DOI: 10.5433/1679-0359.2020v41n3p829

Regionalization of reference streamflows for the Araguaia River basin in Brazil

Regionalização de vazões de referência para bacias do rio Araguaia, Brasil

Marco Antonio Vieira Morais[1*]; Marcelo Ribeiro Viola[2]; Carlos Rogério de Mello[2]; Jéssica Assaid Martins Rodrigues[1]; Vinícius Augusto de Oliveira[3]

Highlights:
Statistical modelling for flow estimation.
Cross-validation for selection of hydrological models.
Probability distribution 4-parameter Kappa and 5-parameter Wakeby.

Abstract

Hydraulic projects and water management require reliable hydrological data. The Araguaia-Tocantins River basin, in addition to agricultural use, has great potential for hydroelectric exploitation. However, the streamflow monitoring network in the Araguaia River basin is composed of only a few stations, resulting in a lack of hydrological data. The regionalization of the reference streamflows is a technique that can help circumvent this lack of data, enabling the estimation of streamflows from easily obtainable explanatory variables. In this context, the objective of this study was to develop regional functions for the maximum streamflow (Qmax) applicable to different Return Periods (RP), the long-term mean streamflow (Qmlt) and the 95% streamflow permanence (Q_{95}) of the upper and middle Araguaia River sub-basins. The dimensionless streamflow methodology was adopted with the drainage area as an explanatory variable. The tested regressive models were the linear, potential and quotient models. Leave-one-out cross-validation was used to assess the quality of the regional models. Ten statistical distributions of 2 to 5 parameters were used. (i) Satisfactory results were obtained for all reference streamflows. (ii) The cross-validation technique proved to be essential for the selection of the most robust model. (iii) The quotient model was shown to be superior to the potential linear model in most cases.
Key words: Cerrado. Statistical Hydrology. Hydrological Modelling.

Resumo

Projetos hidráulicos e a gestão da água demandam dados hidrológicos confiáveis. A bacia hidrográfica do rio Araguaia-Tocantins, além do uso agrícola, apresenta grande potencial para exploração hidroelétrica. No entanto, a rede de monitoramento fluviométrico na bacia hidrográfica do rio Araguaia apresenta densidade reduzida de estações, o que implica na falta de dados hidrológicos. A regionalização de vazões de referência é uma técnica que pode ajudar a contornar essa insuficiência de dados, propiciando

[1] Discentes de Doutorado, Programa de Pós-Graduação em Recursos Hídricos em Sistemas Agrícolas, Universidade Federal de Lavras, PPGRHASA/UFLA, Lavras, MG, Brasil. E-mail: marco.morais@bag.ifmt.edu.br; je_assaid@yahoo.com.br
[2] Profs. Drs., PPGRHASA/UFLA, Lavras, MG, Brasil. E-mail: marcelo.viola@deg.ufla.br; crmello@deg.ufla.br
[3] Pesquisador de Pós-Doutorado, PPGRHASA/UFLA, Lavras, MG, Brasil. E-mail: aovinicius@gmail.com
[*] Author for correpondence

Received: July 14, 2019 - Approved: Ago. 23, 2019

a estimativa de vazões a partir de variáveis explicativas de fácil obtenção. Neste contexto, objetivou-se desenvolver funções regionais para vazão máxima (Qmax) aplicáveis a diferentes Períodos de Retorno (RP), vazão média em longo prazo (Qmlt) e vazão com 95% de permanência (Q_{95}) para as sub-bacias de alto e médio curso do rio Araguaia. Adotou-se a metodologia da vazão adimensional e a área de drenagem como variável explicativa. Os modelos regressivos testados foram o linear, potencial e quociente. Empregou-se para verificação da qualidade dos modelos regionais a validação-cruzada *leave-one-out*. Utilizou-se 10 distribuições estatística de 2 a 5 parâmetros. (i) Obtiveram-se resultados satisfatórios para todas as vazões de referência. (ii) A técnica de validação cruzada mostrou-se essencial para a seleção do modelo mais robusto. (iii) O modelo de quociente mostrou-se superior ao modelo potencial e linear na maioria dos casos.
Palavras-chave: Cerrado. Hidrologia Estatística. Modelagem Hidrológica.

Introduction

The Araguaia River basin has a calculated drainage area of 381,508 km². The establishment of a streamflow monitoring network in this basin with an adequate density of stations to support the development of hydrological studies is still a challenge, especially considering the high costs of implementing, operationalizing and maintaining stream gauging stations. Thus, several tributaries of the Araguaia River lack historical runoff series for use in hydrological analyses.

Thus, to enable a more efficient water resource management, tools capable of meeting this monitoring deficiency must be sought. One of the tools used for this purpose is streamflow regionalization, which is a technical possibility that meets the demand for reliable data on maximum, long-term mean, and minimum streamflows. Its use can contribute to both the design of hydraulic projects and the management of water resources, and it can be used as a reference for the development of academic research, especially for hydrological modelling.

However, caution should be exercised regarding its use because on one hand, for hydraulic works, there is the possibility of overestimation, which can lead to unnecessary construction costs. On the other hand, there is the risk of underestimation, which puts the integrity of the hydraulic work at risk (Cassalho et al., 2017b). The same concern applies to the minimum streamflows: overestimation will result in assigning a greater volume than the capacity of the water body, which may lead to conflicts, while underestimation may lead to non-optimal use.

The basin under study has multiple uses for surface water resources, namely, public water supply and irrigation, in addition to hydroelectric power generation with an installed capacity of 2,483 MW (http://www.epe.gov.br), as well as numerous other inventoried uses awaiting concession or environmental licensing. Because this basin is an important drainage basin that lacks data and is highly relevant in terms of its ability to generate hydroelectric power, it was decided that maximum and long-term mean streamflows should be regionalized because they are important references when planning hydraulic works and that the minimum streamflow 95% of the time (Q_{95}) should be regionalized because it is the reference when granting the rights to water use in the states of Goiás and Mato Grosso.

Regarding the minimum streamflows of the reference for granting the rights to the use of water resources, the states of Goiás and Mato Grosso set the run-of-river use limits to 70% of Q_{95}, as defined in Resolution No. 09 from May 04 of 2005 (*Conselho Estadual de Recurso Hídricos de Goiás* [State Council for Water Resources of Goiás] - CERH) and Resolution No. 27 from July 9 of 2009 (*Conselho Estadual de Recurso Hídricos do Mato Grosso* [State Council for Water Resources of Mato Grosso] - Cehidro). Q_{95} is the reference streamflow

adopted by the Brazilian National Water Agency (ANA) in most of the water bodies under control of the federal government in the Araguaia basin, as stated in the ANA Rights Granting Manual (www3.ana.gov.br).

To regionalize the maximum and minimum streamflows, it is essential to fit a probability distribution of the extreme values. There are several distributions recommended for this purpose, including 2-parameter Log-Normal, 3-parameter Log-Normal, Gumbel, and Generalized Extreme Values, among others. Several studies have been conducted that use different distributions (Ahn & Palmer, 2016; Basu & Srinivas, 2015; Cassalho, Beskow, Mello, & Moura, 2018). Thanks to computational advances, 4 or 5-parameter distributions are now being used (Cassalho et al., 2017b; Kjeldsen, Ahn, & Prosdocimi, 2017) and should be preferred whenever possible.

The aim of the present study was to regionalize the maximum, minimum and long-term mean streamflows of the upper and middle Araguaia River sub-basins. Specifically, the study sought to check the suitability of 10 distributions that use between 2 and 5 widely used parameters in hydrological studies, in order to identify those with the best fit for the maximum streamflows.

Materials and Methods

Database

The Araguaia River basin has a drainage area of 381,508 km², which covers part of the states of Mato Grosso, Goiás, Pará and Tocantins, and is the fourth largest hydrographic basin of South America, namely, the Tocantins-Araguaia River Basin. The basin is classified in Upper, Middle and Lower Araguaia (Latrubesse & Stevaux, 2002).

