Water Supply

Streamflow forecasting in Tocantins river basins using machine learning

Victor Braga Rodrigues Duarte [a,*], Marcelo Ribeiro Viola[b], Marcos Giongo[a], Eduardo Morgan Uliana[c] and Carlos Rogério de Mello[b]

[a] Center for Environmental Monitoring and Fire Management, Forest Engineering Department, Federal University of Tocantins, Gurupi, TO 77404-970, Brazil
[b] Water Resources Department, Federal University of Lavras, Lavras, MG 37200-900, Brazil
[c] Institute of Agrarian and Environmental Sciences, Federal University of Mato Grosso, Sinop, MT 78557-267, Brazil
*Corresponding author. E-mail: victorbrduarte@gmail.com

VBRD, 0000-0002-4958-6810

ABSTRACT

Understanding the behavior of the river regime in watersheds is fundamental for water resources planning and management. Empirical hydrological models are powerful tools for this purpose, with the selection of input variables as one of the main steps of the modeling. Therefore, the objectives of this study were to select the best input variables using the genetic, recursive feature elimination, and vsurf algorithms, and to evaluate the performance of the random forest, artificial neural networks, support vector regression, and M5 model tree models in forecasting daily streamflow in Sono (SRB), Manuel Alves da Natividade (MRB), and Palma (PRB) River basins. Based on several performance indexes, the best model in all basins was the M5 model tree, which showed the best performances in SRB and PRB using the variables selected by the recursive feature elimination algorithm. The good performance of the evaluated models allows them to be used to assist different demands faced by the water resources management in the studied river basins, especially the M5 model tree model using streamflow lags, average rainfall, and evapotranspiration as inputs.

Key words: artificial intelligence, feature selection, hydrological forecasting, hydrology

HIGHLIGHTS

- The Recursive Feature Elimination was the best input feature selection algorithm.
- The machine learning models were efficient in the daily streamflow forecasting.
- The performance of the M5 model tree was better than the other models.
- The models are potential tools to assist water resources management.

GRAPHICAL ABSTRACT
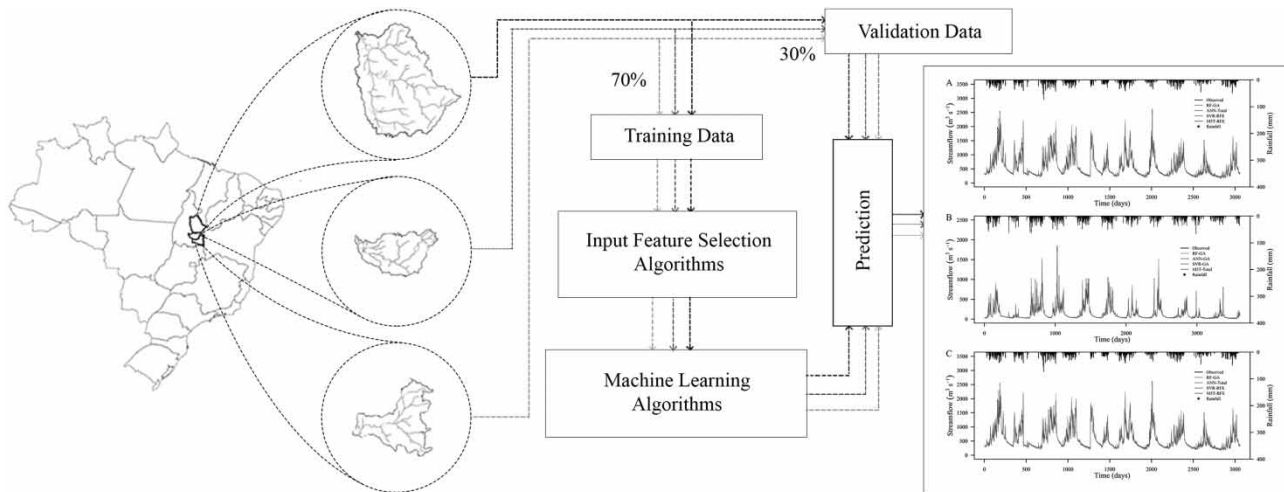
1. INTRODUCTION

Brazil is the country with the highest amount of fresh water in the world. This resource has been used mainly in agriculture, energy production, economic growth, and sanitation (Cantelle *et al.* 2018). Due to population growth and economic development, it is estimated that in the last two decades there has been an increase of 80% in total water withdrawal from water bodies, and that there will be an increase of 23% by the year 2030 (ANA 2020). Based on the foregoing, understanding the river system behavior in basins is fundamental to hydroelectricity planning, characterization of grant flows, flood forecasting, and assessment of impacts of climate change and soil use, among others (Bourdin *et al.* 2012; Tongal & Booij 2018; Yaseen *et al.* 2018).

The streamflow forecasting and prognoses of hydrological behavior are often performed using hydrological models, which can be classified as conceptual or empirical (Debastiani *et al.* 2019). Conceptual hydrological models are based on physical characteristics of basins and require large amounts of data, which in many contexts are difficult to be acquired, are unavailable or insufficient for covering all the spatial and time variability (Yang *et al.* 2019). Empirical models, on the other hand, use a system's data series for mathematical functions to establish connections between the target variable of the estimate and the input variables in the system, disregarding the intervening physical processes (Uliana *et al.* 2019).

Among the empirical models are the machine learning algorithms, which have been used in basins around the world due to their relatively easy adjustment and practicality, and they are robust tools for analyzing complex systems (Tongal & Booij 2018). Jimeno-Sáez *et al.* (2018) compared the performance of the Soil and Water Assessment Tool (SWAT) conceptual model to the artificial neural networks for daily streamflow forecasting in the Miño-Sil and Segura watersheds, Spain, and concluded that both are efficient tools for forecasting. Further, they concluded that the conceptual model was superior in the lower streamflows forecasting, whereas the empirical model showed better performance in the highest streamflows forecasting. Adnan *et al.* (2018) analyzed the least square support vector machine (LSSVM), fuzzy genetic algorithm, and M5 model tree models in the daily and monthly streamflow forecasting in the Hunza River basin, Pakistan, and concluded that these models are efficient for forecasting in the studied river basin, especially the LSSVM model. Kabir *et al.* (2020) evaluated the wavelet-based artificial neural networks, support vector regression, and deep belief network models for hourly streamflow prediction in three locations in the United Kingdom and observed good performance of the models for a 2-hour forecast horizon.

The use of empirical hydrological models requires the appropriate input variables selection to make the learning process less complex, the interpretation of the results simpler, and reduce the computational cost (Dariane *et al.* 2019). Because of a large amount of hydrometeorological data available for some basins, the variables selection process is essential to eliminate correlated variables, noisy or non-significant relationships with the dependent variable (Jain & Zongker 1997; Bowden *et al.* 2005; Prasad *et al.* 2017; Hadi *et al.* 2019; Zhu *et al.* 2019; Afan *et al.* 2020; Ren *et al.* 2020).

According to the National Water Agency (2009), one of the threats to the Tocantins-Araguaia hydrographic region is the entry of large enterprises, which can pressure the environment if implemented without proper planning. In view of the need to support the planning and management of water resources in the Sono, Manuel Alves da Natividade and Palma River basins, the objectives of this study were to: (i) select the best sets of hydrometeorological variables using selection algorithms, namely genetic algorithm, recursive feature elimination and vsurf, and (ii) evaluate the performance of the random forest, artificial neural networks, support vector regression, and M5 model tree models in forecasting daily streamflow of the basins.

