RBRH

Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 23, e51, 2018
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.231820180079

Regionalization of precipitation with determination of homogeneous regions via fuzzy c-means

Regionalização de precipitação com determinação de regiões homogêneas via agrupamento fuzzy c-means

Evanice Pinheiro Gomes[1], Claudio José Cavalcante Blanco[2] and Francisco Carlos Lira Pessoa[2]

[1]Programa de Pós-graduação em Engenharia Civil, Instituto de Tecnologia, Universidade Federal do Pará, Belém, PA, Brasil

[2]Faculdade de Engenharia Sanitária e Ambiental, Instituto de Tecnologia, Universidade Federal do Pará, Belém, PA, Brasil
E-mails: gomesevanice@ufpa.br (EPG), blanco@ufpa.br (CJCB), fclpessoa@ufpa.br (FCLP)

Received: June 11, 2018 - Revised: August 06, 2018 - Accept: August 27, 2018

ABSTRACT

Knowledge about precipitation is indispensable for hydrological and climatic studies because precipitation subsidizes projects related to water supply, sanitation, drainage, flood and erosion control, reservoirs, agricultural production, hydroelectric facilities, and waterway transportation and other projects. In this context, methodologies are used to estimate precipitation in unmonitored locations. Thus, the objectives of this work are to i) identify homogeneous regions of precipitation in the Tocantins-Araguaia Hydrographic Region (TAHR) via the fuzzy c-means method, ii) regionalize and estimate the probability of occurrence of monthly and annual average precipitation using probability distribution models, and iii) regionalize and estimate the precipitation height using multiple regression models. Three homogeneous regions of precipitation were identified, and the results of the performance indices from the regional models of probability distribution were satisfactory for estimating average monthly and annual precipitation. The results of the regional multiple regression models showed that the annual mean precipitation was satisfactorily estimated. For the average monthly precipitation, the estimates of multiple regression models were only satisfactory when the months used were distributed in the dry and rainy seasons. Therefore, our results show that the methodology developed can be used to estimate precipitation in unmonitored locations in the TAHR.

Keywords: PBM index; Probability distribution models; Multiple regression models; Tocantins-Araguaia Hydrographic Region.

RESUMO

O conhecimento da precipitação é indispensável para estudos hidrológicos e climáticos, que subsidiam projetos de sistemas de abastecimento de água, saneamento e drenagem; controle de inundações, erosão e reservatórios; produção agrícola e hidrelétrica, transporte hidroviário, entre outros. Nesse contexto, buscam-se metodologias para estimar a precipitação em locais sem monitoramento. Assim, os objetivos do trabalho são: i) identificar regiões homogêneas de precipitação na Região Hidrográfica Tocantins Araguaia (RHTA) via método fuzzy c-means; ii) regionalizar e estimar a probabilidade de ocorrências de precipitações médias mensais e anuais através de modelos de distribuição de probabilidades; e iii) regionalizar e estimar lâminas de precipitação através de modelos de regressão múltipla. Nesse caso, foram identificadas 3 regiões homogêneas de precipitação e os resultados dos parâmetros de desempenho dos modelos regionais de distribuição de probabilidades foram satisfatórios para estimativas de precipitações médias mensais e anuais. Os resultados dos modelos regionais de regressão múltipla revelaram que as precipitações médias anuais são estimadas satisfatoriamente. Já no caso de precipitações médias mensais, as estimativas dos modelos de regressão múltipla só foram satisfatórias quando os meses foram distribuídos em secos e chuvosos. Assim, constata-se que a metodologia desenvolvida pode ser aplicada para estimativas de precipitação em locais sem monitoramento da RHTA.

Palavras-chave: Índice PBM; Modelos de distribuição de probabilidades; Modelos de regressão múltipla; Região Hidrográfica Tocantins-Araguaia.

INTRODUCTION

Precipitation is one of the most important hydrological variables. Its scarcity or excess directly affects society, influencing water supply, drainage, flood control and erosion systems, agricultural production, generation of energy, etc. However, precipitation monitoring is generally confined to scattered points, leaving gaps in more isolated and difficult to access areas, which highlights the importance of methods that allow hydrological information to be obtained. Thus, the development of techniques for estimating precipitation has become relevant. Regionalization is a possible technique that can provide hydrological data at low cost. Several works, such as Arellano-Lara and Escalante-Sandoval (2014), Asong, Khaliq and Wheater (2015), Shahana Shirin and Thomas (2016) and Fazel et al. (2018), are examples of the application of precipitation estimates in several regions. Regionalization is a well-known methodology and its importance is related to the obtainment of hydrological information in places without monitoring. In addition, using this technique, the zoning of the earth based on physical and hydrological characteristics can generate a greater understanding of the distribution and intensity of rainfall and streamflow in a specific region.

According to Samuel, Coulibaly and Metcalfe (2011), regionalization consists of the use of a set of methods that attempt to transfer information from one place to another in river basins, for the purpose of filling in missing information in a given region considered homogeneous. To apply precipitation regionalization, mathematical and statistical procedures are applied to the historical data series and to the physical and climatic characteristics of the river basins using hydrological models, which, after being calibrated and validated, are able to estimate the precipitation in the homogeneous regions.

The best known models of precipitation estimates are those created through spatial interpolation, statistical and satellite estimation methods. Models of spatial interpolation include the polygon of Thiessen, the kriging and the isohyetal methods. Among the statistical models, we highlight the probability distribution functions (PDF) and the multiple regression analysis (MRA). Satellite estimates are obtained from observations of the atmosphere, captured by micro waves and transformed into precipitation data by specific algorithms that require advanced technology. Spatial interpolation methods mainly consider precipitation. Mathematical and statistical models, such as those derived from multiple regression models, correlate several of the variables that exert some influence on the element studied to improve the results.

Numerous studies related to the estimation of precipitation and its probability of occurrence, through MRA and PDF, have been published. Chifurira and Chikobvu (2014) developed a simple, predictive model of precipitation using multiple regression, using climatic determinants (southern oscillation and sea level pressures) from Zimbabwe, Africa. This model had a reasonable adjustment at a significance level of 5% and is easily applied. Chatzithomas, Alexandris and Karavitis (2015) used multiple regression models to estimate the annual and monthly means of precipitation in the Viotikos Kefissos basin in Ecuador. In this study, the authors used 17 rainfall gauge stations, three independent variables (elevation, location and direction of storms), verifying that the regression models had excellent results when compared with the kriging method. Das and Umamahesh (2016) used a multiple regression model constructed with main components and fuzzy clusters that estimated the behavior of precipitation between 2008 and 2100, and found good results for the Godavari basin in India.

Li, Brissette and Chen (2014) evaluated the performance of six distributions of precipitation probability (exponential, gamma, Weibull, normal, mixed exponential and hybrid exponents) from the Loess Plateau in China, identifying the normal function as the best with which to simulate the distributions of monthly and annual frequency. Yuan et al. (2018) tested five different probability distribution functions to predict the distribution of the occurrence of the maximum hourly annual precipitation. The quality of the fit was assessed using the chi-square test, which indicated that the log-Pearson function had the best overall fit for the maximum hourly annual precipitation from most regions of Japan.

Thus, regionalization and precipitation estimates are the main objectives of this study, which is motivated by the regions of the Amazon that still lack rainfall gauge stations with long series of records. An example of one of these regions is the TAHR. In this case, the homogeneous regions were determined via the fuzzy c-means clustering technique. Probability distribution functions and regional models, determined through multiple regression models, were employed for precipitation height estimates.

MATERIAL AND METHODS

Study area

The TAHR is located between 0º 30 'and 18º 05' south and 45º 45 'and 56º 20' west (Figure 1). It has an elongated configuration, with a south-north direction, following the predominant direction of the main watercourses, the Tocantins and Araguaia Rivers, which intersect in the northern part of the region, from which point they are called the Tocantins River, which empties into the Marajó Bay. The total area of the TAHR is 918,822 km^2, covering part of the midwestern, northern and northeastern regions. The TAHR occupies 11% of the national territory and includes the states of Goiás (21.4%), Tocantins (30.2%), Pará (30.3%),

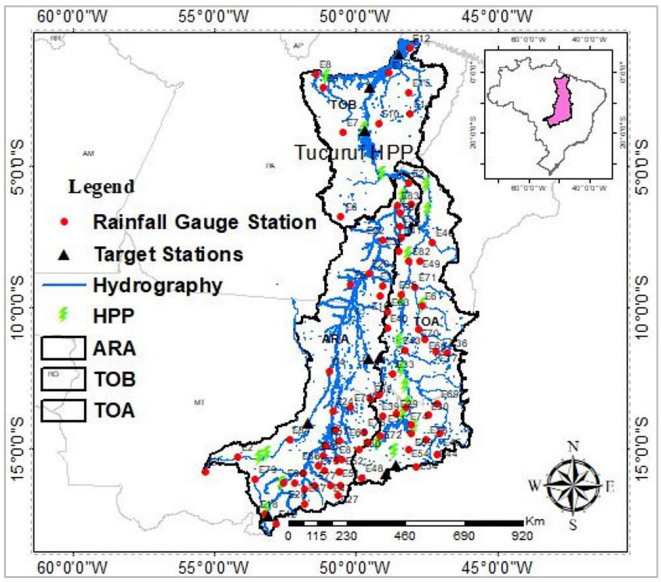


Figure 1. Tocantins Araguaia Hydrographic Region (TAHR).

Maranhão (3.3%), Mato Grosso 14.7%) and the Federal District (0.1%). This region is divided into three subbasins: Alto Tocantins (TOA), Baixo Tocantins (TOB) and Araguaia (ARA), a division adopted by the National Council of Water Resources.

The TAHR has great importance for the development of the country since it provides electricity for the Brazil, through the Hydroelectric Power Plant (HPP) of Tucuruí, and is important for mining, agribusiness, agriculture and livestock farming. According to studies conducted by the National Water Agency (ANA, 2006), the average annual precipitation is approximately 1,837 mm, and the rate of flow is approximately 13,624 m^3/s; the evapotranspiration is 1,371 mm, representing 75% of the precipitation (the average annual evapotranspiration of the country is 1,134 mm or 63% of the precipitation); and the average coefficient of the surface flow is 0.30. According to ANA (2016a), 109.5 thousand hectares of irrigable areas were registered in this region in 2014 (Figure 2). The most relevant land use and occupation activities are categorized into urbanized areas, crops, forests, pastures and agricultural establishments (Figure 3).

Data sources

Precipitation data from 92 stations located at TAHR in the ANA database (ANA, 2016b) were used (Table 1). The stations were chosen based on the historical series; the chosen stations had the largest data series. Despite flaws found in the daily series, the annual and monthly accumulated data was not compromised. The data consistency methodology adopted by ANA (2012) prioritizes the degree of homogeneity of the data, correcting possible errors.

To calibrate the models used in the regionalization, 83 stations were used and in the validation, 9 target stations were used (Figure 1). Altitude information and station coordinates are available in the ANA database. The mean annual precipitation (P), altitude (H), latitude (la) and longitude (lo) of each rainfall gauge station were used to identify the homogeneous regions of precipitation and to develop regional models of precipitation estimation. Of the 92 stations used, 70 have 30 years of data (1975-2004), and the remaining 22 include 17 and 28 years.

