Revista Brasileira de Meteorologia, **v. 37, n. 4,** 503−516, 2022
DOI: http://dx.doi.org/10.1590/0102-77863440052

rbmet.org.br

Article

Regional Frequency Analysis for the Prediction of Maximum Flows in Ungauged Basins of the Peruvian Amazon

Efrain Lujano[1], German Belizario[1], Apolinario Lujano[2]

[1]Escuela Profesional de Ingeniería Agrícola, Universidad Nacional del Altiplano, Puno, Perú.
[2]Autoridad Nacional del Agua, Puno, Perú.

Recebido em: 28 de Maio de 2022 - Aceito em: 2 de Outubro de 2022

Resumo

A estimativa da vazão máxima de projeto e importante para o gerenciamento de inundações. No entanto, a existência limitada de sítios calibrados e a escassez de medições hidrológicas impossibilitam sua estimativa em bacias não calibradas. Neste estudo, a análise de frequência regional (RFA) foi realizada para a previsão de vazões máximas em bacias não calibradas da Amazônia peruana. A metodologia consistiu na identificação de regiões homogêneas, seleção da função de distribuição regional, estimação de quantis regionais, regionalização do índice de inundação e previsão de vazões máximas em bacias não calibradas. Os resultados identificaram uma região homogênea bem definida chamada região 1. A distribuição de valores extremos generalizados (GEV) mostrou-se mais adequada para representar a amostra de dados da região 1, e a área da bacia explicou a variabilidade do cheia-índice em 99,4% (R^2 = 0,994). A previsão de vazões máximas em bacias não calibradas apresentou amplas faixas de incerteza, principalmente para períodos de retorno alto. Conclui-se que o RFA fornece estimativas confiáveis para a previsão de vazões máximas desde que sejam consideradas as faixas de incerteza em cada frequência.

Palavras-chave: análise multivariada, vazões máximas, bacia amazônica, cheia-índice, momentos L, regionalização hidrológica.

Análise de Frequência Regional para a Previsão de Vazões Máximas em Bacias não Calibradas da Amazônia Peruana

Abstract

The estimation of the maximum design flow is important for flood management. However, the limited existence of gauged sites and the scarcity of hydrological measurements make it impossible to estimate them in ungauged basins. In this study, the regional frequency analysis (RFA) was carried out for the prediction of maximum flows in ungauged basins of the Peruvian Amazon. The methodology consisted of the identification of homogeneous regions, selection of the regional distribution function, estimation of regional quantiles, regionalization of the index-flood, and the prediction of maximum flows in ungauged basins. The results have identified a well-defined homogeneous region called region 1. The generalized extreme value (GEV) distribution proved to be more adequate to represent the data sample of region 1, and the basin area explained the variability of the index-flood in 99.4% (R^2 = 0.994). The prediction of maximum flows in ungauged basins presented wide ranges of uncertainty, mainly for high return periods. It is concluded that the RFA provides reliable estimates for the prediction of maximum flows as long as the uncertainty ranges are considered at each frequency.

Keywords: multivariate analysis, maximum flows, amazon basin, index-flood, L-moments, hydrological regionalization.

Corresponding autor: Efrain Lujano, elujano28@gmail.com.

1. Introduction

Design maximum flows are of importance for flood management, disaster risk management, hydraulic planning, and design of hydraulic structures, however, they must be predicted from a theoretical distribution function that best fits sample data from a site of interest.

An adequate technique to make predictions of maximum design flows when information is available is the local frequency analysis (LFA) (OMM, 2011). Although the methodology of the LFA is well established (Viglione, 2007; Hosking and Wallis 1997) and allows to make reliable predictions associated with high return periods (Campos-Aranda, 2016), the information that is required is not always available in time and space, so it is necessary to use other methodologies.

In the Amazon basin of Peru, the limited existence of hydrometric stations and the short length of measurements make it impossible to estimate the maximum design flow in ungauged basins. Consequently, the lack of information on maximum design flows has led to inadequate flood management in the study area, generating a lot of damage. Given this problem, the regional frequency analysis (RFA) is an appropriate method that uses records from several sites to characterize the study variable in non-instrumented sites, thus allowing flow rates to be obtained through unsupervised data such as the morphometric parameters of the basins, with more precise inferences when they are in the same homogeneous region (Hosking and Wallis, 1997).

The ARF has been successfully applied in flood modeling, for example in Canada (Msilini *et al.*, 2020; Desai and Ouarda, 2021), Switzerland (Le *et al.*, 2022), Australia (Zaman *et al.*, 2012), Turquía (Saf, 2009), Pakistán (Khan *et al.*, 2017), Brazil (Rezende de Souza *et al.*, 2021), Southeastern Europe (Lescesen *et al.*, 2022), likewise, it has been applied for the evaluation of hydrological drought in the Czech Republic (Strnad *et al.*, 2020) and analysis of meteorological drought in Indonesia (Kuswanto *et al.*, 2021). A detailed comparison of various RFA methodologies can be found in Cunnane (1988) and GREHYS (1996). The RFA includes the use of L-moments (Hosking and Wallis, 1997) together with the index-flood method (Dalrymple, 1960), a useful alternative for the transfer of information to sites that make up a supposed homogeneous region (Rodriguez and Marreno de León, 2011).

The stages of the RFA are based on the identification of homogeneous regions, selection of the regional frequency distribution, estimation of regional quantiles, and regionalization of the index-flood (Hosking and Wallis, 1997; Viglione, 2007). An exhaustive RFA requires the identification of homogeneous regions (IHR) (Hosking and Wallis, 1997; Rodriguez and Marreno de León, 2011) and for this, Ward's cluster analysis and principal components can be applied (Gottschalk, 1985).

Ward's method has been widely studied for the classification of different climatic and hydrological data (Domroes *et al.*, 1998; Jackson and Weinand, 1995; Nathan and McMahon, 1990; Ramachandra Rao and Srinivas, 2006; Hosking and Wallis, 1997). The use of cluster analysis with hydrological variables is based on the similarity of hydrological characteristics, such as geographical, physical, statistical, or stochastic properties (Hassan and Ping, 2012). Hosking and Wallis (1997) recommend that the IHR be based on site statistics, however, Viglione (2007) notes that it is preferable to use site features rather than site statistics.

In the process of selecting the regional frequency distribution, it is best to rely on the sample mean and not on a line of best fit through the data points (Naghettini and Pinto, 2007; Peel *et al.*, 2001). Hosking and Wallis (1997) recommend using the statistic Z goodness-of-fit test (Z^{DIST}) for the selection of regional frequency distribution. Consequently, the adjusted regional parameters can be transferred to the specific sites with confidence (Parida and Moalafhi, 2008).

ARF applications in Peru were also studied with satisfactory results in estimating extreme events such as maximum rainfall (Fernández and Lavado, 2016; Lujano and Obando, 2015), determination of drought maps (Acuña *et al.*, 2015; Acuña *et al.*, 2011) and regionalization of monthly and annual average flows (Lujano *et al.*, 2016; Lujano *et al.*, 2017). However, the prediction of maximum design flows in ungauged basins through ARF has not yet been reported for the Amazon basin of Peru.

Since ARF can be a suitable method for predicting maximum design flows, we focus our analysis on the main research question: Is it possible to obtain accurate results of maximum design flows for different return periods in ungauged basins of the Peruvian Amazon? To provide reasonable answers, this study aimed to perform a regional frequency analysis for the prediction of maximum flows in ungauged basins of the Peruvian Amazon. With the results of the research, it is expected to contribute to the estimation of maximum flows for different return periods, in a short, fast, reliable way and at a low economic cost in ungauged basins within the Amazon basin of Peru. In addition, the results will serve as input for environmental management, disaster risk management, flood control, hydraulic planning, and the design of hydraulic structures.

2. Materials and Methods

2.1. Study area and data

The study area comprised 10 basins within the Amazon hydrographic region in Peru (Fig. 1 and Table 1). The smallest and largest basin has an approximate area of 360.4 km^2 and 877,478. 6 km^2 respectively (Table 2). It is

Table 1 - Characteristics of hydrometric stations.