2. METHODS

2.1. Basins and hydrometeorological data

The study hydrographic basins are the Sono (SRB), Manuel Alves da Natividade (MRB), and Palma (PRB), located in the Tocantins-Araguaia hydrographic region, on the right side of the Tocantins River (Figure 1). The Tocantins-Araguaia hydrographic region has an area of approximately 920,087 km^2, which starts in the Center-West region and downstream to the bay of Ilha de Marajó, in the north. It is the largest hydrographic region fully inserted in the Brazilian territory (ANA 2015). Due to its great water availability, which is equivalent to approximately 6% of the national total, the region has potential for hydro-electricity, mining, livestock, irrigation, fishing, agriculture, transport, and tourism. The main consumptive water uses are irrigation and human, animal, and industrial consumption (ANA 2009). The SRB (45,042 km^2), MRB (14,344 km^2), and PRB (17,468 km^2) are important hydrological and environmental units of the upper and middle courses of the Tocantins River. Entirely inserted in the Cerrado biome, they also stand out in non-consumptive water uses for the ecosystem's
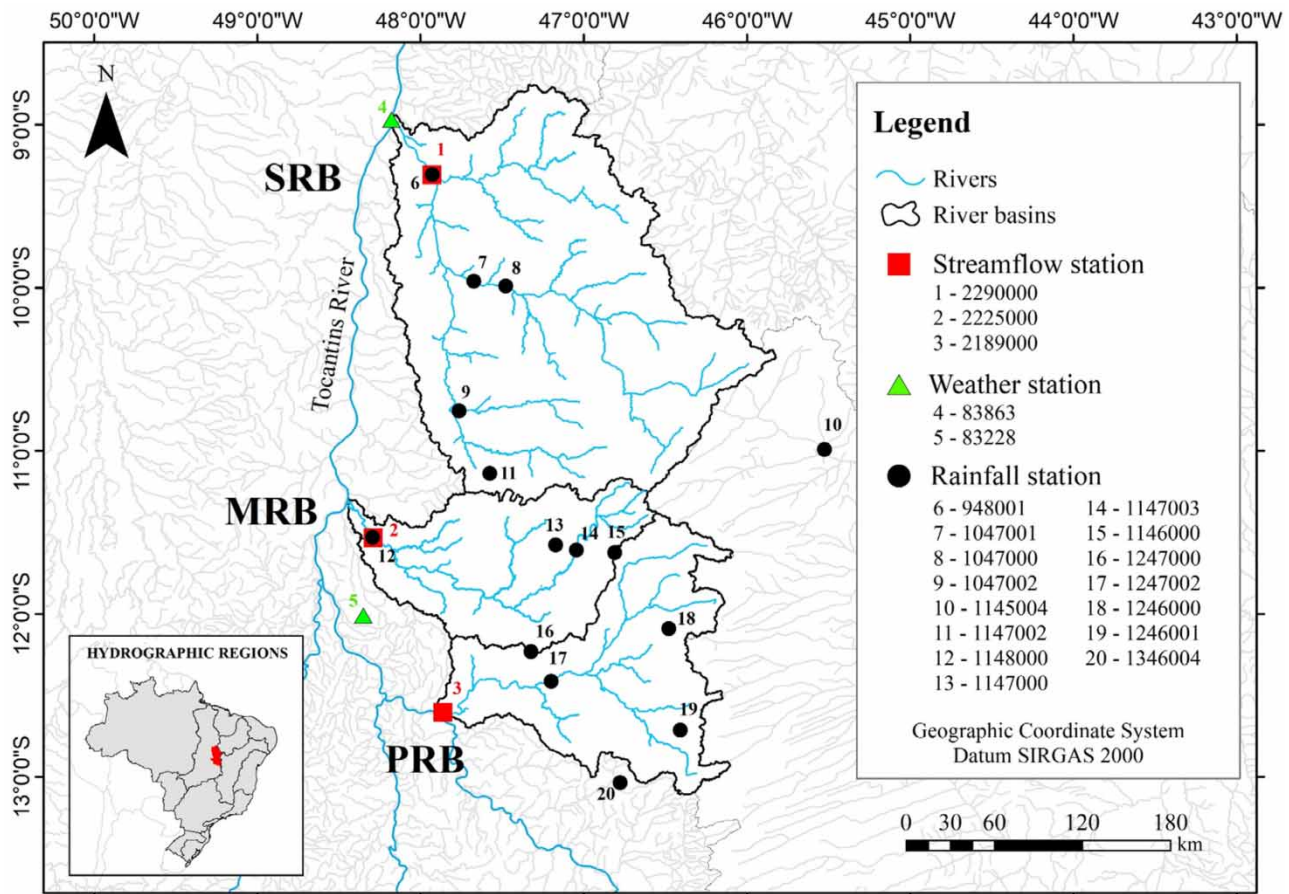

Figure 1 | Geographical location of the Sono (SRB), Manuel Alves da Natividade (MRB), and Palma (PRB) River basins in the Tocantins-Araguaia hydrographic region, Brazil.

conservation, supply of four hydroelectric reservoirs along the Tocantins River (Peixe Angical, Luís Eduardo Magalhães, Estreito, and Tucuruí), and tourism (Mauriz 2008; ANA 2009; Rodrigues *et al.* 2020).

Hydrometeorological data were obtained from the Hydrological Information System (HIDROWEB) of the National Water Agency (ANA) and the Meteorological Database for Teaching and Research (BDMEP) of the National Institute of Meteorology (INMET). The codes and location of the stations used for obtaining the daily streamflow series ($m^3 s^{-1}$), average rainfall (mm), and potential evapotranspiration (mm) are presented in Figure 1. In the gauged stations 1, 2, and 3 streamflow series used in the study of SRB, MRB, and PRB, respectively, were monitored. The spatial average rainfall in the basins was obtained by the Thiessen Polygon method (Macêdo *et al.* 2013), considering the rainfall stations n.6 to n.11, n.13 and n.15 for SRB; n.11 to n.15 for MRB; and n.15 to n.20 for PRB. The weather station data n.4 was used for SRB and n.5 for MRB and PRB for calculating the evapotranspiration by the Penman-Monteith equation (Allen *et al.* 1998).

As input variables in the selection and modeling process the Julian day (JD), ordered from January 1, and the observed values in the previous 3 days of the respective streamflow ($Q_{t-1}, Q_{t-2}, Q_{t-3}$), average rainfall ($R_{t-1}, R_{t-2}, R_{t-3}$) and evapotranspiration ($ET_{t-1}, ET_{t-2}, ET_{t-3}$), totaling 10 variables were considered. The use of lagged variables in time, mainly from the rainfall series ($R_{t-1}, R_{t-2}, R_{t-3}$), is important in empirical models, because, among other reasons, it covers the time of concentration of the basins in the modeling, which for this study are 63.35 (SRB), 37.82 (MRB) and 61.72 hours (PRB) according to the Giandotti equation (Giandotti 1940).

The periods available and used in the evaluations were from 7/1/1977 to 5/28/2017, 12/30/2017, and 9/25/2006 for SRB, MRB, and PRB, respectively. Because of the existence of gaps in some series and the large amount of observed data, dates with missing data were removed from the database, leaving a total of 10,194 (SRB), 12,019 (MRB), and 5830 observations (PRB). The dataset was divided into training (70%) and validation (30%) sets to assess the efficiency of the models. All statistical procedures were performed in the R Statistical Environment (R Development Core Team 2019).

2.2. Feature selection

The feature selection algorithms used (genetic algorithm, recursive feature elimination, and vsurf) belong to the wrapper methods, whose approach consists of the search for subsets of variables that minimize the estimation error of a machine learning algorithm (Kohavi & John 1997). The random forest (RF) machine learning model was used in all feature selection approaches due to its accurate forecasting and resistance to overfitting (Breiman 2001; Cutler *et al.* 2012). The selected algorithms are presented below.

2.2.1. Genetic algorithm (GA)

Inspired by evolutionary biology, the GA for feature selection considers a set of variables to be a chromosome/individual, and a set of individuals to be a population. By a fitness function, the algorithm evaluates and ranks the individuals according to a specific metric. The ordered individuals go through genetic operators (elitism, selection, crossing, and mutation) for creating new individuals that will integrate to a new generation, which will be evaluated and ordered by the fitness function recursively until the stopping criterion is met and the best individual is selected (Holland 1975; Bento & Kagan 2008; Xue *et al.* 2018). The algorithm was fitted using the *gafs* function of the 'caret' package (Kuhn *et al.* 2019), with the RF model as fitness function and the root mean square error (RMSE) as the objective function to be minimized. The internal parameters used in the GA configuration were defined through a trial-and-error process, guided by related studies (Dariane & Azimi 2016, 2018; Afan *et al.* 2020), and the parameters are the number of generations (20), population size (30), and probability of crossing (0.8), mutation (0.1), and elitism (0.0).

