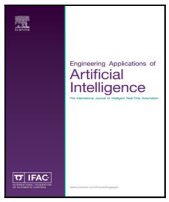
Engineering Applications of Artificial Intelligence 95 (2020) 103910

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Multi-objective ensembles of echo state networks and extreme learning machines for streamflow series forecasting

Victor Henrique Alves Ribeiro [a,*], Gilberto Reynoso-Meza [a], Hugo Valadares Siqueira [b]

[a] Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS), Pontifícia Universidade Católica do Paraná (PUCPR), Rua Imaculada Conceição, 1155, Zip code 80215-901, Curitiba, PR, Brazil

[b] Programa de Pós-Graduação em Ciência da Computação (PPGCC), Programa de Pós-Graduação em Engenharia de Produção (PPGEP), Universidade Tecnológica Federal do Paraná (UTFPR), Rua Doutor Washington Subtil Chueire, 330, Zip code 84017-220, Ponta Grossa, PR, Brazil

ARTICLE INFO

Keywords:
Streamflow
Time series forecasting
Extreme learning machine
Echo state network
Ensemble learning
Multi-objective optimization

ABSTRACT

Streamflow series forecasting composes a fundamental step in planning electric energy production for hydroelectric plants. In Brazil, such plants produce almost 70% of the total energy. Therefore, it is of great importance to improve the quality of streamflow series forecasting by investigating state-of-the-art time series forecasting algorithms. To this end, this work proposes the development of ensembles of unorganized machines, namely Extreme Learning Machines (ELMs) and Echo State Networks (ESNs). Two primary contributions are proposed: (1) a new training logic for ESNs that enables the application of bootstrap aggregation (bagging); and (2) the employment of multi-objective optimization to select and adjust the weights of the ensemble's base models, taking into account the trade-off between bias and variance. Experiments are conducted on streamflow series data from five real-world Brazilian hydroelectric plants, namely those in *Sobradinho*, *Serra da Mesa*, *Jiraú*, *Furnas* and *Água Vermelha*. The statistical results for four different prediction horizons (1, 3, 6, and 12 months ahead) indicate that the ensembles of unorganized machines achieve better results than autoregressive (AR) models in terms of the Nash–Sutcliffe model efficiency coefficient (NSE), root mean squared error (RMSE), coefficient of determination (R^2), and RMSE-observations standard deviation ratio (RSR). In such results, the ensembles with ESNs and the multi-objective optimization design procedure achieve the best scores.

1. Introduction

Many steps are involved in the operation planning of a power generation system. The process aims to use the sources efficiently, considering the cost reduction of the whole operation (Malfatti et al., 2018). In Brazil, around 66.6% of the entire electric power generation is performed by hydroelectric plants (EPE - Energy Research Company, 2019).

There is an evident necessity for providing a suitable use of water as a power source, given its significant impact on the cost of energy production (Siqueira et al., 2014; Francelin et al., 1996; Sacchi et al., 2007). In this context, an important step is to perform accurate streamflow predictions related to the rivers in which the hydroelectric plants are located (Siqueira et al., 2018). One of the main examples of this kind is to accurately predict the monthly seasonal streamflow (Siqueira et al., 2018). The mentioned series presents a seasonal component because their water volume follows the rainfall behavior over the year (Sacchi et al., 2007).

Many studies have been developed to tackle this task. The linear Box & Jenkins methodology (Box et al., 2015) is still widely applied in the Brazilian Electric Sector, as discussed by Francelin et al. (1996), Siqueira et al. (2012), and Siqueira et al. (2014). The highlight is the autoregressive (AR) model due to its simplicity in the implementation and optimization of the free parameters. The autoregressive moving average (ARMA) model is also largely employed (Francelin et al., 1996). Notwithstanding, some investigations have presented the benefits of using artificial neural networks (ANNs) in time series forecasting (de Matto Neto et al., 2014; Firmino et al., 2014). ANNs are nonlinear methodologies with high mapping and generalization capability inspired on the behavior of the nervous system of superior organisms (Haykin et al., 2009).

Many works have been developed to solve the aforementioned prediction task, and the results show their advantages in comparison to linear models (Siqueira et al., 2012, 2018; Siqueira and Luna, 2019). In such studies, the authors show that the classic multilayer perceptron (MLP) and the radial basis function (RBF) network are good alternatives to replace linear autoregressive models. However, unorganized machines - ELMs and ESNs - are relatively novel methods that have also shown good comparative performances for the streamflow forecasting task, as reported by Siqueira et al. (2012, 2014, 2018).

* Corresponding author.
E-mail addresses: victor.henrique@pucpr.edu.br (V.H.A. Ribeiro), g.reynosomeza@pucpr.br (G. Reynoso-Meza), hugosiqueira@utfpr.edu.br (H.V. Siqueira).

https://doi.org/10.1016/j.engappai.2020.103910
Received 21 September 2019; Received in revised form 11 July 2020; Accepted 18 August 2020
Available online 25 August 2020

In addition, Lima et al. (2016) employed an online sequential ELM to forecast daily streamflow at two watersheds in Canada. Such a model outperforms online sequential multiple linear regression. ELM also performed faster than ANNs, support vector regression (SVR), and random forests (RF) in several different forecasting scenarios (Lima et al., 2015). Finally, Yaseen et al. (2019b) achieved remarkable results for river flow forecasting using an enhanced ELM.

In recent years, the use of ensemble learning to deal with time series prediction has gained increasing attention (Sharkey, 1999; de Matto Neto et al., 2014; Firmino et al., 2014; Ribeiro and dos Santos Coelho, 2020). Such a technique combines the output of multiple different predictors, which leads to an improvement in predictive performance. The most common approach is the use of non-trainable methods such as the average or median of the outputs, as the ensemble combiner (Rokach, 2010). Other works employ trainable approaches as the combiners, such as the stacking ensemble (Ribeiro and dos Santos Coelho, 2020). Recent works have shown the benefits of using ensembles in similar tasks (Kasiviswanathan and Sudheer, 2013; Fan et al., 2016; Thober et al., 2018).

Additional research has also employed hybrid models combining global optimization and machine learning algorithms. Yaseen et al. (2019a) combined different bio-inspired optimization algorithms to tune membership functions of an adaptive neuro-fuzzy inference system (ANFIS), achieving reliable results in forecasting the streamflow of the Pahang River. Al-Sudani et al. (2019) developed adaptive multivariate spline regression integrated with differential evolution, which provided reliable streamflow forecasting results for semi-arid patterns. Moreover, the combination of ANFIS with the firefly algorithm presented high predictive accuracy for the monthly streamflow forecasting of the Malaysian Pahang River (Yaseen et al., 2017). Finally, Ghorbani et al. (2018) forecast monthly streamflows at three stations in the Zarrineh River using different ensemble strategies for SVR trained with and without the firefly algorithm.

1.1. Motivation

Nevertheless, there is still room for improvement in the development and application of ensemble models for river flow series forecasting. Many important studies have been developed recently in similar subjects considering databases from different countries. However, most of them are focused on the use of just one model to perform the predictions.

Zhu et al. (2016) used ELMs with the discrete wavelet transform (DWT) and empirical mode decomposition (EMD) to predict monthly streamflows from the Jinsha River in China. The work by Yaseen et al. (2017) addresses the database from the Pahang River, the major system in Malaysia. The authors introduced a hybrid proposal between the ANFIS and the bio-inspired firefly algorithm.

Siqueira et al. (2014) developed an investigation considering the application of unorganized machines for the monthly seasonal streamflow series forecasting of four Brazilian plants. Siqueira et al. (2018) extended that study utilizing improved versions of unorganized machines and variable selection methods. In both cases, the authors showed the advantages of using these proposals, which surpassed traditional linear methods and neural networks.

Three databases from Quebec, Canada, were addressed by Arsenault and Côté (2018) to predict hydrological streamflows from hydroelectric stations using ensemble streamflow prediction (ESP). This is a dynamic method that receives, as input, climate data such as future weather scenarios. The authors addressed exogenous variables, as in work by Hailegeorgis and Alfredsen (2017), which uses, for example, precipitation, evapotranspiration, and infiltration as inputs. The streamflows are from Norway, and the predictor is the precipitation-runoff model developed for these scenarios.

Hernandez-Ambato et al. (2017) used a MLP to predict streamflows from Ecuador. Stojković et al. (2017) investigated the annual trend of river flows in Serbia employing spectral analysis. Dilini et al. (2016) presented a study using the streamflow series from the Mahaweli hydropower plant located in Sri Lanka. They applied a hybrid between a dynamic neural network and a non-linear autoregressive with exogenous input (NARX-ANN). As one may observe, the use of single models stands out in the streamflow forecasting context. This is a strong indication that the development of a new tool provided by parallel processing such as ensembles should be encouraged due to the importance of the theme.

Regarding such an approach, diversity is one of the most critical characteristics of ensemble models (Rokach, 2010). Hence, the current ensemble learning methodology in time series forecasting focus on building a pool of heterogeneous base models (Ribeiro and dos Santos Coelho, 2020). Few studies rely on the creation of homogeneous base model pools (Rigamonti et al., 2018). However, such models need sequential input data, which hamper the generation of diverse models by input data manipulation (Rokach, 2010). To solve this problem, this work employs the bootstrap aggregating (bagging) method to create pools of ESN and ELM base models combined using the average rule. To enable the bagging procedure without the need for sequential data for the ESN, a novel training logic employing sub-time series is proposed.

Moreover, a multi-objective optimization design (MOOD) procedure is proposed to improve further the performance of the forecasting models, which have shown excellent results in both time series classification and forecasting problems (Ribeiro and Reynoso-Meza, 2019; Ribeiro et al., 2019; Ribeiro and Reynoso-Meza, 2020). The proposed procedure selects and weights each base model in a weighted average ensemble. By using evolutionary multi-objective optimization (MOO), the error bias and variance may be optimized simultaneously, enabling the generation of ensembles with a more preferable trade-off between the two conflicting objectives.

Finally, experiments are conducted with unorganized machines as single models, bagged ensembles, and optimized ensembles. A comparison is performed against autoregressive models, namely autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA). The experiments are conducted using data from the streamflow series of five real-world Brazilian hydroelectric plants located in *Sobradinho*, *Serra da Mesa*, *Jiraú*, *Furnas* and *Água Vermelha*. The results are related to multi-step forecasting (1, 3, 6, and 12 months ahead) and indicate the advantage of using the optimized ensembles in terms of the Nash–Sutcliffe model efficiency coefficient (NSE), root mean squared error (RMSE), coefficient of determination (R^2), and RMSE-observations standard deviation ratio (RSR). More specifically, statistical results confirm the benefits of using optimized ESN ensembles for the given task.

The remainder of this work is organized as follows. Section 2 details the background regarding the used techniques of ELM, ESN, and MOOD. The proposed ensemble techniques are explained in Section 3. Section 4 presents the case studies and experimental procedure. The results are shown in Section 5 and discussed in Section 6. Lastly, Section 7 concludes the paper with the final remarks.

2. Background

This section brings the background necessary for this work. First, the ELM and ESN models are explained. Subsequently, the MOOD steps are introduced, namely the multi-objective problem (MOP) formulation, MOO, and multi-criteria decision making (MCDM).

2.1. Extreme Learning Machine (ELM)

ELMs are feedforward neural networks with only one intermediate layer (Huang et al., 2004). This architecture is quite similar to the traditional MLP in its conception. However, the main difference lies in the definition of the weights of the neurons in the hidden layer, given that they remain untuned during the adjustment phase (Siqueira et al.,

2014). Similar to feedforward neural networks, ELMs are universal function approximators when enough neurons are present in the hidden layer (Huang et al., 2006a).

The training process of an ELM is summarized in finding the best set of weights for output layer $\mathbf{W^{out}}$, in a minimum squared error sense. In this case, the target is to reduce the difference between the desired response $\mathbf{d}$ and the output of the network. According to Huang et al. (2006b), a good alternative to solve this task is to use the Moore–Penrose pseudo-inverse operator, because it brings to the ELM a fast training process and ensures the minimum mean squared error. In addition, the aforementioned operator guarantees the minimum norm of the weights, which leads to an increase in the generalization capability of the model (Siqueira et al., 2018). This solution is described in Eq. (1), where $\mathbf{X_{hid}} \in \mathbb{R}^{T_S \times N}$ is the matrix containing the outputs of the hidden layer for the training set, T_S is the number of training samples, and N_k is the number of neurons on the output layer.

$$\mathbf{W^{out}} = (\mathbf{X_{hid}^T X_{hid}})^{-1} \mathbf{X_{hid}^T d} \tag{1}$$

2.2. Echo State Network (ESN)

ESNs are recurrent neural networks (RNNs) since they present feedback loops in the hidden layer. Proposed by Jaeger (2001), they provide different outputs for the same set of inputs depending on the internal state of the network. The main characteristic that differs this method from traditional RNNs is that the intermediate layer, called the dynamic reservoir, remains untrained during the adjustment process (in the same way as the ELM).

The term *echo* comes from the idea that previous outputs and recent samples entered into the ESN cause more considerable influence in the formation of the output than other past signals (Siqueira et al., 2018). The *echo state property* presented in the pioneer paper by Jaeger shows the mathematical proof of the existence of the echo states (Jaeger, 2001), which means that an intrinsic memory endows the model.

An ESN, in general, presents three layers: input, dynamic reservoir, and output. In the reservoir, the neurons are fully connected, generating a non-linear characteristic. The output layer, as in the ELM, combines the outputs of the reservoir, thus being, in turn, the linear stage of the network (Siqueira et al., 2014).

Consider a generic ESN. For each input sample on time $t+1$, the internal states are updated following Eq. (2), where $\mathbf{u_t}$ is the input vector, $\mathbf{x_{t+1}}$ are the states in time $t+1$, $\mathbf{f}(\cdot)$ is the activation functions of reservoir neurons, $\mathbf{W^{in}}$ are the input layer coefficients, and $\mathbf{W}$ is the reservoir matrix.

$$\mathbf{x_{t+1}} = \mathbf{f}(\mathbf{W^{in}} \mathbf{u_{t+1}} + \mathbf{W} \mathbf{x_t}) \tag{2}$$

The network output vector $\mathbf{y_{t+1}}$ is generated by Expression (3), in which $\mathbf{W^{out}} \in \mathbb{R}^{L \times N}$ is the matrix with the synaptic weights of the output layer and L is the number of network outputs.

$$\mathbf{y_{t+1}} = \mathbf{W^{out}} \mathbf{x_{t+1}} \tag{3}$$

Again, the Moore–Penrose pseudo-inverse operator is the obvious manner to adjust the output weights of $\mathbf{W^{out}}$, as described in Section 2.1.

In addition, we consider the proposal by Jaeger (2001) to create reservoir $\mathbf{W}$. In this case, the weight matrix may present three possible values, which are randomly chosen according to the probabilities described below in Eq. (4).

$$\mathbf{W^{in}} = \begin{cases} 0,4 \text{ with probability of } 0,025 \\ -0,4 \text{ with probability of } 0,025 \\ 0 \text{ with probability of } 0,95 \end{cases} \tag{4}$$

Table 1
Preference matrix example. Five preference ranges have been defined: highly desirable (HD), desirable (D), tolerable (T) undesirable (U) and highly undesirable (HU).

