RBRH

Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 30, e23, 2025
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.302520240087

Evaluating different strategies for machine learning training applied to flow forecasting based on clustering of flood events

Avaliação de diferentes estratégias para treinamento de aprendizado de máquina aplicado à previsão de vazão com base no agrupamento de eventos de inundação

Patrícia Cristina Steffen[1], Júlio Gomes[2], Eloy Kaviski[2] & Daniel Henrique Marco Detzel[2]

[1]Universidade Federal de Mato Grosso, Barra do Garças, MT, Brasil
[2]Universidade Federal do Paraná, Curitiba, PR, Brasil
E-mails: patricia.steffen@ufmt.br (PCS), jgomes.dhs@ufpr.br (JG), eloy.dhs@ufpr.br (EK), detzel@ufpr.br (DHMD)

Received: November 30, 2024 - Revised: March 13, 2025 - Accepted: March 15, 2025

ABSTRACT

This paper presents a new hydrological modeling approach for discharge prediction based on flood clustering. Combined with Machine Learning techniques, river flow simulation is optimized through increased data similarity within clusters. Using daily mean discharge from 1964 to 2015 in União da Vitória (Iguaçu River basin, Paraná State, Brazil), the Fuzzy C-Means algorithm clustered flood events into three groups. So, five models were trained: one for the complete series, one for all flood events, and one for each cluster. The Support Vector Regression algorithm was used to develop Artificial Intelligence (AI) models, that had better performance in predicting discharge for each group they were trained and showed similar efficiency to the model trained for the entire series for a 1-day forecast time. The present paper discusses only the results from the training and testing phases. A future paper (in elaboration) will present the development and evaluation of the flow forecast models based on the proposed methodology.

Keywords: Artificial Intelligence techniques; Data-driven intelligent models; Fuzzy C-Means algorithm; Hydrological modeling; Support Vector Regression algorithm.

RESUMO

O presente trabalho apresenta uma nova abordagem para a modelagem hidrológica, baseada no agrupamento de cheias. Combinada às técnicas de Aprendizagem de Máquina, a pesquisa visa otimizar a simulação de vazões, por meio do aumento da similaridade dos dados de cada grupo. Usando as vazões médias diárias de 1964 a 2015 em União da Vitória (bacia hidrográfica do rio Iguaçu, Paraná, Brasil), o algoritmo Fuzzy C-Means agrupou os eventos de cheia em três grupos. Assim, foram treinados cinco modelos: um para a série completa, um para todas as cheias e um para cada grupo. O algoritmo de Regressão por Vetores de Suporte foi aplicado para obtenção dos modelos de Inteligência Artificial (IA), que apresentaram melhor desempenho ao prever vazões para os grupos em que foram treinados e apresentaram desempenho similar ao modelo treinado com a série completa para um horizonte de previsão de um dia. Este artigo discute somente os resultados obtidos nas fases de treinamento e teste. Um trabalho futuro (em elaboração) apresentará o desenvolvimento e a avaliação dos modelos de previsão, com base na metodologia proposta.

Palavras-chave: Técnicas de Inteligência Artificial; Modelos inteligentes de análise de séries temporais. Algoritmo Fuzzy C-Means; Modelagem hidrológica; Algoritmo de Regressão por Vetores de Suporte.

INTRODUCTION

Hydrological modeling has achieved significant breakthroughs due to the computational revolution (Singh, 2018). In this context, hydrological forecasting is generally conducted using two primary approaches (Mohammadi et al., 2020; Snieder et al., 2020; Wang et al., 2021; Difi et al., 2023): (i) traditional models and (ii) data-driven models.

In classical hydrological modeling, traditional models analyze hydrological characteristics and physically describe the confluence of flows, giving their parameters a certain degree of physical relevance (Ding et al., 2020). However, the calibration of these models is based on observed data and requires a high level of familiarity with the model from the researcher (Ding et al., 2020; Ebtehaj & Bonakdari, 2022). Moreover, incorrect parameter estimation can significantly increase errors in this class of models (Luppichini et al., 2019, 2022). Luppichini et al. (2022) further note that physically based models face challenges due to inherent heterogeneity of natural systems.

Ebtehaj & Bonakdari (2022), while acknowledging that traditional models have produced good results for flood forecasting, also highlighted several limitations. These include the manual collection of data, which is often stored in a disorganized manner, leading to insufficiently sized data series – especially problematic in remote regions where data availability is already limited. Obviously, scarcity of data is also an important limitation for data-driven models.

On the other hand, with the growth of hydrological infrastructure, increased data availability, and advances in computational development (Singh, 2018; Bai et al., 2021), the capacity of empirical models has expanded, leading to the emergence of so-called data-driven models. These models use system data to identify connections between variables (input, internal, and output) without explaining the system's physical behavior. In other words, the information source for these models is the data series itself (Adnan et al., 2019; Fathian et al., 2019; Ding et al., 2020; Bai et al., 2021; Ebtehaj & Bonakdari, 2022).

Time series models, which represent a specific type of data-driven model, capture both linear and non-linear relationships between discharges and their mathematical parameters (Zhang et al., 2018; Adnan et al., 2019). Commonly used approaches include the least squares method, multiple linear regression, and models such as Autoregressive (AR), Moving Average (MA), and Autoregressive Moving Average (ARMA). However, these methods typically only account for the linear relationships between inputs and outputs, often resulting in outputs that are insufficient and fail to address the inherent non-linearity in the systems (Khosravi et al., 2021).

To address these disadvantages, automated time series models using Artificial Intelligence (AI) have been developed and explored for hydrological modeling and discharge forecasting (Bai et al., 2021; Khosravi et al., 2021). These efforts have focused on Computational Intelligence (CI), which can complement or replace traditional physical models (Fathian et al., 2019; Bai et al., 2021). Such models, often referred to as data-driven intelligent models (Ding et al., 2020; Khosravi et al., 2021), are notable for their ability to handle large data series and accommodate data at different scales (Bai et al., 2021; Khosravi et al., 2021). Among the most popular CI techniques are neural networks, fuzzy rule-based systems, and genetic algorithms (Mahdavi-Meymand et al., 2023).

Among CI methods, Machine Learning (ML) techniques involve the automatic detection of significant patterns in data series. These techniques are categorized based on different learning objectives (Xu & Liang, 2021), including classification, regression, and clustering (Ibrahim et al., 2022; Mahdavi-Meymand et al., 2023). Over the last two decades, these techniques have yielded promising results for hydrological problems, including: (i) Support Vector Machines (SVM) and Support Vector Regression (SVR) (Li et al., 2019; Chen et al., 2021; Niu & Feng, 2021; Kim et al., 2022; Difi et al., 2023; Sharma et al., 2023); (ii) Random Forest (RF) (Fathian et al., 2019; Saadi et al., 2019; Schoppa et al., 2020; Desai & Ouarda, 2021; Sharma et al., 2023); (iii) Artificial Neural Networks (ANN) (Zhang et al., 2018; Fathian et al., 2019; Snieder et al., 2020; Brêda et al., 2021; Chen et al., 2021; Lima & Scofield, 2021); (iv) Long Short-Term Memory (LSTM) (Ding et al., 2020; Chen et al., 2021; Kim et al., 2022; Sharma et al., 2023; Zakhrouf et al., 2023); (v) Extreme Learning Machine (ELM) (Li et al., 2019; Ribeiro et al., 2020; Yaseen et al., 2020; Niu & Feng, 2021; Ebtehaj & Bonakdari, 2022; Feng et al., 2022; Difi et al., 2023); (vi) Multivariate Adaptive Regression Splines (MARS) (Niu & Feng, 2021; Sharma et al., 2023); and (vii) Adaptive Neuro-Fuzzy Inference System (ANFIS) (Mohammadi et al., 2020; Shukla et al., 2022; Samantaray et al., 2023).

Most applications involve Artificial Neural Networks (ANNs) and their variations, such as Extreme Learning Machines (ELM) and Long Short-Term Memory (LSTM), as well as Support Vector Machines (SVM) and its regression variant, Support Vector Regression (SVR). Generally, research has focused on comparing these techniques. Recently, LSTM, ELM, and SVM (or SVR) have been frequently compared in studies (Li et al., 2019; Chen et al., 2021; Niu & Feng, 2021; Kim et al., 2022; Difi et al., 2023; Sharma et al., 2023). Other applications of these models in hydrology can be found in Adnan et al. (2019), Kurian et al. (2020), Islam et al. (2021), and Ebtehaj & Bonakdari (2022). Additionally, Mahdavi-Meymand et al. (2023) provide a literature review of the last decade's publications applying Machine Learning techniques to various hydrological objectives.

For discharge forecasting, Adnan et al. (2019) highlight that recent decades have seen successful applications of methods such as ANN, ANFIS, MARS, SVM, genetic algorithms (GA), and their hybrid models. These methods are effective in identifying non-linearity in flood formation processes. In this work, SVR algorithm was chosen for its satisfactory results and straightforward computational implementation. However, selecting the model is only one aspect; it also requires consideration of other details in forecasting hydrological modeling.

Thus, this manuscript evaluates a hydrological modeling approach for discharge prediction, supported by previous flood events clustering and flood events characterization. In this approach, events within the same cluster share similar characteristics and differ from those in other clusters. By clustering events with similar features, the approach constrains the predicted event's characteristics to those of its respective group (Joo et al., 2021), thereby reducing prediction variability.

In this context, Xu & Liang (2021) noted that machine learning methods for clustering and classification are often employed to support regression methods. However, these approaches

typically serve as alternatives to regression techniques rather than integrating group hydrological modeling.

No application like the one proposed in this paper was found in the literature. However, Joo et al. (2021) noted that cluster analysis has been widely used in hydrology. They presented various studies that utilized cluster analysis for applications such as forecasting ungauged hydrological data, flood frequency analysis, hydrological modeling, flood forecasting, and classifying hydrological and catchment similarities.

According to Tarasova et al. (2019), classifying flood events provides a better understanding of flood generation mechanisms, which are not always clearly defined at the catchment scale. There are three main categories for flood generation mechanisms: hydroclimatic, characterized by large-scale circulation patterns and atmospheric conditions at the event's onset; hydrological, defined by the catchment's precipitation patterns and its antecedent conditions; and hydrograph-based, which considers formation mechanisms through their effects on hydrograph characteristics (Tarasova et al., 2019). Tarasova et al. (2019) reviewed existing classifications of flood events, discussing their validity, limitations, and transferability across temporal and spatial scales.

To aid in understanding the mechanisms of flood formation, clustering emphasizes similar characteristics among events within the same group. When used as a preliminary step in discharge forecasting, this two-step framework narrows the focus of forecasts typically produced by a single model. Consequently, the further an event is from a cluster of observed floods, the less it is likely to resemble the model associated with that group. In other words, when predicting the occurrence of an event, it is expected that the event will closely match its own model and differ more significantly from others.

Thus, the primary goal of cluster analysis is to use data characteristic variables to group data into classes in the most appropriate way, ensuring that similar objects are in the same class, thereby reducing the number of observations in a sample (Morettin & Singer, 2023). This process also requires defining the similarity within a class and the dissimilarity between classes, which is a complex task (Ezugwu et al., 2022). In this study, we use the Fuzzy C-Means (FCM) algorithm to assign each element to a group from a predefined number of groups (Mosavi et al., 2021).

Hence, the objective of this paper is to develop a new hydrological modeling approach for discharge prediction based on the previous characterization and clustering of flood events. This approach aims to optimize the simulation of extreme maximum and minimum river flows by reducing the amplitude of the clustered data. Additionally, the forecasting process is automated, thereby eliminating the inherent disadvantages of statistical time series models and reducing processing time.