2.2.2. Recursive feature elimination (RFE)

The RFE algorithm consists of the retroactive process of feature selection, which occurs in the fitting of a machine learning model, allocation importance to the available features, and removing the least relevant feature in the predictive process. The recursive elimination of features is gradual, starting from the total set of features for a predetermined maximum number (Guyon *et al.* 2002; Granitto *et al.* 2006). The execution of the algorithm was performed using the *rfe* function of the 'caret' package (Kuhn *et al.* 2019), with the maximum number of features regressing from 10 to 1.

2.2.3. Vsurf (VS)

The VS selection algorithm operates in three steps: the first one is the classification and ordering of features as to their importance, according to the results of the RF model, eliminating those considered irrelevant to the process; the second one selects

the features related to the dependent features that lead to lower forecasting error; and in the third step there is the fitting of models with the sequential introduction of the features selected in the previous step, evaluating the reduction of the error for the decision regarding the permanence of each one (Genuer *et al.* 2015). The algorithm was implemented using the 'VSURF' package (Genuer *et al.* 2019), with 500 trees developed in the RF (ntree) and the standard number of features randomly sampled for the growth of each tree (mtry). The selected features in the third step were adopted.

2.3. Description of the models and respective parameters

Machine learning algorithms must be able to generalize their estimates from the input data, and an important step is the appropriate choice of parameters of the model. One of the main methods for selecting the best model with optimization of parameters is the cross-validation, applied in the models training step (Elshorbagy *et al.* 2010; Shortridge *et al.* 2016; Yang *et al.* 2017). The cross-validation k-fold was implemented in this study with the 'caret' package (Kuhn *et al.* 2019), with 10 resamples. The evaluated models were: random forest, artificial neural networks, support vector regression, and M5 model tree.

2.3.1. Random forest (RF)

Introduced by Breiman (2001), the RF is a classification or regression algorithm that aggregates the results of a set of decision trees, grown with random variables/features, independently sampled, and equally distributed. For regression problems, the result of the model is given by the average of the predictions made by all the trees in the forest (Breiman 2001). The RF models were fitted using the *rf* function of the 'randomForest' package (Liaw & Wiener 2002), in which the optimized parameter was mtry, which varied from 2 to 10, with ntree equal to 500.

2.3.2. Artificial neural networks (ANN)

ANN are algorithms inspired by the human nervous system. The basic structure has an input layer containing the independent variables, one or more hidden layers for processing the data, and an output layer with the results of the iterations. The neurons are interconnected by weights so that the receiving neuron aggregates the weights of the previous layer, adds a bias, and forwards the result through a transfer function (Dastorani *et al.* 2018; Dariane *et al.* 2019). To fit the ANN, the *nnet* function of the 'nnet' package were used (Venables & Ripley 2002) with feedforward multilayer perceptron architecture, one intermediate layer, and linear output units. The input data were normalized by the min-max method to be between 0 and 1 (Riad *et al.* 2004; Selvi & Huseyinov 2020). The optimized parameters were the number of neurons in the hidden layer (size), from 1 to 10, and the decay of the weights (decay), used mainly to avoid overfitting the model to the data, considering 0.00, 0.01, 0.05, 0.10, 0.50, 1.00, and 2.00.

2.3.3. Support vector regression (SVR)

The SVR, introduced as support vector machine for classification by Vapnik (1995), is the projection of the features in high-dimensional space to map data sets by fitting a curve (kernel function) between two marginal hyperplanes, to minimize the regression error. Among the available kernel functions are linear, polynomial, radial, and sigmoid (Vapnik 1995; Yang *et al.* 2017; Dastorani *et al.* 2018; Shamshirband *et al.* 2020). The function used to fit the SVR was the *svmRadialSigma* from the 'kernlab' package (Karatzoglou *et al.* 2004), in which the kernel function used is radial and the parameters regularization (C) and kernel function (sigma) are required. The values for C optimization ranged from 1 to 20 and were tested for sigma $1\ 10^{-5}$, $1\ 10^{-4}$, $1\ 10^{-3}$, $1\ 10^{-2}$, and $1\ 10^{-1}$.

2.3.4. M5 model tree (M5T)

The M5T algorithm, proposed by Quinlan (1992), presents a tree structure developed in two stages. The first step consists of growing a decision tree using independent variables to homogenize the responses to the variable of interest, using the reduction of the standard deviation as a criterion for dividing the nodes. Successive divisions generate excessive branches, which may cause overfitting of the data. Therefore, the second step is pruning the tree grown by replacing the branches with linear regression functions according to the independent variables used in the divisions of the pruned branch. Thus, for each homogenized subspace of the target variable, a linear model is fitted (Quinlan 1992; Pal & Deswal 2009; Shamshirband *et al.* 2020). The 'Cubist' package (Kuhn *et al.* 2020) was used to fit the models, with the number of interactive model trees (committees) ranging from 10 to 100 adding 10.

2.4. Performance metrics

To evaluate the performance of the models, the root mean square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe efficiency index (NSE), and its logarithmic version (LNSE) were used as objective functions, according to the equations presented below. The models were also compared using hydrographs and flow duration curves (FDC):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{2}$$

$$NSE = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{3}$$

$$LNSE = 1 - \frac{\sum_{i=1}^{n}[log(y_t) - log(\hat{y}_t)]^2}{\sum_{i=1}^{n}[log(y_t) - log(\bar{y})]^2} \tag{4}$$

where y_t is the observed streamflow at time t, $\hat{y}_t$ is the predict streamflow at time t, $\bar{y}$ is the average observed streamflow, and n is the total number of observations in the series analyzed.

The NSE proposed by Nash & Sutcliffe (1970) is one of the main metrics for evaluating the performance of hydrological models. The classification of the performance of the fitted models with data on a daily scale proposed by Moriasi *et al.* (2015), suggests: NSE>0.80 as 'very good'; 0.70<NSE≤0.80 'good'; 0.50<NSE≤0.70 'satisfactory'; and NSE≤0.50 as 'not satisfactory'.

3. RESULTS AND DISCUSSION

Table 1 presents the sets of total independent variables/features and the selected ones by the selection algorithms for each basin, with the latter being ordered in descending order according to the importance attributed to them during the selection processes. The VS algorithm selected a smaller number of variables, followed by the GA, which removed one variable from the total set of the SRB and MRB and four from the PRB, and RFE which selected nine in each basin. The variables Q_{t-1} and

Table 1 | Sets of total variables and selected by the genetic algorithm (GA), recursive feature elimination (RFE), and vsurf (VS), in decreasing order according to the importance attributed by each selection algorithm (SA), for hydrological modeling in the Sono (SRB), Manuel Alves da Natividade (MRB), and Palma (PRB) River basins

Basin	SA	Input variables	Total
SRB	Total	JD, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_{t-1}, R_{t-2}, R_{t-3}, ET_{t-1}, ET_{t-2}, ET_{t-3}	10
	GA	Q_{t-1}, Q_{t-2}, Q_{t-3}, JD, R_{t-1}, R_{t-2}, R_{t-3}, ET_{t-1}, ET_{t-3}	9
	RFE	Q_{t-1}, Q_{t-2}, JD, Q_{t-3}, R_{t-1}, R_{t-2}, ET_{t-1}, ET_{t-2}, R_{t-3}	9
	VS	Q_{t-1}, Q_{t-2}, JD, R_{t-1}	4
MRB	Total	JD, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_{t-1}, R_{t-2}, R_{t-3}, ET_{t-1}, ET_{t-2}, ET_{t-3}	10
	GA	Q_{t-1}, Q_{t-2}, Q_{t-3}, R_{t-1}, R_{t-2}, R_{t-3}, ET_{t-1}, ET_{t-2}, ET_{t-3}	9
	RFE	Q_{t-1}, Q_{t-2}, R_{t-1}, Q_{t-3}, JD, ET_{t-2}, ET_{t-1}, R_{t-2}, ET_{t-3}	9
	VS	Q_{t-1}, Q_{t-2}, JD, R_{t-1}, R_{t-3}, ET_{t-1}	6
PRB	Total	JD, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_{t-1}, R_{t-2}, R_{t-3}, ET_{t-1}, ET_{t-2}, ET_{t-3}	10
	GA	Q_{t-1}, Q_{t-3}, R_{t-1}, JD, ET_{t-1}, ET_{t-2}	6
	RFE	Q_{t-1}, Q_{t-2}, Q_{t-3}, JD, R_{t-1}, ET_{t-3}, R_{t-2}, ET_{t-1}, R_{t-3}	9
	VS	Q_{t-1}, Q_{t-2}, R_{t-1}, ET_{t-1}	4

JD, Julian day; and Q_{t-n}, R_{t-n}, ET_{t-n}, streamflow (m³ s⁻¹), average rainfall (mm), and reference evapotranspiration (mm) n days before the forecasting date, respectively.

