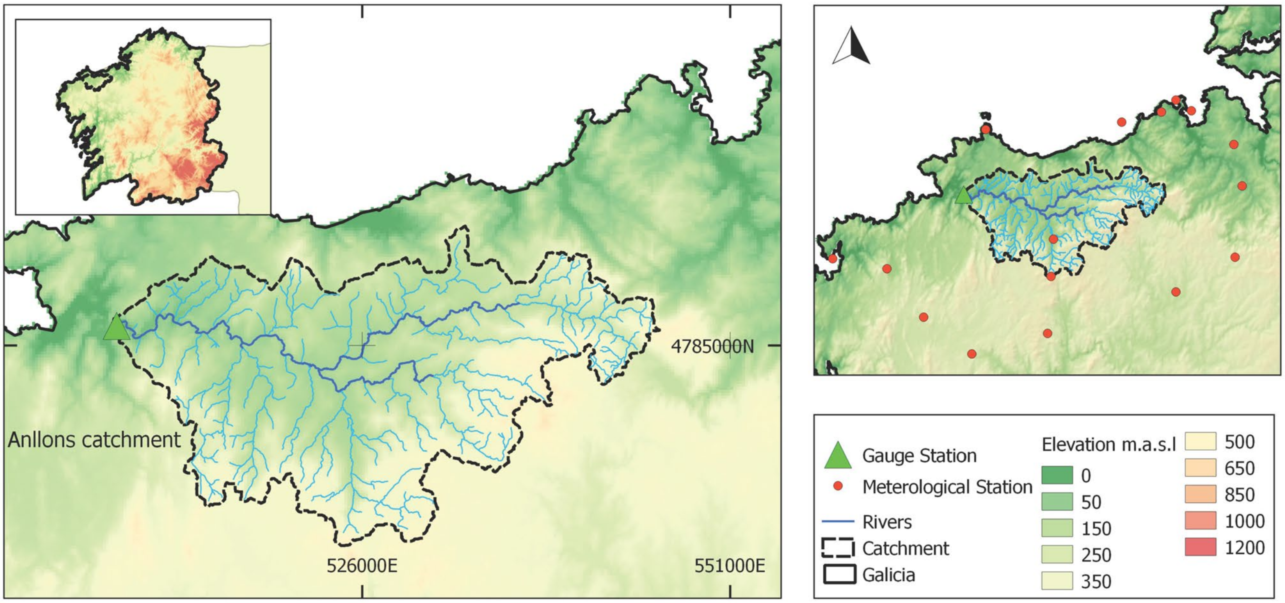

Fig. 1 Anllóns catchment, including the location of the stream gauge and meteorological stations

Table 1 Representative discharges in the calibration and validation periods for the Anllóns stream gauge station

Gauge	Period	Minimum	Q25	Q50	Mean	Q75	Maximum	Standard deviation
Anllóns	Calibration (2008 - 2013)	1.37	2.53	7.70	11.51	15.80	103.62	12.22
	Validation (2013 - 2015)	1.45	2.41	7.20	13.11	18.42	112.74	15.63

Computed from the hourly discharge time series. Q_p represents the $p-th$ percentile

work on a 1 h scale, we focus on the application of single model ensembles. The individual models that compose the ensemble are obtained from one hydrological model run with a number of different parameter sets. The model output obtained with each parameter set is treated as a member of the ensemble.

The hydrological model MHIA (acronym for *Lumped Hydrological Model* in Spanish) is used in two processes: (1) To generate a number of individual models using the Monte Carlo method and (2) To generate an individual model using a gradient-based method.

The Monte Carlo method has been run in the calibration and validation period with 5000 parameter sets sampled from its feasible parameter space using the latin hypercube sampling technique (Audze 1977; McKay et al. 2000).

Once the model has been run, 3 samples of hydrographs have been extracted: (1) The 50 hydrographs with the highest Nash-Sutcliffe coefficient in the calibration period, (2) The 50 simulations with the highest Pearson coefficient in the calibration period and (3) The complete set of 5000 simulations is treated as a sample. These samples are used to define the criteria for the construction of the ANN-based ensembles as explained later in the Sect. 3.2.

In the case of the gradient-based method, it is used to identify a reference model which we assume to be the best single model to compare the results of different ANN-based ensemble configurations. The objective function optimised with the gradient-based method is a modification of the Nash-Sutcliffe coefficient (Nash and Sutcliffe 1970) focused on prioritising a correct simulation of peak flows which is detailed in Sect. 3.4.

Hydrological model

MHIA is a continuous lumped hydrological model. The model performs a balance of the volume of water in the soil taking into account the following processes: precipitation, infiltration, percolation, evapotranspiration and exfiltration. Based on these processes, the model evaluates the hydrograph at the outlet of the modelled basin. The input data required by the model are the time series of precipitation and temperature (spatially averaged over the whole basin) with any resolution in time. For this study, we have worked with a time resolution of 1 h.

The balance of water content in the soil is performed by representing the basin as a reservoir with a maximum storage capacity S_{max} (in *mm*) and a volume of water V (in *mm*) that varies over time. At each time step the aforementioned hydrological processes are calculated, and the water content of the soil is updated according to the following equation:

$$V^{i+1} = V^i + \left(f^i - p^i - e^i - q^i\right)\Delta t \tag{1}$$

where V is the volume of water stored in the soil expressed in mm, f is the infiltration rate in mm/h, p is the percolation rate in mm/h, e is the evapotranspiration rate in mm/h, q is the exfiltration rate of water from the soil to the surface water streams, also expressed in mm/h, and Δt is the calculation time step in h. The super index i in Eq. (1) refers to the time step of the computation. The parameters of the model are shown in table 2. The reader can access the code and the conceptualisation manual of the MHIA model in Farfán and Cea (2022) for further details.

Ensemble model based on artificial neural networks

An ensemble model consists of the combination of the outputs of different numerical simulations obtained with different parameter sets, initial conditions, or model structures (Najafi and Moradkhani 2016; Li and Sankarasubramanian 2012; Viney et al. 2009). In the case of ANN-based ensembles, the ANN inputs are the discharge time series obtained with the hydrological model using different parameter sets, and the ANN output are the observed discharge time series. Figure 2 shows the general topology of the ANN that have been used in this work to build ensemble models. It consists of an input layer, one hidden layer (formed by 3 neurons) and an output layer (Payal et al. 2013). In the applied ANN topology, a neuron is represented as the weighted sum of N external stimuli plus a constant bias followed by a non-linear activation function which controls the activity at the output of the neuron (Tkacz and Hu 1999). The training of the ANN starts with a random initialization of a set of weights, which is updated iteratively to find an optimal set that minimizes the error between the ANN output and the observed data (Alados et al. 2004). Notice that, even if in Fig. 2 the number of ANN inputs is 5, the number of individual models that has been used to build the ensemble was varied between 3 and 10, as it is mentioned in table 3.