The proposed hydrological modeling approach is demonstrated through a real-world application. Daily mean discharges from 1964 to 2015 in União da Vitória gauging station (Iguaçu River basin, Paraná State, Brazil) were used, and the SVR algorithm was applied to develop AI models for 1-day discharge forecasting. The FCM algorithm clustered flood events into three groups, a number previously defined. Hence, five models were trained: one model for the Complete Series, one model for All Flood Events, and one model for each of the three Clusters. The results depend on the clustering, the chosen hydrological model, and the forecast horizon. So, the proposed approach is allowed for the use of various hydrological models.

CASE STUDY

The chosen study area encompasses the Iguaçu River basin up to the União da Vitória gauging station (65310000), selected due to the historically recurrent flood events in the basin, particularly in the city of União da Vitória, Paraná State, Brazil (Figure 1).

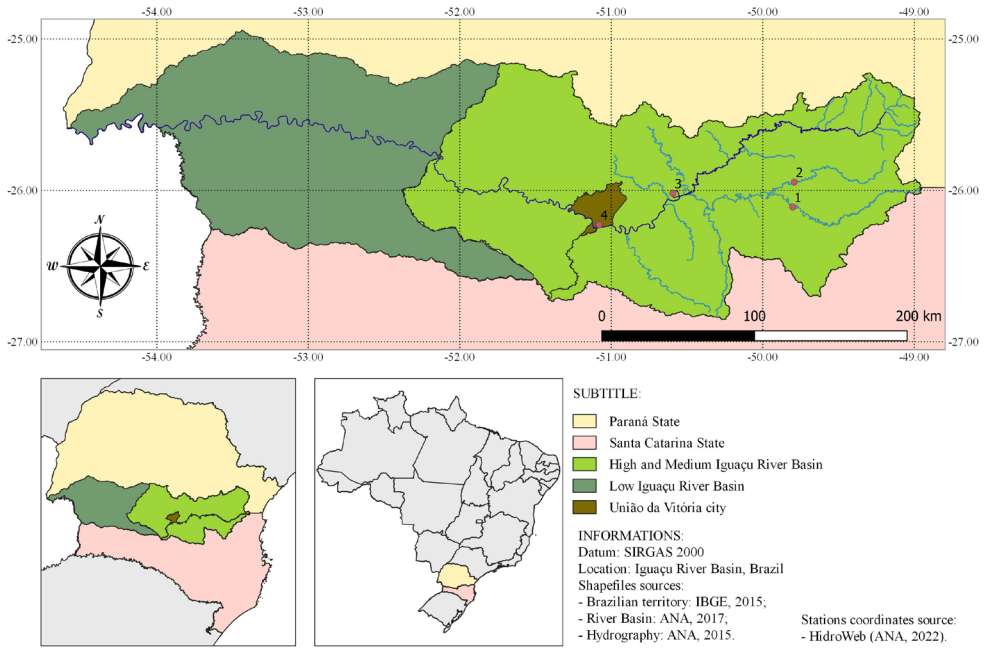

Figure 1. Location of União da Vitória city in Paraná State and Iguaçu River Basin, along with the positions of gauging stations.

The Iguaçu River flows almost entirely across the state of Paraná, in an east-west direction, spanning 1,320 km before discharging into the Paraná River (Mine & Tucci, 2002). The Iguaçu River basin covers a total drainage area of 70,800 km², divided between the states of Paraná and Santa Catarina in Brazil, with a small portion extending into Argentina (Rocha, 2012).

The Iguaçu River is formed by the confluence of the Iraí and Atuba Rivers in Curitiba Metropolitan Region. Upstream of União da Vitória, the main contributors to the Iguaçu River are the Potinga River on the right bank, and the Negro and Timbó Rivers on the left bank. Other tributaries, including the Verde, Itaqui, da Vargem, and Claro Rivers (on the right bank), and the Passa Dois and Paciência Rivers (on the left bank), provide smaller contributions to the Iguaçu River upstream of the União da Vitória gauging station. At this point, the drainage area is approximately 24,200 km² (Agência Nacional de Águas e Saneamento Básico, 2023), accounting for more than one-third of the entire basin area.

Based on the identified tributaries of the Iguaçu River, existing gauging stations in the river basin were investigated using data from the National Water and Sanitation Agency database (Agência Nacional de Águas e Saneamento Básico, 2023). The selection of stations primarily considered factors such as sample size, location, data quality, and the consistency of their records (Tozzi & Fill, 2020). In addition to União da Vitória (65310000), data were collected from Fluviópolis (65220000) on the Iguaçu River, Rio Negro (65100000) on the Negro River, and São Bento (65155000) on the Várzea River. São Bento station data were filled in using information from Rio da Várzea dos Lima station, which is located on the same river.

Figure 1 displays these stations with their IDs: Rio Negro (ID 1), São Bento (ID 2), Fluviópolis (ID 3), and União da Vitória (ID 4). Table 1 provides their descriptive statistics.

The Iguaçu River has a cascade of hydroelectric Power plant reservoirs for electric energy production. However, all the selected gauging stations are located upstream of the reservoirs. Furthermore, it is worth mentioning that União da Vitória is subject to the backwater originating from the Foz do Areia Hydroelectric Power Plant reservoir, which was implemented in 1980. Steffen & Gomes (2018) corrected the flow data for União da Vitória, from 1980 to 2015, using water level data from the R5 Porto Vitória fluviometric station (65365800), located between União da Vitória and Foz do Areia.

To account for the travel time of water from stations ID 1, ID 2, and ID 3 to station ID 4, a correlation analysis was performed (Table 2). Among all stations and lags, including lag 0, ID 3 exhibited the highest correlation with ID 4, although this correlation gradually decreases as the lag increases. Station ID 3 is the closest to ID 4 (Figure 1) and is in the Iguaçu River. However, with about 100 km distance between those stations, the estimated travel time is less than 1 day and cannot be precisely defined on a daily scale.

Station ID 1 exhibited a higher correlation with ID 4 at lag 2, which can be attributed to the distance between the stations and its location on the Negro River. ID 1 consistently showed higher correlations with ID 4 than ID 2 across all lags. This is because ID 2 is situated on a tributary of the Negro River, which explains its lower correlation.

MATERIAL AND METHODS

The methodological processes are summarized in Figure 2 and are detailed in the following sections.

Table 1. Descriptive statistics of the gauging stations (Agência Nacional de Águas e Saneamento Básico, 2023).

Code stations	ID	Stations name	Periods	n (years)	$\bar{Q}$ (m³/s)	q (L/s/km²)	Min (m³/s)	Max (m³/s)	S (m³/s)	a
65100000	1	Rio Negro	1931-2015	72	73.0	21.2	7.7	953.9	68.7	3.24
65155000	2	São Bento	1931-2015	84	35.8	17.9	6.0	482.3	28.8	2.84
65220000	3	Fluviópolis	1964-2015	51	399.7	21.5	34.5	3819.4	346.5	2.53
65310000	4	União da Vitória	1931-2015	84	492.2	20.3	46.2	5156.7	439.5	2.51

n – series size; $\bar{Q}$ – mean discharge; q – mean specific flow rate; S – standard-deviation; a – skewness.

Table 2. Correlations between the mean daily discharges at União da Vitória (ID 4) and the upstream gauging stations for different time lags. The highest correlations are highlighted.

Lag	Gauging Stations (ID)		
	1	2	3
0	0.8411	0.8246	0.9781
1	0.8572	0.8300	0.9696
2	0.8617	0.8266	0.9523
3	0.8588	0.8172	0.9286

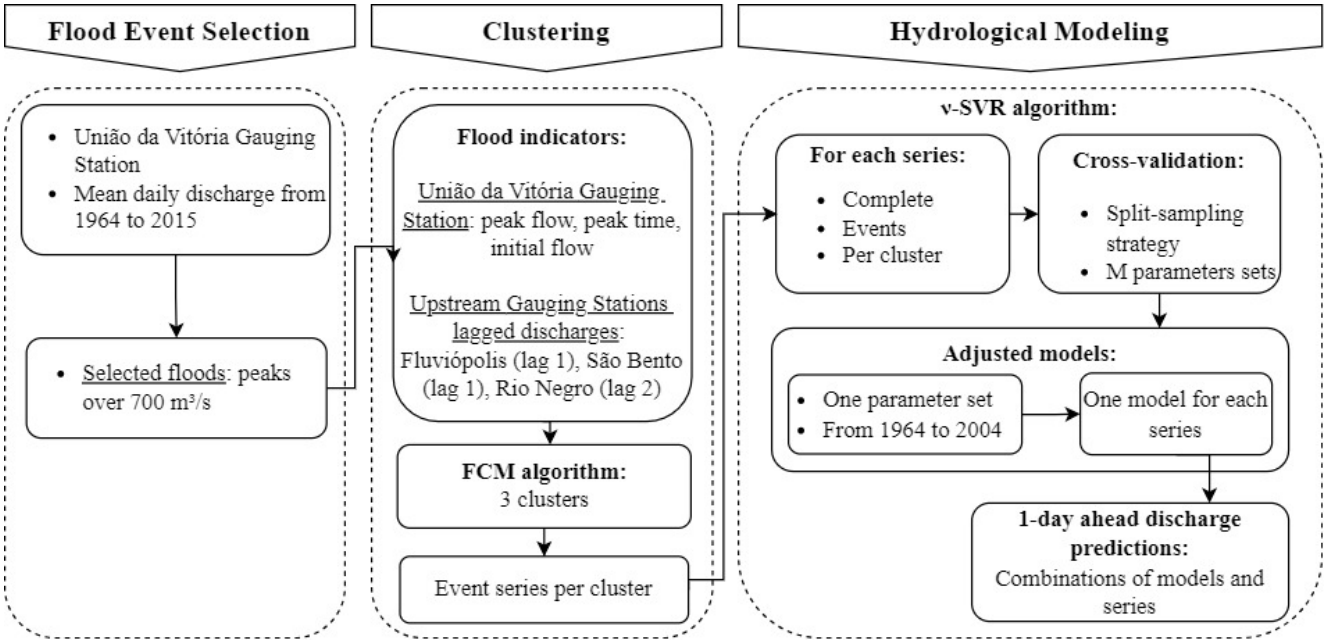

Figure 2. Workflow for 1-day-ahead discharge prediction at União da Vitória.

Flood events selection

Flood events were selected from the mean daily discharge series of the ID 4 gauging station, observed from 1964 to 2015, which corresponds to the simultaneous data record period among all the river stations considered in the study. All observed floods with peaks exceeding 700 m^3/s were considered, which corresponds to 20% on the flow duration curve. The mean daily discharge at ID 4 has a flow duration of 35% (about 500 m^3/s), whereas the overflow threshold of the main channel of Iguaçu River at this location is 1,387 m^3/s, corresponding to 5% (Steffen & Gomes, 2018).

The events were delineated by identifying their respective start, end, and peak dates, along with their peak and initial discharges. The start of an event is defined by the biggest lag previously the date on which the average daily flow equals or exceeds the long-term average flow in the ID4 gauging station. The end of an event is defined by the date on which the average daily flow is equal to or less than the long-term average flow in the ID4 gauging station.

This selection yielded a total of 194 events with these characteristics. Over 52 years of observation, nearly 30% of the events (57 events) resulted in overbank flows (peak discharge > 1,387 m^3/s), while 54.1% of the peak discharges were equal to or exceeded 1,000 m^3/s. Additionally, almost 46% of the events (89 events) had peak discharges between 700 m^3/s and 1,000 m^3/s.

Clustering

Fuzzy C-Means (FCM) algorithm application requires flood indicators, which are the characteristic variables of the events used to group the data. For the events selected at ID 4 and the upstream stations, six indicators were determined: peak flow ($\mathbf{Q}_p$), peak time ($\mathbf{t}_p$), and initial flow ($\mathbf{Q}_0$) at ID 4, as well as observed discharges at the upstream stations, which are adjusted for the peak time ($\mathbf{t}_p$) at ID 4 and the lags defined from the correlation analysis – lag 1 to ID 3 $(\mathbf{Q}_{tp-1})_{ID-3}$, lag 1 to ID 2 $(\mathbf{Q}_{tp-1})_{ID-2}$, and lag 2 to ID 1 $(\mathbf{Q}_{tp-2})_{ID-1}$.