P_{t-1} were selected by all algorithms, followed by Q_{t-2} and ET_{t-1}, selected in eight of the nine selections (3 basins$\times$3 selection), and JD selected in seven. The variables that were more frequently removed were ET_{t-2} and ET_{t-3}.

The most important variables by most algorithms were the Q_{t-1} and Q_{t-2} streamflows, with only the selection made by the GA in the PRB presenting the Q_{t-3} as the second most important variable. Previous streamflows are widely used in empirical hydrological models due to their strong correlations with the target streamflow (Yang *et al.* 2017; Liu *et al.* 2019; Yaseen *et al.* 2019; Zhu *et al.* 2019). Following the streamflows, overall, the JD and the lagged average rainfall were the most important, since the JD temporarily marks the changes in the streamflow, and rainfall provides the main water entry into the basins (Dinpashoh *et al.* 2019).

The model parameters, optimized during the cross-validation process, are shown in Table 2, and the results of the performance of the models using different sets of variables, in the training and validation periods, for the SRB, MRB, and PRB are present in Tables 3–5, respectively. The models' performances were classified as very good, according to the classification proposed by Moriasi *et al.* (2015), with NSE ranging from 0.849 to 0.926 in the validation. Only the ANN model, fitted with the set of variables selected by VS for PRB presented a performance classified as good (0.769). The results of the NSE elucidate the precision of the models in the peak flows forecasting, which has been a challenge for hybrid hydrological models, such as SWAT, LASH, MGB-IPH among others. Beyond this, the LNSE values ranged from 0.782 to 0.969, revealing a good accuracy in the forecasting of the baseflow in the basins.

The models presented better performance when fitted with all available variables or with those selected by the GA and RFE algorithms, corroborating the increasing use of selection algorithms in hydrological studies. Dariane & Azimi (2016) evaluated the GA to select input variables in the ANN model, which was applied for forecasting the monthly streamflows in the Ajichai sub-basin, Iran, and obtained superior performance using the selected variables (NSE=0.870) instead of all the variables (NSE=0.780). Khan *et al.* (2020), for forecasting droughts in Pakistan, used the RFE to select variables for the ANN, support vector machine, and k-nearest neighbor models and obtained results considered satisfactory. Dariane *et al.* (2019) using ANN for forecasting rainfall at six stations in Iran, evaluating the selection methods GA, wingamma, and self-organized map, concluded that the GA is the most reliable for input feature selection.

According to the results presented in Tables 3–5, the performance of the M5T model using the set of variables selected by the RFE was superior to the other models for SRB and PRB, with lower RMSE (99.254 and 59.671 m^3 s^{-1}) and higher NSE

Table 2 | Optimized parameters of the random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA), recursive feature elimination (RFE), and vsurf (VS) for the Sono (SRB), Manuel Alves da Natividade (MRB), and Palma (PRB) River basins

| | RF | | ANN | | SVR | | M5T |
Set of variables	*mtry*	*ntree*	*size*	*decay*	*cost*	*sigma*	*committees*
SRB							
Total	6	500	1	0.0	20	$1\ 10^{-3}$	50
GA	5	500	6	0.0	9	$1\ 10^{-2}$	80
RFE	5	500	1	0.0	18	$1\ 10^{-3}$	60
VS	2	500	6	0.0	20	$1\ 10^{-2}$	40
MRB							
Total	6	500	4	1.0	6	$1\ 10^{-2}$	90
GA	5	500	7	0.5	9	$1\ 10^{-2}$	30
RFE	5	500	10	1.0	6	$1\ 10^{-2}$	80
VS	4	500	4	0.5	6	$1\ 10^{-2}$	90
PRB							
Total	8	500	4	0.0	15	$1\ 10^{-2}$	70
GA	3	500	9	0.0	17	$1\ 10^{-2}$	60
RFE	6	500	7	0.0	20	$1\ 10^{-2}$	100
VS	2	500	3	0.0	20	$1\ 10^{-2}$	40

Table 3 | Performance of the random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA), recursive feature elimination (RFE), and vsurf (VS) in the streamflow forecasting in the Sono River basin

Model	Set of variables	Training (07/1977–09/2004)				Validation (09/2004–05/2017)			
		RMSE	MAE	NSE	LNSE	RMSE	MAE	NSE	LNSE
RF	Total	147.923	66.247	0.928	0.960	121.415	60.721	0.889	0.916
	GA	146.483	65.863	0.929	0.960	120.981	60.654	0.890	0.914
	RFE	146.756	65.681	0.929	0.960	121.595	60.956	0.889	0.915
	VS	149.996	68.365	0.926	0.958	122.519	63.982	0.887	0.908
ANN	Total	143.861	66.965	0.932	0.956	103.046	53.124	0.920	0.949
	GA	143.169	65.336	0.932	0.958	108.200	67.864	0.912	0.925
	RFE	143.729	67.591	0.932	0.955	103.123	53.634	0.920	0.944
	VS	142.115	64.400	0.933	0.959	129.661	95.562	0.874	0.843
SVR	Total	140.105	66.368	0.935	0.957	100.422	55.144	0.924	0.948
	GA	136.128	64.026	0.939	0.959	100.639	55.471	0.924	0.948
	RFE	140.404	66.257	0.935	0.957	100.361	54.343	0.924	0.949
	VS	143.963	67.897	0.932	0.956	104.608	57.842	0.918	0.945
M5T	Total	135.366	59.307	0.940	0.963	99.131	46.794	0.926	0.959
	GA	135.096	59.393	0.940	0.963	99.277	46.822	0.926	0.959
	RFE	134.045	58.878	0.941	0.963	99.254	46.705	0.926	0.959
	VS	142.258	62.681	0.933	0.960	102.942	48.939	0.920	0.957

RMSE, root mean square error ($m^3 \, s^{-1}$); MAE, mean absolute error ($m^3 \, s^{-1}$); NSE, Nash-Sutcliffe efficiency index; LNSE, Nash-Sutcliffe efficiency index logarithmic.

Table 4 | Performance of the random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA), recursive feature elimination (RFE), and vsurf (VS) in the streamflow forecasting in the Manuel Alves da Natividade River basin

Model	Set of variables	Training (07/1977–08/2005)				Validation (08/2005–12/2017)			
		RMSE	MAE	NSE	LNSE	RMSE	MAE	NSE	LNSE
RF	Total	68.556	28.425	0.928	0.973	61.954	28.470	0.885	0.909
	GA	68.679	28.553	0.928	0.971	58.967	24.850	0.896	0.956
	RFE	69.068	28.515	0.927	0.973	61.948	28.429	0.885	0.910
	VS	69.693	28.856	0.926	0.972	62.855	28.505	0.882	0.910
ANN	Total	80.578	36.345	0.901	0.949	67.316	28.061	0.865	0.939
	GA	76.341	34.996	0.911	0.945	64.777	25.090	0.875	0.964
	RFE	80.663	36.682	0.900	0.947	67.477	27.960	0.864	0.941
	VS	76.373	34.375	0.911	0.954	65.049	27.050	0.874	0.934
SVR	Total	63.355	31.251	0.939	0.942	58.134	30.269	0.899	0.881
	GA	63.104	30.867	0.939	0.944	58.372	29.165	0.898	0.899
	RFE	63.771	31.691	0.938	0.939	58.679	30.742	0.897	0.872
	VS	65.909	33.388	0.934	0.928	60.470	32.559	0.891	0.846
M5T	Total	56.327	23.690	0.951	0.976	58.039	22.857	0.899	0.969
	GA	56.617	23.893	0.951	0.976	57.994	22.954	0.900	0.969
	RFE	57.223	24.078	0.950	0.975	57.999	22.951	0.900	0.969
	VS	58.903	24.700	0.947	0.974	59.528	23.187	0.894	0.969

RMSE, root mean square error ($m^3 \, s^{-1}$); MAE, mean absolute error ($m^3 \, s^{-1}$); NSE, Nash-Sutcliffe efficiency index; LNSE, Nash-Sutcliffe efficiency index logarithmic.