Homogeneous regions

One of the conditions necessary for the application of regionalization is the identification of homogeneous regions, which are associated with regions that have hydrological similarities. The identification of hydrologically homogeneous regions has two purposes: to impose boundaries between regions and to hydrologically characterize the regions. The identification of homogeneous regions can be performed in several ways. However, the most widely adopted method in hydrological and environmental studies is cluster analysis. The applications developed by Satyanarayana and Srinivas (2011), Dikbas et al. (2011), Santos, Lucio and Silva (2014),

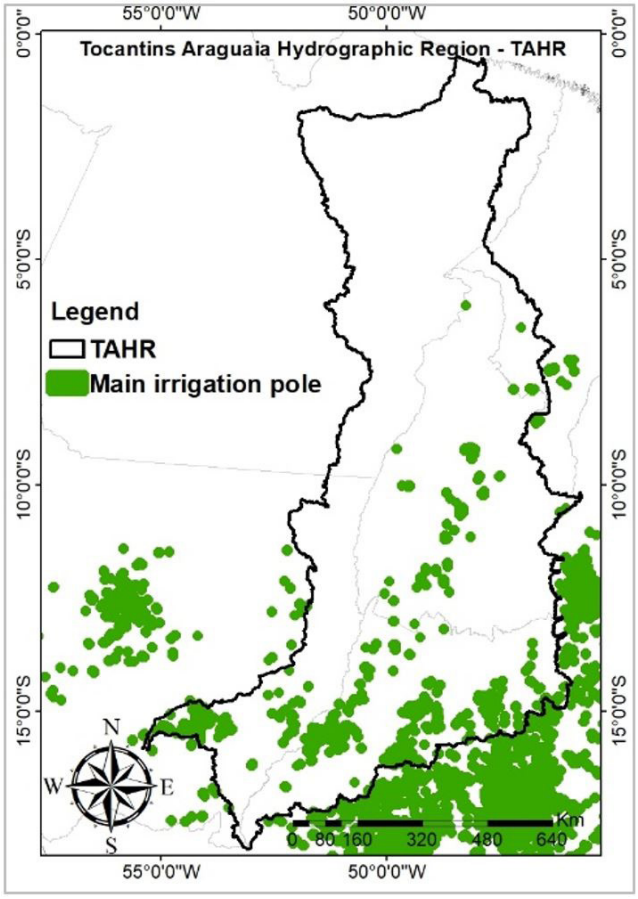


Figure 2. Irrigable Areas on TAHR (Source: ANA, 2016a).

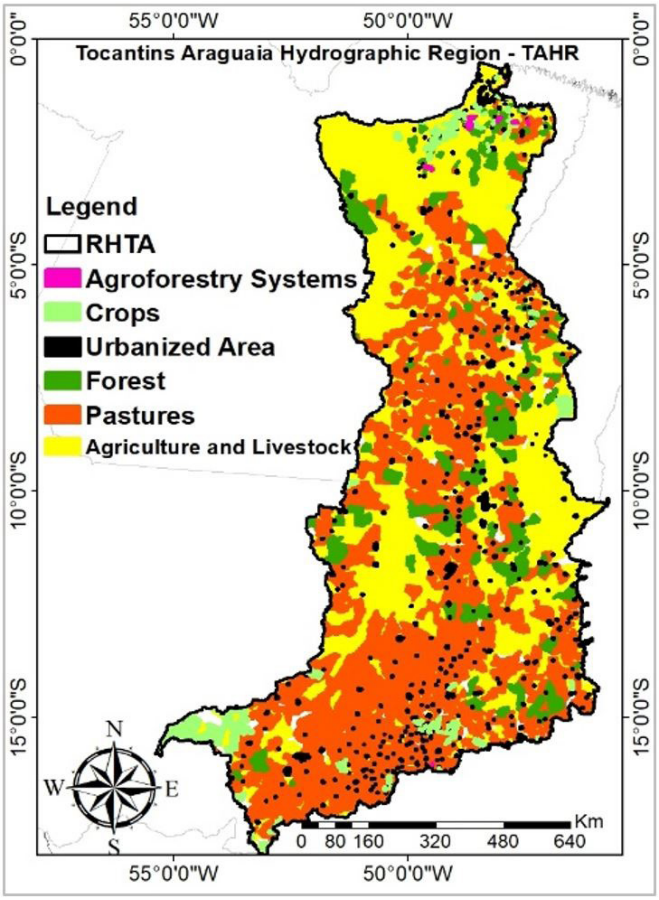


Figure 3. TAHR Soil Uses (Source: IBGE, 2014).

Table 1. TAHR rainfall gauge stations considered in the study.

ID	Station	Code	P (mm)	Alt (m)	Lat	Lon
E1	São José da Serra	01555005	1614	797	-15.8	-55.3
E2	Rio das Mortes	01554005	1690	551	-15.3	-54.2
E3	Alô Brasil	01251000	1658	339	-12.2	-51.7
E4	Santo Antônio Leverger	01250001	1581	205	-12.3	-51
E5	Barreira do Campo	00950001	1417	195	-9.2	-50.2
E6	Fazenda Caiçara	00650001	1730	95	-6.8	-50.5
E7	Faz. Estrela Norte	00350000	1931	22	-3.9	-50.5
E8	Acampamento IBDF	00151001	2012	11	-1.8	-51.4
E9	Maracacuera Florestal	00251000	2620	20	-2.2	-51.2
E10	Cachoeira Tracambeua	00349001	2382	50	-3.5	-49.2
E11	Abaetetuba	00148010	2584	13	-1.7	-48.9
E12	Vigia	00048006	2843	15	-0.9	-48.1
E13	Faz. Maringa	00348001	1933	20	-3.2	-48.1
E14	Tomé-Açu	00248003	2553	45	-2.4	-48.1
E15	Abreulandia	00949000	2117	240	-9.6	-49.2
E16	Almas	01147000	1524	427	-11.6	-47.2
E17	Alto Araguaia	01753000	1681	659	-17.3	-53.2
E18	Ananas	00648001	1562	191	-6.4	-48.1
E19	Araguacema	00849002	2048	203	-8.8	-49.6
E20	Araguatins	00548000	1552	122	-5.6	-48.1
E21	Arapoema	00749000	1867	215	-7.7	-49.1
E22	Aruanã	01451000	1537	200	-14.9	-51.1
E23	Bandeirantes	01350000	1456	276	-13.7	-50.8
E24	Bom Jardim de Goiás	01652000	1650	402	-16.2	-52.2
E25	Britânia	01551000	1417	297	-15.2	-51.2
E26	Cachoeira GO	01650000	1514	766	-16.7	-50.6
E27	Caiaponia	01651000	1632	713	-16.9	-51.8
E28	Campinaçu	01348000	2441	683	-13.8	-48.6
E29	Cavalcante	1347000	1850	821	-13.8	-47.5
E30	Colinas do Sul	01448000	1573	530	-14.2	-48.1
E31	Colinas TO	00848000	1801	229	-8.1	-48.5
E32	Colonha	01248001	1420	264	-12.4	-48.7
E33	Contagem	01547010	1570	1242	-15.7	-47.9
E34	Córrego do ouro	01650001	1544	569	-16.3	-50.6
E35	Dianópolis	01146000	1449	679	-11.6	-46.8
E36	Dois Irmãos Tocantins	00949001	2029	264	-9.3	-49.1
E37	Entroncamento S M	01349003	1635	345	-13.1	-49.2
E38	Estrela do Norte	01349000	1751	467	-13.9	-49.1
E39	Fátima	01048000	1897	352	-10.8	-48.9
E40	Faz Primavera	00748002	1816	257	-7.6	-48.4
E41	Faz São Bernardo	01752002	1674	750	-17.7	-52.8
E42	Faz. Lobeira	01148000	1556	243	-11.5	-48.3
E43	Faz. Santa sé	01547001	1684	573	-15.2	-47.2
E44	Flores GO	01447001	1144	200	-14.5	-47
E45	Goiantins	00747001	1572	185	-7.7	-47.3
E46	Israelândia	01650002	1597	406	-16.3	-50.9
E47	Itaberaí	01649007	1828	726	-16	-49.8
E48	Itacaja	00847001	1845	250	-8.4	-47.8
E49	Itapirapua	01550000	1589	343	-15.8	-50.6
E50	Itapuranga	01549002	1645	646	-15.8	-50.6
E51	Jeroaquara	01550001	1780	400	-15.4	-50.5
E52	Lagoa da Flexa	01450000	1436	200	-14.3	-50.7
E53	Mimoso	01548001	1308	687	-15.1	-48.2
E54	Miracema Tocantins	00948000	1707	210	-9.6	-48.4
E55	Monte Carlos GO	01551001	1543	400	-15.6	-51.4
E56	Mozarlandia	01450001	1654	400	-14.7	-50.6

*Target Stations; TAHR – Tocantins-Araguaia Hydrographic Region.

Table 1. Continued...

ID	Station	Code	P (mm)	Alt (m)	Lat	Lon
E57	Muricilandia	00748003	1671	393	-7.2	-48.5
E58	Niquelândia	01448001	1704	568	-14.5	-48
E59	Nova América	01549004	1606	800	-15	-49.9
E60	Novo Acordo	01047001	1598	300	-10	-47.7
E61	Novo Planalto	01349001	1588	286	-13.2	-49.5
E62	Paraíso do TO	01048001	2281	390	-10.2	-48.9
E63	Perez	01551002	1499	299	-15.9	-51.9
E64	Pilar de Goiás	01449000	1948	765	-14.8	-49.6
E65	Pindorama do Tocantins	01147002	1615	444	-11.1	-47.6
E66	Piranhas	01651002	1583	356	-16.4	-51.8
E67	Piraquê	00648002	1761	184	-6.7	-48.5
E68	Ponte Paranã	01347001	1245	363	-13.4	-47.1
E69	Porto Gilandia	01047002	1656	220	-10.8	-47.8
E70	Porto Real	00948001	1599	200	-9.3	-47.9
E71	Porto Uruaçu	01449001	1468	572	-14.6	-49.1
E72	Rio Pintado	01350001	1444	200	-13.5	-50.2
E73	Sama	01348001	1411	375	-13.5	-48.2
E74	Santa fé	01551003	1615	400	-15.8	-51.1
E75	Santa Terezinha GO	01449002	1505	400	-14.4	-49.7
E76	São Ferreira	01651003	1673	361	-16.3	-51.5
E77	São João Aliança	01447002	1499	1009	-14.7	-47.5
E78	Tesouro	01653000	1715	389	-16.1	-53.5
E79	Torixoreu	01652002	1406	307	-16.2	-52.6
E80	Travessão	01550002	1517	450	-15.4	-50.7
E81	Tupiratins	00848003	1740	192	-8.4	-48.1
E82	Xambioá	00648000	1695	148	-6.4	-48.5
E83	Xavantina	01452000	1526	263	-14.7	-52.4
E84	Tucuruí*	01449000	2422	40	-3.8	-49.7
E85	Cametá*	01147002	2590	24	-2.2	-49.5
E86	Belém*	01651002	2943	10	-1.5	-48.5
E87	Trecho Médio*	00648002	1555	232	-14.1	-51.7
E88	Gurupi*	01347001	1497	353	-11.7	-49.1
E89	Formosa* do Araguaia	01047002	1708	247	-11.8	-49.5
E90	Faz. Marajá*	00948001	1498	666	-15.6	-48.6
E91	Pirenópolis*	01449001	1687	740	-15.9	-49
E92	Faz. Babilônia*	01350001	1632	699	-17.4	-53.1

*Target Stations; TAHR – Tocantins-Araguaia Hydrographic Region.

Farsadnia et al. (2014), Parracho, Melo-Gonçalves and Rocha (2015), Awan, Bae and Kim (2015), Latt, Wittenberg and Urban (2015) and Pessoa, Blanco and Gomes (2018) are examples of the successful use of cluster analysis to identify hydrologically homogeneous regions, demonstrating their significant efficacy.