The digital elevation model (DEM) was obtained from ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) images downloaded directly from the Internet (https://earthexplorer.usgs.gov/). The digital processing of the DEM and the automatic delimitation of the hydrographic sub-basins were performed using the ArcGIS 9 geographic information system (Environmental Systems Research Institute [ESRI], 2002).

Historical streamflow series were obtained from the Hydrological Information System of ANA (HidroWEB-ANA). Figure 1 shows the location of the 27 selected stream-gauging stations, which had nine or more complete years of data, and their respective sub-basins were numerically delimited.

Based on the complete historical series, a reduced series of annual maximum streamflows (Q_{max}), annual mean streamflows (Q_{mean}) and the annual 95% streamflow permanence (Q_{95a}) were tabulated. For each station, the 95% streamflow permanence (Q_{95}) and the long-term mean streamflow (Q_{mlt}) were also obtained.

As recommended by (Tucci, 2002), it was determined whether the reduced series are stationary, adopting the non-parametric Mann-Kendall test (Cassalho et al., 2018; Cassalho et al., 2017b), where $\alpha = 0.05$ was the level of significance for the test. The series accepted by the Mann-Kendal test for each stream-gauging station were used to define the homogeneous regions from the dimensionless streamflow curves. Once it was determined which dimensionless series showed the same trends, the homogeneous regions were defined both graphically and by eye. In this sense, the hypothesis of the adopted method states that in a homogeneous region, the frequency distributions in the stations in that region are similar, except for a local scaling factor called the index-flood. The means of the respective streamflow series were adopted as a nondimensionalization factor (Naghettini & Pinto, 2007; Tucci, 2002).

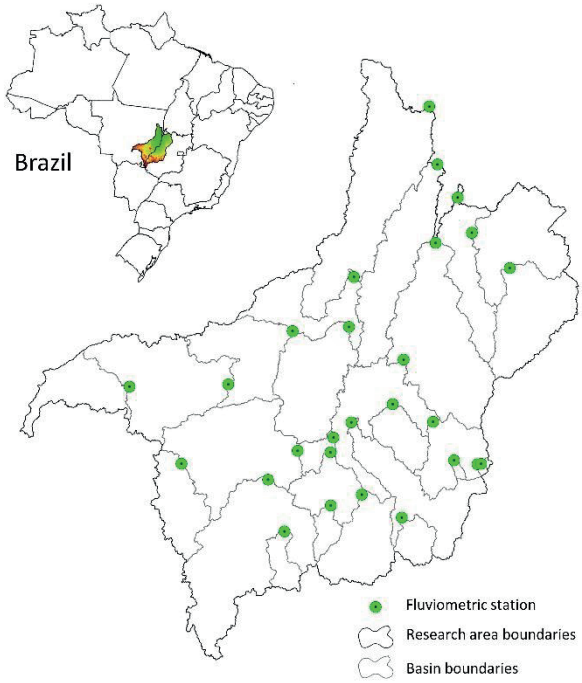


Figure 1. Location of river stations and delimitation of the respective sub-basins.

Probability distribution and parameter estimation

The probability distributions used were 2-parameter Log-Normal (LN2), 3-parameter Log-Normal (LN3), Person type III (PE-III), Gumbel, Generalized Extreme Values (GEV), Gamma (GAM), Generalized Logistic (GLO) and Generalized Pareto (GPA), 4-parameter Kappa (KAP), and 5-parameter Wakeby (WAK) (Beskow, Caldeira, Mello, Faria, & Guedes, 2015; Cassalho et al., 2018; Cassalho et al., 2017a; Jeong, Murshed, Seo, & Park, 2014; Kjeldsen et al., 2017).

The parameter estimation method used was L-Moments (Hosking, 1990), which was estimated using the R language in the R Studio programming environment with the support of the "lmomoc" library. The Anderson-Darling (AD) test was used to assess the goodness of fit of the distributions because it enables a comparison of different distributions or even methods (Cannarozzo, Noto, Viola, & La Loggia, 2009; Mello & Silva, 2013). The AD test was applied using a routine developed with support of the "nsRFA" library for the R language.

As recommended by Mello and Silva (2013), the fit of the distribution parameters was considered to be adequate when AD < 0.765, considering that the lower the value of the test, the better the fit.

Models and regionalization

The explanatory variable used in the regional models was the drainage area, since it is an important variable that explains hydrological behaviour (Tucci, 2002). Linear and non-linear models were used as described in Table 1.

Cross-validation was used either to evaluate the performance of the predictive models regarding their generalization capacity or to select the best characteristics for models (Bergstra & Bengio, 2012). This is an empirical technique that has gained ground in hydrological studies, including recent studies on streamflow regionalization (Cassalho et al., 2017a; Vezza, Comoglio, Rossoe, & Viglione, 2010). The method consists of dividing the dataset into k subsets of the same size; hence, it is widely known as *k-folds*. The fitting procedure

will be performed k times. For each round, a subset is separated for testing, and the other k-1 subsets are regrouped and used to estimate the model parameters. The model quality is calculated on the prediction errors of the test subsets. In this study, a specific case of cross-validation was adopted, namely, the leave-one-out, and the sample size was k (Cheng, Garrick, & Fernando, 2017; Mikshowsky, Gianola, & Weigel, 2017).

Table 1
Mathematical models used for regionalization of reference streamflows

Acronym	Type	Model
LM	Linear	$F(A) = aA + b$
PM	Potential	$F(A) = aA^b$
QM	Quotient	$F(A) = aA(b + A)^{-1}$

Note: "a" and "b" are estimated coefficients, "A" is the drainage area.

The model's goodness-of-fit and evaluation methods used were RMSE (Root Mean Square Error), R^2 (Pearson's Coefficient of Determination) and the index c proposed by Camargo and Sentelhas (1997). The performance of the models can be classified according to c as follows: $c \geq 0.85$, Great; $0.85 > c \geq 0.76$, Very good; $0.76 > c \geq 0.66$, Good; $0.66 > c \geq 0.61$, Fair; $0.61 > c \geq 0.51$, Poor; $0.51 > c \geq 0.41$, Very poor; and $c < 0.41$, Extremely poor.

The regional function for the quantiles Q_{95} and Qmlt was obtained via linear and non-linear regression. The regional function to estimate the maximum streamflow (m^3s^{-1}) for a given return period (RP; years) via the dimensionless curve method (Cannarozzo et al., 2009; Cassalho et al., 2017b; Tucci, 2002) is given by Eq. 1:

$$Qmax(A, TR) = f(A) * h(RP) \quad (1)$$

where f(A) is the regression between the dimensionless factors of the series and the selected explanatory variable. In this case, the dimensionless factor of the series is the mean maximum streamflow ($Q_{max\text{-}mean}$), the explanatory variable is the drainage area (A), and h(RP) is the inverse function of the distribution selected for regionalization. Each of the regional parameters (θj,r) associated with h(RP) are estimated parametrically via the mean of the parameters weighted by the length of the series for each homogeneous region (Naghettini & Pinto, 2007), as explained in Eq. 2:

$$\theta_{j,r} = \left(\sum_{i=1}^{N} n_i * \theta_{j,i}\right) * \left(\sum_{i=1}^{N} n_i\right)^{-1} \quad (2)$$

where N is the number of series of the region, ni is the amount of data in the series i, and θj,i is the parameter j of the distribution fitted for series i to be regionalized.

Results and Discussion

Delimitation of the homogeneous regions

For each reference streamflow, the slopes of the dimensionless streamflow curves plotted together can be observed in Figure 2A (Q_{max}), Figure 3A (Q_{mean}) and Figure 4A (Q_{95a}). The homogeneous regions were defined by eye, as observed in Figures 2 to 4. Table 2 presents the systematization of the homogeneous regions for the maximum streamflow (Q_{max}), long-term mean streamflow (Q_{mlt}) and 95% streamflow permanence (Q_{95}).