(0.926 and 0.915) and LNSE (0.959 and 0.964) in the validation periods, respectively. For MRB, the total set of input variables provided the best fit of the M5T model, considered the best for the basin (RMSE=58.039 $m^3 \, s^{-1}$, MAE=22.857 $m^3 \, s^{-1}$, NSE=0.899 and LNSE=0.969).

Yin *et al.* (2018) fitted the SVR, multivariate adaptive regression splines, and M5T models to streamflow forecasting in the Pailugou River basin, China. They concluded that the M5T model was superior to the others, with an average NSE of 0.890 in

Table 5 | Performance of the random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA), recursive feature elimination (RFE), and vsurf (VS) in the streamflow forecasting in the Palma River basin

Model	Set of variables	Training (07/1977–05/1999)				Validation (05/1999–09/2006)			
		RMSE	MAE	NSE	LNSE	RMSE	MAE	NSE	LNSE
RF	Total	86.295	28.316	0.922	0.958	62.687	24.671	0.906	0.954
	GA	85.855	28.524	0.923	0.960	63.552	25.589	0.903	0.957
	RFE	86.711	28.457	0.922	0.958	61.347	24.331	0.910	0.956
	VS	86.659	29.257	0.922	0.957	62.392	24.743	0.907	0.949
ANN	Total	84.351	31.464	0.926	0.947	72.497	50.712	0.874	0.811
	GA	84.406	32.818	0.926	0.947	77.110	51.321	0.858	0.823
	RFE	81.714	29.990	0.930	0.951	79.465	54.016	0.849	0.806
	VS	80.246	30.536	0.933	0.953	98.263	52.479	0.769	0.782
SVR	Total	74.401	27.999	0.942	0.958	59.729	27.313	0.915	0.953
	GA	79.298	31.206	0.935	0.953	60.540	30.234	0.912	0.945
	RFE	74.799	28.194	0.942	0.957	59.738	27.520	0.915	0.952
	VS	80.316	31.797	0.933	0.951	61.066	29.601	0.911	0.945
M5T	Total	69.306	23.565	0.950	0.967	59.880	22.593	0.914	0.964
	GA	71.986	24.473	0.946	0.965	61.413	23.641	0.910	0.962
	RFE	70.282	23.809	0.949	0.967	59.671	22.642	0.915	0.964
	VS	74.716	25.273	0.942	0.964	60.596	23.153	0.912	0.962

RMSE, root mean square error (m^3 s^{-1}); MAE, mean absolute error (m^3 s^{-1}); NSE, Nash-Sutcliffe efficiency index; LNSE, Nash-Sutcliffe efficiency index logarithmic.

the validation. Diverging from this study, Adnan *et al.* (2019) observed, when evaluating six models in the daily forecasting of the monitored streamflows in two stations on the Fujiang River, China, lower performance of the M5T model, with average NSE values of 0.652 and 0.670, evidencing the need to study the models for different basins. Tongal & Booij (2018) verified the accuracy of the SVR, ANN, and RF models for four river basins in the United States, and they observed that the NSE varied between 0.880 and 0.980 in the validation, which was evaluated as very good by the researchers and similar to those obtained in this study. Liu *et al.* (2019) fitted the Gaussian mixture regression, SVM, and ANN models for forecasting streamflows in two sections of the Jinsha River basin, China, and obtained NSE indexes between 0.850 and 0.860, corroborating the results found for the river basins in this study. Regarding the above results, the good performance of statistical models fitted in this study was confirmed, especially the M5T model.

Figure 2 shows the hydrographs observed and simulated by the models using the variables that provided the least statistical errors for the SRB, MRB, and PRB basins in the validation. Good adherence of the forecasted hydrographs to those observed can be seen, mainly in the periods of streamflow recession. According to Rodrigues *et al.* (2021), the best models for these periods should be based on the relationship between the groundwater flow, which is predominant in the recession periods, and the discharge from the aquifer. Peak flows were underestimated by the models, mainly due to the difficulty in capturing the many variables involved in the streamflow process, the difficult representation of the spatial variability of rainfall over the basin due to the low density of rain-gauge stations in the basins, and even errors from the extrapolation of the stage-curve of the river (Viola *et al.* 2009; Silva Neto *et al.* 2020).

Figure 3 presents the flow duration curves (FDC) of the observed streamflows and forecasted by the models using the variables that provided the best fitting for each river basin, considering the validation period. This analysis allows inferring about the percentage of time in which a given streamflow is exceeded or equalized and to establish reference values for water management and projects. The FDC forecasted by the M5T and ANN models presented a great agreement with the observed FDC for the SRB and MRB, as well as by the M5T and RF models for the PRB. The inferior performances were observed for the RF and SVR models for MRB, and ANN for PRB, mainly with overestimates of the lowest streamflows.