Basin number	River	Station	Latitude [°]	Longitude [°]	Data range
1	Amazonas	Tamishiyacu (TAM)	-4.000	-73.160	1985-2010
2	Ucayali	km 105 (KM)	-13.183	-72.534	1958-2012
3	Chotano	Chotano Lajas (CHL)	-6.560	-78.741	1979-2008
4	Llaucano	Llaucano Corellama (LLC)	-6.687	-78.518	1980-2011
5	Napo	Bellavista (BE)	-3.480	-73.080	1990-2009
6	Ucayali	Pucallpa (PU)	-8.378	-74.533	1988-2009
7	Amazonas	Tabatinga (TAB)	-4.250	-69.933	1983-2017
8	Amazonas	Nazareth (NAZ)	-4.121	-70.036	1990-2004
9	Marañon	Borja (BO)	-4.470	-77.548	1987-2016
10	Marañon	San Regis (SR)	-4.516	-73.908	1999-2014

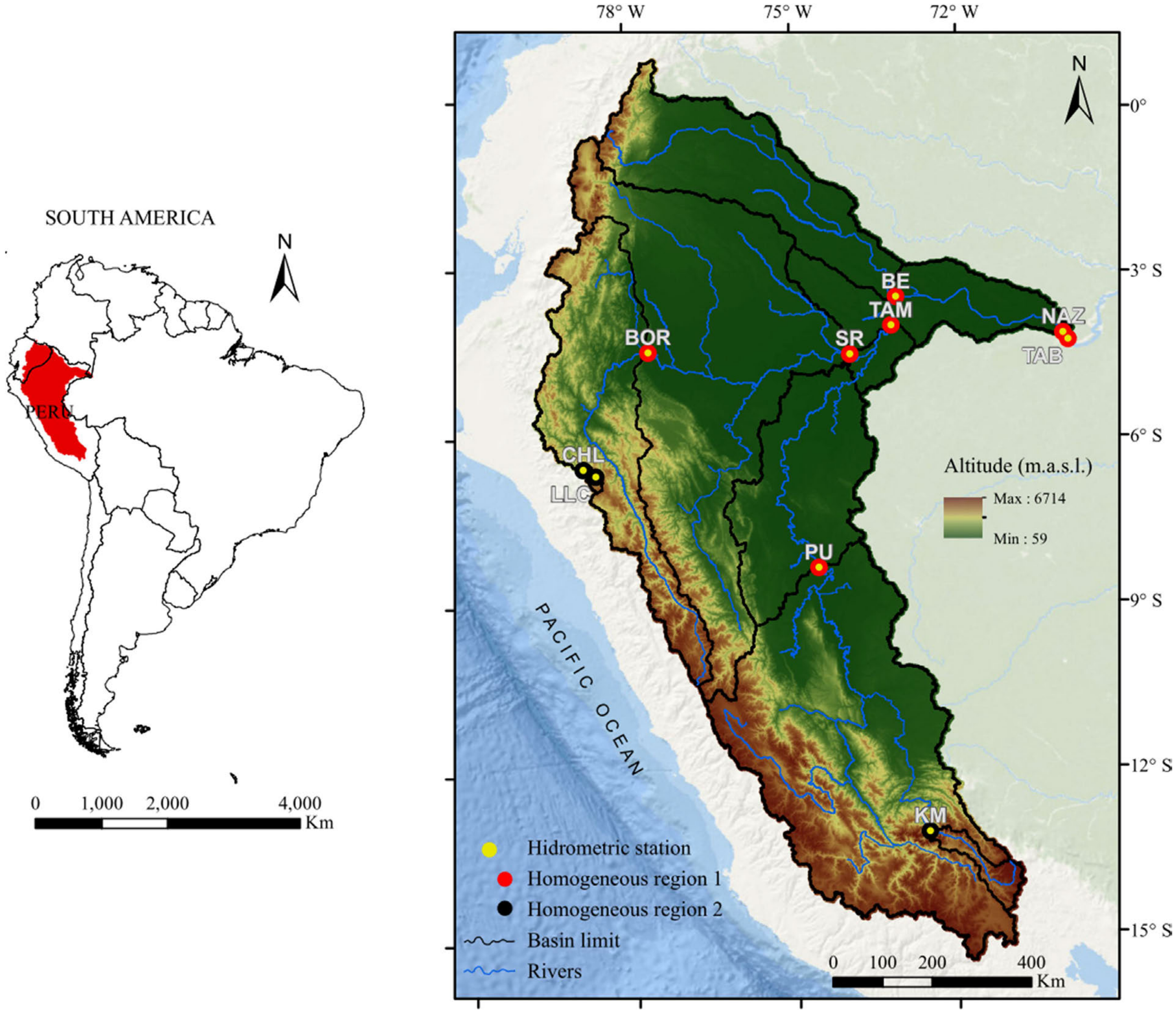

Figure 1 - Location of the study area and spatial distribution of hydrometric stations.

characterized as an exorheic basin system, bordering the western limit with the hydrographic region of the Pacific, to the north with Colombia, while to the east it borders Brazil. The hydrographic region of Titicaca and part of the hydrographic region of the Pacific are its hydrographic limits to the south. The hydrological regime in the southern area of the basin is generalized in the austral summer while in the northern area in autumn.

According to Peruvian Interpolation data of SENAMHI Climatological and hydrological Observations (PISCO) precipitation and temperature climate data (1981-2016), with daily temporal resolution and spatial resolution of 0.1° (Aybar *et al.*, 2017), the mean annual precipitation for the basins varies between 737 to 2080.5 mm, while the mean annual temperature varies between 10.0 and 25.7 °C. On the other hand, based on the hydrometric record, the mean annual flow for the basins varied between 74.4 m^3/s and 55092.9 m^3/s (Table 2).

To define the basin area, we used the NASA Shuttle Radar Topographic Mission (SRTM) Digital Elevation Model (DEM), obtained from the Google Earth Engine (GEE) platform, image ID CGIAR/ SRTM90_V4 (Jarvis *et al.*, 2008), with a spatial resolution of ~90 m. The maximum average daily flows were collected from 10 stations located within the study area, of which 5 stations (Tamishiyacu, Chotano Lajas, Llaucano Corellama, Bellavista, and Pucallpa) belong to the National Service of Meteorology and Hydrology of Peru (SENAMHI), 1 station (km 105) to the Machupicchu Electric Generation Company (EGEMSA), while 4 stations (Tabatinga, Nazareth, Borja and San Regis) correspond to the SO-HYBAM Observation Service (formerly Environmental Research Observatory) “Geodynamic, hydrological and biogeochemical control of erosion/alteration and transport of materials in the Amazon, Orinoco and Congo basins”, which has been operational since 2003, responding to an invitation from the French Ministry of Higher Education and Research, which aims to provide the research community with the high-quality scientific data necessary to understand and model the behavior of systems and their long-term dynamics (Table 1).

Table 2 - Basin area (A), functions to estimate instantaneous maximum flows (Q_p) and index-flood ($\overline{Q}$), mean annual precipitation (MAP), mean annual temperature (MAT).

Station	A [km^2]	Q_p [m^3/s]	$\overline{Q}$ [m^3/s]	MAP [mm]	MAT [°C]
TAM	719917.8	$Q_p = 1.047Q_m$	48594.2	1669.0	21.4
KM	9613.3	$Q_p = 1.170Q_m$	640.8	737.0	10.0
CHL	360.4	$Q_p = 1.455Q_m$	74.4	897.8	15.1
LLC	608.7	$Q_p = 1.389Q_m$	117.8	839.5	12.7
BE	99779.4	$Q_p = 1.084Q_m$	11820.2	2080.5	25.7
PU	260890.0	$Q_p = 1.063Q_m$	20164.4	1456.1	17.3
TAB	877478.6	$Q_p = 1.044Q_m$	55092.9	1759.5	22.2
NAZ	877066.5	$Q_p = 1.044Q_m$	54278.8	1759.5	22.2
BO	114529.8	$Q_p = 1.081Q_m$	13178.4	1205.3	19.6
SR	356882.9	$Q_p = 1.057Q_m$	29042.3	1757.8	23.0

2.2. Identification of homogeneous regions

2.2.1. Multivariate analysis

A fundamental step in the RFA is the IHR. Ward's Method (Ward, 1963), k-means (Hartigan and Wong, 1979), and Andrews curves (Andrews, 1972), are some of the processes used for IHR. Although there are some variations of the Andrews equations, we used the function suggested in Khattree and Naik (2002). Cluster analysis based on Ward's method helps in the preliminary formation of homogeneous regions and takes into account site characteristics (basin area, elevation, latitude and longitude of the measurement site) (Hosking and Wallis, 1997). In this study, site characteristics (latitude and longitude) and site statistics (coefficient of L-variation (L-CV), L-skewness, and L-kurtosis) were considered. Taking into account that the heterogeneity measure is defined in terms of L-CV and the goodness-of-fit measure of the regional frequency distribution is defined in terms of L- kurtosis, Lucas *et al.* (2017) in their study considered L-CV and L-kurtosis for the identification of homogeneous regions.

To define a stable hydrological frequency distribution that allows probabilistic predictions to be estimated at a site, it is necessary that the sample size be large enough (OMM, 2011). However, Hosking and Wallis (1997) indicate that in the RFA the sample size should be ≥ 15 years, moreover, using the parameter estimation method (L-moments) they can produce very reliable results with small sample sizes and even with outliers. Under these premises, we consider hydrometric stations that have a data record ≥ 15 years. From the instantaneous maximum flows, the L-CV, L-skewness, and L-kurtosis site statistics were calculated based on the L-moment relationships. The maximum instantaneous flows (Q_p) in m^3/s, were estimated based on the relationship proposed by Fuller (1914). The equation is based on the area of the basin (A) in km^2 and the daily average maximum flow (Q_m) in m^3/s:

$$Q_p = Q_m\left(1 + \frac{2.66}{A^{0.3}}\right) \quad (1)$$

The values of 2.66 and 0.3 are dimensionless parameters obtained by Fuller from the study of 24 basins of different sizes in the USA.

