An Acad Bras Cienc (2021) 93(Suppl. 4): e20210538 DOI 10.1590/0001-3765202120210538
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

ENGINEERING SCIENCES

Minimum streamflow regionalization in a Brazilian watershed under different clustering approaches

CARINA K. BORK, HUGO A.S. GUEDES, SAMUEL BESKOW, MICAEL DE S. FRAGA & MYLENA F. TORMAM

Abstract: Estimating the minimum streamflows in rivers is essential to solving problems related to water resources. In gauged watersheds, this task is relatively easy. However, the spatial and temporal insufficiency of gauged watercourses in Brazil makes researchers rely on the hydrological regionalization technique. This study's objective was to compare different hierarchical and non-hierarchical clustering approaches for the delimitation of hydrologically homogeneous regions in the state of Rio Grande do Sul, Brazil, aiming to regionalize the minimum streamflow that is equaled or exceeded in 90% of the time (Q_{90}). The methodological development for the regionalization of Q_{90} consisted of using regression analysis supported by multivariate statistics. With respect to independent variables for regionalization, this study considered the morphoclimatic attributes of 100 watersheds located in southern Brazil. The results of this study highlighted that: (i) the clustering techniques had the potential to define hydrologically homogeneous regions, in the context of Q_{90} in the Rio Grande do Sul State, mostly the Ward algorithm associated with the Manhattan distance; (ii) drainage area, perimeter, centroids X and Y, and mean annual total rainfall aggregated important information that increased the accuracy of the cluster; and (iii) the refined mathematical models provided excellent performance and can be used to estimate Q_{90} in ungauged rivers.

Key words: drought indicator, hydrological regionalization, multivariate statistics, Rio Grande do Sul State, ungauged watersheds.

INTRODUCTION

Streamflow estimates are a requirement for solving various engineering problems, such as designing or sizing water control structures, assessment of water availability for different uses (e.g., irrigation, urban and industrial supply), planning and land use management, water quality control, river habitat assessment, among others (Agarwal et al. 2016). Streamflow estimate is relatively easy in gauged watersheds, where a long period of historical records is available. However, the assessment of water availability is challenging in watersheds that are not adequately gauged (Athira et al. 2016). According to Beskow et al. (2016a), the lack of historical records is one of the main difficulties for managing water resources in developing countries such as Brazil. Several researchers have tried to transfer information of streamflows from gauged watersheds to other ungauged watersheds, which corresponds to a process commonly known as hydrological regionalization (Blöschl & Sivapalan 1995, Rao & Srinivas 2006).

The lack of hydrological data on watersheds is troublesome in the analysis of minimum streamflows (Sadri & Burn 2011). According to Li et al. (2010), hydrological regionalization allows

for estimating hydrological indicators without calibration. This convenience is beneficial for water resource planners, who often need to make decisions about ungauged watersheds or watersheds with a short historical series for the recurrence period of interest (Beskow et al. 2016b). The quantiles associated with the minimum streamflows, obtained from the flow duration curves (FDC), are adopted in the state of Rio Grande do Sul to guide water resources planners in the decision making regarding projects that need this analysis. FDC is a graphical representation of the cumulative distribution of the streamflow percentiles in a watershed (Pugliesi et al. 2016), thus allowing the identification of the streamflow that is equaled or exceeded in a given percentage of time (Fouad et al. 2018).

Hydrological regionalization involves two phases: the delimitation of hydrologically homogeneous regions, and the determination of regional equations (Lin & Wang 2006). The homogeneous regions are derived from a set of data representing the characteristics of the watersheds that help explain the hydrological indicators of interest (Haddad et al. 2015). Cluster analysis (CA) has been widely accepted as an essential tool to support hydrological regionalization (Rao & Srinivas 2006), as this technique assists hydrologists in forming homogeneous regions. Jain et al. (1999) stated that CA is a process by which a data set, formed by several objects characterized by different attributes, is divided into groups so that the objects in the same group are more similar, while objects from different groups are considered different.

CA is divided into hierarchical and non-hierarchical methods. The non-hierarchical methods have a specific number of groups with iterative computing algorithms (for example, K-means and genetic algorithms based on artificial intelligence) (Beskow et al. 2016b, Cupak 2017). The hierarchical methods are well known and investigate the data structure at various levels (for example, a single bond, full bond, and Ward) (Elesbon et al. 2015, Farhan & Al-Shaik 2017, Fouad et al. 2018). These approaches provide different results, depending on the area of study, and it is not possible to determine with a certain degree of certainty which approach is more indicated. However, the definition of a clustering methodology for the study region is crucial to obtain reliable regional equations.

Different regionalization methodologies of the minimum streamflow that is equaled or exceeded in 90% of the time (Q_{90}) have been used in different states in Brazil. The regionalization model indicated by Liazi et al. (1988) divided the state of São Paulo into 21 hydrologically homogeneous regions and proposed that the variables annual mean precipitation and drainage area directly affected the Q_{90} estimates. Wolff et al. (2014) proposed a method of regionalization of streamflows for the state of São Paulo based on the spatial interpolation of the specific mean streamflows by the inverse to the square of the distance. In this case, the streamflow was calculated by the mean annual precipitation, without, however, obtaining hydrologically homogeneous regions. The hydrological regionalization model adopted by the state of Minas Gerais divides the state into 25 homogeneous regions, with the drainage area being adopted as an independent variable to estimate the Q_{90} streamflows (IGAM 2012). However, in the Rio Grande do Sul state (Southern Brazil), few researchers studied the hydrological regionalization of minimum streamflows. The most recent study was conducted by Beskow et al. (2016b). The authors divided the state into six homogeneous regions using artificial intelligence techniques coupled with seasonality measures associated

with minimum streamflows, generating regional equations to calculate the Q_{90} based on the watersheds' drainage area. The scarcity of studies on hydrological regionalization has been impairing management bodies' actions in the state of Rio Grande do Sul.

This study aimed to compare the performance of different hierarchical and non-hierarchical clustering approaches to delineate hydrologically homogeneous regions in the state of Rio Grande do Sul, Brazil, with a view to regionalizing the Q_{90}. The study's premises, characterized by different sizes watersheds, were: i) the Q_{90} corresponds uniquely to baseflow, that is, it is little affected by isolated rainfall events; and ii) the watersheds evaluated do not present an unstable regime within a hydrological year.

MATERIALS AND METHODS

The methodological development for the hydrological regionalization of Q_{90} in the state of Rio Grande do Sul consisted of using regression analysis supported by multivariate statistics. The methodology used a Geographic Information System (GIS) software to facilitate the management, reproducibility, and analysis of spatial data (Fraga et al. 2019).

In the study, different clustering approaches were applied in order to increase the reliability of homogeneous regions' design. Subsequently, multivariate regression analysis was applied in order to determine the regional models. The flowchart of the methodology used is illustrated in Figure 1. The following topics detail the main steps for applying the methodology.

Study watershed

This study was carried out in watersheds located in the state of Rio Grande do Sul, in the extreme south of Brazil. The state of Rio Grande do Sul has a subtropical climate with hot and humid summer, corresponding to the Cfa and Cfb type according to the the Köppen classification. The humid subtropical climate with mild summers (Cfb) occurs in the Serra do Sudeste and Serra do Nordeste, where the mean temperatures of the summer months are below 22°C; the Cfa type is predominant in other regions of the state, where the mean temperature of the hottest month exceeds 22°C (Alvares et al. 2013). According to Mello et al. (2013), the rainfall has a

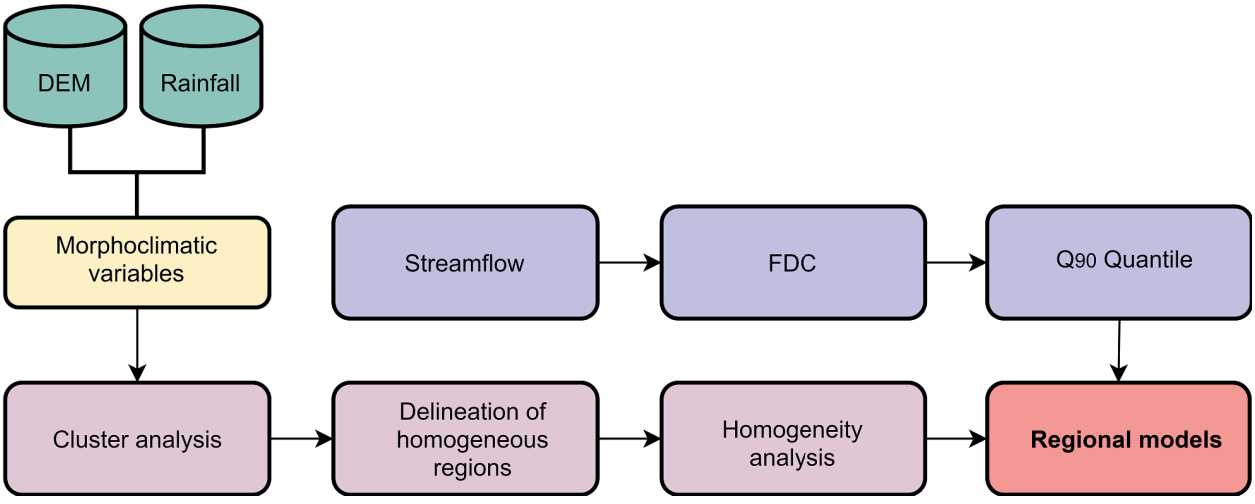

Figure 1. Flowchart of the methodology used for hydrological regionalization of the Q_{90} in the Rio Grande do Sul State, Brazil.

well-defined seasonality, with rainfall relatively well distributed throughout the year.

Data selection and processing criteria

The hydrometeorological series used were obtained from the HidroWeb Portal - Hydrological Information Systems of the National Water Agency (ANA) for the base period until 2017, presenting a mean of 30 years of data. Initially, 120 streamflow gauging stations and 1109 rain gauges were considered. The streamflow gauging stations were selected following the indication by Vezza et al. (2010) and Cupak (2017), who highlighted the use of a minimum period of 10 years of observed daily data. The criteria of Garcia et al. (2017) emphasize that the limit of 10% of missing data for these series should not be exceeded. The rain gauges' selection was performed according to the criterion indicated by Caldeira et al. (2015), adopting stations with a minimum of 10 years of records without missing data.

After applying the initial criteria, hydrological stations were evaluated for trends, according to the Mann-Kendall test (Mann 1945, Kendall 1975), and for homogeneity, according to the Mann-Whitney test (Mann & Whitney 1947). Both tests were performed at a significance level of 5% ($p < 0.05$). A similar methodological procedure was carried out by Uliana et al. (2015), Salviano et al. (2016), Beskow et al. (2016b), and Guedes et al. (2019).

In this study, data from the Brazilian geomorphometric database (TOPODATA) were used, consisting of scenes from the Shuttle Radar Topography Mission (SRTM) with a spatial resolution of 30 meters. The digital elevation model (DEM) was consisted in accordance with the methodology described by Guedes & Silva (2012). ESRI's ArcGIS / ArcMap v.10.5 software was used to define the subwatersheds.