Fuzzy c-means (FCM)

The nonhierarchical fuzzy c-means method was initially proposed by Dunn (1973) and then generalized by Bezdek (1981). Known as fuzzy clustering, it is based on the premise that a set can be grouped into *p* groups by the degree of membership that each element has to one or more sets. The fuzzy c-means group is generated by minimizing the objective function (Equation 1) and by iteratively performing the algorithm (FCM), which indicates the degree of membership of an element to a given cluster group. Therefore, technique, each element belongs to a group with a certain degree of pertinence, which requiring an initial estimate of the number of groups.

$$J = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(u_{ij}\right)^{m} d\left(X_i, C_j\right)^2 \quad (1)$$

where n is the number of data points; p is the number of groups; u_{ij} is the degree of relevance of the sample Xi to the j-th cluster; m is the fuzzy parameter; d is the Euclidean distance between Xi and Cj; Xi is data vector, with $i = 1, 2,..., n$, representing a data attribute; and Cj is the center of a fuzzy cluster.

The fuzzy parameter (m) is also known as the fuzzy weight exponent, and is the parameter that controls the level of diffusivity in the classification process. The cluster decision is defined by the greater degree of relevance presented for each element analyzed. Thus, for a given Xi, its greater degree of pertinence, determines which group this object belongs to.

PBM index

The PBM index proposed by Pakhira, Bandyopadhyay and Maulik (2004), which is an acronym of the initials of the authors' names, serves to validate the number of clusters or subsets formed from a set of data by evaluating whether the clusters are well defined and separated. The PBM index is a maximization parameter; therefore, the higher its value, the better the quality of the partition is. It is defined as the product of three factors (Equation 2) and its maximization ensures that the partition has a small number of compact groups with a large separation between at least two of them.

$$PBM(K) = \left(\frac{1}{k} . \frac{E_1}{E_k} . D_k \right)^2 \quad (2)$$

where K is the number of clusters; E_1 is the sum of the distances of each sample to the geometric center of all samples; E_k is the sum of the distances between the groups and D_k represents the maximum separation of each pair of groupings.

Heterogeneity test (H)

The measurement of H (Equation 3) which is used in hydrology and meteorology, was proposed by Hosking and Wallis (1993) and aims to verify the degree of heterogeneity of a region by comparing the observed variability to the expected variability of a homogeneous region based in L-moment statistics. H helps verify the homogeneity of the regions formed in the cluster.

$$H = \frac{(V - \mu_v)}{\sigma_v} \quad (3)$$

where V is the weighted standard deviation, μv is the arithmetic mean of the statistics Vj, obtained by simulation and σv is the standard deviation of the dispersion measure of the estimated samples. According to a test of significance, if $H < 1$, the region is considered to be "acceptably homogeneous," if $1 \leq H < 2$, the region is "possibly homogeneous" and finally if $H \geq 2$, the region must be classified as "definitely heterogeneous."

Probability Distribution Functions – PDF

In hydrology, the PDFs produces a projection of what will happen in the future, based on the frequency of past occurrences. Thus, to model the frequency of hydrological data, it is necessary to study its occurrence and to establish whether the variable can be larger or smaller than a given value. Several probability distribution functions have been used to verify precipitation behavior and variability. Among these, we use the normal, gamma two parameters, log-normal and Weibull distributions because they show good adjustments of monthly and annual precipitation totals and some of them are highlighted in the publications of Li, Brissette and Chen (2014), Caldeira et al. (2015), Yuan et al. (2018).

The chi-square test (X^2) was used to select the PDF that best fit the probability values of monthly and annual precipitation. The choice of this test is justified because it is the most commonly used to test frequency distributions. In the calibration of the PDF, simulations were carried out using a computer code called PDF, created to generate the occurrence frequencies of annual and monthly average precipitation heights of each station in the homogeneous regions formed by the fuzzy c-means cluster. The PDFs selected in the calibration evaluated by their fit in the 9 target stations, which were not adopted in the calibration step. Thus, the frequency distribution of the target stations was determined by the best PDF obtained in the calibration.

Adhesion test - Chi-square (X^2)

The chi-square test (Equation 4) was used to select the best probability function, adjusted to the observed data. The test is based on the comparison of the sum of the square of the deviations to the observed and estimated frequencies. In this work, the application of the chi-square test considered the number of degrees of freedom to be equal to two; and the level of significance to be equal to 5%, since these are the most usual values used in the application of this test. Thus, the value of the X^2 is equal to 5.99 for all functions. For the probability distribution to be considered adequate, the calculated value of X^2 must be smaller than the table (CORDER; FOREMAN, 2009).

$$X^2 = \Sigma \left[\frac{(f_0 - f_e)^2}{f_e} \right] \quad (4)$$

where f_o is the frequency observed (mm); and f_e is the frequency (mm) estimated by the probability function.

Multiple regression models

According to Hair et al. (2005), this technique can be used to verify the relationship between a single dependent variable and several independent variables. The objective of this method is to use the independent variables, whose values are known, to predict the values of the dependent variable studied. The relationship between the dependent variable and the independent variables can be represented by a linear model (Equation 5).

$$Y = \beta_o + \beta_1 . X_1 + \beta_2 . X_2 + \beta_i . X_i + \varepsilon \quad (5)$$

where Y is the dependent or predicted variable, X_1, X_2,…Xi, are the independent or explanatory variables. β_o, β_1, β_2...β_3, are the regression coefficients, and ε denotes the residuals of the regression. In the determination of the dependent variable (Y), represented by the precipitation (P), the multiple regression method was applied between the independent variables (elevation - H, latitude - la, and longitude - lo). For the determination of the parameters β_o, β_1, β_2 and β_3, the least squares method was adopted. Thus, precipitation was determined by the following regression models: linear (Equation 6), potential (Equation 7), exponential (Equation 8) and logarithm (Equation 9).

$$P = \beta_o + \beta_1 . H + \beta_2 . la + \beta_3 . lo \quad (6)$$

$$P = \beta_o + H^{\beta_1} + la^{\beta_2} + lo^{\beta_3} \quad (7)$$

$$P = e^{\beta_o + \beta_1 .H + \beta_2 .la + \beta_3 .lo} \quad (8)$$

$$P = \beta_o + \beta_1 .\ln(H) + \beta_2 .\ln(la) + \beta_3 .ln(lo) \quad (9)$$

These models were chosen because they are successful in estimating hydrological variables. In most studies involving regression models, we only observe the use of the variables latitude, longitude and altitude, which are most often available. However, this does not inhibit the success of satisfactory results in the estimation of precipitation, as in, for example, the work of Teixeira-Gandra, Damé and Simonete (2015) and Chatzithomas, Alexandris and Karavitis (2015).

Performance criteria

In the calibration of the regression models, the mean annual and monthly precipitation values at the rainfall stations of the formed groups were used. To evaluate the proposed regression models, we chose the performance criteria presented in Table 2. According to Nash and Sutcliffe (1970) and Rencher and Christensen (2012), the coefficient of determination (R^2) and Nash are equivalent, and the R^2 value varies between 0 and 1. An R^2 value of 9 indicates that 90% of the total variability in the response variable is accounted for by the independent variables. The root mean squared error (RMSE) corresponds to the mean magnitude of the estimated errors. According to Chai and Draxler (2014), the closer the value is to zero, the higher the quality of the estimated values. The percentage relative error, E (%), and the mean relative root square error, ε (%), are coefficients used in several areas of science. According to Jose (2017), the first evaluates the performance of the model, considering the percentage difference between the values of the observed estimated variables, and the second prioritizes the adjustment of the relative values, using the weight of values higher or lower. These coefficients are the most used in the applications of prediction models of hydrological variables, as observed in Mekanik et al. (2013), Chifurira and Chikobvu (2014), Supriya, Krishnaveni and Subbulakshmi (2015), Chatzithomas, Alexandris and Karavitis (2015) and Das and Umamahesh (2016).

For validation, 9 target stations were adopted. Based on the location and altitude data, the precipitation was estimated by applying the regression model, defined in the calibration. Thus, it was possible to compare observed and estimated mean annual and monthly precipitation data of each target station. The estimated data were obtained by the regression model. The mean percentage relative error, E (%) (Table 2) was used as a reference in the validation of the performance of the regression models since the evaluation considers the observed and estimated values, allowing a more direct and objective analysis.

RESULTS AND DISCUSSION

Homogeneous regions

In the formation of homogeneous regions, 63 clusters were performed, changing the fuzzification parameter to the range of 1.2 to 2.0 and the number of groups to 2 to 15. However, it was observed that the larger the number of groups was, the lower the value of the PBM index. Tests with up to 8 groups were considered since the PBM index would tend to decrease with clusters larger than 8. The choice of the best cluster was decided by the PBM index, which presented a higher index (Figure 4) in the formation of three groups with a fuzzing parameter equal to 1.9.

The groups formed represent the homogeneous regions of precipitation (Figure 5). Region I is formed by 52 stations, Region II is formed by 21 stations and Region III is formed by 10 stations. Regions I and II present average annual precipitation ratios of

Table 2. Performance criteria of multiple regression models.

Coefficients	
F - Test F of Significance. If it is < 5% the model is useful.	
$R^2 = \frac{SQReg}{SQT} = \frac{[\hat{\beta}]^T [Y]^T [Y] - n.\hat{Y}^2}{[Y]^T [Y] - n.\hat{Y}^2}$	Determination coefficient[1]
$R_a^2 = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right)$	Adjusted coefficient of determination[2]
$E = \left(\frac{p_o - p_e}{p_o}\right) * 100$	Percentage relative error[3]
$\hat{a} = N^{-1}\left[\sum_{i=1}^{N}(\frac{p_o - p_e}{p_o})^2\right]^{1/2}$	Mean relative root square error[4]
$NASH = 1 - \frac{\Sigma(p_o - p_e)^2}{\Sigma(p_o - \bar{X})^2}$	Nash coefficient[5]
$RMSE = \sqrt{\frac{1}{n} * \Sigma(p_o - p_e)^2}$	Root mean Ssquared error[6]

[1]SQReg = sum of the regression squares, SQT = sum of total squares, n = sample size, Y = independent variables; [2,6] n = sample size, p = number of independent variables; [3,4,5,6] p_o = observed precipitation (m^3/s); p_e = estimated precipitation (m^3/s); [4] N = sample size; [5] $\bar{X}$ = average observed precipitation.

1,600 and 1,700 mm, respectively, while Region III presents an index of approximately 2,400 mm.

Studies by Loureiro, Fernandes and Ishihara (2015), which used geostatistical interpolation in the region, identified that the precipitation totals decrease from north to south but did not define homogeneous regions. In the present work, in addition to confirming this result, it was possible to define three homogeneous regions by the fuzzy c-means clustering. In the verification of the heterogeneity test (H), the value of 0.047 was obtained for Region I, -0.0049 for Region II and -0.7874 for Region III, conferring acceptably homogeneous regions, since H <1.

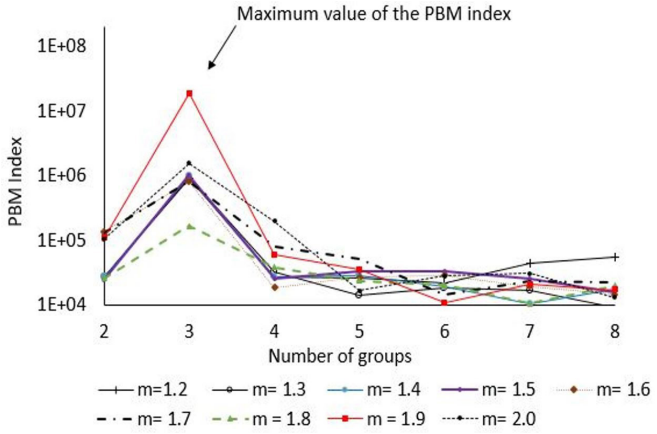


Figure 4. PBM index as a function of the number of groups.

