Received: 16 March 2018 | Accepted: 18 December 2018
DOI: 10.1002/hyp.13388

RESEARCH ARTICLE

WILEY

Artificial intelligence for identifying hydrologically homogeneous regions: A state-of-the-art regional flood frequency analysis

Felício Cassalho[1] | Samuel Beskow[1] | Carlos Rogério de Mello[2] | Maíra Martim de Moura[1] | Leroi Floriano de Oliveira[3] | Marilton Sanchotene de Aguiar[3]

[1]Center for Technological Development/Water Resources Engineering, Federal University of Pelotas. Hydrology and Hydrological Modeling Laboratory, Pelotas, Brazil

[2]Water Resources Department, Federal University of Lavras, Lavras, Brazil

[3]Center for Technological Development/Computer Science Graduate Program, Federal University of Pelotas, Pelotas, Brazil

Correspondence
Felício Cassalho, Center for Technological Development/Water Resources Engineering, Federal University of Pelotas. Hydrology and Hydrological Modeling Laboratory, 1 Gomes Carneiro Street, Pelotas, RS 96010-610, Brazil.
Email: felicioufpel@gmail.com

Funding information
Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Grant/Award Numbers: 16/2551-0000 247-9 and 2082-2551/13-0; Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Grant/Award Number: PPM VIII 071/2014; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Grant/Award Number: 88887.178100/2018-00; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant/Award Numbers: 485279/2013-4, 301556/2017-2, 308645/2017-0

Present Address: Felício Cassalho, National Institute for Space Research/Remote Sensing Graduate Program. 1758 dos Astronautas Avenue, Jardim da Granja, São José dos Campos, SP 12227-010, Brazil.

Abstract

Due to the severity related to extreme flood events, recent efforts have focused on the development of reliable methods for design flood estimation. Historical streamflow series correspond to the most reliable information source for such estimation; however, they have temporal and spatial limitations that may be minimized by means of regional flood frequency analysis (RFFA). Several studies have emphasized that the identification of hydrologically homogeneous regions is the most important and challenging step in an RFFA. This study aims to identify state-of-the-art clustering techniques (e.g., K-means, partition around medoids, fuzzy C-means, K-harmonic means, and genetic K-means) with potential to form hydrologically homogeneous regions for flood regionalization in Southern Brazil. The applicability of some probability density function, such as generalized extreme value, generalized logistic, generalized normal, and Pearson type 3, was evaluated based on the regions formed. Among all the 15 possible combinations of the aforementioned clustering techniques and the Euclidian, Mahalanobis, and Manhattan distance measures, the five best were selected. Several watersheds' physiographic and climatological attributes were chosen to derive multiple regression equations for all the combinations. The accuracy of the equations was quantified with respect to adjusted coefficient of determination, root mean square error, and Nash–Sutcliffe coefficient, whereas, a cross-validation procedure was applied to check their reliability. It was concluded that reliable results were obtained when using robust clustering techniques based on fuzzy logic (e.g., K-harmonic means), which have not been commonly used in RFFA. Furthermore, the probability density functions were capable of representing the regional annual maximum streamflows. Drainage area, main river length, and mean altitude of the watershed were the most recurrent attributes for modelling of mean annual maximum streamflow. Finally, an integration of all the five best combinations stands out as a robust, reliable, and simple tool for estimation of design floods.

KEYWORDS

cluster analysis, evolutionary computation, fuzzy logic, heterogeneity measure, index-flood, L-moments

1 | INTRODUCTION

Among all types of natural disasters, those related to flooding are often considered the costliest and the most devastating to life (Lam, Thompson, & Croke, 2017; Rahman, Charron, Ouarda, & Chebana, 2018). Thus, advances in techniques for estimation of design streamflows (quantiles) are necessary, as these represent the main hydrological input to the sizing of hydraulic structures (e.g., bridges, channels, dams, highway culverts, levees, and spillways) and to the conduction of environmental studies such as effects of climate change, flood risk assessment, and land use planning (Agarwal, Maheswaran, Kurths, & Khosa, 2016).

According to Haddad and Rahman (2012), the most reliable design streamflow estimates are obtained directly from observed streamflow historical series, but these have intrinsic temporal and spatial limitations. The temporal limitations can be attributed to the need of estimates for return periods (RPs) longer than the length of the historical series under analysis (Aydoğan, Kankal, & Onsoy, 2016). When a long-term series is available for the site of interest, at-site flood frequency analysis (at-site FFA) can be generally applied. Nonetheless, at-site FFA does not overcome the spatial limitations. In other words, it does not provide estimates for ungauged sites, which is often necessary in water resources engineering (Basu & Srinivas, 2015). In this case, a broader approach that allows the joint use of at-site data, that is, regional flood frequency analysis (RFFA), stands out as a powerful tool (Sabourin & Renard, 2015).

Several RFFA techniques have been proposed throughout the years. Smith, Sampson, and Bates (2015) highlighted those based on direct regression, geostatistical procedure, and index-flood method, which have the common goal of transferring information from a group of gauged watersheds to ungauged sites (Haddad & Rahman, 2012). The index-flood method coupled with L-moments, as presented by Hosking and Wallis (1997), has been extensively used worldwide (Abida & Ellouze, 2008; Aydoğan et al., 2016; Haddad & Rahman, 2012; Hussain & Pasha, 2009; Noto & La Loggia, 2009; Saf, 2009; Seckin, Haktanir, & Yurtal, 2011; Smith et al., 2015).

Hosking and Wallis (1997) structured the application of the index-flood method in conjunction with L-moments in four steps: (a) screening of data, (b) identification of homogeneous regions, (c) choice of a probability density function (PDF), and (d) estimation of the PDF. Farsadnia et al. (2014) stated that the identification of homogeneous regions is considered the most difficult task among all the aforementioned steps because it often depends on subjective decisions. There are several methods for identifying homogeneous regions in RFFA (Goyal & Gupta, 2014): (a) canonical correlation analysis, (b) cluster analysis, (c) hierarchical approach, (d) method of residuals, and (e) region of influence. Rao and Srinivas (2008a) reported that regions have been frequently delineated considering administrative, physiographic, or political boundaries; however, these criteria do not necessarily ensure hydrological homogeneity. In this context, cluster analysis for identification of hydrologically homogeneous regions arises as a state-of-the-art technique that has potential to reduce the process' subjectivity and to generate a more appropriate grouping under a hydrological point of view.

According to Rao and Srinivas (2008a), algorithms used for cluster analysis in regionalization studies can be categorized into hard (e.g., hierarchical, partitional, or hybrid) and fuzzy clustering. Hierarchical clustering algorithms, such as the Ward's algorithm, have been widely employed in hydrological regionalization (Farsadnia et al., 2014; Latt, Wittenberg, & Urban, 2015; Noto & La Loggia, 2009). Likewise, partitional clustering algorithms (e.g., K-means and partitional around medoids [PAMs]) have been commonly applied for identifying hydrologically homogeneous regions (Beskow et al., 2016; Farsadnia et al., 2014). Algorithms that create a nested sequence of partitions (hierarchical clustering) and those that partition data in a predefined number of clusters (partitional clustering) can be combined to form a hybrid cluster, which may present better performance in forming regions, as observed by Rao and Srinivas (2008b).

There has been a recent increase in the use of artificial intelligence (AI) to solve engineering and science problems, including those related to the environment and hydrology (Yaseen, El-shafie, Jaafar, Afan, & Sayl, 2015). In hydrological regionalization studies, AI has been used in the form of fuzzy logic (Basu & Srinivas, 2015; Beskow et al., 2016; Chavoshi, Sulainman, Saghafian, Sulainman, & Manaf, 2013; Farsadnia et al., 2014; Rao & Srinivas, 2008c), artificial neural networks (ANNs; Farsadnia et al., 2014), and evolutionary computation (Beskow et al., 2016; Chavoshi et al., 2013; Shu & Burn, 2004). It should be pointed out that such AI techniques often provide superior results when compared with partitional clustering algorithms, such as K-means and PAM (Beskow et al., 2016; Goyal & Gupta, 2014). Fuzzy C-means (FCM) and K-harmonic means (KHM) are examples of fuzzy logic-based algorithms, which have been recently used in some hydrological regionalization studies (Beskow et al., 2016; Sadri & Burn, 2011). In addition, a few studies have focused on genetic algorithms for clustering applications in hydrology (Beskow et al., 2016) or as an optimization process of soft algorithms (Chavoshi et al., 2013; Shu & Burn, 2004).

Several recent studies concerning RFFA have compared results obtained from different categories of clustering algorithms. Kingston, Hannah, Lawler, and McGregor (2011) appraised a set of hierarchical, partitional, and hybrid clustering algorithms for 112 coastal watersheds in the northern North Atlantic region over the United States, Canada, Iceland, Scotland, Norway, Sweden, Denmark, and Finland. These authors identified seven regions in the best solution combination (Ward's algorithm). In a study applied to 117 watersheds across western United States, Agarwal et al. (2016) found that robust AI techniques coupled with wavelet analysis outperformed traditionally used clustering algorithms with respect to data's temporal and dimensional limitations. Chérif and Bargaoui (2013) used watershed physiographic and climatological attributes in a RFFA for 40 maximum annual streamflow (MAS) historical series in Tunisia. The authors obtained reliable regional curves for the two regions delineated according to hierarchical and partitional clustering algorithms. Farsadnia et al. (2014) compared ANNs, Fuzzy logic based, and hard cluster regionalization techniques for identifying hydrologically homogeneous regions considering 47 watersheds in northern Iran. These authors found that the best results were obtained when combining self-organizing feature map and hierarchical clustering algorithms. Kumar, Goel, Chatterjee, and Nayak (2015) compared soft computing techniques with the regional relationships derived from the L-moments approach for 17 gauged sites in the lower Godavari subzone—India. The authors

concluded that fuzzy inference system and ANN outperformed the L-moments approach when designing floods.

It is worth noting that to the best of the authors' knowledge, some of the algorithms used in the present study—Fuzzy logic based KHM and Evolutionary computation derived genetic K-means algorithm (GKA)—have not been used in RFFA yet. Moreover, after an extensive literature review, no previous RFFA coupled with cluster analysis was found for South America, reinforcing the need of a more robust regional approach in this continent whose hydrological monitoring is scarce.

In this context, the present study has as main objective to assess the superiority of several state-of-the-art clustering techniques based on AI (FCM, KHM, and GKA) over traditionally used techniques (K-means and PAM) for identifying hydrologically homogeneous regions for flood regionalization in Southern Brazil. In addition, this study has as specific objectives to (a) identify hydrological attributes that best describe the flood behaviour in the regions formed from AI techniques; (b) evaluate the applicability of the index-flood coupled with L-moments method; (c) appraise PDFs for regional analysis, such as generalized extreme value (GEV), generalized logistic (GLO), generalized normal (GNO), and Pearson type 3 (PE3); (d) provide a reliable tool for managers and engineers in a region of scarce streamflow monitoring.

2 | MATERIAL AND METHODS

2.1 | Study area

The proposed approach was assessed taking as reference 106 MAS historical series corresponding to watersheds situated in Rio Grande do Sul State, Southern Brazil (Figure 1a,b). They were obtained from the Hydrological Information System platform (www.snirh.gov.br/hidroweb/) of the National Water Agency of Brazil. According to the nonparametric test of Mann–Kendall (Kendall, 1975; Mann, 1945) for a significance level of 5%, all the 106 series presented stationarity with respect to monotonic trend (Cassalho et al., 2018).