2.2.2. L-moments

They constitute an alternative system to the traditional method of conventional moments (Hosking, 1990) and arise from linear combinations of the probability weighted moments (PWMs) introduced by Greenwood *et al.* (1979). For Hosking and Wallis (1997) the estimate is based on a sample of size n, organized in ascending order $x_{1:n} \leq x_{2:n} \leq \ldots \leq x_{n:n}$. It is convenient to start with an estimator of PWMs β_r. An impartial estimator of β_r is:

$$\beta_r = n^{-1}\binom{n-1}{r}\sum\nolimits_{j=r+1}^{n}\binom{j-1}{r}x_{j:n} \qquad (2)$$

Alternatively, it can be written as:

$$\beta_0 = n^{-1}\sum\nolimits_{j=1}^{n} x_{j:n} \qquad (3)$$

$$\beta_1 = n^{-1}\sum\nolimits_{j=2}^{n}\frac{(j-1)}{(n-1)}x_{j:n} \qquad (4)$$

$$\beta_2 = n^{-1}\sum\nolimits_{j=3}^{n}\frac{(j-1)(j-2)}{(n-1)(n-2)}x_{j:n} \qquad (5)$$

$$\beta_3 = n^{-1}\sum\nolimits_{j=4}^{n}\frac{(j-1)(j-2)(j-3)}{(n-1)(n-2)(n-3)}x_{j:n} \qquad (6)$$

Similarly, the L-moments of the sample are defined by:

$$l_1 = \beta_0 \qquad (7)$$

$$l_2 = 2\beta_1 - \beta \qquad (8)$$

$$l_3 = 6\beta_2 - 6\beta_1 + \beta_0 \qquad (9)$$

$$l_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \qquad (10)$$

where l_1 is the L-location or mean of the distribution, and l_2 is the L-scale. L-CV can be defined as:

$$t = \frac{l_2}{l_1} \qquad (11)$$

While the coefficients of L-skewness and L-kurtosis as:

$$t_3 = \frac{l_3}{l_2} \qquad (12)$$

$$t_4 = \frac{l_4}{l_2} \qquad (13)$$

2.2.3. Regional discordancy measure

In this stage, the entire data set was analyzed to verify the existence of incorrect values, outliers, trends, and changes in the sample mean using the measure of discordancy (D_i). A site is discordant if $D_i > D_c$. The discordancy measure (D_i) for station i is defined as:

$$D_i = \frac{1}{3}N(u_i - \bar{u})^T A^{-1}(u_i - \bar{u}) \qquad (14)$$

where $\bar{u} = N^{-1}\sum_{i=1}^{N} u_i$ is the unweighted regional average of vectors u_i, $A = \sum_{i=1}^{N}(u_i - \bar{u})(u_i - \bar{u})^T$ is the matrix of sums of squares and cross products, N is the number of stations in the study region, $u_i = \left[t^{(i)}, t_3^{(i)}, t_4^{(i)}\right]^T$, is a vector of the L-moments ratios for the i-th site, $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$ are the L-CV, L-skewness and L-kurtosis for station i respectively.

2.2.4. Regional heterogeneity test

The homogeneity of each region is assessed using measures of heterogeneity $H1$, $H2$, and $H3$, each based on a different measure of the spread between sites of the L-moments ratios (L-CV, L-skewness and L-kurtosis). Hosking and Wallis (1993) found that $H2$ and $H3$ lacked the power to discriminate between homogeneous and heterogeneous regions and that $H1$ based on L-CV had a much better discriminant power. Therefore, $H1$ was considered as a main indicator of heterogeneity, denoted by H.

$$H = \frac{(V - \mu_V)}{\sigma_V} \qquad (15)$$

where μ_V and σ_V are the mean and standard deviation of V, derived from a large number of simulated values (N_{sim}) of the region under study. The weighted standard deviation V, is calculated as:

$$V = \left\{\frac{\sum_{i=1}^{N} n_i\left(t^{(i)} - t^R\right)^2}{\sum_{i=1}^{N} n_i}\right\}^{1/2} \qquad (16)$$

where N is the number of sites in a homogeneous region, n_i the sample size for station i, $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$ denotes the ratio of L-moments of the sample, $t^{(R)}$, $t_3^{(R)}$ and $t_4^{(R)}$ are expressed as the regional average of L-CV, L-skewness and L-kurtosis, weighted proportionally to the record length of the sites.

$$t^R = \frac{\sum_{i=1}^{N} n_i t^{(i)}}{\sum_{i=1}^{N} n_i} \qquad (17)$$

It fits a kappa distribution as its frequency distribution the average regional L-moments ratios 1, $t^{(R)}$, $t_3^{(R)}$ and $t_4^{(R)}$. A large number ($N_{sim} = 500$) of realizations of a region with N sites are simulated. A region is declared “acceptably homogeneous” if $H < 1$, “possibly heterogeneous” if $1 \leq H < 2$ and “definitely heterogeneous” if $H \geq 2$ (Hosking and Wallis, 1997), or alternatively “acceptably homogeneous” if $H < 2$, “possibly heterogeneous” if $2 \leq H < 3$ and “definitely heterogeneous” if $H \geq 3$ (Wallis *et al.*, 2007).

2.3. Selection of regional frequency distribution

Five three-parameter probabilistic distribution functions were evaluated, namely generalized logistic (GLO), generalized extreme value (GEV), generalized Pareto (GPA), generalized normal (GNO), and Pearson type III (PE3). Parameter estimation was performed using the L-moments method. According to Hosking and Wallis (1997) two-parameter distributions can cause bias in the tail of the estimated quantiles if the shape of the true frequency distribution is not well approximated by the fitted distribution. The best fit distribution is one that gives robust estimates for the regional growth curve as well as for the quantiles at each site. For more detail on the distribution functions, we refer the reader to Hosking and Wallis (1997).

2.3.1. Goodness-of-fit test

The regional frequency distribution is chosen based on the goodness-of-fit test Z^{DIST} (Hosking and Wallis, 1997). For each candidate distribution Z^{DIST} is defined:

$$Z^{DIST} = \frac{\left(\tau_4^{DIST} - t_4^R + B_4\right)}{\sigma_4} \qquad (18)$$

where τ_4^{DIST} is the L-kurtosis coefficient of the fitted distribution, $DIST$ refers to GLO, GEV, GPA, GNO and PE3, the standard deviation of t_4^R is calculated with:

$$\sigma_4 = \left[\left(N_{sim} - 1\right)^{-1}\left\{\sum\nolimits_{m=1}^{N_{sim}}\left(t_4^{[m]} - t_4^R\right)^2 - N_{sim}B_4^2\right\}\right]^{1/2} \qquad (19)$$

and the bias of t_4^R is defined by:

$$B_4 = N_{sim}^{-1}\sum\nolimits_{m=1}^{N_{sim}}\left(t_4^{[m]} - t_4^R\right) \qquad (20)$$

N_{sim} is the simulated regional data set, using a kappa distribution. The fit is adequate if Z^{DIST} is sufficiently close to zero, a reasonable criterion being $|Z^{DIST}| \leq 1.64$. This criterion corresponds to the acceptance of the hypothetical distribution at a confidence level of 90%.

2.4. Regional quantile estimation

The index-flood (Dalrymple, 1960), was used to estimate maximum flow quantiles for different return periods. The key assumption of an index-flood procedure is that sites that form a homogeneous region have an identical frequency distribution called the regional growth curve, but a site-specific scale factor (Hosking and Wallis, 1997), the index-flood. The equation used to estimate the quantiles of maximum flows was:

$$Q_i(T) = \overline{Q}_i q(T) \qquad (21)$$

where $Q_i(T)$ is the maximum flow estimate for site i for a given return period of T years in m^3/s, $\overline{Q}_i$ is the site-dependent scale factor, the index-flood in m^3/s, and q (T) is the dimensionless regional growth curve estimated from the regional distribution function of a supposedly homogeneous region, i = 1, 2, …, N denotes the sites and N the number of sites. The sample mean of the maximum flow series (l_1) is used as the index-flood (Hosking and Wallis, 1997; OMM, 2011; Viglione, 2007).

2.5. Prediction in ungauged basins

The most used characteristics in the regionalization of flows are the drainage area, the length of the main river, the average slope of the main river, the drainage density, and the unevenness of the basin (Tucci, 2002). Morphometric characteristics of the basins, including area, mean elevation, mean slope, length of the main river, slope of the main river, area above 2000 m.a.s.l., the orientation of the basin, center of gravity, the radius of circularity, and climatic characteristics such as the Thornthwaite index and the Budyko index (Viglione, 2007). In this research, to regionalize the index-flood, the area of the basin was used, first verifying the correlation between both variables. The statistical significance of the correlation coefficient was evaluated using the t'Student test at a significance level of 5%.

Multiple regression models are the most used to estimate the index-flood in sites without measured data (Viglione, 2007), this approach links the index-flood with the characteristics of the basin. In this study, the linear regression equation was used between the area of the basin and the index-flood.

$$\overline{Q} = \beta_0 + \beta_1(A) + \varepsilon \qquad (22)$$

where $\overline{Q}$ is the index-flood in m^3/s, A is the basin area in km^2, β_0 and β_1 are the regression parameters and ε is an error term.

For the estimation of the regression parameters of the model β_0 and β_i, the method of least squares was used and the statistical significance was evaluated by means of the t'Student test with a level of significance of 5%. The results of the regression were also evaluated through the coefficient of determination R^2, defined by:

$$R^2 = \frac{\left(\sum\nolimits_{i=1}^{n}\left(O_i - \overline{O}\right)\left(S_i - \overline{S}\right)\right)^2}{\left(\sum\nolimits_{i=1}^{n}\left(O_i - \overline{O}\right)^2\left(\sum\nolimits_{i=1}^{n}\left(S_i - \overline{S}\right)\right)\right)^2} \qquad (23)$$

where n is the number of sites, S_i is the simulated value, $\overline{S}$ is the mean of simulated values; O_i is the observed value and $\overline{O}$ is the mean of observed values.

