ISSN 1807-1929
Revista Brasileira de Engenharia Agrícola e Ambiental
Brazilian Journal of Agricultural and Environmental Engineering
v.28, n.11, e277005, 2024
Campina Grande, PB – http://www.agriambi.com.br – http://www.scielo.br/rbeaa

DOI: http://dx.doi.org/10.1590/1807-1929/agriambi.v28n11e277005

Original Article

Cluster analysis and hydrological regionalization for Brazilian states[1]

Análise de cluster e regionalização hidrológica para Estados brasileiros

Thaís da S. Charles[2], Tárcio R. Lopes[3], Sérgio N. Duarte[2*], Jéssica G. Nascimento[4], Hugo de C. Ricardo[2], Adriano B. Pacheco[5] & Fernando C. Mendonça[2]

[1] Research developed at Universidade de São Paulo/Escola Superior de Agricultura "Luiz de Queiroz"/Departamento de Biossistemas, Piracicaba, SP, Brazil
[2] Universidade de São Paulo/Escola Superior de Agricultura "Luiz de Queiroz"/Departamento de Biossistemas, Piracicaba, SP, Brazil
[3] Universidade de Maringá, Maringá, PR, Brasil
[4] University of Nebraska/Daugherty Water For Food Global Institute, Lincoln, USA
[5] Universidade Federal Rural da Amazônia/Campus Tomé-Açu, Tomé-Açu, PA, Brazil

HIGHLIGHTS:

Seven homogeneous hydrological regions were obtained using cluster analysis.
Five models showed excellent performance, while two showed satisfactory to good performance based on classification indices.
GIS models are useful for water agencies in Goiás state and the Brazilian Federal District to grant water use rights.

ABSTRACT: Streamflow data from gauging stations are essential for effective water resources management. However, some regions in Brazil lack the necessary data. Hydrological regionalization is an alternative technique for obtaining data such regions. However, not all regions in Brazil have defined hydrological regionalization models, including the state of Goiás and the Brazilian Federal District. The objective of this study was to develop a hydrological regionalization methodology based on the separation of hydrologically homogeneous regions and multiple linear regression, using a Geographic Information System (GIS) program. Historical series data were used to calculate reference flows with 90 or 95% duration over time in the watercourse (Q90 and Q95) and the mean flow ($\bar{Q}$). Rain gauge station data were used to calculate the mean annual rainfall in each watershed through spatial interpolation by ordinary kriging. Subsequently, the physiographic characteristics of each watershed were calculated. The hydrologically homogeneous regions were delimited based on these data using cluster analysis, which identified seven hydrologically homogeneous regions in Goiás, two of them belonging to the Federal District. Multiple regression allowed the development of seven regionalization models. Models for regions 1, 3, 4, 5, and 7 showed better performance.

Key words: homogeneous hydrological regions, hydrological modeling, multiple regression

RESUMO: Para gerenciar os recursos hídricos de forma eficaz, os dados de vazão das estações hidrométricas são cruciais. No entanto, no Brasil, algumas regiões não dispõem dos dados necessários. A regionalização hidrológica é uma técnica alternativa para obter dados para essas regiões. Infelizmente, nem todas as regiões do Brasil têm modelos de regionalização hidrológica definidos, incluindo Goiás e o Distrito Federal. O objetivo deste estudo foi desenvolver uma metodologia de regionalização hidrológica, baseada na separação de regiões hidrologicamente homogêneas e na regressão multivariada, utilizando o programa Sistema de Informações Geográficas (SIG). Com base na série histórica, foram calculadas as vazões de referência com 90 ou 95% de permanência ao longo do tempo no curso d'água (Q90 e Q95) e a vazão média ($\bar{Q}$). Com os dados das estações pluviométricas, foi calculada a precipitação média anual de cada uma das bacias hidrográficas fluviométricas, por meio de interpolação espacial por krigagem ordinária. Posteriormente, foram calculadas as características fisiográficas de cada bacia. Com esses dados, as regiões hidrologicamente homogêneas foram delimitadas por meio da análise de agrupamento. A análise de cluster identificou sete regiões hidrologicamente homogêneas em Goiás, sendo que duas delas pertencem ao Distrito Federal. A regressão multivariada levou ao desenvolvimento de sete modelos de regionalização. Os modelos para as regiões 1, 3, 4, 5 e 7 apresentaram melhor desempenho.

Palavras-chave: regiões hidrológicas homogêneas, modelagem hidrológica, regressão multivariada

• Ref. 277005 – Received 27 Jul, 2023
* Corresponding author - E-mail: snduarte@usp.br
• Accepted 05 Jun, 2024 • Published 02 Jul, 2024
Editors: Lauriane Almeida dos Anjos Soares & Carlos Alberto Vieira de Azevedo

Introduction

Hydrological models have been used as alternatives for predicting hydrological events in ungauged locations. Effective estimates of hydrological variables, such as flow, in these locations assist in objective decision-making in water resources management (Macedo et al., 2023).

The regionalization of hydrological variables is one of those alternatives (Beskow et al., 2016; Gomes et al., 2018). It consists of transferring data from a gauged (donor) to an ungauged location (target) (Lelis et al., 2020).

Parametric regression is among the most widely used regionalization methods (Cassalho et al., 2019), in which the values of the desired parameter are determined through multiple regression between dependent (flow) and independent variables (morphometrics characteristics) (Beskow et al., 2016; Manke et al., 2022).

Wolff et al. (2014), Nascimento et al. (2021), and Wolff & Duarte (2021) conducted hydrological regionalization for the states of São Paulo, Paraná, and Santa Catarina, Brazil, based on data from 176, 81, and 74 gauging stations, respectively. The present study faced challenges with only 77 gauging stations in a larger area, resulting in regionalization issues, especially northern Goiás state due to fewer stations, requiring interpolation from distant regions.

Thus, the general objective of the present study was to develop a methodology for hydrological regionalization based on the use of a Geographic Information System (GIS), with separation of hydrologically homogeneous regions by cluster analysis, using multiple linear regression within each homogeneous region.

Material and Methods

The state of Goiás, in the Central-West Region of Brazil, is between the parallels 12° 23’ S and 19° 29’ S and meridians 45° 54’ W and 53° 14’ W, with a mean altitude of 496 m (Figure 1). Its area is approximately 340,111 km^2, representing 4% of the total area of the country. The state comprises 246 municipalities, including its capital, Goiânia (IMB, 2014).

The Brazilian Federal District is also in the Central-West Region of Brazil, between the parallels 15° 30’ S and 16° 03’ S and the meridians 47° 25’ W and 48° 12’ W, with an altitude of 1,000 m, surrounded by 10 municipalities of Goiás, representing a small territorial strip to the east that borders the state of Minas Gerais (Figure 1). It is the smallest autonomous territory in Brazil, with an area of approximately 5,783 km^2 (IMB, 2014).

The region has four climate types (Am, Aw, Cwa, and Cwb), according to the Köppen-Geiger classification, with predominance of Aw (Rohli et al., 2015), a tropical climate with two well-defined seasons (dry and rainy seasons), with an mean annual rainfall depths ranging from 1,200 to 2,500 mm (Cardoso et al., 2014).

The region’s relief is characterized by low altimetric amplitude, mostly flat terrains. The soils found in the region are predominantly classified as Latossolo Vermelho-Amarelo, according to the Brazilian Soil Classification System