Considering all the Brazilian states, Rio Grande do Sul has the fifth largest population and ninth largest area, with 11,286,500 inhabitants and 282,000 km^2, respectively (Instituto Brasileiro de Geografia e Estatística, 2016). Based on Köppen type-climate classification, the state is classified as humid subtropical and has an oceanic without dry season and with hot summer (Cfa) climate in most of its area. Its mountainous region north-eastwards is classified as oceanic climate without dry season and with temperate summer (Cfb; Alvares, Stape, Sentelhas, Gonçalves, & Sparovek, 2014). According to the normal climatology for Rio Grande do Sul, monthly average temperature and total annual rainfall vary from 10.5 to 27.1°C and 1,229 and 1,823 mm, respectively (Instituto Nacional de Meteorologia, 2017).

2.2 | Regional flood frequency analysis

2.2.1 | Screening of data by means of discordancy measure

An important step in RFFA is the identification of sites containing gross errors and outliers. Due to the fact that incorrect values, outliers, shifts, and trends are reflected in the sample L-moments, Hosking and Wallis (1997) proposed the use of a single statistic defined as discordancy measure (D_i) to identify erroneous data. According to these authors, D_i can be used in two cases: (a) at the beginning of the RFFA for identifying grossly erroneous data in a large data set, such as errors in recording and/or transcription of data, and (b) to aid the decision making towards discordant sites within a predefined region (e.g., moving a site to another region). The D_i of site i is defined as follows:

$$D_i = \frac{1}{3}N(u_i - \bar{u})^T A^{-1}(u_i - \bar{u}) \quad (1)$$

$$\bar{u} = N^{-1}\sum_{i=1}^{N} u_i \quad (2)$$

$$A = \sum_{i=1}^{N}(u_i - \bar{u})(u_i - \bar{u})^T \quad (3)$$

where N is the number of sites, u_i is a vector composed by sample coefficients of L-variation, L-skewness, and L-kurtosis; T refers to

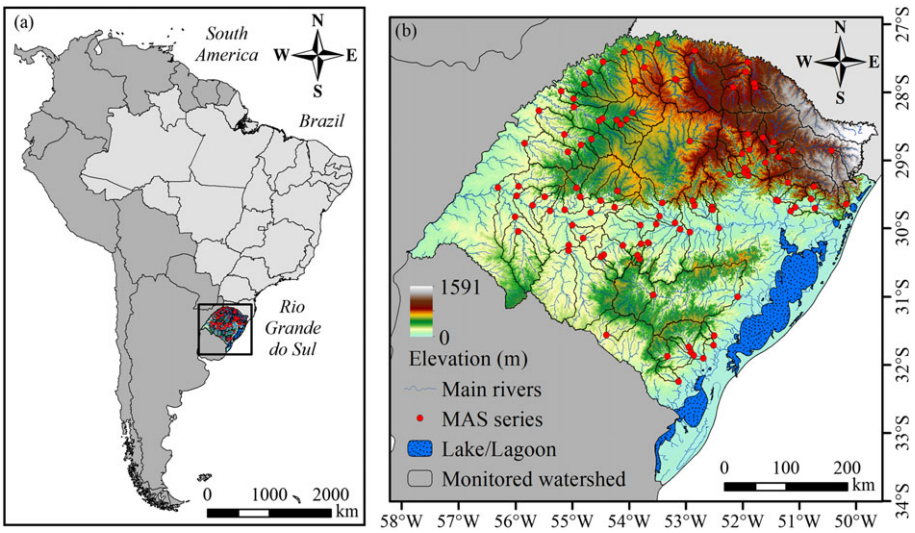

FIGURE 1 Location of the 106 gauged watersheds in Rio Grande do Sul State, Brazil

the transposition of the vector derived from the difference between vector u_i and its respective regional average $\overline{u}$, and A is the matrix of sums of squares and cross products, by means of u_i and $\overline{u}$. Threshold values for discordant sites depend on N and can be found in Hosking and Wallis (1997).

2.2.2 | Attributes

The methodological sequence used in this study followed the work routine recommended by Rao and Srinivas (2008a) for applying clustering algorithms in hydrological regionalization: (a) identification of explanatory variables related to flood, (b) rescaling of the data, (c) selection of the clustering algorithm to be used, (d) determination of the number of clusters, (e) evaluation of the formed regions according to a homogeneity criterion, (f) adjustment of the regions classified as heterogeneous, and (g) estimation of quantiles.

Among the possible clustering attributes, those related to watershed physiographic, location, and climatological characteristics have been widely used (Basu & Srinivas, 2015; Chavoshi et al., 2013; Chérif & Bargaoui, 2013; Farsadnia et al., 2014; Haddad & Rahman, 2012; Lam et al., 2017; Latt et al., 2015; Razavi & Coulibaly, 2013; Saf, 2009; Srinivas, Tripathi, Rao, & Govindaraju, 2008). These attributes were tested against flood related variables, for example, mean and median annual flood (MEF), mean and MEF per area, and mean and MEF per mean annual rainfall, and then those with the greatest correlations were selected, as suggested by Rao and Srinivas (2008b). It is worth mentioning that these attributes must be rescaled in order to nullify their magnitudes and the influence of the variance over the clusters' formation (Basu & Srinivas, 2015). Considering that x'_{sj} is the transformed vector (e.g., logarithmic or square root) containing j values attributes from site s; and $\overline{x}'_j$ and $\widehat{\sigma}_j$, its mean and standard deviation, respectively. Thus, the rescaled vector x_{sj} is given by (Basu & Srinivas, 2015):

$$x_{sj} = \frac{\left(x'_{sj} - \overline{x}'_j\right)}{\widehat{\sigma}_j} \tag{4}$$

Once all attributes have been properly identified, transformed, and rescaled, clustering algorithms can be used to form hydrologically similar regions.

2.2.3 | Artificial intelligence techniques

Due to the simplicity and efficiency of K-means, it is one of the most used clustering algorithms (Jain, 2010). Through an iterative reallocation process, this centroid-based algorithm minimizes its objective function at each iteration by moving objects between groups (Beskow et al., 2016). Chang, Tsai, Tsai, and Herricks (2008) summarized the application of K-means in six steps: (a) determination of a k number of clusters ($NC = \{nc_1, nc_2, \ldots, nc_k\}$); (b) location of k points at the initial group's centroids; (c) assignment of objects to the closest centroid's cluster; (d) recalculation of the centroids, once all objects have been assigned; (e) replication of steps c and d until the k centroids stop moving; (f) calculation of the precision statistical index to be minimized for all the formed groups. Taking the sum of squared errors (SSE) as objective function for a data set $X = \{x_1, x_2, \ldots, x_n\}$, the objective function can be written as follows (Beskow et al., 2016):

$$SSE = \sum_{p=1}^{k} \sum_{x_i \in nc_p} \left(x_i - \overline{x}_{nc_p}\right)^2 \tag{5}$$

where the n-dimension feature vector x_i ($i = 1, \ldots, n$) is an object of cluster $nc_p \in NC$ of centroid $\overline{x}_{nc_p}$.

In order to minimize the noise and/or the influence of outliers in the clustering process, a clustering procedure based on the data median (*K*-medoids) instead of its mean (K-means) was proposed (Beskow et al., 2016; Han, Kamber, & Pei, 2006). PAM stands out among the K-medoids based algorithms. Considering the same attributes used in Equation 5, the average distance intra-cluster AD_{IC} can be written as follows (Beskow et al., 2016; Sadri & Burn, 2011):

$$AD_{\text{IC}} = \left(\frac{1}{k}\right) \sum_{x_i \in nc} d(x_i, \overline{x}_{nc}) \tag{6}$$

where d is the distance between objects according to the used distance measure.

Due to recent computational developments, attempts of applying fuzzy logic based algorithms in hydrological regionalization have been performed (Chavoshi et al., 2013). As a soft clustering method, one of the main characteristics of fuzzy logic-based algorithms is that according to a certain degree of membership (i.e., strength in which a feature vector belongs to a cluster), a feature vector can be simultaneously considered as part of all clusters formed (Goyal & Gupta, 2014; Rao & Srinivas, 2008c).

FCM algorithm is a variation of K-means, where the watersheds are partitioned into a number of fuzzy clusters following an iterative optimization of a fuzzy objective function, which is written as follows (Rao & Srinivas, 2008c):

$$FCM_{\text{cost}} = \sum_{p=1}^{k} \sum_{i=1}^{N} u_{\text{ip}}^{r_{\text{fuz}}} d\left(x_i - \overline{x}_{nc_p}\right)^2 \tag{7}$$

where N is the number of objects and k the number of clusters; u_{ip} refers to the degree of membership of object ith in the cluster pth; r_{fuz} corresponds to the fuzzification parameter that represents the degree of overlapping between clusters (Farsadnia et al., 2014), and d is the distance between the ith object and the pth cluster's centroid $\overline{x}_{nc_p}$. These authors stated that in its minimum value ($r_{\text{fuz}} = 1$), the result is equivalent of that of a hard cluster (i.e., no overlap), then it should always be greater than 1.

Although FCM has been reported as a flexible clustering algorithm, it is strongly affected by outliers, noise, and the subjectivity on defining initial clusters' centres (Beskow et al., 2016). To overcome these problems, KHM algorithm contains a weighted function that makes distant objects more important, reducing the importance of the initial clusters' centres (Güngör & Ünler, 2008). Considering the same variables described for Equation 7, the objective function for KHM, which uses the harmonic average between the formed clusters' centres and all remaining objects, is written as follows (Beskow et al., 2016):

$$KHM_{\text{cost}} = \sum_{i=1}^{N} \frac{k}{\sum_{p=1}^{k} \left[d\left(x_i - \bar{x}_{nc_p}\right)^{r_{fuz}} \right]^{-1}} \tag{8}$$

It should be mentioned that the fuzzification parameter must be always greater than 2 for Equation 8.

The use of genetic algorithms in hydrological regionalization is still incipient. The present study applies a hybrid algorithm known as GKA (Krishma & Murty, 1999), which combines concepts of genetic algorithm and K-means for identifying possible hydrologic homogeneous regions, replacing the crossover operator of traditional genetic algorithm on each chromosome by K-means algorithm (Krishma & Murty, 1999). At each iteration, GKA aims at keeping good solutions throughout generations, according to the ranked values provided by the fitness function:

$$F(S_i) = \begin{cases} f(S_i) - (\bar{f} - ct.\sigma), & for \quad F(S_i) \geq 0 \\ 0, & for \quad < 0 \end{cases} \tag{9}$$

where $F(S_i)$ is the fitness function of the S_ith population chromosome; $\bar{f}$ and σ are the mean and standard deviation of $f(S_i)$, which is an intermediary function for ith chromosome given by the expression $f(S_i) = -SSE(S_i)$; and ct is a constant that varies from one to three.