Table 6 shows the observed streamflows with 95% (Q_{95}), 90% (Q_{90}), 50% (Q_{50}), 10% (Q_{10}), and 5% (Q_5) of exceedance and forecasted by the models in the validation periods. The Q_{95} and Q_{90} streamflows indicate the values that are equaled or exceeded for 95 and 90% of the time, respectively, and have been used as a reference for water rights in several Brazilian
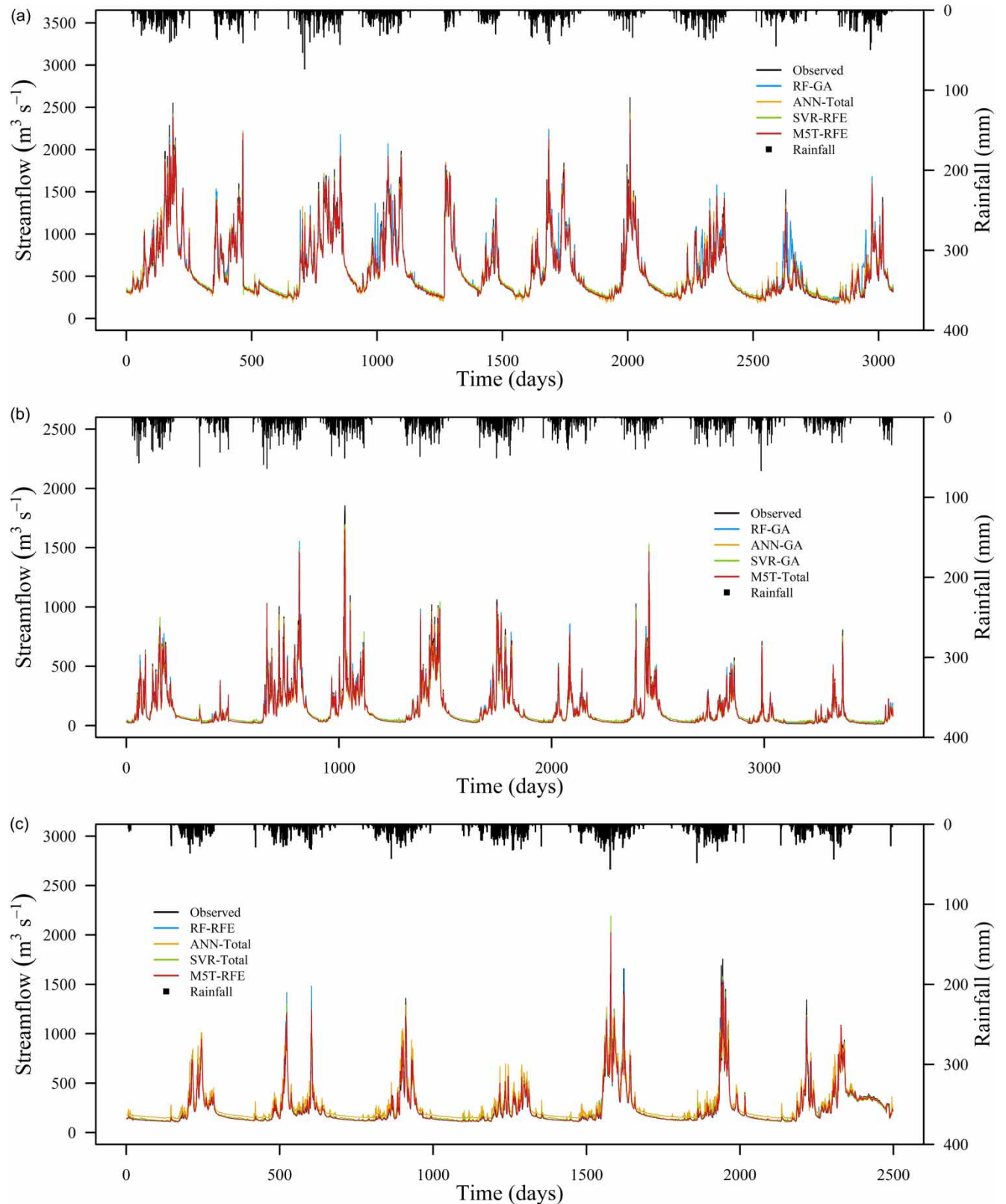
Figure 2 | Hydrographs of the observed and forecasted streamflows by the random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA) and recursive feature elimination (RFE), validation periods, for the Sono (a), Manuel Alves da Natividade (b) and Palma (c) River basins.
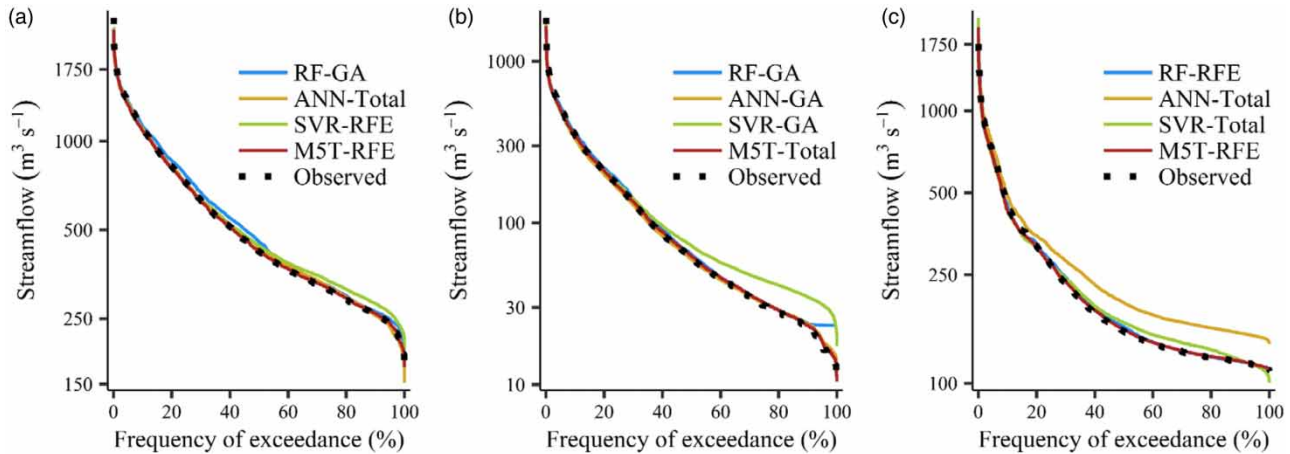
Figure 3 | Flow duration curves on a logarithmic scale of the observed streamflows and forecasted by the random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA) and recursive feature elimination (RFE) for the Sono (a), Manuel Alves da Natividade (b) and Palma (c) River basins, validation periods.

Table 6 | Observed streamflows ($m^3 s^{-1}$) with 95% (Q_{95}), 90% (Q_{90}), 50% (Q_{50}), 10% (Q_{10}), and 5% (Q_5) of permanence and percentage errors of the streamflows forecasted by random forest (RF), artificial neural networks (ANN), support vector regression (SVR), and M5 model tree (M5T) models, fitted with the total set of variables and sets selected by the genetic algorithm (GA) and recursive feature elimination (RFE), validation periods, for the Sono (A), Manuel Alves da Natividade (B) and Palma (C) River basins

Permanence streamflow	SRB				
	Observed	RF-GA	ANN-Total	SVR-RFE	M5T-RFE
Q_{95}	241.44	2.97%	−3.18%	8.40%	0.07%
Q_{90}	258.27	2.19%	−0.76%	9.16%	0.73%
Q_{50}	421.05	8.11%	3.36%	4.72%	0.51%
Q_{10}	1100.89	1.85%	0.14%	−1.62%	−1.67%
Q_5	1365.69	−0.34%	−2.44%	−2.22%	−3.14%
	MRB				
	Observed	RF-GA	ANN-GA	SVR-GA	M5T- Total
Q_{95}	17.02	36.90%	10.69%	83.20%	7.64%
Q_{90}	22.78	4.87%	4.35%	52.81%	1.67%
Q_{50}	60.49	5.47%	1.42%	19.74%	2.35%
Q_{10}	353.60	1.90%	−6.45%	−2.45%	−1.57%
Q_5	532.73	−1.76%	−5.72%	−7.95%	−6.45%
	PRB				
	Observed	RF- RFE	ANN-Total	SVR-Total	M5T-RFE
Q_{95}	116.00	1.07%	28.53%	0.45%	0.84%
Q_{90}	119.85	0.11%	27.43%	1.26%	0.82%
Q_{50}	155.20	3.86%	27.01%	7.49%	1.17%
Q_{10}	449.50	0.26%	8.57%	−4.76%	−4.04%
Q_5	696.80	−1.97%	6.30%	−5.19%	−3.22%

states and for planning irrigation systems, dams, and hydroelectric plants since they represent the minimum flows (recession period). The estimated Q_{95} were close to those observed for the SRB and PRB, with errors ranging from 0.07 to 28.53%, while for MRB the best forecasts were obtained by the ANN and M5T models, with errors of 10.69 and 7.64%, respectively.

According to Decree No. 2432 of June 6, 2005, the Q_{90} streamflow is the reference used for granting in the Tocantins state. The streamflow forecast with 90% of exceedance was performed satisfactorily by the models, highlighting M5T. Such value was forecasted for SRB as equal to 260.15 m^3 s^{-1} (observed=258.27 m^3 s^{-1}) and for MRB, 23.16 m^3 s^{-1} (22.78 m^3 s^{-1}). The RF model forecasted this streamflow as 119.98 m^3 s^{-1} for PRB (observed=119.85 m^3 s^{-1}). These results, together with those found for Q_{95}, showed the precision of the streamflow forecasting by the models for the recession periods and its potential use as a tool in the management of water resources.

The other reference streamflows (Q_{50}, Q_{10}, and Q_5) represent the values with medium and highest values (low durations) in the basins. These streamflows (Q_{10} and Q_5) were underestimated by most models; however, the forecasts for this complex hydrological process were satisfactory, with errors between 0.14 and 8.57% for Q_{10} and −0.34 and −7.95% for Q_5. The forecast for the Q_{50} streamflow resulted in average errors of 4.18% (SRB), 7.25% (MRB), and 9.88% (PRB), with better performances of the M5T model for the SRB and PRB, and ANN for the MRB. Accurate forecasting of peak flows is necessary to plan actions to mitigate floods and support the drainage and dam projects (Tucci 2004).