The drainage area (A, in km^2), perimeter (P, in km), centroid X (X, in km), centroid Y (Y, in km), and mean slope (D, in %) of each subwatershed were derived individually from the DEM. Centroids X and Y were used to obtain geographically continuous regions (Rao & Srinivas 2006, Calegario et al. 2020). The mean annual total rainfall (p, in mm) was calculated with the aid of the Thiessen Polygon method using datasets from the rain gauges (Cabral et al. 2016). This method originates from the Voronoi diagrams (Aurenhammer 1991) assuming that at any point in the watershed, the rainfall value is equal to the weighted mean of the nearest rain gauges, being possible to trace the areas of influence of the stations to characterize the spatial variability of rainfall (Souza et al. 2017). These variables are easily obtained and are commonly used in streamflow regionalization studies (Gubareva 2012, Xu et al. 2014, Elesbon et al. 2015).

From the daily streamflow records, the quantile Q_{90} was calculated, in other words, the streamflow that is equalled or excedeed in 90% of the time. In this study, Q_{90} calculation was made for each streamflow gauging station, that is, for each subwatershed (Beskow et al. 2016b). The software SisCAH 1.0 was used to calculate Q_{90}, as described by Vogel & Fennessey (1994). The authors classified the streamflows in descending order in classes according to the magnitude and associated them with the empirical frequencies of exceedance.

Dissimilarity measures

All variables used in this study were standardized to eliminate the effects of dependence on the units and scales in which they were obtained (Gulgundi & Shetty 2018). The Anderson Darling test ($p < 0.05$) assessed the normality of the standardized variables. Following the indication of Elesbon et al. (2015), Spearman's

correlation coefficient (r) was applied toanalyze the correlation of the variables (Gauthier 2001), eliminating the least correlated variables.

Different measures of dissimilarity were assessed to quantify similarity between two objects or clusters: Euclidean Distance (Harris 1955), Manhattan Distance (Sokal & Michener 1957), and Mahalanobis (Mahalanobis 1936). These equations are respectively defined as:

$$d(x,y)_{euclidean} = \sqrt{\sum_{j=1}^{d}\left(x_j - y_j\right)^2} \quad (1)$$

$$d(\mathbf{x},\mathbf{y})_{manhattan} = \sum_{j=1}^{d}\left|x_j - y_j\right| \quad (2)$$

$$d(x,y)_{mahalanobis} = \sqrt{(x-y)^T S^{-1}(x-y)} \quad (3)$$

where d is the number of attributes of each object; x_j and y_j are the variables that represent two objects of the same dataset with d attributes each; $(x - y)^T$ is the vector transposition of values obtained through the difference between the attributes of the objects x and y; S^{-1} is the inverse covariance matrix of the attributes existing in the dataset under analysis.

Clustering approaches

The next step was to cluster watersheds with similar hydrological behavior according to Q_{90}, based on the morphoclimatic variables, evaluating the following hierarchical clustering techniques: simple-linkage, complete-linkage, and Ward. In addition, the K-means non-hierarchical clustering technique was used. All statistical analyzes of clustering by hierarchical and non-hierarchical algorithms were implemented in the program R v.3.3.3, with the aid of the "Biotools" package for calculating the Mahalanobis distance (Silva 2017).

Single-linkage and Complete-linkage

This study applied the single-linkage and complete-linkage clustering algorithms according to Florek et al. (1951) and Lance & Willians (1967), respectively. These algorithms are hierarchical and agglomerative. They build a hierarchy of sets into groups, each following group formed by merging a pair from the collection of previously defined groups (Wilks 2006). The ideal result is a division of data that minimizes the differences between individuals in a given group and maximizes differences between individuals in different groups (Hair Jr et al. 2009). According to Wilks (2006), the distances between pairs of points can be defined unambiguously and stored in a distance matrix. However, even after calculating a distance matrix, there are alternative definitions for distances between groups of points. The choice made for the distance measurement and the criteria used to define the cluster-to-cluster distances essentially define the cluster method.

In the scope of this study, the simple-linkage method consists of a distance matrix (d) (dissimilarity) between the fluviometric stations (individuals). The distance between the G_1 and G_2 clusters is the shortest distance between a G_1 individual and a G_2 individual, defined as:

$$d_{G_1,G_2} = \min\left(d_{i,j}\right); \text{where } i \in G_1 \text{ and } j \in G_2 \quad (4)$$

The complete-linkage clustering method presents a similar procedure to the simple-linkage method. The difference is in the calculation of the distance (d) between the groups (G_1 and G_2), in which the highest value gives it between the groups (Elesbon et al. 2015). The method is defined as:

$$d_{G_1,G_2} = \max\left(d_{i,j}\right); \text{where } i \in G_1 \text{ and } j \in G_2 \quad (5)$$

Ward's method

Ward's method was applied in this study according to the methodology described by Ward Júnior (1963) and used successfully in several other studies (Melo Júnior et al. 2006, Srinivas 2009, Yang et al. 2010, Hassan & Ping 2012, Sharghi et al. 2018). According to Wilks (2006), this clustering method is also hierarchical and agglomerative; however, it does not operate with the distance matrix. As the clustering method is agglomerative, it divides individuals into a dedicated number of groupings into several stages (Eszergár-Kiss & Caesar 2017). Initially, each individual is independent and, step by step, more elements are ordered for a grouping. The method includes the closest individuals to the existing clusters at each stage, minimizing the sum of the square distances between the individuals and the centroid of their respective groups (Wilks 2006). Among all the possible ways of performing the grouping, this algorithm seeks to find the ideal number of storage steps in groups (G), minimizing the objective function:

$$W = \sum_{g=1}^{G}\sum_{i=1}^{n_g} x_i - \overline{x_g}^2 = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{K}\left(x_{i,k} - \overline{x_{g,k}}\right)^2 \quad (6)$$

Ward's method is conservative, monotone, and creates approximately more regular groups but is sensitive to extreme values (Almeida et al. 2007). On comparing this algorithm with other, Eszergár-Kiss & Caesar (2017) reported that Ward's algorithm offers greater precision in the results and minimizes variation between individuals.

In this study, dendograms were generated for each hierarchical clustering algorithms (Wilks 2006, Elesbon et al. 2015). The Calinski and Harabasz index (Calinski & Harabasz 1974) defined the number of homogeneous groups to characterize minimum streamflows.

K-means algorithm

The K-means partition algorithm was employed in this study in accordance with the procedures reported by Hartigan & Wong (1979). K-means is a centroid-based algorithm and has an objective function that is minimized at each iteration along an optimization process known as the iterative relocation technique (Wilks 2006, Beskow et al. 2016b) (Equation 7).

$$F = \sum_{i=1}^{n}\sum_{j=1}^{k} d\left(x_i, y_j\right) \quad (7)$$

In the context of this study, d (x_i, y_j) represented the Euclidean distance. Contrary to hieralchical algorithms, this method does not generate dendrograms. Thus, the number of homogeneous regions formed by the hierarchical algorithms was used as initial assumptions (k values) in the K-means algorithm. According to Beskow et al. (2016b), the algorithm is highly dependent on the initial configuration of the number of clusters. Therefore, poor initializations lead to unrealistic solutions. In this way, the number of regions was changed until reaching a reasonable solution. Primary initializations were also reevaluated as needed.

Regional modeling of Q_{90}, homogeneity analysis, and cross-validation

A mathematical model was adjusted for each homogeneous region relating the Q_{90} to the subwatersheds' morphoclimatic variables by using a potential mathematical model, as suggested by Beskow et al. (2016b) and Uliana et al. (2015). The study used the stepwise backward method and the F test for the selection of variables, both with a 5% significance level ($p < 0.05$), as suggested by Mohamoud (2008), Booker & Snelder (2012), and Aissia et al. (2017). The Nash-Sutcliffe logarithmic efficiency (NSElog) and the R^2_{adjust} coefficient were applied to quantify the performance of the adjusted models. The NSE_{log}

was calculated using Equation (8) and the R^2_{adjust} coefficient was calculated using Equation (9).

$$NSE_{log} = 1 - \frac{\sum_{i=1}^{n}\left(Q_{90o,i,log} - Q_{90e,i,log}\right)^2}{\sum_{i=1}^{n}\left(Q_{90o,i,log} - Q_{90e,m,log}\right)^2} \quad (8)$$

$$R^2_{adjust} = 1 - \frac{(n-1)}{(n-p-1)} \times \left(1 - R^2\right) \quad (9)$$

where $Q_{90o,i,log}$ are the logarithmic values of observed Q_{90}; $Q_{90e,i,log}$ are the logarithmic values of estimated Q_{90}; $Q_{90e,m,log}$ are the mean logarithmized values of estimated Q_{90}; n is the quantity of observed Q_{90} values in the region under analysis; p is the number of independent variables; R^2 is the determination coefficient.

According to Sadri & Burn (2011), it is expected that groups formed by the clustering process do not meet the criteria of homogeneity. Hosking & Wallis (1993) developed the regional test (H) to verify the homogeneity of harmonized regions, which was used in this study. The H test is based on the L-moments and has been used in several studies on hydrological regionalization (Abdolhay et al. 2012, Beskow et al. 2016b, Sharghi et al. 2018, Lelis et al. 2020). The H test classifies the region as follows: homogeneous (|H| < 1), possibly heterogeneous (1 ≤ |H| < 2), and heterogeneous (|H| ≥ 2).

Regional models were also evaluated using the cross-validation procedure, which, according to Vezza et al. (2010), presents advantages to other techniques for evaluation of predictive errors, such as robustness and applicability to all regionalization models. Cross-validation was analyzed only for the best scenario with respect to regions formed, using the confidence index (c), proposed by Camargo & Sentelhas (1997), determination coefficient (R^2), and Mean Absolute Error (MAE), according to the recommendations of Guilhon et al. (2007), Vezza et al. (2010), Elesbon et al. (2015), Beskow et al. (2016b), and Razavi & Coulibaly (2013).

The value of (c) was calculated using the correlation coefficient (r_{correl}) and the accuracy coefficient (d) using Equations (10) and (11): The confidence index (c) was assessed according to the classification proposed by Camargo & Sentelhas (1997): Excellent (c > 0.85); Very Good (0.76 ≤ c ≤ 0.85); Good (0.66 ≤ c ≤ 0.75); Average (0.61 ≤ c ≤ 0.65); Tolerable (0.51 ≤ c ≤ 0.60); Bad (0.41 ≤ c ≤ 0.50); and Terrible (c ≤ 0.40). The mathematical expressions of R^2 and MAE are described in the Equation (12) and Equation (13), respectively.

$$d = 1 - \left[\frac{\Sigma\left(Q_{90e,i} - Q_{90o,i}\right)^2}{\Sigma\left(\left|Q_{90e,i} - Q_{90o,m}\right| + \left|Q_{90o,i} - Q_{90o,m}\right|\right)^2}\right] \quad (10)$$

$$c = r_{correl} \times d \quad (11)$$

$$R^2 = b^2 \times \frac{S^2_{Q_{90o}}}{S^2_{Q_{90e}}} \quad (12)$$

$$MAE = \frac{\sum_{i=1}^{n}\left(Q_{90e,i} - Q_{90o,i}\right)}{n} \quad (13)$$

where $Q_{90e,I}$ are estimated Q_{90} values; $Q_{90o,I}$ are observed Q_{90} values; $Q_{90o,m}$ is the observed mean Q_{90} value. b is the angular coefficient of the regression line; S^2_{Q90o} is the sample variance of the observed Q_{90} values; S^2_{Q90e} is the sample variance of the estimated Q_{90} values.