The regression model must satisfy general assumptions such as homoscedasticity (variation of the residual is constant) and normality of residuals (residuals are normally distributed) (Vezza *et al*., 2010; Vezza *et al*., 2009). To detect heteroscedasticity, residuals were plotted against fitted values and were also verified using Harrison and McCabe (1979) test. On the other hand, to evaluate the normality of the residuals, we used the Anderson Darling (AD) normality test. Although in order to avoid heteroscedasticity and the non-normality of the regression residuals and obtain greater model efficiency, different transformations of the index-flood $(\overline{Q})$ can be used, such as $\sqrt{\overline{Q}}$, $\sqrt[3]{\overline{Q}}$ or $ln(\overline{Q})$ (Viglione, 2007; Viglione *et al.*, 2007). In this study, the untransformed index-flood was considered.

The regional frequency analysis procedure for predicting maximum flows in ungauged basins is summarized in the flow diagram of Fig. 2.

3. Results and Discussions

3.1. Identification of homogeneous regions

3.1.1. Multivariate analysis

The functions to estimate the instantaneous maximum flows (Q_p) of 10 hydrometric stations (Table 2) deduce that with an increase in the size of the basin, the coefficient to estimate Q_p decreases, while in small basins the coefficient increases. For all floods, Q_p must be greater than Q_m, this is because, in large basins, the runoff rate is high for at least 24 h because the storm that generates it is of considerable duration, while in small basins a storm can

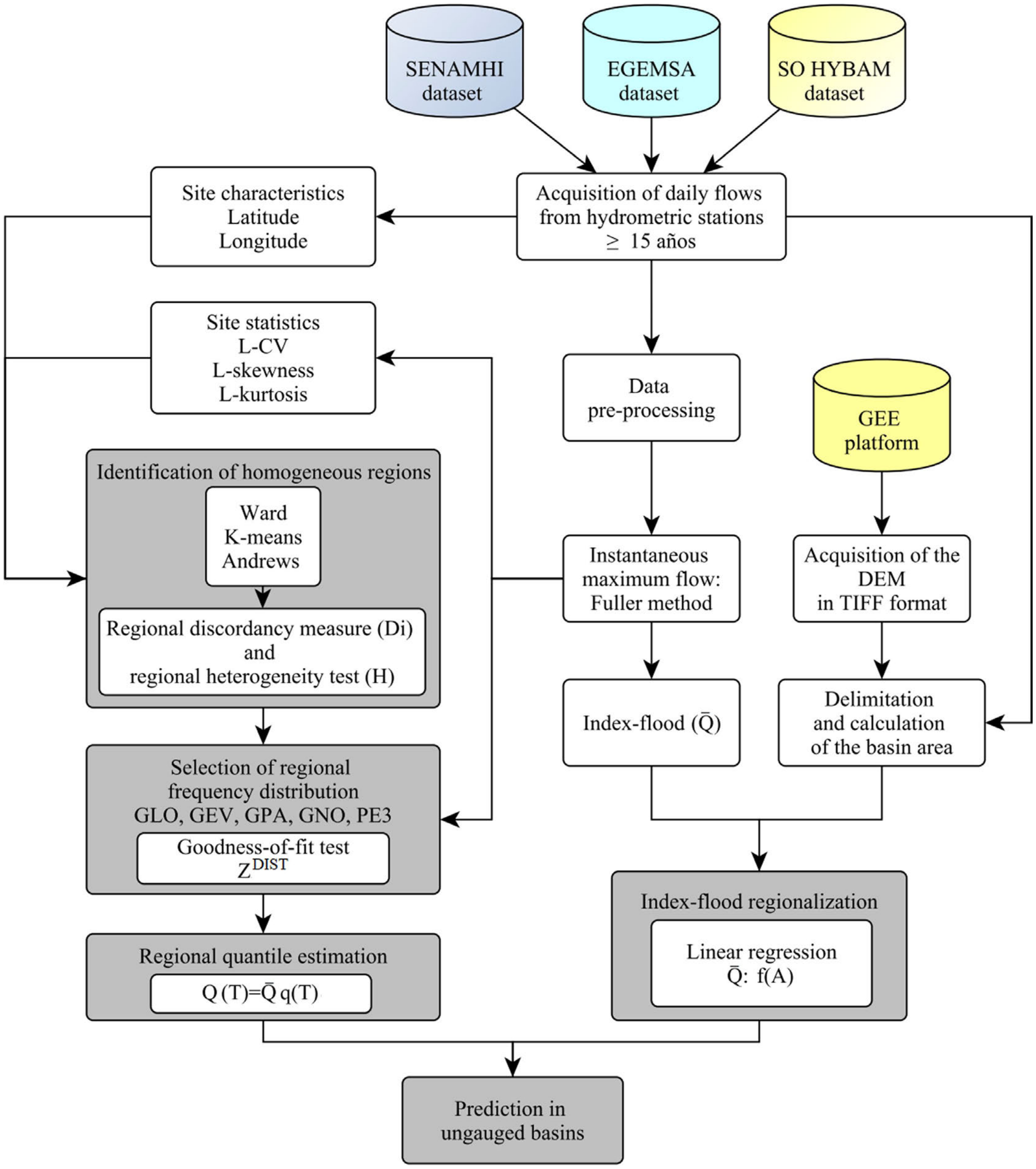
SENAMHI dataset
EGEMSA dataset
SO HYBAM dataset
Site characteristics Latitude Longitude
Acquisition of daily flows from hydrometric stations ≥ 15 años
Site statistics L-CV L-skewness L-kurtosis
Data pre-processing
GEE platform
Identification of homogeneous regions
Ward K-means Andrews
Regional discordancy measure (Di) and regional heterogeneity test (H)
Instantaneous maximum flow: Fuller method
Acquisition of the DEM in TIFF format
Index-flood (Q̄)
Delimitation and calculation of the basin area
Selection of regional frequency distribution GLO, GEV, GPA, GNO, PE3
Goodness-of-fit test Z^DIST
Index-flood regionalization
Linear regression Q̄: f(A)
Regional quantile estimation
Q (T)=Q̄ q(T)
Prediction in ungauged basins

Figure 2 - Flow diagram of the regional frequency analysis for the prediction of maximum flows in ungauged basins.

cause flooding in few hours resulting in a large Q_p and moderate Q_m (Fuller, 1914).

The Ward, k-means, and Andrews methods agreed on the formation of homogeneous regions Figs. 3a-3c respectively. Region 1 includes the TAM, BE, PU, TAB, NAZ, BO, and SR sites, while region 2 may not be well defined due to the low number of sites. Although Ward's method is considered a suitable procedure for a preliminary determination of homogeneous regions (Domroes *et al*., 1998; Hosking and Wallis, 1997; Jackson and Weinand, 1995; Gottschalk, 1985), however, the k-means method and Andrews curves also allowed the identification of basins with similar hydrological behavior.

Although there is no correct number of groups, a balance must be struck between using regions that are too small or too large. Homogeneous regions containing few sites will achieve little improvement in the precision of quantile estimates over *in situ* analysis. However, as you increase the sites in the region, the precision is higher, but little gain in quantile estimation precision is obtained by using more than about 20 sites in a homogeneous region (Hosking and Wallis, 1997; Viglione, 2007). Under these premises, in the following analysis, only one homogeneous region (region 1) was considered, made up of 7 TAM, BE, PU, TAB, NAZ, BO, and SR sites.

3.1.2. Regional discordancy measure and regional heterogeneity test

Results of regional and site L-moments relationships for 7 homogeneous basins are given in Fig. 4, L-skewness with L-CV (Fig. 4a) and L-skewness with L-kurtosis (Fig. 4b).

The statistics of the discordancy measure for each site (D_i), indicate that the values are lower than the critical discordancy (D_c), deducing that the region of 7 sites is not discordant, with D_i of each site less than 2.76 (Table 3). For the results of frequency analysis in hydrology to be theoretically valid, each data sample must satisfy certain basic assumptions, such as randomness, independence, homogeneity, and seasonality (OMM, 2011). Hosking and Wallis (1997), indicate that a site is discordant if D_i exceeds the D_c of the group, also in the context of the RFA using L-moments, they found that when comparing the relations of the L-moments of the samples from different sites, incorrect data values, outliers, trends, and changes in the mean may be reflected in the L-moments of the data from each site. Apparently, when records are short, climate variability can easily give rise to a trend and can disappear when as much information as possible has been collected (Kundzewicz and Robson, 2004).