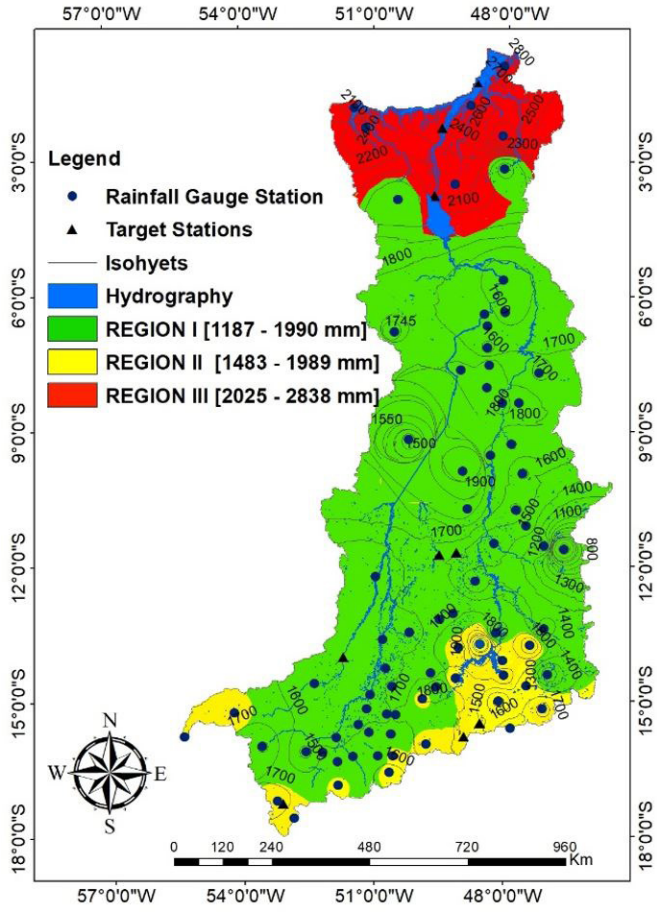


Figure 5. Homogeneous Regions of TAHR Precipitation.

PDF applied to annual average precipitation

The PDFs from normal, log-normal, gamma (two parameters) and Weibull distributions had good adherence in the chi-square test since their values were all below the table value of 5.99, as can be observed in Table 3.

However, the log-normal distribution showed better graphic adjustment between the frequencies observed and estimated. Thus, the log-normal function is the most appropriate model for estimating the probability of occurrence of annual precipitation in homogeneous regions I, II and III of the TAHR.

To validate the log-normal function in homogeneous regions, 9 target stations, three per homogeneous region, were tested using the chi-square test. The test values are below 5.99 (Table 4), validating the log-normal function. The graphical analysis of Figure 6 shows the good adjustment of the probability of occurrence of annual mean precipitation at the target stations in the TAHR. According to Naghettini and Pinto (2007), because the log-normal variable is positive and has a nonfixed asymmetry coefficient greater than zero, this distribution has a parametric form that is adequate to estimate precipitation heights monthly, quarterly or annually.

PDF applied to monthly average precipitation

The average monthly precipitation probabilities of each region were evaluated for adherence to the probability models (normal, log-normal, gamma and Weibull) by the chi-square test. The results of the chi-square test (Table 5) show that the gamma

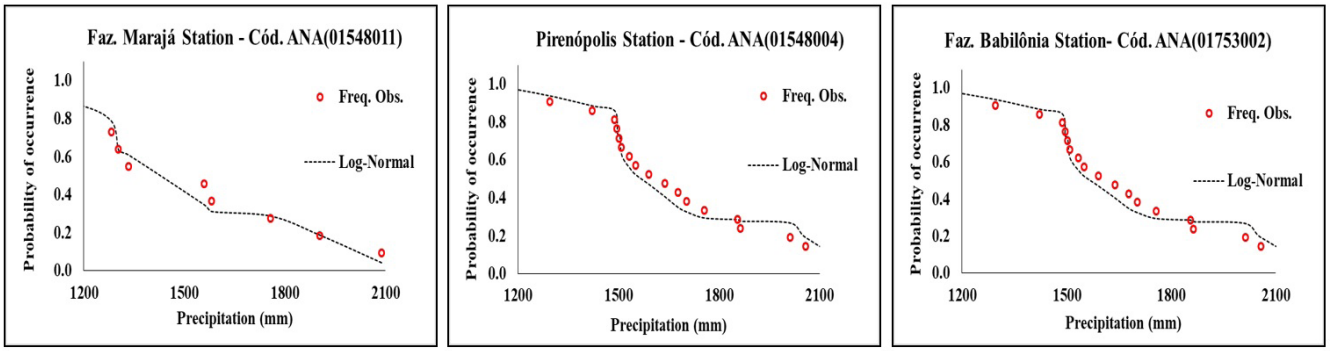
Table 3. Chi-square test for the mean annual precipitation probability functions.

H. R.	Result of the chi-square			
	Normal	Log-Normal	Gamma	Weibull
I	0.79	0.54	3.57	2.93
II	0.46	0.36	4.39	3.95
III	0.09	0.04	5.18	3.08

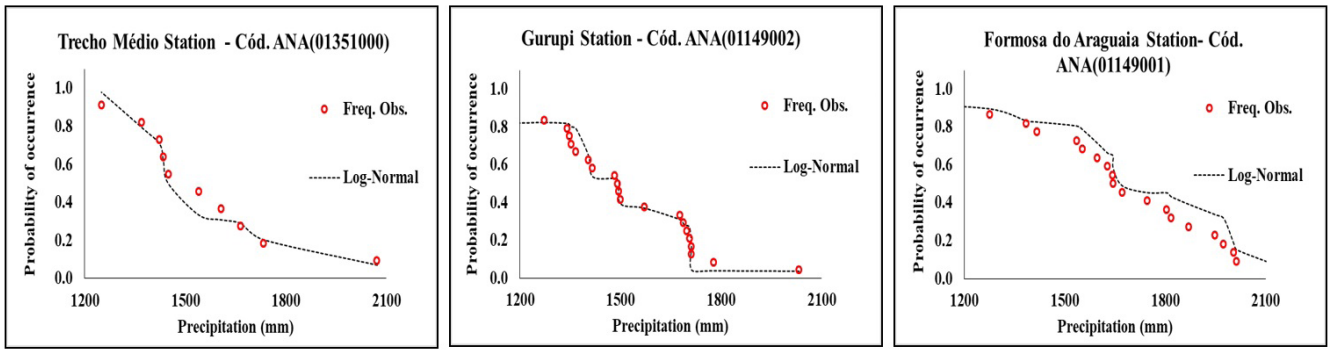
H. R. – Homogeneous Region.

Table 4. Chi-square values in the validation of the log-normal function for the annual series.

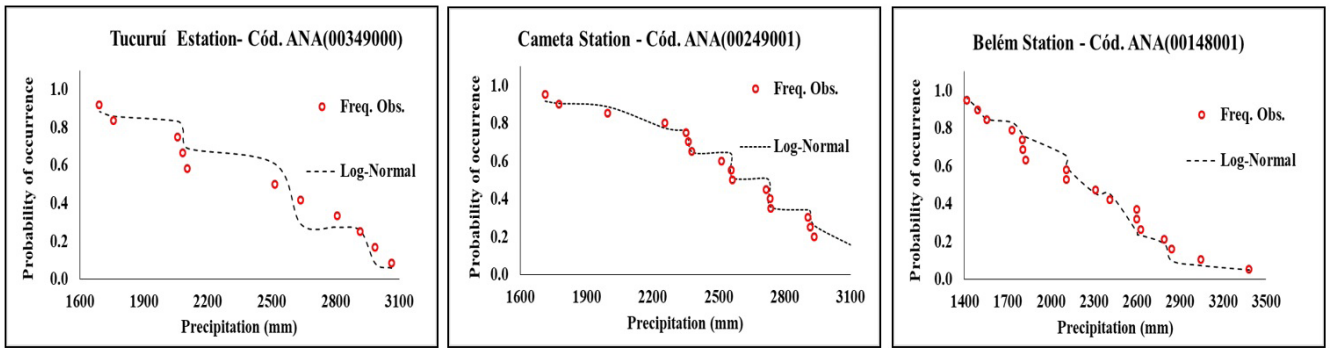
Region	Target Stations	X^2
		Log-Normal
I	Faz. Marajá	1.70
	Pirenópolis	0.82
	Faz. Babilônia	1.12
II	Trecho Médio	2.95
	Gurupi	0.76
	Formosa do Araguaia	1.80
III	Tucuruí	2.37
	Cametá	0.59
	Belém	2.96



(a) Target Stations of the Homogeneous Region I



(b) Target Stations of the Homogeneous Region II



(c) Target Stations of the Homogeneous Region III

Figure 6. Probability of occurrence of observed and estimated annual mean precipitation at the target stations.

Table 5. Chi-square test with PDFs – probability distribution functions.

	Normal			Log Normal			Gamma			Weibull		
	HR I	**HR II**	**HR III**	**HR I**	**HR II**	**HR IIII**	**HR I**	**HR II**	**HR III**	**HR I**	**HR II**	**HR III**
Jan	0.89	0.85	0.083	0.52	0.52	0.22	0.72	0.56	0.17	2.32	2.16	0.005
Feb	3.85	2.17	0.016	3.1	2.43	0.06	3.31	2.36	0.04	6.7*	1.97	1.23
Mar	1.06	1.29	0.05	1.61	1.14	0.23	1.02	1.21	0.14	1.31	1.92	0.09
Apr	10.11*	3.01	6.5*	6.15*	2.76	8.07*	5.67	2.48	7.7*	5.74	3.86	5.94
May	15.0*	1.68	0.37	4.64	2.54	5.58	1.16	2.14	1.9	1.63	1.3	1.52
June	46.58*	3.41	0.196	2.23	1.84	8.26*	2.14	1.83	1.76	1.44	1.53	1.56
July	49.58*	7.59*	0.38	2.52	7.94*	12.5*	2.02	6.9*	3.22	1.35	6.1*	3.04
Aug	12.15*	0.75	3.9	1.61	1.41	0.42	5.11	0.89	0.43	6.5*	0.58	0.47
Sept	7.75*	2.91	2.53	2.11	2.05	1.43	5.19	2.03	1.62	8.1*	2.57	2.08
Oct	2.88	0.91	0.16	5.18	1.03	0.75	2.52	0.96	0.31	3.86	0.88	0.12
Nov	4.86	0.37	0.12	15.4*	0.36	1.17	4.09	0.37	0.55	4.33	0.65	0.3
Dec	1.09	1.07	0.13	3.19	1.2	1.20	2.41	1.1	0.73	0.13	1.57	0.3

*Inappropriate Values; H. R. – Homogeneous Region.

function had only 2 unsuitable values, while the normal, log-normal and Weibull function had 8, 7 and 5 values without adherence, respectively. This result indicates that, with the exception of the months of April and July (RH II and RH III), the gamma function offered lower values than the table value (5.99), indicating it adjusted well to the frequencies of occurrence of the monthly precipitation observed. Thus, the PDF gamma had the best adherence to the chi-square test for monthly precipitation.

In a general evaluation of the adjusted graphs, in the November, December and January, the most adequate adjustments occur, whereas in the months of April, June and July, less adequate adjustments occurred. This result was observed based on the number of times the Chi-square values were above the chosen threshold (5.99), with a significance level of 5% and degree of freedom equal to 2. To validate the gamma function, the probabilities of occurrence of monthly average precipitation at the target stations were generated by this function. The results of this validation indicate a good adjustment of the gamma function, since the values of the chi-square test were all adequate, as can be observed in Table 6 and in the adjustment of the graphs that represent the probabilities of observed and estimated occurrence of average monthly precipitation (Figures 7, 8 and 9).