The NSE_{log} was evaluated according to the classification proposed by Motovilov et al. (1999): Adequate and Good (NSE_{log} > 0.75); Acceptable (0.36 < NSE_{log} ≥ 0.75); and Unsatisfactory (0.36 < NSE_{log}). Regional modeling, cross-validation, as well as the application of the H test were performed using software R v.3.3.3, with the aid of the Lmom (Hosking 2017a) and LmomRFA packages (Hosking 2017b).

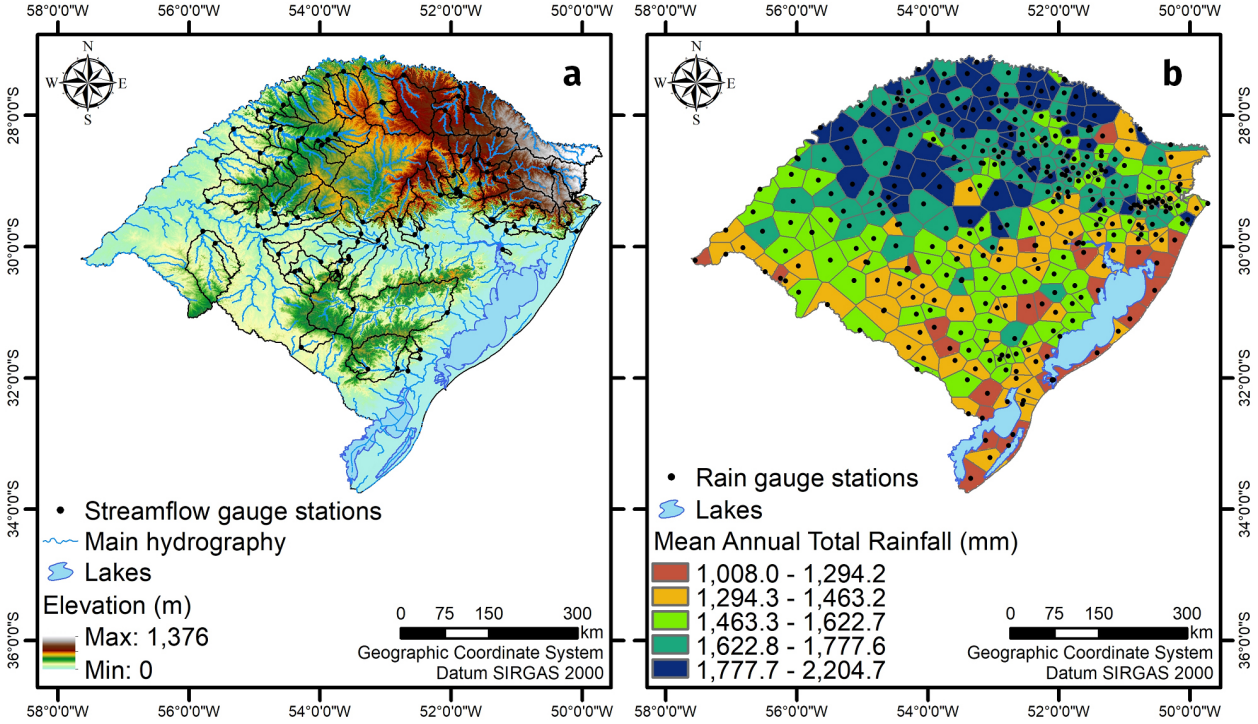

Figure 2. Location of streamflow gauging stations (a) and rain gauges (b) located in the state of Rio Grande do Sul, which met the pre-established criteria. In (a) the digital elevation model and the individualized subwatersheds are presented. In (b), the spatial variability of total annual rainfall is presented using the Thiessen polygons.

RESULTS AND DISCUSSION

Data analysis and dissimilarity measures

Of the 1109 rain gauges initially selected, only 365 contained a historical series with an extension of 10 years or more without missing data. These series were accepted according to the Mann-Kendall and Mann-Whitney tests ($p < 0.05$). For the 305 streamflow gauging stations found in the ANA database, only 170 had data and only 120 series presented length equal to or greater than ten years and a maximum of 31 days of failure. The Mann-Kendall test highlighted the need to exclude series from three streamflow gauging stations. The Mann-Whitney test indicated the acceptance of the other stations. As a result of the analyzes, it was still necessary to remove 17 subwatersheds (corresponding to the streamflow gauging stations) whose drainage areas exceeded the limit of the state of Rio Grande do Sul. At the end of the analyzes, 365 rain gauges and 100 streamflow gauging stations were available and employed for this study (see Figure 2).

In a regionalization study of Q_{90} in the state of Rio Grande do Sul, Beskow et al. (2016b) used 78 stations, which underwent similar selection criteria to that of this study. This difference in the number of stations for this study highlights the importance of frequent updating the series. However, despite the gain in the number of streamflow gauging stations, Figure 2a highlights that the southwest and southeast regions have more scarcity of streamflow gauging stations than the other regions of the state. Overall, the lack of hydrological information is the main difficulty in conducting reliable hydrological studies in Brazilian watersheds.

The spatial distribution of mean annual total rainfall in the state (see Figure 2b) allowed us to identify that the lowest rainfall depths occur in the northeast region. In the opposite

Table I. Sperman (r) correlation matrix between the independent variables analyzed.

Variable	A	P	X	Y	D	p
A	1.00	0.98**	0.90	0.08	0.05	-0.03
P	-	1.00	0.07	0.02	0.10	-0.07
X	-	-	1.00	0.20*	-0.05	0.25*
Y	-	-	-	1.00	0.06	0.66**
D	-	-	-	-	1.00	-0.10
p	-	-	-	-	-	1.00

A – drainage area; P – perimeter; X – centroid X; Y – centroid Y; D – mean slope; p – mean annual total rainfall.
**** significance at 0.01 and * significance at 0.05.**

region, to the northwest, there are the highest rainfall depths. One can observe in Figure 2 that these higher rainfall depths occur in more rugged relief regions, characteristic of orographic rains coming from the central part of Brazil and Argentina (Guedes et al. 2019).

Table I shows the correlation matrix of the independent variables, which were extracted and/or derived from the DEM (see Figure 3a). In this study, Spearman's correlation coefficient was applied when realizing that the variables drainage area, perimeter, and slope did not present a normal distribution. According to Elesbon et al. (2015), this stage of the study is critical since it is possible to assess the importance of each of the variables, promoting the elimination of those that will contribute less, in terms of variability, in the homogeneous regions formed in the flow regionalization process.

In this study, the A and P variables were strongly correlated ($r = 0.98$, $p < 0.01$). The p variable had a significant correlation with Y variable ($r = 0.66$, $p < 0.01$) and X variable ($r = 0.25$, $p < 0.05$), while X variable showed a significant correlation with Y variable ($r = 0.20$, $p < 0.05$). All of these results agree with the results found by Gubareva (2012) and Elesbon et al. (2015). The D variable did not correlate with the other variables analyzed, indicating the possibility of exclusion in the study, as it shows little contribution in the homogeneous regions to be formed. A similar result was found by Elesbon et al. (2015), who observed the lack of correlation between this variable and the other independent variables, excluding it from their study of regionalization of minimum streamflows in the Doce River basin, state of Minas Gerais.

After obtaining and analyzing the independent variables (A, P, X, Y, p), the dendrograms were derived from the application of the clustering algorithms single-linkage (Figure 3), complete-linkage (Figure 4) and Ward (Figure 5) for the Q_{90}, combined with the Euclidean, Mahalanobis and Manhattan distances. Figure 3 displays that the Euclidean (Figure 3a) and Manhattan (Figure 3c) distances did not provide groups' formation by the simple connection algorithm, suggesting that a group was formed with stations 60 and 65, which are characterized by having the subwatersheds with the largest areas and Q_{90}. The other streamflow gauging stations would form a large group. The Mahalanobis distance (Figure 3b) suggests the formation of groups with only one station. According to Kumar et al. (2013), there is no distinction between dependent and independent variables in the cluster analysis, with statistical techniques applied to the same standardized data matrix. Patidar & Verma (2017) highlight that each linking

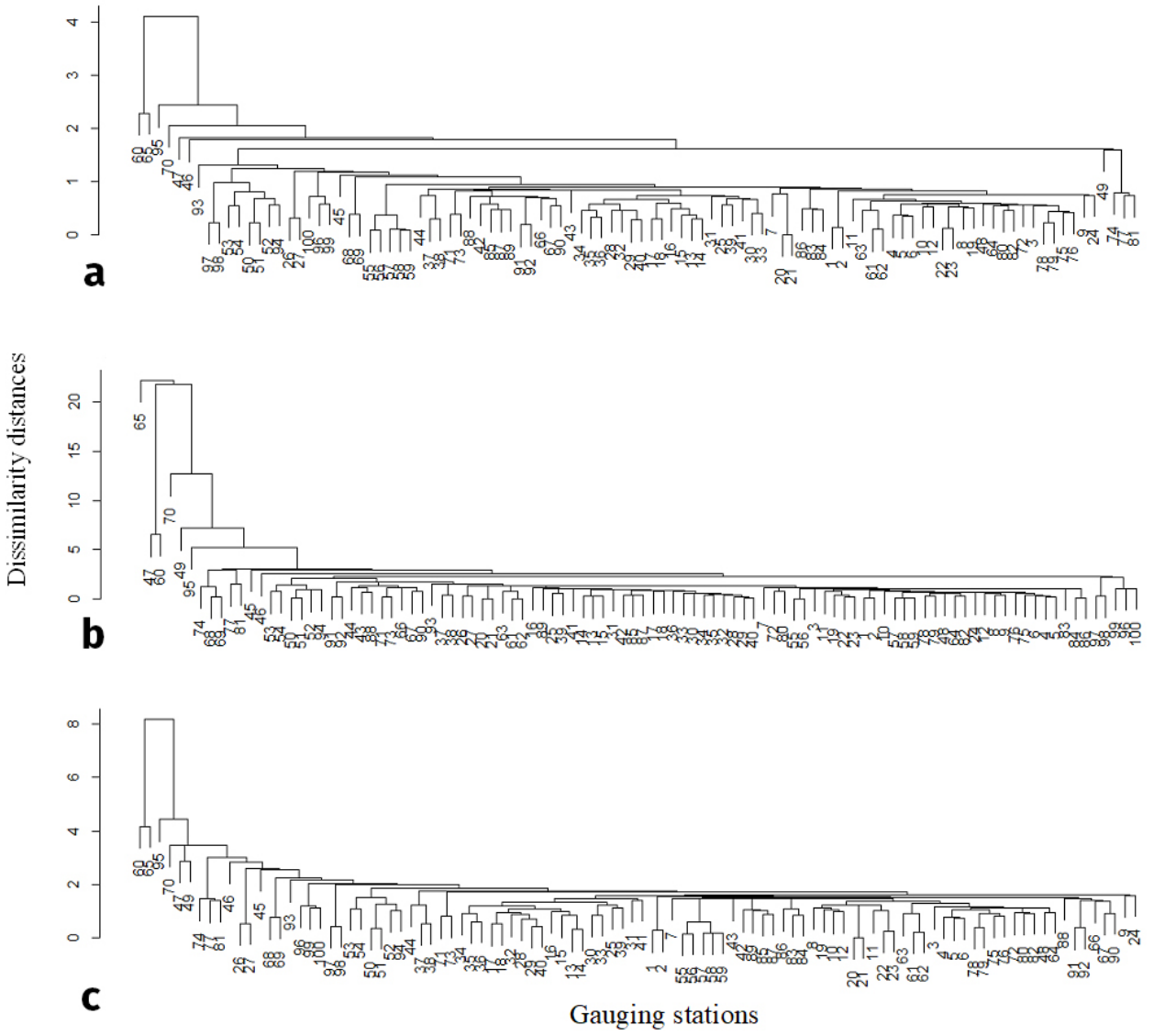

Figure 3. Dendrograms obtained from the combination of the simple-linkage clustering algorithm with the independent variables for Q_{90} taking into account the distances: (a) Euclidean, (b) Mahalanobis, and (c) Manhattan.

algorithm produces different results when used in the same data set. According to Bivand et al. (2017) and Melo et al. (2017), the ideal is that the formed groups present more homogeneous numbers of stations, characterizing outlier groups formed with only one or two stations. In a regionalization study performed by Rao & Srinivas (2006), the groups obtained by the simple-linkage algorithm consisted of a large group and several small branches, indicating that the simple-linkage algorithm was not suitable for the regionalization of streamflows. Melo Júnior et al. (2006) also observed that this algorithm presented irregular clusters and was discarded by the authors. Thus, this algorithm was also discarded from this study.