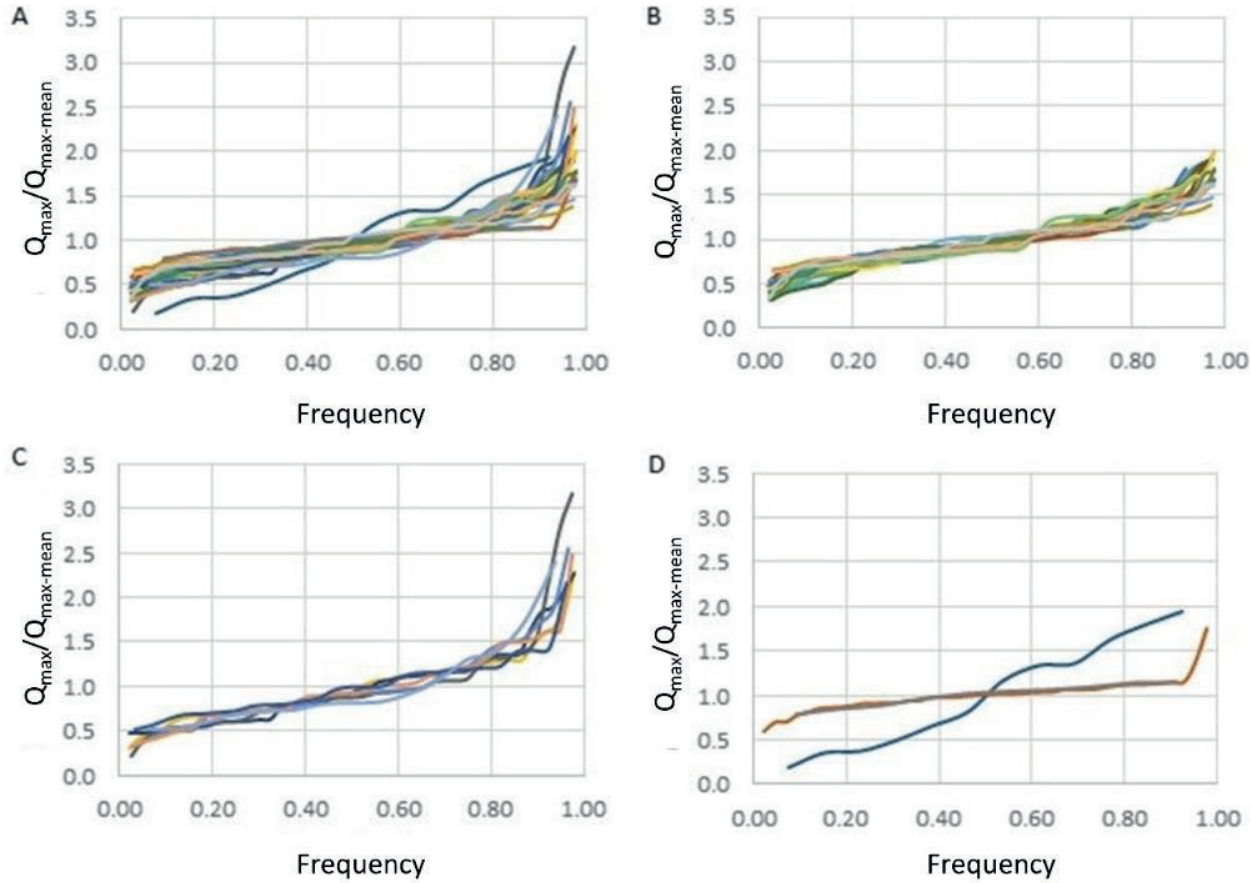


Figure 2. Sentiment trend analysis of the maximum flow series (Q_{max}) dimensionless by the mean maximum flow ($Q_{max\text{-}average}$) where in A are all series plotted together, B and C are the series grouped in homogeneous regions and D the series that showed no similar tendency to the defined regions or enough to compose another region.

Observing the trend of the dimensionless Q_{max} series plotted together, it was possible to define two homogeneous regions ($R1_{max}$ and $R2_{max}$). In this situation, three series were excluded (25090000, 2595000 and 26015000) for not showing similar trends to either of the two regions and because they did not comprise the minimum number of series for defining a third region.

Analysing the behaviour of the dimensionless Q_{95a} series plotted together, it was possible to define three homogeneous regions ($R1_{Q95}$, $R2_{Q95}$ and $R3_{Q95}$), with four series excluded (24750000, 25750000, 2580000 and 26150000) for not showing similar trends to the other series and not comprising the minimum number of series for forming a new region.

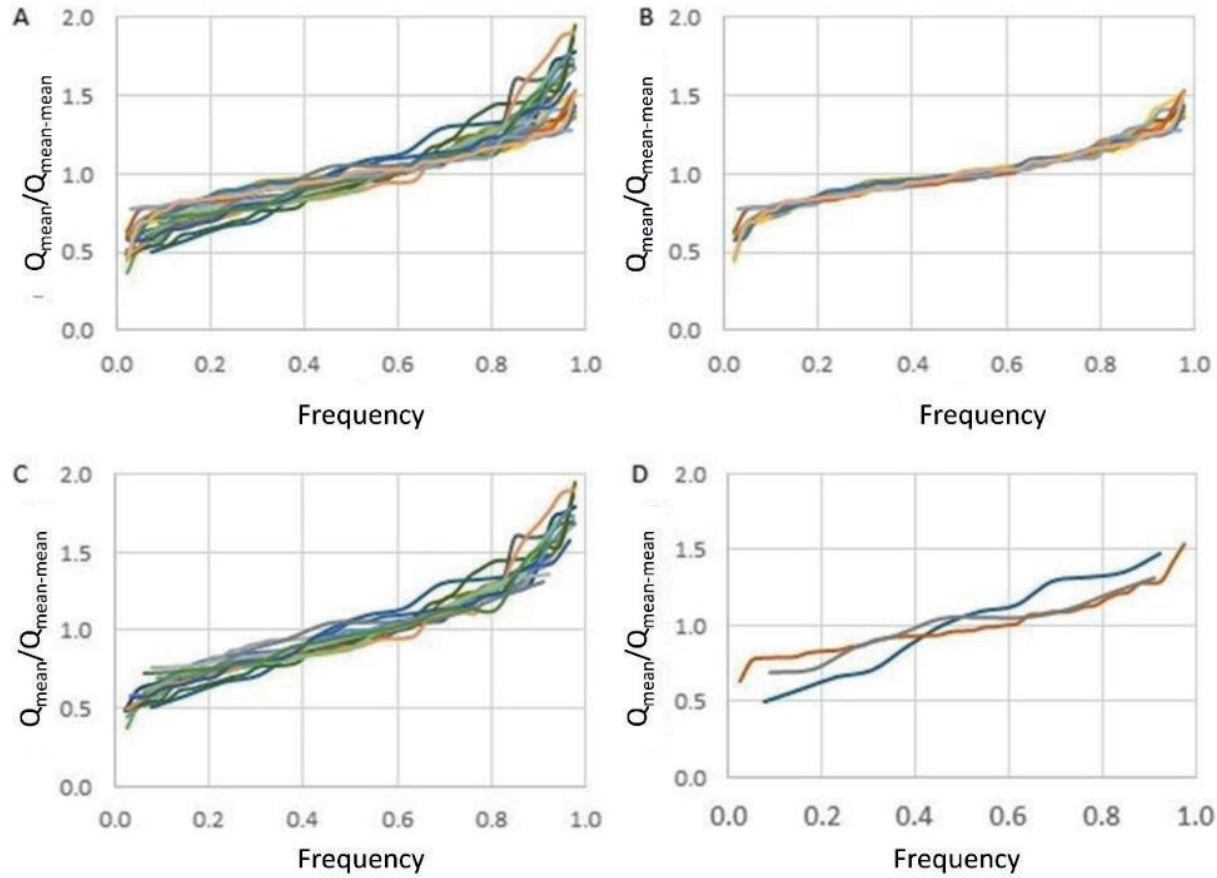


Figure 3. Sentiment trend analysis of long-term mean flowrates (Q) dimensionless by mean mean flow ($Q_{mean-mean}$) where in A are all series plotted together, B and C are series grouped in homogeneous regions and D the series that did not show a tendency similar to the defined regions or enough to compose another region.