2.2.4 | Distance measures

There are several methods for calculating the distance measure d used in the evaluation of dissimilarity among clusters or between clusters and feature vectors. Following recommendations of Beskow et al. (2016), the present study considered three distance measures: Euclidean, Mahalanobis, and Manhattan, respectively:

$$d(x,y)_{\text{Euclidean}} = \sqrt{\sum_{j=1}^{n} \left(x_j - y_j\right)^2} \tag{10}$$

$$d(x,y)_{\text{Mahalanobis}} = (x-y)^T S^{-1} (x-y) \tag{11}$$

$$d(x,y)_{\text{Manhattan}} = \sum_{j=1}^{n} \left| x_j - y_j \right| \tag{12}$$

where, $x_j - y_j$ is the difference between two objects (x_j and y_j) that contain n attributes; T is transpose of matrix derived from the difference ($x - y$); and S^{-1} is the inverse covariance matrix of all considered attributes.

2.2.5 | Hydrological homogeneity

Hosking and Wallis (1997) presented three different approaches for calculating the heterogeneity measure (H, H_2, H_3). However, the authors stated that both H_2 and H_3 tend to be more sensitive for smaller regions, and often give false homogeneity classification. Thus, only the heterogeneity measure H was adopted in this study.

The H measure has been extensively used in the identification of hydrological homogeneous regions (Aydoğan et al., 2016; Beskow et al., 2016; Seckin et al., 2011; Smith et al., 2015). Given a group of sites, which can be generated using a clustering algorithm, H compares the sample L-moments variation between sites that would be expected from a simulated homogeneous region (Hosking & Wallis, 1997). The heterogeneity measure H is written as follows:

$$H = \frac{(V - \mu_V)}{\sigma_V} \tag{13}$$

$$V = \left\{ \frac{\sum_{i=1}^{N} n_i \left[t^{(i)} - t^R \right]^2}{\sum_{i=1}^{N} n_i} \right\}^{\frac{1}{2}} \tag{14}$$

$$t^R = \frac{\sum_{i=1}^{N} n_i t^{(i)}}{\sum_{i=1}^{N} n_i} \tag{15}$$

where V is the weighted standard deviation of the coefficient of L-variation; μ_V and σ_V are the mean and standard deviation of N_{sim} simulations for a region containing N sites i with a length n_i and sample coefficient of L-variation, L-skewness and L-kurtosis $t^{(i)}, t_3^{(i)}, t_4^{(i)}$, respectively; and t^R is the regional average coefficient of L-variation. A region can be considered hydrologically homogeneous if $H < 1$, possibly homogeneous when $1 \leq H < 2$, and definitely heterogeneous when $H \geq 2$ (Hosking & Wallis, 1997).

2.2.6 | Choice of a probability density function

Several PDFs have been proposed for frequency modelling of extremes in hydrology. Among them, a few have been extensively used in RFFA worldwide (Aydoğan et al., 2016; Noto & La Loggia, 2009; Saf, 2009; Seckin et al., 2011). However, whenever choosing the PDFs to be considered one should make sure they are adequate to the proposed analysis. The three-parameter distributions GEV (Equation 16), GLO (Equation 17), GNO (Equation 18), and PE3 (Equation 19) were appraised in this study.

$$f(x) = \alpha^{-1} e^{-(1-k)y - e^{-y}}, \quad y = \begin{cases} -k^{-1} \log\left[1 - k\frac{(x-\xi)}{\alpha}\right], & k \neq 0 \\ \frac{(x-\xi)}{\alpha}, & k = 0 \end{cases} \tag{16}$$

$$f(x) = \frac{\alpha^{-1} e^{-(1-k)y}}{\left(1 + e^{-y}\right)^2}, \quad y = \begin{cases} -k^{-1} \log\left[1 - k\frac{(x-\xi)}{\alpha}\right], & k \neq 0 \\ \frac{(x-\xi)}{\alpha}, & k = 0 \end{cases} \tag{17}$$

$$f(x) = \frac{e^{ky - y^2/2}}{\alpha\sqrt{2\pi}}, \quad y = \begin{cases} -k^{-1} \log\left[1 - k\frac{(x-\xi)}{\alpha}\right], & k \neq 0 \\ \frac{(x-\xi)}{\alpha}, & k = 0 \end{cases} \tag{18}$$

$$\frac{f(x) = (x-\xi)^{\alpha-1} e^{-\frac{(x-\xi)}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \tag{19}$$

where, ξ, α, and k, are the position, scale, and shape parameters, respectively, which can be estimated from the L-moments λ_1, λ_2, and L-moment ratios τ_3, τ_4. A detailed description of the aforementioned PDFs' parameters can be found in the study of Asquith (2011).

The PDF that provides the best regional fitting was then identified by means of the goodness-of-fit measure Z^{DIST} proposed by Hosking

and Wallis (1997). The bias (B_4) and standard deviation (σ_4) of the regional L-kurtosis (t_4^R) are estimated by

$$B_4 = N_{sim}^{-1} \sum_{m=1}^{Nsim} \left(t_{4^{[m]}} - t_4^R\right) \tag{20}$$

$$\sigma_4 = \left\{ (N_{sim}-1)^{-1} \left[\sum_{m=1}^{Nsim} \left(t_{4^{[m]}} - t_4^R\right)^2 - N_{sim} B_4^2 \right] \right\}^{\frac{1}{2}} \tag{21}$$

where $t_4^{[m]}$ is the sample L-kurtosis for the mth simulated kappa region having N sites. Thus, Z^{DIST} of a distribution is calculated as follows:

$$Z^{\text{DIST}} = \frac{\left(\tau_4^{DIST} - t_4^R + B_4\right)}{\sigma_4} \tag{22}$$

where τ_4^{DIST} is the distribution's L-kurtosis. According to Hosking and Wallis (1997), for an adequate fit $|Z^{\text{DIST}}| \leq 1.64$.

The aforementioned cluster analysis was performed with the aid of the software clustering tool developed by Corrêa (2014) in Delphi programming language. The statistical analyses were carried out with the support of the System of Hydrological Data Acquisition and Analysis, which has been applied for at-site and regional hydrological frequency analysis (Beskow et al., 2016; Caldeira et al., 2015; Cassalho et al., 2018).

2.2.7 | Index flood

Once a region has been checked with respect to hydrological homogeneity and a PDF has been properly fitted, dimensionless estimations can be made applying the growth curve $\widehat{q}(F)$. As F is the nonexceedance frequency ($0 < F < 1$) and $\widehat{Q}_i(F)$ is the quantile function for site i, flood estimations associated with RP (design flood) are made by

$$\widehat{Q}_i(F) = \widehat{q}(F)\widehat{\mu}_i \tag{23}$$

where $\widehat{\mu}_i$ is the index flood, which can be obtained by means of empirical regression equations of watershed characteristics that exert influence on mean annual maximum streamflow (Costa, 2017; Noto & La Loggia, 2009) and can be written as follows (Noto & La Loggia, 2009):

$$\widehat{\mu}_i = b_0 \prod_{i=1}^{N} C_i^{b_i} \tag{24}$$

where b_0 and b_i are constants; and N refers to the number of explanatory variables C_i. Following the recommendations of Noto and La Loggia (2009), a stepwise regression was performed between watersheds' mean annual maximum streamflow and their physical and climatological characteristics. These parameters were also checked for multicollinearity considering the coefficient of determination (R^2), as proposed by Costa (2017).

The regional regression equations were assessed according to Nash-Sutcliffe coefficient (NSE), Root Mean Square Error (RMSE), and R_i^2, by using a routine developed in R language (R Core Team, 2016). Furthermore, a leave-one-out cross-validation was performed as this cross-validation technique is the most commonly used method to assess the validity of regionalization studies (Razavi & Coulibaly, 2013). The error statistics for assessment of cross-validation were the same as those used for appraisal of index-flood equations.

3 | RESULTS AND DISCUSSION

3.1 | Identification of explanatory attributes

Several attributes were extracted for all the 106 watersheds, which were then used in the clustering procedure and the index-flood method. Watershed's physiographic and location attributes, for example, area (A), compactness coefficient (Kc), drainage density (Dd), main river length (L), main river slope (Ls), mean altitude (E), and mean slope (S), were estimated with the aid of a geographical information system using the cartographic database provided by Hasenack and Weber (2010) for Rio Grande do Sul State. Climatological attributes (Table 1), such as maximum annual daily rainfall (Pd) and mean annual rainfall (Pa), were derived for all the 106 watersheds from datasets of 342 rain gauges used by Caldeira et al. (2015; Supplementary Material 1).

Rao and Srinivas (2008b) suggested that the attributes used in the clustering procedure should be selected according to their correlation with flood-related variables, for example, mean annual flood (MAF), MAF per watershed's area (MAF/A), MAF per mean annual rainfall (MAF/P), MEF, mean annual flood per watershed's area (MEF/A),

TABLE 1 Watershed's attributes and their respective range

		Range		
Attributes	**Units**	**Minimum**	**Mean**	**Maximum**
Area (A)	km^2	62.06	5,049.53	42,595.55
Compactness coefficient (Kc)	dimensionless	1.34	1.76	2.30
Drainage density (Dd)	$km.km^{-2}$	0.94	1.42	1.86
Main river length (L)	km	14.58	150.69	492.41
Main river slope (Ls)	$m.m^{-1}$	0.00074	0.00484	0.04854
Mean altitude (E)	m	116.76	397.79	902.61
Mean slope (S)	%	2.54	10.65	33.89
Maximum annual daily rainfall (Pd)	mm	75.49	96.27	113.23
Mean annual rainfall (Pa)	mm	1,334.46	1,651.35	1,907.64

and MEF per mean annual rainfall (MEF/P). The correlation matrix among attributes presented in Table 1 and their correlation with flood-related variables are described in Table 2.

According to the correlation matrix (Table 2), Dd, E, S, and Pd had the weakest correlation with flood-related variables. Therefore, they were not used in the clustering analysis. A and L were found to have strong correlation with flood-related variables. However, such variables also presented high collinearity because they had an R^2 of 0.86. To guarantee the use of independent variables, only L was selected, as this attribute had the highest correlation with flood-related variables. Kc was also selected because it presented some degree of correlation with flood-related variables and no strong collinearity with any other selected attributes. It should be mentioned that any pair of variables with $R^2 > 0.80$ was considered having strong collinearity, as suggested by Costa (2017).

Rao and Srinivas (2008b) recommended the use of meteorological variables, even when no strong correlation with flood-related variables is found. The most commonly used meteorological variable in cluster analysis for flood regionalization is Pa (Chavoshi et al., 2013; Chérif & Bargaoui, 2013; Haddad & Rahman, 2012; Smith et al., 2015), which was also considered in the present study. Finally, following recommendations of Beskow et al. (2016), location attributes such as latitude and longitude of watersheds' centroid were selected. Alternatively, the same approach was performed without the inclusion of location attributes, resulting in the loss of more than half of all sites for being discordant or for belonging to a hydrologically heterogeneous region in all the 15 possible combinations. Therefore, only the results obtained when using latitude and longitude were considered in the following analysis.