In comparison with other probability functions, the gamma function has presented good adjustments in the predictions of the probability of occurrence of monthly precipitation. Sampaio et al. (2006) and Amburn, Lang and Buonaiuto (2015), for example, used different PDFs to estimate the occurrences of precipitation probabilities, and the gamma function had the best result for monthly precipitation data.

The results of Table 5 show that there are many values with adherence in the normal, log-normal and Weibull models. However, according to Kist and Virgem Filho (2015), the adherence of a distribution to the data does not necessarily mean that the adjustment is good, only that there was not enough evidence in the series for rejection. Thus, because four different distributions were tested, and some presented values considered adherent, we cannot totally rule out the use of these functions in the studied region, and thus, the other PDFs could be adopted in this region if they pass other measures of calibration and validation. This analysis is also valid for the annual data series, in which the probability functions were also determined to be adequate by the Chi-square test (Table 3).

According to Murta et al. (2005), the gamma function, from the statistical point of view, does not behave as if evenly distributed around the mean value, but rather shows irregular and large deviations around the mean value. This function could guarantee a better result in the study of average monthly precipitation if the average value of the series is not influenced by the results. Thus, the adhesion test (Table 6) and the graph adjustment (Figure 7, 8 and 9) confirm that the Gamma model is valid for application in TAHR.

Multiple regression models for annual mean precipitation estimates

The multiple regression models were tested considering three independent variables (altitude, latitude and longitude) from the set of stations representing each homogeneous region. Thus, using the results of the performance criteria, we determined the best model for estimating the dependent variable.

In homogeneous regions I and II, in relation to R^2, R^2_a and *NASH*, the models were not significant, with a R^2 value varying from 0.39 to 0.46 (Table 7). In homogeneous region III, the models were more significant, with R^2 values of 0.67 to 0.74. In terms of percentage, this coefficient represents how much of the variability in precipitation is explained by the independent variables (altitude, latitude and longitude). Thus, the linear model represents 46% and 41% (0.46 and 0.41 - Table 7) of the variability in precipitation that occurred in regions I and II, respectively, presenting the highest R^2 value among the models for these regions. In homogeneous region III, this percentage was much better, at 74%. Considering E (%), ε (%) and RMSE, the models would perform well in the estimation of precipitation, since the errors obtained are less than 7% and 0.7%, and the RMSE presented minimum values. Therefore, the linear model is the most significant for the estimation of the annual precipitation in regions I, II and III, as it also presents higher R^2 and *Nash* values (Table 7).

To validate the linear model, the percentage relative error, E (%), between the observed precipitations (Po) of the target stations and the estimated precipitations (Pe) of the linear model (Figure 10) was calculated. The percentage errors obtained by the linear model were lower than 9% for almost all of the target stations. Only for the Fazenda Marajá station, which belongs to the homogeneous region II, was the error greater than 10%. However, for the Pirenopolis station located in the homogeneous

Table 6. Chi-square test with frequencies observed and estimated by the gamma function at the target stations.

HR	Station	Months											
		Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
I	Faz. Marajá	0.07	0.2	1.46	0.63	0.52	0.53	2.71	1.64	0.14	1.13	2.71	0.06
	Pirenópolis	3.8	2.84	0.03	1.36	0.96	0.46	2.21	1.04	0.19	0.77	0.81	0.5
	Faz. Babilônia	0.57	0.71	0.83	1.1	2.41	0.54	0.29	3.6	0.95	0.07	0.78	0.65
II	Trecho Médio	2.19	0.26	2.31	3.21	1.74	0.03	2.12	3.52	1.63	0.11	2.8	2.25
	Gurupi	0.28	1.17	2.64	0.16	0.59	0.57	0.26	2.52	0.3	0.31	0.61	1.23
	Formosa do Araguaia	3.19	0.92	0.98	3.28	2.39	0.09	3.12	0.75	3.4	3.36	2.58	2.36
III	Tucuruí	3.69	1.52	2.23	3.61	4.38	3.59	1.35	1.99	0.44	1.08	3.38	2.65
	Cametá	1.13	1.12	0.79	1.07	2.14	3.35	0.19	1.52	0.43	1.21	0.23	2
	Belém	0.21	1.66	1.63	1.67	1.86	3.39	1.97	4.58	2.40	0.29	0.17	0.32

H. R. – Homogeneous Region.

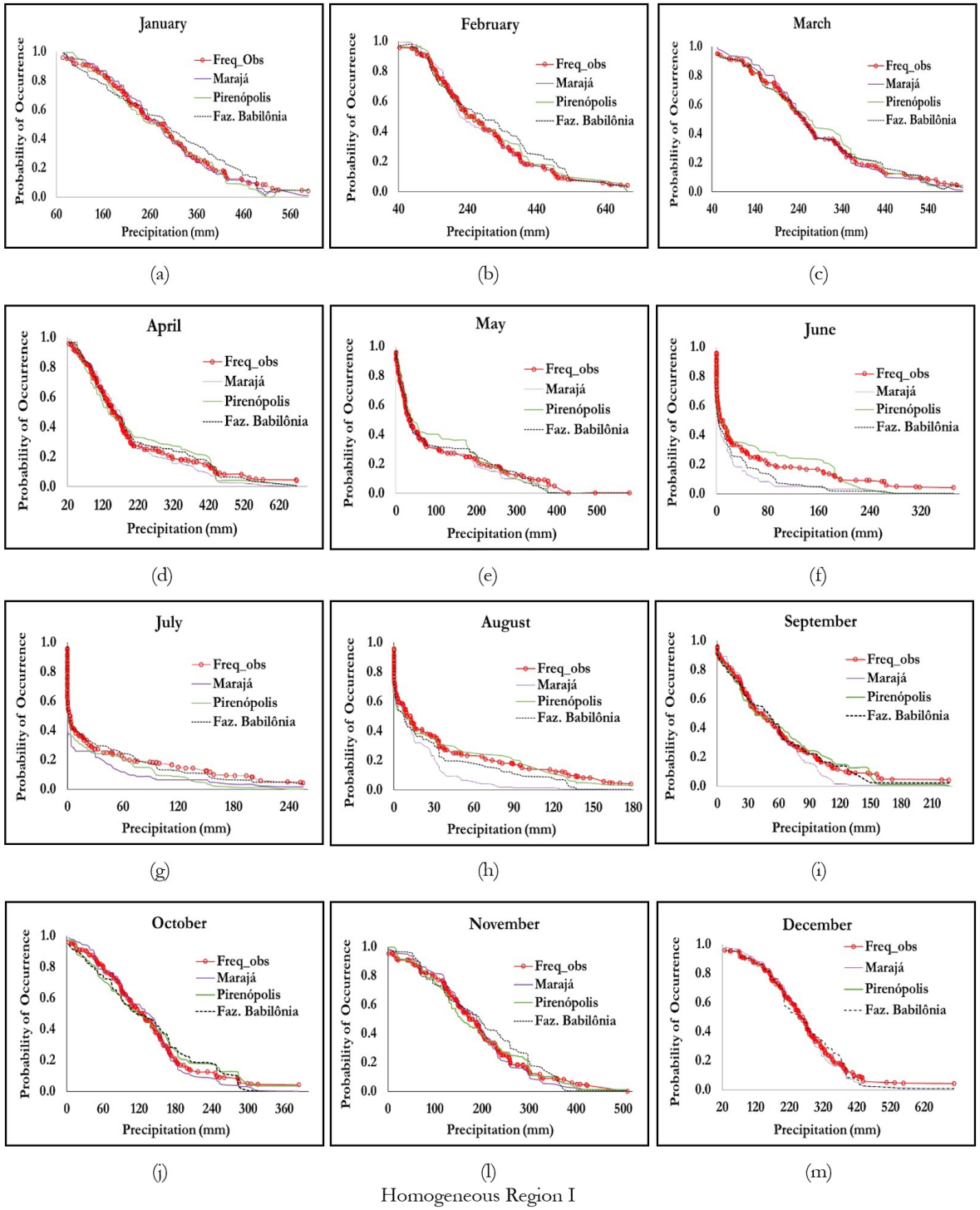


Figure 7. Probability of occurrence of observed and estimated monthly mean precipitation at the target stations – Homogeneous Region I.

region II, the error was at least 0.16% (Figure 10). In general, the errors between the observed and estimated heights were acceptable.

Regression models for the rainy and dry season

The multiple regression models did not perform well in estimates of monthly mean precipitation. The highest relative percentage errors occurred in the dry months, and the lowest errors occurred in the rainy season. Thus, the multiple regression was conducted on the dry and rainy season, in an attempt to obtain more representative and adequate models of the estimation of average monthly precipitation. Following this method, rainy months were considered, i.e., the months of November, December, January, February, March and April. The dry months contain May, June, July, August, September and October. This analysis was performed using the monthly average values of the rainy and dry months from each station in the homogeneous regions formed

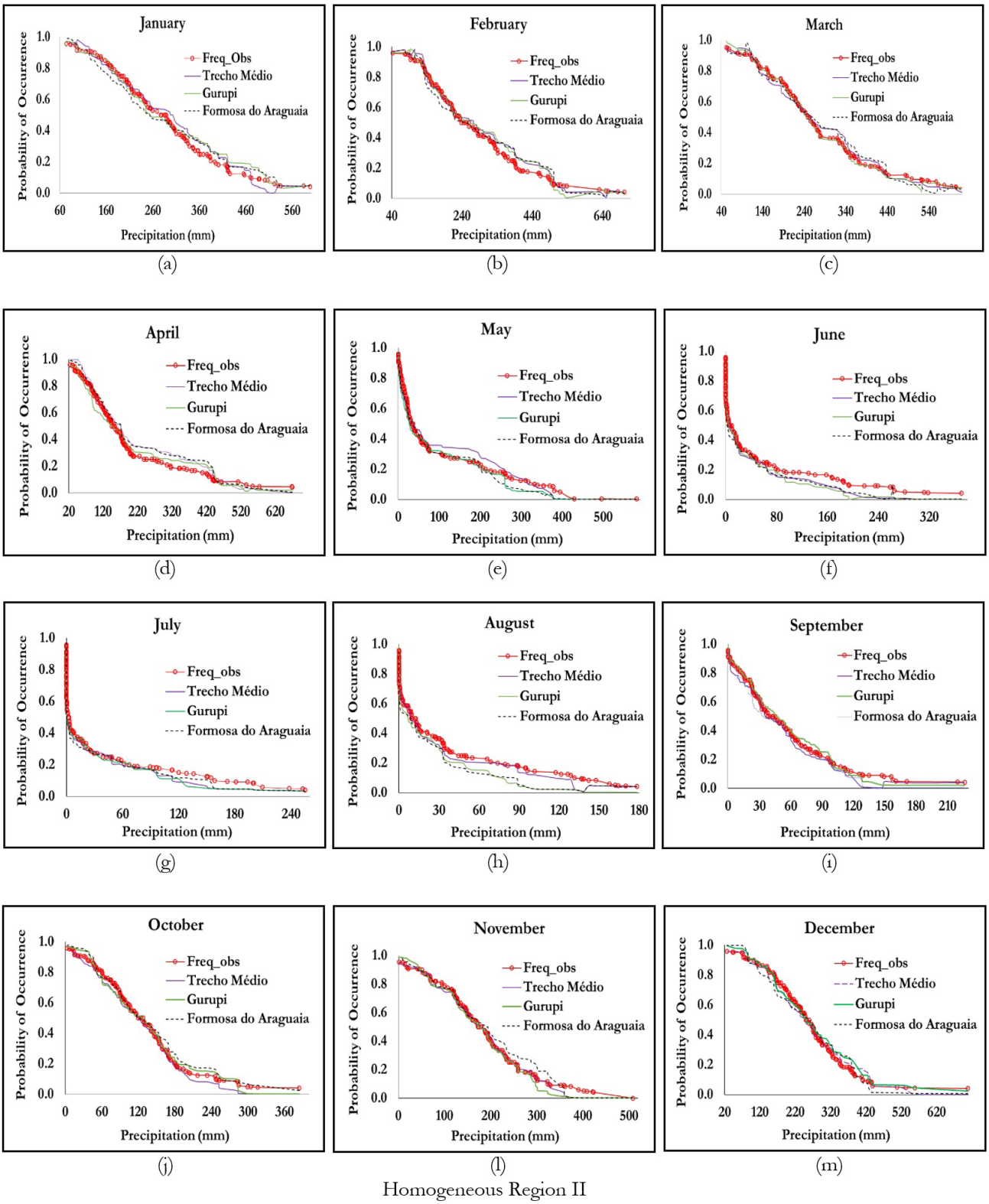


Figure 8. Probability of occurrence of observed and estimated monthly mean precipitation at the target stations - Homogeneous Region II.

from the fuzzy c-means clustering. Thus, a multiple regression model was applied with the linear, potential, exponential and logarithm models, adopting the mean precipitation of the rainy and dry season as a dependent variable. For the rainy months, the R^2 and *Nash* values obtained from the regression models were all below 0.39 in homogeneous regions I and II (Table 8), indicating that there is a weak relationship between the independent variables.