Rodrigues et al. (2021) evaluated the performances of the SWAT and ANN models for forecasting daily streamflow of the MRB and obtained NSE indexes in the validation step of 0.610 and 0.910, respectively, with errors of −14.10% in estimating Q_{90} using SWAT and −10.60% by ANN. For the same basins, Rodrigues et al. (2020) found, by the SWAT model, NSE values of 0.730 (SRB), 0.810 (MRB), and 0.700 (PRB) in the validation period. In view of these results for the same basins in this study, the good performance of the evaluated models was confirmed, since they presented results superior to those of the conceptual SWAT model and values close to that of the empirical ANN, mainly the M5T model using the sets of variables selected by the RFE for SRB and PRB and total set for MRB. However, although machine learning algorithms showed good streamflow simulation capabilities, the models are based on pattern recognition and do not consider the hydrological processes involved. Therefore, although new data is incorporated, it only operates according to previously learned patterns and is not recommended for situations in which there are alterations of the physical characteristics of the basin (Rodrigues et al. 2021).

The analyzed models proved to be suitable for application in the management and planning of water resources. Considering the short forecast horizon (1 day) the models can be useful tools in the operation of hydraulic works, river navigation, irrigation, hydroelectric power generation, water supply, flood control, among other purposes requiring short-time streamflow predictions in the SRB, MRB, and PRB basins.

4. CONCLUSIONS

The random forest, artificial neural networks, support vector regression, and M5 model tree models, fitted with the total set of variables and sets selected by the genetic algorithm, recursive feature elimination, and vsurf, showed good performances in the daily streamflow forecasting of the Sono, Manuel Alves da Natividade and Palma River basins.

The performance of the M5 model tree model stood out from that of the other models in the calibration and validation periods using the variables selected by the recursive feature elimination algorithm for the Sono (Q_{t-1}, Q_{t-2}, JD, Q_{t-3}, R_{t-1}, R_{t-2}, ET_{t-1}, ET_{t-2}, R_{t-3}) and Palma (Q_{t-1}, Q_{t-2}, Q_{t-3}, JD, P_{t-1}, ET_{t-3}, R_{t-2}, ET_{t-1}, R_{t-3}) River basins, and total set of variables for the Manuel Alves da Natividade River basin (JD, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_{t-1}, R_{t-2}, R_{t-3}, ET_{t-1}, ET_{t-2}, ET_{t-3}), considered the best sets of variables for fitting the M5 model tree algorithm in the studied basins.

The hydrographs, FDC, and forecasted reference streamflows confirmed the satisfactory performance of the models, considering the agreement between the hydrographs (observed and forecasted) and the forecasted and observed FDC, and the small estimation errors in the peak and recession streamflows periods. In view of these results, it can be concluded that the evaluated models are potential tools to assist in the different demands faced by the water resources management in the river basins, especially the M5 model tree model using streamflow lags, average rainfall, and evapotranspiration as inputs.

The main limitation of this study was the difficult representation of the spatial variability of rainfall over the basin due to the low density of rain-gauge stations in the basins. Due to the successful application of the machine learning algorithms in this study, it is suggested for further studies cover a longer forecast horizon.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Adnan, R. M., Yuan, X., Kisi, O., Adnan, M. & Mehmood, A. 2018 Stream flow forecasting of poorly gauged mountainous watershed by least square support vector machine, fuzzy genetic algorithm and M5 model tree using climatic data from nearby station. *Water Resources Management* **32**, 4469–4486.

Adnan, R. M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B. & Kisi, O. 2019 Daily streamflow prediction using optimally pruned extreme learning machine. *Journal of Hydrology* **577**, 123981.

Afan, H. A., Allawi, M. F., El-Shafie, A., Yaseen, Z. M., Ahmed, A. N., Malek, M. A., Koting, S. B., Salih, S. Q., Mohtar, W. H. W., Lai, S. H., Sefelnasr, A., Sherif, M. & El-Shafie, A. 2020 Input attributes optimization using the feasibility of genetic nature inspired algorithm: application of river flow forecasting. *Scientific Reports* **10**, 4684.

Agência Nacional de Águas – ANA 2009 *Plano estratégico de recursos hídricos dos rios Tocantins e Araguaia: relatório síntese (Strategic Plan for the Water Resources of the Tocantins and Araguaia Rivers: Summary Report)*. Brasília.

Agência Nacional de Águas – ANA 2015 *Conjuntura dos recursos hídricos no Brasil: regiões hidrográficas brasileiras (Situation of Water Resources in Brazil: Brazilian Hydrographic Regions)*. Brasília.

Agência Nacional de Águas – ANA 2020 *Conjuntura dos recursos hídricos no Brasil 2020: informe anual (Situation of Water Resources in Brazil 2020: Annual Report)*. Brasília.

Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. 1998 *Crop Evapotranspiration – Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. United Nations FAO, Rome.

Bento, E. P. & Kagan, N. 2008 Algoritmos genéticos e variantes na solução de problemas de configuração de redes de distribuição (Genetic algorithms and variants in the solution of distribution network configuration problems). *Sba: Controle & Automação* **19**, 302–315.

Bourdin, D. R., Fleming, S. W. & Stull, R. B. 2012 Streamflow modelling: a primer on applications, approaches and challenges. *Atmosphere-Ocean* **50**, 507–536.

Bowden, G. J., Maier, H. R. & Dandy, G. C. 2005 Input determination for neural network models in water resources applications: part 2 - case study: forecasting salinity in a river. *Journal of Hydrology* **301** (1/4), 93–107.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Cantelle, T. D., Lima, E. C. & Borges, L. A. C. 2018 Panorama dos recursos hídricos no mundo e no Brasil (Survey of hydric resouces worldwide and in Brazil). *Revista em Agronegócio e Meio Ambiente* **11**, 1259–1282.

Cutler, A., Cutler, D. R. & Stevens, J. R. 2012 Random forests. In: Zhang, C. & Ma, Y. (eds). *Ensemble Machine Learning*. Springer, New York, pp. 157–175.

Dariane, A. B. & Azimi, S. 2016 Forecasting streamflow by combination of a genetic input selection algorithm and wavelet transforms using ANFIS models. *Hydrological Sciences Journal* **61**, 585–600.

Dariane, A. B. & Azimi, S. 2018 Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection. *Journal of Hydroinformatics* **20**, 520–532.

Dariane, A. B., Gol, M. A. & Karami, F. 2019 Forecasting of rainfall using different input selection methods on climate signals for neural network inputs. *Journal of Hydraulic Structures* **5**, 42–59.

Dastorani, M. T., Mahjoobi, J., Talebi, A. & Fakhar, F. 2018 Application of machine learning approaches in rainfall-runoff modeling (case study: Zayadeh_Rood Basin in Iran). *Civil Engineering Infrastructures Journal* **51**, 293–310.

Debastiani, A. B., Rafaeli Neto, S. L. & Dalagnol, R. 2019 Árvore modelo frente a uma rede neural artificial para a modelagem chuva-vazão (Model tree in comparison to artificial neural network for rainfall-runoff modeling). *Nativa* **7**, 527–534.

Dinpashoh, Y., Singh, V. P., Biazar, S. M. & Kavehkar, S. 2019 Impact of climate change on streamflow timing (case study: Guilan Province). *Theoretical and Applied Climatology* **138**, 65–76.

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. P. 2010 Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – part 1: concepts and methodology. *Hydrology and Earth System Sciences* **14**, 1931–1941.

Genuer, R., Poggi, J. M. & Tuleau-Malot, C. 2015 VSURF: an R package for variable selection using random forests. *R Journal* **7**, 19–33.

Genuer, R., Poggi, J. M. & Tuleau-Malot, C. 2019 *VSURF: Variable Selection Using Random Forests*. R package version 1.1.0.

Giandotti, M. 1940 Previsione empirica delle piene in base alle precipitazioni meteoriche, alle caratteristiche fisiche e morfologiche dei bacini; Applicazione del metodo ad alcuni bacini dell'Appennino Ligure (Empirical flood forecasting based on the meteoric precipitations, the physical and morphological characteristics of the basins; application of the method to some of the Ligurian basins). *Memorie e Studi Idrografici* **10**, 5–13.