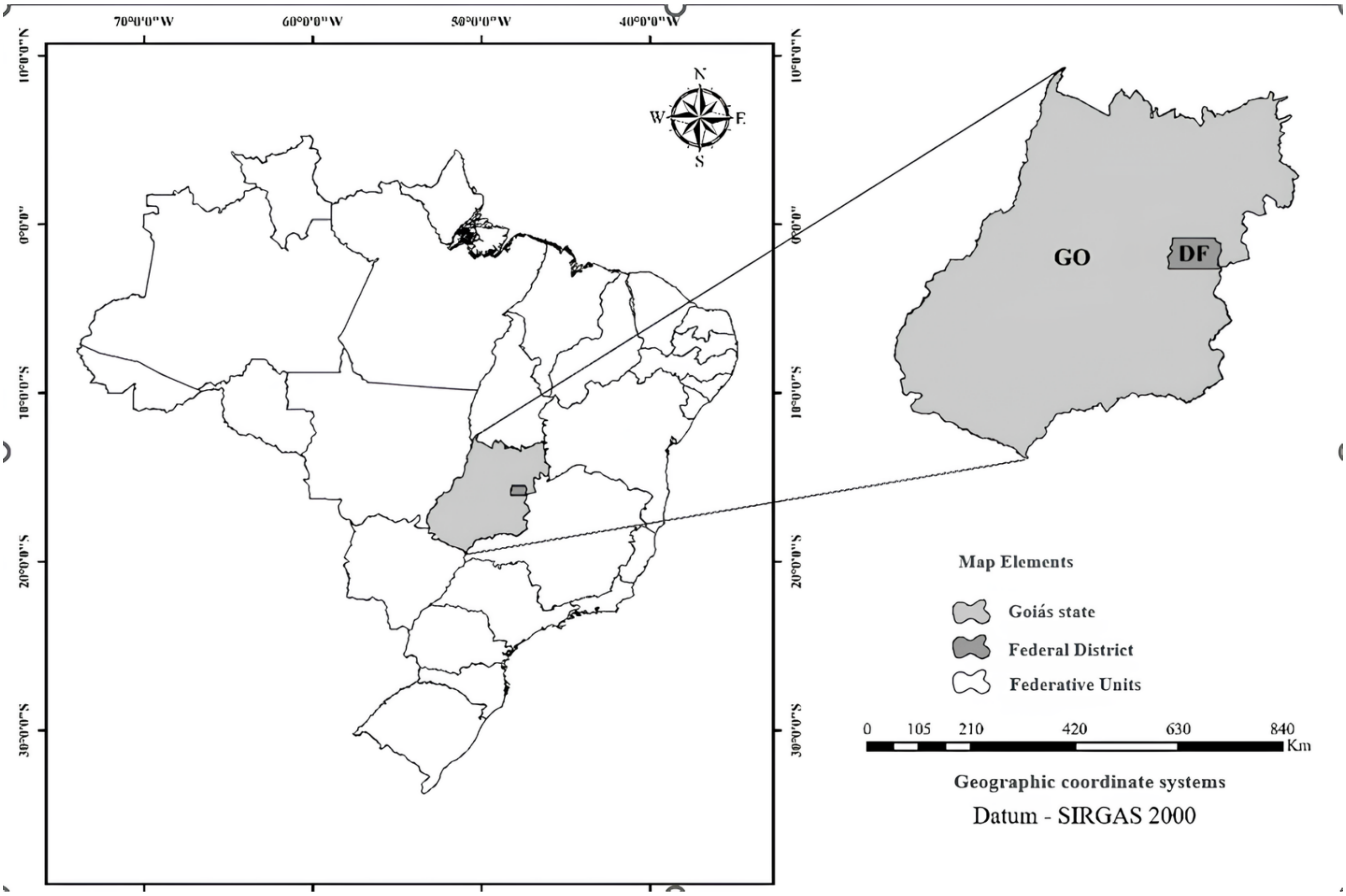

Figure 1. Geographic localization of the state of Goiás (GO) and the Federal District (DF), Brazil

(EMBRAPA, 2018), which corresponds to Typic Hapludox (United States, 2014). Vegetation is predominantly composed of savannas (Cerrado vegetation), with large anthropic areas occupied by agriculture and pasture (IMB, 2014).

The region has a hydrography system composed of rivers that feed three important hydrographic regions of the country: Tocantins River, Paraná River, and São Francisco River basins. It also has a dense drainage network formed by medium and large rivers (ANA, 2022).

The methodology of the study was characterized by continuous steps to achieve the proposed objectives. Each step and software used are shown in Figure 2. The methodology applied in each phase will be described in the following items.

Flow data were obtained through historical series from stream gauging stations implemented and managed by the Brazilian National Water Agency (ANA) in the state of Goiás and the Federal District, as well as from stations geographically close to the study areas, available on the HidroWeb portal of the Hydrological Information System, located at http://www.snirh.gov.br/hidroweb/serieshistoricas. Stations with data periods from 1948 to 2018 were selected, with at least 15 years of information, but not necessarily with a common baseline period for all stations (Wolff & Duarte, 2021). Stations were submitted to data absence analysis in the historical series, selecting those that contained at least 80% of annual data.

Thus, 77 stream gauging stations in Goiás, 25 in the Federal District, and 12 in the surroundings of these units were selected (Figure 3). Data from the surrounding stations were used to fill gaps in data from the other stations, maximizing the number

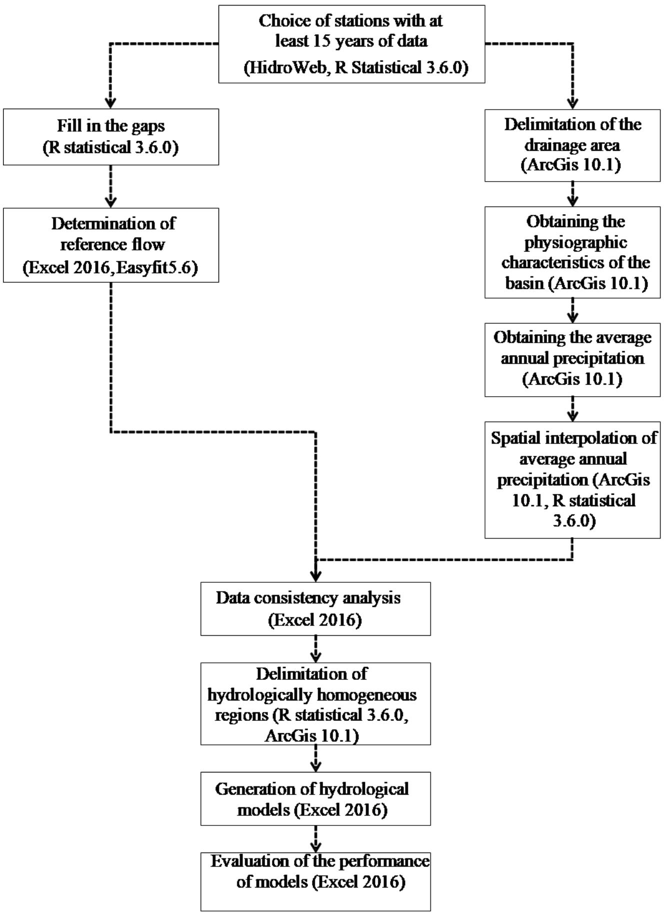
Choice of stations with at least 15 years of data (HidroWeb, R Statistical 3.6.0)
Fill in the gaps (R statistical 3.6.0)
Determination of reference flow (Excel 2016, Easyfit5.6)
Delimitation of the drainage area (ArcGis 10.1)
Obtaining the physiographic characteristics of the basin (ArcGis 10.1)
Obtaining the average annual precipitation (ArcGis 10.1)
Spatial interpolation of average annual precipitation (ArcGis 10.1, R statistical 3.6.0)
Data consistency analysis (Excel 2016)
Delimitation of hydrologically homogeneous regions (R statistical 3.6.0, ArcGis 10.1)
Generation of hydrological models (Excel 2016)
Evaluation of the performance of models (Excel 2016)