Table 2 Calibration parameters for the MHIA model

Parameter	Symbol	Unity	Lower limit	Upper limit
Curve number	CN		20	70
Exponent of drainage	m_1		5	60
Parameter of infiltration and drainage	$K s$	mm/h	0.1	20
Coefficient lag-time relationship for SF	k_1		0.1	6.5
Parameter for scaling gamma functiont SF	n_1		1	10
Base flow bias correction	m_2		15	75
Parameter of exfiltration and BF	K_b	mm/h	0.1	20
Coefficient lag-time relationship for BF	k_2		1	6.5
Parameter for scaling gamma functiont BF	n_2		1	10
Parameter of potential evapotranspiration	b		0.4	2
Initial abstraction coefficient	α		0.0001	0.2
Decaying coefficient for P_{acum}	d		0.01	1
Correction coefficient for S	a		1	4

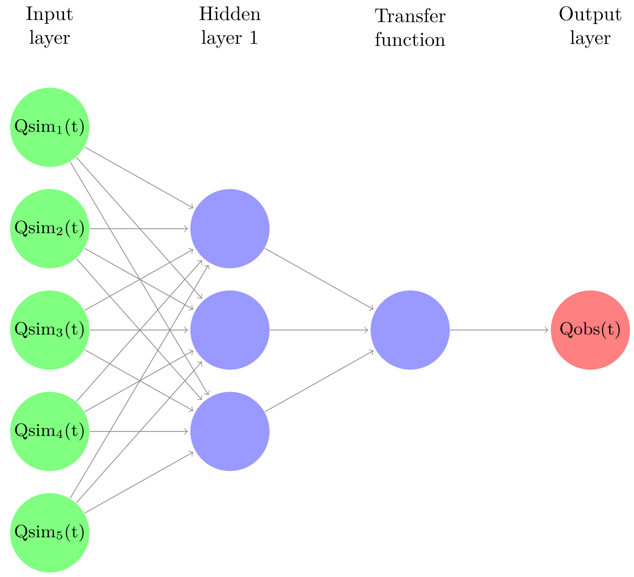

Fig. 2 Architecture of the applied ANN-based ensemble models. $Qsim_i(t)$ represents a value of the simulated discharge series at a time t. $Qobs(t)$ represents a value of the observed discharge series at the same time t

The application of ANN-based ensembles requires to take the following decisions for its configuration: (1) the number of neurons in the hidden layer, (2) the transfer function, (3) the number of hidden layers, (4) the number of inputs, (5) the training algorithm and (6) the criteria used to select the inputs. We use the proposed combinatorial approach to analyze the effect of some of these factors in the performance of the ANN-based ensemble, we have explored the application of different ANN configurations for building ensemble models.

For the purposes of the present study and to avoid overfitting oh the ANN, we have fixed the number of neurons in the hidden layer to 3 in all the configurations, the hyperbolic tangent was used as the transfer function (Farfán et al. 2020) and the number of hidden layers was fixed to 1, with the aim of avoiding the loss of effectiveness in the training algorithm that could be caused by a greater number of hidden layers (Raut and Dani 2020; Günther and Fritsch 2010). In addition, since in the present study the ANNs are trained in an iterative process, an ANN was trained 5 times in each iteration and the average of these was taken as the output to reduce the sensitivity of the ANN to the initial weights. On the other hand, the effect of the number of inputs (i.e individual models), the training algorithm and the criteria used to select the inputs was explored by means of the 24 different ANN configurations, which are summarised in Table 3.

The configurations tested include the 3 different samples of individual models that are used to build the ANN-based ensemble (NSE, Pearson or Random). Regarding the number of individual models used to build the ensemble, we have evaluated ANN configurations with 3, 5, 7 and 10 individual models as input.

The combinatorial optimisation approach is also used with two different ANN training algorithms: the Levenberg-Marquardt algorithm (LM) (Marquardt 1963) and the Bayesian regularisation algorithm (BR) (MacKay 1992). The Levenberg-Mardquart (LM) algorithm was designed for a fast convergence of back-propagation processes and to deal with moderate-sized problems (Kayri 2016; Jazayeri et al. 2016). A detailed description of the mathematical basis of the LM back-propagation algorithm is available in Jazayeri et al. (2016); Wilamowski and Yu (2010); Marquardt (1963). The Bayesian Regularisation (BR) proposed by MacKay (1992) is another well-known algorithm used in the training of ANN. It is a variation of the LM algorithm designed to minimise a convex combination of the mean-square-error and the mean-square-weights (Burden and Winkler 2008). This algorithm is capable of producing robust predictions provided that there are a sufficient number of observations.

Table 3 Nomenclature for the 24 evaluated ANN-based ensemble configurations

Configuration number	Configuration name	Sample of individual models	Number of individual models (Inputs)	Training algorithm
1	NSE_3_LM	50 simulations with highest NSE	3	LM
2	NSE_5_LM	50 simulations with highest NSE	5	LM
3	NSE_7_LM	50 simulations with highest NSE	7	LM
4	NSE_10_LM	50 simulations with highest NSE	10	LM
5	Pearson_3_LM	50 simulations with highest Pearson	3	LM
6	Pearson_5_LM	50 simulations with highest Pearson	5	LM
7	Pearson_7_LM	50 simulations with highest Pearson	7	LM
8	Pearson_10_LM	50 simulations with highest Pearson	10	LM
9	Complete_3_LM	Complete set of 5000 simulations	3	LM
10	Complete_5_LM	Complete set of 5000 simulations	5	LM
11	Complete_7_LM	Complete set of 5000 simulations	7	LM
12	Complete_10_LM	Complete set of 5000 simulations	10	LM
13	NSE_3_BR	50 simulations with highest NSE	3	BR
14	NSE_5_BR	50 simulations with highest NSE	5	BR
15	NSE_7_BR	50 simulations with highest NSE	7	BR
16	NSE_10_BR	50 simulations with highest NSE	10	BR
17	Pearson_3_BR	50 simulations with highest Pearson	3	BR
18	Pearson_5_BR	50 simulations with highest Pearson	5	BR
19	Pearson_7_BR	50 simulations with highest Pearson	7	BR
20	Pearson_10_BR	50 simulations with highest Pearson	10	BR
21	Complete_3_BR	Complete set of 5000 simulations	3	BR
22	Complete_5_BR	Complete set of 5000 simulations	5	BR
23	Complete_7_BR	Complete set of 5000 simulations	7	BR
24	Complete_10_BR	Complete set of 5000 simulations	10	BR

Additionally, it works by eliminating weights that make no relevant impact on the solution and is capable of escaping to local minima. The reader is referred to MacKay (1992); Buntine and Weigend (1991); Jazayeri et al. (2016) for further details.

To apply the LM algorithm, the calibration period was divided into 3 data sets: a training set, a verification set, and a testing set. These sets were defined as follows: (1) 50 % of the data was allotted to the training set to compute the adjustments of the ANN weights, (2) 25 % of the data was used in the verification set to compute the generalisation capabilities of the trained ANN, and (3) the remaining 25 % of the data was used as a testing set to measure the performance of the ANN. The division of the calibration data into these three sets was carried out by an interleaving process, assigning 2 data to the training set, 1 to the verification test set and 1 to the testing set in a sequential manner until the end of the time series, as it is shown schematically in Fig. 4a. The objective function used to train the ANN was the Mean Squared Error (MSE). Once the ensemble model was trained with the data from the calibration period, it was validated with the data from the validation period [Fig. 3a]. Numerous studies suggest using to use the cross-validation method because it provides the opportunity for the ANN on multiple train-test split. However, with the aim of robustness, as can be seen in Fig. 4, the interleaved method allows us to make a continuous sample of the entire regime of the discharge series that provides homogeneous information to the 3 data sets. Therefore, we have considered this method more appropriate for the specific case of ensemble models.