In FCM, fuzzy logic is incorporated through an uncertainty parameter, which determines the degree of similarity of events to each group. Clustering is considered fuzzy when the uncertainty parameter exceeds 1.0 (Askari, 2021; Mosavi et al., 2021). Using linearly standardized indicators (Mosavi et al., 2021), FCM is defined by the following iterative process (Steffen & Gomes, 2018; Askari, 2021):

1. The application of linear normalization to the indicators is performed using Equation 1, where x_{ij} and x_{ij}^o are the normalized and non-normalized values of the *i*-th observation of the *j*-th indicator, respectively, and $x_{j\,min}^0$ and $x_{j\,max}^0$ are the minimum and maximum values of the *j*-th indicator, respectively.

$$x_{ij} = \frac{x_{ij}^o - x_{j\,min}^0}{x_{j\,max}^0 - x_{j\,min}^0} \quad (1)$$

2. The number of clusters was defined as $c = 3$, and the degree of uncertainty was set to $r = 1.25$.
3. An initial random fuzzy partition matrix ($\mathbf{U}_0$) is defined. This matrix, with *j* rows and *c* columns, represents the degree of membership of the *j*-th sample observation to each of the *c* predefined clusters.
4. The cluster centroids ($\mathbf{V}$) are calculated, with their coordinates representing a weighted average of the indicators for each cluster. Equation 2 defines the cluster centroids, where v_{kj} is the coordinate of the *j*-th indicator of the *k*-th centroid, and μ_{ki} is the *i*-th element of the membership matrix $\mathbf{U}$ for the *k*-th cluster.

$$v_{kj} = \frac{\sum_{k=1}^{c} \left(\mu_{ki}\right)^{r} x_{ij}}{\sum_{k=1}^{c} \left(\mu_{ki}\right)^{r}} \quad (2)$$

5. The Euclidean distances from the events to the centroids (**D**) are calculated using Equation 3, where d_{ik} represents the Euclidean distance of the *i*-th observation to the *k*-th cluster, and d denotes the total number of flood indicators, which defines the clustering dimension.

$$d_{ik} = \| x_i - v_k \|^2 = \sqrt{\sum_{j=1}^{d} \left(x_{ij} - v_{kj}\right)^2} \quad (3)$$

6. The objective function (**J**) is calculated using Equation 4, where n represents the total number of input observations.

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} \left(\mu_{ki}\right)^{r} \left(d_{ik}\right)^2 \quad (4)$$

7. The fuzzy partition matrix (**U**) is updated using Equation 5.

$$\mu_{ki} = \left[\sum_{t=1}^{c} \left(\frac{d_{ik}}{d_{tk}} \right)^{\frac{2}{r-1}} \right]^{-1} \quad (5)$$

8. Verification of the algorithm's stopping criteria: the process stops when the absolute difference between the maximum degrees of similarity of the observations (between the last two stages of the iterative process) is less than the maximum tolerated ($\varepsilon_t = 10^{-5}$), indicating that the minimum value of the objective function has been reached.
9. If the stopping criterion is not met, return to Step 4 and repeat the process until the condition is satisfied.

The definition of the number of groups (*c*) and the degree of uncertainty (*r*) was performed empirically. For each combination of *c*, integer and belonging to the interval [2, 10], and *r* from 1.25 to 2.00, incremented by 0.25, the FCM method algorithm was executed and the results compared. The values of $c = 3$ and $r = 1.25$ were adopted, which resulted in more homogeneous behavior of the fuzzy partition matrix and avoided groups with a small number of events.

Goodness of fit measures

To evaluate the adjustments and forecasts, the following metrics were applied: Nash-Sutcliffe Efficiency (NS) (Unduche et al., 2018; Lawin et al., 2019; Liang et al., 2019; Althoff et al., 2021; Brêda et al., 2021; Mosavi et al., 2021; Moura et al., 2022), Kling-Gupta Efficiency (KG) (Gupta et al., 2009; Althoff et al., 2021; Moura et al., 2022; Lappicy & Lima, 2023) and Mean Absolute Relative Error (MARE) (Staudinger et al., 2011; Unduche et al., 2018; Liang et al., 2019; Althoff et al., 2021; Moura et al., 2022; Letessier et al., 2023). These metrics were chosen to evaluate different phases of the hydrographs, ensuring a comprehensive assessment of the model's performance.

ν-SVR algorithm

Support Vector Machines (SVM) have gained popularity over the last four decades due to their efficiency and processing speed in the realm of modern statistical learning (Ibrahim et al., 2022). The evolution of the SVM approach is well-documented, with Freitas (2016) providing a comprehensive history of its development – from the early algorithms for pattern recognition in the 1950s to its current form, grounded in Statistical Learning Theory.

Support Vector Machines (SVM) is a classification method rooted in the Vapnik-Chervonenski theory, which was developed to address supervised learning problems. SVM is part of the generalized linear classification family (Ibrahim et al., 2022; Roy & Chakraborty, 2023), and, when adapted for non-linear regression, is known as Support Vector Regression (SVR). This adaptation is particularly effective for flow forecasting tasks. Based on Vapnik's ε-indifferent cost function, Schölkopf et al. (2000) introduced a significant modification to the SVR algorithm, known as ν-SVR. This version automatically minimizes the ε parameter, making the model less dependent on manually adjusting this parameter through optimization algorithms.

In ν-SVR algorithm, ν is the parameter responsible for controlling the number of support vectors – or the number of errors in the solution (Schölkopf et al., 2000); considering that $\nu \in (0,1]$, the closer to 1, the greater the number of support vectors and, consequently, the greater the tolerance to errors (Chang & Lin, 2001, 2022). From the set of parameters, the ν-SVR algorithm aims to minimize Equation 6 (Schölkopf et al., 2000; Chang & Lin, 2001, 2022):

$$\tau\left(\mathbf{w}, \xi^*, \varepsilon\right) = \frac{1}{2} \| \mathbf{w}^2 \| + C\left(\nu\varepsilon + \frac{1}{m} \sum_{i=1}^{m} \left(\xi_i + \xi_i^*\right) \right) \quad (6)$$

subject to:

$$\left(\left(\mathbf{w} \cdot x_i\right) + b\right) - y_i \leq \varepsilon + \xi_i \quad (7)$$

$$y_i - \left(\left(\mathbf{w} \cdot x_i\right) + b\right) \leq \varepsilon + \xi_i^*, \quad (8)$$

where parameter C is the constant that defines the degree of cost penalty when there is a training error – C is also called as the regularization parameter and must be defined a priori; ξ_i and ξ_i^* are greater than zero and correspond to the slack variables that, respectively, specify the highest and lowest training errors, subject to tolerance; and, ε is the tolerance – greater than zero.

Therefore, to obtain the regression equation capable of predicting flows, NuSVR class of *sklearn.svm* module of Python was used, which was implemented using the LIBSVM library by Chang & Lin (2022). The parameters for calibration in this paper are: (i) the kernel function, (ii) the kernel coefficient, called Gamma (γ), (iii) the ν (Nu) parameter defined in the interval (0, 1], delimiting an upper limit for the training errors fraction, and a lower limit for the support vectors fraction, and (iv) the regularization parameter C.

The Radial Basis Function (RBF) was selected as kernel function. RBF is less difficult numerically, allows non-linear samples to be mapped into a higher dimensional space and has fewer parameters to be calibrated – if compared to the polynomial function (Freitas, 2016).

In the calibration phase, C and γ parameters were combined, one by one, to identify the best combination among the 64 possibilities, considering:

$$C = [0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0] \quad (9)$$

$$\gamma = [0.001, 0.01, 0.05, 0.1, 0.33, 0.5, 1.0, 5.0] \quad (10)$$

To the other parameters of the NuSVR class, their respective standardized specifications were assigned, including the ν parameter, whose default value is 0.5.

Given the parameters and correlations (Table 2) previously discussed, the SVR model for predicting the discharge at União da Vitória one day ahead uses the following input variables: the discharges from Rio Negro using lag 2 $(\mathbf{Q}_{tp-2})_{RN}$, and the discharges from São Bento $(\mathbf{Q}_{tp-1})_{SB}$, Fluviópolis $(\mathbf{Q}_{tp-1})_{FL}$, and União da Vitória $(\mathbf{Q}_{tp-1})_{UV}$, using lag 1.

Hyperparameters tuning

For data series categories under analysis, SVR algorithm trained and validated 5 different models: one model for the complete series, one model for all events series; and one model for each event of the three clusters series.

The period from 1964 to 2004 was subdivided into training and validation sets using a split sampling strategy based on a 20-year window (Hallouin et al., 2020). This approach divides the data into approximately equal intervals for training and validation stages, ensuring that the largest flood events are included in at least one of these stages. Eight samples were generated to account for the largest historical flood event in 1983 as shown in Figure 3.

For each sample and each of the 64 combinations from C and γ values, the NuSVR algorithm was executed for the training period, generating the models later applied to the corresponding validation period. In the training stage, the 32 best parameter sets for each model were selected, and their models were applied to the validation period. The metrics for evaluating the training and validation results were calculated, along with their main statistics: mean, maximum, and minimum flows; standard deviation; skewness; and volume ratio (simulated/observed). Figure 4 presents a summary of the training and test stages.

Adjusting models

The parameters set selected to validate the 5 models was obtained by averaging the quality measures results from the eight training and validation samples. For each parameter set, the median of the averages was determined. Then, for each model, the parameter sets with performance lower than the median were identified. Finally, the parameter set that was repeated across all models was selected. This same set was used for all 5 models.

Based on Hallouin et al. (2020), to avoid overfitting, the selected parameter set was the one with metrics close to the medians but capable of satisfactorily estimating discharges at ID 4, considering all five categories of models. When the performances of different parameter sets were close to each other, the set with the lowest C and γ values was adopted. Increasing these parameter values can speed up processing but does not necessarily improve performance.

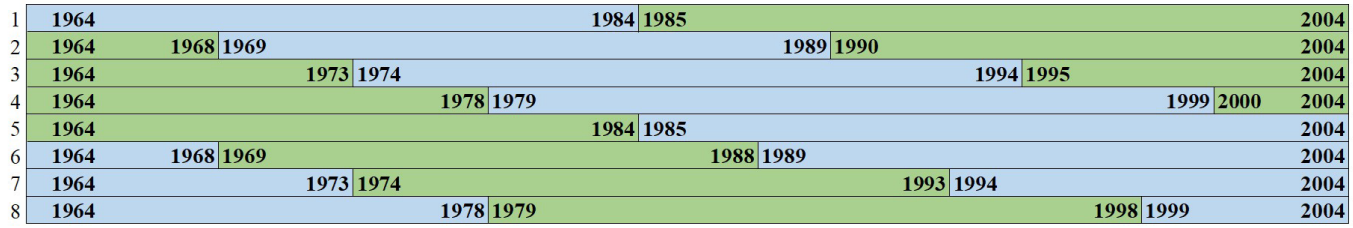

Figure 3. Sampling strategy for the training and validation stages. (training sets in blue, validation sets in green).