Objective	Preference Matrix									
	$\leftarrow$	HD	$\rightarrow\leftarrow$	D	$\rightarrow\leftarrow$	T	$\rightarrow\leftarrow$	U	$\rightarrow\leftarrow$	HU	$\rightarrow$
	J_i^0		J_i^1		J_i^2		J_i^3		J_i^4		J_i^5
J_1	$-.--$		$-.--$		$-.--$		$-.--$		$-.--$		$-.--$
$\vdots$											
J_m	$-.--$		$-.--$		$-.--$		$-.--$		$-.--$		$-.--$

2.3. Multi-objective optimization design

A MOP with m objectives may be stated as follows (Miettinen, 1999):

$$\min_{\boldsymbol{x}} \boldsymbol{J}(\boldsymbol{x}) = [J_1(\boldsymbol{x}), \ldots, J_m(\boldsymbol{x})], \tag{5}$$

subject to

$$\boldsymbol{K}(\boldsymbol{x}) \leq 0 \tag{6}$$

$$\boldsymbol{L}(\boldsymbol{x}) = 0 \tag{7}$$

$$\underline{x_i} \leq x_i \leq \overline{x_i}, i \in [1, \ldots, n], \tag{8}$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ is defined as the decision vector with $\dim(\boldsymbol{x}) = n$, $\boldsymbol{J}(\boldsymbol{x})$ as the objective vector, $\boldsymbol{K}(\boldsymbol{x})$ and $\boldsymbol{L}(\boldsymbol{x})$ as the inequality and equality constraint vectors, respectively, and $\underline{x_i}$ and $\overline{x_i}$ as the lower and upper bounds of the decision space.

It has been noticed that there is no single solution in MOPs, because there is generally not a better solution in all the objectives. Therefore, a set of solutions is defined: the Pareto set. Each solution in the Pareto set determines an objective vector in the Pareto front (see Fig. 1). All solutions in the Pareto front are a set of Pareto optimal and non-dominated solutions. The dominance definition is described below (Miettinen, 1999):

- Pareto optimality: an objective vector $\boldsymbol{J}(\boldsymbol{x}^1)$ is Pareto optimal if there is no other objective vector $\boldsymbol{J}(\boldsymbol{x}^2)$ such as $J_i(\boldsymbol{x}^2) \leq J_i(\boldsymbol{x}^1)$ for all $i \in [1, 2, \ldots, m]$ and $J_j(\boldsymbol{x}^2) < J_j(\boldsymbol{x}^1)$ for at least one j, $j \in [1, 2, \ldots, m]$.
- Dominance: an objective vector $\boldsymbol{J}(\boldsymbol{x}^1)$ is dominated by another vector $\boldsymbol{J}(\boldsymbol{x}^2)$ if $J_i(\boldsymbol{x}^2) \leq J_i(\boldsymbol{x}^1)$ for all $i \in [1, 2, \ldots, m]$ and $J_j(\boldsymbol{x}^2) < J_j(\boldsymbol{x}^1)$ for at least one j, $j \in [1, 2, \ldots, m]$. This is denoted as $\boldsymbol{J}(\boldsymbol{x_2}) \preceq \boldsymbol{J}(\boldsymbol{x_1})$.

To successfully implement the MOO approach, three fundamental steps are required: the MOP definition, the MOO process and the MCDM stage. This integral and holistic process will be denoted hereafter as a MOOD procedure (Reynoso-Meza et al., 2014a). In the MOP statement, the design objectives are defined, as well as decision variables (with their bounds) and constraints for feasibility or suitability. In the MOO process, the primary goal is to calculate a useful Pareto front approximation. Regarding the MCDM stage, an analysis of the approximated Pareto front and trade-offs is carried out according to a set of preferences so to select the final solution to implement.

In this work the algorithm selected for the MOO process is the sp-MODEII (Reynoso-Meza et al., 2014b) algorithm. It is a multi-objective version of the Differential Evolution algorithm (Storn and Price, 1997; Das et al., 2016), which uses a spherical pruning on the objective space to maintain diversity along the Pareto front approximation. For the MCDM stage, the global physical programming index (GPPI) (Messac, 1996; Reynoso-Meza et al., 2014a) technique is used. To this end, a matrix is defined stating preferences for each design objective in a linguistic manner (see Table 1).

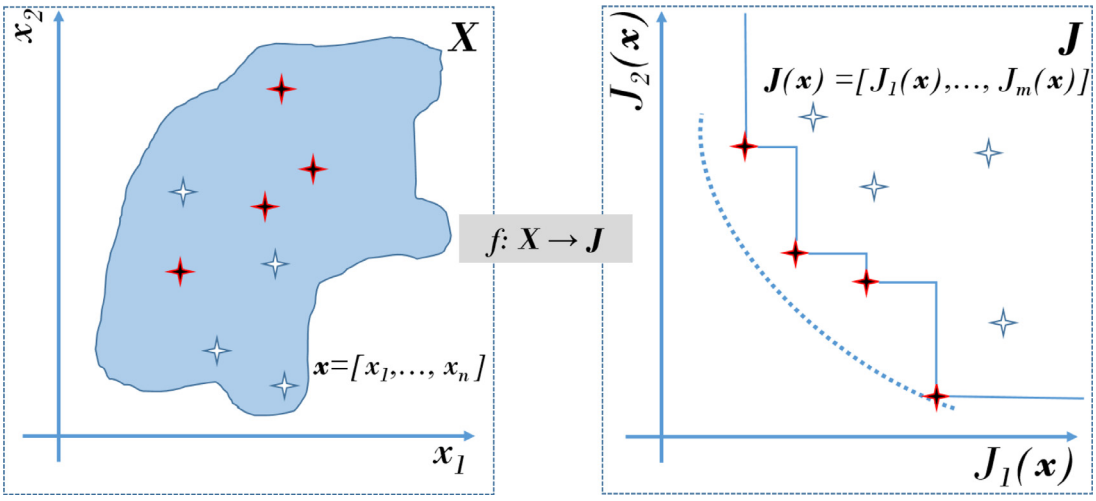


Fig. 1. Pareto optimality and dominance concepts for a min–min problem. Dark solutions are the subset of non-dominated solutions that approximate a Pareto front (right) and a Pareto set (left). The other solutions are dominated solutions, given that it is possible to find at least one solution with better values in all design objectives.
Source: Reynoso-Meza et al. (2016).

3. Optimized ensemble of unorganized machines

To predict the streamflow of hydroelectric plants, a novel ensemble methodology is proposed in this work. First, a pool of base models is generated using the bagging procedure for ELMs and ESNs. Subsequently, the selection and combination of such models are performed employing a MOOD method. The combination of bagging and MOO in an ensemble learning task has been demonstrated to have superior performance over other methods by Ribeiro and Reynoso-Meza (2019).

3.1. Bagging unorganized machines

Breiman (1996) proposed the procedure for bagging predictors, where bootstrap replicates of the training set are built for training the base learners. Such replicates are generated by uniformly sampling the training set with replacement. However, this procedure removes important time-related information when dealing with time-dependent models such as ESNs and other RNNs. Therefore, a new process for building bootstrapped time series and training the ESN is detailed in this section.

Firstly, given a time series $\boldsymbol{u} \in \mathbb{R}^{L\times T}$ with T samples and L lagged inputs, a new structure $\boldsymbol{v} \in \mathbb{R}^{L\times S\times Ts}$ with S samples and $Ts = T - S + 1$ sub-time series is created according to Algorithm 1. Such a procedure enables the selection of random sub-time series when generating ESN base models through bagging.

Algorithm 1 Procedure for generating multiple sub time series.

Inputs: $\boldsymbol{u}$ - Original time series with T samples and L lagged inputs; S - Number of samples to be used in each sub time series
Output: $\boldsymbol{v}$ - Set of Ts sub time series with S samples and L lagged inputs.

```
for t ← S; t ≤ T; t ← t + 1 do
    for s ← t − S + 1; s ≤ t; s ← s + 1 do
        for l ← 1; l ≤ L; l ← l + 1 do
            v_{l,s,t} ← u_{l,s}
        end for
    end for
end for
```

Once the training set is ready, bagging may be performed for generating the base models (Breiman, 1996). On the one hand, the original time series $\boldsymbol{u}$ is employed for building an ensemble of ELMs, where each of the T samples may be uniformly sampled with replacement for training the base learners. On the other hand, the new set of sub-time series $\boldsymbol{v}$ is employed for building an ensemble of ESNs in which each of the Ts sub-time series may be uniformly sampled with replacement for training the base learners.

To enable the training of the ESNs using multiple sub-time series, the logic behind computing the reservoir states must be modified. The new logic is performed according to Algorithm 2, in which the states are computed according to Eq. (2) for each sub-time series. However, only the last state of each sub-time series is stored for calculating the output weights $\mathbf{W}^{\mathbf{out}}$ as per Eq. (1).

Algorithm 2 Procedure for computing the echo state network (ESN) reservoir states.

Inputs: $\boldsymbol{v}$ - Set of Ts sub time series with S samples and L lagged inputs.
Output: $\boldsymbol{x}$ - Reservoir states.

```
for t ← 1; t ≤ Ts; t ← t + 1 do
    vv ← v_{*,*,t}
    for s ← 1; s ≤ S − 1; s ← s + 1 do
        for l ← 1; l ≤ L; l ← l + 1 do
            xx_{s+1} = f(W^in vv_{s+1} + W xx_s)
        end for
        x_t ← xx_{s+1}
    end for
end for
```

This section detailed how to build the pool of base models according to the bagging algorithm (Breiman, 1996). Afterward, the ensemble member selection and weighting are conducted using a MOOD procedure, which is explained in the following subsection, in which the MOP formulation, MOO, and MCDM are performed.

3.2. Multi-objective ensemble member selection and combination

As presented by Ribeiro and Reynoso-Meza (2020), multi-objective ensemble learning (MOEL) is a field of research in which the models are created through MOO. Several different applications make use of this field, such as remaining useful life prediction (Rigamonti et al., 2018), fault diagnostics of rotor-bearing systems (Ma and Chu, 2019), and image identification (Albukhanajer et al., 2017). This work employs multi-objective ensemble member selection and combination (MOEMSC), a

Table 2
Preference matrix for model selection. Five preference ranges have been defined: highly desirable (HD), desirable (D), tolerable (T) undesirable (U) and highly undesirable (HU).

Objective	Preference Matrix					
	← HD →← J_i^0	J_i^1	D →← J_i^2	T →← J_i^3	U →← J_i^4	HU → J_i^5
J_{MAE}	0.00	$0.25 \cdot T_{MAE}$	$0.50 \cdot T_{MAE}$	T_{MAE}	$1.5 \cdot T_{MAE}$	$2.0 \cdot T_{MAE}$
J_{σ^2}	0.00	$0.25 \cdot T_{\sigma^2}$	$0.50 \cdot T_{\sigma^2}$	T_{σ^2}	$1.5 \cdot T_{\sigma^2}$	$2.0 \cdot T_{\sigma^2}$

type of MOEL in which the selection and weighting of candidate ensemble members are performed through MOO (Ribeiro and Reynoso-Meza, 2020).

Once the pool of base models is generated, the selection and weighting are optimized by an evolutionary MOO algorithm. To this end, an MOP is formulated to minimize the mean absolute error and error variance of the ensemble, as shown in Eq. (9), where $J_{MAE}(\boldsymbol{x})$ is the mean absolute error of the ensemble and $J_{\sigma^2}(\boldsymbol{x})$ is the error variance.

$$\min_{\boldsymbol{x}} \boldsymbol{J}(\boldsymbol{x}) = [J_{MAE}(\boldsymbol{x}), J_{\sigma^2}(\boldsymbol{x})], \tag{9}$$

subject to

$$\boldsymbol{x} = [\boldsymbol{m}, \boldsymbol{w}] \tag{10}$$

$$m_i \in \{0, 1\}, i \in [1, \ldots, n] \tag{11}$$

$$0 \leq w_i \leq 1, i \in [1, \ldots, n], \tag{12}$$

given

$$\hat{\boldsymbol{y}} = \sum_{i=0}^{n} m_i \cdot w_i \cdot \hat{\boldsymbol{y}}^i \tag{13}$$

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{14}$$

$$\bar{e} = \sum_{t=1}^{T} e_t / T \tag{15}$$

$$J_{MAE} = \sum_{t=1}^{T} |e_t| / T \tag{16}$$

$$J_{\sigma^2} = \sum_{t=1}^{T} (e_t - \bar{e})^2 / T \tag{17}$$

The decision vector $\boldsymbol{x}$ is composed of the selection of each of the n base models (m_i) and their weights (w_i). The ensemble's prediction ($\hat{\boldsymbol{y}}$) is computed based on the decision variables and each base model's predictions ($\hat{\boldsymbol{y}}^i$), composed of T samples. The prediction error ($\boldsymbol{e}$) and its mean ($\bar{e}$) are used to calculate the two objectives of the two problems ($J_{MAE}(\boldsymbol{x})$ and $J_{\sigma^2}(\boldsymbol{x})$).

To optimize the given problem and achieve a desirable ensemble, multi-objective differential evolution with spherical pruning (spMODE-II) is used (Reynoso-Meza et al., 2014a). To this end, the algorithm is configured with a crossover ratio of 0.2, a scaling factor of 0.5, a population of 200 individuals, and 100 generations. Since this procedure results in a set of non-dominated solutions, MCDM is employed for assisting the selection of a final preferable model.

In this work, the GPPI (Reynoso-Meza et al., 2014a) is computed for each solution according to the preferences defined in Table 2, where T_{MAE} and T_{σ^2} are the mean absolute error (MAE) and error variance, respectively, calculated based on a single ESN or ELM model. The non-dominated solutions with the lowest GPPI values are usually the most preferable ones. Also, additional visualization tools may be used by the decision maker to select a final preferred solution.

4. Experiments

This section details the experimental procedure employed to evaluate the proposed method using unorganized machines ensembles with multi-objective ensemble member combination and selection. To this end, five different streamflow series data sets are tested under four forecasting configurations, namely one, three, six, and twelve months ahead. Moreover, eight models are compared: ARIMA, SARIMA, ELM, ESN, ELM ensemble, ESN ensemble, optimized ELM ensemble, and optimized ESN ensemble. Additionally, the NSE, RMSE, MAE, R^2, percentage of bias (PBIAS), and RSR metrics are computed for each solution, and the Friedman's test and the Nemenyi post hoc test with significance values of 0.01 are used for statistical comparison after 31 runs on each scenario (Demšar, 2006). Finally, the CD plots (Demšar, 2006), scatter plots, and Taylor diagrams (Taylor, 2001) are shown for each solution. The following subsections detail each step of the experimental procedure.

4.1. Datasets

The experiment uses streamflow series data collected from five Brazilian hydroelectric plants, namely those in *Sobradinho*, *Serra da Mesa*, *Jiraú*, *Furnas*, and *Água Vermelha*. In total, this work analyzes 85 years of monthly water inflow data, ranging from 1935 to 2010. Table 3 depicts the descriptive characteristics of each data set in terms of the minimum, maximum, median, mean, and standard deviation. Skewness and kurtosis are also detailed (Corder and Foreman, 2009), along with the resulting p-value of the Wald–Wolfowitz runs test for randomness (Wald and Wolfowitz, 1940).

Serra da Mesa and *Furnas* have the lowest inflows in terms of all descriptive values, while *Jiraú* has the highest. For all data sets, the skewness and kurtosis values are higher than zero, thus indicating a lower right-hand tail in the distributions and a leptokurtic distribution (heavier tails). Finally, the Wald–Wolfowitz runs test rejects the null hypothesis of data randomness at a significance value of 0.01 for all data sets.