The application of the BR training algorithm consists on set the calibration period divided into the following 2 sets: (1) a training set with 75% of the data and (2) a testing set with the remaining 25% of the data. The verification subset is not required in the BR algorithm [Fig. 3b]. (Jazayeri et al. 2016; MacKay 1992; Buntine and Weigend 1991). As in the case of the LM algorithm, the data division is carried out by means of a interleaving procedure as it is shown in Fig. 4(b). The objective function is the MSE.

Random-restart hill-climbing

Once the topology of the ANN-based ensemble models is defined, in the next step we propose to apply the Random-Restart Hill-Climbing algorithm (RRHC) (Russell and Norvig 2010) to overcome the problem of identification

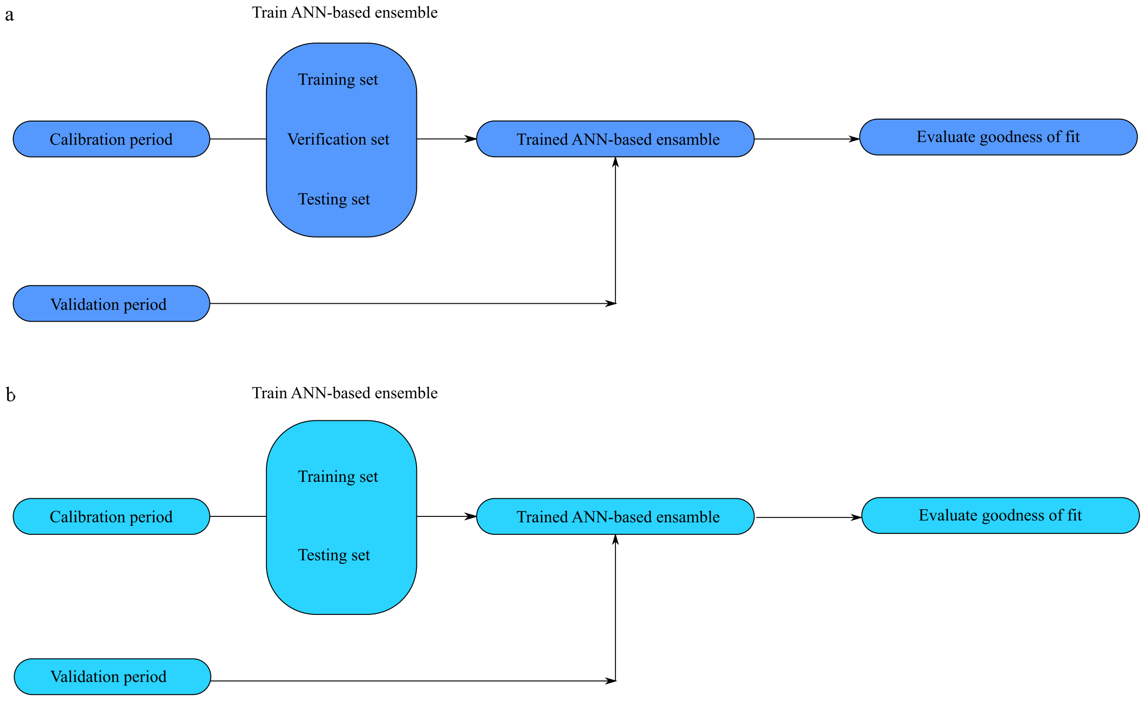

Fig. 3 Scheme of the training process of the ANN-based ensembles using the LM algorithm (**a**) and for the BR algorithm (**b**)

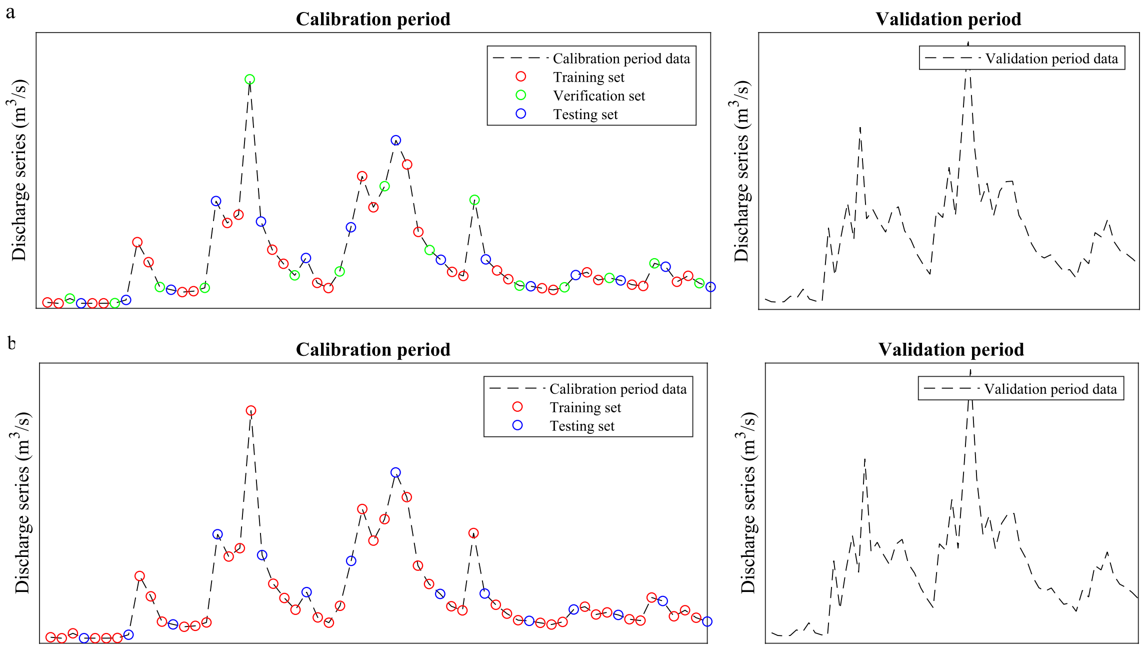

Fig. 4 Schematic representation of the data division using a interleaving process for the LM algorithm (**a**) and for the BR algorithm (**b**)

of optimal combinations of individual models to conform ensembles. The algorithm starts by generating a random state (i.e. a combination of individual models) and evaluating its goodness of fit. Then, an iterative process is started in which a new random state is generated and its goodness of fit is computed. If the new ensemble outperforms the previous one in terms of an array of objective functions, it is saved and the previous one is removed. Otherwise, the new ensemble is rejected. This procedure is repeated iteratively until a stop criterion is reached. In the present work, we have set the stop criterion as 5 consecutive iterations without improvement with respect to the previous ensemble (i.e. 5 consecutive rejections). To maximise the outcome of the algorithm, this iterative process is repeated a number of times defined by the user, starting each time from a different random ensemble. We have repeated the process using 50 different initial positions, in order to identify 50 different ensembles for each of the configurations detailed in Table 3.

The stop criterion in the previous algorithm needs to define an array of objective functions to identify if the new ensemble outperforms the previous one. For this purpose we have established a criterion based on 4 goodness-of-fit coefficients, Nash-Sutcliffe Efficiency (NSE), High Flows Weighted Nash-Sutcliffe Efficiency (HF-WNSE), Low Flows Weighted Nash-Sutcliffe Efficiency (LF-WNSE) and Kling & Gupta (KGE), all of which are described in Sect. 3.4. At every iteration, each new ensemble is trained in the calibration period and validated in the validation period. If 3 out of the 4 goodness-of-fit coefficients outperform those of the previous ensemble in both, calibration and validation, it is saved and the previous one is removed, otherwise it is rejected. This criterion has been set with the aim of saving only the best fitting ensembles and ensuring that the trained ensembles can generalise the modeling process and to avoid saving those that may have been overfitted.