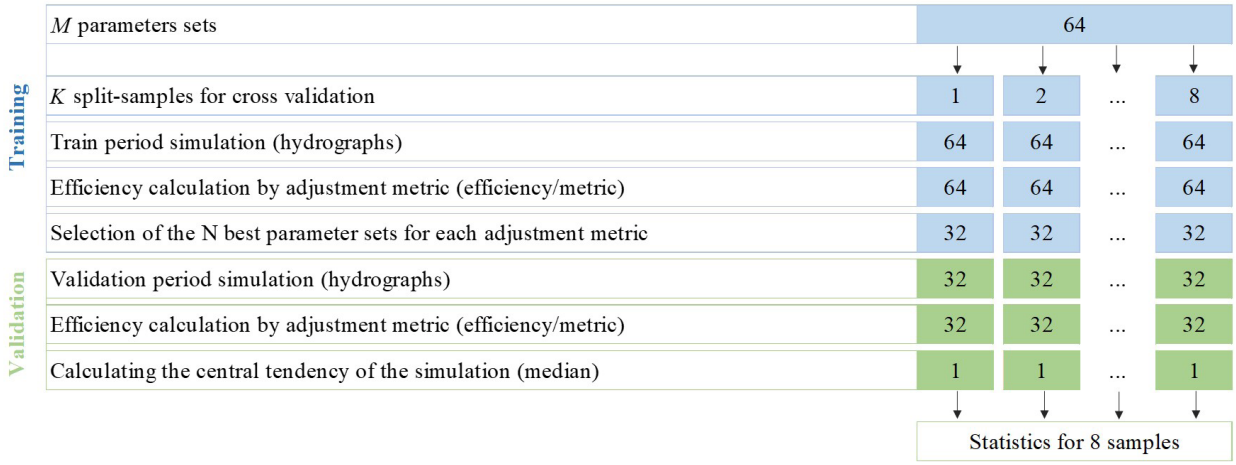

Figure 4. Summary of the training and validation stages.

Based on the established values of C and γ, each type of model was calibrated for the entire period of the sampling strategy (from 1964 to 2004). Simulations were then carried out for all series using each model, i.e., each type of calibrated model was applied to all samples: Complete, All Events, and by group.

Descriptive statistics and goodness-of-fit measures were calculated for each combination of model and simulated series. However, since each series has different sizes, proper comparisons were only possible by considering one series obtained from different models, rather than comparing simulations of different series using a single model.

Testing stage

Each calibrated model was applied to all discharge samples for the period from 2005 to 2015. Once again, descriptive statistics and goodness-of-fit measures were calculated for each combination of model and predicted series. A comparison analysis was possible only for one series due to size limitations encountered during the model adjustment phase.

RESULTS AND DISCUSSIONS

Flood clustering

Table 3 shows the statistical parameters of the maximum similarity degrees of events, with an ideal value of 1.0. The position of the events in the groups is determined by identifying the event's highest similarity degree (U) with the group. The largest number of events is found in Cluster 2; Cluster 3 had the lowest minimum value and average of the maximum similarity degrees, as well as the highest amplitude and standard deviation.

Due to the minimum and maximum values observed in Table 3, Table 4 shows the distribution of the similarity degrees (U) between these extremes, through frequencies of occurrence for certain U intervals for each event series. Overall, 153 events (more than 78%) exhibited a degree of belonging to their group greater than 0.95, while 17 events (around 9%) were less than 0.70. For the events in Clusters 1 and 2, more than 80% presented U values over 0.95. In contrast, events in Cluster 3 showed lower results, with around 54% having a degree of belonging greater than 0.95, 18% below 0.70, and approximately 9% in each of the other intervals.

Table 3. Statistical parameters of the maximum similarity degrees (U).

Parameter	Cluster 1	Cluster 2	Cluster 3
Number of observations	70	102	22
Minimum	0.55	0.53	0.49
Maximum	1.00	1.00	1.00
Amplitude	0.45	0.47	0.51
Mean	0.96	0.95	0.88
Standard-deviation	0.10	0.11	0.15

Table 4. Frequencies of the intervals of maximum similarity degrees (U).

		U intervals					
		[0.95; 1.00]	**[0.90; 0.95)**	**[0.80; 0.90)**	**[0.70, 0.80)**	**< 0.70**	**Sum**
All Events	**Absolut frequency**	153	6	12	6	17	194
	Relative frequency	78.9%	3.1%	6.2%	3.1%	8.7%	100%
Cluster 1	**Absolut frequency**	59	1	5	0	5	70
	Relative frequency	84.3%	1.5%	7.1%	0.0%	7.1%	100%
Cluster 2	**Absolut frequency**	82	3	5	4	8	102
	Relative frequency	80.4%	2.9%	4.9%	3.9%	7.9%	100%
Cluster 3	**Absolut frequency**	12	2	2	2	4	22
	Relative frequency	54.5%	9.1%	9.1%	9.1%	18.2%	100%

Lastly, the coordinates of the centroids of each group are shown in Table 5. The observed peak flow at União da Vitória ($\mathbf{Q_p}$) and the observed lagged flows at upstream stations [($\mathbf{Q_{tp-1}}$)$_{FL}$, ($\mathbf{Q_{tp-1}}$)$_{SB}$, ($\mathbf{Q_{tp-2}}$)$_{RN}$)] in Cluster 3 were higher than the respective coordinates in Clusters 1 and 2. Except for the initial flow ($\mathbf{Q_0}$) coordinate values, the other indicators remained close between their pairs for Clusters 1 and 2. The initial flow coordinate ($\mathbf{Q_0}$) for Cluster 2 was higher than the same coordinate for the other groups.

Considering the entire data period (1964 to 2015), the statistical parameters (Table 6) were determined for the different data sets. It is observed that Clusters 1 and 2 present similar values for both mean and standard deviation. Additionally, the mean and standard deviation for "all events" data set fall between the values observed for Clusters 2 and 3.

In summary, clustering effectively identified similar patterns and characteristics within the events of each group. This resulted in reduced mean, standard deviation, and skewness for Clusters 1 and 2, while increasing the mean and standard deviation for Cluster 3 due to the greater magnitude of its events.

Cross-validation

Considering all eight samples from cross-validation (training and validation phases) to tune hyperparameters, mean values of KG, NS, and MARE were calculated for the 64 parameter sets. For training stage, Figure 5 displays the results variation as heatmaps. Generally, C parameter above 2.0 yielded reasonable results across all γ values, except for Complete Series Model (Figure 5a), which showed reasonable indexes for a broader range of parameter sets, including those with C values below 2.0.

Figure 5 shows that lower values of C and γ generally result in worse performance, with better results observed towards the bottom right corner, indicating increased C and γ parameters. The diagonal from the bottom left corner (C = 10.0 and γ = 0.001) to the top right corner (C = 0.01 and γ = 5.0) represents more uniform results. However, while increasing C and γ improves performance up to a point, further increases lead to minimal improvement and only enhance processing speed.

Figure 5 shows that lower values of C and γ generally result in worse performance, with better results observed towards the bottom right corner, indicating increased C and γ parameters. The diagonal from the bottom left corner (C = 10.0 and γ = 0.001) to the top right corner (C = 0.01 and γ = 5.0) represents more uniform results. However, while increasing C and γ improves performance up to a point, further increases lead to minimal improvement and only enhance processing time.

The variability between parameter sets and samples tends to decrease as the data availability increases. For Complete Series Model (Figure 5a), more data, the range of results across different parameter sets was smaller compared to Cluster 3 Model (Figure 5e), less data. Additionally, the results for Complete Series Model showed more uniformity across samples than those for Cluster 3.

Since Figure 5 is based on the average of the sample metrics, the results are significantly influenced by the specific events included in each sample. In other words, as the differences among the model's input data (i.e., different data samples) increase, the KG, NS, and MARE coefficients tend to diverge from their averages.

In this context, the poorest performance of parameter sets was observed in models by group (Figures 5c, 5d, and 5e), particularly for Clusters 1 and 3. This trend, along with the heatmaps' behavior, underscores the importance of cross-validation and the need to

Table 5. Centroid coordinates of Clusters 1 to 3.

Cluster	$t_{p\ UV}$ (days)	$Q_{p\ UV}$ (m³/s)	$Q_{0\ UV}$ (m³/s)	$Q_{tp\text{-}1\ FL}$ (m³/s)	$Q_{tp\text{-}1\ SB}$ (m³/s)	$Q_{tp\text{-}2\ RN}$ (m³/s)
1	12	1075.2	246.6	669.7	64.3	148.3
2	14	1087.9	434.3	730.4	66.9	150.0
3	41	2490.1	365.9	1934.8	130.2	353.2

Table 6. Descriptive statistics of the data series from the entire period (1964 – 2015).

	Data set				
	Complete	**All events**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Number of events	-	194	70	102	22
Series size	18991	7356	2078	3508	1770
Minimum (m³/s)	6.9	114.0	114.0	314.2	148.1
Maximum (m³/s)	5156.7	5156.7	1939.5	1915.1	5156.7
Mean (m³/s)	513.0	961.5	836.6	842.3	1344.3
Standard-deviation (m³/s)	461.40	511.54	349.56	315.55	744.30
Skewness	3.24	2.29	0.59	0.92	1.52

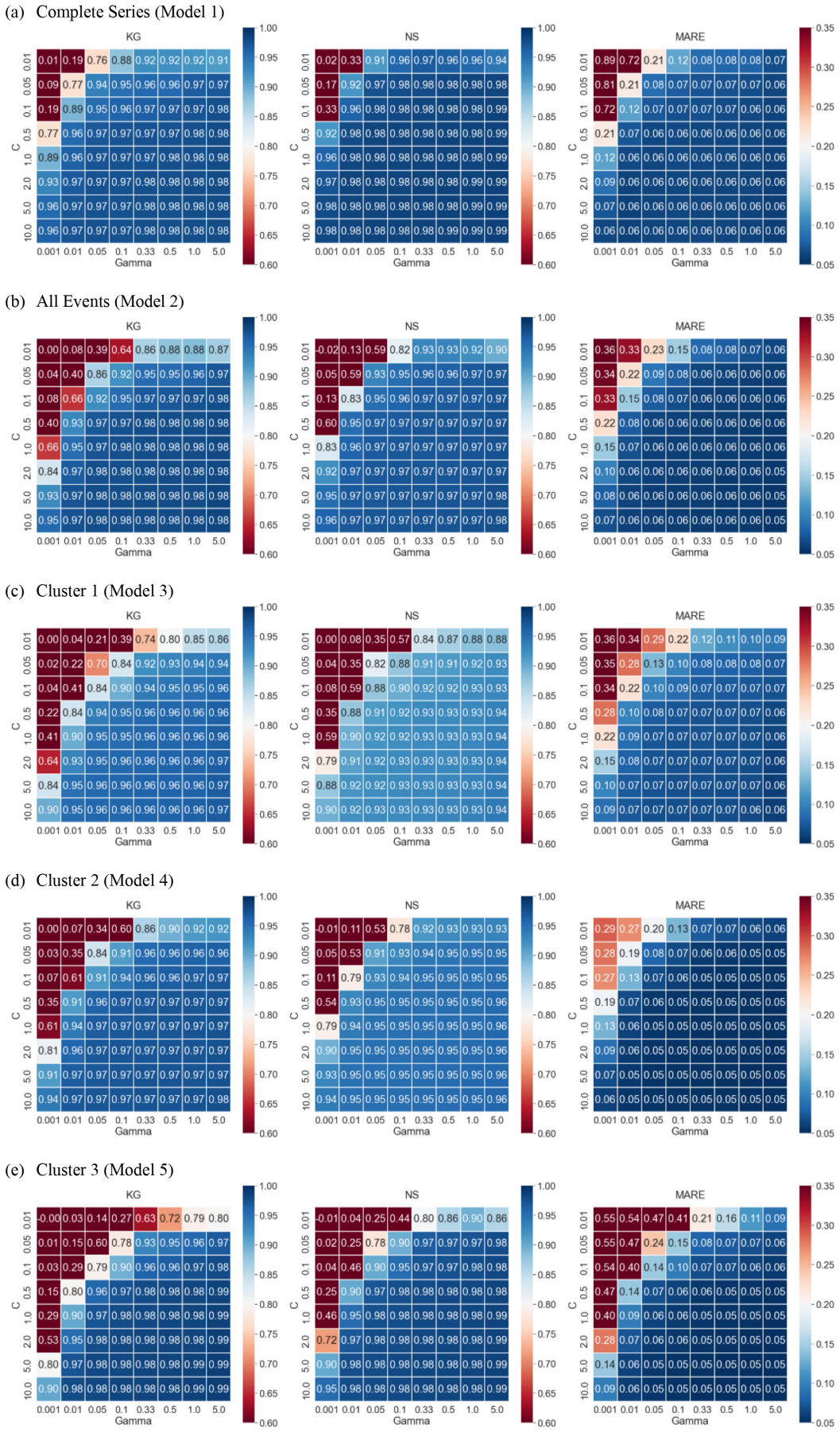

Figure 5. Average variation of KG, NS, and MARE coefficients for parameter sets of data series during the training phase.

minimize the risk of underfitting or overfitting. According to the literature, large γ values tend to produce overfitting, while large *C* values lead to models that are less tolerant to misclassifications, resembling overfitting. Consequently, parameter sets with values located in the center of the heatmaps generally produce better results in the training stage.