Figure 2. Methodology flowchart describing steps and software used in the study

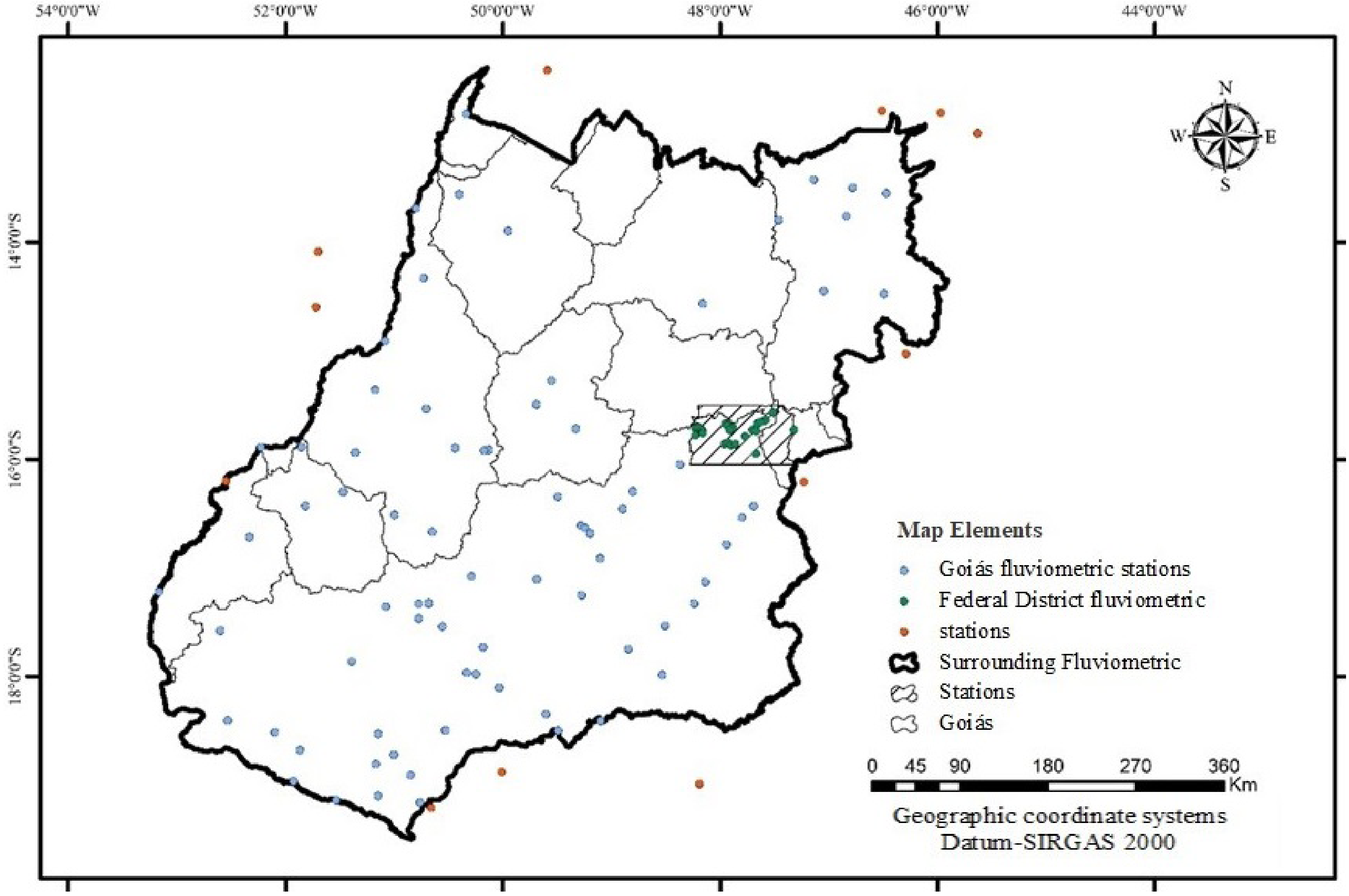

Figure 3. Stream gauging stations in Goiás state, the Federal District, and surrounding areas (Charles et al., 2022), Brazil

of data, thus reducing the effect of missing data. Therefore, gaps in the daily streamflow data were filled using the R 3.6.0 statistical software with the mtsdi package; this package uses software specifically adapted for climate data, based on the Maximized Hope algorithm to impute missing data in multiple time series (Junger & Leon, 2015).

Data from the following physiographic characteristics were used: watershed drainage area, average slope, drainage density, and length of the main thalweg. The data for these characteristics were obtained from the delimitation of the watersheds for each stream gauging station (Figure 4), using ArcHydro (Musselman & Aguilar, 2016), which provides a set of tools for hydrological analysis within the ArcGIS 10.1 software, based on the Digital Elevation Model (DEM).

The DEM used in the study was from Shuttle Radar Topography Mission (SRTM) satellite images, with a spatial resolution of 30 m; these images are available on the Google Earth Engine platform and reprojected to UTM 22S plane coordinate.

The largest drainage area measured was approximately 91,819 km^2 and is near the border with the state of Minas Gerais. The station measures the flow of the Paranaíba River, which belongs to the Paraná River basin. The smallest drainage area (33 km^2) is the Paraná River basin.

Monthly arithmetic means for each stream gauging station were calculated based on the historical series of daily flows using an Excel 2016 spreadsheet. Monthly average flow values were then arranged in decreasing order, using the EasyFit 3.6 software, to fit the probability of the flow duration curve (FDC) for each station (Costa & Fernandes, 2021); thus, the values of long-term mean flow ($\bar{Q}$) and reference flows with 90 or 95% duration over time (Q_{90} and Q_{95}) were obtained. The $\bar{Q}_{esp}$, used in the model as a dependent variable, was calculated using Eq. 1.

$$\bar{Q}_{esp} = \frac{\bar{Q}}{A} \tag{1}$$

where:

$\bar{Q}_{esp}$ - specific mean flow (m^3 s^{-1} km^{-2})
$\bar{Q}$ - long-term mean flow (m^3 s^{-1}); and,
A - drainage area (km^2).

The ANA's HidroWeb platform was also used to survey the rain gauge stations, as was done in obtaining the streamflow data. Similarly, Excel 2016 and R statistical 3.6.0 software were used to identify historical series with a period longer than 15 years. Subsequently, 112 stations in Goiás state, 26 in Federal District, and 29 nearby stations with data periods ranging from 1944 to 2018 were observed (Figure 5). The neighboring stations were used to maximize the amount of data for the study, using the same gap filling method explained previously (Tanim et al., 2021).

The interpolation of mean annual rainfall was performed using ordinary kriging, with an adjustment to the experimental Gaussian semivariogram, applying the Kriging tool in the ArcGis 10.1 software. The hydrological modeling was bases on the mean rainfall depths in the watersheds, obtained through the Zonal Statistical as a Table tool (Figure 6).

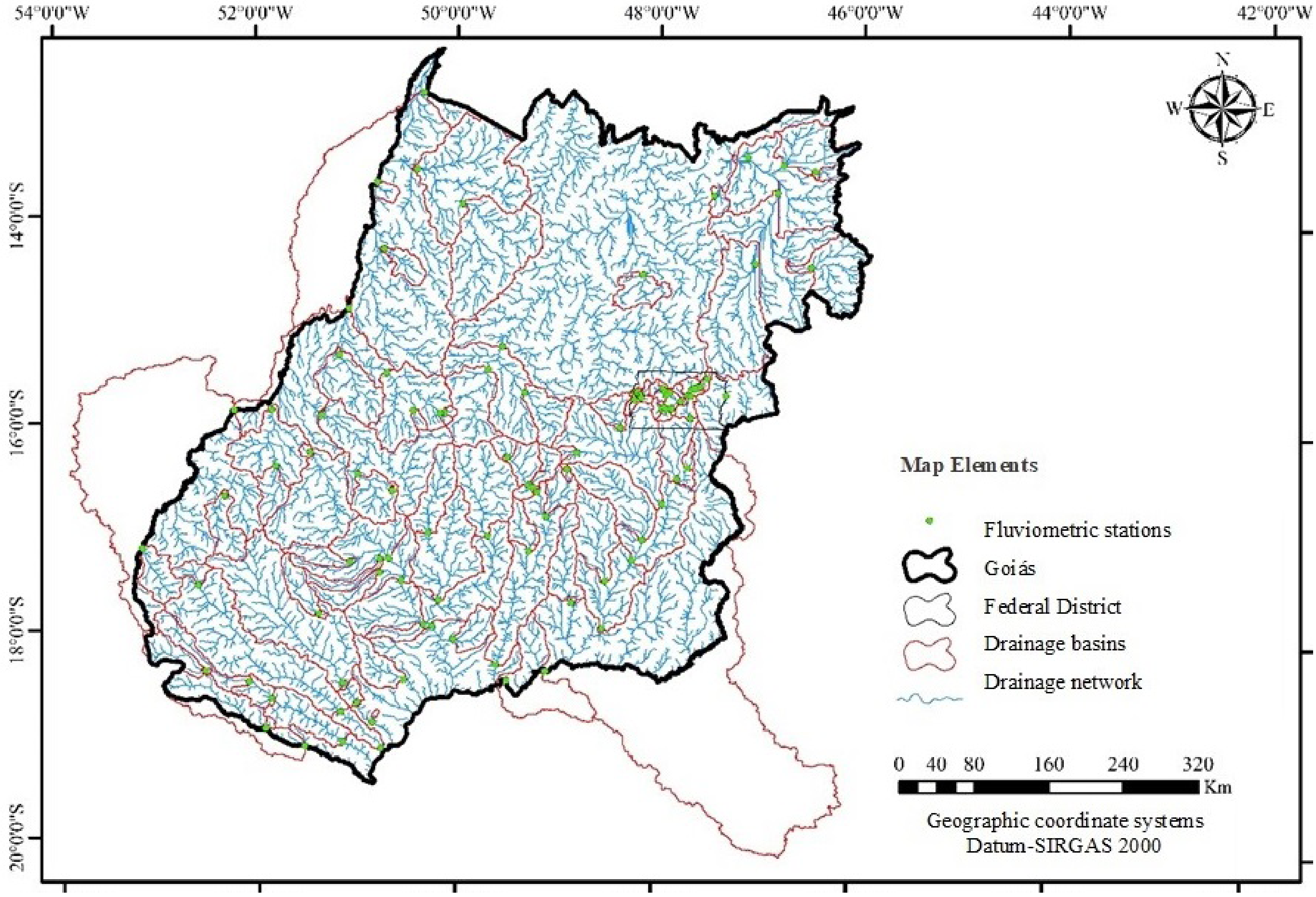

Figure 4. Watersheds located upstream of the stream gauging stations in Goiás and the Federal District, Brazil (Charles et al., 2022)