The heterogeneity statistic for the 7-site group was H = -1.25 (Table 3). According to Hosking and Wallis (1997), a region is declared "acceptably homogeneous" if $H < 1$, "possibly heterogeneous" if $1 \leq H < 2$, and "definitely heterogeneous" if $H \geq 2$. For their part, Wallis *et al.* (2007) establish that a region is "acceptably homogeneous" if $H < 2$, "possibly heterogeneous" if $2 \leq H < 3$, and "definitely heterogeneous" if $H \geq 3$. Under these deductions, the set of 7 sites belongs to a supposed homogeneous region. if $H > 1$ Hosking and Wallis (1997) suggest that further subdivision of the region should be considered, as it could improve the precision of the quantile estimates. As it is a purely statistical criterion Wallis *et al.* (2007) and Schaefer *et al.* (2006) consider that a region

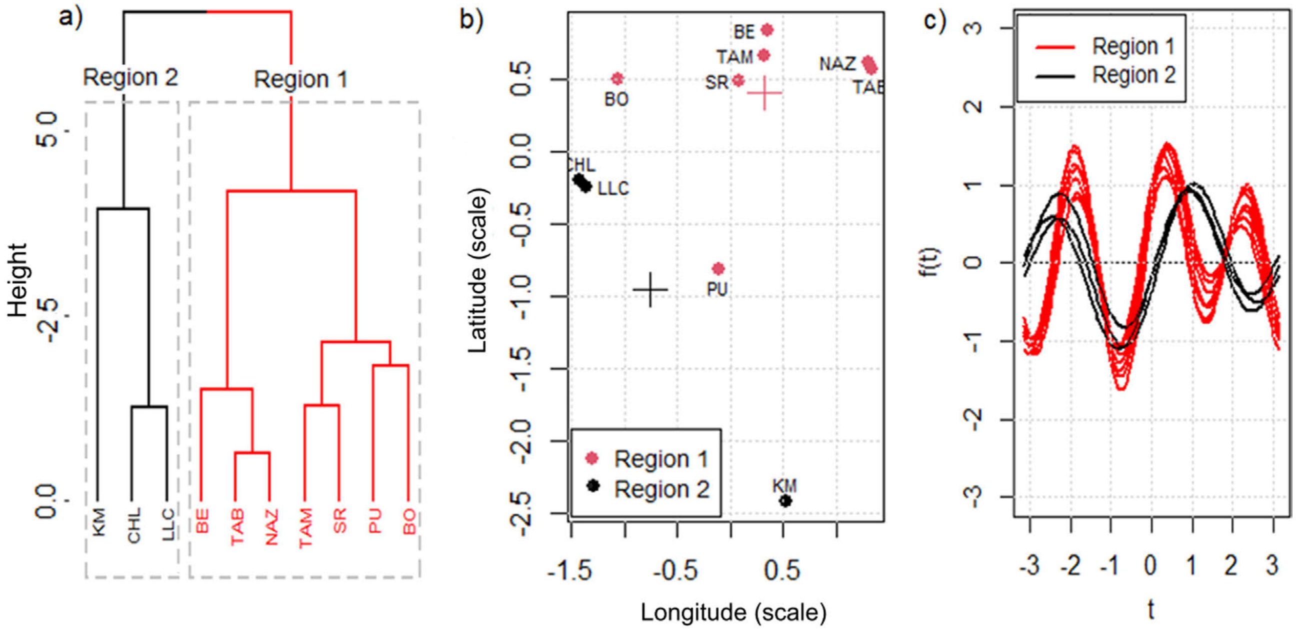

Figure 3 - Identification of homogeneous regions a) Ward's dendrogram, b) k-means and c) Andrews curves.

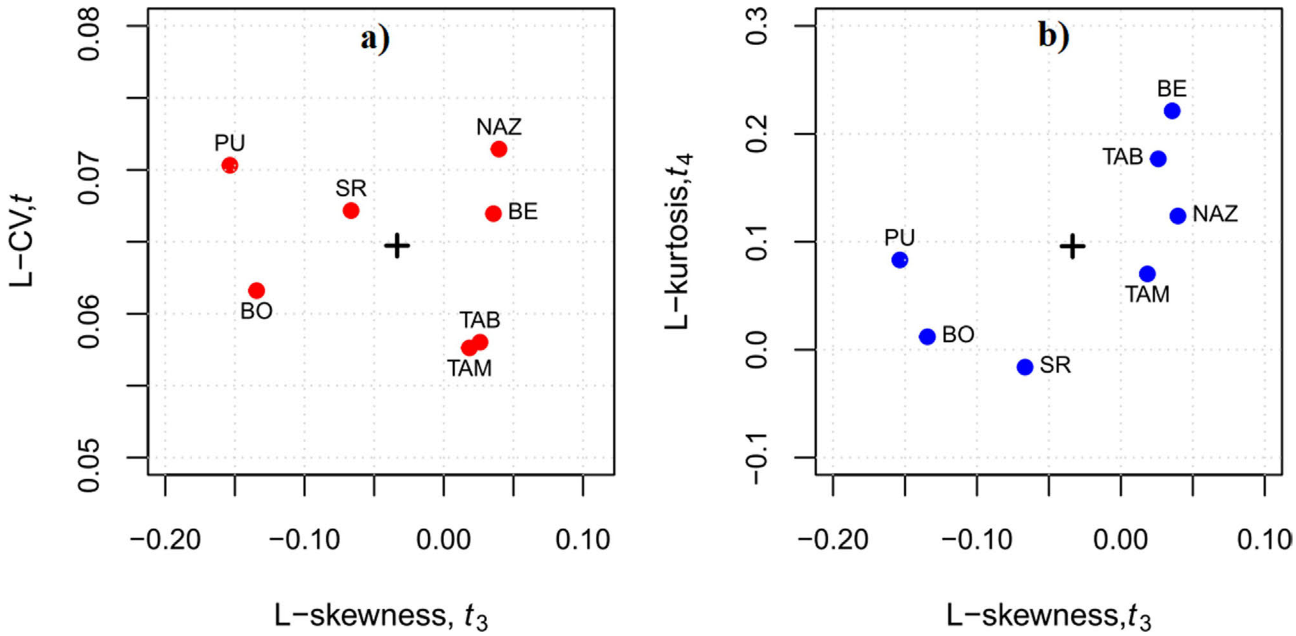

Figure 4 - Dispersion of the L-moments ratios of the samples, a) L-skewness with L-CV and b) L-skewness with L-kurtosis.

Table 3 - Summary of discordancy statistics and regional heterogeneity test for 7 homogeneous basins.

Station	D_i	D_c	H
TAM	0.58		
BE	0.98		
PU	1.32		
TAB	0.50	2.76	-1.25
NAZ	0.25		
BO	0.84		
SR	1.15		

is considered homogeneous if $H < 2$. Viglione (2007) also suggests accepting H values less than 2 as homogeneous to avoid forming groups that are too small. It is confirmed that region 1 remains made up of stations TAM, BE, PU, TAB, NAZ, BO, and SR.

3.2. Selection of regional frequency distribution

GEV, PE3, and GNO respectively are the appropriate distributions for the group of 7 sites. However, the GEV distribution is the best-fitted distribution to the regional average (+ symbol) (Fig. 5a). Peel *et al.* (2001) indicates that the selection of the regional frequency distribution of homogeneous groups is best based on the average of the sample and not on a line of best fit through the data points.

Hosking and Wallis (1997) used the statistical Z goodness-of-fit test ($|Z^{DIST}| \leq 1.64$), to select the best regional frequency distribution, which means that the true distribution of the region should be accepted approximately 90% of the time. Consequently, the Z statistic confirms that the regional distribution function that best fits the group of 7 homogeneous stations is the GEV distribution, followed by PE3 and GNO (Fig. 5a). Then, the estimated parameters of location, scale, and shape of the regional GEV distribution were ε = 0.9638, α = 0.1156 and k = 0.3458, respectively. These regional parameters could be confidently transferred to specific sites (Parida and Moalafhi, 2008). The regional growth curve and error limits for different return periods were elaborated using the GEV distribution function (Fig. 5b). The results indicate that there are higher uncertainty limits when the return period is high. This is also seen in the regional growth curve and peak flows for TAM, TAB, NAZ, BE, BO, PU and SR sites (Figs. 6a-6g) respectively. In periods return high, the uncertainties were larger due to the extrapolation of observed events, both for the regional analysis or on the site. This can represent great risks in hydraulic planning; however, the estimates are within the confidence intervals (Rezende de Souza *et al.*, 2021).

3.3. Regional quantile estimation

GEV was the selected regional distribution function. Consequently, the regional equation to estimate the maximum flows in m^3/s for different return periods in basins with information (Fig. 6) and without information for region 1 is deduced as:

$$Q(T) = \left[0.9638 + \frac{0.1156}{0.3458}\left\{1 - \left(-ln\left(1-\frac{1}{T}\right)\right)^{0.3458}\right\}\right] * \overline{Q} \qquad (24)$$