Based on the results of the training phase (Figure 5), parameter sets within the first and third quantiles were selected for validation phase. For each training sample, 32 parameter sets were selected, and then applied to their respective validation samples. The parameter sets that were common across all samples within a single model are shown in Figure 6.

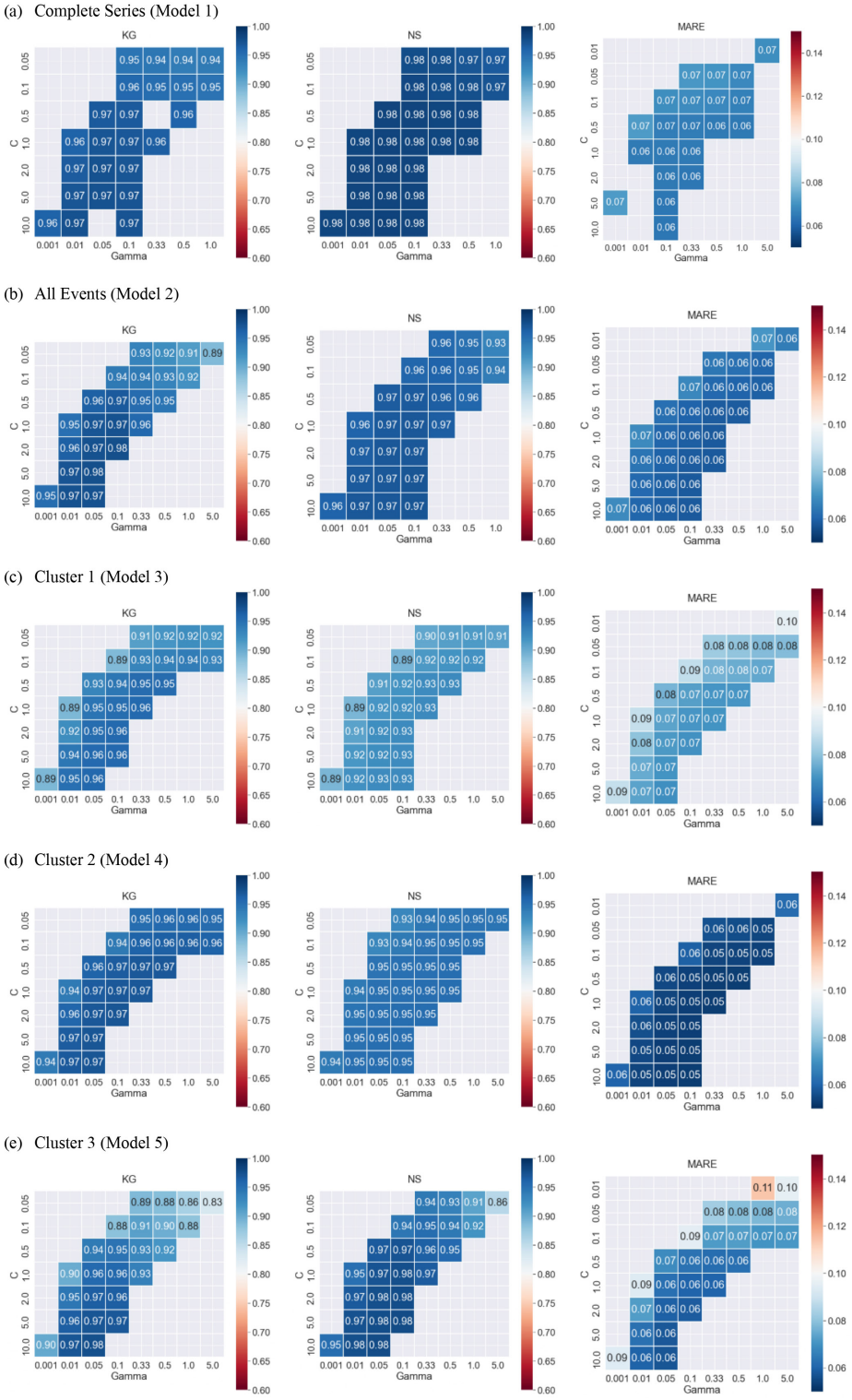

Figure 6. Average variation of KG, NS, and MARE coefficients for parameter sets of data series during the validation phase.

As expected, this procedure led to a reduction in the number of outliers. Despite the smaller number of combinations compared to training phase, validation period exhibited greater variability in the results for all models. For Clusters 1 and 3 Models (Figures 6c and 6e), the variability between the corresponding parameter sets in both phases was even more pronounced, indicating that, for instance, Cluster 3 Model – dealing with extreme flow events – is more sensitive to the selection of events considered at each stage. To illustrate susceptibility to the choice of events at each stage, Table 7 ranks the peak discharges of Cluster 3 events at União da Vitória for the period from 1964 to 2004.

Table 8 shows the peak dates of the events included in each sample and helps interpret the sample metrics by indicating which Cluster 3 events were used in each validation sample. Highlighted colors in both tables represent the presence of events from Table 7 in the samples listed in Table 8.

The highest observed peak (on July 18th, 1983) appeared in Samples 5 to 8, while the lowest peak (on February 04th, 1997) was present in Samples 1, 2, 3, and 8. In other words, both the maximum and minimum extreme values are found simultaneously only in Sample 8 during the validation stage phase. This variability in the samples was also observed during the training phase.

Adjusted models

The set of parameters with C = 0.5 and γ = 0.33 was centered in the heatmaps and appears in all models shown in Figure 6. For the reasons previously mentioned, these parameter values were adopted across all models. Model evaluation measures for the period from 1964 to 2004 (training and validation phases) are shown in Table 9.

The results along with the diagonals (in bold) indicate the outcomes when simulating the same series used to train the model. The results underlined represent models with metrics lower than 0.90 for KG and NS, and higher than 0.10 for MARE. As before, better results were expected when a series was estimated using its own model – trained with that specific category of

Table 7. Ranking of Cluster 3 peak discharges at União da Vitória from 1964 to 2004.

Ord.	Peak date	Peak discharge (m^3/s)	Ord.	Peak date	Peak discharge (m^3/s)
1st	*07/18/1983*	5156.7	9th	*11/27/1982*	2263.0
2nd	*06/08/1992*	3953.6	10th	*01/27/1990*	2196.0
3rd	*10/06/1993*	2786.2	11th	*05/28/1987*	2167.4
4th	*10/12/1998*	2751.7	12th	*10/11/2001*	1975.6
5th	*01/22/1995*	2453.3	13th	*01/01/1981*	1887.1
6th	05/02/1998	2447.2	14th	*07/14/1999*	1832.9
7th	01/13/1971	2428.5	15th	*02/04/1997*	1808.7
8th	09/25/2000	228.0			

Table 8. Sampling strategy for Cluster 3 Events during the validation phase.

Cluster 3 peak date events							
Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**	**Sample 8**
05/28/1987	01/27/1990	01/13/1971	01/13/1971	01/13/1971	01/13/1971	01/01/1981	01/01/1981
01/27/1990	*06/08/1992*	*01/22/1995*	09/25/2000	01/01/1981	01/01/1981	11/27/1982	11/27/1982
06/08/1992	*10/06/1993*	*02/04/1997*	10/11/2001	11/27/1982	11/27/1982	*07/18/1983*	*07/18/1983*
10/06/1993	*01/22/1995*	05/02/1998		*07/18/1983*	*07/18/1983*	05/28/1987	05/28/1987
01/22/1995	*02/04/1997*	*10/12/1998*			05/28/1987	01/27/1990	01/27/1990
02/04/1997	05/02/1998	*07/14/1999*				*06/08/1992*	*06/08/1992*
05/02/1998	*10/12/1998*	09/25/2000				*10/06/1993*	*10/06/1993*
10/12/1998	*07/14/1999*	10/11/2001					*01/22/1995*
07/14/1999	09/25/2000						*02/04/1997*
09/25/2000	10/11/2001						05/02/1998
10/11/2001							*10/12/1998*

series – compared to simulations conducted with models trained on different data classes (off-diagonal values).

First, analyzing the results for each data series, Table 9 shows that: (i) the Complete Series was best simulated by Models 1 (its own model), 2, and 5; (ii) the All Events Series was well estimated by all models, though Model 3 showed poorer results; (iii) Cluster 1 Series were well represented by all models; (iv) Cluster 2 Series were also well represented by all models; and (v) Cluster 3 Series were well represented by all models except Model 3. Second, analyzing the results for each model, Table 9 shows that: (i) Models 1, 2, and 5 provided good results for all series; (ii) Model 3 provided good results only for Clusters 1 and 2; and (iii) Model 4 provided good results for all series except the Complete Series.

Table 10 shows the ratio between statistics from the simulated and observed series. Models' behavior can also be statistically observed, as shown in Table 10, where simulated minimum discharges differ significantly from the observed ones, even in series forecasted by their own models. Simulated maximum discharges remained close to the observed ones. For the other statistics, the main errors were identified in simulations performed by Models 3 and 4 for the Complete and All Events series.

Table 9. KG, NS, and MARE coefficients for final models' simulations from 1964 to 2004.

		Simulated series				
		Complete	**All Events**	**Cluster 1**	**Cluster 2**	**Cluster 3**
KG	**M1**	**0.975**	0.984	0.949	0.971	0.982
	M2	0.954	**0.978**	0.934	0.975	0.983
	M3	0.585	0.817	**0.960**	0.930	0.882
	M4	0.885	0.949	0.958	**0.972**	0.962
	M5	0.948	0.977	0.933	0.973	**0.980**
NS	**M1**	**0.985**	0.974	0.923	0.952	0.982
	M2	0.983	**0.973**	0.919	0.951	0.982
	M3	0.766	0.905	**0.931**	0.953	0.963
	M4	0.968	0.971	0.930	**0.953**	0.979
	M5	0.982	0.973	0.917	0.950	**0.981**
MARE	**M1**	**0.063**	0.062	0.084	0.058	0.053
	M2	0.139	**0.058**	0.088	0.059	0.055
	M3	0.896	0.205	**0.069**	0.067	0.144
	M4	0.275	0.075	0.070	**0.052**	0.064
	M5	0.150	0.066	0.088	0.060	**0.053**

Table 10. Statistical relationships between simulated series by final models and observed series, from 1964 to 2004.