Some weaknesses of the attribute selection approach proposed by Rao and Srinivas (2008b) and used in the present study should be mentioned. First, the correlation matrix only counts for linear correlation between physiographic/climatological attributes and flood-related variables. Therefore, although linear correlations have been observed in this study, nonlinear correlations might occur without being identified by the proposed approach. In addition, flood-related variables per se may not be representative for the floods in the region. As a matter of fact, different approaches are found in the literature for selection of clustering attributes, that is, no selection (using all available attributes), selection based on empirical knowledge, and nonlinear correlation (Basu & Srinivas, 2015; Farsadnia et al., 2014; Rao & Srinivas, 2008b; Shu & Burn, 2004). Another alternative would be the use of principal components, as performed by Jones, Blenkinsop, Fowler, and Kilsby (2014) and Forestieri et al. (2018). Kingston et al. (2011) reported that any chosen method involves several subjective decisions. Thus, one should select the procedure that best addresses the problem.

3.2 | Cluster analysis for identifying hydrologically homogeneous regions

The most subjective decision when using clustering algorithms for hydrological regionalization is to define the ideal number of clusters. According to Sadri and Burn (2011), there is not a widely accepted standardized method for defining the optimum number of clusters. Hosking and Wallis (1997) affirm that one must look for a balance in the number of sites among clusters because small clusters provide little accuracy gain and large clusters might fail to be homogeneous. Rao and Srinivas (2008a) pointed out two alternatives for aiding the cluster formation: (a) visual interpretation and (b) the use of cluster validity indices. Chang et al. (2008) applied common statistics, such as the R^2 and root mean square standard deviation, for supporting the choice of the ideal number of clusters. Goyal and Gupta (2014) used cluster validity indices, for instance, Dunn Index and Average Silhouette Width. Latt et al. (2015) applied the

TABLE 2 Correlation matrix among watershed's attributes and flood-related variables

	Attributes								
	A	Kc	Dd	L	Ls	E	S	Pd	Pa
A	1								
Kc	0.57	1							
Dd	0.04	0.02	1						
L	0.86	0.68	0.07	1					
Ls	−0.29	−0.28	−0.30	−0.43	1				
E	−0.14	0.15	−0.08	0.06	0.15	1			
S	−0.18	0.17	−0.35	−0.11	0.60	0.50	1		
Pd	0.08	−0.20	0.10	0.02	−0.40	−0.62	−0.59	1	
Pa	−0.19	−0.15	−0.21	−0.06	−0.13	0.21	0.00	0.43	1
MAF	0.79	0.60	0.07	0.88	−0.32	0.13	−0.02	−0.10	−0.17
MAF/A	−0.38	−0.35	0.06	−0.41	0.27	0.10	0.15	−0.01	0.14
MAF/P	0.79	0.60	0.08	0.86	−0.31	0.12	−0.02	−0.12	−0.23
MEF	0.80	0.60	0.07	0.88	−0.32	0.12	−0.02	−0.10	−0.18
MEF/A	−0.38	−0.35	0.08	−0.42	0.29	0.09	0.18	−0.02	0.10
MEF/P	0.80	0.60	0.08	0.86	−0.31	0.11	−0.02	−0.12	−0.23

Note. A: area; Dd: drainage density; E: mean altitude; Kc: compactness coefficient; L: main river length; Ls: main river slope; MAF: mean annual flood; MEF: median annual flood; Pa: mean annual rainfall; Pd: maximum annual daily rainfall; S: mean slope.

hierarchical Ward's algorithm to obtain a first approach of the ideal number of clusters.

It should be mentioned that one of the main restrictions of RFFA applicability for estimating design flood is associated with the temporal limitations imposed by the historical series length and the total number of sites considered. For practical purposes, Rao and Srinivas (2008b) and Jones et al. (2014) adopted that the maximum RP applicable for estimating the design flood should be five times lower than the sum of all series lengths within the region of interest. Various hydraulic structures (e.g., bridges, highway culverts, levees, small to large dams, and urban storm sewers) are designed for RPs between 50 and 200 years (Rao & Srinivas, 2008a). Because the total length of historical series used in this study for Rio Grande do Sul State is about 3,844 years, six equally sized regions (i.e., about 600 years of records) would be the ideal number, thus allowing the estimation of design floods associated with RPs over 100 years. In addition, according to Hosking and Wallis (1997), little accuracy gain is observed in regions consisting of more than 20 sites, thus corroborating the choice of six clusters as 106 MAS series were considered. The authors also affirm that some of their statistics, for example, D_i, are not recommended for very small regions (i.e., less than seven sites); thereby, only regions with at least seven sites were considered. In a low streamflow assessment using the same gauging stations as those of the present study, Beskow et al. (2016) also indicated that six regions would provide good results for the State of Rio Grande do Sul.

The results provided by the cluster analysis are not final and may be enhanced through subjective adjustments, such as those suggested by Hosking and Wallis (1997): (a) moving sites between regions, (b) deleting sites, (c) subdividing regions, (d) merging a fraction or the whole region with another region, (e) merging regions and then redefining groups, and (f) adding more data. After deleting all the discordant and the most discordant sites in accordance with the Discordancy measure D_i (Equation 1), the final results were obtained (Table 3) for all possible combinations of clustering algorithms and distance measures. The removal of discordant sites as a criterion for reaching hydrological homogeneity has been widely used in regionalization studies despite often resulting in a considerable loss of information (Abida & Ellouze, 2008; Forestieri et al., 2018). It is valid to mention that by definition, different clustering algorithms provide different groupings, which do not necessarily share the same geographical locations. Thus, one should not make comparisons between regions derived from different clustering algorithms. Instead, readers are encouraged to observe the performance of such algorithms in properly extracting the largest amount of information from the data set. This characteristic is reflected in terms of the total number of sites considered, the total length of historical series, the heterogeneity measure, and regional statistics.

TABLE 3 Main regional flood frequency analysis characteristics for all the five clustering algorithms and their respective regions (R1–R6), combined with (a) Euclidean, (b) Mahalanobis, and (c) Manhattan distance measures

		K-means			PAM			FCM			KHM			GKA		
		a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
R1	NS	35	17	23	11	14	16	24	25	13	10	29	20	24	12	20
	ND	5	0	1	0	0	2	2	6	1	0	4	1	5	0	2
	RS	19	10	8	4	7	7	5	12	5	3	8	6	12	4	11
	T	370	226	486	280	337	224	517	290	311	304	759	380	297	215	291
	H	1.09	0.57	−0.81	−0.56	4.88	2.20	1.74	2.96	3.87	2.82	1.49	1.32	3.88	1.41	4.21
R2	NS	20	16	30	11	13	19	14	19	14	26	20	31	21	21	25
	ND	3	0	3	0	0	0	0	0	0	4	0	5	2	3	6
	RS	10	7	15	1	6	10	7	12	2	12	13	14	12	11	12
	T	235	324	530	361	286	383	332	284	481	427	302	522	291	277	286
	H	4.05	1.82	1.52	1.29	4.67	1.26	5.73	3.95	1.49	0.11	3.88	1.67	4.26	3.25	0.26
R3	NS	7	23	13	22	18	13	9	20	24	20	11	10	24	13	24
	ND	0	0	0	1	0	1	0	0	4	1	1	0	2	1	2
	RS	0	9	0	14	9	4	2	13	7	6	0	3	5	3	5
	T	235	555	488	257	385	296	304	323	543	380	377	304	517	307	517
	H	4.03	1.28	1.82	−0.84	1.57	0.19	3.05	2.81	1.92	1.46	1.72	2.84	1.81	−0.17	1.78
R4	NS	12	40	21	33	14	17	24	13	9	29	19	24	9	31	9
	ND	3	2	1	2	0	1	5	1	0	3	3	1	0	2	0
	RS	2	24	6	26	7	9	12	0	2	15	4	15	2	22	2
	T	263	647	528	319	332	287	297	405	304	504	469	303	304	305	304
	H	−2.49	1.47	1.86	1.54	4.19	2.26	3.82	1.34	2.78	1.90	1.24	4.51	2.78	2.49	2.79
R5	NS	19	10	19	19	14	20	13	19	24	13	14	14	15	14	15
	ND	1	0	3	3	0	1	0	1	2	0	0	0	0	1	0
	RS	6	3	6	6	3	12	0	11	5	0	7	2	8	6	8
	T	303	203	389	443	389	309	488	321	517	488	349	481	338	263	338
	H	−0.64	1.81	0.74	1.71	1.19	2.29	1.83	0.80	1.76	1.79	2.97	1.45	2.52	−2.45	2.43
R6	NS	13	—	—	10	33	21	22	10	22	8	13	7	13	15	13
	ND	0	—	—	0	9	3	1	0	1	0	0	0	0	1	0
	RS	6	—	—	3	16	11	14	3	6	1	6	0	0	7	0
	T	311	—	—	269	330	361	291	203	643	238	203	190	488	259	488
	H	3.02	—	—	3.87	1.71	5.66	4.09	1.90	1.15	3.25	2.01	3.29	1.75	5.26	1.82

Note. FCM: Fuzzy c-means; GKA: genetic K-means algorithm; KHM: K-harmonic means; PAM: partitional around medoid. NS is the initial number of sites, ND is the total number of further excluded sites for being discordant; RS are the number of sites further removed for being the most discordant, thus not meeting the homogeneity criterion; T is the total lengths of all the remaining historical series, and H is the heterogeneity measure.

The interpretation of Table 3 is nontrivial. One may decide which combinations should be maintained taking into account specific criteria (e.g., best local spatial coverage and greater homogeneity to a specific location) instead of an overall statewide representation. Out the five clustering algorithms, K-means combined with Mahalanobis distance resulted in the most discrepant regions with respect to both the total number of sites and the lowest total record length within a region (Table 3). In addition, the combination of K-means with Euclidean and Mahalanobis distances generated clusters with less than 250 years of records, which means they do not allow for design flood estimates with RP over 50 years. Discrepant clusters in terms of total number of sites were also observed by Chang et al. (2008) when comparing K-means with other clustering techniques. Small regions with less than 250 years of records were also identified for the combinations of PAM with Manhattan, FCM with Mahalanobis, KHM with Euclidean, KHM with Mahalanobis, KHM with Manhattan, and GKA with Mahalanobis, thus affecting the index-flood method applicability. Also, K-means with the Euclidean distance presented the greatest number of discordant sites and K-means in conjunction with Mahalanobis and Manhattan distance measures were able to form only five regions.

The fuzzification factors equal to 1.3 and 3.5 were used for the FCM and KHM, respectively. Moreover, 500 generations were established to form clusters in the GKA. This may explain the similarity between GKA regions because after so many generations, little influence from starting condition is expected. One of the main characteristics of genetic algorithms is that only good solutions tend to be kept throughout generations, thereby leading to acceptable results (Beskow et al., 2016).

Of all the 15 possible combinations, those combining Euclidian distance with K-means, FCM, and GKA; Mahalanobis distance with PAM, FCM, KHM, and GKA; and Manhattan with PAM, KHM, and GKA, contained more than two heterogeneous regions ($H \geq 2$). Therefore, such combinations were excluded from further considerations as they would significantly limit the study's spatial applicability.

It should be highlighted that GKA resulted in groups with various desirable characteristics, such as large regions in terms of total record length, as it was possible to estimate design floods for RPs over 50 years for almost all its combinations. However, on the contrary of what was observed by Beskow et al. (2016), this algorithm was responsible for the exclusion of the largest amount of discordant series, representing a significant loss of information for a region where the number of gauged sites is limited.