The logarithm model, for example, can explain only 21% and 17% of the precipitation variability in the homogeneous regions I and II, simultaneously (0.21 and 0.17 - Table 8). The percentage

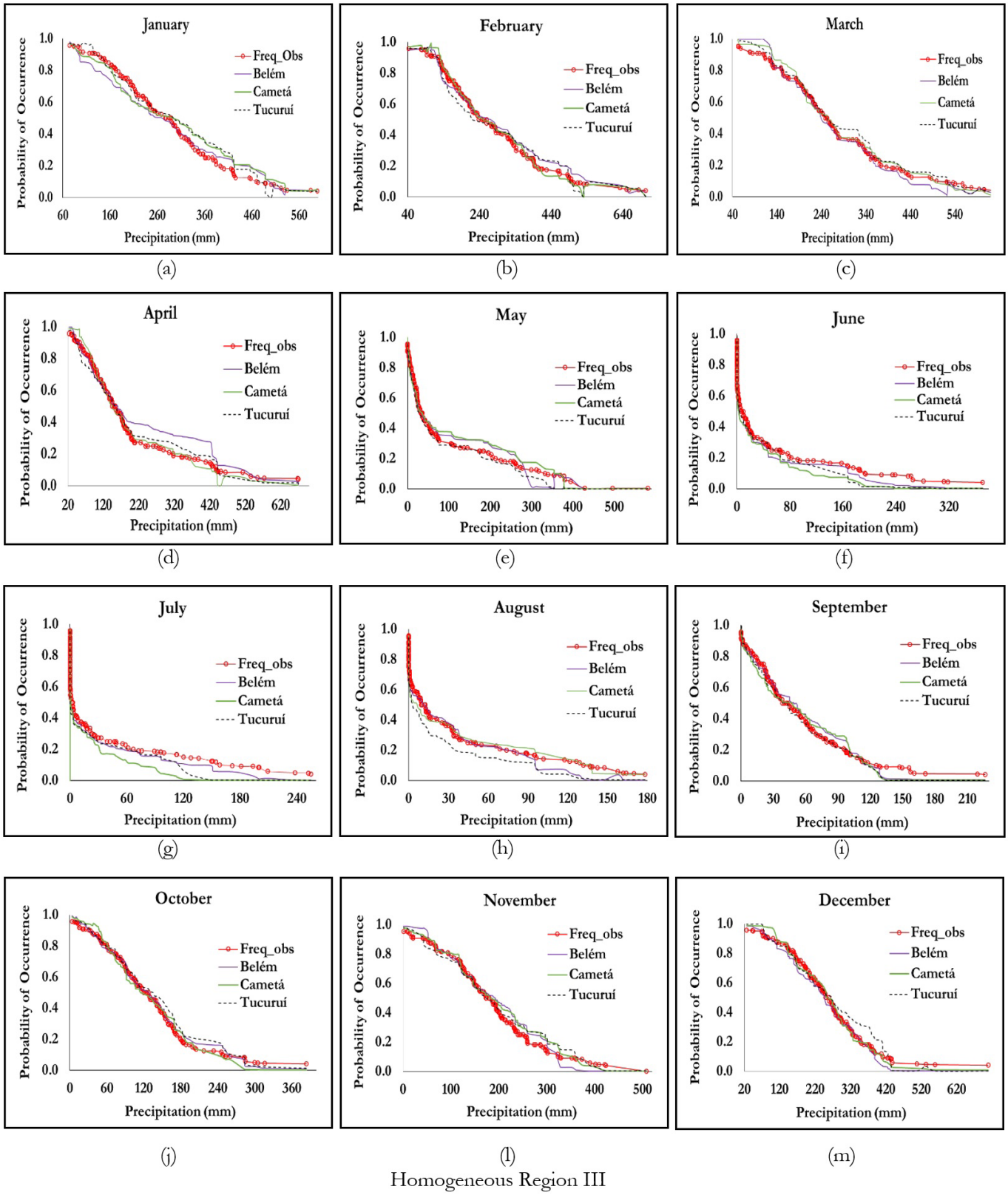


Figure 9. Probability of occurrence of observed and estimated monthly mean precipitation at the target stations - Homogeneous Region III.

errors (E, ε) were below 6.4% and 0.46%, respectively, and the *RMSE* was minimal, indicating that the models may be useful, even though the R^2 is low. In homogeneous region III, for the rainy season, all models presented values of 0.99 for the Nash coefficient, which indicates that they are excellent estimators. The R^2 was approximately 0.64 to 0.73. The percentage errors

Table 7. Regression models performance criteria for annual mean precipitation height estimation.

Homogeneous Region I						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.46	0.32	6.09	0.138	0.46	0.00017
Potential	0.42	0.28	5.61	0.138	0.42	0.00017
Exponential	0.45	0.30	5.61	0.139	0.45	0.00017
Logarithm	0.42	0.28	6.08	0.137	0.42	0.00017
Homogeneous Region II						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.41	0.26	3.89	0.58	0.41	0.00075
Potential	0.39	0.29	3.81	0.58	0.40	0.00076
Exponential	0.40	0.28	3.90	0.58	0.40	0.00075
Logarithm	0.39	0.22	3.80	0.58	0.39	0.00076
Homogeneous Region III						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.74	0.61	4.73	0.64	0.74	0.00071
Potential	0.67	0.51	5.75	0.63	0.68	0.00064
Exponential	0.72	0.58	4.51	0.65	0.75	0.00072
Logarithm	0.70	0.55	5.64	0.63	0.70	0.00046

R^2 - determination coefficient; R^2_a - adjusted coefficient of determination; *E* (%) – the average percentage relative error; ε (%) – mean relative root square error; NASH – coefficient of Nash Sutcliffe; RMSE – root mean squared error.

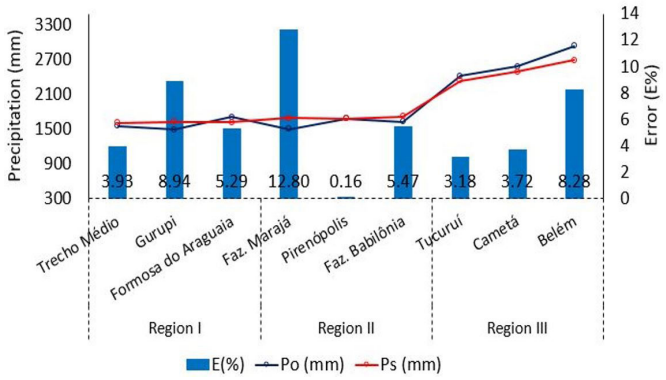


Figure 10. Percent errors in annual mean precipitation by homogeneous region and target station.

were below 5% and 0.63%, giving an acceptable percentage with which to estimate the average precipitation of the rainy season in this region.

In Figure 11d, e, f, which compares the observed and estimated precipitation from the stations of each region to the rainy season values, the linear model shows a better fit in the three regions, as indicated by the small variability of the points around the 1:1 line, and provides a better estimation of the data, suggesting that the model simulates values close to the observed precipitation.

In the dry season, in homogeneous region II and homogeneous region III, although the percentage relative error, *E* (%), was greater than 10%, the R^2 and Nash values range from 0.59 to 0.80 and 0.59 to 0.89, respectively (Table 9), indicating that the models explain precipitation variability well. The *RSME* and the mean relative root square error, ε (%), were low, confirming the good fit of the models. However, the potential model presented

Table 8. Performance criteria of the models for the rainy season.

Homogeneous Region I						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.38	0.23	6.00	0.11	0.38	0.0011
Potential	0.32	0.18	6.30	0.11	0.32	0.00104
Exponential	0.29	0.25	6.10	0.11	0.26	0.00109
Logarithm	0.21	0.16	6.20	0.11	0.21	0.00105
Homogeneous Region II						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.37	0.340	4.70	0.44	0.37	0.00341
Potential	0.17	0.027	4.80	0.44	0.17	0.00338
Exponential	0.18	0.037	4.70	0.45	0.18	0.0034
Logarithm	0.17	0.024	4.80	0.44	0.17	0.00339
Homogeneous Region III						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.73	0.44	4.60	0.60	0.996	0.00568
Potential	0.66	0.49	4.61	0.62	0.997	0.006
Exponential	0.64	0.41	4.63	0.60	0.996	0.00562
Logarithm	0.69	0.53	4.64	0.62	0.997	0.00427

R^2 - determination coefficient; R^2_a - adjusted coefficient of determination; *E* (%) – the average percentage relative error; ε (%) – mean relative root square error; NASH – coefficient of Nash Sutcliffe; RMSE – root mean squared error.

Table 9. Performance criteria for the models of the dry season.

Homogeneous Region I						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.80	0.78	9.65	0.03	0.8	0.00391
Potential	0.80	0.78	9.22	0.04	0.79	0.00387
Exponential	0.80	0.79	9.57	0.04	0.81	0.00403
Logarithm	0.80	0.78	9.58	0.03	0.80	0.00391
Homogeneous Region II						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.60	0.53	11.09	0.17	0.6	0.00953
Potential	0.62	0.61	10.94	0.17	0.61	0.00948
Exponential	0.60	0.53	10.85	0.18	0.60	0.00965
Logarithm	0.59	0.52	11.21	0.16	0.59	0.00935
Homogeneous Region III						
Models	**R^2**	**R^2_a**	***E*(%)**	**ε(%)**	**Nash**	**RMSE**
Linear	0.85	0.77	15.73	0.16	0.85	0.00527
Potential	0.89	0.87	14.08	0.27	0.80	0.00457
Exponential	0.87	0.82	13.78	0.23	0.87	0.00561
Logarithm	0.85	0.77	15.11	0.21	0.85	0.0036

R^2 - determination coefficient; R^2_a - adjusted coefficient of determination; *E* (%) – the average percentage relative error; ε (%) – mean relative root square error; NASH – coefficient of Nash Sutcliffe; RMSE – root mean squared error.

higher coefficients of determination (0.62 and 0.89 - Table 9), and the data points of the scatter plot in Figure 11g, h, f are very close to line 1:1 when compared to the observed and estimated precipitation, thus indicating that the potential model is the most acceptable for estimating the mean precipitation in the dry season.