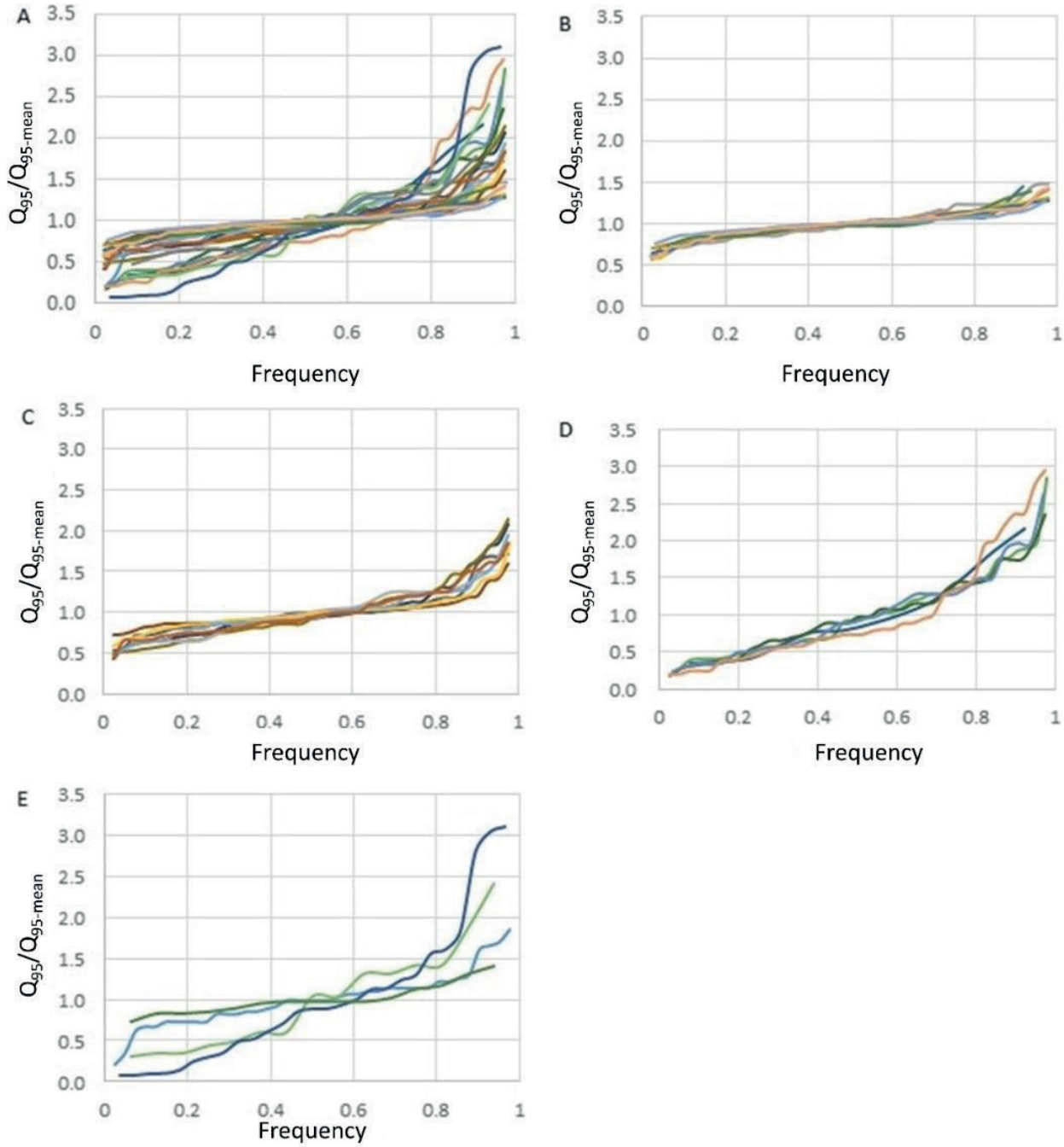


Figure 4. Sentiment trend analysis of 95% permanence series ($Q_{95)}$ dimensionless by the average Q_{95} ($Q_{95\text{-mean}}$) where in A are all series plotted together, B, C and D are the series grouped in homogeneous regions. and E the series that did not show a similar tendency to the defined regions or sufficient amount to compose another region.

Table 2
Homogeneous regions for the regionalization of maximum streamflow (Q_{max}), long-term mean streamflow (Q_{mlt}) and 95% streamflow permanence (Q_{95})

Station	Area (km²)	Q_{max}	Q_{mlt}	Q_{95}	Station	Area (km²)	Q_{max}	Q_{mlt}	Q_{95}
24196000	1788.6	1	2	1	25140000	3211.5	1	2	2
24200000	18306.3	1	1	1	25200000	76507.9	1	1	2
24500000	5201.2	1	2	1	25700000	92288.0	1	1	2
24700000	36675.2	2	1	1	25750000	8848.2	1	2	EF
24750000	6466.0	1	2	EF	25800000	18332.2	2	2	EF
24780000	1347.5	1	2	3	25950000	117053.0	EF	1	2
24800000	12021.4	2	2	2	26015000	10215.7	EF	1	1
24850000	49947.3	1	1	2	26040000	5312.5	1	ET	ET
24900000	2045.6	2	2	2	26050000	17768.3	1	1	1
24950000	10166.2	1	2	2	26100000	25323.4	1	1	1
25090000	114.2	EF	2	3	26150000	9463.1	2	2	EF
25100000	246.5	1	2	3	26200000	40826.6	1	1	1
25120000	225.0	2	2	3	26300000	58911.3	1	1	1
25130000	5342.90	2	2	3					

Note: EF – Excluded from the delineation of homogeneous regions ET – Excluded from the Mann-Kendal test. 1, 2 and 3 are the homogeneous regions.

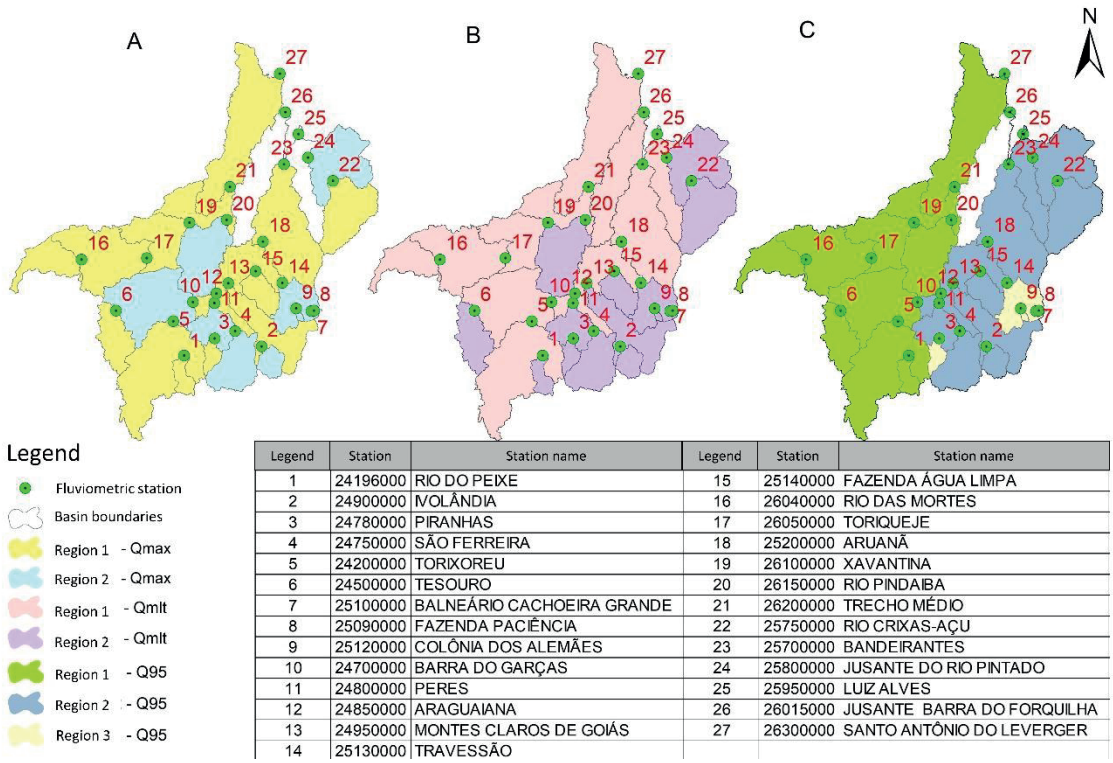
Observing the trend of the dimensionless Q_{mlt} series plotted together, it was possible to define two homogeneous regions ($R1_{mlt}$ and $R2_{mlt}$), with 3 series excluded (24750000, 25800000 and 26015000) for not displaying a similar trend to the other series and not comprising the minimum number of series for forming a new region.