3.3 | Choice of the best combinations

Five combinations remained for the sequence of this study: K-means with Mahalanobis and Manahattan distances, PAM with Euclidean distance, FCM with Manhattan distance, and KHM with Euclidean distance. These were then assessed considering the fittings of different PDFs (GEV, GLO, GNO, and PE3), having their respective goodness-of-fit evaluated with respect to the Z^{DIST} measure (Table 4).

Despite the fact that the five remaining combinations have been classified as hydrologically homogenous, they caused the loss of at least one region. This occurred because no PDF was satisfactory adjusted since $|Z^{DIST}|$ should be less than 1.64. PE3 was the PDF that had the best fit to most of the regions (nine regions), followed by GEV, GNO, and GLO, for regions 4, 3, and 1, respectively (Table 5). Saf (2009) and Aydoğan et al. (2016), similarly to the results found in the present study, also found that PE3 was the PDF with the best fitting for most of their regions.

TABLE 4 Goodness-of-fit Z^{DIST} measure for all the hydrological homogeneous regions (R1–R6)

		K-means		PAM	FCM	KHM
	PDF	b	c	a	c	a
R1	GEV	0.59	1.39	1.88	—	—
	GLO	2.22	5.31	4.26	—	—
	GNO	0.43	2.29	1.94	—	—
	PE3	−0.05	2.27	1.58	—	—
R2	GEV	1.64	1.07	1.95	1.31	0.98
	GLO	4.84	3.87	5.24	4.34	3.82
	GNO	2.17	0.96	2.42	1.34	1.06
	PE3	2.07	0.33	2.27	0.84	0.65
R3	GEV	1.59	1.64	4.20	1.55	1.50
	GLO	5.29	4.90	7.52	4.39	4.97
	GNO	2.11	1.93	5.00	1.36	2.27
	PE3	1.94	1.63	5.00	0.62	2.25
R4	GEV	0.46	0.45	0.21	—	0.47
	GLO	3.47	2.52	2.17	—	2.97
	GNO	0.37	0.03	−0.01	—	0.24
	PE3	−0.28	−0.90	−0.64	—	−0.48
R5	GEV	1.78	1.97	1.01	1.73	1.73
	GLO	4.17	4.84	3.88	5.62	5.04
	GNO	2.19	2.14	0.98	2.50	2.01
	PE3	2.13	1.81	0.44	2.45	1.72
R6	GEV	—	—	—	−1.89	—
	GLO	—	—	—	0.13	—
	GNO	—	—	—	−2.17	—
	PE3	—	—	—	−2.90	—

Note. FCM: Fuzzy C-mean; GEV: generalized extreme value; GLO; generalized logistics; GNO: generalized normal; KHM: K-harmonic means; PAM: partitional around medoid; PDF: probability density function; PE3: Pearson type 3. The a, b, and c parameters represent Euclidean, Mahalanobis, and Manhattan distance measures, respectively.

In order to identify the best out of the five selected combinations, several statistical measures were calculated for the index flood ($\hat{\mu}_i$) in each resulting region as well as multiple regression equations were proposed through stepwise regression procedure (Table 6). Only those attributes that culminated in an increase of at least 0.05 in the R^2_{adj} were considered in the present study, where R^2_{adj} refers to the adjusted coefficient of determination with respect to the number of independent variables in the multiple regression equation.

Of all the remaining combinations, K-means with Mahalanobis and Manahattan distances were the ones with the overall highest R^2_{adj} and NSE. These combinations also had either the lowest number of discordant sites or the most equally sized clusters in terms of total length of records. PAM with Euclidean distance was also able to model index-flood; however, its region number 5 (R5) presented low R^2_{adj} and NSE, indicating a larger uncertainty in estimating design floods. Also, PAM with Euclidean distance measure as well as FCM with Manhattan and KHM with Euclidean, had limited applicability since half of the initial number of regions had to be excluded (Table 3 and Table 4). Although KHM with Euclidean had neither the highest R^2_{adj} nor NSE,

TABLE 5 Growth curve's parameters along with their respective growth curve factors for commonly used return periods

Combination	Region	Best fit PDF	Growth curve parameters			Growth curve factors (RP)					
			ξ^R or μ	α^R or σ	k^R or γ	2	5	10	20	50	100
K-means b	R1	PE3	1.000	0.509	0.979	0.918	1.387	1.682	1.953	2.288	2.531
	R2	GEV	0.843	0.372	0.183	0.975	1.331	1.529	1.695	1.879	1.999
	R3	GEV	0.825	0.401	0.162	0.968	1.358	1.580	1.769	1.983	2.123
	R4	PE3	1.000	0.392	0.855	0.945	1.303	1.524	1.725	1.971	2.147
K-means c	R1	GEV	0.850	0.398	0.246	0.990	1.349	1.538	1.688	1.848	1.946
	R2	PE3	1.000	0.392	0.872	0.944	1.303	1.524	1.726	1.974	2.151
	R3	PE3	1.000	0.460	0.582	0.956	1.368	1.610	1.824	2.081	2.261
	R4	GNO	0.903	0.471	−0.396	0.903	1.373	1.689	1.994	2.396	2.701
PAM a	R1	PE3	1.000	0.460	0.735	0.944	1.362	1.613	1.839	2.114	2.309
	R4	GNO	0.933	0.380	−0.343	0.933	1.304	1.544	1.773	2.066	2.285
	R5	PE3	1.000	0.399	0.810	0.947	1.311	1.534	1.735	1.981	2.157
FCM c	R2	PE3	1.000	0.494	0.764	0.938	1.387	1.660	1.904	2.203	2.415
	R3	PE3	1.000	0.468	0.928	0.929	1.358	1.626	1.872	2.174	2.392
	R6	GLO	0.937	0.212	−0.175	0.937	1.269	1.504	1.752	2.117	2.430
KHM a	R2	PE3	1.000	0.372	0.723	0.956	1.293	1.496	1.678	1.900	2.056
	R3	GEV	0.840	0.418	0.239	0.987	1.367	1.568	1.730	1.902	2.008
	R4	GNO	0.927	0.433	−0.329	0.927	1.347	1.617	1.872	2.197	2.440

Note. GEV: generalized extreme value; GLO: generalized logistics; GNO: generalized normal; probability density function; PE3: Pearson type 3; RP: return period. The a, b, and c parameters represent Euclidean, Mahalanobis, and Manhattan distance measures, respectively.

TABLE 6 Index-flood equations and their respective statistics and cross-validations for all the five remaining clustering combinations

Combination	Region	Index flood	Index-flood statistics			Cross-validation		
			R^2_{adj}	RMSE	NSE	R^2	RMSE	NSE
K-means b	R1	$\widehat{\mu}_i = 0.255 \cdot L^{1.613}$	0.89	559.86	0.81	0.48	1178.36	0.15
	R2	$\widehat{\mu}_i = 1.554 \cdot L^{1.225}$	0.91	86.50	0.92	0.91	108.12	0.88
	R3	$\widehat{\mu}_i = 0.083 \cdot L^{1.817}$	0.95	586.98	0.92	0.89	679.78	0.89
	R4	$\widehat{\mu}_i = 36.855 \cdot A^{0.704} \cdot Ls^{1.182} \cdot L^{0.880}$	0.90	94.64	0.93	0.86	143.96	0.84
K-means c	R1	$\widehat{\mu}_i = 2.460 \cdot A^{0.758} \cdot S^{0.812}$	0.90	273.74	0.86	0.86	423.12	0.67
	R2	$\widehat{\mu}_i = 0.023 \cdot A^{1.344}$	0.83	221.93	0.67	0.66	290.69	0.44
	R3	$\widehat{\mu}_i = 0.199 \cdot L^{1.686}$	0.96	369.71	0.97	0.97	470.29	0.95
	R4	$\widehat{\mu}_i = 1.090 \cdot 10^{-15} \cdot A^{0.673} \cdot Pa^{4.817}$	0.78	114.78	0.74	0.38	191.90	0.27
PAM a	R1	$\widehat{\mu}_i = 0.271 \cdot E^{0.709} \cdot A^{0.550}$	0.83	638.41	0.93	0.34	1517.62	−0.12
	R4	$\widehat{\mu}_i = 20.952 \cdot 10^{3} \cdot L^{2.638} \cdot Ls^{2.755}$	0.96	65.77	0.95	0.53	2650.05	−81.49
	R5	$\widehat{\mu}_i = 7.576 \cdot L^{0.927}$	0.59	238.16	0.63	0.38	333.85	0.27
FCM c	R2	$\widehat{\mu}_i = 0.369 \cdot L^{1.578}$	0.93	463.04	0.95	0.95	600.71	0.91
	R3	$\widehat{\mu}_i = 3.367 \cdot 10^{-26} \cdot L^{1.420} \cdot E^{1.169} \cdot Pa^{6.815}$	0.93	177.19	0.77	0.48	370.21	−0.02
	R6	$\widehat{\mu}_i = 13.833 \cdot 10^{5} \cdot A^{1.103} \cdot Ls^{1.696} \cdot E^{-1.053}$	0.90	114.75	0.86	0.79	159.51	0.73
KHM a	R2	$\widehat{\mu}_i = 0.203 \cdot A^{1.038}$	0.84	115.36	0.81	0.60	298.79	−0.28
	R3	$\widehat{\mu}_i = 1.635 \cdot 10^{-5} \cdot A^{0.879} \cdot E^{1.988}$	0.85	258.43	0.80	0.19	742.56	−0.69
	R4	$\widehat{\mu}_i = 1.507 \cdot L^{1.487} \cdot Kc^{-3.580} \cdot S^{0.338}$	0.91	150.68	0.85	0.41	484.69	−0.52

Note. FCM: Fuzzy C-mean; KHM: K-harmonic means; NSE: Nash–Sutcliffe coefficients; PAM: partitional around medoid; RMSE: root mean square error. The a, b, and c parameters represent Euclidean, Mahalanobis, and Manhattan distance measures, respectively.

which would indicate a good model adjustment, this combination presented the lowest overall RMSE; that is, this would result in lower errors for design flood estimation. The large variety of possible clustering solutions, along with the outstanding results obtained when using algorithms that have not been frequently used for such applications (e.g., KHM), reinforce the importance of the approach proposed in the present study. This is important to be highlighted because no single clustering algorithm can encompass all input data and/or provide the best solution for all hydrological applications.

3.4 | Estimation of design floods

The spatial distribution of clusters for the five best combinations (Table 5) is depicted in Figure 2. It should be mentioned that similar sites within a given grouping assessed by heterogeneity measure do not necessarily need to be geographically contiguous (Srinivas et al., 2008).