Moreover, the datasets are split for training and testing. Of all the 85 years, the first 60 years are used for training and validation, while the later 25 years are employed for testing the models. Additional processing of the datasets is applied for removing seasonal information (Siqueira et al., 2014), according to Eq. (18), where $s_{y,m}$ is the inflow value in year y and month m, while μ_m and σ_m are the mean and standard deviation for month m computed with $N_{train} = 60$ years of training data, as follows.

$$s_{y,m} = \frac{s_{y,m} - \mu_m}{\sigma_m}, \quad y = 1, \ldots, 85, \quad m = 1, \ldots, 12 \tag{18}$$

$$\mu_m = \sum_{y=1}^{N_{train}} s_{y,m} / N_{train}, \quad m = 1, \ldots, 12 \tag{19}$$

$$\sigma_m = \sqrt{\sum_{y=1}^{N_{train}} (s_{y,m} - \mu_m) / (N_{train} - 1)}, \quad m = 1, \ldots, 12 \tag{20}$$

4.2. Models

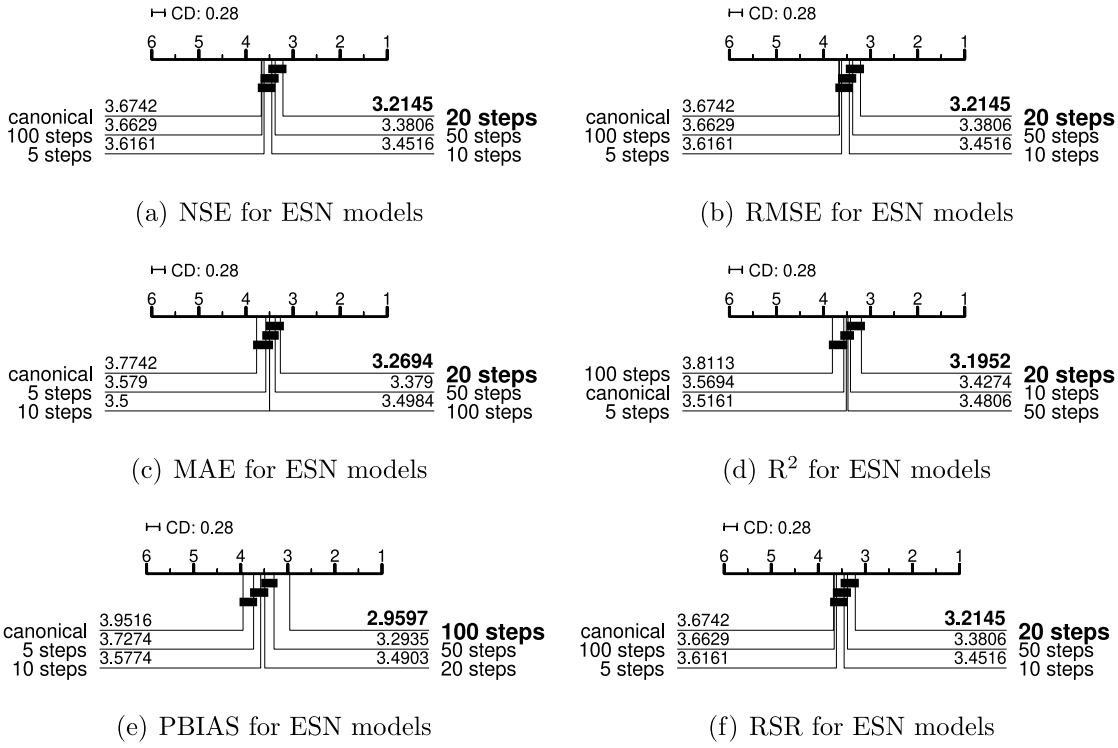
In total, the proposed experiment compares eight models: two autoregressive (ARIMA and SARIMA), two unorganized machines (ELM and ESN), two ensembles of unorganized machines (ELM-B and ESN-B), and two optimized ensembles of unorganized machines (ELM-MOB and ESN-MOB). Such models and their configurations are detailed in the following subsections.

4.2.1. Autoregressive models

For each dataset, ARIMA(p, d, q) models are estimated using the training data for $p = 1, \ldots, 6$, $d = 0, \ldots, 6$, and $q = 0, \ldots, 6$. Subsequently, the Bayesian information criterion (BIC) is measured to perform model selection (Burnham and Anderson, 2002). Finally, an additional SARIMA model ARIMA$(p, d, q)(p, d, q)_{12}$ is estimated considering the selected parameters and a seasonality of twelve months.

Table 3
Characteristics of each data set.

Data Set	Minimum	Maximum	Median	Mean	Standard deviation	Skewness	Kurtosis	Randomness (p-value)
Sobradinho	167.0	10 504.0	1 010.0	1 473.2	1 145.2	2.42	15.38	0.00
Serra da Mesa	97.0	3 235.0	466.0	662.9	544.3	1.52	5.71	0.00
Jiraú	2586.0	53 724.0	16 734.5	18 813.5	11 627.8	0.41	2.07	0.00
Furnas	102.0	3 621.0	647.5	840.3	591.8	1.63	6.07	0.00
Água Vermelha	364.0	7 144.0	1 607.0	2 036.3	1 237.2	1.35	4.79	0.00



(a) NSE for ESN models

(b) RMSE for ESN models

(c) MAE for ESN models

(d) R^2 for ESN models

(e) PBIAS for ESN models

(f) RSR for ESN models

Fig. 2. Resulting critical differences (CD) plots for different echo state network (ESN) configurations.

4.2.2. Unorganized machines

ELMs and ESNs are built for each scenario considering six inputs/lags and twelve neurons in the hidden layer/reservoir. For both models, input weights $\boldsymbol{W}^{in}$ are selected using a random uniform distribution $unif(-1, 1)$.[1] Moreover, the direct prediction approach is adopted (Sorjamaa et al., 2007), in which each model is trained considering the desired number of months ahead (1, 3, 6 and 12). Additionally, different sub-time series sizes $S = 5, 10, 20, 50, 100$ are tested for the new ESN logic (Algorithm 1).

4.2.3. Ensemble of unorganized machines

The ELM and ESN ensembles are built according to Section 3, following the bagging procedure (Breiman, 1996). For both ensembles, fifty base models are trained using the total number of available training samples but considering random selection with replacement. The ensembles also employ the direct prediction approach (Sorjamaa et al., 2007). In the results section, such models are denominated ELM-B and ESN-B.

4.2.4. Multi-objective ensemble of unorganized machines

Finally, multi-objective ensembles of unorganized machines are built according to Section 3. Of the 60 years of available training data, 50% is randomly chosen to train the ensembles, while the remaining 50% is employed to compute the J_{MAE} and J_{σ^2} scores during optimization. The direct prediction approach is also employed (Sorjamaa et al., 2007). In the results section, such ensembles are denominated ELM-MOB and ESN-MOB.

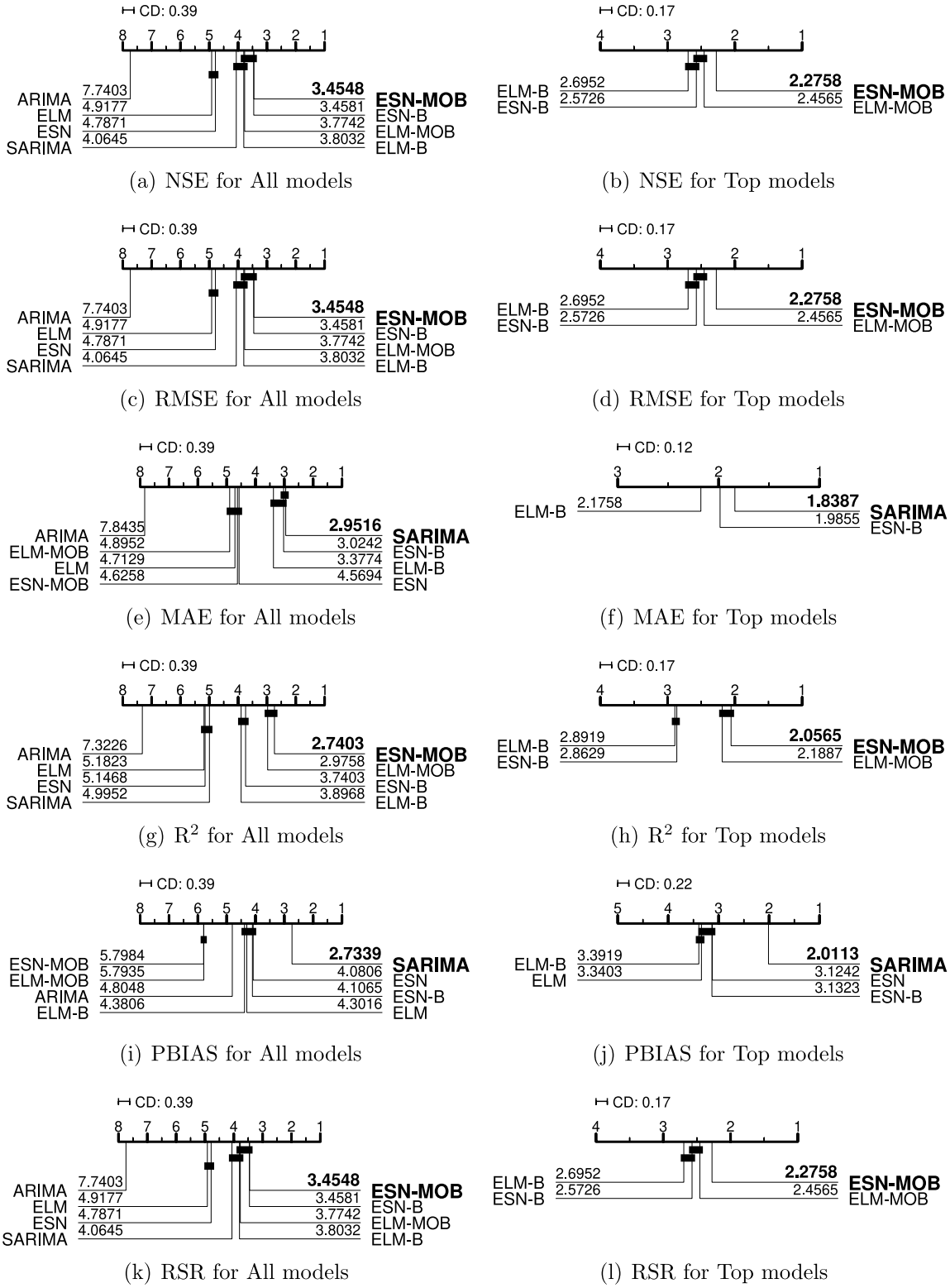
[1] Reservoir weights $\boldsymbol{W}$ and output weights $\boldsymbol{W}^{out}$ are computed according to Section 2.

4.3. Evaluation metrics

Six evaluation metrics are selected to assess the predictive performance of the models: NSE, RMSE, MAE, R^2, PBIAS, and RSR. NSE ranges from $-\infty$ to 1, with 1 indicating a perfect match, values higher than 0 presenting acceptable performance (Golmohammadi et al., 2014), and values lower than 0 showing unacceptable performance. RMSE ranges from 0 to ∞, where 0 indicates a perfect match. Similar to RMSE, MAE computes the error, ranging from 0 to ∞. R^2 describes the portion of variance explained by the model, ranging from 0 to 1, with values above 0.5 being considered acceptable (Golmohammadi et al., 2014). PBIAS measures the tendency of the prediction data, ranging from $-\infty$ (under-estimation) to ∞ (over-estimation), with an optimal value of zero. For comparison purposes, the absolute value of PBIAS is considered. Finally, RSR ranges from the optimal value of 0 to ∞. Such metrics are computed as follows for the observed $\boldsymbol{y}$ and predicted $\hat{\boldsymbol{y}}$ values.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{21}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{22}$$



(a) NSE for All models (b) NSE for Top models

(c) RMSE for All models (d) RMSE for Top models

(e) MAE for All models (f) MAE for Top models

(g) R^2 for All models (h) R^2 for Top models

(i) PBIAS for All models (j) PBIAS for Top models

(k) RSR for All models (l) RSR for Top models

Fig. 3. Statistical results considering all different scenarios and models.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{23}$$

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y}) \cdot (\overline{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2} \cdot \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2}} \tag{24}$$

$$PBIAS = \frac{\sum_{i=1}^{n} y_i - \hat{y}_i}{\sum_{i=1}^{n} y_i} \tag{25}$$

$$RSR = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{26}$$

4.4. Visualization tools

Data visualization of the results is performed in four different manners. First, overall statistical results are shown with the CD plots

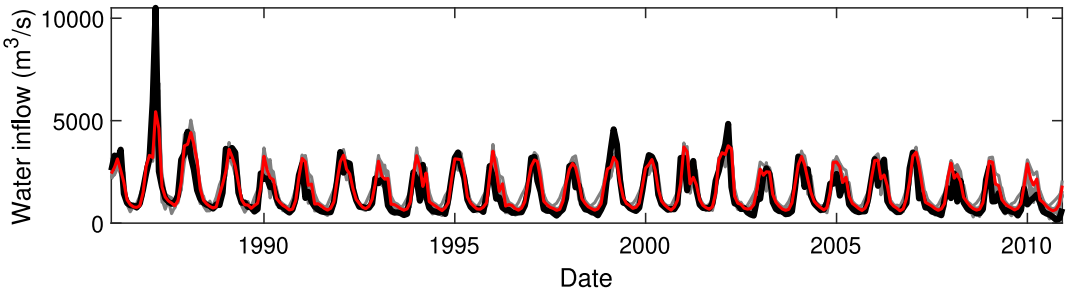


Fig. 4. Water inflow values for the *Sobradinho* data set. The black line shows the observed values, while the red line shows the ESN-MOB. The gray lines depict all the other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

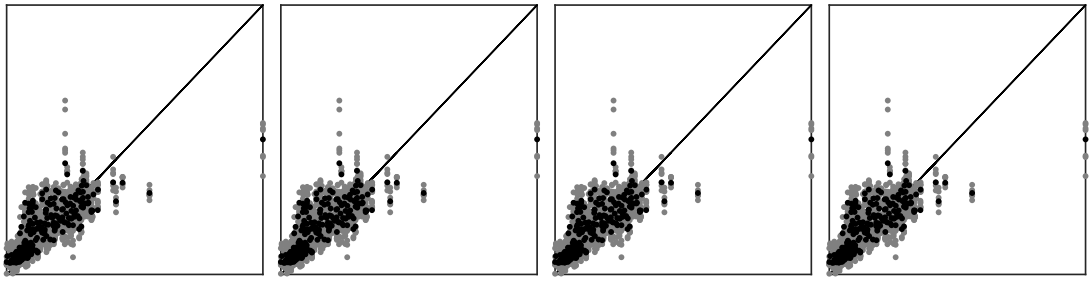


Fig. 5. Scatter plots for the *Sobradinho* data set. Black points show the predictions for the ESN-MOB model, while gray points show the predictions for all other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Demšar, 2006). For each data set, tables show the mean values for the six evaluation metrics on each model. Time series and scatter plots are employed to visualize the data dispersion. Finally, Taylor diagrams (Taylor, 2001) are shown to compare models for each data set and step scenario.

5. Results

This section presents the results for the experiments described previously. First, an analysis of the new ESN logic is performed to validate if it is possible to build bagged ESN ensembles. Next, statistical results are shown for the compared models. Finally, a detailed analysis is performed on each data set.

5.1. Analysis of the new echo state network logic

The first analysis evaluates if the proposed ESN training logic affects the model's performance. In total, six models are compared: the canonical learning algorithm and the modified ESN (Algorithm 2) using sub-time series with 5, 10, 20, 50, and 100 steps. If no significant differences are found, the new proposed training logic enables the application of the bagging ensemble strategy. Fig. 2 displays the CD plot with the statistical results for the six evaluation criteria: NSE, RMSE, MAE, R^2, PBIAS, and RSR.