The steps for the adaptation of the RRHC algorithm for the identification of ANN-based ensembles are summarised as a pseudo code as follows:

```
Pseudo-code for the adaptation of the RRHC Algorithm
for identification of ANN-based ensembles
  $i = 0$
  while $i <= X$ do
      Set an initial random ensemble:
      $current$ = initial random ensemble,
      $current \in Sample$
      $best = current$
      $istop = 0$
      while $istop < 5$ do
          Train ANN with current ensemble
          Calculate $fit(current)$
          if $fit(current) >= fit(best)$ then
              $best = current$
              $istop = 0$
          else
              $current$ = random ensemble,
              $current \in Sample$
              $istop = istop + 1$
          end if
      end while
      $i = i + 1$
  end while
```

where*X* denotes the number of ensembles we want to identify (in this study *X*= 50), hence the number of restarts of the RRHC algorithm, fit() denotes the goodness of fit measure, **current** is the combination of individual models used as input to the ANN, **best** is the best combination of models identified, and **Sample** denotes the group of individual models from which the ensembles are built.

The reader is referred to the work carried out by O'Neil and Burtscher (2015), Russell and Norvig (2010) and Kato et al. (2018) for further details about the Random-Restart Hill-Climbing algorithm.

Goodness-of-fit and performance measures

The results of the different ANN-based ensemble models have been evaluated using an array of goodness-of-fit coefficients, each one focused on a specific zone of the output hydrograph. These are the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe 1970), Kling & Gupta efficiency (KGE) (Gupta et al. 2009), and two modifications of the NSE focused on high and low discharges respectively. Previous studies have already used modifications of *NSE* applied to surface runoff and flood studies (Hundecha and Bárdossy 2004) in which an adequate evaluation of peak flows is of

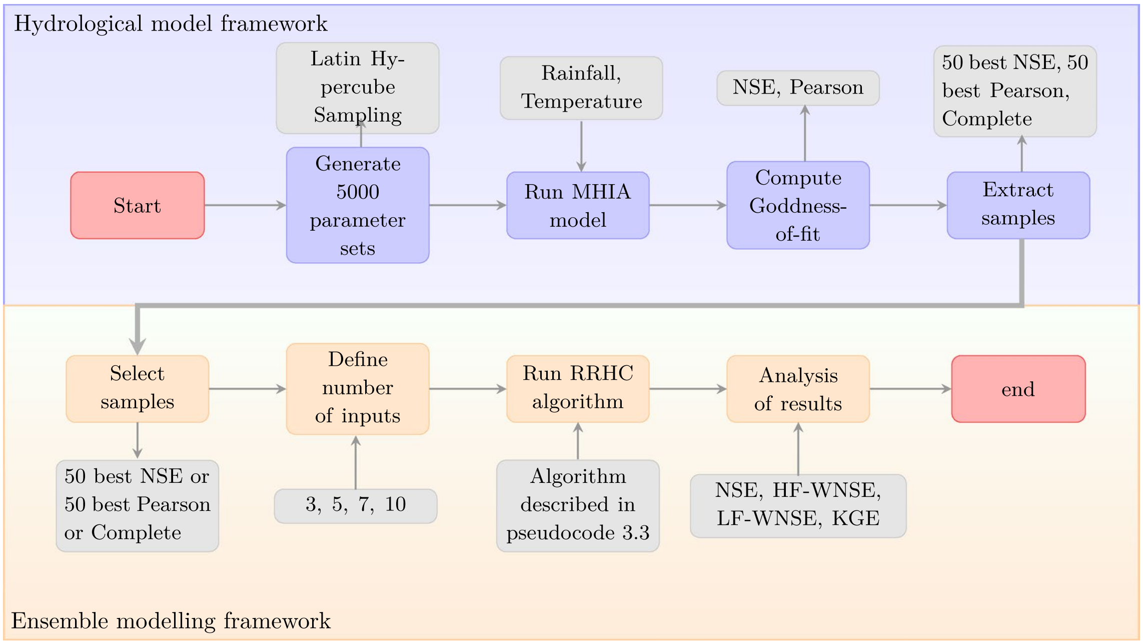
Hydrological model framework
Latin Hypercube Sampling
Rainfall, Temperature
NSE, Pearson
50 best NSE, 50 best Pearson, Complete
Start
Generate 5000 parameter sets
Run MHIA model
Compute Goddness-of-fit
Extract samples
Select samples
Define number of inputs
Run RRHC algorithm
Analysis of results
end
50 best NSE or 50 best Pearson or Complete
3, 5, 7, 10
Algorithm described in pseudocode 3.3
NSE, HF-WNSE, LF-WNSE, KGE
Ensemble modelling framework

Fig. 5 Overall description of the methodology

utmost importance. In this work, we have used the following modification of NSE:

$$WNSE = 1 - \frac{\sum w_i(Q_{obs,i} - Q_{sim,i})^2}{\sum w_i(Q_{obs,i} - \overline{Q_{obs}})^2}, \tag{2}$$

$$\overline{Q_{obs}} = \sum w_i Q_{obs,i}$$

where w_i is a vector of weights evaluated as:

$$w_i = \frac{Q^p_{obs,i}}{\sum Q^p_{obs,i}}$$

If the exponent p is positive, the high discharges have a higher effect on the value of WNSE, while if p is negative, the low discharges will dominate the value of WNSE. In the particular case that $p = 0$, the standard NSE is recovered.

In the present work, the WNSE was used with two different exponents, $p = 1$ and $p = -0.5$, to give a higher weight to high flows ($HF - WNSE$) and low flows ($LF - WNSE$) respectively. Thus, 4 different goodness-of-fit coefficients (NSE, $HF - NSE$, $LF - NSE$ and KGE) were used to analyse the performance of the proposed methodology.

Finally, the proposed methodology is summarised in the Fig. 5.

Results and discussion

In the present section, we analyse the results of the proposed technique by means of (1) The effectiveness of the combinatorial optimisation approach to identify reliable ANN-based ensembles, (2) the goodness of fit of the identified ensembles, compared to the individual models, and to the reference model optimised with a gradient-based method, (3) The computation cost of the proposed technique.

Effectiveness of the combinatorial optimisation approach

Table 4 shows the percentage of the 50 ensemble models obtained with each ANN configuration that outperformed the reference individual model. In the table, the NSE, $HF - WNSE$, $LF - WNSE$, KGE columns refer to the percentage of ensembles that exceeded the reference individual model for the specified coefficient in both, calibration and validation stages, while the Total column refers to the percentage of ensembles that outperformed the reference individual model in at least 3 of the 4 coefficients in both, calibration and validation stages.