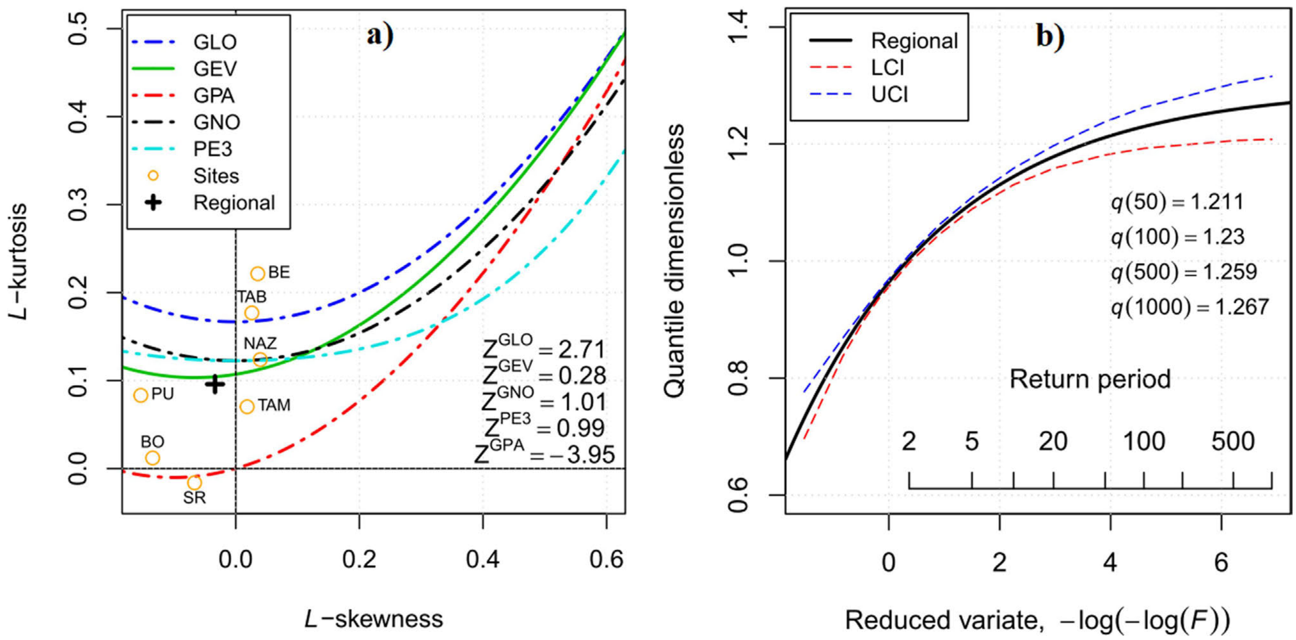

Figure 5 - a) Relationship diagram of L-moments for basins based on site and regional L-moments and b) regional growth curve.

While for the dimensionless quantiles for different return periods it is:

$$\frac{Q(T)}{\overline{Q}} = q(T) = 0.9638 + \frac{0.1156}{0.3458}\left(1 - \left(-ln\left(1 - \frac{1}{T}\right)\right)^{0.3458}\right) \tag{25}$$

3.4. Prediction in ungauged basins

The results of the correlation analysis between the area of the basin and the index-flood, resulted in a statistically significant correlation (r = 0.997) with a confidence level of 95% and a significance level of 5%. Consequently, the basin area explained the variability of $\overline{Q}$ in 99.4% (R^2 = 0.994). These results were corroborated with the significance of the parameter β_1 evaluated on the basis of the t'Student statistic. Thus, considering a significance level of 5%, the result for β_1 gave a p-value < 0.05 (p-value = 7.84E-07), which implies that the changes in A are related to the changes in $\overline{Q}$. On the other hand, the results of homoscedasticity and normality of the residuals evaluated by means of the HMC and AD tests, indicate that the variance of the residuals is constant (p-value = 0.494) and the residuals are normally distributed (p-value = 0.881) since the p-value is greater than the significance level of 5% (p-value > 0.05), therefore, there is evidence to explain the fulfillment of homoscedasticity (Fig. 7a) and normality (Fig. 7b) of the residuals respectively. Viglione (2007) indicates that the presence of particular patterns in the arrangement of the points can be an index of heteroskedasticity (diversity in variance).

In the analysis, climatic variables such as precipitation and mean annual temperature of the basin were also considered, however, no significant correlation with $\overline{Q}$ was found. From the above, the basin area would become the explanatory variable to estimate $\overline{Q}$ in ungauged basins. Viglione (2007) obtains better relationships between the average altitude of the basin, the center of gravity of the basin, and the Budiko index with the flows transformed into logarithms. However, Tucci (2002) finds a better relationship between the area of the basin and the untransformed flows. The regional index-flood equation for region 1 is defined by:

$$\overline{Q} = 6987.6393 + 0.0554(A) \tag{26}$$

The standard error for $\overline{Q}$ was 1565.95 m^3/s. The confidence interval for the adjusted model parameter β_0, varies between 4271.421 and 9703.857 m^3/s while for β_1 it varies between 0.0507 and 0.0602, with a 95% confidence level. Lower and upper confidence intervals for ungauged basins can then be calculated based on variations in these parameters (Fig. 8a). The index-flood estimate turned out to be higher for basins with a larger drainage area, but the estimates are within the upper (UCI) and lower (LCI) confidence intervals, which makes its estimation feasible. The index-flood model is only applicable for basins with areas between 99779.4 km^2 and 877478.6 km^2 as long as they are in region 1, established as a homogeneous region.

On the other hand, making use of Eq. 24, the prediction of maximum design flows (Fig. 8b) presents

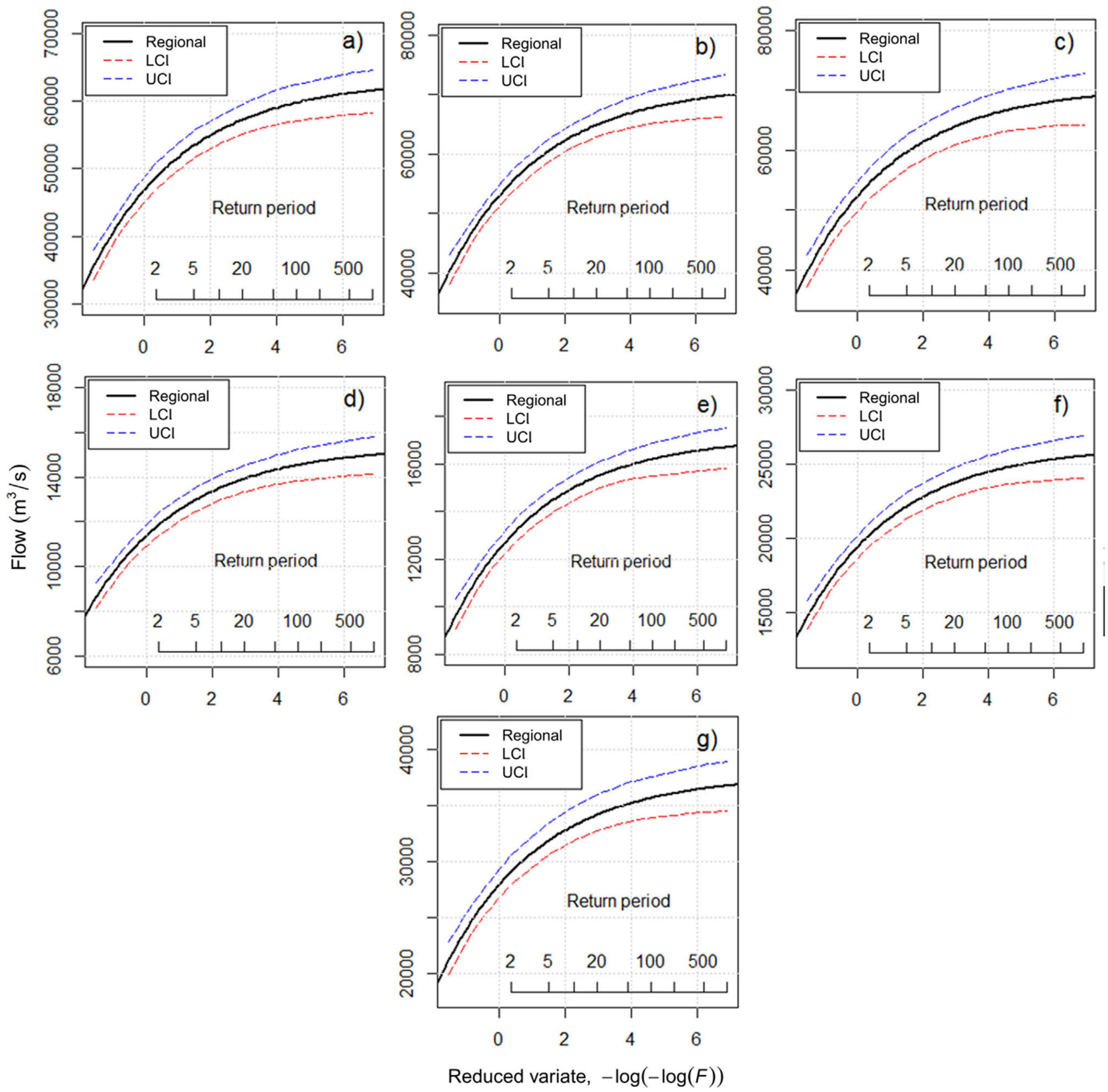

Figure 6 - On-site estimated quantiles with uncertainty limits for the GEV distribution, a) TAM, b) TAB, c) NAZ, d) BE, e) BO, f) PU and g) SR.

wide ranges of uncertainty, mainly for high return periods. However, together with uncertainty limits, they can be useful for hydraulic planning in ungauged basins, limiting their use of the equation to basins with areas between 99779.4 $km^2 \leq A\ (km^2) \leq$ 877478.6 km^2 and could be used with special caution for basins with areas greater than 877478.6 km^2 and less than 99779.4 km^2 within the Amazon basin of Peru. According to Rezende de Souza *et al*. (2021) for flood control and hydraulic structure protection, the most important thing is to consider the upper limit where the maximum flow value is shown, the lower confidence interval being negligible for this approach.