Figure 4 presents the results for the complete-linkage algorithm. The complete-linkage algorithm represented the groups better, compared to those obtained with the

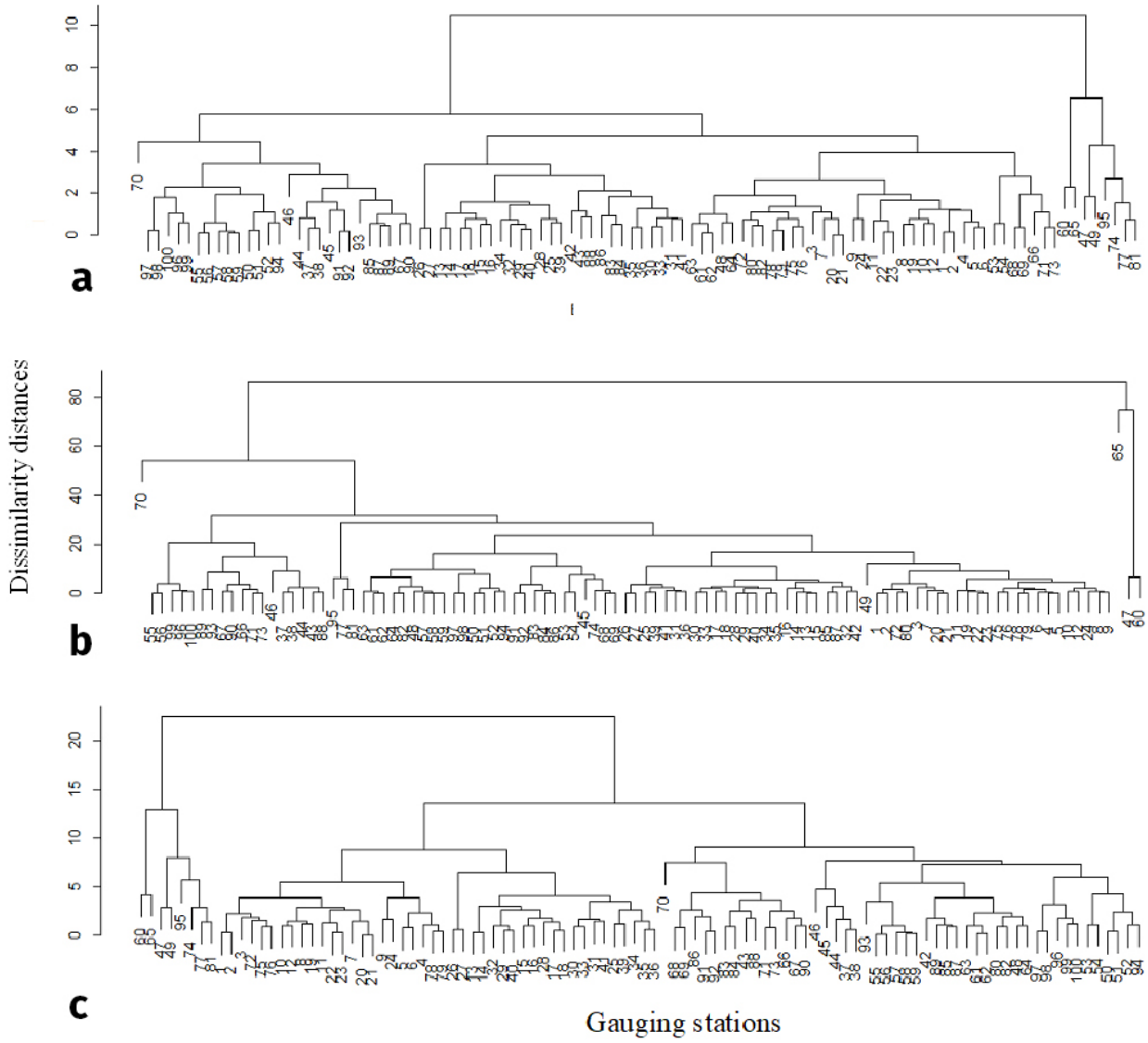

Figure 4. Dendrograms obtained from the combination of the complete-linkage clustering algorithm with the independent variables for Q_{90} taking into account the distances: (a) Euclidean, (b) Mahalanobis, and (c) Manhattan.

simple-linkage algorithm. For the Euclidean distance (Figure 4a), two clusters were formed, according to the methodology described by Calinski & Harabasz (1974). Depending on the cut criterion for the determination of dendograms, up to three clusters were formed for the Manhattan distance (Figure 4c). However, for the Mahalanobis distance (Figure 4b), the results were not satisfactory. In the study by Elesbon et al. (2015), the complete-linkage algorithm, associated with the Mahalanobis distance, was the one that presented the best results. According to Patidar & Verma (2017), the complete-linkage grouping method tends to produce more compact groups and more useful hierarchies than the simple linkage method. However, the simple linkage method is more versatile. The authors, comparing the two algorithms, observed that the precision of the algorithms increases as the number of

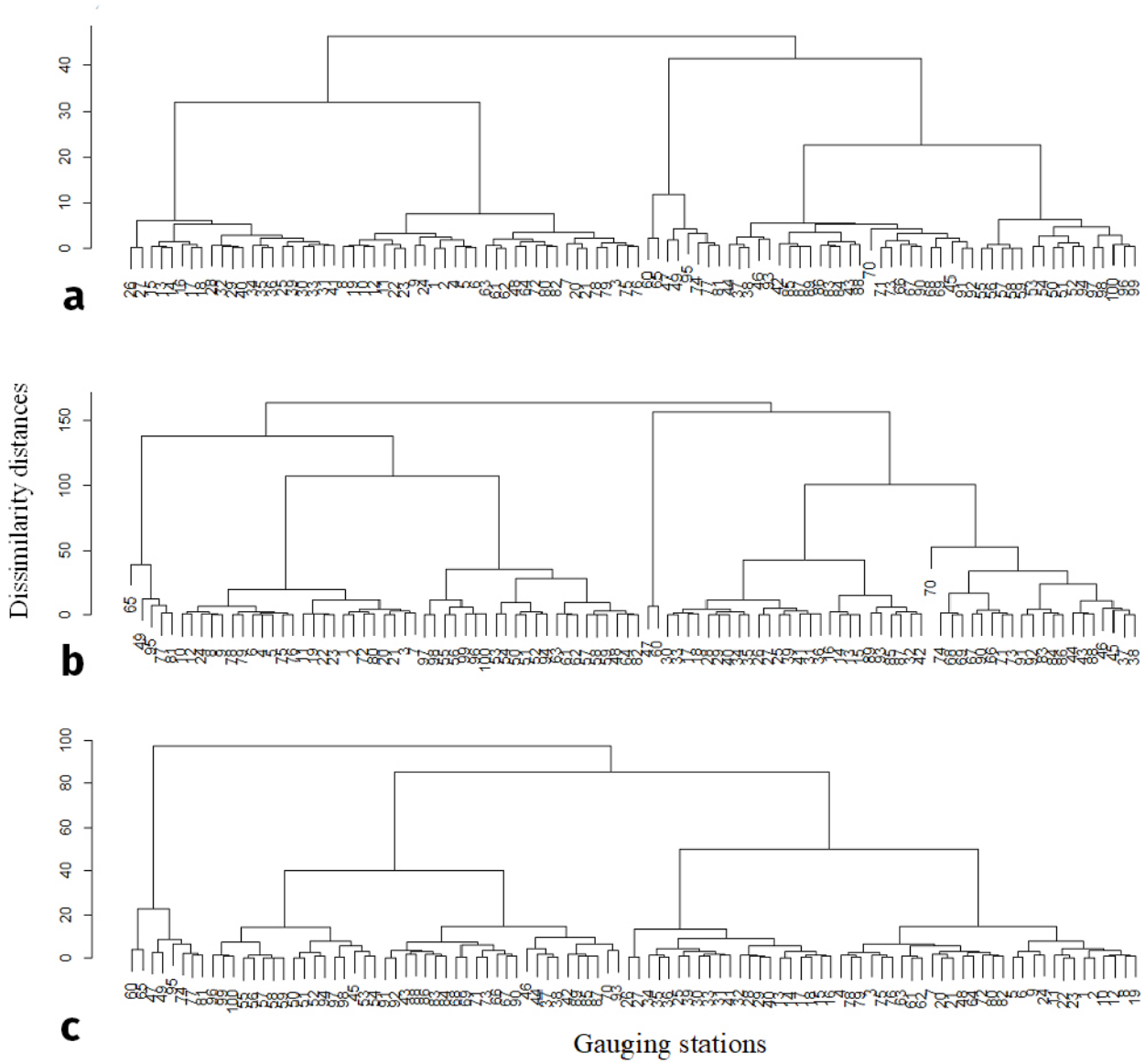

Figure 5. Dendrograms obtained from the combination of the Ward clustering algorithm with the independent variables for Q_{90} taking into account the distances: (a) Euclidean, (b) Mahalanobis, and (c) Manhattan.

individuals grows, reaching a certain level. After this level, there is less clustering efficiency, with the complete linkage algorithm being less susceptible to this loss of efficiency.

Ward's method (Figure 5) provided more possibilities for cutting the dendrogram. The dendrograms could be cut in different positions when using Euclidean (Figure 5a) and Manhattan (Figure 5c), generating up to 5 groups. On the other hand, the Mahalanobis distance (Figure 5b) resulted in only two groups, therefore, this was the most restrictive distance in all algorithms. Rao & Srinivas (2006) observed that the Ward's method presents the groups differently. According to Yang et al. (2010), Ward's method tends to produce small groups with an equal number of individuals. This observation agrees with the results obtained in this study since it was the algorithm that resulted in the largest number of groups. Hassan & Ping (2012)

Table II. Mean values of the confidence index c, considering different homogeneous regions, clustering algorithms, and dissimilarity measures.