The spatializations of the homogeneous regions for Q_{max}, Q_{mlt} and Q_{95} are shown in Figure 5A, 5B and 5C, respectively. Figure 5A reveals that the Mortes River sub-basin is partially contained in $R1_{max}$, excluding only one tributary along the right bank (Pindaíba River), which was inserted in $R2_{max}$. Also included in $R1_{max}$ are the headwater sub-basins of the Garças and Araguaia rivers. $R2_{max}$, in turn, was composed of the sub-basins of the lower Graças, Diamantino and Caiapó rivers and the headwater of the Claro River. The sub-basin of the Cristalino and Crixas Mirim rivers did not fall into any of the homogeneous regions; however, one tributary along the right bank of the Crixas Mirim River had its headwater drainage area inserted in $R1_{max}$, and the other sub-basins were inserted in $R2_{max}$.

According to Figure 5A, the different homogeneous regions did not present specific characteristics in terms of the drainage area of the sub-basins that they were composed of. $R1_{max}$ is composed of sub-basins between 246.55 and 92288.00 km², while $R2_{max}$ has areas between 225.09 and 36675.20 km². This indicates that sub-basins inserted in the same spatial scale may have different hydrological behaviours, and these behaviours may occur due to pedological, physiographic, geological, climatic, and land use differences, among others. This result demonstrates the importance of the establishment of homogeneous hydrological regions prior to the regionalization of streamflows.

Figure 5B depicts the spatialization of the homogeneous regions defined for the regionalization of the long-term mean streamflows (Q_{mlt}). The das Mortes River sub-basin is partially contained in $R1_{mlt}$, with the exception of the Pindaíba River

sub-basin. $R1_{mlt}$ is also composed of the headwater and the main course of the Araguaia River, the Diamantino River sub-basin, and the middle and lower Garças River. Notably, the Garças River headwaters, the Cristalino River sub-basin and the headwaters of several tributaries of the Araguaia River along the right bank are inserted in $R2_{mlt}$.



Legend	Station	Station name	Legend	Station	Station name
1	24196000	RIO DO PEIXE	15	25140000	FAZENDA ÁGUA LIMPA
2	24900000	IVOLÂNDIA	16	26040000	RIO DAS MORTES
3	24780000	PIRANHAS	17	26050000	TORIQUEJE
4	24750000	SÃO FERREIRA	18	25200000	ARUANÃ
5	24200000	TORIXOREU	19	26100000	XAVANTINA
6	24500000	TESOURO	20	26150000	RIO PINDAIBA
7	25100000	BALNEÁRIO CACHOEIRA GRANDE	21	26200000	TRECHO MÉDIO
8	25090000	FAZENDA PACIÊNCIA	22	25750000	RIO CRIXAS-AÇU
9	25120000	COLÔNIA DOS ALEMÃES	23	25700000	BANDEIRANTES
10	24700000	BARRA DO GARÇAS	24	25800000	JUSANTE DO RIO PINTADO
11	24800000	PERES	25	25950000	LUIZ ALVES
12	24850000	ARAGUAIANA	26	26015000	JUSANTE BARRA DO FORQUILHA
13	24950000	MONTES CLAROS DE GOIÁS	27	26300000	SANTO ANTÔNIO DO LEVERGER
14	25130000	TRAVESSÃO			

Figure 5. Spatialization of homogeneous regions for Q_{max} (A), Q_{mlt} (B) and Q_{95} (C).

Figure 5C depicts the spatial distribution of the homogeneous regions for Q_{95} regionalization. The das Mortes River sub-basin is completely inserted in $R1_{Q95}$, which is also composed of the Araguaia River headwater sub-basins and the Diamantino and Garças River sub-basins. The sub-basins along the right bank of the Araguaia River make up $R2_{Q95}$, except for the headwater of a tributary of the Caiapó River sub-basin and the headwaters of the Vermelho River, which are inserted in $R3_{Q95}$.

Fitting of the probability density functions and the regionalization of Qmax

The GEV probability distribution was the only accepted for all series according to the Anderson-Darling (AD) goodness-of-fit test presented in Table 3. Considering that the goal is to define a regional function that is capable of estimating the maximum streamflow for different RPs, the GEV distribution was adopted here. Cassalho et al. (2017a), using a similar regionalization methodology, adopted the GEV distribution, asserting that it is a robust model. Cassalho et al. (2017a), when fitting 10 distributions for basins in the state of Rio Grande do Sul, observed the GEV to have the best fit, corroborating the findings of the present study. Table 4 presents a summary of the parameters fitted for the GEV and the length of the series that were used to estimate the regional parameters of the distribution. The regional parameters of the GEV distribution obtained via the mean weighted by the length of the series for $R1_{max}$ were $\xi = 0.870$, $\alpha = 0.269$ and $\kappa = 0.112$, and for $R2_{max}$ were $\xi = 0.786$, $\alpha = 0.318$ and $\kappa = -0.079$. Thus, defining the term h(RP) of Eq. 1.

Table 3
Results of the Anderson-Darling goodness-of-fit test for the following distributions: 2-parameter Log-Normal (LN2), 3-parameter Log-Normal (LN3), Person type 3 (PE3), Gumbel (GUM), Generalized Extreme Values (GEV), Gamma (GAM), Generalized Logistics (GLO), Generalized Pareto (GPA), four-parameter Kappa (KAP) and five-parameter Wakeby (WAK)