The percentage of successful ensembles depends on the selection criterion for the application of RRHC algorithm (whether it is NSE, Pearson or Random). The most successful criteria are those based on the Pearson and NSE coefficients, while the Random selection criterion performs clearly worse. Within each of these criteria, the number of

Table 4 Percentages of ensembles capable of outperforming the best individual model in the different coefficients in both stages of calibration and validation for the Anllóns basin

Number	Configuration name	NS	HF-NSE	LF-NSE	KGE	Total
1	NSE_3_LM	0	0	0	96	0
2	NSE_5_LM	8	6	12	100	8
3	NSE_7_LM	30	36	32	100	30
4	NSE_10_LM	**76**	**78**	**64**	**100**	**76**
5	Pearson_3_LM	4	6	6	100	4
6	Pearson_5_LM	26	22	32	100	26
7	Pearson_7_LM	54	50	62	100	54
8	Pearson_10_LM	**88**	**86**	**94**	**100**	**90**
9	Complete_3_LM	0	0	0	60	0
10	Complete_5_LM	10	6	8	88	8
11	Complete_7_LM	18	22	20	84	18
12	Complete_10_LM	**32**	**46**	**20**	**90**	**34**
13	NSE_3_BR	2	0	0	98	2
14	NSE_5_BR	10	16	4	100	10
15	NSE_7_BR	34	30	32	100	34
16	NSE_10_BR	**74**	**80**	**56**	**100**	**74**
17	Pearson_3_BR	6	8	8	98	6
18	Pearson_5_BR	38	36	36	98	38
19	Pearson_7_BR	70	72	66	100	70
20	Pearson_10_BR	**78**	**80**	**70**	**100**	**78**
21	Complete_3_BR	0	0	2	56	0
22	Complete_5_BR	2	6	2	80	2
23	Complete_7_BR	16	18	14	80	16
24	Complete_10_BR	**28**	**56**	**20**	**82**	**28**

successful ensembles increases proportionally to the number of ANN inputs. The most efficient ANN configuration was Pearson-10-LM (sample based on the Pearson coefficient with 10 inputs and training the ANN with the LM algorithm) in which 90% of the ensembles outperformed the reference individual model in at least 3 of the 4 coefficients (Total). Regarding the most effective training algorithm, both LM and BR algorithms produced similar results.

It is also interesting to note that the KGE coefficient was exceeded in a higher percentage of ensembles than the rest of coefficients. This might be explained by the fact that the KGE coefficient is obtained from a decomposition of the NSE coefficient into 3 components that measure respectively the linear correlation, bias and variability, with an equal weighting for all of these 3 components. Thus, it is possible to have a high KGE value if any 2 of its components are close to their optimal value, even if the third one is slightly lower (Gupta et al. 2009).

Goodness of fit

To analyze the results obtained by applying the proposed algorithm, Fig. 6 shows the probability distribution corresponding to the goodness-of-fit coefficients of the individual models of one of the samples described in Sect. 3 and the ensembles of the configuration NSE_10_BR. This analysis has been carried out for all 24 configurations, although in this section we limit to showing one configuration for the sake of simplicity.

The ensembles have outperformed the individual models significantly in terms of frequency distribution in the 4 goodness-of-fit coefficients. The average of the sample with the 50 best individual models in terms of NSE is 0.65 with a standard deviation of 0.04, while the ensembles composed of combinations of 10 of these individual models reach an average of 0.86 with a standard deviation of 0.01, these results are approximately equal for both calibration and validation periods. The improvement in the predictive skills using the algorithm RRHC is most evident in the validation period, where even negative values of $HF - NSE$ were obtained for the individual models. In this period, the mean value of $HF - NSE$ for the individual models is 0.38 with a standard deviation of 0.21, while the $HF - NSE$ ensambles reach an average of 0.81 with a standard deviation of 0.16. This indicates that there is an important uncertainty in the individual models related to the model parameters. Thus, an individual model that presents a high $HF - NSE$ may present a low $LF - NSE$ and vice versa. Despite this, the applied algorithm has shown to be able to adequately identify combinations of individual models that produce ensembles with very goods

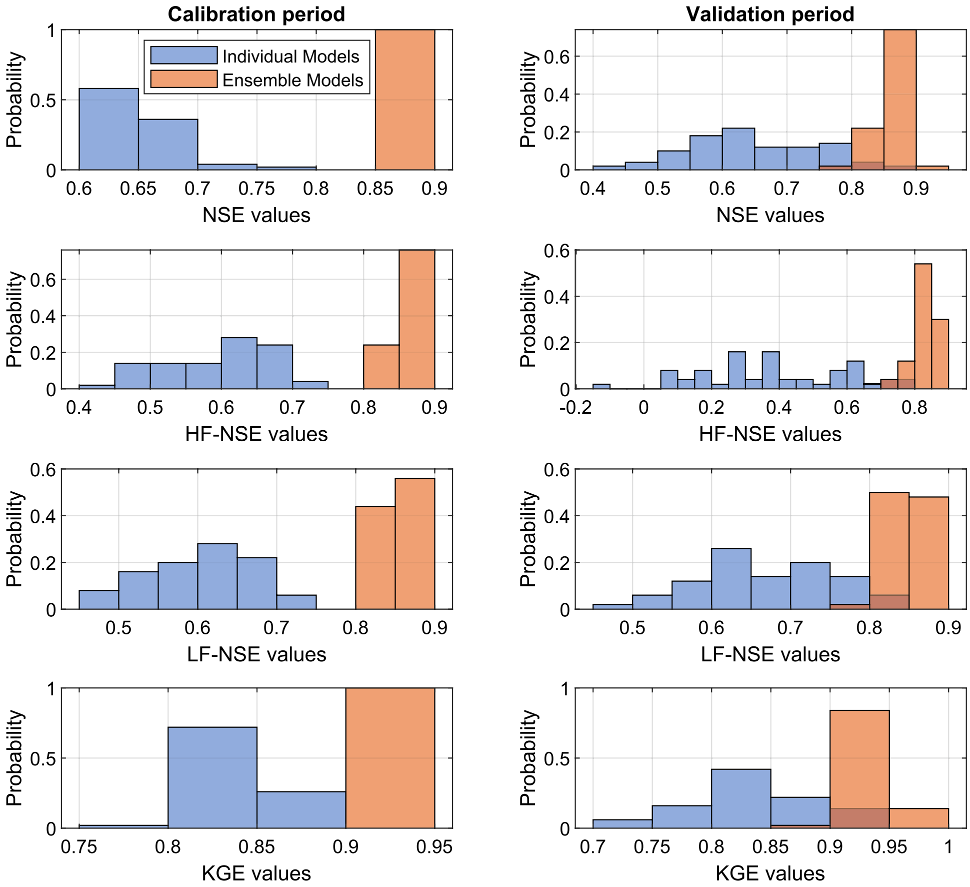

Fig. 6 Probability distribution of the NSE coefficient of the individual models and of the ensembles obtained from the RRHC algorithm for the Pearson_10_BR configuration

values for all coefficients, which implies a reduction in the parametric uncertainty that is dominant at fine time scales, such as that of the present study. In addition, the good results obtained in the validation period indicates that the ANN has not been overfitted when the RRHC algorithm was applied.

Table 5 shows the 4 goodness-of-fit coefficients corresponding to the best ensembles obtained for each configuration tested. The first row includes the goodness-of-fit coefficients obtained the reference hydrological model optimised with the gradient descent algorithm. To prioritise a correct generalisation of the ANN-based ensembles, the criterion taken into account to select the best ensemble among the different configurations is based on the coefficient results and has been complemented with a visual analysis of the simulated hydrographs (Ritter and Muñoz-Carpena 2013).

The most efficient configuration is Pearson_10_BR, which achieves values of the NSE, $LF-WNSE$ and $HF-WNSE$ near to 0.9, the KGE coefficient being even higher. The ANN-based ensemble has been successful in generalising the results, obtaining similar performance in both, calibration and validation periods. Regarding the ANN training algorithm, there are no relevant differences between the LM and BR algorithms, having both of them a similar performance.