Homogeneous regions	Algorithm	Dissimilarity measures*		
		Euclidean	Mahalanobis	Manhattan
2	Complete-linkage	0.71 – 0.82 (0.77)	-	0.71 – 0.82 (0.77)
	Ward	0.73 – 0.79 (0.76)	0.65 – 0.97 (0.81)	0.73 – 0.79 (0.76)
	K-means	0.74 – 0.77 (0.76)		
3	Complete-linkage	-	-	-
	Ward	0.54 – 0.82 (0.70)	-	0.54 – 0.82 (0.70)
	K-means	0.56 – 0.82 (0.69)		
4	Complete-linkage	-	-	-
	Ward	0.54 – 0.82 (0.71)	-	0.54 – 0.82 (0.71)
	K-means	0.58 – 0.82 (0.70)		
5	Complete-linkage	-	-	-
	Ward	0.52 – 0.85 (0.72)	-	0.53 – 0.86 (0.73)
	K-means	0.52 – 0.82 (0.74)		
6	Complete-linkage	-	-	-
	Ward	-	-	-
	K-means	0.47 – 0.81 (0.63)		

*** Values outside the parentheses represent the statistic c minimum and maximum respectively, whereas values inside the parentheses represent the statistic c mean.**

and Melo Júnior et al. (2006) also found that Ward's method provided the best results than other classification algorithms.

The analysis of dendrogram cutting is essential, as it will influence the possibilities of forming homogeneous regions (Melo Júnior et al. 2006). This step's purpose was to select the most versatile algorithm and grouping distance to increase the possibilities of forming different homogeneous regions. Modarres (2010) points out that in studies of regionalization of minimum streamflows for planning water resources, a greater number of groups is recommended to make it possible theinterpretation of the characteristics of the regions and use them with greater confidence in the applicability of the equations. According to Vezza et al. (2010), a single homogeneous region is the simplest way to regionalize. However, these researchers pointed out that the hypothesis of applying a general model generalizes all the different vital processes for the analysis of minimum streamflows, thereby compromising somewhat the management of water resources.

Homogeneous regions

Table II summarizes the minimum, mean and maximum values of the index c considering different homogeneous regions and their regional equations relating Q_{90} to different morphoclimatic variables, following a potential mathematical model for all analyzed scenarios, which combine different grouping algorithms and measures of dissimilarity.

Based on the results presented in Table II, it is possible to observe that the simple-linkage algorithm was not considered in the analysis. This was necessary because the characteristic dendrogram, generated due to the dissimilarity measures, only presented the possibility of forming a group. According to Beskow et al. (2016b), there is no general rule for defining the number of groups formed. In their study on the regionalization of Q_{90} in the state of Rio Grande do Sul using different artificial intelligence algorithms combined with seasonality measures for minimum streamflows, they chose six homogeneous regions as an initial criterion. This decision was derived from the recommendation of Ribeiro et al. (2005). These authors indicated that at least six monitoring stations are needed per region in the regionalization process when studying the formation of homogeneous regions in the Doce River basin (Brazil).

In the context this study, the formation of homogeneous regions was based on the dendrograms generated and analyzed by the Calinski and Harabasz index (Azam et al. 2018). Therefore it was possible to form up to five homogeneous regions by hierarchical algorithms. From the third homogeneous region, only Ward's hierarchical algorithm was used. The non-hierarchical method K-means does not depend on the formation of dendrograms and can form several regions. In this study, K-means was applied to up to six regions, aiming to improve the results obtained by forming five

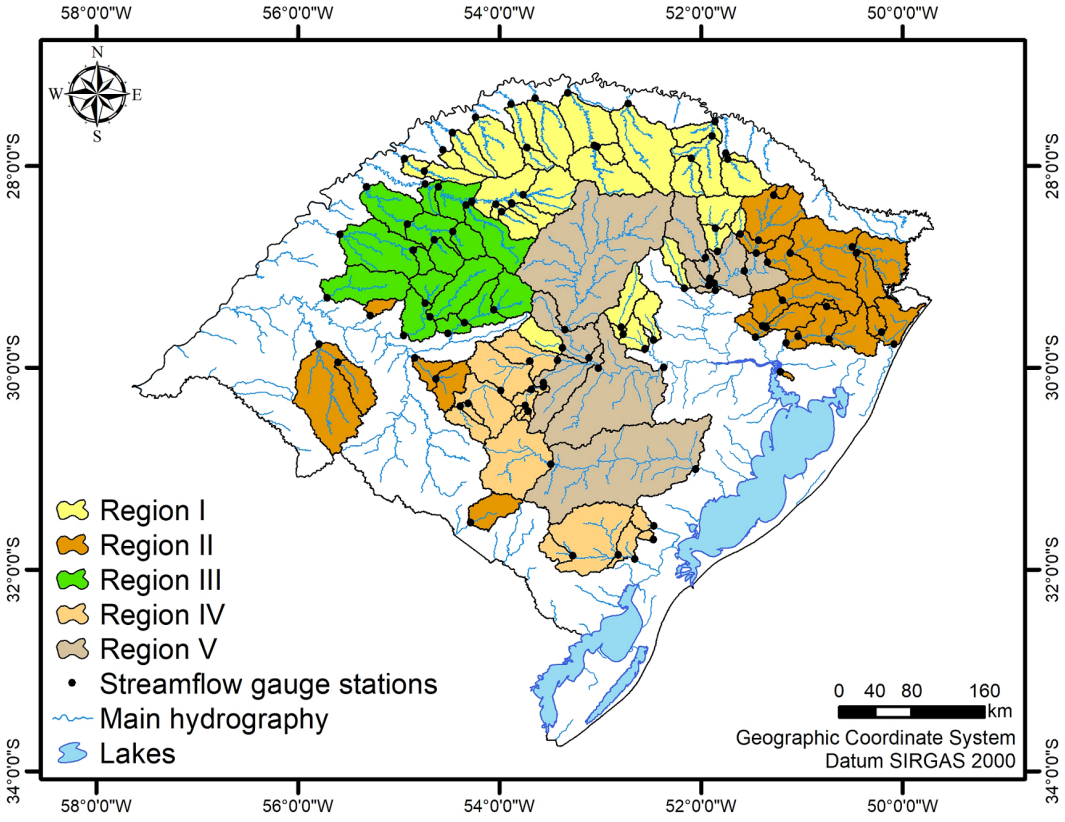

Figure 6. Spatial variability of homogeneous regions determined for the state of Rio Grande do Sul using the K-means algorithm.

regions (c_{mean} = 0.74). However, the result was not satisfactory (c_{mean} = 0.63), concluding that five homogeneous regions were the best for the study area. The methodology used in the studies diversifies the results of the present work from those presented by Beskow et al. (2016b), who adopted six homogeneous regions for the regionalization of Q_{90} in the state of Rio Grande do Sul.

Unlike the results obtained by Beskow et al. (2016b), who found several algorithms classified as Excellent (c > 0.85) according to the confidence index c indicated by Camargo & Sentelhas (1997), in this study, the c values were between Good (0.66 ≤ c ≤ 0.75) and Very Good (0.76 ≤ c ≤ 0.85). The exception was the value resulting from the K-means algorithm for six regions (c = 0.63), classified as Median (0.61 ≤ c ≤ 0.65). The best results found were for two regions, with index c classified as Very Good, highlighting Ward's algorithm with the measure of Mahalanobis dissimilarity (c = 0.81). In turn, this result was the most restrictive measure in this study, presenting satisfactory results for the formation of only two regions with Ward's method. This result agrees with that reported by Elesbon et al. (2015) in the study of hydrological regionalization in the Doce River basin (Brazil) that agrees to define the Mahalanobis measure like the one that presented the best results but using the full link algorithm.

Even with the formation of two homogeneous regions having the best mean "c" indexes, there were discrepancies in each region. For example, considering the best result (c_{mean} = 0.81), region I presented c_{mean} = 0.97 (Excellent); region II, on the other hand, presented c_{mean} = 0.65 (Median). Other combinations also displayed the same contrasts. Thus, with the increase in the number of regions, the regions' division sought to better balance the performance. Figure 6 shows the spatial distribution of homogeneous regions under the best grouping condition, that is, five regions and the K-means algorithm.

Several studies used the K-means algorithm to analyze the formation of homogeneous regions and hydrological regionalization. Rao & Srinivas (2006) compared hierarchical clustering algorithms (single-linkage, complete-linkage and Ward), K-means, and hybrid (K-means with different initializations), for the definition of homogeneous regions and regionalization of maximum annual streamflows in watersheds located in Indiana (USA). They obtained a better overall performance for the hybrid clustering algorithm when combining the K-means algorithm initialized by Ward. To regionalize streamflows monthly in the United States, Agarwal et al. (2016) used the multiscale entropy method based on wavelets coupled with the K-means algorithm for the regionalization of watersheds. The authors concluded that the system used surpasses some existing limitations, especially when data are limited. For a study on hydrological regionalization in India, Swain & Patra (2019) concluded that the K-means algorithm provided robustness in the definition of homogeneous regions.

In this study, however, the K-means algorithm, even with the highest mean values of the c index, showed some inconsistencies in the formation of homogeneous regions. Figure 6 points out that some of the subwatersheds that form region II are distant geographically, indicating a low adjustment. This low adjustment was demonstrated by the confidence index c_{mean}, which was equal to 0.52, classified as "Tolerable" by Camargo & Sentelhas (1997). Furthermore, the R^2_{adjust} for this region was 0.49, considered low by Elebon et al. (2015). According to the authors, this index must be greater than 0.70 for the adjustment to be considered satisfactory. In general, Arsenault & Brissette (2016) consider that, regionalization methods have

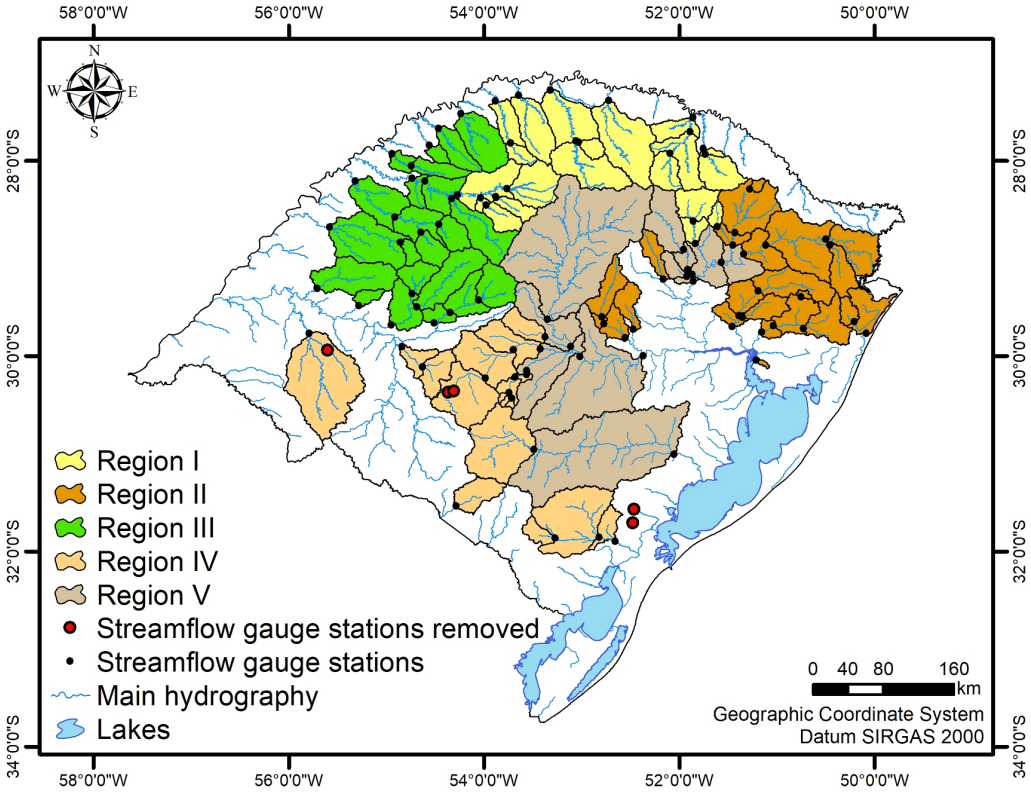

Figure 7. **Homogeneous regions for the state of Rio Grande do Sul determined using Ward's algorithm combined with the Manhattan distance.**

difficulties in defining homogeneous regions with similar hydrological characteristics, mainly in complex regions with different drainage areas, topography, soils, and geometry of the drainage network. This scenario is even more challenging when carrying out the hydrological regionalization of extreme events (Elesbon et al. 2015), that is, the case of Q_{90}. However, additional efforts must be made to ensure homogeneity in the region. This homogeneity can be achieved by excluding watersheds or regions, relocating watersheds, subdividing a region, and merging regions (Hosking & Wallis 1993, Azam et al. 2018). For Singh et al. (2016), spatial proximity is a good indicator of the similarity between the possible groups formed, presenting a greater connectivity degree. However, Corduas (2011) believes that the geographical proximity of the grouped locations is usually inferred as a restriction on the study of regionalization. This previous result may indicate the necessity for another region.