		Simulated / Observed series										
		Complete	**All Events**	**Cluster 1**	**Cluster 2**	**Cluster 3**			**Complete**	**All Events**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Minimum	**M1**	1.19	0.90	0.85	1.05	1.07	**Standard - deviation**	**M1**	0.98	0.99	1.04	0.98	0.99
	M2	2.16	1.07	0.94	1.07	1.36		**M2**	0.96	0.99	1.05	0.99	0.99
	M3	8.26	3.03	1.49	1.34	3.20		**M3**	0.80	0.86	0.99	0.93	0.89
	M4	3.52	1.60	1.09	1.14	1.77		**M4**	0.91	0.95	1.02	0.99	0.97
	M5	2.31	1.25	0.98	1.10	1.35		**M5**	0.95	0.98	1.05	0.99	0.99
Maximum	**M1**	1.00	1.00	0.95	0.94	1.00	**Skewness**	**M1**	1.02	0.98	0.79	0.93	0.99
	M2	0.99	0.99	0.95	0.94	0.99		**M2**	1.07	1.00	0.80	0.93	0.99
	M3	0.95	0.96	0.97	0.96	0.96		**M3**	1.32	1.20	1.04	1.14	1.17
	M4	0.99	0.99	0.96	0.96	0.99		**M4**	1.19	1.11	0.95	1.03	1.10
	M5	0.98	0.98	0.95	0.94	0.98		**M5**	1.08	1.01	0.79	0.95	0.99
Mean	**M1**	0.98	1.00	1.00	0.99	0.99	**Volume**	**M1**	0.98	1.00	1.00	0.99	0.99
	M2	1.02	0.99	1.01	1.00	0.99		**M2**	1.02	0.99	1.01	1.00	0.99
	M3	1.36	1.12	0.98	1.01	1.04		**M3**	1.36	1.12	0.98	1.01	1.04
	M4	1.07	1.00	0.98	0.99	0.98		**M4**	1.07	1.00	0.98	0.99	0.98
	M5	1.02	1.00	1.01	1.00	0.99		**M5**	1.02	1.00	1.01	1.00	0.99

Thus, the estimations for Complete and All Events Series using all five models were compared. Model 3 performed poorly in predicting both series, such that:

1. KG values were reduced due to the greater distances between the simulated mean discharges and standard deviations compared to the respective observed parameters. This is because the KG coefficient is calculated based on the mean discharges, variability (standard deviation), and dynamics (correlation).
2. NS values were lower due to the difficulty in predicting mean discharges, as the coefficient prioritizes maximum flows over mean and low flows. This behavior is also reflected in the volume ratio, which is greater than 1.0, indicating an overestimation of the low flows, as shown in Table 10.
3. MARE values increased due to the greater differences between simulated minimum and mean flows compared to the observed parameters, as MARE prioritizes low and mean flows.

Discharge prediction

All five data samples were predicted by the five models for the 2005 to 2015 period (testing phase) with a one-day lead time. The measures used to evaluate the forecasts of the series are shown in Table 11.

In Table 11, the main diagonals – highlighted in bold – indicate the results obtained when forecasting the same series used to train the model. The underlined results represent models with metrics lower than 0.90 for KG and NS, and higher than 0.10 for MARE. The same combinations of series and models highlighted in Table 9 are also prominent in Table 11; the forecast measures performed similarly to the simulations in previous stage.

As observed in final models, the performance of the forecasts for the main diagonals (Table 11) was satisfactory when compared within their matching columns. Simulating a series with a model other than its own was expected to perform worse than using the model that was trained with that specific series. Once again, Models 3 and 4 stand out for having the worst performances when forecasting Complete and All Events Series.

Models' behavior can also be statistically observed (Table 12). A small difference between the results of the final validation and the forecast phases was noted, indicating that the observed differences in statistical relationships between those two phases were minimal. In other words, the simulated statistics well represented the observed ones, except for the minimum discharges. Greater variability was observed in mean discharges, standard deviations, skewness, and volumes for simulations made by Model 3, particularly for the Complete and All Events Series.

Thus, as observed in previous stage (cross-validation), the estimations for Complete and All Events Series using all five models were exclusively compared and highlighted in Table 12. Model 3 performed poorly in predicting Complete and All Events Series, such that: (i) the relationship between means and standard deviations explains the behavior of KG; (ii) the relationship between volumes and means accounts for NS values; and (iii) the relationship between minimum and mean discharges suggests changes in

Table 11. KG, NS, and MARE coefficients for series predictions from 2005 to 2015 – Testing phase.

		Predicted series				
		Complete	**All Events**	**Cluster 1**	**Cluster 2**	**Cluster 3**
KG	**M1**	**0.973**	0.976	0.974	0.968	0.967
	M2	0.949	**0.936**	0.970	0.970	0.964
	M3	0.619	0.820	**0.958**	0.925	0.875
	M4	0.892	0.949	0.972	**0.974**	0.960
	M5	0.944	0.967	0.969	0.967	**0.967**
NS	**M1**	**0.984**	0.968	0.950	0.948	0.964
	M2	0.982	**0.967**	0.949	0.948	0.964
	M3	0.794	0.903	**0.955**	0.951	0.942
	M4	0.968	0.964	0.954	**0.951**	0.958
	M5	0.981	0.968	0.948	0.948	**0.963**
MARE	**M1**	**0.068**	0.068	0.082	0.067	0.058
	M2	0.141	**0.064**	0.084	0.068	0.060
	M3	0.873	0.201	**0.072**	0.074	0.113
	M4	0.275	0.082	0.071	**0.062**	0.072
	M5	0.156	0.073	0.085	0.069	**0.057**

MARE. Therefore, improvements in statistical relationships also indicate improvements in quality metrics.

Based on the previous results, the hydrographs obtained from forecasts for the Complete and All Events Series – simulated by their own models and by the models per group – are shown in Figures 7 and 8. For the Complete Series (Figure 7), as observed for the 1964 to 2004 period, Model 5's forecasts were very close to those of Model 1, and consequently, to the observed flows – all three hydrographs overlapped. In contrast, Models 3 and 4, while not diverging significantly from peak discharges, displayed a notable offset from low flows.

For the All Events Series (Figure 8), Model 5's forecasts closely matched Model 2's for both peak and low flows, with both hydrographs overlapping the observed data. However, Model 3 overestimated smaller peaks and underestimated larger ones. Compared to the final validation stage, the All Events Series predicted by Model 3 showed the greatest behavioral difference, along with a slight decline in performance metrics.

Table 12. Statistical relationships between predicted and observed series, from 2005 to 2015 – Testing phase.

		Predicted / observed series										
		Complete	All Events	Cluster 1	Cluster 2	Cluster 3			Complete	All Events	Cluster 1	Cluster 2	Cluster 3
Minimum	**M1**	1.24	1.22	1.16	1.11	1.04	**Standard - deviation**	**M1**	0.98	0.98	1.01	0.98	0.98
	M2	2.23	1.43	1.28	1.13	1.21		**M2**	0.95	0.98	1.02	0.99	0.97
	M3	8.62	3.73	1.91	1.46	2.36		**M3**	0.80	0.85	0.97	0.93	0.88
	M4	3.64	2.04	1.45	1.25	1.45		**M4**	0.91	0.95	1.01	1.00	0.97
	M5	2.44	1.60	1.29	1.18	1.19		**M5**	0.95	0.97	1.02	0.98	0.98
Maximum	**M1**	0.96	0.97	1.00	0.95	0.96	**Skewness**	**M1**	1.00	0.95	0.74	0.92	0.93
	M2	0.96	0.96	1.00	0.95	0.96		**M2**	1.05	0.98	0.72	0.91	0.98
	M3	0.99	0.99	0.97	0.97	0.99		**M3**	1.27	1.27	1.00	1.09	1.46
	M4	1.01	1.02	0.99	0.98	1.01		**M4**	1.17	1.17	0.95	1.00	1.31
	M5	0.97	0.97	0.99	0.95	0.97		**M5**	1.06	1.01	0.71	0.92	1.02
Mean	**M1**	0.99	1.00	1.00	1.00	0.99	**Volume**	**M1**	0.98	1.00	1.00	1.00	0.99
	M2	1.02	0.98	1.01	1.00	0.99		**M2**	1.02	0.98	1.01	1.00	0.99
	M3	1.32	1.10	0.98	1.01	1.04		**M3**	1.32	1.10	0.98	1.01	1.04
	M4	1.06	1.00	0.98	0.99	0.98		**M4**	1.06	1.00	0.98	0.99	0.98
	M5	1.02	1.00	1.01	1.01	0.98		**M5**	1.02	1.00	1.01	1.01	0.98

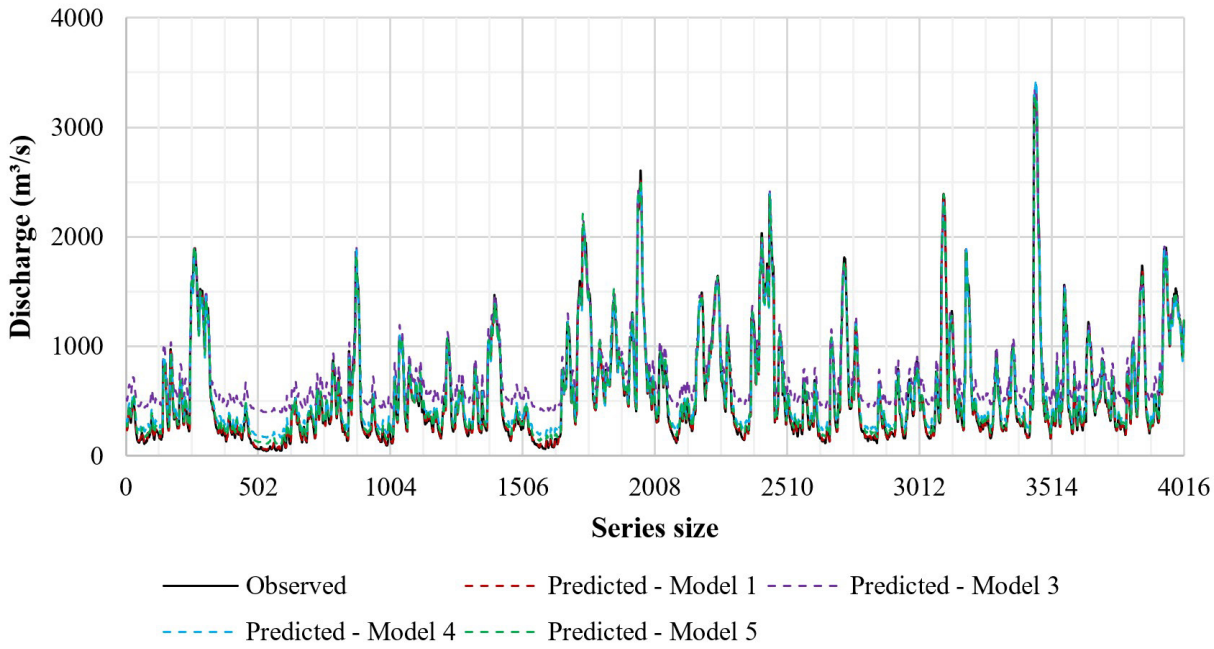

Figure 7. Predicted and observed hydrographs from 2005 to 2015 – Complete Series – Testing phase.