4. Conclusions

The regional frequency analysis was carried out for the prediction of maximum flows in ungauged basins of the Peruvian Amazon. The main conclusions are summarized below:

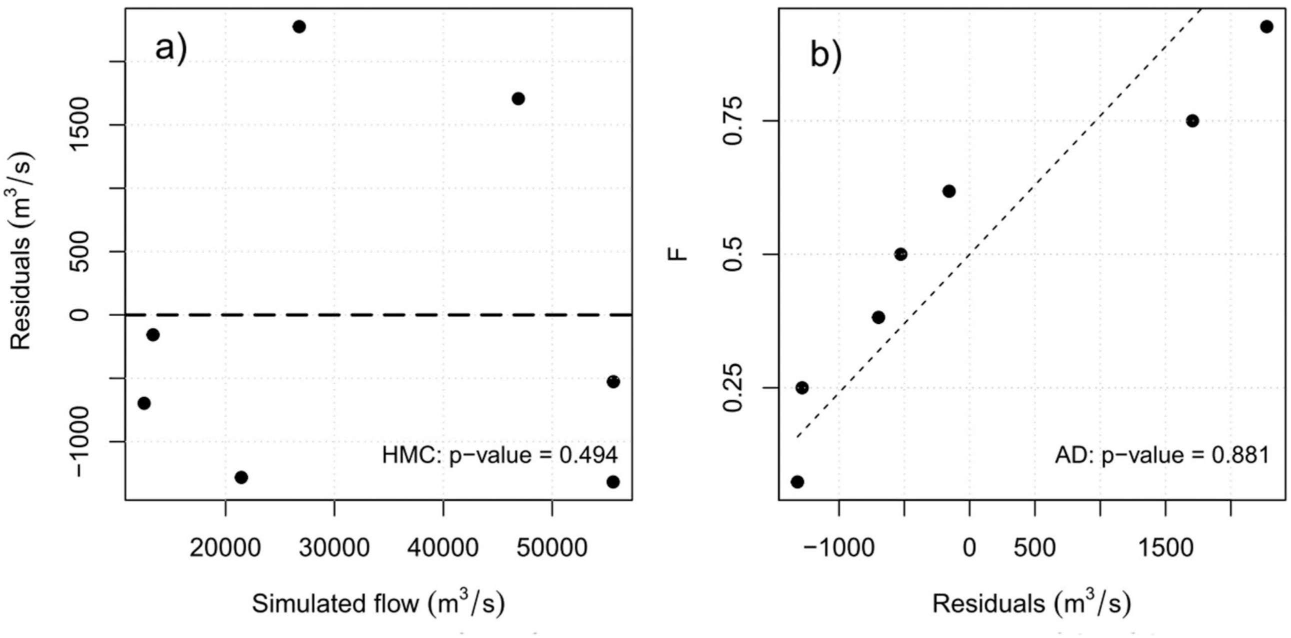

Figure 7 - Verification of assumptions of a) homoscedasticity and b) normality of residuals.

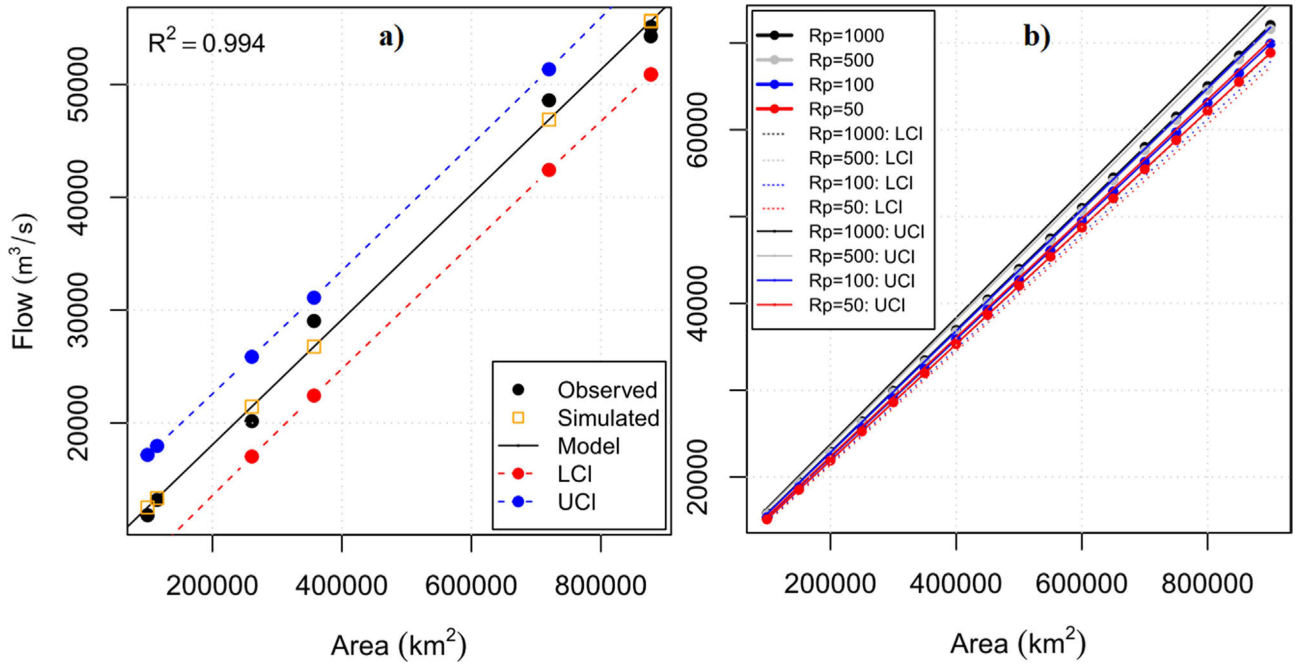

Figure 8 - a) Linear regression relationships between the index-flood and the basin area, b) prediction of maximum flow quantiles in ungauged basins as a function of the basin area.

It was found that, of 10 hydrographic basins analyzed, 7 belong to a region defined as homogeneous. Ward's methods, k-means, Andrews curves, and the heterogeneity test, coincided in the identification of homogeneous regions.

The selection of the regional frequency distribution indicates that the GEV function proved to be more adequate to represent the data sample of region 1, presenting a lower Z^{DIST} value with respect to PE3, GNO, GLO, and GPA.

The index-flood proved to be related to the area of the basin and is the most significant independent variable for the prediction of the index-flood in ungauged basins within region 1. The model found allows the index-flood to be obtained as a function of the basin area and is valid for the area ranges for which they were established.

The maximum flows for different return periods are a function of the regional growth curve and the index-flood. They represent important information that can be used in environmental management, disaster risk management, flood control, hydraulic planning and the design of hydraulic structures in ungauged basins within the Peruvian Amazon for areas within which they were established.

The prediction of maximum design flows in ungauged basins presents wide ranges of uncertainty, mainly for high return periods, therefore, in the estimation of maximum flows, uncertainty limits must be incorporated at all frequencies. Although this can represent great risks in hydraulic planning, the estimates are within the confidence intervals.

Acknowledgments

Our sincere thanks to the National Meteorology and Hydrology Service (SENAMHI) - Peru, EGEMSA, and the SO-HYBAM Observation Service for providing the hydrometric information for the realization of this research study.

References

ACUÑA, J.; FELIPE, O.G.; FERNÁNDEZ, C. Análisis regional de frecuencia de precipitación anual para la determinación de mapas de sequías en las cuencas Chillón, Rímac, Lurín y Alto Mantaro. **Revista Peruana Geo-Atmosférica RPGA**, v. 4, p. 104-115, 2015.

ACUÑA, J.; FELIPE, O.; ORDOÑEZ, J.; ARBOLEDA, F. Análisis regional de frecuencia de precipitación anual para la determinación de mapas de sequías. **Revista Peruana Geo-Atmosférica RPGA**, v. 3, p. 104-115, 2011.

ANDREWS, D.F. 1972. Plots of high-dimensional data. **Biometrics**, v. 28, n. 1, p. 125-36, 1972.

AYBAR, C.; LAVADO-CASIMIRO, W.; HUERTA, A.; FERNÁNDEZ, C.; VEGA, F.; SABINO, E. **Uso del Producto Grillado "PISCO" de Precipitación en Estudios, Investigaciones y Sistemas Operacionales de Monitoreo y Pronóstico Hidrometeorológico**. Lima: SENAMHI-DHI, 2017.

CAMPOS-ARANDA, D.F. Ajuste de las distribuciones GVE, LOG y PAG con momentos L de orden mayor. **Ingeniería, Investigación y Tecnología**, v. 17, n. 1, p. 131-142, 2016.

CUNNANE, C. Methods and merits of regional flood frequency analysis. **Journal of Hydrology**, v. 100, n. 1-3, p. 269-290, 1988.

DALRYMPLE, T. **Flood-Frequency Analyses. Water Supply Paper 1534-A**. Washington: Geological Survey, 1960.

DESAI, S.; OUARDA, T.B.M.J. Regional hydrological frequency analysis at ungauged sites with random forest regression. **Journal of Hydrology**, v. 594, p. 125861, 2021.

DOMROES, M.; KAVIANI, M.; SCHAEFER, D. An Analysis of regional and intra-annual precipitation variability over iran using multivariate statistical methods. **Theoretical and Applied Climatology**, v. 61, n. 3, p. 151-159, 1998.