Therefore, to obtain geographically closer regions and with acceptable statistical indexes, some subwatersheds were relocated, as indicated by Farsadnia et al. (2014) and Azam et al. (2018). However, the reallocation alone was not enough for region II to be homogeneous. In this way, five subwatersheds were excluded in order to achieve the best result. All combinations were applied, but Ward's method combined with the Manhattan distances presented the best statistical result (Figure 7). Rao & Srinivas (2006), Farsadnia et al. (2014), and Beskow et al. (2016b) performed the same procedure, confirming that, even after the definition of homogeneous regions by different clustering techniques, some regions may not meet the homogeneity requirements. Therefore, their studies emphasized that the review of the formed regions is critical. Ward's method is widely used in hydrological regionalization, as the studies by Malekinezhad et al. (2011), Ilorme & Griffis (2013), and Farhan & Al-Shaikh (2017), allowing to achieve satisfactory results of homogeneous grouping regions.

Table III. Regional equations determined for the state of Rio Grande do Sul, with statistical indices used to adjust functions, cross-validation, and regions' homogeneity.

Region	n[a]	Regional Equation[b] (Q_{90})	Function Adjustment		Cross-Validation			H[d]
			NSE_{log}	R^2_{adjust}	c[c]	R^2	MAE[b]	
I	24		0.76	0.75	0.81	0.76	3.22	-0.11
II	23		0.64	0.72	0.66	0.74	4.87	-0.20
III	23		0.78	0.77	0.66	0.78	3.32	-0.09
IV	17		0.82	0.80	0.78	0.82	1.09	0.64
V	8		0.81	0.73	0.82	0.81	24.14	-0.18

[a] n is the number of gauging stations.
[b] Q_{90}, in $m^3 s^{-1}$; A, in km^2; P, in km, p, in mm; and MAE in $m^3 s^{-1}$.
[c] Confidence coefficient c results, Camargo & Sentelhas (1997).
[d] H test results, Hosking & Wallis (1993).

Regional modeling of Q_{90}

After acquiring the homogeneous regions, the next step was to determine the regional equations for each region. We opted to analyze morphoclimatic variables as descriptors of Q_{90} because they are easy to obtain, making the planning and management of water resources in the state of Rio Grande do Sul more executable. Table III shows the regional Q_{90} equations for each homogeneous region, the number of monitoring stations by region, and the statistical indices used in the adjustment of functions, cross-validation, and the regions' homogeneity. Lisboa et al. (2008) also used the potential mathematical model to regionalize minimum streamflows in the Paracatu River watershed (Brazil). The authors concluded that the function adequately represented the Q_{90}. The same finding was reported by Elesbon et al. (2015) and Beskow et al. (2016b).

In this study, three variables represented the regional equations: drainage area (regions III and IV), perimeter (regions I, II, and V), and mean annual total rainfall (region V). The drainage area variable is the most present in the flow regionalization equations. In the comprehensive review by Razavi & Coulibaly (2013), the drainage area is one of the most used attributes and presents more satisfactory models. Several studies verified the application of this variable, such as those by Lisboa et al. (2008), Pruski et al. (2012), Elesbon et al. (2015), Beskow et al. (2016b), Farhan & Al-Shaikh (2017), Fouad et al. (2018) and Pagliero et al. (2019). Although the drainage area was not used as independent variable for some regional equations in this study, it was somehow represented by the perimeter. According to Gasques et al. (2018), the drainage area correlates well with the other physical characteristics of the watershed and influences water availability throughout the hydrography.

The incorporation of different independent variables provides strength to regional equations, as performed by Pruski et al. (2012) in a study conducted in the Pará River watershed, a tributary of the São Francisco River (Brazil). The authors added the variable annual mean rainfall to regionalize minimum streamflows. Similarly, Elesbon et al. (2015) incorporated the

precipitation variable in the regionalization equation of the Q_{90} in the Doce River watershed (Brazil). According to Melati & Marcuzzo (2016), the rainfall that occurs in the watershed directly interferes with the behavior of minimum streamflows, explaining why the inclusion of precipitation as an explanatory variable can represent a significant improvement in the streamflow regionalization model.

This study's regional equations were generated using the streamflow gaugingc stations from the ANA database until 2017. It is noteworthy that this database is updated continuously, and the water resources manager should pay attention to the temporal variability of the data. The equations have application limits for each variable, which is given according to the gauging stations present in each subwatershed. Pruski et al. (2015) stressed that extrapolating the regression equations beyond the limits of the sample data used to estimate the regression model's parameters is not recommended. Thus, the limits of application of the equations by homogeneous region are: perimeter between 97.78 to 468.82 km (region I), 96.40 to 633.71 km (region II) and 767.68 to 1533.71 km (region V); a drainage area between 70.68 to 8292.29 km^2 (region III) and 68.61 to 7891.90 km^2 (region IV); and mean annual total rainfall between 1455.2 to 1693.6 mm (region V).

Table III highlights that the equations for the homogeneous regions I, II, III, IV, and V displayed Adequate and Good adjustments ($NSE_{log} > 0.75$), according to Motovilov et al. (1999). The equation for region II was Acceptable ($0.36 < NSE_{log} \geq 0.75$). All the equations had $R^2_{adjust} > 0.70$, the adjustment being considered satisfactory. This result agrees with that reported by Elesbon et al. (2015), who recommended $R^2_{adjus} > 0.70$. Besides, all regions were homogeneous, according to Hosking & Wallis (1993) ($|H| < 1$). It should be noted that this final configuration for the regionalization of the Q_{90} in the state of Rio Grande do Sul was the only one to achieve these results.

Cross-validation (Table III) indicated that the predictive capacity of regions I ($c = 0.81$; $R^2 = 0.76$), IV ($c = 0.78$; $R^2 = 0.82$) and V ($c = 0.82$; $R^2 = 0.81$) was Very Good, according to the classification proposed by Camargo & Sentelhas (1997). The predictive capacity of regions II and III was classified as Good ($c = 0.66$). It can be observed on Table III that the MAE of region V presented the higest value in this study, probably because of the small number of fluviometric stations grouped. Results of cross-validation can be considered an impotant instrument of evaluation of regionalization function generated for each homogenous region, promoting a greater consistency to the analysis on the regional point of view (Beskow et al. 2016a). This conclusion demonstrates that the cross-validation allows to verify the predictive capacity of regionalization function in each watershed of the region of interest. Other studies on hydrological regionalization area used the cross-validation to evaluate mathematical models adjusted and proved the efficiency of this procedure, suh as the studies of Li et al. (2010), Vezza et al. (2010) and Beskow et al. (2016a).

Figure 8 shows the Q-Q plots graphs, which compare the estimated and observed Q_{90} values for the Rio Grande do Sul state. The gauging stations more distant from the 1:1 line were at the bottom, which characterizes an underestimation of the Q_{90}. Thus, in practice, the application of the regional equations obtained in this study tends to benefit the conservation of water resources. However, this underestimation can be high, causing the conflicts of different users of water resources to increase. Within this perspective, for Wolff et al. (2014), it is plausible to establish less conservative grant criteria and the adoption of minimum seasonal streamflows, which better

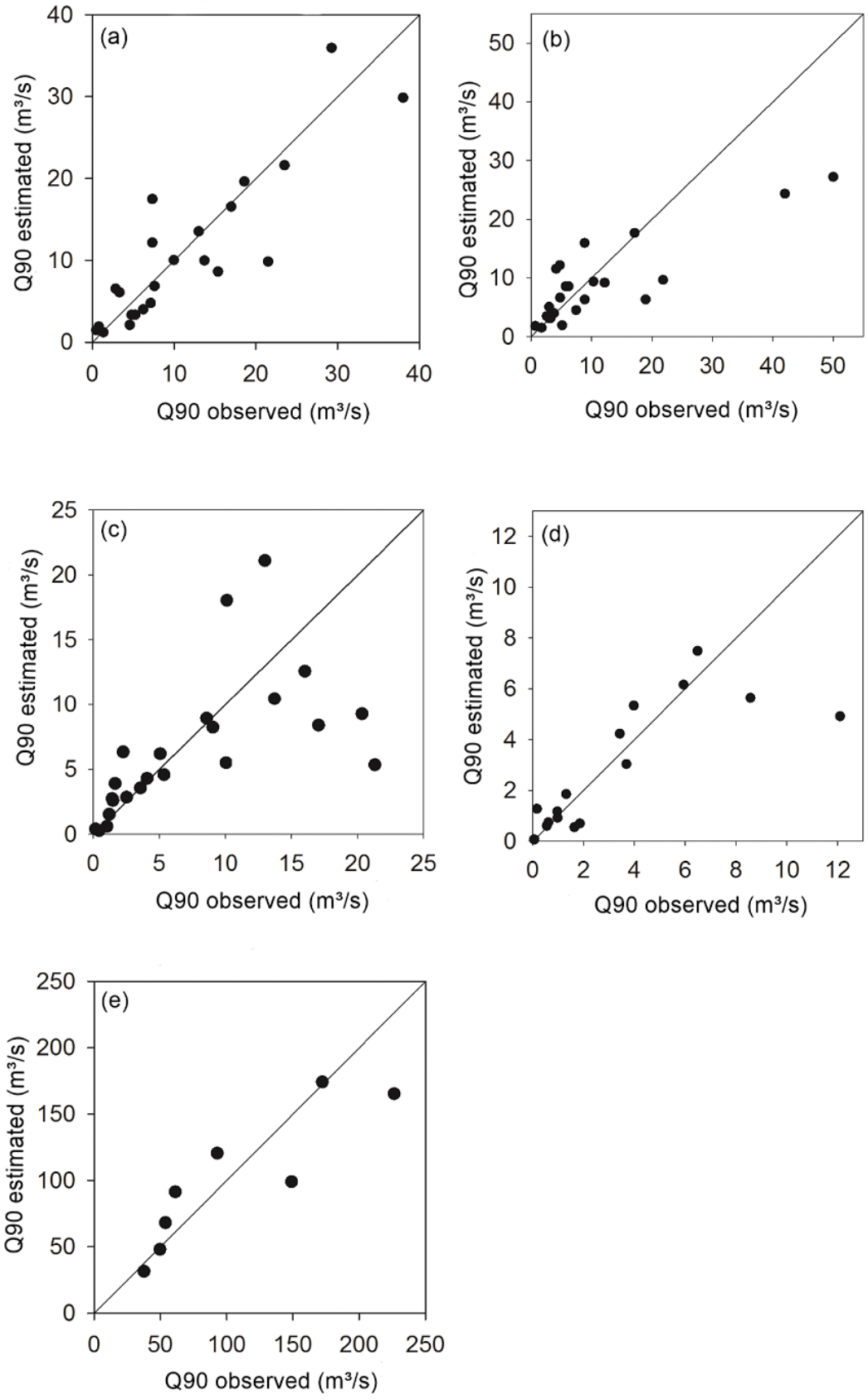

Figure 8. Co-relation between observed and estimated Q_{90}: a) Region I; b) Region II; c) Region III; d) Region IV; e) Region V.

correspond to the quarterly conditions of greater water scarcity. According to the authors, in this way, greater water availability would be likely to be guaranteed in rainy periods.