Regarding the criterion used for the selection of individual models, when using the Pearson and Random criteria there is a direct relation between the number of ANN inputs and the achieved goodness of fit coefficients. The ensembles built with just 3 inputs presented the lowest goodness of fit coefficients, and most of them failed to outperform the reference individual model. In case of using the NSE to select the individual models, there is also a direct relation between number of inputs and performance when training the ANN with the LM algorithm. However, if the ANN is trained with the BR algorithm, the best ensemble was obtained with 7 inputs (NSE_7_BR configuration). This could be due to the fact that the individual models found in this configuration are a better combination than those found in the Nash_BR_10 configuration due to the randomness component of the RRHC algorithm. However, all of the other cases show that the ensembles with 10 inputs are more reliable, both in terms of success rate and goodness of fit of the best ensemble. Regarding the number of inputs, it is important to remark that in the development of the present study, the proposed methodology was also evaluated with more than 10 inputs. However, these configurations are not included in the results because marginal or null improvement in the results was obtained with respect to configurations with 7 or 10 inputs. This is explained by taking into account

Table 5 Best goodness-of-fit ensembles for each configuration in the Anllóns basin

Number	Configuration name	Calibration				Validation			
		NS	HF-NSE	LF-NSE	KGE	NS	HF-NSE	LF-NSE	KGE
	Best individual model	0.86	0.84	0.82	0.87	0.86	0.79	0.84	0.84
1	NSE_3_LM	0.84	0.82	0.79	0.91	0.87	0.86	0.82	0.94
2	NSE_5_LM	0.86	0.84	0.83	0.93	0.89	0.85	0.87	0.95
3	NSE_7_LM	0.86	0.84	0.83	0.93	0.90	0.87	0.89	0.96
4	NSE_10_LM	0.89	0.87	0.87	0.94	0.90	0.88	0.89	0.96
5	Pearson_3_LM	0.86	0.85	0.83	0.93	0.89	0.86	0.86	0.96
6	Pearson_5_LM	0.89	0.87	0.88	0.94	0.88	0.86	0.86	0.95
7	Pearson_7_LM	0.88	0.86	0.87	0.94	0.89	0.86	0.88	0.95
8	Pearson_10_LM	0.90	0.88	0.88	0.95	0.90	0.89	0.88	0.96
9	Complete_3_LM	0.84	0.83	0.80	0.92	0.87	0.84	0.86	0.94
10	Complete_5_LM	0.86	0.84	0.85	0.93	0.87	0.84	0.86	0.95
11	Complete_7_LM	0.88	0.88	0.85	0.94	0.89	0.86	0.87	0.95
12	Complete_10_LM	0.89	0.88	0.87	0.94	0.89	0.87	0.86	0.95
13	NSE_3_BR	0.86	0.83	0.83	0.93	0.87	0.86	0.84	0.94
14	NSE_5_BR	0.86	0.84	0.82	0.93	0.89	0.87	0.88	0.95
15	NSE_7_BR	0.88	0.86	0.86	0.94	0.91	0.88	0.90	0.96
16	NSE_10_BR	0.88	0.87	0.85	0.94	0.90	0.88	0.89	0.95
17	Pearson_3_BR	0.86	0.85	0.83	0.93	0.88	0.87	0.86	0.95
18	Pearson_5_BR	0.88	0.87	0.85	0.94	0.88	0.85	0.86	0.95
19	Pearson_7_BR	0.90	0.87	0.89	0.95	0.89	0.87	0.87	0.96
20	Pearson_10_BR	0.91	0.89	0.89	0.95	0.91	0.88	0.89	0.96
21	Complete_3_BR	0.84	0.83	0.80	0.92	0.87	0.82	0.86	0.95
22	Complete_5_BR	0.87	0.84	0.85	0.93	0.88	0.85	0.88	0.95
23	Complete_7_BR	0.88	0.86	0.85	0.94	0.89	0.87	0.85	0.95
24	Complete_10_BR	0.89	0.88	0.87	0.94	0.88	0.85	0.86	0.95

that, after exceeding a certain number of inputs to the ANN, the incorporated information starts to be redundant and no enhancement of the simulations was observed.

Figure 7 shows in grey the 50 ensembles built using the Pearson_10_BR configuration, in red the best model ensemble identified, in green the result of the optimized hydrological model and in black the observed discharge series. The best ensemble model has a better fit to the observed series than the best individual model both, in calibration and validation. The more accurate fit is specially remarkable regarding the peak discharges. In addition, the best ensemble model has a lower dispersion with respect to the 1:1 line, which indicates that it is able to enhance the simulations in terms of correlation and variability.

Computational cost

In this section, we analyze the efficiency of the proposed algorithms in terms of CPU time. The computational time was measured each time the RRHC algorithm was run with each ANN configuration (Table 3). From the CPU times reported in Table 6, 2 factors can be identified as those that mainly control the computational time of the proposed method: (1) the number of inputs that conform each ensemble model and (2) the sample from which the individual models are chosen. These 2 factors influence the computational times differently depending on the training algorithm.

The computation times for the NSE_LM_3 configuration are considerably longer than the configurations with 3, 5 and 7 inputs. This may be due to the configuration of the stop criteria, which are reset to 0 each time the algorithm finds an ensemble better than the previous one and to the fact that the number of 3-element combinations implies a larger number of ensembles to evaluate that can outperform the previous one, regardless of whether or not this one is better than the best individual model in terms of the goodness-of-fit coefficients.

This behaviour is also observable in the Complete_LM_3 and Complete_LM_5 configurations. This indicates that in cases where a sufficient number of inputs are not selected, not only a few ensembles that outperform the best individual model are obtained, but the computational cost may increase, making the proposed technique inefficient. On the other hand, in the case of the sample based on the coefficient of Pearson, the times maintain similar values for the

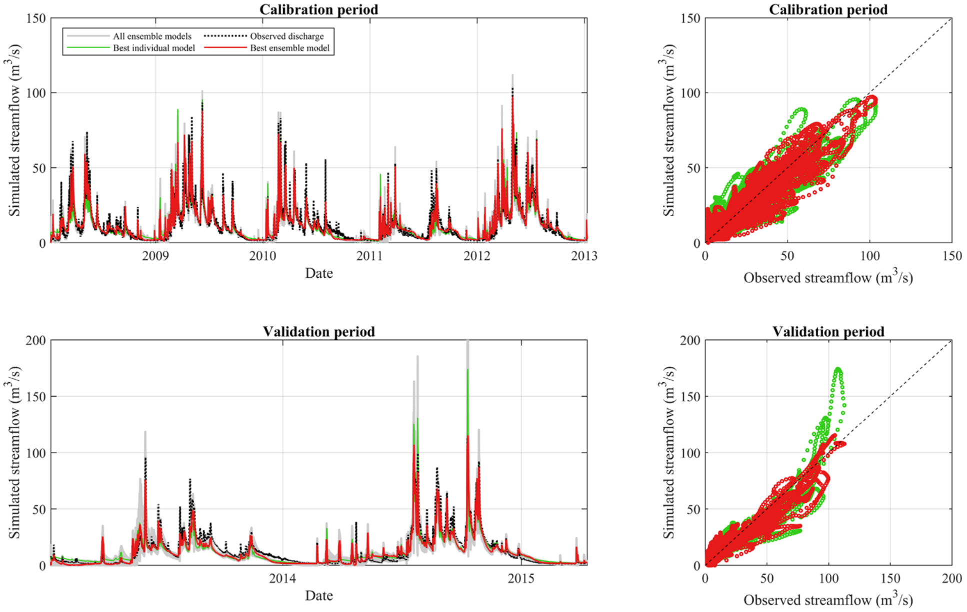

Fig. 7 Results for the case of Anllóns river for the configuration Pearson_10_BR

Table 6 Comparison of run time of the different configurations of the RRHC algorithm for the different configurations of ANN-based ensembles

Inputs	NSE			Pearson			Complete set		
	LM (h)	BR (h)	Ratio	LM (h)	BR (h)	Ratio	LM (h)	BR (h)	Ratio
3	3.20	2.99	0.93	3.50	3.82	1.09	4.55	5.83	1.28
5	2.60	4.09	1.57	3.43	4.63	1.35	4.78	6.05	1.26
7	2.70	4.16	1.54	3.56	5.07	1.42	3.86	6.68	1.73
10	2.46	4.74	1.93	3.56	6.21	1.74	3.92	7.09	1.81

LM algorithm, varying between 3.43 to 3.56 h to identify 50 ensembles.