Station	LN2	LN3	PE3	GUM	GEV	GAM	GLO	GPA	KAP	WAK
24196000	0.297	0.259	1.906	0.283	0.261	0.347	0.273	0.357	0.267	0.285
24200000	0.366	0.314	0.315	0.334	0.316	0.318	0.377	0.545	0.319	-
24500000	0.301	0.205	0.225	0.239	0.202	0.191	0.313	0.822	0.167	0.179
24700000	0.64	-	0.224	0.637	0.139	0.354	0.306	0.942	0.086	0.111
24750000	0.496	0.496	0.925	0.54	0.535	0.558	0.651	1.049	0.365	0.319
24780000	0.34	0.106	0.137	0.211	0.104	0.138	0.158	0.856	0.098	0.094
24800000	0.211	0.218	0.234	0.246	0.201	0.287	0.194	0.411	0.191	0.177
24850000	0.32	0.216	0.254	0.227	0.215	0.222	0.208	0.974	0.206	0.191
24900000	0.404	0.383	0.412	0.406	0.387	0.534	0.442	0.511	0.407	-
24950000	0.789	-	0.281	1.28	0.288	0.569	0.306	2.052	0.277	0.144
25100000	0.355	0.359	0.351	0.364	0.365	0.382	0.459	0.417	-	-
25120000	0.247	0.245	0.264	0.25	0.249	0.293	0.29	0.418	0.249	0.219
25130000	0.524	0.545	0.544	0.601	0.5	0.52	0.772	0.655	-	0.298
25140000	0.438	0.274	0.273	0.66	0.29	0.311	0.205	1.384	0.205	0.135
25200000	0.534	0.37	0.376	0.654	0.393	0.41	0.215	2.381	-	0.152
25700000	0.377	0.182	1.359	0.503	0.183	0.534	0.199	0.232	0.187	0.278
25750000	0.459	0.335	0.336	0.369	0.345	0.568	0.45	0.468	-	0.462
25800000	0.626	0.398	0.401	0.815	0.411	0.477	0.456	0.682	0.402	-
26040000	0.297	0.259	1.906	0.283	0.261	0.347	0.273	0.357	0.267	0.285
26050000	0.366	0.314	0.315	0.334	0.316	0.318	0.377	0.545	0.319	-
26100000	0.301	0.205	0.225	0.239	0.202	0.191	0.313	0.822	0.167	0.179
26150000	0.64	-	0.224	0.637	0.139	0.354	0.306	0.942	0.086	0.111
26200000	0.496	0.496	0.925	0.54	0.535	0.558	0.651	1.049	0.365	0.319
26300000	0.34	0.106	0.137	0.211	0.104	0.138	0.158	0.856	0.098	0.094

Table 4
Parameters of position (ξ), scale (α) and shape (κ) fitted for the GEV distribution and the length of the data series in years (N)

Station	Parameters			N	Station	Parameters			N
	ξ	α	κ			ξ	α	κ	
24196000	0.839	0.224	-0.126	11	25130000	0.806	0.352	0.027	38
24200000	0.831	0.259	-0.070	43	25140000	0.870	0.182	-0.120	12
24500000	0.870	0.281	0.129	48	25200000	0.830	0.274	-0.041	44
24700000	0.827	0.322	0.041	44	25700000	0.929	0.243	0.384	40
24750000	0.821	0.355	0.080	40	25750000	0.833	0.292	0.005	15
24780000	0.879	0.347	0.292	40	25800000	0.841	0.272	-0.008	27

continue

continuation

24800000	0.783	0.327	-0.082	42	26040000	0.919	0.201	0.207	32
24850000	0.857	0.216	-0.080	40	26050000	0.893	0.246	0.166	46
24900000	0.722	0.313	-0.243	39	26100000	0.882	0.261	0.142	45
24950000	0.886	0.268	0.178	43	26150000	0.726	0.258	-0.333	15
25100000	0.827	0.374	0.130	35	26200000	0.879	0.195	-0.042	33
25120000	0.767	0.337	-0.105	30	26300000	0.901	0.263	0.250	45

Table 5 presents the statistics associated with the fit of the different regression models tested to estimate Qmax-mean, allowing for the term f(A) of Eq. 1 to be defined.

For $R1_{max}$, the quotient model showed the best results, being accepted in the cross-validation, with R^2 equal to 0.87, a lower RMSE and a higher "c" in the fitting stage. In $R2_{max}$, the potential and quotient models were not accepted in the cross-validation. Thus, the only model that proved to be robust was the linear model. The selected models are classified as Great ($c > 0.85$) according to the confidence index proposed by Camargo and Sentelhas (1997).

Table 5
Analysis of the quality of the regression models and the goodness-of-fit for estimating $Q_{max\text{-}mean}$

Index by Region	Cross-validation			Goodness-of-fit		
	LM	PM	QM	LM	PM	QM
Region 1						
R^2	0.8	0.8	0.8	0.85	0.87	0.87
RMSE	739.84	687.88	702.22	623.7	599.87	599.85
c	0.85	0.85	0.84	0.89	0.89	0.9
Region 2						
R^2	0.69	0.34	0.2	0.87	0.9	0.87
RMSE	670.61	1027.05	1084.04	455.21	451.36	498.66
c	0.73	0.38	0.24	0.9	0.92	0.9

Based on the cross-validation statistics, it can be seen that even the model with the best fit may not be sufficiently robust for different possible situations, as was the case of the potential model (PM) and the quotient model (QM) for estimation of Qmax-mean in region 2 ($R2_{max}$). It should therefore be rejected.

The regression models selected to estimate $Q_{max\text{-}mean}$ of region 1 and region 2, $f_{r1max}(A)$ and $f_{r2max}(A)$, respectively, are presented in Eq. 3 and Eq. 4, respectively:

$$f_{r1\,max}(A) = \frac{10641.39 * A}{(121018.22 + A)} \quad (3)$$

$$f_{r2\,max}(A) = 0.092 * A \quad (4)$$

Figure 6 shows graphs of the models used to estimate $Q_{max\text{-}mean}$ in $R1_{max}$ and $R2_{max}$ relative to the observed data. An adequate fit can be observed, as measured by the aforementioned statistical coefficients.

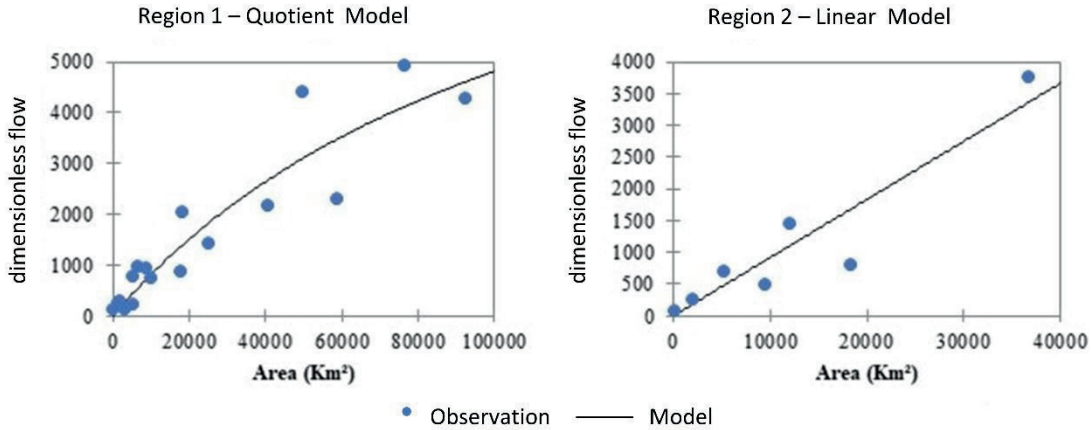


Figure 6. Regression models graphs for $Q_{max\text{-}mean}$ estimates (dimensionless flow) in Araguaia river sub-basins.

Therefore, based on the regional parameters of the GEV distribution and the regressions for estimating the dimensionless streamflow (scaling factor), the regional functions for estimating the maximum streamflow in m3s-1 as a function of the RP (years) and of the drainage area (km2) are given by Eq. 5 and Eq. 6 for $R1_{max}$ and $R2_{max}$, respectively:

$$Q_{R1_{max}}(A, RP) = \left[\frac{10641.39 * A}{(121018.22 + A)}\right] * \left\{0.870 + 2.402 * \left[1 - \left(-LN\left(1 - \frac{1}{RP}\right)\right)^{0.112}\right]\right\} \quad (5)$$

$$Q_{R2_{max}}(A, RP) = [0.092 * A] * \left\{0.786 + 4.025 * \left[1 - \left(-LN\left(1 - \frac{1}{RP}\right)\right)^{-0.079}\right]\right\} \quad (6)$$

The quality of the regional models for RP values of 2, 10, 50 and 100 years from the values of Qmax obtained directly from the GEV distribution for the stream-gauging stations are shown in Table 6.