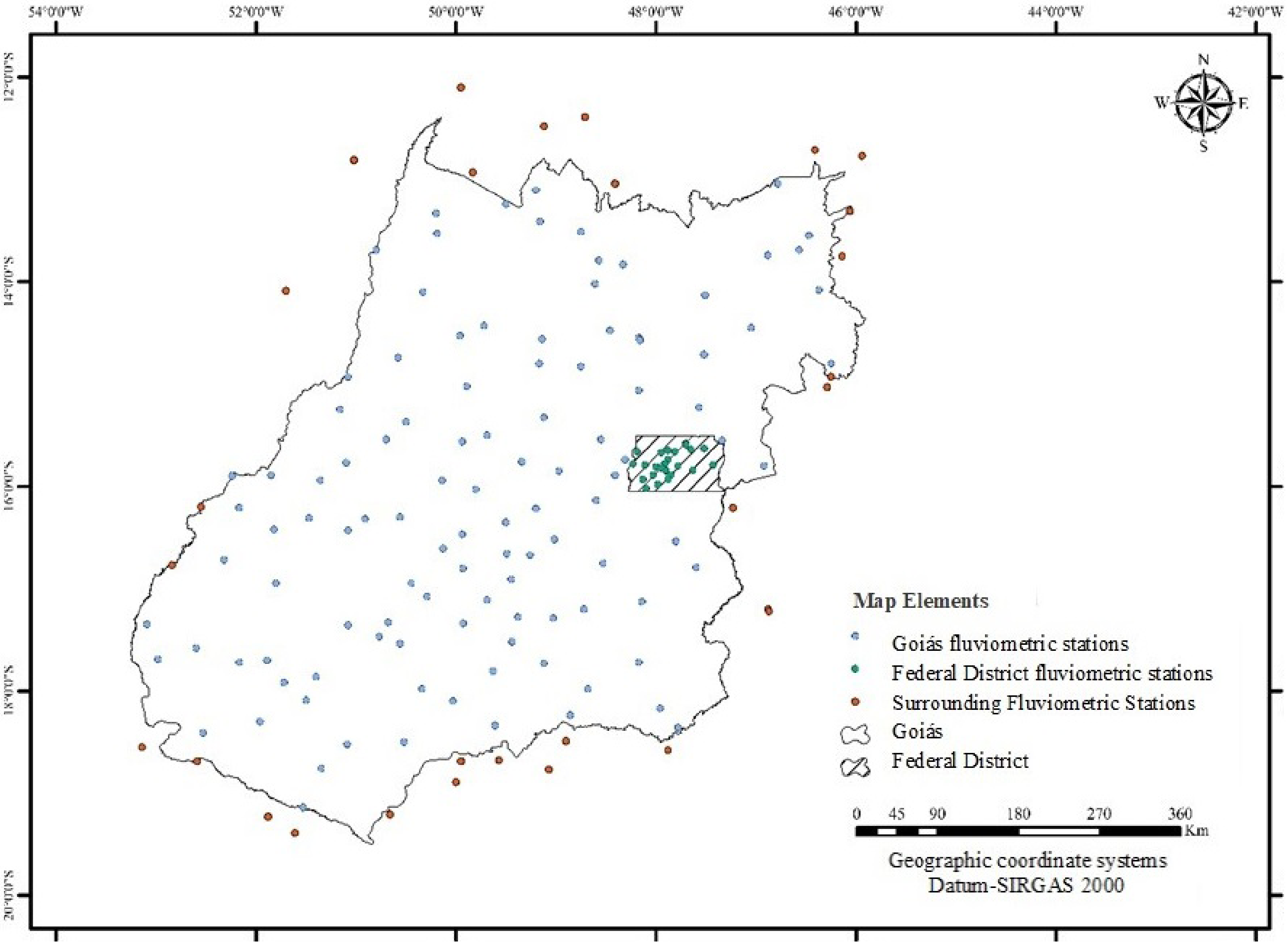

Figure 5. Rain gauge stations in Goiás state, the Federal District, and surrounding areas used in the study (Charles et al., 2022)

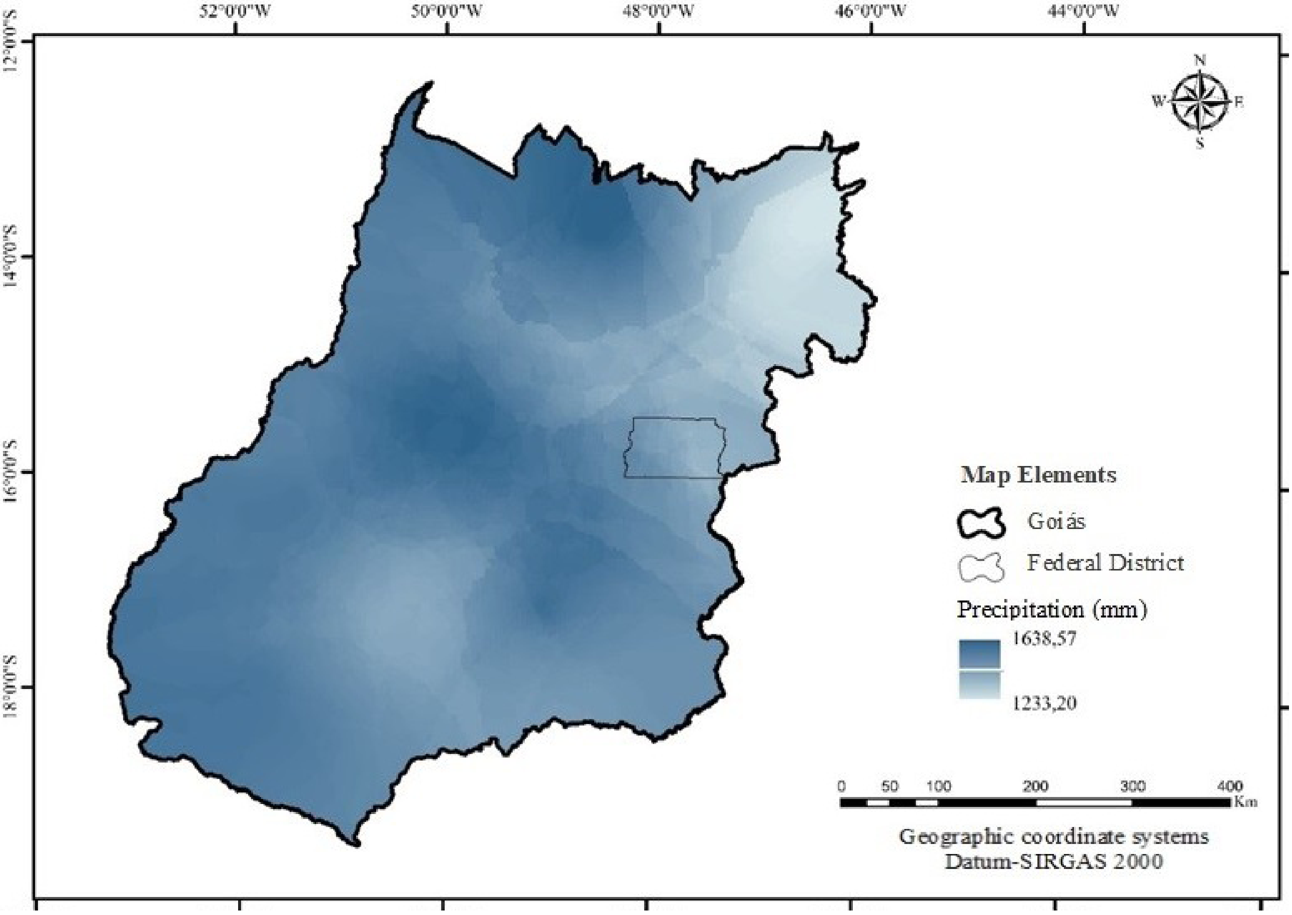

Figure 6. Spatial distribution of rainfall in Goiás state and the Federal District, Brazil

Cluster analysis was performed to delimit hydrologically homogeneous regions (Freitas et al., 2013) using the k-means algorithm in the R 3.6.0 statistical software. Cluster analysis is a technique for grouping similar components and identifying differences among elements. The following independent variables were considered: latitude and longitude of the watershed centroid, mean annual rainfall in the watershed, average slope, drainage density, and length of the main thalweg. The hierarchical method was used to conduct the cluster analysis, as it is flexible in the total number of groups, considering the researcher's knowledge of the region.