Another point that must be evaluated is the watersheds' anthropic activity, which alters the availability of low streamflows. According to Pruski et al. (2011), the minimum streamflows are much more susceptible to changes than the mean streamflows, for example. Thus, studying the seasonal behavior of minimum streamflows is extremely important because, in addition to influencing the use of water resources, they also influence the maintenance of aquatic ecosystems (Queiroz et al. 2010).

Besides, there is the possibility of the results found in this study be used in order to best calibrate the models of hydrological forecast and decision making related to the minimum flows (Mishra & Desai 2005, Goyal & Sharma 2016). In general, hydrological models require input data in space and time scale, usually expansive and of difficult access, what can influence on its performance concerning the simulation of hydrological processes (Beskow et al. 2016a). Each input parameter needs to go over an exhaustive process of calibration which, depending on the model used, can take some minutes or several hours. Thus, having a regional equation that represents local conditions in a study minimizes possible uncertainties in the modelling process.

CONCLUSIONS

The results of this study are consistent with the following conclusions:

I) Clustering techniques have the potential to define hydrologically homogeneous regions for the regionalization of Q_{90} in southern Brazil, mainly the Ward method associated with the Manhattan distance, which was the most effective.
II) Morphoclimatic attributes (drainage area, perimeter, centroids X and Y, mean annual total rainfall) added important information related to Q_{90} facilitating the clustering.
III) The adjusted regional models had an excellent performance to estimate the Q_{90}, requiring explanatory variables (drainage area, perimeter and mean annual total rainfall) easy to obtain, which can be justified by the statistics used in this study to assess the accuracy of the adjustment and the predictive capacity (cross-validation). Such models can be applied in the Rio Grande do Sul state, southern Brazil, to estimate Q_{90} in watercourses without streamflow monitoring.
IV) The results found permit to guide the choice of the best method to be used in regionalization studies in the Rio Grande do Sul State. Furthermore, the present study can be used for the best calibration of the models of hydrological forecast and decision making.

Acknowledgments

The authors wish to thank the Agência Nacional de Águas e Saneamento Básico (ANA) for the hydro-meteorological data used.

REFERENCES

ABDOLHAY A, SAGHAFIAN B, SOOM MAM & GHAZALI AHB. 2012. Identification of homogeneous regions in Gorganrood basin (Iran) for the purpose of regionalization. Nat Hazards 61: 1427-1442.

AGARWAL A, MAHESWARAN R, SEHGAL V, KHOSA R, SIVAKUMAR B & BERNHOFER C. 2016. Hydrologic regionalization using wavelet-based multscale entropy method. J Hydrol 538: 22-32.

AISSIA M-AB, CHEBANA F & OUARDA TBMJ. 2017. Multivariate missing data in hydrology – Review and applications. Adv Water Resour 110: 299-309.

ALMEIDA J, BARBOSA LMS, PAIS A & FORMOSINHO S. 2007. Improving hierarchical cluster analysis: A new method

with outlier detection and automatic clustering. Chemometr Intell Lab 87: 208-217.

ALVARES CA, STAPE JL, SENTELHAS PC, DE MORAES GONÇALVES JL & SPAROVEK G. 2013. Köppen's climate classification map for Brazil. Meteorol Z 22: 711-728.

ARSENAULT R & BRISSETTE F. 2016. Analysis of continuous streamflow regionalization methods within a virtual setting. Hydrolog Sci J 61: 2680-2693.

ATHIRA P, SUDHEER KP, CIBIN R & CHAUBEY I. 2016. Predictions in ungauged basins: an approach for regionalization of hydrological models considering the probability distribution of model parameters. Stoch Environ Res Risk Assess 30: 1131-1149.

AURENHAMMER F. 1991. Voronoi diagrams - a survey of fundamental geometric data structure. ACM Comput Surv 23: 345-405.

AZAM M, PARK HK, MAENG SJ & KIM HS. 2018. Regionalization of Drought across South Korea Using Multivariate Methods. Water 10: 24.

BESKOW S, MELLO CR, VARGAS MM, CORRÊA LL, CALDEIRA TL, DURÃES MF & AGUIAR MS. 2016b. Artificial intelligence techniques coupled with seasonality measures for hydrological regionalization of Q_{90} under Brazilian conditions. J Hydrol 541: 1406-1419.

BESKOW S, TIMM LC, TAVARES VEQ, CALDEIRA TL & AQUINO LS. 2016a. Potential of the LASH model for water resources management in data-scarce basins: a case study of the Fragata River basin, southern Brazil. Hydrol Sci J 61: 2567-2578.

BIVAND R, WILK J & KOSSOWSKI T. 2017. Spatial association of population pyramids across Europe: The application of symbolic data, cluster analysis and join-count tests. Spat Stat-Neth 21: 339-361.

BLÖSCHL G & SIVAPALAN M. 1995. Scale issues in hydrological modelling: a review. Hydrol Process 9: 251-290.

BOOKER DJ & SNELDER TH. 2012. Comparing methods for estimating flow duration curves at ungauged sites. J Hydrol 434-435: 78-94.

CABRAL SL, CAMPOS JNB, DA SILVEIRA CS & PEREIRA JMR. 2016. O intervalo de tempo para uma máxima previsibilidade da precipitação sobre o semiárido brasileiro. Rev Bras Meteorol 31: 105-113.

CALDEIRA TL, BESKOW S, MELLO CR, FARIA LC, SOUZA MR & GUEDES HAS. 2015. Modelagem probabilística de eventos de precipitação extrema no estado do Rio Grande do Sul. Rev Bras Eng Agr Amb 19: 197-203.

CALEGARIO AT, PRUSKI FF, RIBEIRO RB, RAMOS MCA & REGO FS. 2020. Physical analysis of regionalized flow as an aid in the identification of hydrologically homogeneous regions. Eng Agric 40: 334-343.

CALINSKI T & HARABASZ J. 1974. A dendrite method for cluster analysis. Commun Stat-Theor M 3: 1-27.

CAMARGO AP & SENTELHAS PC. 1997. Avaliação do desempenho de diferentes métodos de estimativas da evapotranspiração potencial no Estado de São Paulo, Brasil. Rev Bras Agrometeorol 5: 89-97.

CORDUAS M. 2011. Clustering streamflow time series for regional classification. J Hydrol 407: 73-80.

CUPAK A. 2017. Initial results of nonhierarchical cluster methods use for low flow Grouping. J Ecol Eng 18: 44-50.

ELESBON AAA, DA SILVA DD, SEDIYAMA GC, GUEDES HAS, RIBEIRO CAAS & RIBEIRO CBM. 2015. Multivariate statistical analysis to support the minimum streamflow regionalization. Eng Agríc 35: 838-851.

ESZERGÁR-KISS D & CAESAR B. 2017. Definition of user groups applying Ward's method. Transp Res Proc 22: 25-34.

FARHAN Y & AL-SHAIKH N. 2017. Quantitative Regionalization of W. Mujib-Wala Sub-Watersheds (Southern Jordan) Using GIS and Multivariate Statistical Techniques. Open J Mod Hydrol 7: 165-199.

FARSADNIA F, KAMROOD MR, NIA AM, MODARRES R, BRAY MT, HAN D & SADATINEJAD J. 2014. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. J Hydrol 509: 387-397.

FLOREK K, LUKASZEWICZ J, PERKAL J, STEINHAUS H & ZUBRZYCKI S. 1951. Taksonomia Wroclawska. Prz Antropol 17: 193-211.

FOUAD G, SKUPIN A & TAGUE CL. 2018. Regional regression models of percentile flows for the contiguous United States: Expert versus data-driven independent variable selection. J Hydrol Reg Stud 17: 64-82.

FRAGA MS, DA SILVA DD, ELESBON AAA & GUEDES HAS. 2019. Methodological proposal for the allocation of water quality monitoring stations using strategic decision analysis. Environ Monit Assess 191: 776-795.

GARCIA F, FOLTON N & OUDIN L. 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? Hydrolog Sci J 62: 1149-1166.

GASQUES ACF, NEVES GL, SANTOS JD, MAUAD FF & OKAWA CMP. 2018. Regionalização de vazões mínimas: breve revisão teórica. Rev Eletrônica Eng Civ 14: 60-70.

GAUTHIER TD. 2001. Detecting Trends Using Spearman's Rank Correlation Coefficient. Environ Forensics 2: 359-362.

GOYAL MK & SHARMA A. 2016. A fuzzy c-means approach regionalization for analysis of meteorological drought homogeneous regions in western India. Nat Hazards 84: 1831-1847.

GUBAREVA TS. 2012. Classification of River Basins and Hydrological Regionalization (as Exemplified by Japan). Geogr Nat Resour 33: 74-82.

GUEDES HAS & DA SILVA DD. 2012. Comparison between hydrographically conditioned digital elevation models in the morphometric characterization of watersheds. Eng Agric 32: 932-943.

GUEDES HAS, PRIEBE PS & MANKE EB. 2019. Tendências em séries temporais de precipitação no Norte do Estado do Rio Grande do Sul, Brasil. Rev Bras Meteorol 34: 283-291.

GUILHON LGF, ROCHA VF & MOREIRA JC. 2007. Comparação de Métodos de Previsão de Vazões Naturais Afluentes a Aproveitamentos Hidroelétricos. Rev Bras Recur Hídr 12: 13-20.

GULGUNDI MS & SHETTY A. 2018. Groundwater quality assessment of urban Bengaluru using multivariate statistical techniques. Appl Water Sci 8: 1-15.

HADDAD K, JOHNSON F, RAHMAN A, GREEN J & KUCZERA G. 2015. Comparing three methods to form regions for design rainfall statistics: two case studies in Australia. J Hydrol 527: 62-76.

HAIR JR JF, BLACK WC, BABIN BJ, ANDERSON RE & TATHAM RL. 2009. Análise multivariada de dados 6. ed. Porto Alegre: Bookman, 688 p.

HARRIS CW. 1955. Characteristics of two measures of profile similarity. Psychometrika 20: 289-297.

HARTIGAN JA & WONG MA. 1979. A K-means clustering algorithm. J R Stat Soc C-Appl 28: 100-108.