For the K-means based combinations, the sites are well distributed over the Rio Grande do Sul State. However, PAM with Euclidean

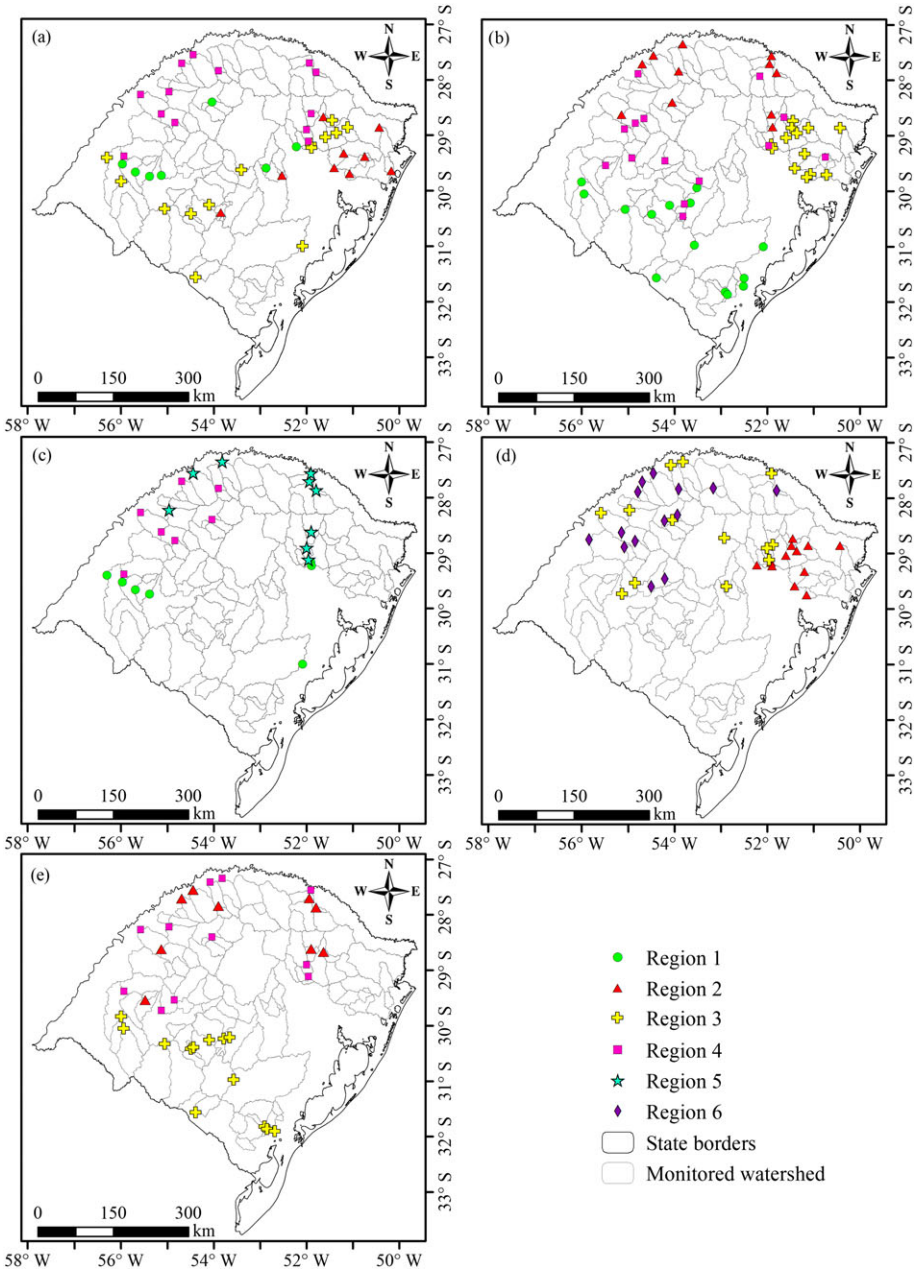

FIGURE 2 Spatial distribution of clustered sites in watersheds of the Rio Grande do Sul State, Brazil. (a and b represent K-means with Mahalanobis and Manhattan distance measures, respectively; c and e represents the algorithms partitional around medoid and K-harmonic means respectively associated with Euclidean distance measure; and d represents Fuzzy C-means with Manhattan distance measure)

measure and FCM with Manhattan measure were not able to provide a cluster in the most southern part of the State. Hydrological monitoring in Brazil is not evenly distributed, and the majority of gauged sites are found where there is an economical interest in monitoring streamflow (e.g., irrigated areas, potential for hydroelectric energy generation, and urban water supply). In this context, the southern Rio Grande do Sul State, which is less developed and has an inferior population density, experiences a lower number of gauging stations and these have shorter temporal series, as reported by Cassalho et al. (2018). Thus, longer distances between sites, as well as a lower representativeness as a consequence of short historical series, might have exerted a negative influence not only on the cluster formation process but also on the adjustment of regional multiple regression equations.

On the other hand, KHM was able to form a hydrologically homogeneous region in the southern part of the state, allowing the fit of one of the PDFs used. However, the most eastern regions formed by such clustering algorithm failed to be fitted to a PDF. For practical purposes, a composition of these five combinations is recommended for practitioners interested in estimating design floods within the state because both the temporal limitation (Table 3) and the watershed's parameters applicable ranges (Table 1) are taken into consideration.

The index flood based on L-moments method is categorized as an intuitive approach for combining at-site information (e.g., sample L-moments and L-moment ratios), although it does not provide a superior performance if compared with the maximum likelihood method (Hosking & Wallis, 1997). According to these authors, the fit of a

single regional PDF to all sites instead of fitting a different PDF to each site individually is what makes RFFA more accurate than at-site FFA. As presented in Equation 23, two equations are necessary for design flood estimates in ungauged sites. The first equation refers to the regional quantile function (i.e., regional growth curve), whereas, the second equation associates watershed's parameters with the mean annual maximum streamflows of the watershed (Razavi & Coulibaly, 2013). The quantile functions for the best fitting distributions are presented in Equations 25 and 26 for GLO and GEV, respectively.

$$\widehat{q}(F) = \xi^R + \frac{\alpha^R}{k^R}\left[1 - (-\log F)^{k^R}\right] \tag{25}$$

$$\widehat{q}(F) = \xi^R + \frac{\alpha^R}{k^R}\left\{1 - \left[\frac{(1-F)}{F}\right]^{k^R}\right\} \tag{26}$$

where $\widehat{q}(F)$ is the regional quantile function, ξ^R, α^R, and k^R are the regional parameters defined for all the regions (Figure 2 and Table 5), and F is the nonexceedance frequency. The quantile functions GNO and PE3 whose parameters are written in terms of ξ, α, and k and μ, σ, and γ, respectively, have no explicit analytical form (for more details, see Hosking & Wallis, 1997).

The regression equations developed for this study were obtained adopting some of the steps presented by Razavi and Coulibaly (2013) for hydrological regionalization studies as follows: (a) selection of watershed attributes to be used as explanatory variables (Table 1); (b) choice of variables of interest (Table 2); (c) establishment of a relationship between these attributes and maximum streamflows (Table 6) ; and (d) assessment of errors by means of R^2_{adj}, RMSE, and NSE and leave-one-out cross-validation (Table 6).

The fitting quality expressed in Table 5 can be also visually assessed in Figure 3, where estimated annual maximum streamflows derived from the regression equations were compared with the observed values (Table 6). Considering PAM along with Euclidean distance, region 5 was that with the lowest R^2_{adj} and NSE when compared with any region formed by the different clustering algorithms, thus resulting in scattered points as illustrated in Figure 3c. On the contrary, region 3 derived from K-means with Manhattan distance and region 4 from PAM with Euclidean distance presented the highest R^2_{adj} and NSE values, leading to a more linear dispersion of points (Figure 3b and c).

In addition to the fitting quality, a leave-one-out cross-validation technique was performed to evaluate the predictive power of the regression equations (Table 6). In the smallest regions with respect to the number of nondiscordant sites, a greater decrease in the R^2 and NSE and an increase in the RMSE were evidenced (Table 3). This may reflect the strong sensitivity of these statistics to small samples. A detailed discussion about the sensitivity of *NSE* to sample size and outliers in power models used in water resources engineering can be found in the study of McCuen, Knight, and Cutter (2006). For combination PAM, region 4 presented an unusually small NSE value (−81.49). This value might have occurred due to the fact that besides having the smallest number of sites possible (seven sites), this region presented an atypical MAS that was not identified in the screening step, corroborating the conclusions of McCuen et al. (2006) regarding the applicability of the NSE. Although NSE has been widely used in hydrology, there is no standardized statistical value that defines whether a validated index-flood equation is acceptable or not. The appraisal of statistical tests linked to validation goes beyond the scope of the present study. Nevertheless, values greater than 0.5 for the R^2 and NSE have been traditionally used as indicators of satisfactory results for hydrological applications, as suggested by Moriasi et al. (2007).

Practitioners interested in estimating design floods using the tools developed in this study first need to identify the site where the ungauged watershed's outlet of interest is located (Figure 1) and then extract the position, physiographic, and meteorological attributes used for clustering (latitude, longitude, L, Kc, Pa). Once these attributes have been extracted, one must check whether this feature vector falls within one of the predefined clusters (Supplementary Material 1). If that is the case, they can identify whether this cloud in the attribute space (i.e., cluster) corresponds to a hydrologically homogeneous region (Table 3). On the contrary, if the feature vector does not fall within the cloud of attributes of any proposed combinations, the ungauged watershed may not be assigned to any region; therefore, the proposed approach cannot be applied. Based on Table 3 and following the recommendations of Rao and Srinivas (2008b), it is possible to determine the temporal limitations for the chosen index-flood equation. Once its temporal limitations are considered, the index-flood method (Equation 23) should be applied by replacing the parameters of the best fit quantile function by the values presented in Table 5 along with the RP of interest. This function should be then multiplied by the index-flood resulting from the corresponding regression equation (Table 6), which was formulated considering the watershed of interest's physiographical and/or climatological attributes. One should also consider the quality of fit between observed and estimated streamflows along with the cross-validation results.

For instance, consider an engineer interested in designing a bridge in southern Rio Grande do Sul State, which will be built over an ungauged river. After extracting the necessary attributes to the corresponding watershed, the engineer found the following values (hypothetical) for latitude, longitude, *L*, *Kc*, *Pa*, respectively: −30.8, −53.7, 129.5, 1.7, and 1,465.2. On plotting these values against the cloud of attributes in the 5-dimensional attribute space for all the combinations, the engineer would observe that the feature vector (i.e., ungauged watershed) would fall within those derived from K-means with Manhattan and KHM with Euclidian (Supplementary Material 1). Alternatively, the engineer may calculate the distance, in terms of dissimilarity measure, between this feature vector and the clusters in all combinations, finally assigning it to the most appropriate one. As one can notice in Table 3, combination K-means with Manhattan distance resulted in a hydrologically homogenous region ($H < 1$) and this consists of 15 sites, allowing design flood estimations for RPs of almost 100 years as suggested by Rao and Srinivas (2008b). On the other hand, KHM with Euclidian distance generate a region classified as possibly homogeneous ($1 \leq H < 2$) composed of 14 sites, allowing design flood estimations for RPs of up to 76 years. Therefore, one may easily conclude that in this context, combination K-means with Manhattan culminated in a region