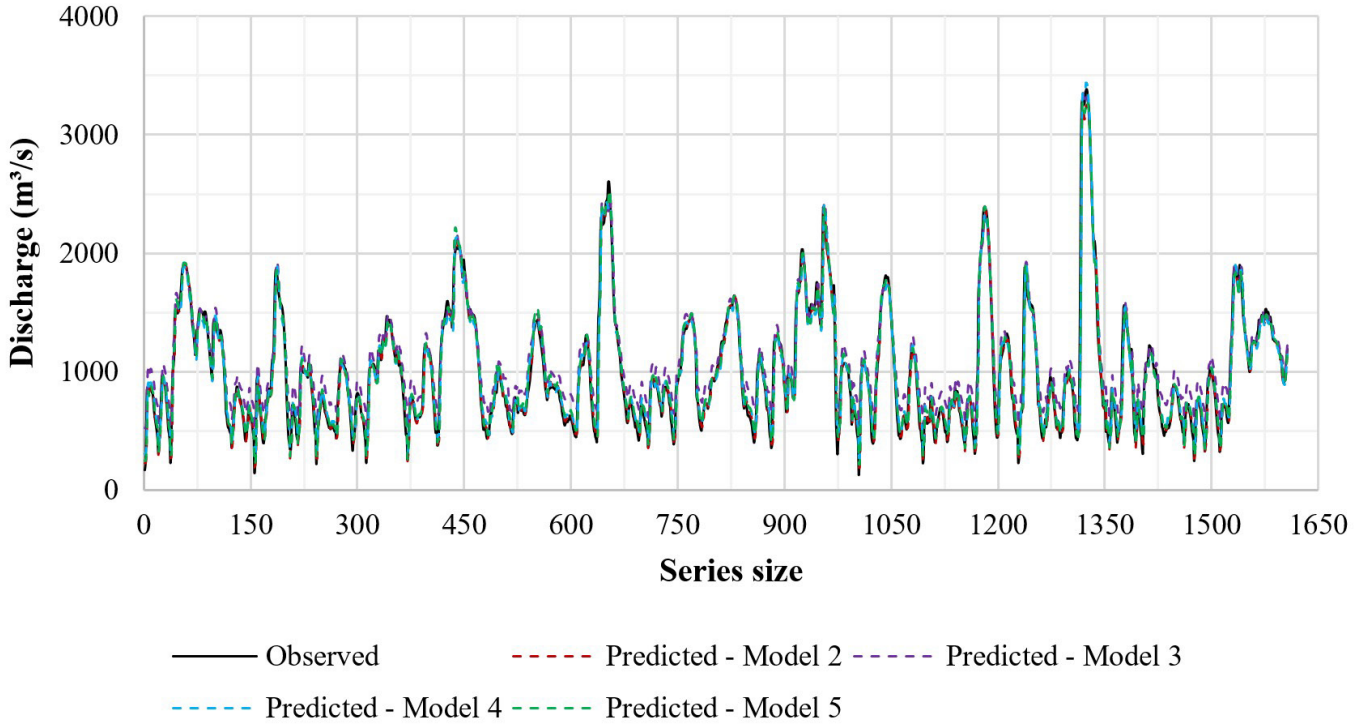

Figure 8. Predicted and observed hydrographs from 2005 to 2015 – All Events Series – Testing phase.

CONCLUSIONS

This paper proposes a new hydrological modeling approach for discharge prediction, based on previous flood events characterization through flood events clustering. The proposed hydrological modeling approach is illustrated with a real-world application. The present paper discusses only the results from the training and testing phases and does not present an operational forecasting flow model, which will be the subject of a future paper (in elaboration).

Daily mean discharges from 1964 to 2015 in União da Vitória (Iguaçu River basin, Paraná State, Brazil) were used, and the SVR algorithm was applied to create AI models for 1-day discharge forecasts. The results obtained are a function of the clustering itself, the chosen hydrological model, and the adopted forecast horizon. Once again, we highlight that the proposed approach allows the use of different hydrological models.

Due to the prior application of the FCM algorithm, the results revealed that all events share similarities with each cluster to varying degrees. This overlap may negatively impact the performance of the cluster models.

Therefore, clustering effectively helps identify similar patterns and characteristics within event groups. It reduces averages, standard deviations, and skewness for the first clusters while increasing these parameters for Cluster 3, due to its events' greater magnitude. The centroid coordinates further highlight the larger scale of events in the last group.

Training, validation, and testing periods were subdivided to distribute major events across these periods. The metric results, facing the sampling strategy, revealed that the presence or absence of certain events in the training or validation phases often led to significant differences between samples. During the validation stage, due to the smaller number of parameters sets, some samples showed less variability in their results, except for Models 1, 2, and 5, particularly in Samples 5 to 8.

As expected, both in the definition of final models and during testing stage, models performed better when simulating series from the same category used for calibration. Model 3 performed worse with Complete and All Events Series, which is justified by the disparity between the mean discharges of simulated and observed flows. This disparity impacted all three metrics, with the standard deviation having a greater influence on KG.

However, these results may have been influenced by the events within each group generated during clustering, a limitation of the research. To reduce these impacts, one option is to investigate the final fuzzy partition matrix from FCM to weight events with low membership degrees, increasing the sharing information with other clusters. In the same context, another approach could be to eliminate these events from the analysis, keeping only those events that predominantly look like a single cluster. However, the latter consideration could significantly reduce the number of events available for defining the calibration and forecasting models by group, especially for Cluster 3.

Other research limitations include the process of defining the start and end of events, which is not well established in literature. In this work, it was done empirically based on the local characteristics of the União da Vitória gauging station. This issue directly influences the determination of flood indicators, which could lead to changes in clustering and, consequently, in the modeling process. Additionally, incorporating rainfall-related indicators could result in similar changes.

Topics related to the characteristics of the Iguaçu River can also be discussed as limitations, including: (i) the topographical

conditions of the Iguaçu River Basin, particularly between Fluviópolis and União da Vitória, where the river flows through a main channel constrained by low and flat banks, that are more prone to overflow during high runoff volumes (Sociedade de Estudos Contemporâneos – Comissão Regional Permanente de Prevenção Contra Enchentes do Rio Iguaçu, 1999); and (ii) the proximity of both gauging stations, which necessitates more frequent discharge data than daily measurements.

On the other hand, all the research limitations mentioned here can be seen as opportunities for further development. The approach of combining clustering with hydrological modeling can be applied to both traditional and data-driven models. In future studies, for example, we suggest applying different scaling methods to the flow series and applying the Principal Component Analysis (PCA) to define the indicators used in both the clustering procedure and the machine learning process.

In the context of flow data scale, we suggest applying the methodological processes presented here to a study area with compatible data availability. This means that the smaller the river basins under analysis, the smaller the data interval should be.

Based on flood indicators and a certain number of clusters as suggested in this research, a forecasting system based on groups of events could perform by reducing the forecasting possibilities, such that: (i) at the initial instant (t_0), a flood event is identified; (ii) at the next time point (t_1), the peak time and flow of the ongoing event are determined, and forecasts for the following time instant are made using the three models per group; (iii) as the event progresses (t_2, t_3, and the next ones), peak time and flow are updated when there is an overflow, and the forecasts tend to be closer to a particular group, restricting the forecasting possibilities.

Therefore, studying the process of forecasting by groups in real-time is necessary and represents a very intriguing challenge, so that, once the start of an event has been identified, the observed flows during the flood development can be used to identify the similarities of the events with groups, restricting the characteristics of the flood in progress. One possibility is using the models by group integrated with a fuzzy analysis to forecast the discharges.

REFERENCES

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 (Amsterdam)*, *577*, 123981. http://doi.org/10.1016/j.jhydrol.2019.123981.

Agência Nacional de Águas e Saneamento Básico – ANA. (2023). *Portal HidroWeb*. Brasília: ANA. Retrieved in 2023, August 15, from: https://www.snirh.gov.br/hidroweb/serieshistoricas

Althoff, D., Rodrigues, L. N., & Bazame, H. C. (2021). Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble. *Stochastic Environmental Research and Risk Assessment*, *35*(5), 1051-1067. http://doi.org/10.1007/s00477-021-01980-8.

Askari, S. (2021). Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: review and development. *Expert Systems with Applications*, *165*, 1-27. http://doi.org/10.1016/j.eswa.2020.113856.

Bai, P., Liu, X., & Xie, J. (2021). Simulating runoff under changing climatic conditions: a comparison of the long short-term memory network with two conceptual hydrologic models. *Journal of Hydrology (Amsterdam)*, *592*, 1-11. http://doi.org/10.1016/j.jhydrol.2020.125779.

Brêda, J. P. L. F., Paiva, R. C. D., Pedrollo, O. C., Passaia, O. A., & Collischonn, W. (2021). Modeling coordinated operation of multiple hydropower reservoirs at a continental scale using artificial neural network: the case of Brazilian hydropower system. *Revista Brasileira de Recursos Hídricos*, *26*, e12. http://doi.org/10.1590/2318-0331.262120210011.

Chang, C.-C., & Lin, C.-J. (2001). Training ν-support vector classifiers: theory and algorithms. *Neural Computation*, *13*(9), 2119-2147. http://doi.org/10.1162/089976601750399335.

Chang, C.-C., & Lin, C.-J. (2022). *LIBSVM: a library for Support Vector Machines*. Retrieved in 2023, August 15, from: https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf

Chen, S., Ren, M., & Sun, W. (2021). Combining two-stage decomposition based machine learning methods for annual runoff forecasting. *Journal of Hydrology (Amsterdam)*, *603*(Pt B), 126945. http://doi.org/10.1016/j.jhydrol.2021.126945.

Desai, S., & Ouarda, T. B. M. J. (2021). Regional hydrological frequency analysis at ungauged sites with random forest regression. *Journal of Hydrology (Amsterdam)*, *594*, 125861. http://doi.org/10.1016/j.jhydrol.2020.125861.

Difi, S., Elmeddahi, Y., Hebal, A., Singh, V. P., Heddam, S., Kim, S., & Kisi, O. (2023). Monthly streamflow prediction using hybrid extreme learning machine optimized by bat algorithm: a case study of Cheliff watershed, Algeria. *Hydrological Sciences Journal*, *68*(2), 189-208. http://doi.org/10.1080/02626667.2022.2149334.

Ding, Y., Zhu, Y., Feng, J., Zhang, P., & Cheng, Z. (2020). Interpretable spatio-temporal attention LSTM model for flood forecasting. *Neurocomputing*, *403*, 348-359. http://doi.org/10.1016/j.neucom.2020.04.110.

Ebtehaj, I., & Bonakdari, H. (2022). A reliable hybrid outlier robust non-tuned rapid machine learning model for multi-step ahead flood forecasting in Quebec, Canada. *Journal of Hydrology (Amsterdam)*, *614*, 1-21. http://doi.org/10.1016/j.jhydrol.2022.128592.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. http://doi.org/10.1016/j.engappai.2022.104743.

Fathian, F., Mehdizadeh, S., Sales, A. K., & Safari, M. J. S. (2019). Hybrid models to improve the monthly river flow prediction:

integrating artificial intelligence and non-linear time series models, and future research prospects. *Journal of Hydrology (Amsterdam)*, *575*, 1200-1213. http://doi.org/10.1016/j.jhydrol.2019.06.025.

Feng, Z., Shi, P., Yang, T., Niu, W., Zhou, J., & Cheng, C. (2022). Parallel cooperation search algorithm and artificial intelligence method for streamflow time series forecasting. *Journal of Hydrology (Amsterdam)*, *606*, 1-14. http://doi.org/10.1016/j.jhydrol.2022.127434.

Freitas, C. (2016). *Uso de técnicas de classificação de dados na operação de sistemas de reservatórios de usinas hidrelétricas durante cheias* (Tese de doutorado). Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental, Universidade Federal do Paraná, Curitiba.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology (Amsterdam)*, *377*(1-2), 80-91. http://doi.org/10.1016/j.jhydrol.2009.08.003.

Hallouin, T., Bruen, M., & O'Loughlin, F. E. (2020). Calibration of hydrological models for ecologically relevant streamflow predictions: a trade-off between fitting well to data and estimating consistent parameter sets? *Hydrology and Earth System Sciences*, *24*(3), 1031-1054. http://doi.org/10.5194/hess-24-1031-2020.

Ibrahim, K. S. M. H., Huang, Y. F., Ahmed, A. N., Koo, C. H., & El-Shafie, A. (2022). A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alexandria Engineering Journal*, *61*(1), 279-303. http://doi.org/10.1016/j.aej.2021.04.100.