Once the optimal number of clusters was established, i.e., the number of hydrologically homogeneous regions in the areas of Goiás and the Federal District, the spatialization of the clusters was performed using the k-nearest neighbor (knn) algorithm in the R 3.6.0 statistical software. Considering that the independent variables have different units, they were standardized using Eq. 2, which consists of the normalization of the z-score, which resizes each variable in terms of the ratio between the standard deviation and its mean (Pandey & Jain, 2017).

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

where:

z - standard value;
x - observed variable;
μ - mean; and,
σ - standard deviation.

The knn algorithm considers similarity measures between the data to separate the groups into their corresponding class. Thus, the Euclidean distance of the centroids of the watersheds was considered, which corresponds to the straight-line distance between the points and was calculated using Eq. 3 (Pandey & Jain, 2017).

$$d_{i,j} = \sqrt{\sum \left(x_{in} - x_{jn}\right)^2} \tag{3}$$

where:

d - Euclidean distance between the centroids of the watersheds i and j (km);
x - observed variable (km); and,
n - number of independent variables.

Hydrologically homogeneous regions were delimited using cluster analysis, considering the independent variables in each watershed and the latitude and longitude of the centroid. A plot of the j and k relationship was generated to estimate the optimal number of classes, using the elbow algorithm (Kassambra, 2017; Setiawan et al., 2020), where the k value corresponding to the "elbow" is the number of classes (input value in the knn algorithm), i.e., the number of hydrologically homogeneous regions.

The regionalization of reference flows consisted of using multiple linear regression (Eq. 4) to identify the correlation between the dependent variable (specific mean flow; $\bar{Q}_{esp}$) and the independent variables (mean annual rainfall, average slope, drainage density, and main thalweg length) within each hydrologically homogeneous region. Analysis of variance (ANOVA) Student's t-test ($p \leq 0.05$) was then applied to determine the statistical significance of the coefficients used in each model.

$$\bar{Q}_{espr} = a + bP + cI + dDd + eL \tag{4}$$

where:

$\bar{Q}_{espr}$ - regionalized specific mean flow ($m^3\ s^{-1}\ km^{-2}$);
P - mean annual rainfall (mm per year);
I - average watershed slope (%);
Dd - drainage density ($km\ km^{-2}$);
L - length of the main thalweg (km); and,
a, b, c, d, e - model coefficients

After the hydrological models were defined, the mean flows were determined for each gauging station were using Eq. 5 (Tucci, 2012). The estimated reference flows were calculated to validate the hydrological models.

$$\bar{Q}_r = \bar{Q}_{espr} \cdot A \tag{5}$$

where:

$\bar{Q}_r$ - long-term regionalized mean flow ($m^3\ s^{-1}$);
$\bar{Q}_{espr}$ - regionalized specific mean flow ($m^3\ s^{-1}\ km^{-2}$); and,
A - area, m^2.

The regionalized Q_{90} and Q_{95} reference flows were determined based on linear coefficients obtained from the relationship between Q_{90} and Q_{95} with $\bar{Q}$ for each hydrologically homogeneous region, according to Eq. 6 and Eq. 7, respectively (Tucci, 2012).

$$Q_{90r} = a_{q90} \cdot \bar{Q}_r \tag{6}$$

$$Q_{95r} = a_{q95} \cdot \bar{Q}_r \tag{7}$$

where:

Q_{90r} - regionalized minimum flow for 90% of the time ($m^3\ s^{-1}$);
Q_{95r} - regionalized minimum flow for 95% of the time ($m^3\ s^{-1}$);
a_{q90} and a_{q95} - angular coefficients from the relationship between Q_{90} and Q_{95} with $\bar{Q}$, dimensionless; and,
$\bar{Q}_r$ - long-term regionalized mean flow ($m^3\ s^{-1}$).

The performance of the hydrological regionalization models proposed in this study was evaluated by comparing the data estimated by each model with those obtained by the stream gauging stations, using performance indices (Moriasi et al., 2015).

Initially, four performance indices recommended by Moriasi et al. (2015) were used: coefficient of determination (R^2) (Eq. 8), Nash-Sutcliffe efficiency test (NSE) (Eq. 10), percent bias (PBIAS) (Eq. 12), and Willmott index of agreement

(d) (Eq. 11). Additionally, the Pearson correlation coefficient (r) (Eq. 9) and the confidence index (c) (Eq. 13) were also determined (Camargo & Sentelhas. 1997).

$$R^2 = \left[\frac{\left[\sum_{i=1}^{n}(O_i - O)(S_i - S)\right]}{\left[\sqrt{\left(\sum_{i=1}^{n}(|O_i - O|)\right)^2}\right]\sqrt{\sqrt{\sum_{i=1}^{n}(|S_i - O|)}^{\,2}}} \right]^2 \qquad (8)$$

$$r = \frac{\sum_{i=1}^{n}(O_i - O)(S_i - S)}{\sqrt{\left(\sum_{i=1}^{n}(O_i - O)^2\right)\left(\sum_{i=1}^{n}(S_i - S)^2\right)}} \qquad (9)$$

$$NSE = 1 - \frac{\sum(O_i - S_i)^2}{\sum(O_i - O)^2} \qquad (10)$$

$$d = 1 - \frac{\sum(S_i - O_i)^2}{\sum\left(|S_i - O| + |O_i - O|\right)^2} \qquad (11)$$

$$PBIAS = \left[\frac{\sum_{i=1}^{n}(O_i - S_i)}{\sum_{i=1}^{n}O_i} \times 100 \right] \qquad (12)$$

$$c = r \times d \qquad (13)$$

where:

Oi - observed value ($m^3\ s^{-1}$);
Si - estimated value ($m^3\ s^{-1}$);
O - mean observed values ($m^3\ s^{-1}$);
S - mean estimated values ($m^3\ s^{-1}$); and,
n - final sum index.

Finally, the performance of the models was also evaluated using qualitative classifications related to each statistical index. These criteria are shown in Tables 1 and 2, according to the recommendations of Moriasi et al. (2015) and Camargo & Sentelhas (1997).

Results and Discussion

The annual rainfall depths in the studied watersheds varied spatially between 1,233.20 and 1,638.57 mm, with a mean of 1,480.16 mm. This result is corroborated by the study of Cardoso et al. (2014), who reported a mean annual rainfall of 1,485.3 mm for the region.

The highest mean annual rainfall depth was found in the north (Althoff et al., 2021) and southeast regions of Goiás state, as shown in the Atlas of the State of Goiás (IMB, 2014). The lowest rainfall volumes occurred in the northeast of the state, which can be explained by the influence of the Atlantic tropical air mass; this air mass carries moisture but loses it due to the occurrence of orographic rainfall on the coast of the Northeast Region of Brazil (Freitas et al., 2013).

The data showed the number of stations covered by the probability density functions (pdf_s) that best fit the 102 generated flow duration curves (FDCs). These FDCs were used to obtain reference flows ($\bar{Q}$, Qmax, Qmin, Q_{90}, and Q_{95}) and showed better fits for 34 distributions when performing Kolmogorov-Smirnov, Anderson Darling, and Chi-Square goodness-of-fit tests (Costa & Fernandes, 2021) (Table 3).

The Fatigue Life (3P) distribution had the highest number of FDCs with the best fit to the goodness-of-fit tests, followed by the Johnson SB distribution. Wolff & Duarte (2021) used only the log-normal distribution for all data from 74 gauging stations, but they regionalized a significantly smaller Brazilian state (Santa Catarina).

The hydrologically homogeneous regions were delimited based on cluster analysis using independent variables from each watershed, as well as the latitude and longitude of the centroid; the optimal number of classes (k) correlated to the "elbow" was seven classes for the study region, i.e., seven hydrologically homogeneous regions (clusters) were delimited (Figure 7).

According to the methodology used to group the regions, the state of Goiás has seven hydrologically similar regions, two of them in the Federal District (regions 4 and 7). The exchange of hydrological information should only occur within each hydrologically homogeneous region (Beskow et al., 2016), therefore, seven different hydrological regionalization models were generated.

Table 1. Qualitative classification of model performance adapted by Moriasi et al. (2015) and Camargo & Sentelhas (1997)

Statistical indices	Model's Performance			
	Very good	Good	Satisfactory	Unsatisfactory
R^2	>0.85	$0.75< R^2\leq 0.85$	$0.60< R^2\leq 0.75$	≤0.60
r	>0.85	0.75<r≤ 0.85	0.60<r≤ 0.75	≤0.60
NSE	>0.80	0.75<NSE≤ 0.85	0.50<NSE≤ 0.70	≤0.50
PBIAS (%)	<5	5<PBIAS≤ 10	10<PBIAS≤15	>15
d	>0.90	0.85<d≤ 0.90	0.75<d≤0.85	≤0.75

R^2 - Coefficient of determination; r - Pearson correlation coefficient; NSE - Nash-Sutcliffe efficiency test; PBIAS - Percent bias; d - Willmott index of agreement

Table 2. Qualitative classification of model performance using the confidence index (c), as described in Camargo & Sentelhas (1997)

Performance	Excellent	Very good	Good	Fair	Poor	Very poor
C	c>0.85	0.75<c≤0.85	0.65<c≤0.75	0.50<c≤0.65	0.40<c≤0.50	c<0.40

C = confidence index

Table 3. Number of stream gauging stations most suitable for each probability density function (pdf) related to the flow duration curves obtained by three goodness-of-fit tests.