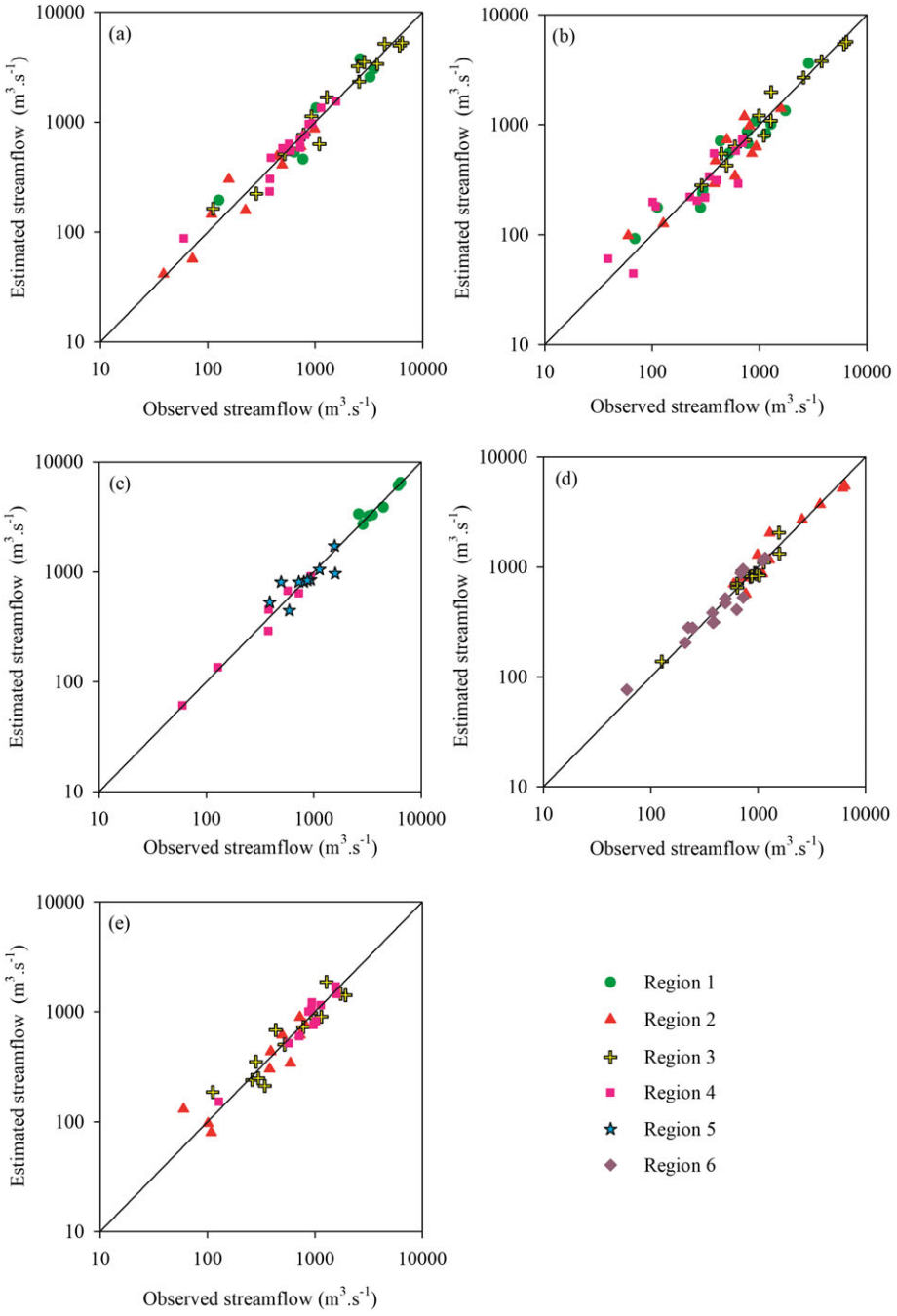

FIGURE 3 Observed and estimated mean maximum annual streamflows for all the five remaining clustering combinations

with superior characteristics for the proposed application. Considering that the best fit distribution for this region is GEV (Table 4) whose quantile function (growth curve) is presented in Equation 26, as well as the RP for which the bridge will be designed, one simply needs to substitute its parameters (ξ^R, α^R, k^R) by the adjusted values (0.850, 0.398, 0.246, respectively; Table 5). Subsequently, this value needs to be multiplied by the result originated from the index-flood equation, which was satisfactorily validated in terms of R^2 and NSE (Moriasi et al., 2007) and written in terms of watershed of interest's area (km^2) and mean slope (%; Table 6). It is worth noting that the growth curve is already calculated for RPs commonly applied to water resources engineering and management applications (Table 5).

Even though the resulting equations can be easily applied, one should always consider the methodological limitations, which may propagate errors and lead to large uncertainties, even after rigorous selection criteria in terms of D_i, H, and Z^{DIST} measures, R^2_{adj}, RMSE, and NSE statistics and cross-validation. Moreover, the index-flood equations should be limited by the lower and upper bounds of their attributes. In other words, one should not use such equations for ungauged watersheds if they extrapolate the magnitude of the used attributes from Table 1, as suggested by Rao and Srinivas (2008a).

In a literature review, Razavi and Coulibaly (2013) identified an increase in the number of regionalization studies in the last decade that address uncertainty analysis. As a matter of fact, recent studies

such as Arsenault and Brissette (2014) have solely focused on evaluating uncertainty of regionalization methods. After a review covering over 70 recent regionalization studies, Razavi and Coulibaly (2013) were able to subdivide streamflow regionalization methods into hydrologic model-dependent and hydrologic model-independent methods. It should be mentioned that the vast majority of regionalization studies that make use of uncertainty analysis listed by these authors, as well as the study of Arsenault and Brissette (2014), were applied to hydrologic model-dependent regionalization methods. Actually, only one study evaluated by Razavi and Coulibaly (2013) was applied to annual maximum series, but such study did not investigate uncertainty estimation. Despite the recent advances in computational procedures, Hsu, Rao, and Srinivas (2008) state that attempts to address the quantification of errors and uncertainties in the context of RFFA based on L-moments, cluster analysis, and index flood have not been properly performed. Hsu et al. (2008) compared three methods used for regional flood quantile estimates to their at-site estimates. The proposed comparative analysis used 75% of all gauged watersheds of Indiana, USA for calibration, whereas the remaining watersheds were used for validation. Errors were estimated by weighting watersheds according to their drainage area. These authors observed significant errors when using regression equations for estimating floods in ungauged locations. Furthermore, the magnitude of errors varied not only as a function of RPs but also between different PDFs, degree of homogeneity, and length of historical series.

It should be pointed out that several assumptions adopted in the present study may limit its applicability, especially for ungauged sites. In such cases, the decision making whether the ungauged station is contained in one of the formed clusters may be labour intensive and subjective (see the example in the Supplementary Material 1). In addition to the existing temporal and spatial limitations as a consequence of the exclusion of sites when forming hydrologically homogenous regions, Hosking and Wallis (1997) emphasize several sources of errors in quantile estimates from RFFA based on L-moments. These different factors are categorized into: (a) estimation procedures (e.g., regional averaging L-moment ratios), (b) specification of the region (e.g., total number of sites within a region and their respective data lengths), and (c) assumptions intrinsic to RFFA (e.g., intersite correlation, hydrological homogeneity, and the fitting of regional PDFs). These authors addressed such factors theoretically by using Monte Carlo simulations in artificial regions. The authors highlighted that even in conditions of moderate heterogeneity, intersite correlation and misspecification of the regional PDF, RFFA is more accurate than at-site FFA. Moreover, the dependence on the length of historical series and the ability of detecting heterogeneity was observed, where the larger the series the easier heterogeneity may be detected (Hosking & Wallis, 1997). Also, a low value of intersite correlation has little effect over quantile estimate bias, and consequently, this should not be a concern. Hosking and Wallis (1997) also stated that errors in quantile estimation are mainly caused by variability in the index-flood estimation. Therefore, practitioners should be aware of the limitation of the proposed method because the detailed evaluation of uncertainty of the flood quantile is beyond the scope of the present study.

4 | CONCLUSIONS

According to the observed results, the proposed method was successfully applied to the study area, providing a robust and reliable tool for practitioners. Thus, some specific conclusions should be highlighted as follows: (a) most of the combinations of clustering algorithms and distances measures resulted in discrepant sized clusters in terms of the total number of sites and total length of the historical series; (b) analysing the number of sites, total length of the historical series, and ability to form homogeneous regions, K-means associated with Mahalanobis and Manhattan distance, PAM and KHM with Euclidean distance, and FCM with Manhattan distance were the best combinations for the study area; (c) such combinations enabled the application of the index-flood method, successfully resulting in appropriate adjustments of PDFs and estimation of their regional parameters; (d) regions with a small number of sites had poor statistics in terms of R^2 and NSE during cross-validation, thus limiting the applicability of the proposed approach to ungauged locations; (e) physiographic attributes, such as A, L, and E, were the most used ones for modelling mean annual maximum streamflow; (f) a composition of all the five best combinations is capable of encompassing all the Rio Grande do Sul State allowing for reliable design flood estimates. In addition, clustering algorithms that have not been so frequently applied to RFFA (e.g., GKA and KHM) have potential to provide useful regionalization alternatives. Thus, the authors reinforce the importance of appraising different clustering algorithms for flood regionalization, as no single solution can fulfil all the practical hydrological applications. Other researchers are encouraged to focus on the identification of better clustering attributes, which have potential to result in more appropriate regions. A special attention should be also given to the identification of the main sources of uncertainty. Such uncertainty analysis should evaluate the data acquisition methods and the reliability of information sources and determine the confidence interval of parameters used for flood estimation and the quantile estimation itself. Moreover, computational efforts should be made towards the development of alternative uses for the discordant and most discordant sites, thus allowing the use of historical series from a greater number of watersheds. In addition, the development of a routine that would automatically assign the ungauged watersheds to the most appropriate region would facilitate the application of the proposed equations.

ACKNOWLEGMENTS

The authors wish to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for scholarships to the second (308645/2017-0) and third (301556/2017-2) authors and for research grant to the second author (485279/2013-4), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for scholarship to the fourth author (88887.178100/2018-00), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for research grant (PPM VIII 071/2014) to the third author, and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) for research grants (2082-2551/13-0; 16/2551-0000 247-9) to the second author.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ORCID

Felício Cassalho https://orcid.org/0000-0001-9496-2910
Samuel Beskow https://orcid.org/0000-0003-3900-0895
Carlos Rogério de Mello https://orcid.org/0000-0002-6033-5342

REFERENCES

Abida, H., & Ellouze, M. (2008). Probability distribution of flood flows in Tunisia. *Hydrology and Earth System Sciences*, *12*, 703–714. https://doi.org/10.5194/hess-12-703-2008

Agarwal, A., Maheswaran, R., Kurths, J., & Khosa, R. (2016). Wavelet spectrum and self-organizing maps-based approach for hydrologic regionalization -a case study in the Western United States. *Water Resources Management*, *30*, 4399–4413. https://doi.org/10.1007/s11269-016-1428-1

Alvares, C. A., Stape, J. L., Sentelhas, P. C., Gonçalves, J. L. M., & Sparovek, G. (2014). Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, *22*(6), 711–728. https://doi.org/10.1127/0941-2948/2013/0507.

Arsenault, R., & Brissette, F. R. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, *50*, 6135–6153. https://doi.org/10.1002/2013WR014898

Asquith, W. H. (2011). *Distributional analysis with L-moments statistics using the R environment for statistical computing* (p. 360). Lubbock, Texas: CreateSpace Independent Publishing Platform.