The CD plots in Fig. 2 show the mean ranks of the models for each evaluation criterion separately. The lower the mean rank is, the better the model. However, models whose mean rank distances to one another are smaller than the CD value (0.28) cannot be rejected as being statistically similar. In such cases, the models are connected by a bold line. The ESNs with sub-time series of 20 and 10 steps achieve the lowest mean ranks for NSE, RMSE, MAE, R^2, and RSR, being outperformed by the model with 100 steps for PBIAS. Therefore, only the ESN with 10 steps is analyzed in the following experiments, since it attains the best performances with fewer steps. Interestingly, the canonical ESN, which uses the whole time series, is among the worst models for all evaluation criteria.

5.2. Model comparison

Next, statistical results are shown considering the eight tested models for all different scenarios. Considering the 5 datasets, 4 different prediction horizons, and 31 runs for each scenario, 620 runs are tested in total. Fig. 3 draws the CD plots for the six evaluation criteria. Moreover, a second round of experiments consisting of the top models is performed to mitigate the problem of higher type-I errors when there are too many tested models (Benavoli et al., 2016; Cruz et al., 2018).

For the NSE in Fig. 3(a), the set of top models is composed of all the ensembles of unorganized machines, which present a difference in mean ranks lower than the CD value of 0.39. Therefore, the second round of statistical tests rejects the null hypothesis that ESN-MOB comes from the same distribution as the other ensembles in Fig. 3(b). Since the ESN-MOB presents a lower mean rank by difference more extensive than the CD value of 0.17, it can be considered as the best model for this criterion. Similar results are obtained for RMSE and RSR, indicating a high correlation among the three metrics.

For MAE, SARIMA and ESN-B present the lowest mean ranks. However, in the second round of tests, SARIMA outperforms the ensemble. For PBIAS, SARIMA outperforms all the other models without the need for a second round of tests. Such results show that SARIMA has the lowest bias among all the models. Finally, for the R^2 score, ELM-MOB and ESN-MOB attain the best mean ranks without the necessity of a second round of tests. It is interesting to notice how ARIMA is never among the top models.

5.3. Results for each dataset

Finally, results are analyzed separately for each dataset. Tables are used to show the mean scores, while scatter plots and Taylor diagrams illustrate the model comparison. Such results are detailed in the following subsections.

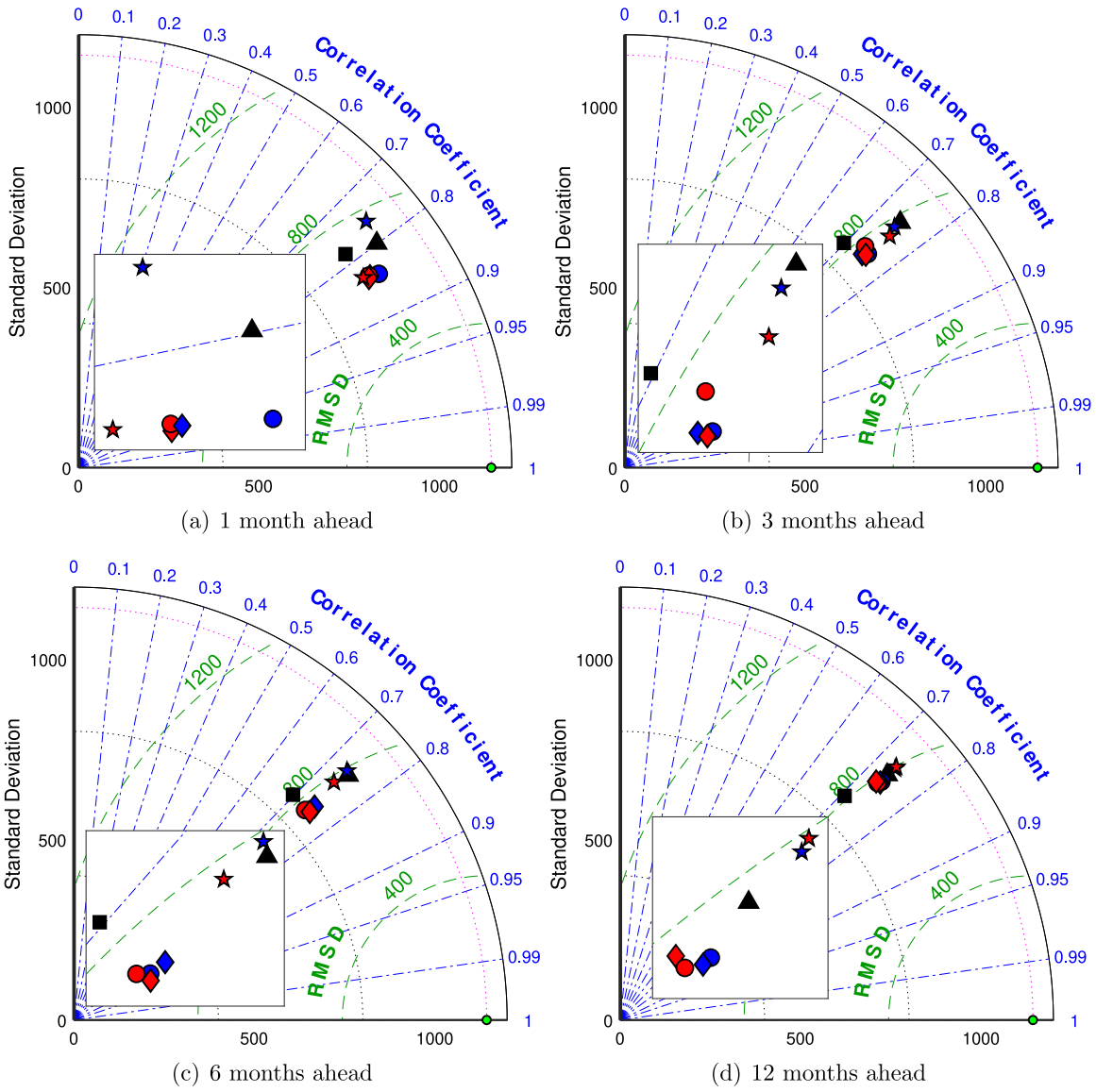


Fig. 6. Taylor diagrams for *Sobradinho* data set. The diagram shows the observed values (green circle), ARIMA (black square), SARIMA (black triangle), ELM (blue circle), ELM-B (blue diamond), ELM-MOB (blue star), ESN (red circle), ESN-B (red diamond), and ESN-MOD (red star). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.3.1. Sobradinho

For the *Sobradinho* dataset, Table 4 shows the mean scores for all six evaluation criteria, with results being computed separately for each prediction horizon (1, 3, 6, and 12 months ahead). Overall, the ELM and ESN ensembles achieve the best mean NSE, RMSE, R^2, and RSR scores for the horizons of 1 and 3 months. However, for 6 and 12 months ahead, the SARIMA model presents the best mean scores for all criteria. In this particular dataset, the optimized ensembles of unorganized machines are not among the best models.

Figs. 4 and 5 show the time series plot and scatter plot for such a dataset in the prediction horizon of 1 month ahead. In both plots, the ESN-MOB model is shown in highlight. In Fig. 4, the red lines show that ESN-MOB tends to follow the observed value closely. However, higher errors are found around the years of 1990, 1993, and 2020. The model comparison is slightly better in Fig. 5, where the black dots show a lower dispersion than several other models. It is interesting to notice how no model was able to forecast the outlier around the year of 1987.

Finally, Fig. 6 shows the Taylor diagrams considering four different prediction horizons. A zoomed view is also incorporated to discriminate the models better. For all horizons, the best RMSE and R^2 values are achieved by the unorganized machines, with little difference between single models and ensembles. However, the optimized ensemble is able to better approximate the observed standard deviation at the cost of the other two criteria. The SARIMA presents a similar scenario to the optimized ensembles for the prediction horizons of 3 and 6 months ahead, but worse trade-offs in the other scenarios.

5.3.2. Serra da mesa

Table 5 shows the mean scores for the *Serra da Mesa* dataset. Overall, SARIMA presents the best mean scores for the longer horizon predictions (6 and 12 months ahead), while unorganized machines show better results for the shorter ones (1 and 3 months ahead). Nevertheless, the optimized ensembles are able to achieve better scores than the other models composed of unorganized machines. It is important to notice that the optimized ensembles have better values for the R^2 score in detriment of worse PBIAS mean scores.

Figs. 7 and 8 show the time series plot and scatter plot for *Serra da Mesa* considering the prediction horizon of 1 month ahead, where ESN-MOB is highlighted. Once again, the presented model tends to follow the observed values. However, higher peaks in some years (1987, 1993,

Table 4
Mean results for the *Sobradinho* data set.

Evaluation metric	Months ahead	Single model		Ensemble		Optimized ensemble		Autoregressive	
		ELM	ESN	ELM	ESN	ELM	ESN	ARIMA	SARIMA
NSE (↑)	1	0.68	0.68	0.70	0.70	0.53	0.61	0.53	0.63
	3	0.53	0.53	0.54	0.54	0.44	0.47	0.21	0.53
	6	0.51	0.51	0.51	0.52	0.42	0.45	0.20	0.53
	12	0.47	0.48	0.47	0.48	0.39	0.38	0.27	0.51
RMSE (↓)	1	$6.5 \cdot 10^2$	$6.5 \cdot 10^2$	$6.3 \cdot 10^2$	$6.3 \cdot 10^2$	$7.8 \cdot 10^2$	$7.1 \cdot 10^2$	$7.8 \cdot 10^2$	$7.0 \cdot 10^2$
	3	$7.8 \cdot 10^2$	$7.8 \cdot 10^2$	$7.8 \cdot 10^2$	$7.8 \cdot 10^2$	$8.6 \cdot 10^2$	$8.3 \cdot 10^2$	$1.0 \cdot 10^3$	$7.8 \cdot 10^2$
	6	$8.0 \cdot 10^2$	$8.0 \cdot 10^2$	$8.0 \cdot 10^2$	$7.9 \cdot 10^2$	$8.7 \cdot 10^2$	$8.5 \cdot 10^2$	$1.0 \cdot 10^3$	$7.8 \cdot 10^2$
	12	$8.3 \cdot 10^2$	$8.3 \cdot 10^2$	$8.3 \cdot 10^2$	$8.2 \cdot 10^2$	$8.9 \cdot 10^2$	$9.0 \cdot 10^2$	$9.8 \cdot 10^2$	$8.0 \cdot 10^2$
MAE (↓)	1	$4.0 \cdot 10^2$	$4.0 \cdot 10^2$	$3.9 \cdot 10^2$	$3.9 \cdot 10^2$	$5.2 \cdot 10^2$	$4.6 \cdot 10^2$	$5.4 \cdot 10^2$	$4.0 \cdot 10^2$
	3	$4.9 \cdot 10^2$	$4.9 \cdot 10^2$	$4.9 \cdot 10^2$	$4.8 \cdot 10^2$	$5.7 \cdot 10^2$	$5.4 \cdot 10^2$	$7.9 \cdot 10^2$	$4.5 \cdot 10^2$
	6	$5.2 \cdot 10^2$	$5.1 \cdot 10^2$	$5.1 \cdot 10^2$	$5.1 \cdot 10^2$	$5.8 \cdot 10^2$	$5.6 \cdot 10^2$	$7.9 \cdot 10^2$	$4.6 \cdot 10^2$
	12	$5.3 \cdot 10^2$	$5.2 \cdot 10^2$	$5.3 \cdot 10^2$	$5.2 \cdot 10^2$	$6.0 \cdot 10^2$	$6.1 \cdot 10^2$	$7.4 \cdot 10^2$	$4.7 \cdot 10^2$
R2 [%] (↑)	1	82.7	82.5	83.6	83.7	79.9	81.9	78.1	79.9
	3	74.5	74.4	75.0	75.1	74.7	74.7	69.8	74.7
	6	74.0	74.1	74.6	74.5	74.1	74.2	69.7	74.6
	12	73.4	73.4	73.6	73.6	73.8	73.8	70.7	73.6
PBIAS [%] (↓)	1	5.1	4.7	4.8	4.8	23.1	17.0	21.6	2.8
	3	11.9	10.7	12.1	11.6	24.6	20.6	40.6	5.1
	6	15.4	14.4	15.9	14.6	26.1	22.9	40.9	5.9
	12	18.1	17.4	18.3	17.0	28.0	28.4	36.8	7.2
RSR (↓)	1	$6.4 \cdot 10^3$	$6.4 \cdot 10^3$	$6.0 \cdot 10^3$	$6.0 \cdot 10^3$	$9.3 \cdot 10^3$	$7.7 \cdot 10^3$	$9.3 \cdot 10^3$	$7.4 \cdot 10^3$
	3	$9.3 \cdot 10^3$	$9.3 \cdot 10^3$	$9.2 \cdot 10^3$	$9.1 \cdot 10^3$	$1.1 \cdot 10^4$	$1.1 \cdot 10^4$	$1.6 \cdot 10^4$	$9.3 \cdot 10^3$
	6	$9.8 \cdot 10^3$	$9.6 \cdot 10^3$	$9.6 \cdot 10^3$	$9.5 \cdot 10^3$	$1.2 \cdot 10^4$	$1.1 \cdot 10^4$	$1.6 \cdot 10^4$	$9.3 \cdot 10^3$
	12	$1.0 \cdot 10^4$	$1.0 \cdot 10^4$	$1.0 \cdot 10^4$	$1.0 \cdot 10^4$	$1.2 \cdot 10^4$	$1.2 \cdot 10^4$	$1.4 \cdot 10^4$	$9.6 \cdot 10^3$

Table 5
Mean results for the *Serra da Mesa* data set.