The shortest times were for the sample based on the NSE coefficient, followed by the sample based on the Pearson coefficient, showing that the NSE sample was a more computationally efficient selection criterion.

Conclusions

In the present study, we have proposed using a combinatorial optimisation approach to identify combinations of individual hydrological models suitable for the construction of hydrological ensemble models. The proposed approach consists of searching for a small combination of individual models that are selected from a large number of individual models when evaluating all of them one by one is not a feasible option. For this purpose, we have adapted the generic optimisation RRHC algorithm to the construction of ANN-based hydrological ensemble models. The proposed methodology has been used to evaluate 24 configurations of ANN-based ensembles including different number of ANN inputs, three criteria to select the individual models and two ANN training algorithms.

The use of ANN-based ensembles has been developed since ANNs involve a higher difficulty for the implementation of the RRHC algorithm, given that ANN have more factors to take into account for application, such as the training algorithm and the possibility of overfitting. However, the proposed methodology can be applied with other techniques to construct the ensembles, such as model averaging, linear regression among others, and their application and comparison can be the objective of future studies.

Furthermore, in the present work we have focused on single-model ensembles, mainly oriented to the reduction of parameter-related uncertainty that is usually dominant at finer time scales (Li et al. 2016) as is the case of the present work. In this sense, it has been observed that the identified ANN ensembles have outperformed the individual model samples in terms of the probability distribution of the goodness-of-fit coefficients, which indicates a significant reduction in parametric uncertainty after its application. Therefore, the application of the proposed technique can constitute an interesting tool for the construction of multimodel ensembles obtained from different models focused on the reduction of the structural uncertainty which is dominant at coarser time scales which should be the purpose of future works.

The proposed method is able to consistently identify ensembles with better goodness of fit than the reference individual model in the case study. This has been made possible because the proposed algorithm allows to search for an object inside a finite collection of a large number of objects, when evaluating all of them one by one is not a feasible option which enhance the application of ensemble models for the modelling of hydrological systems.

Based on our results, the effectiveness of the proposed technique is conditioned by two important factors: (1) the number of individual models used to build the ensemble (i.e. the number of ANN inputs) and (2) the criterion used to select the individual models that are used to build the ensemble. The best results in terms of goodness of fit were obtained with the ensembles formed with 10 individual models, selected by a sampling strategy based on the 50 individual models with the highest Pearson coefficient although the difference with the sample based on NSE are minimal. On the other hand, the worst results were obtained from the ensembles built with just 3 or 5 individual models, selected by a random sampling out of 5000 individual models.

Regarding the ANN training algorithms, the BR method presented slightly higher results in terms of goodness of fit and percentage of models that outperformed the reference individual model. However, this was possible at a much higher computational cost (between 1.75 and 2.5 times higher than that of the LM algorithm). In fact, the CPU time needed by the BR algorithm is quite sensitive to the number of individual models that form the ensemble, while the LM algorithm maintained similar execution times independently of the number of individual models. In this sense, the selection of the training algorithm represented a greater influence on the running times than on the final results of the ensembles.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This study is financed by the Galician government (Xunta de Galicia) as part of its pre-doctoral fellowship program (Axudas de apoio á etapa predoutoral 2019) Register No ED481A-2019/014.

Data availability The meteorological data was obtained from the agency MeteoGalicia at https://www.meteogalicia.gal/. The discharge data has been provided by the regional water administration Augas de Galicia at https://augasdegalicia.xunta.gal/. The code for the developed methodology is available at https://github.com/jfarfand/Ensemble-Modelling.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Juan F. Farfán-Durán : Conceptualisation, Methodology, Code, Writing-Original draft preparation. Luis Cea: Visualisation, Supervision, Writing-Reviewing and Editing.

References

Al-Betar MA, Awadallah MA, Bolaji AL, Alijla BO (2017) β-hill climbing algorithm for sudoku game. In 2017 Palestinian International Conference on Information and Communication Technology (PICICT), pages 84–88. IEEE

Alados I, Mellado JA, Ramos F, Alados-Arboledas L (2004) Estimating uv erythemal irradiance by means of neural networks. Photochem Photobiol 80(2):351–358

Alsukni E, Arabeyyat OS, Awadallah MA, Alsamarraie L, Abu-Doush I, Al-Betar MA (2019) Multiple-reservoir scheduling using β-hill climbing algorithm. J Intell Syst 28(4):559–570

Andraos C, Najem W (2020) Multi-model approach for reducing uncertainties in rainfall-runoff models. Advances in hydroinformatics. Springer, pp 545–557

Arsenault R, Gatien P, Renaud B, Brissette F, Martel J-L (2015) A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J Hydrol 529:754–767

Audze P (1977) New approach to planning out of experiments. Probl Dyn Strengths 35:104–107

Bermúdez M, Farfán J, Willems P, Cea L (2021) Assessing the effects of climate change on compound flooding in coastal river areas. Water Resour Res 57(10):e202002020WR029321

Buntine WL, Weigend AS (1991) Bayesian back-propagation. Complex Syst 5(6):603–643

Burden F, Winkler D (2008) Bayesian regularization of neural networks. Artificial neural networks. Springer, pp 23–42

Cabalar Fuentes M (2005) Los temporales de lluvia y viento en galicia: propuesta de clasificación y análisis de tendencias (1961–2001). Investig Geogr 36:103–118

Ceylan H (2006) Developing combined genetic algorithm-hill-climbing optimization method for area traffic control. J Transp Eng 132(8):663–671

DeChant CM, Moradkhani H (2014) Toward a reliable prediction of seasonal forecast uncertainty: addressing model and initial condition uncertainty with ensemble data assimilation and sequential Bayesian combination. J Hydrol 519:2967–2977

Dong L, Xiong L, Yu K-x (2013) Uncertainty analysis of multiple hydrologic models using the Bayesian model averaging method. J Appl Math. https://doi.org/10.1155/2013/346045.2013

Duan Q, Ajami NK, Gao X, Sorooshian S (2007) Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv Water Resour 30(5):1371–1386

Farfán JF, Cea L (2022) Mhia model ("modelo hidrológico agregado"). https://doi.org/10.4211/hs.d98161b9f3fb4d03a4358f6c8b5f2c04. Accessed 1 June 2022

Farfán JF, Palacios K, Ulloa J, Avilés A (2020) A hybrid neural network-based technique to improve the flow forecasting of physical and data-driven models: methodology and case studies in Andean watersheds. J Hydrol Reg Stud 27:100652