Probability density function	KS	AD	CS	Probability density function	KS	AD	QS
Beta	3	4	3	Hypersecant	0	0	1
Burr	3	5	1	Inv. Gaussian	0	1	2
Burr (4P)	7	9	7	Inv. Gaussian (3P)	7	10	9
Chi-Square (2P)	1	0	0	Johnson SB	21	11	5
Dagum	0	0	3	Kumaraswamy	1	3	1
Dagum (4P)	3	4	0	Log-Gamma	1	1	0
Exponential (2P)	1	0	2	Log-Logistic	0	0	1
Fatigue Life	1	6	6	Log-Logistic (3P)	1	1	2
Fatigue Life (3P)	17	25	15	Log-Normal	3	0	4
Frechet	0	0	3	Log-Normal (3P)	4	2	4
Frechet (3P)	0	1	1	Log-Pearson 3	1	1	1
Gamma	2	0	1	Pearson 5	1	3	6
Gamma (3P)	0	0	5	Pearson 6	1	2	4
Gen. Extreme Value	1	0	1	Pearson 6 (4P)	1	3	2
Gen. Gamma	0	0	4	Pert	0	1	0
Gen. Gamma (4P)	4	6	6	Weibull	3	0	2
Gen. Pareto	14	1	0	Weibull (3P)	0	2	0

KS - Kolmogorov Smirnov; AD - Anderson Darling; CS - Chi-square; Gen - Generalized; Inv - Inverse; P - Parameters

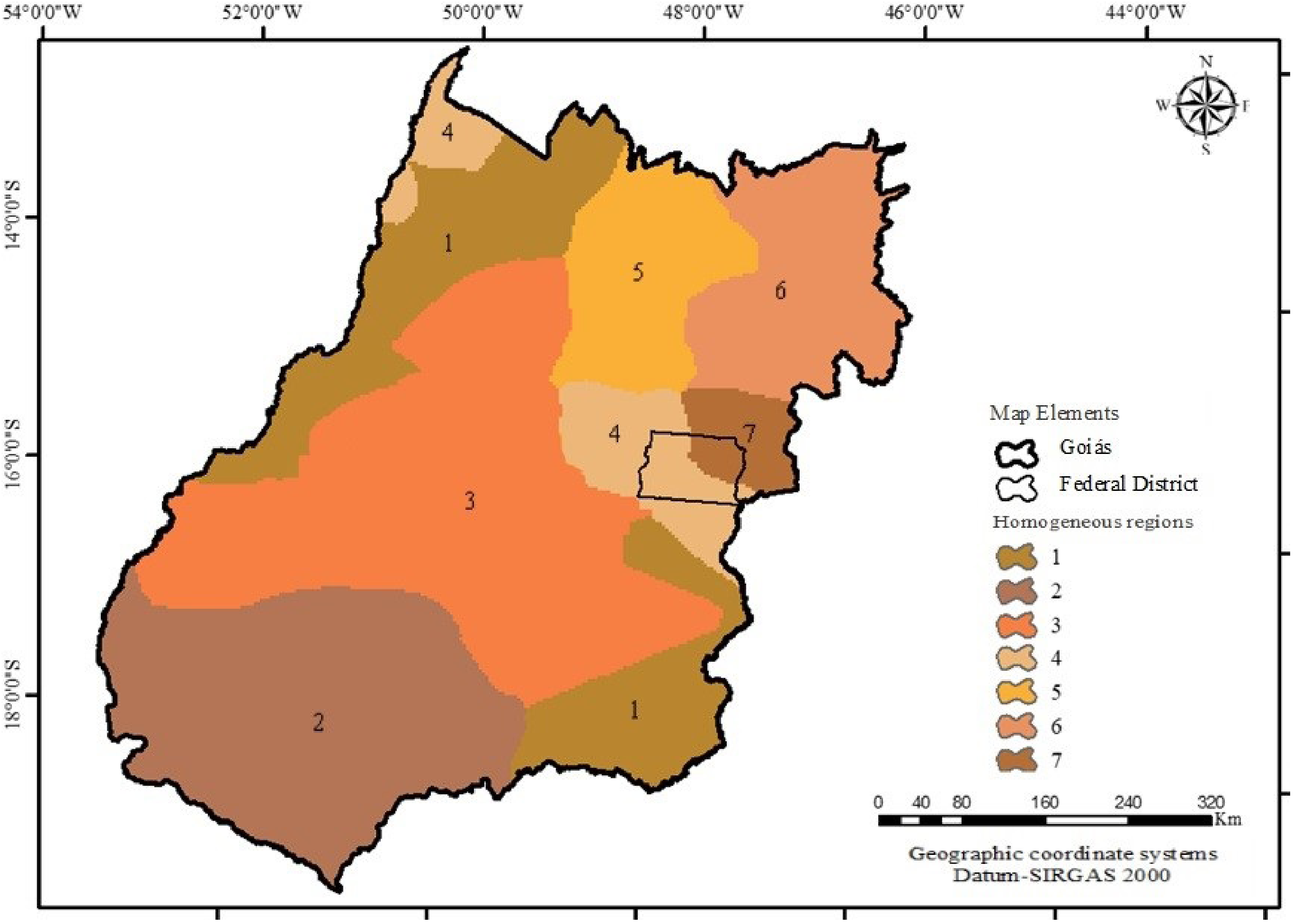

Figure 7. Hydrologically homogeneous regions in Goiás state and the Federal District, Brazil, by cluster analysis

The number of homogeneous regions identified (seven regions) (Figure 7) denotes a dependence on the variation in regional climate and morphometric data used in the cluster analysis. Wolff et al. (2014) and Nascimento at al. (2021) found 5 and 3 homogeneous regions when studying the states of São Paulo and Paraná, Brazil, respectively. Considering the larger territorial extension of the area evaluated in the present study, the number of seven hydrologically homogeneous regions can be considered adequate (Pessoa et al., 2021).

The correlation of the digital elevation model (DEM) with the groupings of homogeneous regions (Figure 7) showed a correlation between the elevation of the terrain and some regions of hydrological similarity. Thus, region 1 (southeast region) showed a grouping of watersheds in flat areas, which have the lowest altitudes in the state of Goiás. However, region 2 (south region) did not show homogeneity in terms of altitude for the grouping of watersheds, which may have contributed to the low performance of the model in this region when compared to the other regions.

Seven regionalization models were generated and their respective indices are shown in Table 4.

All variables were tested in the seven models, with different combinations; however, not all independent variables were significant for all models by the Student's t-test ($p \leq 0.05$) (Musselman & Aguilar, 2016). Only "P" and "I" were explanatory for region 1, while "L" had no significant effect on the model for regions 3 and 6.

$\bar{Q}$ can be calculated by multiplying the regionalized $\bar{q}_{esp}$ by the watershed area (Eq. 9). The regionalized Q_{90} and Q_{95} were calculated using Eq. 6 and Eq. 7, respectively, based on angular coefficients for each homogeneous region, which are shown in Table 5.

Model performance evaluation criteria provide quantitative model rankings, with qualitative thresholds corresponding to each statistical index (Lelis et al., 2020). The qualitative classifications for the R^2, r, NSE, d, and PBIAS indices can be very good, good, satisfactory, and unsatisfactory, whereas the classifications for the confidence index (c) can be excellent, very good, good, fair, poor, and very poor.

According to the statistical indices (R^2, NSE, PBIAS, d, r, and c) for the hydrologically homogeneous regions 1, 3, 4, 5, and 7, the model performance classification ranged from good to excellent (Table 6).

The model's fit to observed data for Q_{90}, Q_{95}, and $\bar{Q}$ in the hydrologically homogeneous region 2 (south region) was classified, in general, as "very good" based on d, r, and c indices and as "good" based on the R^2 index. This indicates that the model had a good fit, with a high degree of dependence, precision, and accuracy. However, the model's fit was classified as "satisfactory" based on the NSE index, denoting some degree of agreement, but not ideal. The PBIAS index, which measured the average tendency for simulated flows to be higher or lower than observed flows, classified the model as "satisfactory" for Q_{90} and Q_{95} and "unsatisfactory" for $\bar{Q}$, with a percent relative error of 15.34% (close to the 15% threshold); this denotes that the model did not perform well in simulating mean magnitudes (Table 6).