FERNÁNDEZ-PALOMINO, C.A.; LAVADO-CASIMIRO, W.S. Regional maximum rainfall analysis using L-moments at the Titicaca Lake drainage, Peru. **Theoretical and Applied Climatology**, v. 129, p. 1295-1307, 2016.

FULLER, W.E. Flood flows. **Transactions of the American Society of Civil Engineers**, v. 77, p. 564- 617, 1914.

GOTTSCHALK, L. Hydrological regionalization of Sweden. **Hydrological Sciences Journal**, v. 30, n. 1, p. 65-83, 1985.

GREHYS, G. Inter-comparison of regional flood frequency procedures for Canadian rivers. **Journal of Hydrology**, v. 186, n. 1-4, p. 85-103, 1996.

HARRISON, M.J.; MCCABE, B.P.M. A test for heteroscedasticity based on ordinary least squares residuals. **Journal of the American Statistical Association**, v. 74, p. 494-499, 1979.

HASSAN, B.G.H.; PING, F. Regional rainfall frequency analysis for the Luanhe Basin by using L-moments and cluster techniques. **APCBEE Procedia**, v. 1, p. 126-135, 2012.

HARTIGAN, J.A.; WONG, M.A. Algorithm AS 136: A K-means clustering algorithm. **Applied Statistics**, v. 28, p. 100-108, 1979.

HOSKING, J. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. **Journal of the Royal Statistical Society**, v. 52, p. 105-124, 1990.

HOSKING, J.R.M.; WALLIS, J.R. Some statistics useful in regional frequency analysis. **Water Resources Research**, v. 29, p. 271-281, 1993.

HOSKING, J.R.M.; WALLIS, J.R. **Regional Frequency Analysis: An Approach Based on L-moments**. Cambridge: Cambridge University Press, 1997.

JACKSON, I.J.; WEINAND, H. Classification of tropical rainfall stations: A comparison of clustering techniques. **International Journal of Climatology**, v. 15, n. 9, p. 985-994, 1995.

JARVIS, A.; REUTER, H.I.; NELSON, A.; GUEVARA, E. **Hole-Filled SRTM for the Globe**. Version 4. CGIAR-CSI SRTM 90 m Database, 2008.

KHAN, S.A.; HUSSAIN, I.; HUSSAIN, T.; FAISAL, M.; MUHAMMAD, Y.S.; MOHAMD SHOUKRY, A. Regional frequency analysis of extremes precipitation using L-moments and partial L-moments. **Advances in Meteorology**, v. 2017, p. 1-20, 2017.

KHATTREE, R.; NAIK, D.N. Andrews plots for multivariate data: Some new suggestions and applications. **Journal of Statistical Planning and Inference**, v. 100, p. 411-425, 2002.

KUNDZEWICZ, Z.W.; ROBSON, A.J. Change detection in hydrological records. A review of the methodology/revue méthodologique de la détection de changements dans les chroniques hydrologiques. **Hydrological Sciences Journal**, v. 49, n. 1, p. 7-19, 2004.

KUSWANTO, H.; PUSPA, A.W.; AHMAD, I.S.; HIBATULLAH, F. Drought analysis in East Nusa Tenggara (Indonesia) using regional frequency analysis. **The Scientific World Journal**, v. 2021, p. 1-10, 2021.

LESCESEN, I.; SRAJ, M.; BASARIN, B.; PAVI, D.; MESAROS, M.; MUDELSEE, M. Regional flood frequency analysis of the Sava River in South-Eastern Europe. **Sustainability**, v. 14, p. 1-19, 2022.

LUCAS, C.; MURALEEDHARAN, G.; GUEDES SOARES, C. Regional frequency analysis of extreme waves in a coastal area. **Coastal Engineering**, v. 126, p. 81-95, 2017.

LUJANO, A.; LUJANO, E.; QUISPE, J.P. Regionalización de caudales anuales en cuencas del altiplano peruano. **Revista de Investigaciones Altoandinas**, v. 18, n. 2, p. 189-194, 2016.

LUJANO, A.; QUISPE, J.P.; LUJANO, E. Regionalización de caudales mensuales en la región hidrográfica del Titicaca Perú. **Revista de Investigaciones Altoandinas**, v. 19, n. 2, p. 219-230, 2017.

LUJANO, E.; OBANDO, O.G. Análisis de frecuencia regional de las precipitaciones máximas diarias en la región hidrográfica del Titicaca. **Revista de Investigaciones Altoandinas**, v. 17, n. 1, p. 53-64, 2015.

MSILINI, A.; MASSELOT, P.; OUARDA, T.B.M.J. Regional frequency analysis at ungauged sites with multivariate adaptive regression splines. **Journal of Hydrometeorology**, v. 21, n. 12, p. 2777-2792, 2020.

NAGHETTINI, M.; PINTO, E.J.D.A. **Hidrologia Estatística**. Belo Horizonte: Serviço Geológico do Brasil - CPRM, 2007.

NATHAN, R.J.; MCMAHON, T.A. Identification of homogeneous regions for the purposes of regionalisation. **Journal of Hydrology**, v. 121, n. 1-4, p. 217-238, 1990.

OMM. **Guía de Prácticas Hidrológicas. Gestión de Recursos Hídricos y Aplicación de Prácticas Hidrológicas**. Ginebra: Organización Meteorológica Mundial, 2011.

PARIDA, B.P.; Y MOALAFHI, D.B. Regional rainfall frequency analysis for Botswana using L-Moments and radial basis function network. **Physics and Chemistry of the Earth, Parts A/B/C**, v. 33, n. 8, p. 614-620, 2008.

PEEL, M.C.; WANG, Q.J.; VOGEL, R.M.; MCMAHON, T.A. The utility of L-moment ratio diagrams for selecting a regional probability distribution. **Hydrological Sciences Journal**, v. 46, n. 1, p. 147-155, 2001.

RAMACHANDRA RAO, A.; SRINIVAS, V.V. Regionalization of watersheds by hybrid-cluster analysis. **Journal of Hydrology**, v. 318, n. 1, p. 37-56, 2006.

REZENDE DE SOUZA, G.; MERWADE, V.; COUTINHO DE OLIVEIRA, L.F.; RIBEIRO VIOLA, M.; DE SÁ FARIAS, M. Regional flood frequency analysis and uncertainties: Maximum streamflow estimates in ungauged basins in the region of Lavras, MG, Brazil. **Catena**, v. 197, p. 104970, 2021.

RODRIGUEZ, Y.; MARRENO DE LEÓN, N. Análisis regional de series de lluvias máximas: consideraciones. **Ingeniería Hidráulica y Ambiental**, v. 32, n. 2, p. 34-45, 2011.

SAF, B. Regional flood frequency analysis using L-moments for the West Mediterranean region of Turkey. **Water Resources Management**, v. 23, n. 3, p. 531-551, 2009.

STRNAD, F.; VOJTECH, M.; MARKONIS, Y.; MÁCA, P.; MASNER, J.; STOCES, M.; HANEL, M. An index-flood statistical model for hydrological drought assessment. **Water**, v. 12, n. 4, p. 1-17, 2020.

SCHAEFER, M.G.; BARKER, B.L.; TAYLOR, G.H.; WALLIS, J.R. **Regional Precipitation-Frequency Analysis and Spatial Mapping of Precipitation for 24-Hour and 2 Hour Durations in Eastern Washington**. Washington: MGS Engineering Consultants, Inc., 2006.

TUCCI, C.E.M. **Regionalização de Vazões**. Porto Alegre: Editora da UFRGS, 2002.

VEZZA, P.; COMOGLIO, C.; ROSSO, M.; VIGLIONE, A. Low flows regionalization in North-Western Italy. **Water Resources Management**, v. 24, n. 14, p. 4049-4074, 2010.

VEZZA, P.; COMOGLIO, C.; VIGLIONE, A.; ROSSO, M. The influence of soil characteristics in low flows regionalization. **American Journal of Environmental Sciences**, v. 5, n. 4, p. 536-546, 2009.

VIGLIONE, A. **Metodi Statistici Non-Supervised Per La Stima di Grandezze Idrologiche in Siti Non Strumentati**. Torino: Politecnico di Torino, 2007.

VIGLIONE, A.; CLAPS, P.; LAIO, F. Mean annual runoff estimation in North-Western Italy. **Water Resources Assessment and Management Under Water Scarcity Scenarios**, v. 2, p. 97-122, 2007.

WALLIS, J.R.; SCHAEFER, M.G.; BARKER, B.L.; TAYLOR, G.H. Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington State. **Hydrology and Earth System Sciences**, v. 11, n. 1, p. 415-442, 2007.

WARD, J. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, n. 301, p. 236-244, 1963.

ZAMAN, M.A.; RAHMAN, A.; HADDAD, K. Regional flood frequency analysis in arid regions: A case study for Australia. **Journal of Hydrology**, v. 475, p. 74-83, 2012.

Internet Resources

GOOGLE EARTH ENGINE, https://earthengine.google.com/.

PERUVIAN INTERPOLATED DATA OF THE SENAMHI'S CLIMATOLOGICAL AND HYDROLOGICAL OBSERVATIONS, https://iridl.ldeo.columbia.edu/SOURCES/.SENAMHI/.HSR/.PISCO/.

SO-HYBAM, http://www.ore-hybam.org/index.php/eng.