Granitto, P. M., Furlanello, C., Biasioli, F. & Gasperi, F. 2006 Recursive feature elimination with random forest for PTR-MS analysis of agro-industrial products. *Chemometrics and Intelligent Laboratory Systems* **83** (2), 83–90.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. 2002 Gene selection for cancer classification using support vector machines. *Machine Learning* **46** (1), 389–422.

Hadi, S. J., Abba, S. I., Sammen, S. S., Salih, S. Q., Al-Ansari, N. & Yaseen, Z. M. 2019 Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. *IEEE Access* **7**, 141533–141548.

Holland, J. H. 1975 *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Jain, A. & Zongker, D. 1997 Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** (2), 153–158.

Jimeno-Sáez, P., Senent-Aparicio, J., Pérez-Sánchez, J. & Pulido-Velazquez, D. 2018 A comparison of SWAT and ANN models for daily runoff simulation in different climatic zones of Peninsular Spain. *Water* **10**, 192.

Kabir, S., Patidar, S. & Pender, G. 2020 Investigating capabilities of machine learning techniques in forecasting stream flow. *Proceedings of the Institution of Civil Engineers: Water Management* **173**, 69–86.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. 2004 kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software* **11** (9), 1–20.

Khan, N., Sachindra, D. A., Shahid, S., Ahmed, K., Shiru, M. S. & Nawaz, N. 2020 Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources* **139**, 103562.

Kohavi, R. & John, G. H. 1997 Wrappers for feature subset selection. *Artificial Intelligence* **97** (1/2), 273–324.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kentel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. & Hunt, T. 2019 caret: *Classification and Regression Training*. R package version 6.0-84.

Kuhn, M., Weston, S., Keefer, C., Coulter, N. & Quinlan, R. 2020 *Cubist: Rule-and Instance-Based Regression Modeling*. R package version 0.2.3.

Liaw, A. & Wiener, M. 2002 Classification and regression by randomForest. *R News* **2** (3), 18–22.

Liu, Y., Ye, L., Qin, H., Ouyang, S., Zhang, Z. & Zhou, J. 2019 Middle and long-term runoff probabilistic forecasting based on Gaussian mixture regression. *Water Resources Management* **33**, 1785–1799.

Macêdo, M. N. C., Dias, H. C. T., Coelho, F. M. G., Araújo, E. A., Souza, M. L. H. & Silva, E. 2013 Precipitação pluviométrica e vazão da bacia hidrográfica do Riozinho do Rôla, Amazônia Ocidental (Rainfall and flow of the Riozinho do Rôla Basin on Western Amazon). *Revista Ambiente & Água* **8**, 206–221.

Mauriz, T. V. M. 2008 *Análise do inventário hidroelétrico da bacia do rio do sono – TO, subsídio para identificação de variáveis socioambientais aplicadas na construção de um modelo de inventário hidrelétrico dinâmico (Analysis of the Hydroelectric Inventory of the Rio do Sono Basin – TO, Subsidy for the Identification of Socio-Environmental Variables Applied in the Construction of a Dynamic Hydroelectric Inventory Model)*. MSc Thesis, Environmental Planning and Management, Catholic University of Brasília, Brasília.

Moriasi, D. N., Gitau, M. W., Pai, N. & Daggupati, P. 2015 Hydrologic and water quality models: performance measures and evaluation criteria. *Transactions of the ASABE* **58**, 1763–1785.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10** (3), 282–290.

Pal, M. & Deswal, S. 2009 M5 model tree-based modelling of reference evapotranspiration. *Hydrological Processes* **23**, 1437–1443.

Prasad, R., Deo, R. C., Li, Y. & Maraseni, T. 2017 Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. *Atmospheric Research* **197**, 42–63.

Quinlan, J. R. 1992 *Learning with continuous classes*. In: *Proceedings AI'92, 5th Australian Joint Conference on Artificial Intelligence* (Adams, S., ed.). World Scientific, in, Singapore, pp. 343–348.

R Development Core Team 2019 *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

Ren, K., Fang, W., Qu, J., Zhang, X. & Shi, X. 2020 Comparison of eight filter-based feature selection methods for monthly streamflow forecasting–three case studies on CAMELS data sets. *Journal of Hydrology* **586**, 124897.

Riad, S., Mania, J., Bouchaou, L. & Najjar, Y. 2004 Rainfall-runoff model using an artificial neural network approach. *Mathematical and Computer Modelling* **40** (7/8), 839–846.

Rodrigues, J. A. M., Viola, M. R., Alvarenga, L. A., Mello, C. R., Chou, S. C., Oliveira, V. A., Uddameri, V. & Morais, M. A. V. 2020 Climate change impacts under representative concentration pathway scenarios on streamflow and droughts of basis in the Brazilian Cerrado biome. *International Journal of Climatology* **40**, 1–16.

Rodrigues, J. A. M., Andrade, A. C. O., Viola, M. R., Ferreira, D. D., Mello, C. R. & Thebaldi, M. S. 2021 Hydrological modeling in a basin of the Brazilian Cerrado biome. *Revista Ambiente & Água* **16**, 1–18.

Selvi, O. & Huseyinov, İ. 2020 A novel algorithm for feature selection based on geographic distance metric: a case study of streamflow forecasting of Austria's water resources. *International Journal of Environmental Science and Technology* **17**, 295–308.

Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., Kargar, K., Mosavi, A., Nabipour, N. & Chau, K. W. 2020 Predicting standardized streamflow index for hydrological drought using machine learning models. *Engineering Applications of Computational Fluid Mechanics* **14**, 339–350.

Shortridge, J. E., Guikema, S. D. & Zaitchik, B. F. 2016 Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences* **20**, 2611–2628.

Silva Neto, V. L., Viola, M. R., Mello, C. R., Alves, M. V. G., Silva, D. D. & Pereira, S. B. 2020 Mapeamento de chuvas intensas para o estado do Tocantins (Heavy rainfall mapping for Tocantins State, Brazil). *Revista Brasileira de Meteorologia* **35**, 1–11.

Tongal, H. & Booij, M. J. 2018 Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology* **564**, 266–282.

Tucci, C. E. M. 2004 *Hidrologia: Ciência e Aplicação (Hydrology: Science and Application)*, 3th edn. UFRGS, ABRH, Porto Alegre.

Uliana, E. M., Silva, D. D., Moreira, M. C., Pereira, D. R. & Almeida, F. T. 2019 Modelo hidrológico híbrido para previsão de vazões na bacia do rio Piracicaba-MG (Hybrid hydrological model for water flow prediction in the Piracicaba River Basin-MG, Brazil). *Revista Brasileira de Meteorologia* **34**, 471–480.

Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Venables, W. N. & Ripley, B. D. 2002 *nnet Package: Modern Applied Statistics with S*. Springer, New York.

Viola, M. R., Mello, C. R., Acerbi Junior, F. W. & Silva, A. B. 2009 Modelagem hidrológica na bacia hidrográfica do rio Aiuruoca, MG (Hydrologic modeling in the Aiuruoca river basin, Minas Gerais State). *Revista Brasileira de Engenharia Agrícola e Ambiental* **13**, 581–590.

Xue, X., Yao, M. & Wu, Z. 2018 A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm. *Knowledge and Information Systems* **57**, 389–412.

Yang, T., Asajan, A. A., Welles, E., Gao, X., Sorooshian, S. & Liu, X. 2017 Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resources Research* **53**, 2786–2812.

Yang, Q., Zhang, H., Wang, G., Luo, S., Che, D., Peng, W. & Shao, J. 2019 Dynamic runoff simulation in a changing environment: a data stream approach. *Environmental Modelling & Software* **112**, 157–165.

Yaseen, Z. M., Allawi, M. F., Yousif, A. A., Jaafar, O., Hamzah, F. M. & El-Shafie, A. 2018 Non-tuned machine learning approach for hydrological time series forecasting. *Neural Computing and Applications* **30**, 1479–1491.

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

Yin, Z., Feng, Q., Wen, X., Deo, R. C., Yang, L., Si, J. & He, Z. 2018 Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. *Stochastic Environmental Research and Risk Assessment* **32**, 2457–2476.

Zhu, S., Luo, X., Xu, Z. & Ye, L. 2019 Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection. *Hydrology Research* **50**, 200–214.