Günther F, Fritsch S (2010) Neuralnet: training of neural networks. R J 2(1):30–38

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol 377(1–2):80–91

Hundecha Y, Bárdossy A (2004) Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. J Hydrol 292(1–4):281–295

Jafarzadeh A, Khashei-Siuki A, Pourreza-Bilondi M (2022) Performance assessment of model averaging techniques to reduce structural uncertainty of groundwater modeling. Water Resour Manag 36(1):353–377

Jazayeri K, Jazayeri M, Uysal S (2016) Comparative analysis of levenberg-marquardt and Bayesian regularization backpropagation algorithms in photovoltaic power estimation using artificial neural network. Industrial Conference on Data Mining. Springer, pp 80–95

Kato ERR, de Aguiar Aranha GD, Tsunaki RH (2018) A new approach to solve the flexible job shop problem based on a hybrid particle swarm optimization and random-restart hill climbing. Comput Ind Eng 125:178–189

Kayri M (2016) Predictive abilities of Bayesian regularization and levenberg-marquardt algorithms in artificial neural networks: a comparative empirical study on social data. Math Comput Appl 21(2):20

Kumar A, Singh R, Jena PP, Chatterjee C, Mishra A (2015) Identification of the best multi-model combination for simulating river discharge. J Hydrol 525:313–325

Li W, Sankarasubramanian A (2012) Reducing hydrologic model uncertainty in monthly streamflow predictions using multimodel combination. Water Resour Res. https://doi.org/10.1029/2011WR011380

Li W, Sankarasubramanian A, Ranjithan R, Sinha T (2016) Role of multimodel combination and data assimilation in improving streamflow prediction over multiple time scales. Stoch Environ Res Risk Assess 30(8):2255–2269

Li Z, Yu J, Xu X, Sun W, Pang B, Yue J (2018) Multi-model ensemble hydrological simulation using a bp neural network for the upper Yalongjiang river basin, china. Proc Int Assoc Hydrol Sci 379:335

Li D, Marshall L, Liang Z, Sharma A (2022) Hydrologic multi-model ensemble predictions using variational Bayesian deep learning. J Hydrol 604:127221

Lim A, Rodrigues B, Zhang X (2006) A simulated annealing and hill-climbing algorithm for the traveling tournament problem. Eur J Oper Res 174(3):1459–1478

Liu J, Yuan X, Zeng J, Jiao Y, Li Y, Zhong L, Yao L (2022) Ensemble streamflow forecasting over a cascade reservoir catchment with integrated hydrometeorological modeling and machine learning. Hydrol Earth Syst Sci 26(2):265–278

Londhe SN, Shah S (2019) A novel approach for knowledge extraction from artificial neural networks. ISH J Hydraul Eng 25(3):269–281

MacKay DJ (1992) Bayesian methods for adaptive models. PhD thesis, California Institute of Technology. https://doi.org/10.7907/H3A1-WM07. https://resolver.caltech.edu/CaltechETD:etd-01042007-131447. Accessed 5 May 2022

Madadgar S, Moradkhani H (2014) Improved bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resour Res 50(12):9586–9603

Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 11(2):431–441

McKay MD, Beckman RJ, Conover WJ (2000) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 42(1):55–61

Najafi MR, Moradkhani H (2016) Ensemble combination of seasonal streamflow forecasts. J Hydrol Eng 21(1):04015043

Najafi MR, Moradkhani H, Piechota TC (2012) Ensemble streamflow prediction: climate signal weighting methods vs. climate forecast system reanalysis. J Hydrol 442:105–116

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part i-a discussion of principles. J Hydrol 10(3):282–290

Nourani V, Kheiri A, Behfar N (2022) Multi-station artificial intelligence based ensemble modeling of suspended sediment load. Water Supply 22(1):707–733

O'Neil MA, Burtscher M (2015) Rethinking the parallelization of random-restart hill climbing: a case study in optimizing a 2-opt tsp solver for gpu execution. In Proceedings of the 8th workshop on general purpose processing using GPUs, pages 99–108

Payal A, Rai C, Reddy B (2013) Comparative analysis of bayesian regularization and levenberg-marquardt training algorithm for localization in wireless sensor network. In 2013 15th International Conference on Advanced Communications Technology (ICACT), pages 191–194. IEEE

Phukoetphim P, Shamseldin AY, Melville BW (2014) Knowledge extraction from artificial neural networks for rainfall-runoff model combination systems. J Hydrol Eng 19(7):1422–1429

Qi W, Chen J, Xu C, Wan Y (2021) Finding the optimal multimodel averaging method for global hydrological simulations. Remote Sens 13(13):2574

Raut P, Dani A (2020) Correlation between number of hidden layers and accuracy of artificial neural network. Advanced computing technologies and applications. Springer, pp 513–521

Ritter A, Muñoz-Carpena R (2013) Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol 480:33–45

Russell SJ, Norvig P (2010) Artificial intelligence-a modern approach. Third International Edition, Pearson

Schäfer Rodrigues Silva A, Weber TK, Gayler S, Guthke A, Höge M, Nowak W, Streck T (2022) Diagnosing similarities in probabilistic multi-model ensembles: an application to soil-plant-growth-modeling. Model Earth Syst Environ. https://doi.org/10.1007/s40808-022-01427-1

Schrijver A (2003) Combinatorial optimization: polyhedra and efficiency, vol 24. Springer Science & Business Media

Shamseldin AY, O'Connor KM, Nasr AE, (2007) A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models. Hydrol Sci J 52(5):896–916

Shen Y, Wang S, Zhang B, Zhu J (2022) Development of a stochastic hydrological modeling system for improving ensemble streamflow prediction. J Hydrol 608:127683

Tkacz G, Hu S (1999) Forecasting GDP growth using artificial neural networks. Technical report, Bank of Canada

Tyralis H, Papacharalampous G, Langousis A (2021) Super ensemble learning for daily streamflow forecasting: large-scale demonstration and comparison with multiple machine learning algorithms. Neural Comput Appl 33(8):3053–3068

Velázquez J, Anctil F, Ramos M, Perrin C (2011) Can a multi-model approach improve hydrological ensemble forecasting? a study on 29 French catchments using 16 hydrological model structures. Adv Geosci 29:33

Viney NR, Bormann H, Breuer L, Bronstert A, Croke BF, Frede H, Gräff T, Hubrechts L, Huisman JA, Jakeman AJ et al (2009) Assessing the impact of land use change on hydrology by ensemble modelling (luchem) ii: Ensemble combinations and predictions. Adv Water Resour 32(2):147–158

Wang A, Bohn TJ, Mahanama SP, Koster RD, Lettenmaier DP (2009) Multimodel ensemble reconstruction of drought over the continental United States. J Clim 22(10):2694–2712

Wang J, Shi P, Jiang P, Hu J, Qu S, Chen X, Chen Y, Dai Y, Xiao Z (2017) Application of BP neural network algorithm in traditional hydrological model for flood forecasting. Water 9(1):48

Wilamowski BM, Yu H (2010) Improved computation for levenberg-marquardt training. IEEE Trans Neural Netw 21(6):930–937

Yaseen ZM, El-Shafie A, Jaafar O, Afan HA, Sayl KN (2015) Artificial intelligence based models for stream-flow forecasting: 2000–2015. J Hydrol 530:829–844

Zhang L, Yang X (2018) Applying a multi-model ensemble method for long-term runoff prediction under climate change scenarios for the yellow river basin, China. Water 10(3):301

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.