Aydoğan, D., Kankal, M., & Onsoy, H. (2016). Regional flood frequency analysis for Çoruh Basin of Turkey with L-moments approach. *Journal of Flood Risk Management*, *9*, 69–86. https://doi.org/10.1111/jfr3.12116

Basu, B., & Srinivas, V. V. (2015). Analytical approach to quantile estimation in regional frequency analysis based on fuzzy framework. *Journal of Hydrology*, *524*, 30–43. https://doi.org/10.1016/j.jhydrol.2015.02.026

Beskow, S., Mello, C. R., Vargas, M. M., Corrêa, L. L., Caldeira, T. L., Durães, M. F., & Aguiar, M. S. (2016). Artificial intelligence techniques coupled with seasonality measures for hydrological regionalization of Q90 under Brazilian conditions. *Journal of Hydrology*, *541*, 1406–1419. https://doi.org/10.1016/j.jhydrol.2016.08.046

Caldeira, T. L., Beskow, S., Mello, C. R., Faria, L. C., Souza, M. R., & Guedes, H. A. S. (2015). Probabilistic modelling of extreme rainfall events in the Rio Grande do Sul state. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *19*(3), 197–203. https://doi.org/10.1590/1807-1929/agriambi.v19n3p197-203

Cassalho, F., Beskow, S., Mello, C. R., Moura, M. M., Kerstner, L., & Ávila, L. F. (2018). At-site flood frequency analysis coupled with multiparameter probability distributions. *Water Resources Management*, *32*(1), 285–300. https://doi.org/10.1007/s11269-017-1810-7

Chang, F., Tsai, M., Tsai, W., & Herricks, E. E. (2008). Assessing the ecological hydrology of natural flow conditions in Taiwan. *Journal of Hydrology*, *354*, 75–89. https://doi.org/10.1016/j.jhydrol.2008.02.022

Chavoshi, S., Sulainman, W. N. A., Saghafian, B., Sulainman, M. N. B., & Manaf, L. A. (2013). Regionalization by fuzzy expert system based approach optimized by genetic algorithm. *Journal of Hydrology*, *486*, 271–280. https://doi.org/10.1016/j.jhydrol.2013.01.033

Chérif, R., & Bargaoui, Z. (2013). Regionalisation of maximum annual runoff using hierarchical and trellis methods with topographic information. *Water Resources Management*, *27*, 2947–2963. https://doi.org/10.1007/s11269-013-0325-0

Corrêa, L. L. (2014) Implementação e análise de técnicas de inteligência artificial aplicadas à clusterização em recursos hídricos. Dissertação. Universidade Federal de Pelotas.

Costa, V. (2017). Correlation and regression. In M. Naghettini (Ed.), *Fundamentals of statistical hydrology* (pp. 391–440). Switzerland: Springer. https://doi.org/10.1007/978-3-319-43561-9

Farsadnia, F., Kamrood, M. R., Nia, A. M., Modarres, R., Bray, M. T., Han, D., & Sadatinejad, J. (2014). Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. *Journal of Hydrology*, *509*, 387–397. https://doi.org/10.1016/j.jhydrol.2013.11.050

Forestieri, A., Lo Conti, F., Blenkinsop, S., Cannarozzo, M., Fowler, H., & Noto, L. V. (2018). Regional frequency analysis of extreme rainfall in Sicily (Italy). *International Journal of Climatology*, *38*(S1), e698–e716. https://doi.org/10.1002/joc.5400

Goyal, M. K., & Gupta, V. (2014). Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis. *Water Resources Management*, *28*, 4491–4511. https://doi.org/10.1007/s11269-014-0699-7

Güngör, Z., & Ünler, A. (2008). K-Harmonic means data clustering with tabu-search method. *Applied Mathematical Modelling*, *32*, 1115–1125. https://doi.org/10.1016/j.apm.2007.03.011

Haddad, K., & Rahman, A. (2012). Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework—Quantile regression vs. parameter regression technique. *Journal of Hydrology*, *430-431*, 142–161. https://doi.org/10.1016/j.jhydrol.2012.02.012

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: Concepts and techniques* (3rd ed.) (p. 703). Waltham-USA: Morgan Kaufmann publishers.

Hasenack, H., & Weber, E. (2010). *Base cartográfica vetorial contínua do Rio Grande do Sul—Escala 1:50.000. DVD-ROM. (Série Geoprocessamento n.3)*. Porto Alegre: UFGRS Centro de Ecologia. ISBN 978–85–63483-00-5 (livreto) e ISBN 978–85–63843-01-2 (DVD)

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis: An approach based on L-moments* (p. 224). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511529443

Hsu, E., Rao, A. R., & Srinivas, V. V. (2008). Effect of regionalization on flood frequency analysis. In A. R. Rao, & V. V. Srinivas (Eds.), *Regionalization of watersheds: An approach based on cluster analysis* (pp. 155–211). Dordrecht: Springer Science+Business Media B.V. https://doi.org/10.1007/978-1-4020-6852-2_5

Hussain, Z., & Pasha, G. R. (2009). Regional flood frequency analysis of the seven sites of Punjab, Pakistan, using L-moments. *Water Resources Management*, *23*(10), 1917–1933. https://doi.org/10.1007/s11269-008-9360-7

Instituto Brasileiro de Geografia e Estatística (2016) Estimativas da população residente no Brasil e unidades da federação com data de referência em 1° de julho de 2016. IBGE. ftp://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2016/estimativa_dou_2016_20160913.pdf. [Accessed 16 August 2017].

Instituto Nacional de Meteorologia. (2017). Normais climatológicas do Brasil 1961–1990. http://www.inmet.gov.br/portal/index.php?r=clima/normaisclimatologicas [Accessed 11 August 2017].

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

Jones, M. R., Blenkinsop, S., Fowler, H. J., & Kilsby, C. G. (2014). Objective classification of extreme rainfall regions for the UK and updated estimates of trends in regional extreme rainfall. *International Journal of Climatology*, *34*, 751–765. https://doi.org/10.1002/joc.3720

Kendall, M. G. (1975). *Rank correlation methods*. London: Charles Griffin.

Kingston, D. G., Hannah, D. M., Lawler, D. M., & McGregor, G. R. (2011). Regional classification, variability, and trends for northern North Atlantic river flow. *Hydrological Processes*, *25*, 1021–1033. https://doi.org/10.1002/hyp.7655

Krishma, K., & Murty, M. N. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems Man And Cybernetics—Part B: Cybernetics*, *29*(3), 433–439. https://doi.org/10.1109/3477.764879

Kumar, R., Goel, N. K., Chatterjee, C., & Nayak, P. C. (2015). Regional flood frequency analysis using soft computing techniques. *Water Resources Management*, *29*(6), 1965–1978. https://doi.org/10.1007/s11269-015-0922-1

Lam, D., Thompson, C., & Croke, J. (2017). Improving at-site flood frequency analysis with additional spatial information: A probabilistic regional envelope curve approach. *Stochastic Environmental Research and Risk Assessment*, *31*, 2011–2031. https://doi.org/10.1007/s00477-016-1303-x

Latt, Z. Z., Wittenberg, H., & Urban, B. (2015). Clustering hydrological homogeneous regions and neural network based index flood estimation for ungauged catchments: An example of the Chindwin River in Myanmar. *Water Resources Management*, *29*, 913–928. https://doi.org/10.1007/s11269-014-0851-4

Mann, H. B. (1945). Non-parametric tests against trend. *Econometrica*, *13*, 245–259. https://doi.org/10.2307/1907187

McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash-Sutcliffe efficiency index. *Journal of Hydrologic Engineering*, *11*(6), 597–602. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597).

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Noto, L. V., & La Loggia, G. (2009). Use of L-moments approach for regional flood frequency analysis in Sicily, Italy. *Water Resources Management*, *23*, 2207–2229. https://doi.org/10.1007/s11269-008-9378-x

R Core Team. (2016). R: A language and environment for statistical computing. R foundation for statistical computing. Vienna. Austria. URL https://www.R-project.org/.

Rahman, A., Charron, C., Ouarda, T. B. M., & Chebana, F. (2018). Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Environmental Research Risk Assessment*, *32*(1), 123–139. https://doi.org/10.1007/s00477-017-1384-1

Rao, A. R., & Srinivas, V. V. (2008a). Introduction. In A. R. Rao, & V. V. Srinivas (Eds.), *Regionalization of watersheds: An approach based on cluster analysis* (pp. 1–16). Dordrecht: Springer Science+Business Media B. V. https://doi.org/10.1007/978-1-4020-6852-2_1

Rao, A. R., & Srinivas, V. V. (2008b). Regionalization by hybrid cluster analysis. In A. R. Rao, & V. V. Srinivas (Eds.), *Regionalization of watersheds: An approach based on cluster analysis* (pp. 17–56). Dordrecht: Springer Science+Business Media B.V. https://doi.org/10.1007/978-1-4020-6852-2_2

Rao, A. R., & Srinivas, V. V. (2008c). Regionalization by fuzzy cluster analysis. In A. R. Rao, & V. V. Srinivas (Eds.), *Regionalization of watersheds: An approach based on cluster analysis* (pp. 57–112). Dordrecht: Springer Science+Business Media B.V. https://doi.org/10.1007/978-1-4020-6852-2_3

Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrological Engineering*, *18*(8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Sabourin, A., & Renard, B. (2015). Combining regional estimation and historical flood: A multivariate semiparametric peaks-over-threshold model with censored data. *Water Resources Research*, *51*, 9646–9664. https://doi.org/10.1002/2015WR017320

Sadri, S., & Burn, D. H. (2011). A fuzzy C-means approach for regionalization using a bivariate homogeneity and discordancy approach. *Journal of Hydrology*, *401*, 231–239. https://doi.org/10.1016/j.jhydrol.2011.02.027

Saf, B. (2009). Regional flood frequency analysis using L-moments for the West Mediterranean region of Turkey. *Water Resources Management*, *23*, 531–551. https://doi.org/10.1007/s11269-008-9287-z

Seckin, N., Haktanir, T., & Yurtal, R. (2011). Flood frequency analysis of Turkey using L-moments method. *Hydrological Processes*, *25*, 3499–3505. https://doi.org/10.1002/hyp.8077

Shu, C., & Burn, D. H. (2004). Homogeneous pooling group delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement. *Journal of Hydrology*, *291*, 132–149. https://doi.org/10.1016/j.jhydrol.2003.12.011

Smith, A., Sampson, C., & Bates, P. (2015). Regional flood frequency analysis at the global scale. *Water Resources Research*, *51*, 539–553. https://doi.org/10.1002/2014WR015814

Srinivas, V. V., Tripathi, S., Rao, A. R., & Govindaraju, R. S. (2008). Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *Journal of Hydrology*, *348*, 148–166. https://doi.org/10.1016/j.jhydrol.2007.09.046

Yaseen, Z. M., El-shafie, A., Jaafar, O., Afan, H. A., & Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000-2015. *Journal of Hydrology*, *530*, 829–844. https://doi.org/10.1016/j.jhydrol.2015.10.038

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Cassalho F, Beskow S, de Mello CR, de Moura MM, de Oliveira LF, de Aguiar MS. Artificial intelligence for identifying hydrologically homogeneous regions: A state-of-the-art regional flood frequency analysis. *Hydrological Processes*. 2019;1–16. https://doi.org/10.1002/hyp.13388