HASSAN BGH & PING F. 2012. Formation of homogenous regions for Luanhe basin – by using L-moments and cluster techniques. Int J Environ Sci Dev 3: 205-210.

HOSKING JRM. 2017a. Package lmom - L-Moments. Available at: < https://CRAN.R-project.org/package=lmom>. Accessed: nov 2017.

HOSKING JRM. 2017b. Package lmomRFA - Regional Frequency Analysis using L-Moments. Available at: < https://CRAN.R-project.org/package=lmomRFA>. Accessed: nov 2017.

HOSKING JRM & WALLIS JR. 1993. Some statistics useful in regional frequency analysis. Water Resour Res 29: 271-281.

IGAM - INSTITUTO MINEIRO DE GESTÃO DAS ÁGUAS. 2012. Estudo de regionalização de vazão para o aprimoramento do processo de outorga no Estado de Minas Gerais. Belo Horizonte: IGAM, 415 p.

ILORME F & GRIFFIS VW. 2013. A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation. J Hydrol 492: 151-162.

JAIN AK, MURTY MN & FLYNN PJ. 1999. Data clustering: a review. ACM Comput Surv 31: 264-323.

KENDALL MG. 1975. Rank Correlation Methods. 4. ed. New York: Hafner Press, 160 p.

KUMAR S, SINGH SK & MISHRA P. 2013. Multivariate Analysis: An Overview. J Dentofac Sci 2: 19-26.

LANCE GN & WILLIAMS WT. 1967. A general theory of classificatory sorting strategies: 1. hierarchical systems. Comput J 9: 373-380.

LELIS LCS, NASCIMENTO JG, DUARTE SN, PACHECO AB, BOSQUILIA RWD & WOLFF W. 2020. Assessment of hydrological regionalization methodologies for the upper Jaguari River basin. J S Am Earth Sci 97: 102402.

LI M, SHAO Q, ZHANG L & CHIEW FHS. 2010. A new regionalization approach and its application to predict flow duration curve in ungauged basins. J Hydrol 389: 137-145.

LIAZI A, CONEJO JL, PALOS JCF & CINTRA PS. 1988. Regionalização hidrológica no estado de São Paulo. Rev Águas Energia Elétrica 5: 4-10.

LIN G-F & WANG C-M. 2006. Performing cluster analysis and discrimination analysis of hydrological factors in one step. Adv Water Resour 29: 1573-1585.

LISBOA L, MOREIRA MC, SILVA DD & PRUSKI FF. 2008. Estimativa e regionalização das vazões mínimas e média na bacia do rio Paracatu. Rev Eng Agric 16: 471-479.

MAHALANOBIS PC. 1936. On the Generalized Distance in Statistics. Proc Nati Inst Sci Indian 2: 49-55. Available at: < http://hdl.handle.net/123456789/6765>. Accessed: jun 2017.

MALEKINEZHAD H, NACHTNEBEL HP & KLIK A. 2011. Comparing the index-flood and multiple-regression methods using L-moments. Phys Chem Earth 36: 54-60.

MANN HB. 1945. Non-parametric test against trend. Econometrica 13: 245-259.

MANN HB & WHITNEY DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 18: 50-60.

MELATI MD & MARCUZZO FFN. 2016. Regressões simples e robusta na regionalização da vazão Q95 na Bacia Hidrográfica do Taquari-Antas. Ciênc Nat 38: 722-739.

MELLO CR, VIOLA MR, BESKOW S & NORTON LD. 2013. Multivariate models for annual rainfall erosivity in Brazil. Geoderma 202-203: 88-102.

MELO CCA, OLIVEIRA KS, ANGÉLICA RS & PAZ SPA. 2017. Análise de agrupamentos associada a difratometria de raios-x: Uma classificação mineralógica prática de bauxitas e seus produtos de digestão Bayer. Holos 6: 32-40.

MELO JÚNIOR JCF, SEDIYAMA GC, FERREIRA PA & LEAL BG. 2006. Determinação de regiões homogêneas quanto à distribuição de frequência de chuvas no leste do Estado de Minas Gerais. Rev Bras Eng Agríc Ambiente 10: 408-416.

MISHRA AK & DESAI VR. 2005. Drought forecasting using stochastic models. Stoch Environ Res Risk Assess 19: 326-339.

MODARRES R. 2010. Regional dry spells frequency analysis by L-Moment and multivariate analysis. Water Resour Manag 24: 2365-2380.

MOHAMOUD YM. 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrolog Sci J 53: 706-724.

PAGLIERO L, BOURAOUI F, DIELS J, WILLEMS P & MCINTYRE N. 2019. Investigating regionalization techniques for large-scale hydrological modelling. J Hydrol 570: 220-235.

PATIDAR D & VERMA VK. 2017. Identify best similarity matrix to find accurate cluster using dendrogram distance. Int J Sci Eng Sci 1: 11-14.

PRUSKI FF, NUNES AA, REGO FS & SOUZA MF. 2012. Extrapolação de equações de regionalização de vazões mínimas: Alternativas para atenuar os riscos. Water Resour Irrig Manag 1: 51-59.

PRUSKI FF, RODRIGUEZ RG, NUNES AA, PRUSKI PL & SINGH VP. 2015. Low-flow estimates in regions of extrapolation of the regionalization equations: A new concept. Eng Agríc 35: 808-816.

PRUSKI FF, RODRIGUEZ RG, SOUZA JF, DA SILVA BMB & SARAIVA IS. 2011. Conhecimento da disponibilidade hídrica natural para a gestão dos recursos hídricos. Eng Agríc 31: 67-77.

PUGLIESI A, FARMER WH, CASTELLARIN A, ARCHFIELD SA & VOGEL RM. 2016. Regional flow duration curves: Geostatistical techniques versus multivariate regression. Adv Water Resour 96: 11-22.

RAO AR & SRINIVAS VV. 2006. Regionalization of watersheds by hybrid-cluster analysis. J Hydrol 318: 37-56.

QUEIROZ MMF, SAMPAIO SC, GOMES BM & LOST C. 2010. Estudo de vazões mínimas $Q_{1,10}$ e $Q_{7,10}$ de rios do Paraná segundo distribuição generalizada. Rev. Verde 5: 32-46.

RAZAVI T & COULIBALY P. 2013. Streamflow prediction in ungauged basins: review of regionalization methods. J Hydrol Eng 18: 958-975.

RIBEIRO CBM, MARQUES FA & SILVA DD. 2005. Estimativa e regionalização de vazões mínimas de referência para a bacia do rio Doce. Rev Eng Agríc 13: 103-107.

SADRI S & BURN DH. 2011. A fuzzy C-means approach for regionalization using a bivariate homogeneity and discordancy approach. J Hydrol 401: 231-239.

SALVIANO MF, GROPPO JD & PELLEGRINO GQ. 2016. Análise de tendência em dados de precipitação e temperatura no Brasil. Rev Bras Meteorol 31: 64-73.

SHARGHI E, NOURANI V, SOLEIMANI S & SADIKOGLU F. 2018. Application of different clustering approaches to hydro-climatological catchment regionalization in mountainous regions, a case study in Utah State. J Mt Sci 15: 461-484.

SILVA AR. 2017. Biotools: Tools for biometry and applied statistics in agricultural science. R package version 3.1. Available at: < http://CRAN.R-project.org/package=biotools>. Accessed: jun 2017.

SINGH SK, MCMILLAN H, BÁRDOSSY A & CHEBANA F. 2016. Non-parametric catchment clustering using the data depth function. Hydrolog Sci J 61: 2649-2667.

SOKAL RR & MICHENER CD. 1957. The effects of different numerical techniques on the phenetic classification of bees of the *Hoplitis* complex (Megachilidae). Proc Linn Soc Lond 178: 59-74.

SOUZA NS, SOUZA WJ & CARDOSO JMS. 2017. Caracterização hidrológica e influência da cobertura do solo nos parâmetros de vazão do Rio das Fêmeas. Eng Sanit Ambient 22: 453-462.

SRINIVAS VV. 2009. Regionalization of watersheds using soft computing techniques. J Hydraul Eng 15: 170-193.

SWAIN JB & PATRA KC. 2019. Impact of catchment classification on streamflow regionalization in ungauged catchments. SN Appl Sci 1: 456-469.

ULIANA EM, SILVA DD, ULIANA EM, RODRIGUES BS & CORRÊDO LP. 2015. Análise de tendência em séries históricas de vazão e precipitação: uso de teste estatístico não paramétrico. Rev Ambient Água 10: 82-88.

VEZZA P, COMOGLIO C, ROSSO M & VIGLIONE A. 2010. Low flows regionalization in North-Western Italy. Water Resour Manag 24: 4049-4074.

VOGEL RM & FENNESSEY NM. 1994. Flow duration curves I: new interpretation and confidence intervals. J Water Res Plan Man 120: 485-504.

WARD JÚNIOR JH. 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58: 236-244.

WILKS DS. 2006. Statistical Methods in the Atmospheric Sciences. 2. ed. San Diego: Academic Press, 648 p.

WOLFF W, DUARTE SN & MINGOTI R. 2014. Nova metodologia de regionalização de vazões, estudo de caso para o Estado de São Paulo. Rev Bras Rec Híd 19: 21-33.

XU C, SHENG S, CHI T, YANG X, AN S & LIU M. 2014. Developing a quantitative landscape regionalization framework integrating driving factors and response attributes of landscapes. Landscape Ecol Eng 10: 295-307.

YANG T, SHAO Q, HAO Z-C, CHEN X, ZHANG Z, XU C-Y & SUN L. 2010. Regional frequency analysis and spatiotemporal pattern characterization of rainfall extremes in the Pearl River Basin, China. J Hydrol 380: 386-405.

How to cite
BORK CK, GUEDES HAS, BESKOW S, FRAGA MS & TORMAM MF. 2021. Minimum streamflow regionalization in a Brazilian watershed under different clustering approaches. An Acad Bras Cienc 93: e20210538. DOI 10.1590/0001-3765202120210538.

Manuscript received on April 10, 2021;
accepted for publication on August 19, 2021

CARINA K. BORK[1]
https://orcid.org/0000-0003-3068-4303

HUGO A.S. GUEDES[1]
https://orcid.org/0000-0002-3592-9595

SAMUEL BESKOW[1]
https://orcid.org/0000-0003-3900-0895

MICAEL DE S. FRAGA[2]
https://orcid.org/0000-0002-1996-9343

MYLENA F. TORMAM[3]
https://orcid.org/0000-0002-6543-4809

[1]Programa de Pós-Graduação em Recursos Hídricos, Universidade Federal de Pelotas (UFPel), Rua Benjamin Constant, 01, 96010-170 Pelotas, RS, Brazil

[2]Instituto Mineiro de Gestão das Águas (IGAM), Cidade Administrativa do Governo de Minas Gerais, Prédio Minas, 31630-900 Belo Horizonte, MG, Brazil

[3]Universidade Federal de Pelotas (UFPel), Centro de Engenharias, Rua Benjamin Constant, 989, 96010-020 Pelotas, RS, Brazil

Correspondence to: **Hugo A.S. Guedes**
E-mail: hugo.hydro@gmail.com

Author contributions

The author Carina Bork is responsible for the research development and execution. Authors Hugo Guedes and Samuel Beskow are the advisor of the research project and for the article writing. Author Micael Fraga is responsible for the preparation of figures and paper review. Authors Mylena Tormam and Carina Bok are responsible for the statistical analysis.