Islam, A. R. M. T., Talukdar, S., Mahato, S., Kundu, S., Eibek, K. U., Pham, Q. B., Kuriqi, A., & Linh, N. T. T. (2021). Flood susceptibility modelling using advanced ensemble machine learning models. *Geoscience Frontiers*, *12*(3), 1-18. http://doi.org/10.1016/j.gsf.2020.09.006.

Joo, H., Lee, M., Kim, J., Jung, J., Kwak, J., & Kim, H. S. (2021). Stream gauge network grouping analysis using community detection. *Stochastic Environmental Research and Risk Assessment*, *35*(4), 781-795. http://doi.org/10.1007/s00477-020-01916-8.

Khosravi, K., Golkarian, A., Booij, M. J., Barzegar, R., Sun, W., Yaseen, Z. M., & Mosavi, A. (2021). Improving daily stochastic streamflow prediction: comparison of novel hybrid data-mining algorithms. *Hydrological Sciences Journal*, *66*(9), 1457-1474. http://doi.org/10.1080/02626667.2021.1928673.

Kim, D., Lee, J., Kim, J., Lee, M., Wang, W., & Kim, H. S. (2022). Comparative analysis of long short-term memory and storage function model for flood water level forecasting of Bokha stream in NamHan River, Korea. *Journal of Hydrology (Amsterdam)*, *606*, 127415. http://doi.org/10.1016/j.jhydrol.2021.127415.

Kurian, C., Sudheer, K. P., Vema, V. K., & Sahoo, D. (2020). Effective flood forecasting at higher lead times through hybrid modelling framework. *Journal of Hydrology (Amsterdam)*, *587*, 124945. http://doi.org/10.1016/j.jhydrol.2020.124945.

Lappicy, T., & Lima, C. H. R. (2023). Enhancing monthly streamflow forecasting for Brazilian hydropower plants through climate index integration with stochastic methods. *Revista Brasileira de Recursos Hídricos*, *28*, e48. http://doi.org/10.1590/2318-0331.282320230118.

Lawin, A. E., Houngue, R., Po, Y., Houngue, N. R., Attogouinon, A., & Afouda, A. A. (2019). Mid-century climate change impacts on ouémé river discharge at bonou outlet (Benin). *Hydrology*, *6*(72), 1-20. http://doi.org/10.3390/hydrology6030072.

Letessier, C., Cardi, J., Dussel, A., Ebtehaj, I., & Bonakdari, H. (2023). Enhancing flood prediction accuracy through integration of meteorological parameters in River Flow observations: a case study Ottawa River. *Hydrology*, *10*(8), 164. http://doi.org/10.3390/hydrology10080164.

Li, X., Sha, J., & Wang, Z.-L. (2019). Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal*, *64*(15), 1857-1866. http://doi.org/10.1080/02626667.2019.1680846.

Liang, Z., Xiao, Z., Wang, J., Sun, L., Li, B., Hu, Y., & Wu, Y. (2019). An improved chaos similarity model for hydrological forecasting. *Journal of Hydrology (Amsterdam)*, *577*, 123953. http://doi.org/10.1016/j.jhydrol.2019.123953.

Lima, G. R. T., & Scofield, G. B. (2021). Feasibility study on operational use of neural networks in a flash flood early warning system. *Revista Brasileira de Recursos Hídricos*, *26*, e7. http://doi.org/10.1590/2318-0331.262120200152.

Luppichini, M., Barsanti, M., Giannecchini, R., & Bini, M. (2022). Deep learning models to predict flood events in fast-flowing watersheds. *The Science of the Total Environment*, *813*, 151885. PMid:34826469. http://doi.org/10.1016/j.scitotenv.2021.151885.

Luppichini, M., Favalli, M., Isola, I., Nannipieri, L., Giannecchini, R., & Bini, M. (2019). Influence of topographic resolution and accuracy on hydraulic channel flow simulations: case study of the Versilia River (Italy). *Remote Sensing (Basel)*, *11*(13), 1630. http://doi.org/10.3390/rs11131630.

Mahdavi-Meymand, A., Sulisz, W., & Zounemat-Kermani, M. (2023). Hybrid and integrative evolutionary Machine Learning in hydrology: a systematic review and metaanalysis. *Archives of Computational Methods in Engineering*, http://doi.org/10.1007/s11831-023-10017-y.

Mine, M. R. M., & Tucci, C. E. M. (2002). Gerenciamento da produção de energia e controle de inundação: Foz de Areia no rio Iguaçu. *RBRH*, *7*(3), 85-107. http://doi.org/10.21168/rbrh.v7n3.p85-107.

Mohammadi, B., Linh, N. T. T., Pham, Q. B., Ahmed, A. N., Vojteková, J., Guan, Y., Abba, S. I., & El-Shafie, A. (2020). Adaptive neuro-fuzzy

inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series. *Hydrological Sciences Journal*, *65*(10), 1738-1751. http://doi.org/10.1080/02626667.2020.1758703.

Morettin, P. A., & Singer, J. M. (2023). *Estatística e ciência de dados*. Rio de Janeiro: LTC.

Mosavi, A., Golshan, M., Choubin, B., Ziegler, A. D., Sigaroodi, S. K., Zhang, F., & Dineva, A. A. (2021). Fuzzy clustering and distributed model for streamflow estimation in ungauged watersheds. *Scientific Reports*, *11*, 8243. http://doi.org/10.1038/s41598-021-87691-0.

Moura, C. N., Jan, S., & Detzel, D. H. M. (2022). Evaluating the long short-term memory (LSTM) network for discharge prediction under changing climate conditions. *Nordic Hydrology*, *53*(5), 657-667. http://doi.org/10.2166/nh.2022.044.

Niu, W., & Feng, Z. (2021). Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management. *Sustainable Cities and Society*, *64*, 1-12. http://doi.org/10.1016/j.scs.2020.102562.

Ribeiro, V. H. A., Reynoso-Meza, G., & Siqueira, H. V. (2020). Multi-objective ensembles of echo state networks and extreme learning machines for streamflow series forecasting. *Engineering Applications of Artificial Intelligence*, *95*, 1-19. http://doi.org/10.1016/j.engappai.2020.103910.

Rocha, P. S. M. (2012). *Gestão em áreas de risco de enchentes: estudo de caso para União da Vitória – Paraná* (Dissertação de mestrado). Universidade Positivo, Curitiba.

Roy, A., & Chakraborty, S. (2023). Support Vector Machine in structural reliability analysis: a review. *Reliability Engineering & System Safety*, *233*, 109126. http://doi.org/10.1016/j.ress.2023.109126.

Saadi, M., Oudin, L., & Ribstein, P. (2019). Random Forest ability in regionalizing hourly hydrological model parameters. *Water (Basel)*, *11*(18), 1540. http://doi.org/10.3390/w11081540.

Samantaray, S., Sahoo, P., Sahoo, A., & Satapathy, D. P. (2023). Flood discharge prediction using improved ANFIS model combined with hybrid particle swarm optimization and slime mould algorithm. *Environmental Science and Pollution Research International*, *30*(35), 83845-83872. PMid:37351742. http://doi.org/10.1007/s11356-023-27844-y.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*(5), 1207-1245. http://doi.org/10.1162/089976600300015565.

Schoppa, L., Disse, M., & Bachmair, S. (2020). Evaluating the performance of random forest for large-scale flood discharge simulation. *Journal of Hydrology (Amsterdam)*, *590*, 1-13. http://doi.org/10.1016/j.jhydrol.2020.125531.

Sharma, R. K., Kumar, S., Padmalal, D., & Roy, A. (2023). Streamflow prediction using machine learning models in selected rivers of Southern India. *International Journal of River Basin Management*, *22*(4), 529-555. https://doi.org/10.1080/15715124.2023.2196635.

Shukla, R., Kumar, P., Vishwakarma, D. K., Ali, R., Kumar, R., & Kuriqi, A. (2022). Modeling of stagedischarge using back propagation ANN, ANFIS, and WANNbased computing techniques. *Theoretical and Applied Climatology*, *147*(3-4), 867-889. http://doi.org/10.1007/s00704-021-03863-y.

Singh, V. P. (2018). Hydrologic modeling: progress and future directions. *Geoscience Letters*, *5*(15), 1-18. http://doi.org/10.1186/s40562-018-0113-z.

Snieder, E., Shakir, R., & Khan, U. T. (2020). A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. *Journal of Hydrology (Amsterdam)*, *583*, 124299. http://doi.org/10.1016/j.jhydrol.2019.124299.

Sociedade de Estudos Contemporâneos – Comissão Regional Permanente de Prevenção Contra Enchentes do Rio Iguaçu – SEC-CORPRERI. (1999). *Conhecendo e Convivendo com Enchentes*. União da Vitória: SEC-CORPRERI.

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., & Tallaksen, L. M. (2011). Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences*, *15*(11), 3447-3459. http://doi.org/10.5194/hess-15-3447-2011.

Steffen, P. C., & Gomes, J. (2018). Clustering of historical floods observed on Iguaçu River, in União da Vitória, Paraná. *Revista Brasileira de Recursos Hídricos*, *23*(38), 1-12. http://doi.org/10.1590/2318-0331.231820170107.

Tarasova, L., Merz, R., Kiss, A., Basso, S., Bloschl, G., Merz, B., Viglione, A., Plotner, S., Guse, B., Schumann, A., Fischer, S., Ahrens, B., Anwar, F., Bardossy, A., Buhler, P., Haberlandt, U., Kreibich, H., Krug, A., Lun, D., Muller-Thomy, H., Pidoto, R., Primo, C., Seidel, J., Vorogushyn, S., & Wietzke, L. (2019). Causative classification of river flood events. *WIREs. Water*, *6*(4), e1353. http://doi.org/10.1002/wat2.1353.

Tozzi, B. K. M., & Fill, H. D. O. A. (2020). Verification of the stationarity of flow series in the Iguaçu River basin. *Revista Brasileira de Recursos Hídricos*, *25*(10), 1-9. http://doi.org/10.1590/2318-0331.252020180171.

Unduche, F., Habtamu, T., Senbeta, D., & Zhu, E. (2018). Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed. *Hydrological Sciences Journal*, *63*(8), 1133-1149. http://doi.org/10.1080/02626667.2018.1474219.

Wang, X., Wang, Y., Yuan, P., Wang, L., & Cheng, D. (2021). An adaptive daily runoff forecast model using VMD-LSTM-PSO hybrid approach. *Hydrological Sciences Journal*, *66*(9), 1488-1502. http://doi.org/10.1080/02626667.2021.1937631.

Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: an introductory overview. *WIREs. Water*, *8*(5), e1533. http://doi.org/10.1002/wat2.1533.

Yaseen, Z. M., Faris, H., & Al-Ansari, N. (2020). Hybridized Extreme Learning Machine Model with Salp Swarm Algorithm: a novel predictive model for hydrological application. *Complexity*, *2020*(1), 8206245. http://doi.org/10.1155/2020/8206245.

Zakhrouf, M., Hamid, B., Kim, S., & Madani, S. (2023). Novel insights for streamflow forecasting based on deep learning models combined the evolutionary optimization algorithm. *Physical Geography*, *44*(1), 31-54. http://doi.org/10.1080/02723646.2021.1943126.

Zhang, Z., Zhang, Q., & Singh, V. P. (2018). Univariate streamflow forecasting using commonly used data-driven models: literature review and case study. *Hydrological Sciences Journal*, *63*(7), 1091-1111. http://doi.org/10.1080/02626667.2018.1469756.

Authors contributions

Patrícia Cristina Steffen: Research development, data analysis and interpretation, article writing, general design, final review.

Júlio Gomes: Guidance, proofreading, project supervisor, final review.

Eloy Kaviski: Proofreading, final review.

Daniel Henrique Marco Detzel: Proofreading, final review.

Editor in-Chief: Adilson Pinheiro

Associated Editor: Carlos Henrique Ribeiro Lima