Evaluation metric	Months ahead	Single model		Ensemble		Optimized ensemble		Autoregressive	
		ELM	ESN	ELM	ESN	ELM	ESN	ARIMA	SARIMA
NSE (↑)	1	0.64	0.64	0.65	0.65	0.66	0.66	0.62	0.61
	3	0.53	0.54	0.54	0.54	0.57	0.57	0.52	0.55
	6	0.52	0.52	0.53	0.53	0.54	0.54	0.52	0.56
	12	0.56	0.57	0.57	0.57	0.55	0.55	0.53	0.57
RMSE (↓)	1	$3.3 \cdot 10^2$	$3.3 \cdot 10^2$	$3.2 \cdot 10^2$	$3.2 \cdot 10^2$	$3.2 \cdot 10^2$	$3.2 \cdot 10^2$	$3.3 \cdot 10^2$	$3.4 \cdot 10^2$
	3	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.8 \cdot 10^2$	$3.6 \cdot 10^2$
	6	$3.7 \cdot 10^2$	$3.8 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.8 \cdot 10^2$	$3.6 \cdot 10^2$
	12	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.7 \cdot 10^2$	$3.6 \cdot 10^2$
MAE (↓)	1	$1.9 \cdot 10^2$	$1.9 \cdot 10^2$	$1.9 \cdot 10^2$	$1.9 \cdot 10^2$	$2.0 \cdot 10^2$	$2.1 \cdot 10^2$	$2.3 \cdot 10^2$	$2.1 \cdot 10^2$
	3	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.8 \cdot 10^2$	$2.4 \cdot 10^2$
	6	$2.6 \cdot 10^2$	$2.6 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.8 \cdot 10^2$	$2.3 \cdot 10^2$
	12	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.8 \cdot 10^2$	$2.3 \cdot 10^2$
R^2 [%] (↑)	1	80.7	80.9	81.6	81.5	82.8	82.7	80.4	78.4
	3	74.8	74.9	75.1	75.0	78.2	78.8	76.3	75.3
	6	76.8	76.4	77.1	76.9	78.4	78.6	76.4	75.9
	12	79.4	79.3	79.5	79.5	79.4	79.4	76.4	76.2
PBIAS [%] (↓)	1	1.3	1.2	1.2	1.2	10.6	10.6	11.7	1.8
	3	5.7	5.8	5.7	5.8	12.7	14.4	20.5	3.2
	6	13.8	13.0	13.7	13.4	16.7	16.9	21.3	3.8
	12	16.6	16.0	16.4	16.1	18.3	18.1	19.3	3.8
RSR (↓)	1	$3.4 \cdot 10^3$	$3.4 \cdot 10^3$	$3.3 \cdot 10^3$	$3.3 \cdot 10^3$	$3.2 \cdot 10^3$	$3.2 \cdot 10^3$	$3.5 \cdot 10^3$	$3.7 \cdot 10^3$
	3	$4.4 \cdot 10^3$	$4.4 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$4.1 \cdot 10^3$	$4.0 \cdot 10^3$	$4.5 \cdot 10^3$	$4.2 \cdot 10^3$
	6	$4.5 \cdot 10^3$	$4.5 \cdot 10^3$	$4.4 \cdot 10^3$	$4.4 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$4.6 \cdot 10^3$	$4.1 \cdot 10^3$
	12	$4.1 \cdot 10^3$	$4.1 \cdot 10^3$	$4.1 \cdot 10^3$	$4.0 \cdot 10^3$	$4.2 \cdot 10^3$	$4.2 \cdot 10^3$	$4.4 \cdot 10^3$	$4.0 \cdot 10^3$

1999, and 2002) and lower values in others (1991, 1994, 2010) hamper the forecasting task, as shown in Fig. 7. Nevertheless, Fig. 8 shows high dispersion for all tested models, with higher inflow values being harder to predict.

Despite the difficult forecasting task considering the *Serra da Mesa* dataset, Fig. 9 indicates that unorganized machines present overall better scores than the autoregressive models. Specifically, the optimized ensembles are able to reduce the RMSE and improve the R^2 scores for 1, 3, and 6 months ahead. In contrast, for the prediction horizon of 12 months ahead, the optimized ensembles are able to improve the observed standard deviation approximation.

5.3.3. *Jiraú*

The mean scores for the *Jiraú* dataset are detailed in Table 6. For this dataset, unorganized machines attain the best values for all scores except the PBIAS, and single ELMs and ESNs are better for the short prediction horizons (1 and 3 months ahead), while the optimized ensembles obtain better scores for the longer horizons (6 and 12 months ahead). Specifically, for the horizon of 1 month ahead, SARIMA presents the best overall performance.

Fig. 10 illustrates how ESN-MOB has trouble forecasting the higher observed inflow values, noticeable around the year of 2009. Nevertheless, Fig. 11 shows that this model does not present a variance different than other models for lower and median inflow values.

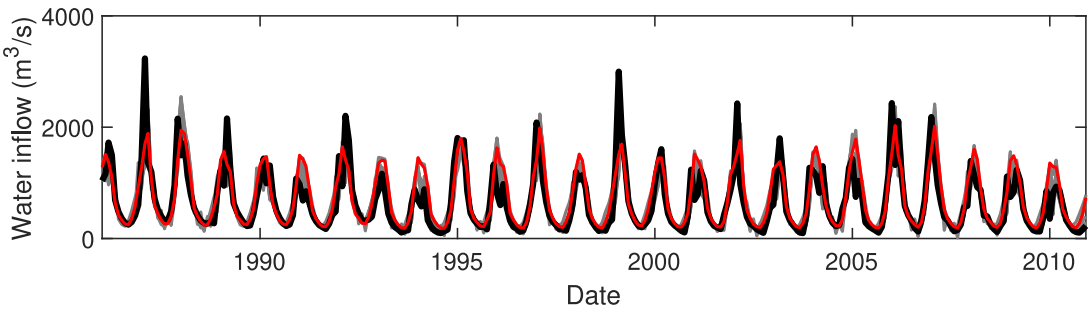


Fig. 7. Water inflow values for the *Serra da Mesa* data set. The black line shows the observed values, while the red line shows the ESN-MOB. The gray lines depict all the other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

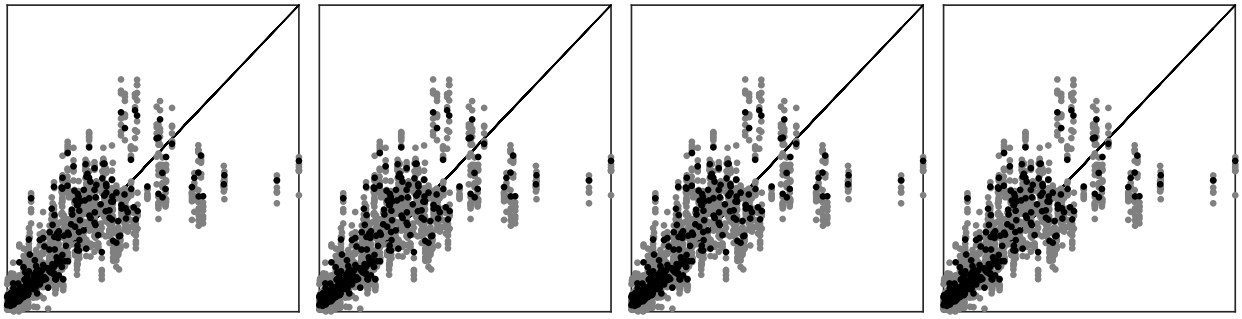


Fig. 8. Scatter plots for the *Serra da Mesa* data set. Black points show the predictions for the ESN-MOB model, while gray points show the predictions for all other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Mean results for the *Jiraú* data set.

Evaluation metric	Months ahead	Single model		Ensemble		Optimized ensemble		Autoregressive	
		ELM	ESN	ELM	ESN	ELM	ESN	ARIMA	SARIMA
NSE (↑)	1	0.95	0.95	0.95	0.95	0.93	0.93	0.94	0.95
	3	0.90	0.90	0.90	0.90	0.89	0.89	0.87	0.90
	6	0.87	0.87	0.87	0.87	0.88	0.88	0.85	0.87
	12	0.86	0.87	0.86	0.87	0.87	0.87	0.86	0.87
RMSE (↓)	1	$2.7 \cdot 10^3$	$2.7 \cdot 10^3$	$2.6 \cdot 10^3$	$2.6 \cdot 10^3$	$3.0 \cdot 10^3$	$3.1 \cdot 10^3$	$2.7 \cdot 10^3$	$2.5 \cdot 10^3$
	3	$3.7 \cdot 10^3$	$3.7 \cdot 10^3$	$3.7 \cdot 10^3$	$3.7 \cdot 10^3$	$3.8 \cdot 10^3$	$3.8 \cdot 10^3$	$4.2 \cdot 10^3$	$3.8 \cdot 10^3$
	6	$4.2 \cdot 10^3$	$4.2 \cdot 10^3$	$4.2 \cdot 10^3$	$4.1 \cdot 10^3$	$4.1 \cdot 10^3$	$4.0 \cdot 10^3$	$4.4 \cdot 10^3$	$4.2 \cdot 10^3$
	12	$4.3 \cdot 10^3$	$4.2 \cdot 10^3$	$4.3 \cdot 10^3$	$4.2 \cdot 10^3$	$4.2 \cdot 10^3$	$4.1 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$
MAE (↓)	1	$2.0 \cdot 10^3$	$2.0 \cdot 10^3$	$2.0 \cdot 10^3$	$2.0 \cdot 10^3$	$2.3 \cdot 10^3$	$2.3 \cdot 10^3$	$2.0 \cdot 10^3$	$1.9 \cdot 10^3$
	3	$2.8 \cdot 10^3$	$2.8 \cdot 10^3$	$2.8 \cdot 10^3$	$2.8 \cdot 10^3$	$2.9 \cdot 10^3$	$2.9 \cdot 10^3$	$3.2 \cdot 10^3$	$2.8 \cdot 10^3$
	6	$3.2 \cdot 10^3$	$3.1 \cdot 10^3$	$3.2 \cdot 10^3$	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.4 \cdot 10^3$	$3.2 \cdot 10^3$
	12	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.1 \cdot 10^3$	$3.3 \cdot 10^3$	$3.2 \cdot 10^3$
R^2 [%] (↑)	1	97.3	97.3	97.5	97.5	96.6	96.4	97.2	97.7
	3	95.0	94.9	95.0	95.0	94.8	94.7	93.2	94.7
	6	93.6	93.7	93.7	93.8	93.9	94.0	92.6	93.4
	12	93.1	93.2	93.1	93.2	93.5	93.6	93.1	93.2
PBIAS [%] (↓)	1	1.5	1.5	1.5	1.4	1.1	1.1	0.2	0.5
	3	1.9	1.9	2.0	1.9	1.4	1.4	0.7	1.4
	6	2.4	2.3	2.4	2.3	1.8	1.6	0.9	2.1
	12	2.1	2.0	2.1	2.1	1.6	1.6	0.8	2.8
RSR (↓)	1	$1.1 \cdot 10^4$	$1.1 \cdot 10^4$	$1.0 \cdot 10^4$	$1.0 \cdot 10^4$	$1.4 \cdot 10^4$	$1.5 \cdot 10^4$	$1.1 \cdot 10^4$	$9.3 \cdot 10^3$
	3	$2.0 \cdot 10^4$	$2.1 \cdot 10^4$	$2.0 \cdot 10^4$	$2.0 \cdot 10^4$	$2.1 \cdot 10^4$	$2.1 \cdot 10^4$	$2.7 \cdot 10^4$	$2.1 \cdot 10^4$
	6	$2.6 \cdot 10^4$	$2.6 \cdot 10^4$	$2.6 \cdot 10^4$	$2.5 \cdot 10^4$	$2.5 \cdot 10^4$	$2.4 \cdot 10^4$	$2.9 \cdot 10^4$	$2.6 \cdot 10^4$
	12	$2.7 \cdot 10^4$	$2.7 \cdot 10^4$	$2.7 \cdot 10^4$	$2.7 \cdot 10^4$	$2.6 \cdot 10^4$	$2.5 \cdot 10^4$	$2.8 \cdot 10^4$	$2.7 \cdot 10^4$

Fig. 12 illustrates how the ESN-MOB achieves poor results for the prediction horizon of 1 month ahead compared to the other models. However, for the horizons of 6 and 12 months ahead, the multi-objective optimization procedure is able to reduce both the RMSE and R^2 scores, thus outperforming other unorganized machines and the autoregressive models. Nevertheless, such a method decreases the standard deviation approximation.

5.3.4. Furnas

Results for the *Furnas* dataset are shown in Table 7 for all six evaluation criteria. In this dataset, the optimized ensembles of unorganized machines tend to outperform the other models for all prediction horizons in terms of NSE, RMSE, R^2, and RSR. Similar to previous datasets, SARIMA obtains the best mean MAE and PBIAS scores.

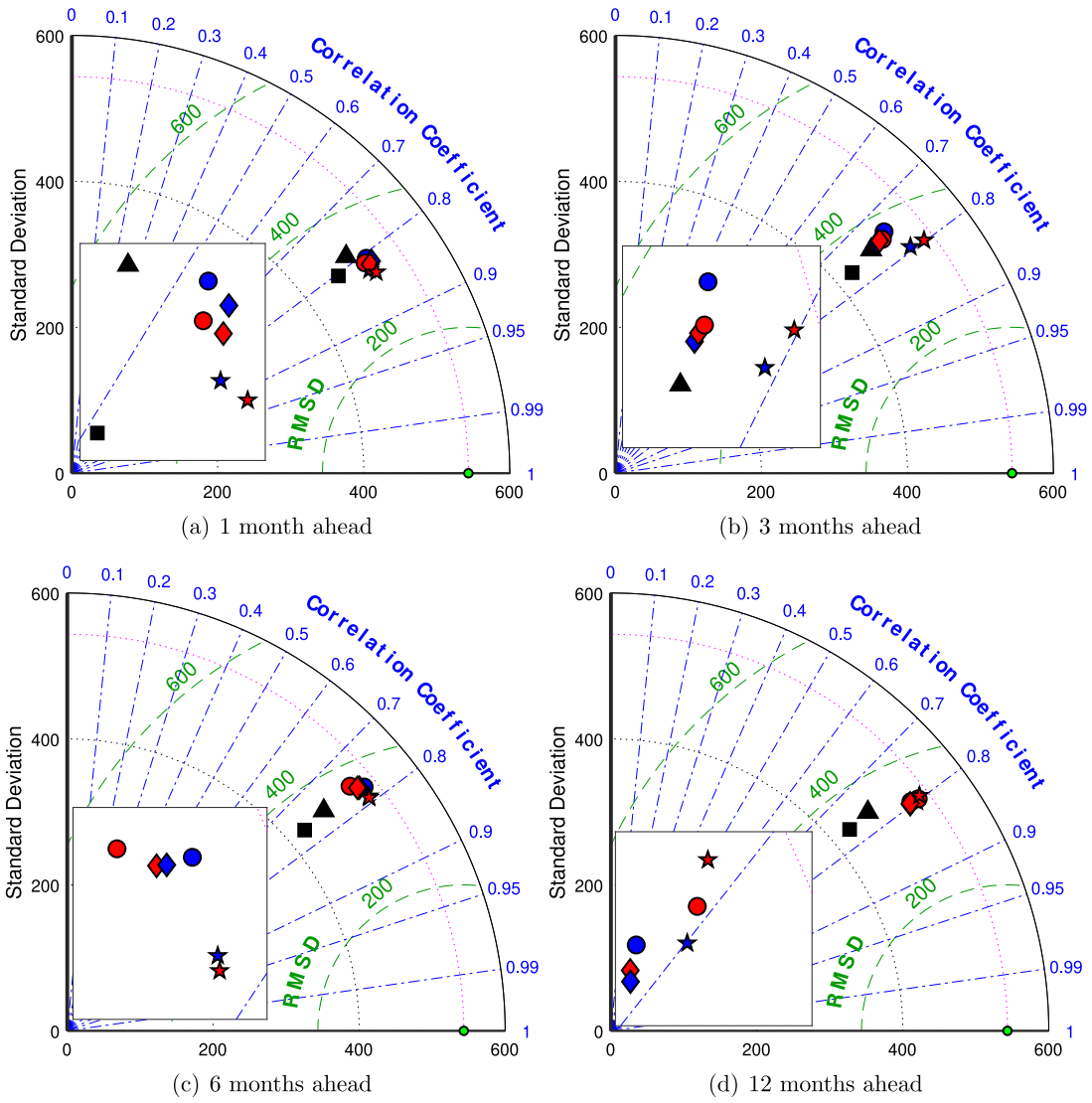


Fig. 9. Taylor diagrams for *Serra da Mesa* data set. The diagram shows the observed values (green circle), ARIMA (black square), SARIMA (black triangle), ELM (blue circle), ELM-B (blue diamond), ELM-MOB (blue star), ESN (red circle), ESN-B (red diamond), and ESN-MOD (red star). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

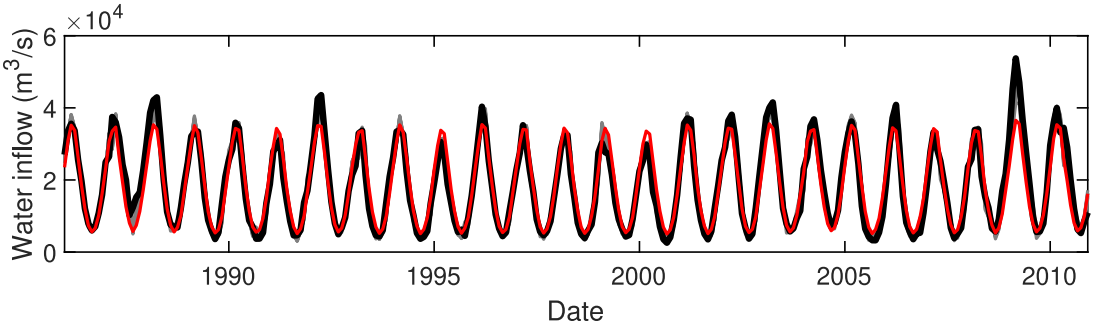


Fig. 10. Water inflow values for the *Jiraú* data set. The black line shows the observed values, while the red line shows the ESN-MOB. The gray lines depict all the other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The prediction plot in Fig. 13 shows that all models have problems detecting peak values, as well as the lower inflow values that occur around the years of 1993, 1996, 2009, and 2010. In terms of dispersion, shown in Fig. 14, the proposed optimized ensemble approach maintains better predictions at lower inflow values. Despite the high variance in median and higher values, ESN-MOB achieves lower variance than other methods.