Table 6
Quality of regional functions statistics for estimating Q_{max} in different return periods

Index	RP (Region 1)				RP (Region 2)			
	2	10	50	100	2	10	50	100
d	0.97	0.95	0.93	0.92	0.96	0.96	0.96	0.96
r	0.94	0.91	0.87	0.85	0.94	0.93	0.92	0.92
c	0.91	0.87	0.81	0.78	0.90	0.89	0.89	0.88
RMSE	495.20	881.55	1364.46	1607.11	402.13	670.61	945.13	1093.27
Confidence	Great	Great	Very good	Very good	Great	Great	Great	Great

The RMSE values are proportional to the return period, which suggests the need for caution in the use of the models for high RP values. However, it is noteworthy that the results are similar to those obtained by Cassalho et al. (2017a) and were considered adequate in the regionalization of the maximum streamflow for the hydrographic sub-basins in the state of Rio Grande do Sul.

Regionalization of Q_{mlt} and Q_{95}

Table 7 presents the statistics of the regression models used to regionalize Q_{mlt}. The models fitted for the two homogeneous regions according to Camargo and Sentelhas (1997) are classified as Great. Regarding the statistical parameters, the quotient model showed the best performance for the two regions.

Pruski, Rodriguez, Pruski, Nunes and Rego (2016), when regionalizing Q_{mlt} for the São Francisco River sub-basins, obtained models with R^2 coefficients equal to 0.898 and 0.893, using the drainage area and the mean annual precipitation as explanatory variables for each model, respectively. Thus, because the results obtained in the present study are comparable, it is considered that the models proposed for the Araguaia River sub-basins are parsimonious and can be considered effective and efficient for estimating Q_{mlt}.

Table 7
Analysis of the quality of the regression models and the goodness-of-fit for estimating Q_{mlt}

Index by Region	Cross-validation			Goodness-of-fit		
	LM	PM	QM	LM	PM	QM
Region 1						
R^2	0.95	0.97	0.99	0.97	0.99	0.99
RMSE	105.49	73.18	53	82.3	61.13	50.1
c	0.96	0.98	0.99	0.98	0.99	0.99
Region 2						
R^2	0.83	0.85	0.86	0.9	0.93	0.93
RMSE	27.64	25.63	25.64	21.62	18.75	17.59
c	0.87	0.89	0.89	0.92	0.94	0.95

The fitted regional functions are presented in Eq. 7 and Eq. 8 for region 1 and region 2, respectively:

$$Q_{r1\ mlt}(A) = \frac{4812.35 * A}{(232224.17 + A)} \quad (7)$$

$$Q_{r2\ mlt}(A) = \frac{466.27 * A}{(23721.36 + A)} \quad (8)$$

Figure 7 reveals the goodness-of-fit of the regional functions to the observed data. An adequate fit is observed, reinforcing the results of the abovementioned accuracy statistics.

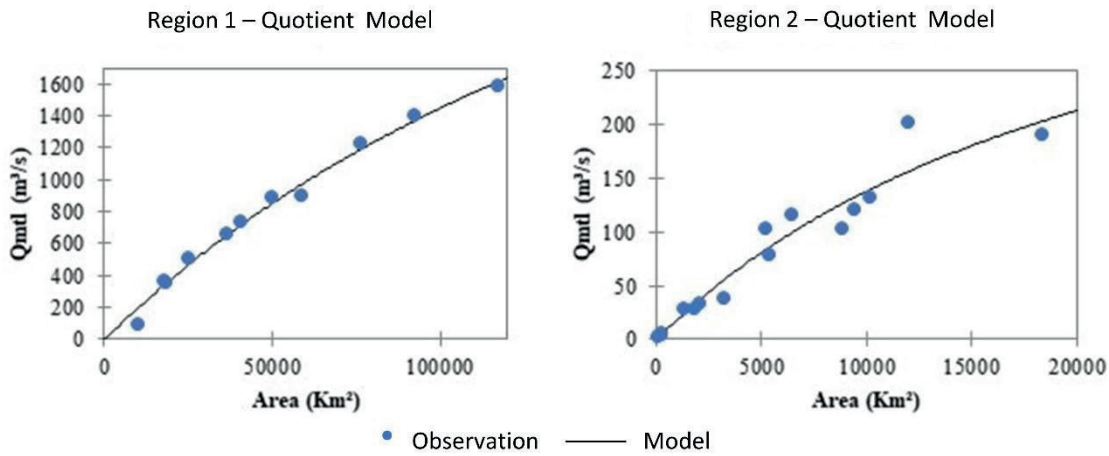


Figure 7 Regression models graphs for Q_{mlt} estimates in Araguaia river sub-basins.

Table 8 presents the statistics of the regression models used to regionalize Q_{95}. The three models fitted for the homogeneous regions according to Camargo and Sentelhas (1997) are classified as Great. However, the best performance was obtained by the quotient model for the three regions. Beskow et al. (2016), when regionalizing Q_{90}, obtained regional functions with a confidence index c between 0.86 and 0.97, and the results were considered adequate by the authors. The regional functions for regions 1, 2 and 3 are shown in Eq. 9, Eq. 10 and Eq. 11, respectively:

$$Q_{r1Q95}(A) = \frac{919.53 * A}{(103720.76 + A)} \quad (9)$$

$$Q_{r2Q95}(A) = \frac{696.89 * A}{(116928.95 + A)} \quad (10)$$

$$Q_{r3Q95}(A) = \frac{9.73 * A}{(5170.80 + A)} \quad (11)$$

Table 8
Analysis of the quality of the regression models and the goodness-of-fit for estimating Q_{95}

Index by Region	Cross-validation			Goodness-of-fit		
	LM	PM	QM	LM	PM	QM
Region 1						
R^2	0.76	0.78	0.80	0.85	0.87	0.88
RMSE	57.93	54.96	50.91	50.03	47.14	45.02
c	0.81	0.83	0.85	0.88	0.90	0.91
Region 2						
R^2	0.87	0.9	0.95	0.93	0.96	0.98
RMSE	52.01	45.45	32.63	41.13	33.99	25.37
c	0.9	0.92	0.96	0.95	0.97	0.98
Region 3						
R^2	0.92	0.94	0.96	0.97	0.99	1.00
RMSE	1.48	1.45	1.38	0.37	0.20	0.06
c	0.87	0.88	0.98	0.98	1.00	1.00

Figure 8 shows the regional functions for Q_{95} estimation with the observed data and reveals an adequate fit of the models to the data for the three homogeneous regions.

Seeking significant improvements in the current management of water resources in the drainage basin of the Araguaia River, with the establishment of drainage basin committees the proposed regionalization models can be important tools for the technical support of such efforts. This is relevant in the context of maximum streamflows, especially for the technical segment related to the design of hydraulic works that involve surface drainage.

Regarding the mean streamflows, a tool is provided that can be widely used in the development of the Hydrographic Basin Master Plan and in the initial stage of the design of small reservoirs. In this context, the Q_{95} prediction models are particularly relevant, as they aim to provide technical support when determining the reference streamflow for granting water use rights in the Araguaia River sub-basins.