For the dry season, in Region I, the values of R^2, R^2_*a* and *Nash* were approximately equal to 0.80, indicating that the models are representative. However, in the potential model, the values of the *RSME* and the percentage error were lower than those of the

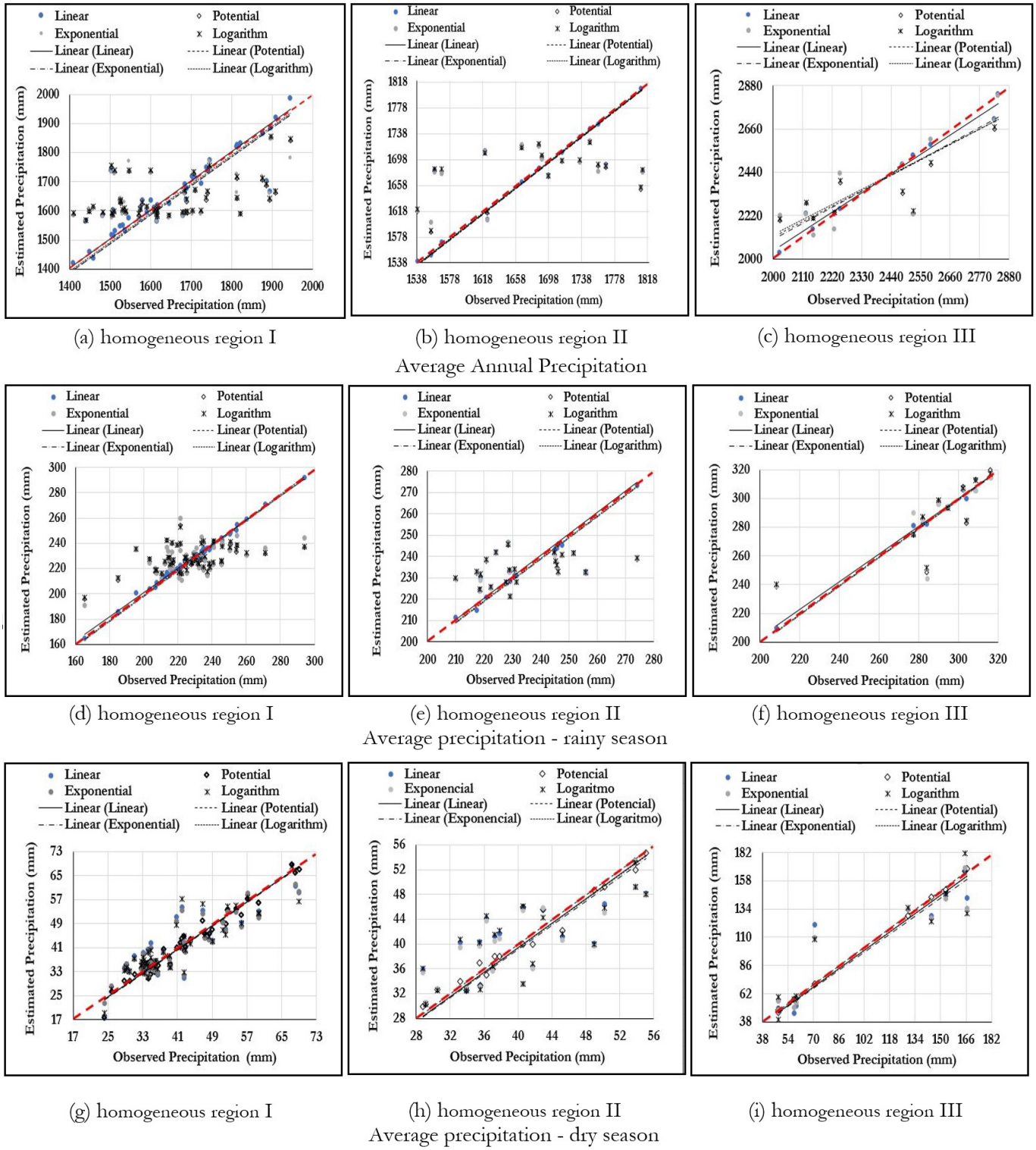


Figure 11. The 1:1 line for average annual precipitation and average monthly precipitation - rainy and dry season.

other models, suggesting that the potential model is best for the prediction of monthly precipitation in this region.

In the validation of the rainy season data, the respective regression parameters were obtained from the calibration with the linear model and the information from the target stations (altitude, latitude and longitude). The percentage relative error was determined between *Po* and *Pe* that was calculated by the linear model. The Tucuruí station presented the maximum error of 13% (Figure 12) in the estimation of monthly precipitation for the rainy season. However, the mean relative error was 5.6%, indicating that the model performed adequately for the rainy season in the 3 homogeneous regions.

In the validation of the dry season data, the observed precipitation (Po) values were compared with the precipitation

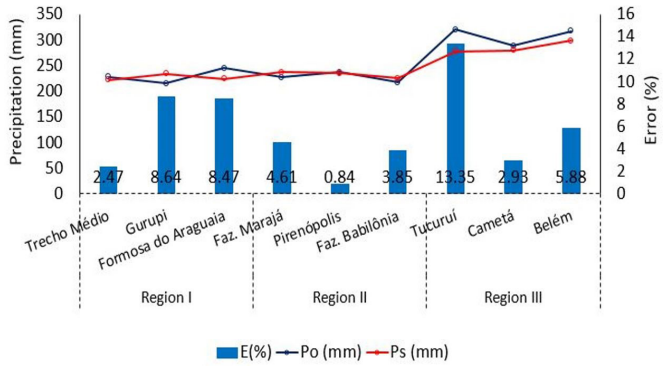


Figure 12. Percent errors by homogeneous region and target station for monthly mean precipitation - rainy season.

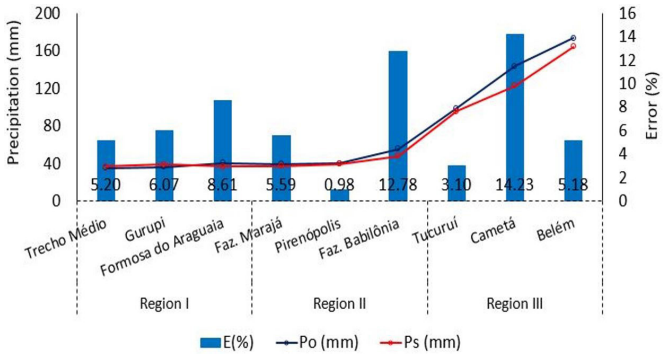


Figure 13. Percent errors by homogeneous region and target station for monthly mean precipitation - dry season.

values obtained by the potential model. The mean errors found were less than 10%. Despite the stations Faz. Babilônia and Cametá presenting errors of 12.78% and 14.23% (Figure 13), respectively, the potential model performed well in estimating the average monthly precipitation, with a mean error of 6.86% for the three homogeneous regions.

By the RMSE values obtained (Tables 7, 8 and 9), all the models evaluated could be considered as good estimators, since all were close to zero. However, when comparing the results of other criteria, the models are not considered satisfactory. To avoid this type of error, other measures were evaluated, such as the *Nash*, R^2, percentage, E (%), and mean, ε (%), errors, and the choice of the most appropriate model was prioritized.

According to Nash and Sutcliffe (1970), the Nash coefficient allows the efficiency of a model to be defined, and its value is analogous to the coefficient of determination (R^2); the closer the value is to 1, the better the model representation. In the results obtained, we can see that the value of R^2 approaches the Nash value. However, in the evaluation of multiple regression models, R^2 is the most important measure, as observed by Fumo and Rafe Biswas (2015), Alexander, Tropsha and Winkler (2015) and Bardak et al. (2016). Thus, R^2 value is the most relevant value to consider for when choosing a regression model; however, its evaluation is more consistent when there is an integration between the other performance criteria.

Table 10. Multiple regression models.

Regression models for the estimation of annual mean precipitation totals	
HR	**Linear**
I	$P = 272.3 + 0.27 * H + 31 * la + 33.4 * lo$
II	$P = 1475 - 0.15 * H + 21.4 * la + 0.38 * lo$
III	$P = 100063 - 1.5 * H + 97 * la + 149 * lo$
Regression models for the estimation of monthly precipitation for the rainy season	
HR	**Linear**
I	$P = -20.14 + 0.11 * H + 5.13 * la + 5.6 * lo$
II	$P = 363.56 - 0.014 * H + 1.7 * la + 1.86 * lo$
III	$P = 1224 + 0.02 * H - 0.6 * la + 19.12 * lo$
Regression models for the estimation of the average monthly precipitation for the dry season	
HR	**Potential**
I	$P = -12.4 + H^{0.089} la^{-0.23} + lo^{1.24}$
II	$P = -8.14 + H^{0.32} la^{0.44} + lo^{2.19}$
III	$P = 11.67 + H^{-0.11} la^{-0.37} + lo^{-1.59}$

H. R. – Homogeneous Region.

The proposed methodology can be considered acceptable for estimating precipitation since it analyzed the results of six performance criteria, evaluated observed and estimated precipitations using the dispersion graph and tested the proposed models with stations that were not considered in the calibration of the models. Through this methodology, estimates of the probability of occurrence of precipitation, as well as estimates of monthly and annual precipitation can be performed in locations without monitoring in a satisfactory way, just knowing the location and altitude data of a certain point within the basin studied. Table 10 shows the multiple regression models for estimating annual and monthly precipitation heights, in dry and rainy seasons, in the three homogeneous regions formed in the TAHR.

CONCLUSION

The grouping techniques, fuzzy c-means, PBM index and H-test were able to form distinct groups, with well-defined precipitation averages and a spatialization of the homogeneous regions appropriate to the rainfall recorded in the homogeneous regions. In the homogeneous regions I and II, formed to the southwest and center-west of the TAHR, respectively, smaller pluviometric volumes were determined. For the homogeneous Region III, located in the north, a higher pluviometric volume was determined, as was to be expected because the Amazon forest exists to the north of the TAHR and the Brazilian cerrado exists to the south.

Annual precipitation estimates performed well, both with the use of the probability distribution functions and through the use of multiple regression models. However, for the estimation of monthly averages, the regression models presented better estimates

when considering dry and rainy seasons. The monthly estimates were estimated satisfactorily using the probability functions without the need to consider dry and rainy seasons.

The performance criteria used in the validation of multiple regression models, provide a better analysis of the results, when used in an integrated way. The multiple regression models obtained use easy-to-obtain input variables, making them a useful tool for locations lacking precipitation data. Thus, the methodology developed can assist in the planning and management of others river basins, in terms of precipitation estimations.

ACKNOWLEDGEMENTS

The authors thank the ANA for the available precipitation data. The first author is grateful for a master's degree scholarship funded by CAPES. The second author is grateful for the research productivity grant funded by CNPq (process number 304936/2015-4). The third author is grateful for a PNPD grant funded by CAPES.

REFERENCES

ALEXANDER, D. L. J.; TROPSHA, A.; WINKLER, D. A. Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling*, v. 55, n. 7, p. 1316-1322, 2015. http://dx.doi.org/10.1021/acs.jcim.5b00206. PMid:26099013.

AMBURN, S. A.; LANG, A. S. I. D.; BUONAIUTO, M. A. Precipitation forecasting with gamma distribution models for gridded precipitation events in Eastern Oklahoma and Northwestern Arkansas. *American Meteorological Society*, v. 30, p. 349-367, 2015. http://dx.doi.org/10.1175/waf-d-14-00054.sl.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. *Caderno da Região Hidrográfica Tocantins Araguaia*. Brasília: ANA, MMA, 2006. 132 p.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. *Orientações para consistência de dados pluviométricos*. Brasília: ANA, SGH, 2012. Available from: < http://arquivos.ana.gov.br/infohidrologicas/cadastro/OrientacoesParaConsistenciaDadosPluviometricos-VersaoJul12.pdf>. Access on: 15 Aug. 2016.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. *Conjuntura dos recursos hídricos:* Informe 2016. Brasília, 2016a. Available from: <http://www.snirh.gov.br/portal/snirh/centrais-de-conteudos/conjuntura-dos-recursos-hidricos>. Access on: 2 Jan. 2018.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. *HidroWeb:* sistemas de informações hidrológicas. Brasília, 2016b. Available from: <http://hidroweb.ana.gov.br/>. Access on: 20 July 2016.