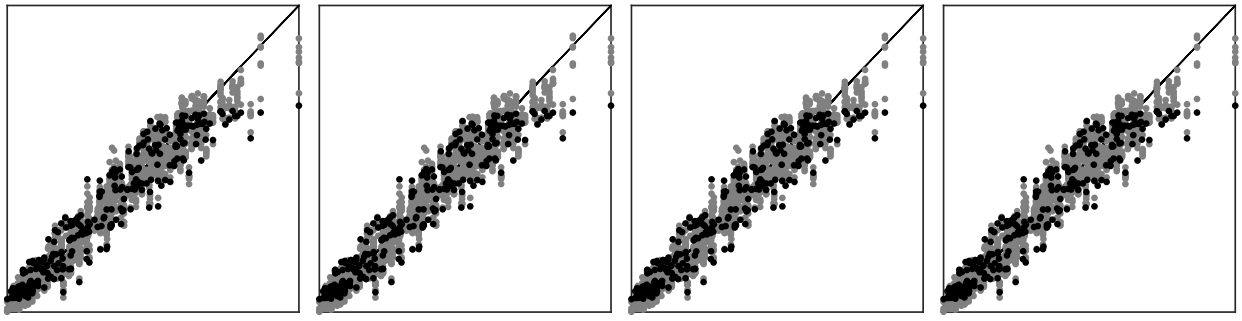


Fig. 11. Scatter plots for the *Jiraú* data set. Black points show the predictions for the ESN-MOB model, while gray points show the predictions for all other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

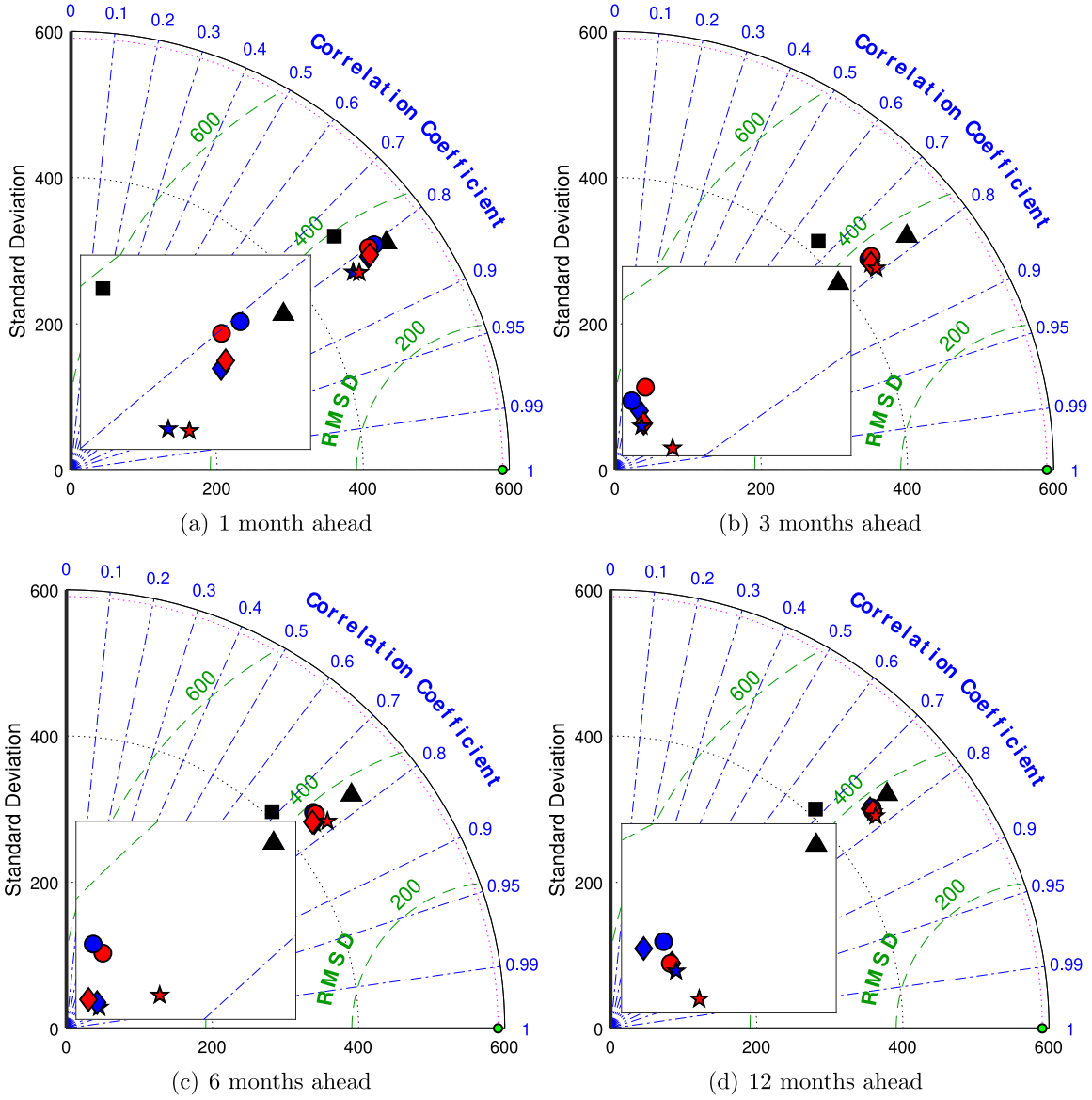


Fig. 12. Taylor diagrams for *Jiraú* data set. The diagram shows the observed values (green circle), ARIMA (black square), SARIMA (black triangle), ELM (blue circle), ELM-B (blue diamond), ELM-MOB (blue star), ESN (red circle), ESN-B (red diamond), and ESN-MOD (red star). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Mean results for the *Furnas* data set.

Evaluation metric	Months ahead	Single model		Ensemble		Optimized ensemble		Autoregressive	
		ELM	ESN	ELM	ESN	ELM	ESN	ARIMA	SARIMA
NSE (↑)	1	0.63	0.63	0.65	0.65	0.66	0.66	0.55	0.65
	3	0.58	0.58	0.59	0.60	0.61	0.61	0.44	0.60
	6	0.57	0.58	0.58	0.58	0.59	0.59	0.47	0.59
	12	0.58	0.57	0.58	0.58	0.58	0.58	0.47	0.58
RMSE (↓)	1	$3.6 \cdot 10^2$	$3.6 \cdot 10^2$	$3.5 \cdot 10^2$	$3.5 \cdot 10^2$	$3.4 \cdot 10^2$	$3.4 \cdot 10^2$	$3.9 \cdot 10^2$	$3.5 \cdot 10^2$
	3	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$3.7 \cdot 10^2$	$4.4 \cdot 10^2$	$3.7 \cdot 10^2$
	6	$3.9 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$4.3 \cdot 10^2$	$3.8 \cdot 10^2$
	12	$3.8 \cdot 10^2$	$3.9 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$3.8 \cdot 10^2$	$4.3 \cdot 10^2$	$3.8 \cdot 10^2$
MAE (↓)	1	$2.2 \cdot 10^2$	$2.2 \cdot 10^2$	$2.1 \cdot 10^2$	$2.1 \cdot 10^2$	$2.2 \cdot 10^2$	$2.2 \cdot 10^2$	$2.5 \cdot 10^2$	$2.1 \cdot 10^2$
	3	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$2.4 \cdot 10^2$	$3.1 \cdot 10^2$	$2.3 \cdot 10^2$
	6	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.9 \cdot 10^2$	$2.4 \cdot 10^2$
	12	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.5 \cdot 10^2$	$2.6 \cdot 10^2$	$2.6 \cdot 10^2$	$2.9 \cdot 10^2$	$2.4 \cdot 10^2$
R^2 [%] (↑)	1	79.7	79.9	81.0	81.0	81.8	81.7	74.8	81.2
	3	76.3	76.4	77.1	77.4	78.8	79.0	66.4	78.0
	6	75.6	76.1	76.6	76.6	77.6	77.7	68.9	77.4
	12	76.7	76.6	76.8	76.8	77.6	77.5	68.3	76.4
PBIAS [%] (↓)	1	1.1	0.8	0.9	0.8	4.8	5.5	0.2	0.7
	3	0.8	1.1	0.8	1.1	6.4	6.4	0.6	1.5
	6	2.9	3.1	3.1	3.3	7.1	7.4	0.9	2.0
	12	7.3	7.3	7.3	7.3	10.1	9.8	1.4	2.9
RSR (↓)	1	$3.8 \cdot 10^3$	$3.7 \cdot 10^3$	$3.5 \cdot 10^3$	$3.5 \cdot 10^3$	$3.4 \cdot 10^3$	$3.5 \cdot 10^3$	$4.6 \cdot 10^3$	$3.6 \cdot 10^3$
	3	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$4.2 \cdot 10^3$	$4.1 \cdot 10^3$	$4.0 \cdot 10^3$	$4.0 \cdot 10^3$	$5.7 \cdot 10^3$	$4.1 \cdot 10^3$
	6	$4.4 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$4.2 \cdot 10^3$	$4.2 \cdot 10^3$	$4.2 \cdot 10^3$	$5.4 \cdot 10^3$	$4.2 \cdot 10^3$
	12	$4.3 \cdot 10^3$	$4.4 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$4.3 \cdot 10^3$	$5.5 \cdot 10^3$	$4.3 \cdot 10^3$

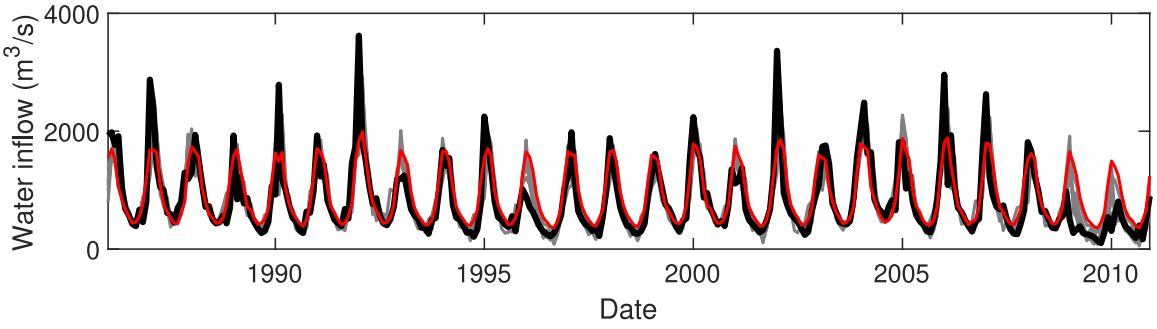


Fig. 13. Water inflow values for the *Furnas* data set. The black line shows the observed values, while the red line shows the ESN-MOB. The gray lines depict all the other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

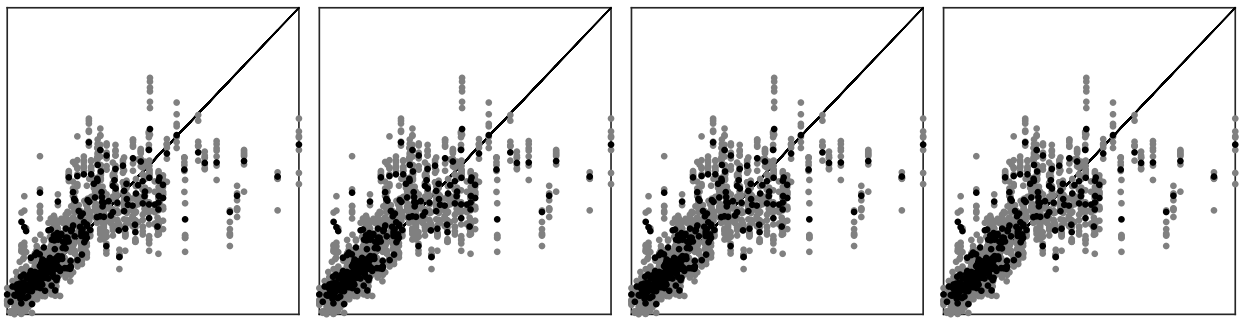


Fig. 14. Scatter plots for the *Furnas* data set. Black points show the predictions for the ESN-MOB model, while gray points show the predictions for all other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Finally, the Taylor diagrams (Fig. 15) show that SARIMA presents the best overall standard deviation approximation. It also attains similar RMSE and R^2 scores than single and bagged ensembles of unorganized machines. However, the multi-objective optimization approach improves both criteria.