The model for hydrologically homogeneous region 2 had the worst performance compared to the other regions, which may be attributed to the small number of gauging stations in this region relative to the area size. This leads to greater uncertainties in model generation, requiring greater caution in its use.

Region 3 was concentrated in the middle of the state of Goiás; it had the largest area among all hydrologically homogeneous regions and showed heterogeneity in terms of altitude. However, it had the highest number of data collection stations, which may have contributed to a better model performance. Similar heterogeneity was observed in region 6 (northeast region). However, this region had more low-altitude areas, which may have contributed to the model performance, as the classification was "good" for $\bar{Q}$ based on R^2 and NSE indices.

Table 5. Reduction slopes of long-term mean flows ($\bar{Q}$) for calculating regionalized Q_{90} and Q_{95}

Homogeneous regions	a_{q90}	a_{q95}
1	0.4850	0.3992
2	0.3135	0.2887
3	0.2511	0.1972
4	0.5501	0.5145
5	0.5714	0.5150
6	0.4306	0.3744
7	0.3627	0.2927

a_{q90}, a_{q95} - Angular coefficients from the relationship between Q_{90} and Q_{95} with $\bar{Q}$.

In region 6, the model's fit was classified as "satisfactory" for Q_{90} and Q_{95} based on R^2 and NSE indices (Table 6), denoting some degree of dependence and agreement, but not ideal. However, the PBIAS index classified the model's as "very good" for Q_{90}, Q_{95}, and $\bar{Q}$, indicating a good performance in simulating mean magnitudes. The model's was classified, in general, as "good" (Q_{90} and Q_{95}) and "very good" ($\bar{Q}$) based on d, r, and c indices, indicating a good degree of dependence, precision, and accuracy.

In general, rainfall was evenly distributed throughout the study area, with the lowest volumes observed in region 6, where the model showed good performance based on PBIAS, d, r, and c indices, but only "satisfactory" for Q_{90} and Q_{95} based on R^2 and NSE indices. This denotes a similarity between rainfall volumes and hydrologically homogeneous regions, which may explain the overall good performance of the models.

Further studies using seasonal flows and considering periods of lower and higher rainfall depths could enable the establishment of less conservative criteria for granting the water use rights. This would allow the use of larger volumes of water resources during rainy periods (Wolff & Duarte, 2021). A more detailed analysis of the effect of rainfall on the models proposed in the present study requires further research on the seasonality of rainfall in the region, which has two well-defined seasons. Seasonality is an essential factor for regional water availability and can be more rational when considered in granting water resources (Beskow et al., 2016).

The performance of these models is not only related to the availability of gauging station data, but also to the physiographic and climatic characteristics used as

Table 4. Angular coefficients of the independent variables of the regionalization equations for the calculation of the specific regionalized mean flow for each hydrologically homogeneous region

Homogeneous regions	a	b	c	d	e
1	0.010914645	-3.17622×10^{-6}	0.002575523	-	-
2	-1.013575685	0.0002731610	0.010471610	1.450755963	-0.000320544
3	-0.071062622	1.89542×10^{-5}	0.003220494	0.115966774	-
4	0.073822880	-5.12397×10^{-5}	-0.002570331	0.072357388	-0.000119287
5	-0.068276192	9.64324×10^{-5}	0.005572445	0.196064381	-1.76075×10^{-5}
6	-0.602081666	0.0002755150	0.026696695	0.432849037	-
7	-0.147393511	4.24326×10^{-5}	0.005936218	0.220347388	4.17005×10^{-6}

a, b, c, d, e - Model coefficients

Table 6. Performance indices and performance classification of the models

Regions	Flows	R^2	NSE	PBIAS (%)	d	r	c
1	Q_{95}	0.9291	0.9060	-0.7718	0.9709	0.9639	0.9358
	Q_{90}	0.9074	0.8822	-1.4234	0.9625	0.9526	0.9169
	$\bar{Q}$	0.9897	0.9826	5.6810	0.9955	0.9948	0.9903
	Q_{95}	Very good	Very good	Very good	Very good	Very good	Excellent
	Q_{90}	Very good	Very good	Very good	Very good	Very good	Excellent
	$\bar{Q}$	Very good	Very good	Good	Very good	Very good	Excellent
2	Q_{95}	0.7617	0.6256	-12.8858	0.9196	0.8728	0.8026
	Q_{90}	0.7787	0.6447	-12.1504	0.9246	0.8824	0.8159
	$\bar{Q}$	0.8085	0.6915	-15.5361	0.9342	0.8992	0.8400
	Q_{95}	Good	Satisfactory	Satisfactory	Very good	Very good	Excellent
	Q_{90}	Good	Satisfactory	Satisfactory	Very good	Very good	Excellent
	$\bar{Q}$	Good	Satisfactory	Satisfactory	Very good	Very good	Excellent
3	Q_{95}	0.8951	0.8737	4.5251	0.9699	0.9461	0.9176
	Q_{90}	0.9134	0.8885	5.1350	0.9742	0.9557	0.9310
	$\bar{Q}$	0.9860	0.9824	-2.9385	0.9958	0.9930	0.9888
	Q_{95}	Very good	Very good	Very good	Very good	Very good	Excellent
	Q_{90}	Very good	Very good	Good	Very good	Very good	Excellent
	$\bar{Q}$	Very good	Very good	Very good	Very good	Very good	Excellent
4	Q_{95}	0.9918	0.9872	-4.7662	0.9966	0.9959	0.9925
	Q_{90}	0.9934	0.9901	-2.2589	0.9974	0.9967	0.9940
	$\bar{Q}$	0.9991	0.9971	2.4240	0.9993	0.9995	0.9995
	Q_{95}	Very good	Very good	Very good	Very good	Very good	Excellent
	Q_{90}	Very good	Very good	Very good	Very good	Very good	Excellent
	$\bar{Q}$	Very good	Very good	Very good	Very good	Very good	Excellent
5	Q_{95}	0.8605	0.8492	3.2966	0.9615	0.9276	0.8919
	Q_{90}	0.8742	0.8632	3.0980	0.9655	0.9350	0.9027
	$\bar{Q}$	0.9686	0.9685	-0.1463	0.9918	0.9995	0.9842
	Q_{95}	Very good	Very good	Very good	Very good	Very good	Excellent
	Q_{90}	Very good	Very good	Very good	Very good	Very good	Excellent
	$\bar{Q}$	Very good	Very good	Very good	Very good	Very good	Excellent
6	Q_{95}	0.6020	0.5733	3.4750	0.8726	0.7922	0.6913
	Q_{90}	0.6276	0.6012	3.0487	0.8832	0.9350	0.8258
	$\bar{Q}$	0.7934	0.7774	-3.9958	0.9418	0.8907	0.9842
	Q_{95}	Satisfactory	Satisfactory	Very Good	Good	Good	Good
	Q_{90}	Satisfactory	Satisfactory	Very Good	Good	Very Good	Very Good
	$\bar{Q}$	Good	Good	Very Good	Very Good	Very Good	Excellent
7	Q_{95}	0.8881	0.8776	-0.2930	0.9699	0.9424	0.9140
	Q_{90}	0.9114	0.9004	0.5831	0.9760	0.9547	0.9317
	$\bar{Q}$	0.9649	0.9557	2.2013	0.9896	0.9823	0.9842
	Q_{95}	Very good	Very good	Very good	Very good	Very good	Excellent
	Q_{90}	Very good	Very good	Very good	Very good	Very good	Excellent
	$\bar{Q}$	Very good	Very good	Very good	Very good	Very good	Excellent

R^2 - Coefficient of determination; NSE - Nash-Sutclife coefficient; PBIAS - percent bias; d - Willmott index; r - Pearson correlation coefficient; c - Confidence index

independent variables. Altitude (a characteristic not used) may have been an important factor in some regions. However, the state of Goiás and the Federal District do not have large variations in altitude, thus lacking significant natural barriers to air masses passing through the region (Cardoso et al., 2014).

Although hydrological models generally perform well, they often do not perform satisfactorily in all regions, denoting that their application for management purposes should be restricted in some regions (Du et al., 2020). However, hydrological models are important tools for water resources management, especially in developing countries such as Brazil, where the long-term data availability throughout the entire territory is low.

Conclusions

1. Cluster analysis identified seven hydrologically homogeneous regions in the state of Goiás, two of them belonging to the Brazilian Federal District.

2. Multiple regression resulted in the development of seven hydrological regionalization models. Models for regions 1, 3, 4, 5, and 7 showed better performance.