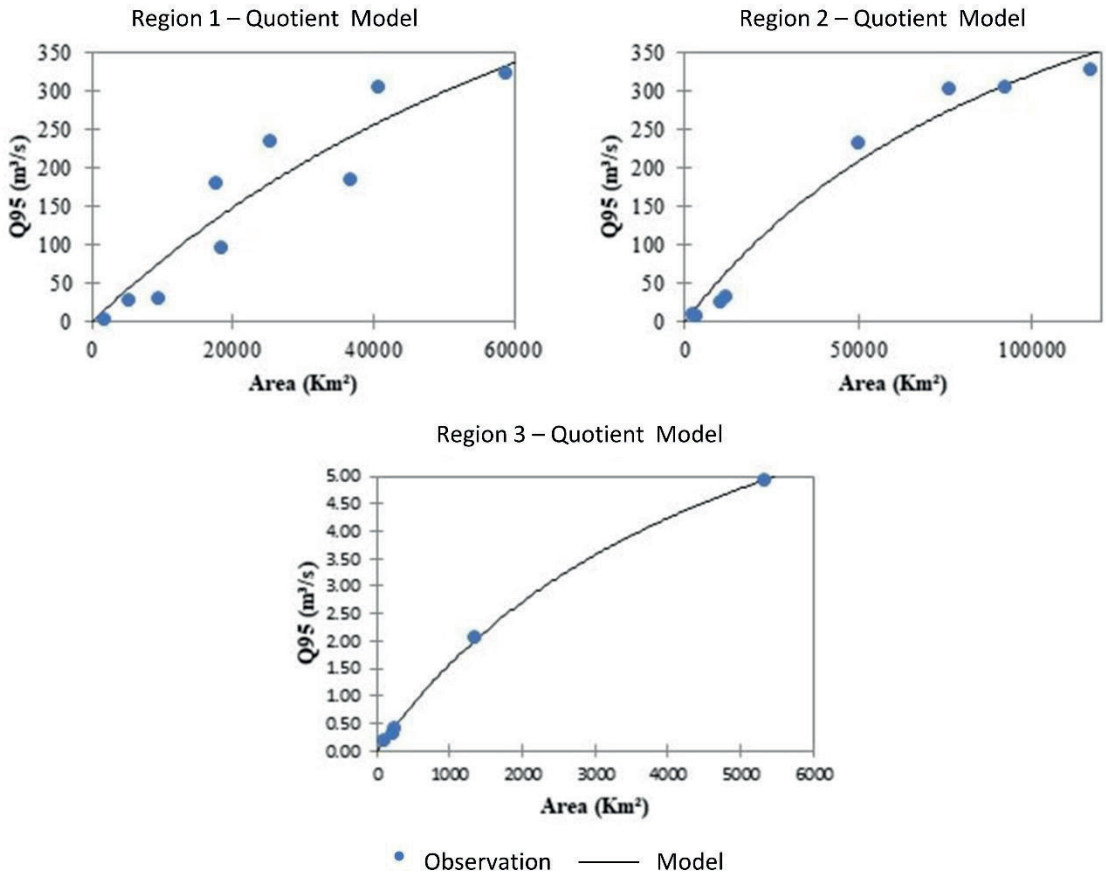


Figure 8. Regression models graphs for Q95 estimates in Araguaia river sub-basins.

Conclusion

The technique for defining homogeneous regions by eye with the use of a trend analysis of dimensionless streamflows proved efficient for the sub-basins of the upper and middle Araguaia River.

The obtained results allow us to conclude that the regional functions fitted for Q_{max}, Q_{mlt} and Q_{95} are suitable for use, respecting the application limits (drainage areas) and observing the exceptions regarding the behaviour of the curves.

The use of cross-validation is recommended for selecting the most robust models of each situation and allowing for the best extrapolation.

Moreover, it is concluded that simple linear and nonlinear models can present satisfactory results. In this case, the quotient model was more suitable than the potential model, which is most commonly employed in streamflow regionalization studies.

Acknowledgements

Federal University of Lavras (Edital PRPG nº 43/2019) and National Council for Scientific and Technological Development (CNPq) for the promotion of project No. 308947/2018-5.

References

Ahn, K.-H., & Palmer, R. (2016). Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique. *Journal of Hydrology, 540*, p. 515-526. doi: 10.1016/j.jhydrol.2016.06.047

Basu, B., & Srinivas, V. V. (2015). A recursive multi-scaling approach to regional flood frequency analysis. *Journal of Hydrology, 529*, Part 1, p. 373-383. doi: 10.1016/j.jhydrol.2015.07.037

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(Feb), p. 281-305. Retrieved from http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Beskow, S., Caldeira, T. L., Mello, C. R. de, Faria, L. C., & Guedes, H. A. S. (2015). Multiparameter probability distributions for heavy rainfall modeling in extreme southern Brazil. *Journal of Hydrology: Regional Studies, 4*, Part B, p. 123-133. doi: 10.1016/j.ejrh.2015.06.007

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

Camargo, A. P. de, & Sentelhas, P. C. (1997). Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no estado de São Paulo, Brasil. *Revista Brasileira de Agrometeorologia, 5*(1), p. 89-97. Retrieved from http://www.leb.esalq.usp.br/agmfacil/artigos/artigos_sentelhas_1997/1997_RBAgro_5(1)_89-97_ETPM%E9todosSP.pdf

Cannarozzo, M., Noto, L. V., Viola, F., & La Loggia, G. (2009). Annual runoff regional frequency analysis in Sicily. *Physics and Chemistry of the Earth, Parts A/B/C, 34,* (10-12), p. 679-687. doi: 10.1016/j.pce.2009.05.001

Cassalho, F., Beskow, S., Mello, C. R., & Moura, M. M. (2018). Regional flood frequency analysis using L-moments for geographically defined regions: An assessment in Brazil. *Journal of Flood Risk Management*, e12453. doi: 10.1111/jfr3.12453

Cassalho, F., Beskow, S., Mello, C. R., Moura, M. M., Kerstner, L., & Ávila, L. F. (2017a). At-site flood frequency analysis coupled with multiparameter probability distributions. *Water Resources Management, 32*(1), p. 285-300. doi: 10.1007/s11269-017-1810-7

Cassalho, F., Beskow, S., Vargas, M. M., Moura, M. M, Ávila, L. F., & Mello, C. R. (2017b). Hydrological regionalization of maximum stream flows using an approach based on L-moments. *RBRH, 22*, e27. doi: 10.1590/2318-0331.021720160064

Cheng, H., Garrick, D. J., & Fernando, R. L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology, 8*(1), p. 1-5. doi: 10.1186/s40104-017-0164-6

Environmental Systems Research Institute. (2002). *REDLANDS. ArcGIS Schematics Documentation. (Version 9)*.

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological), 52*(1), p. 105-124. Retrieved from http://www.jstor.org/stable/2345653

Jeong, B. Y., Murshed, M. S., Seo, Y. A., & Park, J.-S. (2014). A three-parameter kappa distribution with hydrologic application: a generalized gumbel distribution. *Stochastic Environmental Research and Risk Assessment, 28*(8), p. 2063-2074. doi: 10.1007/s00477-014-0865-8

Kjeldsen, T. R., Ahn, H., & Prosdocimi, I. (2017). On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrological Sciences Journal, 62*(9), p. 1354-1363. doi: 10.1080/02626667.2017.1335400

Latrubesse, E. M., & Stevaux, J. C. (2002). Geomorphology and environmental aspects of the Araguaia fluvial basin, Brazil. *Zeitschrift fur Geomorphologie, Supplementband, 129*, 109-127.

Mello, C. R., & Silva, A. M. de. (2013). *Hidrologia:* princípios e aplicações em sistemas agrícolas. Lavras: Ed. UFLA.

Mikshowsky, A. A., Gianola, D., & Weigel, K. A. (2017). Assessing genomic prediction accuracy for Holstein sires using bootstrap aggregation sampling and leave-one-out cross validation. *Journal of Dairy Science, 100*(1), p 453-464. doi: 10.3168/jds.2016-11496

Naghettini, M., & Pinto, É. J. A. (2007). *Hidrologia Estatística*. Belo Horizonte: CPRM.

Pruski, F. F., Rodriguez, R. G., Pruski, P. L., Nunes, A. A., & Rego, F. S. (2016). Extrapolation of regionalization equations for long-term average flow. *Engenharia Agrícola, 36*(5), 830-838. doi: 10.1590/1809-4430-Eng.Agric.v36n5p830-838/2016

Tucci, C. E. M. (2002). *Regionalização de vazões*. Porto Alegre: Ed. UFRGS.

Vezza, P., Comoglio, C., Rosso, M., & Viglione, A. (2010). Low Flows Regionalization in North-Western Italy. *Water Resources Management*, *24*(14), p. 4049-4074. doi: 10.1007/s11269-010-9647-3