5.3.5. Água Vermelha

Lastly, Table 8 shows the results for the *Água Vermelha* dataset. The best NSE scores are achieved by the optimized ELM and ESN ensembles. The same is true for the RMSE, R^2, and RSR, regardless of the prediction horizon. For MAE and PBIAS, SARIMA presents the lowest results for the prediction horizons of 1 and 6 months ahead. However, for 3 and

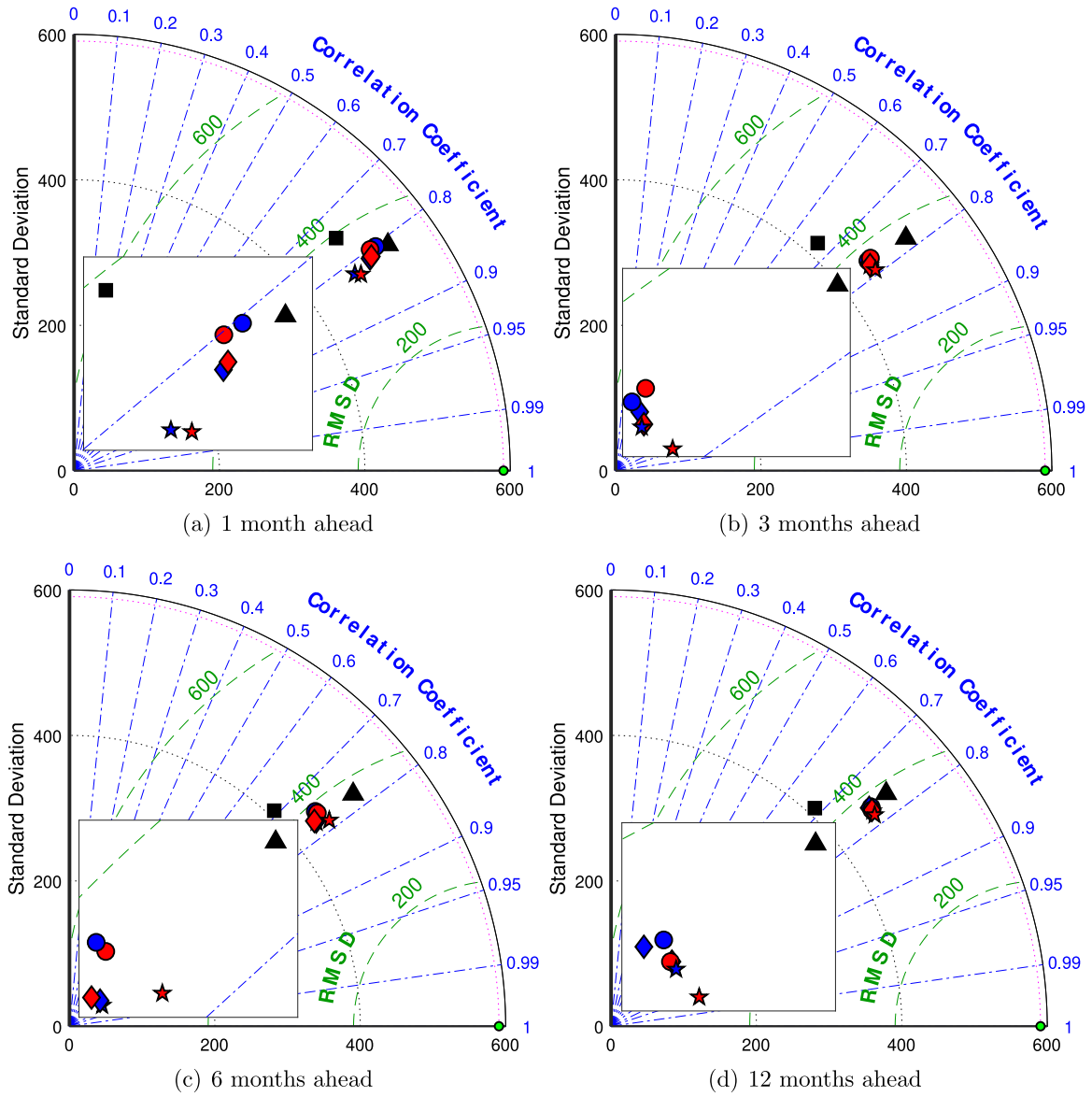


Fig. 15. Taylor diagrams for *Furnas* data set. The diagram shows the observed values (green circle), ARIMA (black square), SARIMA (black triangle), ELM (blue circle), ELM-B (blue diamond), ELM-MOB (blue star), ESN (red circle), ESN-B (red diamond), and ESN-MOD (red star). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

12 months ahead, the non-optimized unorganized machines achieve the lowest scores.

The time series plot in Fig. 16 illustrates once again the difficulty for all models to predict outlier values, such as those around the years of 1990, 1992, 1996, 2002, 2009, and 2010. However, it is interesting to notice how ESN-MOB is able to maintain a better prediction after peak values around 1988, 1993, and 2005. The difficulty in predicting this dataset is also shown in Fig. 17, where a high variance is present for the predictions.

Finally, Fig. 18 shows the Taylor diagram for this final dataset. Yet again, SARIMA has the best approximation of the observed standard deviation. In terms of R^2, the ensemble approaches tend to achieve better results, while the optimization process is more efficient in shorter prediction horizons (1 and 3 months ahead). RMSE values for this dataset tend to show little variation regardless of the model.

6. Discussion

This section discusses the results of the two previous ones. First, the analysis of the new ESN training logic is compared to its canonical algorithm. Next, the proposed ensemble learning techniques are compared to the single unorganized machines and autoregressive models. Finally, the benefits of the proposed method in hydrological problems are detailed.

Regarding the ESN training logic modification, given that the canonical algorithm must receive the whole time series in the correct order for training, the bagging method for ensemble generation would not be possible. However, the new proposed training logic splits the whole time series into multiple sub-time series. Results indicate that it is better to use smaller sub-time series rather than the entire time series in the process. Therefore, the experiments justify and enable the development of ensembles with training data diversity through bagging.

Eight models were compared in terms of the NSE, RMSE, MAE, R^2, PBIAS, and RSR values. First, non-seasonal ARIMA models are among the worst results for all metrics. However, SARIMA models have considerable improvement, achieving the best results for MAE and PBIAS. However, unorganized machines attain better results in terms of NSE, RMSE, R^2, and RSR. Regarding the type of unorganized machines, there is little difference between using ELM and ESN for the proposed task, even when building the ensemble models. Nevertheless,

Table 8
Mean results for the *Água Vermelha* data set.

Evaluation metric	Months ahead	Single model		Ensemble		Optimized ensemble		Autoregressive	
		ELM	ESN	ELM	ESN	ELM	ESN	ARIMA	SARIMA
NSE (↑)	1	0.69	0.70	0.71	0.71	0.72	0.72	0.62	0.70
	3	0.63	0.63	0.64	0.64	0.66	0.66	0.42	0.64
	6	0.63	0.62	0.63	0.63	0.64	0.64	0.43	0.63
	12	0.63	0.63	0.64	0.63	0.64	0.64	0.44	0.62
RMSE (↓)	1	$6.8 \cdot 10^2$	$6.8 \cdot 10^2$	$6.7 \cdot 10^2$	$6.7 \cdot 10^2$	$6.6 \cdot 10^2$	$6.6 \cdot 10^2$	$7.6 \cdot 10^2$	$6.8 \cdot 10^2$
	3	$7.5 \cdot 10^2$	$7.5 \cdot 10^2$	$7.4 \cdot 10^2$	$7.4 \cdot 10^2$	$7.2 \cdot 10^2$	$7.2 \cdot 10^2$	$9.4 \cdot 10^2$	$7.5 \cdot 10^2$
	6	$7.5 \cdot 10^2$	$7.6 \cdot 10^2$	$7.5 \cdot 10^2$	$7.5 \cdot 10^2$	$7.4 \cdot 10^2$	$7.4 \cdot 10^2$	$9.3 \cdot 10^2$	$7.5 \cdot 10^2$
	12	$7.5 \cdot 10^2$	$7.5 \cdot 10^2$	$7.5 \cdot 10^2$	$7.5 \cdot 10^2$	$7.4 \cdot 10^2$	$7.4 \cdot 10^2$	$9.2 \cdot 10^2$	$7.6 \cdot 10^2$
MAE (↓)	1	$4.2 \cdot 10^2$	$4.2 \cdot 10^2$	$4.1 \cdot 10^2$	$4.1 \cdot 10^2$	$4.2 \cdot 10^2$	$4.2 \cdot 10^2$	$5.0 \cdot 10^2$	$4.1 \cdot 10^2$
	3	$4.9 \cdot 10^2$	$4.9 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$4.7 \cdot 10^2$	$4.7 \cdot 10^2$	$6.9 \cdot 10^2$	$4.7 \cdot 10^2$
	6	$4.9 \cdot 10^2$	$4.9 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$6.7 \cdot 10^2$	$4.9 \cdot 10^2$
	12	$4.9 \cdot 10^2$	$4.9 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$4.8 \cdot 10^2$	$6.6 \cdot 10^2$	$4.9 \cdot 10^2$
R^2 [%] (↑)	1	83.5	83.6	84.4	84.3	84.8	84.8	78.8	84.1
	3	79.6	79.4	80.2	80.3	81.3	81.4	65.4	80.3
	6	79.4	79.1	79.8	79.6	80.4	80.4	68.5	79.7
	12	79.5	79.5	79.8	79.7	80.0	80.4	74.0	79.3
PBIAS [%] (↓)	1	1.5	1.3	1.4	1.3	0.8	0.9	0.8	0.3
	3	0.4	0.4	0.3	0.2	1.3	1.3	2.2	0.9
	6	1.3	1.5	1.4	1.4	2.0	2.0	2.7	1.2
	12	1.6	1.7	1.7	1.7	2.2	2.4	2.0	1.9
RSR (↓)	1	$6.6 \cdot 10^3$	$6.5 \cdot 10^3$	$6.2 \cdot 10^3$	$6.2 \cdot 10^3$	$6.0 \cdot 10^3$	$6.0 \cdot 10^3$	$8.2 \cdot 10^3$	$6.4 \cdot 10^3$
	3	$7.9 \cdot 10^3$	$8.0 \cdot 10^3$	$7.7 \cdot 10^3$	$7.6 \cdot 10^3$	$7.3 \cdot 10^3$	$7.2 \cdot 10^3$	$1.2 \cdot 10^4$	$7.8 \cdot 10^3$
	6	$8.0 \cdot 10^3$	$8.1 \cdot 10^3$	$7.8 \cdot 10^3$	$7.9 \cdot 10^3$	$7.6 \cdot 10^3$	$7.6 \cdot 10^3$	$1.2 \cdot 10^4$	$8.0 \cdot 10^3$
	12	$7.9 \cdot 10^3$	$7.9 \cdot 10^3$	$7.8 \cdot 10^3$	$7.8 \cdot 10^3$	$7.7 \cdot 10^3$	$7.6 \cdot 10^3$	$1.2 \cdot 10^4$	$8.1 \cdot 10^3$

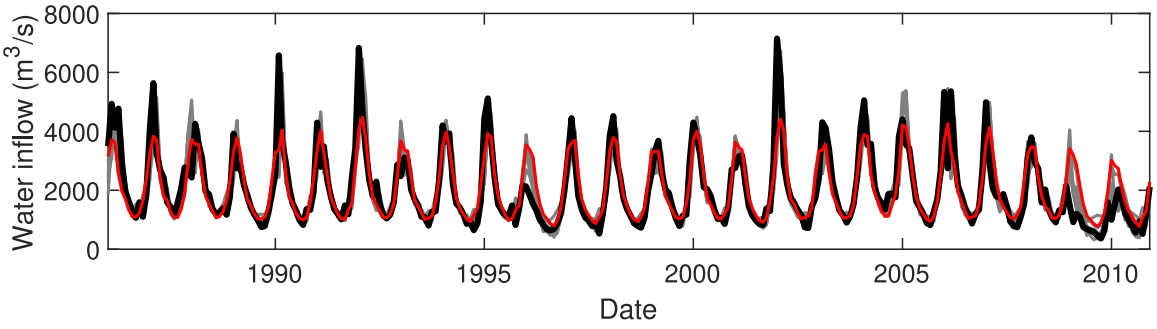


Fig. 16. Water inflow values for the *Água Vermelha* data set. The black line shows the observed values, while the red line shows the ESN-MOB. The gray lines depict all the other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

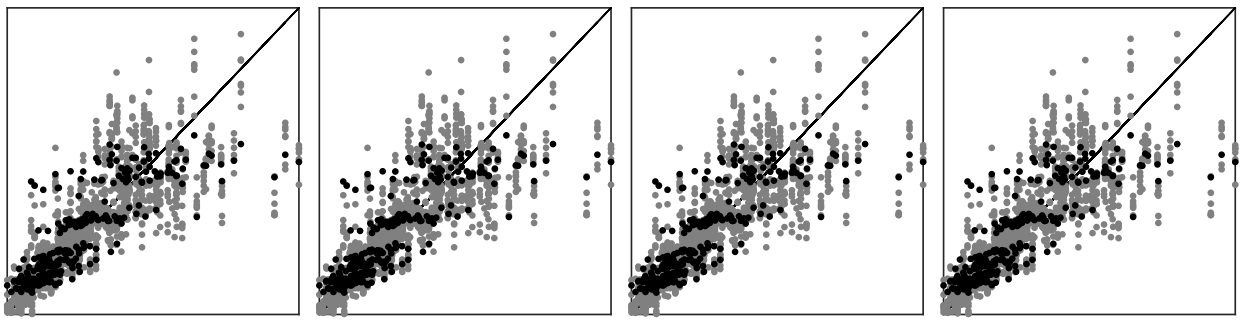


Fig. 17. Scatter plots for the *Água Vermelha* data set. Black points show the predictions for the ESN-MOB model, while gray points show the predictions for all other models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the bagging procedure improves the scores for such models. Moreover, results show that ensembles benefit from the MOEMSC procedure. In particular, the ESN ensemble is able to achieve the best results in terms of NSE, RMSE, and RSR.

Finally, it is possible to present some practical benefits of this investigation for the interested parties of the electric sector, such as government agencies, the private initiative, and the population. The pricing strategies and values paid in energy auctions are directly related to the availability of water in future scenarios. In the Brazilian case, it is even more critical, given that 70% of the power generation is from hydroelectric plants. According to Siqueira et al. (2018), this percentage represents 60,000 MW of installed capacity. Therefore, a single percentage point of energy represents the addition of a plant with 1000 MW of installed capacity.

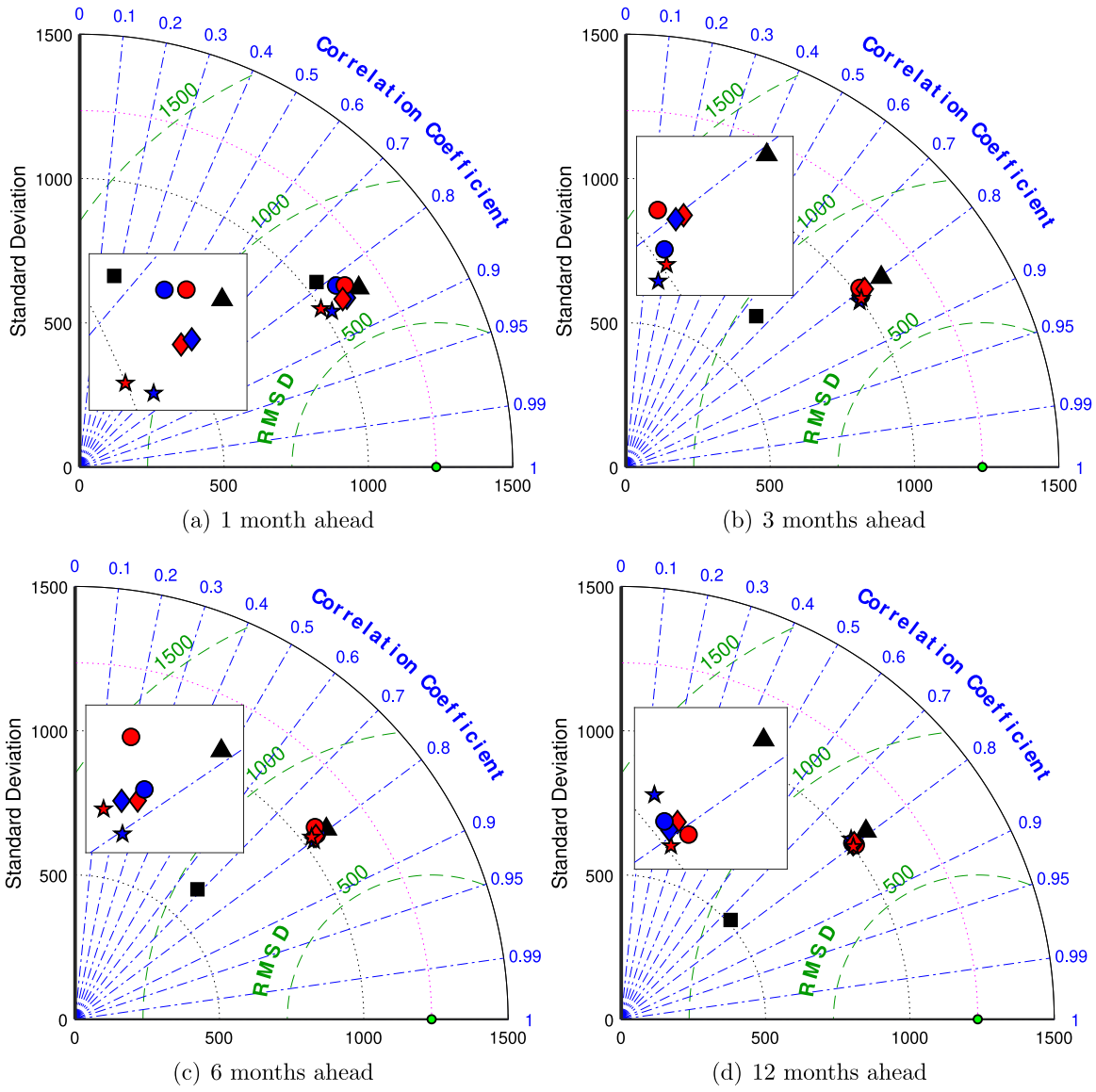


Fig. 18. Taylor diagrams for *Água Vermelha* data set. The diagram shows the observed values (green circle), ARIMA (black square), SARIMA (black triangle), ELM (blue circle), ELM-B (blue diamond), ELM-MOB (blue star), ESN (red circle), ESN-B (red diamond), and ESN-MOD (red star). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The proper management of water facilitates planning the dispatch of the plants, especially for interconnected systems since the water availability may be different depending on the hydrographic basin. Also, for planning the construction of new facilities, it is essential to better know, in advance, the future trends in water availability.