Contribution of authors: Thaís da S. Charles: Article planning, literature review, data collection and analysis, writing the manuscript, drawing conclusions and final review. Tárcio R. Lopes: Writing the manuscript and final review. Sérgio N. Duarte: Article planning, literature review, writing the manuscript, drawing conclusions, supervision and administration. Hugo de C. Ricardo: Writing the manuscript and final review. Adriano B. Pacheco: Writing the manuscript and final review. Fernando C. Mendonça: Final review and supervision.

Supplementary documents: There are no supplementary sources.

Conflict of interest: The authors declare no conflict of interest.

Financing statement: There was no funding for this research.

Acknowledgments: The authors thank the LABHA laboratory (Laboratório de Hidrologia Aplicada e Meio Ambiente) of the Department of Biosystems Engineering at the Luiz de Queiroz College of Agriculture of the University of São Paulo, Brazil. This study was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

Literature Cited

Althoff, D.; Ribeiro, R. B.; Rodrigues, L. N. Gauging and ungauged: Regionalization of flow indices at grid level. Journal of Hydrologic Engineering, v.26, e04021008, 2021. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002067

ANA. Conjuntura dos recursos hídricos no Brasil 2021: Relatório pleno, Agência Nacional de Água e Saneamento Básico. Brasília, 2022. 132p.

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

Camargo, A. P.; Sentelhas, P. C. Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo, Brasil. Revista Brasileira de Agrometeorologia, v.5, p.89-97, 1997. https://doi.org/10.1590/S1415-43662007000600006

Cardoso, M. R. D.; Marcuzzo, F. F. N.; Barros, J. R. Climatic classification of Köppen-Geiger for the state of Goias and Federal District. Acta Geográfica, v.8, p.40-55, 2014. https://doi.org/10.18227/2177-4307.acta.v8i16.1384

Cassalho, F.; Beskow, S.; Mello, C. R. de; Moura, M. M. de; Oliveira, L. F. de; Aguiar, M. S. de. Artificial intelligence for identifying hydrologically homogeneous regions: A state of the art regional flood frequency analysis. Hydrological Processes, v.33, p.1101-1116, 2019. https://doi.org/10.1002/hyp.13388

Charles, T. C.; Lopes, R. L.; Duarte, S. N.; Nascimento, J. G.; Ricardo, H. C.; Pacheco, A. B. Estimating average annual rainfall by ordinary kriging and TRMM precipitation products in midwestern Brazil. Journal of South American Earth Sciences, v.118, e103937, 2022. https://doi.org/10.1016/j.jsames.2022.103937

Costa, V.; Fernandes, W. Regional modeling of long-term and annual flow duration curves: Reliability for information transfer with evolutionary polynomial regression. Journal of Hydrologic Engineering, v.26, p.1-12, 2021. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002051

Du, T. L. T.; Lee, H.; Bui, D. D.; Arheimer, B.; Li, H.; Olsson, J.; Darby, S. E.; Sheffield, J.; Kim, D.; Hwang, E. Streamflow prediction in "geopolitically ungauged" basins using satellite observations and regionalization at subcontinental scale. Journal of Hydrology, v.588, e125016, 2020. https://doi.org/10.1029/2021WR031191

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Sistema Brasileiro de Classificação de Solos, 5.ed. Embrapa, Rio de Janeiro, Brazil, 2018, 356p.

Freitas, J. C.; Andrade, A. R. S.; Braga, C. C.; Godoy Neto, A. H.; Almeida, T. F. Análise de agrupamentos na identificação de regiões homogêneas de índices climáticos no estado da Paraíba, PB – Brasil. Revista Brasileira de Geografia Física, v.6, p.732-748, 2013. https://periodicos.ufpe.br/revistas/rbgfe/article/viewFile/233065/26985

Gomes, E. P.; Blanco, C. J. C.; Pessoa, F. C. L. Regionalization of precipitation with determination of homogeneous regions via fuzzy c-means. Brazilian Journal of Water Resources and Irrigation Management, v.23, e51, 2018. https://doi.org/10.1590/2318-0331.231820180079

IMB - Instituto Mauro Borges. Atlas do Estado de Goiás. Goiânia: Brazil, 2014, 100p.

Junger, W.; De Leon, P. Imputation missing data in time séries for air pollutans. Atmosferic Enviroment, v.35, p.1-16, 2015. https://doi.org/10.1016/j.atmosenv.2014.11.049

Kassambara, A. Practical Guide to Cluster Analysis in R, 1.ed. USA: STHDA, 2017. 189p. https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/book-preview/clustering_en_preview.pdf

Lelis, L. C. S.; Nascimento, J. C.; Duarte, S. N.; Pacheco, A. B.; Bosquilia, R. W. D.; Wolff, W. Assessment of hydrological regionalization methodologies for the upper Jaguari River basin. Journal of South American Earth Sciences, v.97, e102402, 2020. https://doi.org/10.1016/j.jsames.2019.102402

Macedo, P. M.; Schultz, N.; Oliveira, P. T. S.; Pinto, M. F.; Conforto, B. A. A. F.; Carvalho, D.F. Developing na automatic collector of runoff for studies using rainfall simulators. Brazilian Journal of Agricultural and Enviromental Engineering, v.27, p.828-836, 2023. http://dx.doi.org/10.1590/1807-1929/agriambi.v27n10p828-836

Manke, E. B.; Teixeira-Gandra, C. F. A.; Damé, R. de C., F.; Nunes, A. B.; Chagas-Neta, M. C. C.; Karsburg, R. M. Seazonal intensity-duration-frequency relationships for Pelotas, Rio Grande do Sul, Brazil. Brazilian Journal of Agricultural and Environmental Engineering, v.26, p.85-90, 2022. http://dx.doi.org/10.1590/1807-1929/agriambi.v27n10p828-836

Moriasi, D. N.; Gitau, M.; Pai, N.; Daggupati, P. Hydrologic and water quality models: performance measures and evaluation criteria. American Society of Agricultural and Biological Engineers, v.58, p.1763–1785, 2015. https://doi.org/10.13031/trans.58.10715

Musselman, Z. A.; Aguilar A. Investigation of drainage basin geometry near an anomolously straight reach of the Big Black River, Mississipi, USA. Physical Geography, v.37, p.168-187, 2016. https://doi.org/10.1080/02723646.2016.1163482

Nascimento, J. G.; Althoff, D.; Bazame, H. C.; Neale, C. M. U.; Duarte, S. N.; Ruhoff, A. L.; Gonçalves, I. Z. Evaluating the latest IMERG products in a subtropical climate: The case of Paraná State, Brazil. Remote Sensing, v.13, p.1-18, 2021. https://doi.org/10.3390/rs13050906

Pandey, A.; Jain, A. Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security, v.9, p.36-42, 2017. https://doi.org/10.5815/IJCNIS.2017.11.04

Pessoa, F. C. L.; Blanco, C. J. C.; Gomes, E. P. Regionalization of flow duration curves in the Amazon with the definition of homogeneous regions via fuzzy C-means. Engineering Sciences, v.93, e20190747, 2021. https://doi.org/10.1590/0001-3765202120190747

Rohli, R. V.; Joyner, T. A.; Reynolds, S. J.; Shaw, C.; Vázquez, J. Globally extended Köppen-Geiger climate classification and temporal shifts in terrestrial climatic types. Physical Geography, v.36, p.141-157, 2015. https://doi.org/10.1080/02723646.2015.1016382

Setiawan, O.; Sartohadi, J.; Hadi, M. P.; Mardiatno, D. Infiltration characterization using principal component analisys and K-means cluster analysis on quaternary volcanic landscape at southern flank of Rinjani Volcano, Lombok Island, Indonesia. Physical Geography, v.41, p.217-237, 2020. https://doi.org/10.1080/02723646.2019.1620526

Tanim, A. H.; Mullick, R. A.; Sikdar, S. Evaluation of spatial rainfall products in sparsely region using copula uncertainty modeling with triple collocation. Journal of Hydrologic Engineering, v.26, e04021004, 2021. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002071

Tucci, C. E. M. Hidrologia: ciência e aplicação. Porto Alegre, Ed. ABRH/UFRGS, 2012. 944p.

United States. Soil Survey Staff. Keys to Soil Taxonomy (12th ed.) USDA NRCS. 2014. Available at: <http://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/survey/>. Accessed: Feb 16, 2020.

Wolff, W.; Duarte, S. N.; Mingoti, R. Nova metodologia de regionalização de vazões, estudo de caso para o Estado de São Paulo. Revista Brasileira de Recursos Hídricos, v.19, p.21-33, 2014.

Wolff, W.; Duarte, S. N. Toward geostatistical unbiased predictions of flow durations curves at ungauged basins. Advances in Water Resources, v.152, e103915, 2021. https://doi.org/10.1016/j.advwatres.2021.103915.