ARELLANO-LARA, F.; ESCALANTE-SANDOVAL, C. A. Multivariate delineation of rainfall homogeneous regions for estimating quantiles of maximum daily rainfall: a casa study of northwestern Mexico. *Atmosfera*, v. 27, n. 1, p. 47-60, 2014. http://dx.doi.org/10.1016/S0187-6236(14)71100-2.

ASONG, Z. E.; KHALIQ, M. N.; WHEATER, H. S. Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical atributes. *Stochastic Environmental Research and Risk Assessment*, v. 29, n. 3, p. 875-892, 2015. http://dx.doi.org/10.1007/s00477-014-0918-z.

AWAN, A. J.; BAE, D.; KIM, K. Identification and trend analysis of homogeneous rainfall zones over the East Asia monsoon region. *International Journal of Climatology*, v. 35, n. 7, p. 1422-1433, 2015. http://dx.doi.org/10.1002/joc.4066.

BARDAK, S.; TIRYAKI, S.; BARDAK, T.; AYDIN, A. Predictive performance of artificial neural network and multiple linear regression models in predicting adhesive bonding strength of wood. *Streng of Materials*, v. 48, n. 6, p. 811-824, 2016. Available from: <https://doi.org.ez3.periodicos.capes.gov.br/10.1007/s11223-017-9828-x>. Access on: 20 June 2017.

BEZDEK, J. *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981. http://dx.doi.org/10.1007/978-1-4757-0450-1.

CALDEIRA, T. M.; BESKOW, S.; MELLO, R. D.; FARIA, L. C.; SOUZA, M. R.; GUEDES, H. A. S. Modelagem probabilística de eventos de precipitação extrema no estado do Rio Grande do Sul. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 19, n. 3, p. 197-203, 2015. http://dx.doi.org/10.1590/1807-1929/agriambi.v19n3p197-203.

CHAI, T.; DRAXLER, R. R. Root means square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, v. 7, n. 3, p. 1247-1250, 2014. http://dx.doi.org/10.5194/gmd-7-1247-2014.

CHATZITHOMAS, C.; ALEXANDRIS, S.; KARAVITIS, C. Multivariate linear relation for precipitation: a new simple empirical formula. *Studia Geophysica et Geodaetica*, v. 59, n. 2, p. 325-344, 2015. http://dx.doi.org/10.1007/s11200-013-1162-6.

CHIFURIRA, R.; CHIKOBVU, D. A. Weighted multiple regression model to predict rainfall patterns: principal component analysis approach. *Mediterranean Journal of Social Sciences*, v. 5, n. 7, p. 34-52, 2014. http://dx.doi.org/10.5901/mjss.2014.v5n7p34.

CORDER, G. W.; FOREMAN, D. I. *Nonparametric statistics for non-statisticians:* a step-by-step approach. New Jersey: John Wiley and Sons, 2009. 264 p.

DAS, J.; UMAMAHESH, N. D. Downscaling monsoon rainfall over river godavari basin under different climate-change scenarios. *Water Resources Management*, v. 30, n. 15, p. 5575-5587, 2016. http://dx.doi.org/10.1007/s11269-016-1549-6.

DIKBAS, F.; FIRAT, M.; KOC, A. C.; GUNGOR, M. Classification of precipitation series using fuzzy cluster method. *Journal of Climatology*, v. 32, n. 10, p. 1596-1603, 2011. http://dx.doi.org/10.1002/joc.2350.

DUNN, J. C. A. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, v. 3, p. 32-57, 1973. http://dx.doi.org/10.1080/01969727308546046.

FARSADNIA, F.; ROSTAMI KAMROOD, M.; MOGHADDAM NIA, A.; MODARRES, R.; BRAY, M. T.; HAN, D.; SADATINEJAD, J. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. *Journal of Hydrology (Amsterdam)*, v. 509, p. 387-397, 2014. http://dx.doi.org/10.1016/j.jhydrol.2013.11.050.

FAZEL, N.; BERNDTSSON, R.; UVO, C. B.; MADANI, K.; KLØVE, B. Regionalization of precipitation characteristics in Iran's Lake Urmia basin. *Theoretical and Applied Climatology*, v. 132, n. 1-2, p. 363-373, 2018. http://dx.doi.org/10.1007/s00704-017-2090-0.

FUMO, N.; RAFE BISWAS, M. A. Regression analysis for prediction of residential energy consumption. *Renewable & Sustainable Energy Reviews*, v. 47, p. 332-343, 2015. http://dx.doi.org/10.1016/j.rser.2015.03.035.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. *Análise multivariada de dados*. 5. ed. Porto Alegre: Bookman, 2005. 593 p.

HOSKING, J.; WALLIS, J. Some statistic useful in regional frequency analysis. *Water Resources Research*, v. 29, n. 2, p. 271-28, 1993. http://dx.doi.org/10.1029/92WR01980.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Cobertura do uso da terra do Brasil*. Rio de Janeiro: IBGE, 2014. Available from: <https://www.ibge.gov.br/geociencias-novoportal/informacoes-ambientais/cobertura-e-uso-da-terra>. Access on: 13 Sept. 2017.

JOSE, V. R. R. Percentage and relative error measures in forecast evaluation. *Operations Research*, v. 65, n. 1, p. 200-211, 2017. http://dx.doi.org/10.1287/opre.2016.1550.

KIST, A.; VIRGEM FILHO, J. S. Análise probabilística da distribuição de dados diários de chuva no estado do Paraná. *Revista Ambiente & Água*, v. 10, n. 1, p. 172-181, 2015. http://dx.doi.org/10.4136/ambi-agua.1489.

LATT, Z. Z.; WITTENBERG, H.; URBAN, B. Clustering hydrological homogeneous regions end neural network based index flood estimation for ungauged catchments: an example of the Chindwin River in Myanmar. *Water Resources Management*, v. 29, n. 3, p. 913-928, 2015. http://dx.doi.org/10.1007/s11269-014-0851-4.

LI, Z.; BRISSETTE, F.; CHEN, J. Assessing the applicability of six precipitation probability distribution models on the Loess Plateau of China. *International Journal of Climatology*, v. 34, n. 2, p. 462-471, 2014. http://dx.doi.org/10.1002/joc.3699.

LOUREIRO, G. E.; FERNANDES, L. L.; ISHIHARA, J. H. Spatial and temporal variability of rainfall in the Tocantins-Araguaia Hydrographic Region. *Acta Scientiarum*, v. 37, n. 1, p. 89-98, 2015. http://dx.doi.org/10.4025/actascitechnol.v37i1.20778.

MEKANIK, F.; IMTEAZ, M. A.; GATO-TRINIDAD, S.; ELMAHDI, A. Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology (Amsterdam)*, v. 503, p. 11-21, 2013. http://dx.doi.org/10.1016/j.jhydrol.2013.08.035.

MURTA, R. M.; TEODORO, S. M.; BONOMO, P.; CHAVES, M. A. Precipitação pluvial mensal em níveis de probabilidade pela distribuição gama para duas localidades no Sudoeste da Bahia. *Ciência e Agrotecnologia*, v. 29, n. 5, p. 988-994, 2005. http://dx.doi.org/10.1590/S1413-70542005000500011.

NAGHETTINI, M.; PINTO, E. J. A. *Hidrologia estatística*. Belo Horizonte: Ed. CPRM, 2007. 552 p.

NASH, J. E.; SUTCLIFFE, J. V. River flow forecasting through conceptual models part I – a discusssion of principles. *Journal of Hydrology (Amsterdam)*, v. 10, n. 3, p. 282-290, 1970. http://dx.doi.org/10.1016/0022-1694(70)90255-6.

PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, K. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, v. 37, n. 3, p. 481-501, 2004. http://dx.doi.org/10.1016/j.patcog.2003.06.005.

PARRACHO, A. C.; MELO-GONÇALVES, P.; ROCHA, A. Regionalization of precipitation for the Iberian Peninsula and climate change. *Physics and Chemistry of the Earth*, v. 94, p. 146-154, 2015. http://dx.doi.org/10.1016/j.pce.2015.07.004.

PESSOA, F. C. L.; BLANCO, C. J. C.; GOMES, E. P. Delineation of homogeneous regions for streamflow via fuzzy c-means in the Amazon. *Water Practice & Technology*, v. 13, n. 1, p. 210-218, 2018. http://dx.doi.org/10.2166/wpt.2018.035.

RENCHER, A. C.; CHRISTENSEN, W. F. *Methods of multivariate analysis*. New Jersey: John Wiley and Sons, 2012. 768 p. http://dx.doi.org/10.1002/9781118391686.

SAMPAIO, S. C.; LONGO, A. J.; QUEIROZ, M. M. F.; GOMES, B. M.; BOAS, M. A. V.; SUSZEK, M. Estimativa e distribuição da precipitação mensal provável no Estado do Paraná. *Acta Scientiarum Human and Social Sciences*, v. 28, n. 2, p. 267-272, 2006. http://dx.doi.org/10.4025/actascihumansoc.v28i2.169.

SAMUEL, J.; COULIBALY, P.; METCALFE, R. A. Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. *Journal of Hydrologic Engineering*, v. 16, n. 5, p. 447-459, 2011. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000338.

SANTOS, E. B.; LUCIO, S. P.; SILVA, M. S. Precipitation regionalization of the Brazilian Amazon. *Atmospheric Science Letters*, v. 16, n. 3, p. 185-192, 2014. http://dx.doi.org/10.1002/asl2.535.

SATYANARAYANA, P.; SRINIVAS, V. V. Regionalization of precipitation in data sparse areas using large scale atmospheric variables – A fuzzy clustering approach. *Journal of Hidrology*, v. 405, n. 3-4, p. 462-473, 2011. http://dx.doi.org/10.1016/j.jhydrol.2011.05.044.

SHAHANA SHIRIN, A. H.; THOMAS, R. Regionalization of rainfall in Kerala State. *Procedia Technology*, v. 24, p. 15-22, 2016. http://dx.doi.org/10.1016/j.protcy.2016.05.004.

SUPRIYA, P.; KRISHNAVENI, M.; SUBBULAKSHMI, M. Regression analysis of annual maximum daily rainfall and stream flow for flood forecasting in Vellar River Basin. *Aquatic Procedia*, v. 4, p. 957-963, 2015. http://dx.doi.org/10.1016/j.aqpro.2015.02.120.

TEIXEIRA-GANDRA, C. F. A.; DAMÉ, R. C. F.; SIMONETE, M. A. Predição da precipitação a partir das coordenadas geográficas no Estado do Rio Grande do Sul. *Revista Brasileira de Geografia Física*, v. 8, n. 3, p. 848-856, 2015. Available from: <https://periodicos.ufpe.br/revistas/rbgfe/article/view/233264/27096>. Access on: 8 Mar. 2017.

YUAN, J.; EMURA, K.; FARNHAM, C.; ALAM, M. A. Frequency analysis of annual maximum hourly precipitation and determination of best fit probability distribution for regions in Japan. *Urban Climate*, v. 24, p. 276-286, 2018. http://dx.doi.org/10.1016/j.uclim.2017.07.008.

Authors contributions

Evanice Pinheiro Gomes: Initial research study; acquisition, analysis and interpretation of data; preparation of the manuscript.

Claudio José Cavalcante Blanco: Study design and research; contributions to the analysis of results and conclusions; correction and critical review.

Francisco Carlos Lira Pessoa: Participation in the study design and initial research; contributions in the review of results; critical review.