There is no cost for using water as a source to generate energy, and its utilization presents a small environmental impact. It is also remarkable since the alternative to provide electricity is often the dispatching of thermoelectric plants, which are polluting and whose fuel cost (*e.g.*, coal or oil) is much higher. Just the increase of 106 TWh in the hydro-power generation in 2019 compared to 2018 saved at least 80 million metric tons of carbon emissions (Association, 2020).

According to the Hydropower Status Report 2020 from the International Hydropower Association (Association, 2020), in 2019, there was a 2.5% increase in power generation using hydroelectric plants, the most significant contribution from a renewable energy source in history. The report considers 13,000 stations in 150 countries. It seems to be clear that the proper water management may bring substantial benefits in avoiding water or money waste.

7. Conclusions

This work addresses the monthly seasonal streamflow series forecasting problem for Brazilian hydroelectric plants. To this end, unorganized machines, namely ELMs and ESNs, have shown positive results in the literature. Therefore, ensembles of unorganized machines are generated and optimized to forecast the streamflow of five different hydroelectric plants, namely *Sobradinho*, *Serra da Mesa*, *Jirau*, *Furnas*, and *Agua Vermelha*, under four different prediction horizons (1, 3, 6, and 12 months ahead).

To this end, the following two main contributions are presented. First, a modified training logic is proposed for ESN to enable the generation of ensembles using bootstrapped samples. Second, by employing the bagging procedure to generate ESN and ELM ensembles, MOOD is applied to optimize the selection and weight assignment of each base model in the ensemble. The proposed method minimizes both the error bias and variance on the prediction task, enabling the development of more robust time series forecasting models.

Experiments are performed to evaluate both contributions in terms of NSE, RMSE, MAE, R^2, PBIAS, and RSR values. The first experiment

compares the canonical ESN algorithm against the new proposed training logic. Statistical results not only validate the new training logic but also suggest that limiting the time series size in ESNs improves the algorithms performance for all criteria. The second experiment compares eight different models: ARIMA, SARIMA, single ELM, single ESN, bagged ELM ensemble, bagged ESN ensemble, optimized ELM ensemble, and optimized ESN ensemble. Statistical results indicate that the application of ensembles and MOO improves the performance of forecasting models in terms of NSE, RMSE, R^2, and RSR. Moreover, results are visualized for all scenarios using scatter plots and Taylor diagrams, in which it is shown that the optimized ensembles of unorganized machines benefit mostly in terms of RMSE and R^2.

Finally, future work shall focus on the development of boosted ensembles of unorganized machines. Moreover, feature selection may also be explored as a multi-objective optimization problem. Additionally, different time series forecasting problems shall be explored with the proposed contributions.

CRediT authorship contribution statement

Victor Henrique Alves Ribeiro: Performed the experiments, Wrote the manuscript. **Gilberto Reynoso-Meza:** Supervised the experiments, Revised the manuscript. **Hugo Valadares Siqueira:** Curated the datasets for the experiments, Wrote the manuscript.

Declaration of competing 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.

Acknowledgment

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), Brazil - Finance Code: 159063/2017-0-PROSUC, the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), Brazil - Universal Projects: 405580/2018-5 and 437105/2018-0 - Research Productivity Fellowship: 304066/2016-8-PQ2, and the *Fundação Araucária* (FAPPR) - Brazil - Finance Codes: 51497 and PRONEX-042/2018.

References

Al-Sudani, Z.A., Salih, S.Q., Yaseen, Z.M., et al., 2019. Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. J. Hydrol. 573, 1–12.

Albukhanajer, W.A., Jin, Y., Briffa, J.A., 2017. Classifier ensembles for image identification using multi-objective pareto features. Neurocomputing 238, 316–327.

Arsenault, R., Côté, P., 2018. Analysis of the effects of biases in ensemble streamflow prediction (esp) forecasts on electricity production in hydropower reservoir management. Hydrol. Earth Syst. Sci. 23 (6), 2735–2750.

Association, I.H., 2020. Hydropower Status Report. International Hydropower Association, London, UK.

Benavoli, A., Corani, G., Mangili, F., 2016. Should we really use post-hoc tests based on mean-ranks? J. Mach. Learn. Res. 17 (1), 152–161.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Burnham, K.P., Anderson, D.R., 2002. A practical information-theoretic approach. In: Model Selection and Multimodel Inference, second ed. Springer, New York.

Corder, G.W., Foreman, D.I., 2009. Nonparametric Statistics for Non-Statisticians: A Step-By-Step Approach. John Wiley & Sons.

Cruz, R.M., Sabourin, R., Cavalcanti, G.D., 2018. Dynamic classifier selection: Recent advances and perspectives. Inf. Fusion 41, 195–216.

Das, S., Mullick, S.S., Suganthan, P.N., 2016. Recent advances in differential evolution–an updated survey. Swarm Evol. Comput. 27, 1–30.

de Matto Neto, P.S., Madeiro, F., Ferreira, T.A., Cavalcanti, G.D., 2014. Hybrid intelligent system for air quality forecasting using phase adjustment. Eng. Appl. Artif. Intell. 32, 185–191.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7 (Jan), 1–30.

Dilini, W., Attygalle, D.T., Liyanage, L., Nandalal, K., 2016. Ensemble forecast for monthly reservoir inflow: a dynamic neural network approach. In: Proc. Proceedings of the 4th Annual International Conference on Operations Research and Statistics, ORS 2016, pp. 18–19.

EPE - Energy Research Company, 2019. National Energy Balance 2019. Ministry of Mines and Energy - Brazil, Rio de Janeiro.

Fan, F.M., Schwanenberg, D., Alvarado, R., Reis, A.A., Collischonn, W., Naumman, S., 2016. Performance of deterministic and probabilistic hydrological forecasts for the short-term optimization of a tropical hydropower reservoir. Water Resour. Manage. 30 (10), 3609–3625.

Firmino, P.R.A., de Matto Neto, P.S., Ferreira, T.A., 2014. Correcting and combining time series forecasters. Neural Netw. 50, 1–11.

Francelin, R., Ballini, R., Andrade, M.G., 1996. Back-propagation and Box & Jenkins approaches to streamflow forecasting. In: Simpósio Brasileiro de Pesquisa Operacional-SBPO and Latin-Iberian-American Congress on Operations Research and System Engineering-CLAIO, pp. 1307–1312.

Ghorbani, M., Khatibi, R., Karimi, V., Yaseen, Z.M., Zounemat-Kermani, M., 2018. Learning from multiple models using artificial intelligence to improve model prediction accuracies: application to river flows. Water Resour. Manage. 32 (13), 4201–4215.

Golmohammadi, G., Prasher, S., Madani, A., Rudra, R., 2014. Evaluating three hydrological distributed watershed models: Mike-she, apex, swat. Hydrology 1 (1), 20–39.

Hailegeorgis, T.T., Alfredsen, K., 2017. Regional statistical and precipitation–runoff modelling for ecological applications: prediction of hourly streamflow in regulated rivers and ungauged basins. River Res. Appl. 33 (2), 233–248.

Haykin, S.S., et al., 2009. Neural Networks and Learning Machines/Simon Haykin. Prentice Hall, New York.

Hernandez-Ambato, J., Asqui-Santillan, G., Arellano, A., Cunalata, C., 2017. Multistep-ahead streamflow and reservoir level prediction using anns for production planning in hydroelectric stations. In: 2017 16th IEEE International Conference on Machine Learning and Applications. ICMLA, IEEE, pp. 479–484.

Huang, G.-B., Chen, L., Siew, C.K., et al., 2006a. Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans. Neural Netw. 17 (4), 879–892.

Huang, G.H., Zhu, Q.Y., Siew, C.K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of International Joint Conference on Neural Networks. Vol. 2. IEEE International Joint Conference on Neural Networks, pp. 985–990.

Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006b. Extreme learning machine: theory and applications. Neurocomputing 70 (1), 489–501.

Jaeger, H., 2001. The "Echo State" Approach To Analysing and Training Recurrent Neural Networks-with an Erratum Note, Vol. 148. German National Research Center for Information Technology GMD Technical Report, (34), p. 13.

Kasiviswanathan, K.S., Sudheer, K.P., 2013. Quantification of the predictive uncertainty of artificial neural network based river flow forecast models. Stoch. Environ. Res. Risk Assess. 27 (1), 137–146.

Lima, A.R., Cannon, A.J., Hsieh, W.W., 2015. Nonlinear regression in environmental sciences using extreme learning machines: a comparative evaluation. Environ. Model. Softw. 73, 175–188.

Lima, A.R., Cannon, A.J., Hsieh, W.W., 2016. Forecasting daily streamflow using online sequential extreme learning machines. J. Hydrol. 537, 431–443.

Ma, S., Chu, F., 2019. Ensemble deep learning-based fault diagnosis of rotor bearing systems. Comput. Ind. 105, 143–152.

Malfatti, M.G.L., Cardoso, A.O., Hamburger, D.S., 2018. Linear empirical model for streamflow forecast in itaipu hydroelectric dam–parana river basin. Rev. Bras. Meteorol. 33 (2), 257–268.

Messac, A., 1996. Physical programming-effective optimization for computational design. AIAA J. 34 (1), 149–158.

Miettinen, K., 1999. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston, Massachusetts.

Reynoso-Meza, G., Blasco, X., Sanchis, J., Martínez, M., 2014a. Controller tuning using evolutionary multi-objective optimisation: current trends and applications. Control Eng. Pract. 28, 58–73.

Reynoso-Meza, G., Sanchis, J., Blasco, X., García-Nieto, S., 2014b. Physical programming for preference driven evolutionary multi-objective optimization. Appl. Soft Comput. 24, 341–362.

Reynoso-Meza, G., Sanchis, J., Blasco, X., Martínez, M., 2016. Preference driven multi-objective optimization design procedure for industrial controller tuning. Inform. Sci. 339, 108–131.

Ribeiro, V.H.A., Reynoso-Meza, G., 2019. A study of pareto-based methods for ensemble pool generation and aggregation. In: 2019 IEEE Congress on Evolutionary Computation. CEC, IEEE, pp. 2145–2152.

Ribeiro, V.H.A., Reynoso-Meza, G., 2020. Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. Expert Syst. Appl. 147, 113232.

Ribeiro, M.H.D.M., Ribeiro, V.H.A., Reynoso-Meza, G., dos Santo Coelho, L., 2019. Multi-objective ensemble model for short-term price forecasting in corn price time series. In: 2019 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–8.

Ribeiro, M.H.D.M., dos Santos Coelho, L., 2020. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl. Soft Comput. 86, 105837.

Rigamonti, M., Baraldi, P., Zio, E., Roychoudhury, I., Goebel, K., Poll, S., 2018. Ensemble of optimized echo state networks for remaining useful life prediction. Neurocomputing 281, 121–138.

Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33 (1–2), 1–39.

Sacchi, R., Ozturk, M.C., Principe, J.C., Carneiro, A.A.F.M., Silva, I.N., 2007. Water inflow forecasting using the echo state network: a brazilian case study. In: International Joint Conference on Neural Networks. IEEE International Joint Conference on Neural Networks, IJCNN, pp. 2403–2408.

Sharkey, A.J.C., 1999. Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems. Springer London, London.

Siqueira, H., Boccato, L., Attux, R., Lyra, C., 2014. Unorganized machines for seasonal streamflow series forecasting. Int. J. Neural Syst. 24 (03), 1430009.

Siqueira, H.V., Boccato, L., Attux, R.R.F., Lyra Filho, C., 2012. Echo state networks in seasonal streamflow series prediction. Learn. Nonlinear Models 10, 181–191.

Siqueira, H., Boccato, L., Luna, I., Attux, R., Lyra, C., 2018. Performance analysis of unorganized machines in streamflow forecasting of brazilian plants. Appl. Soft Comput. 68, 494–506.

Siqueira, H., Luna, I., 2019. Performance comparison of feedforward neural networks applied to streamflow series forecasting. Math. Eng. Sci. Aerosp. (MESA) 10 (1), 41–53.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., Lendasse, A., 2007. Methodology for long-term prediction of time series. Neurocomputing 70 (16–18), 2861–2869.

Stojković, M., Kostić, S., Prohaska, S., Plavšić, J., Tripković, V., 2017. A new approach for trend assessment of annual streamflows: a case study of hydropower plants in serbia. Water Resour. Manage. 31 (4), 1089–1103.

Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 11 (4), 341–359.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmos. 106 (D7), 7183–7192.

Thober, S., Kumar, R., Wanders, N., Marx, A., Pan, M., Rakovec, O., Samaniego, L., Sheffield, J., Wood, E.F., Zink, M., 2018. Multi-model ensemble projections of european river floods and high flows at 1.5, 2, and 3 degrees global warming. Environ. Res. Lett. 13 (1), 014003.

Wald, A., Wolfowitz, J., 1940. On a test whether two samples are from the same population. Ann. Math. Stat. 11 (2), 147–162.

Yaseen, Z.M., Ebtehaj, I., Bonakdari, H., Deo, R.C., Mehr, A.D., Mohtar, W.H.M.W., Diop, L., El-Shafie, A., Singh, V.P., 2017. Novel approach for streamflow forecasting using a hybrid anfis-ffa model. J. Hydrol. 554, 263–276.

Yaseen, Z.M., Mohtar, W.H.M.W., Ameen, A.M.S., Ebtehaj, I., Razali, S.F.M., Bonakdari, H., Salih, S.Q., Al-Ansari, N., Shahid, S., 2019a. Implementation of univariate paradigm for streamflow simulation using hybrid data-driven model: Case study in tropical region. IEEE Access 7, 74471–74481.

Yaseen, Z.M., Sulaiman, S.O., Deo, R.C., Chau, K.-W., 2019b. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol. 569, 387–408.

Zhu, S., Zhou, J., Ye, L., Meng, C., 2016. Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of yangtze river, china. Environ. Earth Sci. 75 (6), 531.