



# Identification of homogeneous precipitation regions via Fuzzy c-means in the hydrographic region of Tocantins–Araguaia of Brazilian Amazonia

Evanice Pinheiro Gomes<sup>1</sup> · Claudio José Cavalcante Blanco<sup>2</sup> · Francisco Carlos Lira Pessoa<sup>2</sup>

Received: 17 October 2017 / Accepted: 29 November 2018 / Published online: 8 December 2018  
© The Author(s) 2018

## Abstract

Determination of homogeneous regions of precipitation is a major step towards obtaining regional rainfall patterns, which are models for the estimation of total rainfall used in water resources engineering. In this study, homogeneous regions of precipitation were identified within the Hydrographic Region of Tocantins–Araguaia (HRTA) of Brazilian Amazonia. This hydrographic basin is of great importance for Brazil because it has been exploited for the production of hydropower since the 1970s. Currently it is a border of the agribusiness of the country. Therefore, it is important to know the rainfall regime of the region. Thus, three homogeneous regions of precipitation were delimited using the Fuzzy c-means method and physical-climatic variables such as location (latitude and longitude), altitude, and precipitation. These regions were also tested and confirmed for their homogeneity using the Heterogeneity Test  $H$ . The values of total precipitation found for the regions are consistent with the volume of precipitation recorded in the analysed region and that found in the literature. The formation of these regions, in addition to contributing to the understanding of the hydrological behaviour, will aid in studies of the regionalization of rainfall in the region.

**Keywords** Physical-climatic variables ·  $PBM$  index · Heterogeneity Test  $H$

## Introduction

The knowledge of hydrological variables is indispensable for the management of water resources. Among the hydrological variables, precipitation is one of the most important variable. For example, its scarcity affects uses such as irrigation and public supply. Already their surplus can generate floods and erosion of the soil, damaging, respectively, cities and the use of the soil. However, one of the problems presented in precipitation studies is the lack of monitoring and

temporal and spatial information on precipitation. This problem greatly affects the planning and management of water resources, especially in irrigation projects, supply reservoirs, urban drainage design and flood control systems, which need knowledge of precipitation to be planned and operated efficiently, guaranteeing essential services to the population.

In order to obtain information on rainfall in a river basin, researchers have sought techniques to explore existing rainfall data in certain parts of a basin and estimate it for areas in need or lack of hydrological information. In this context, the formation of homogeneous regions of precipitation is a tool capable of providing the spatial and temporal behaviour of precipitation. The term homogeneous regions are associated with regions that have hydrological similarity (Patil and Stieglitz 2011; Wazneh et al. 2013; Swain et al. 2016).

In this context, the Fuzzy c-means method has presented good results in the formation of regional clusters, as, for example, in the studies developed by Dikbas et al. (2011). These authors compared the Fuzzy c-means and k-means clustering methods and noted that the Fuzzy c-means method was the best for homogeneous region formation. Sadri and Burn (2011) adopted the L-moment statistic and

✉ Claudio José Cavalcante Blanco  
blanco@ufpa.br

Evanice Pinheiro Gomes  
gomesevanice@ufpa.br

Francisco Carlos Lira Pessoa  
fclpessoa@ufpa.br

<sup>1</sup> Post-Graduation Program of Civil Engineering, Federal University of Pará – PPGEC/ITEC/UFPA, Belém, Brazil

<sup>2</sup> School of Environmental and Sanitary Engineering, Federal University of Pará – FAESA/ITEC/UFPA, Belém, Brazil

the Fuzzy c-means method for the formation of homogeneous regions of precipitation in the Canadian provinces of Alberta, Saskatchewan, and Manitoba. Satyanarayana and Srinivas (2011) were able to identify and regionalize twenty-four homogeneous precipitation clusters throughout the Chinese territory using the Fuzzy c-means method. Farsadnia et al. (2014) adopted the self-organizing feature map (SOFM) method, along with the Fuzzy c-means, K-means, and Ward methods to identify homogeneous regions of precipitation in Mazandaran Province of northern Iran. Goyal and Gupta (2014) compared the Fuzzy c-means and k-means methods in the definition of homogeneous regions of precipitation in Northeast India and concluded that the Fuzzy c-means method presented better results in the formation of regions.

Existing precipitation studies, carried out by Brazilian Agricultural Research Corporation—EMBRAPA (1994), by the National Water Agency—ANA (2009) and by Loureiro et al. (2015), which adopted geostatistical interpolation in the region, characterized the precipitation, considering only the behaviour of the historical series and identified that the total rainfall decreases in the north–south direction. However, the formation of homogeneous regions by the fuzzy group c-means taking in account the series of precipitation, geographic, and climatic characteristics of the basin, making this method complete, since the precipitation is influenced by several climatic elements, such as altitude and the geographic position, which were adopted in this study. In relation to the formation of homogeneous regions, comparing to other methods, the Fuzzy c-means is based on the concept of pertinence, identifying homogeneous regions with less subjectivity than other methods.

The Hydrographic Region of Tocantins–Araguaia—HRTA is of great importance for Brazil due the hydroelectric production. In addition to hydroelectric potential, the region has excelled in mining, agroindustry, agriculture, and livestock, and especially in irrigation projects for corn, rice, and soybeans. EMBRAPA (2014) registered 109.5 thousand hectares of irrigable areas in the region. Thus, the objective of the paper is to identify homogeneous regions of precipitation in the HRTA, using Fuzzy c-means method, showing the spatial variability of annual rainfall totals in the region. These results can be used by society towards a more sustainable use of water resources, especially for this region, which has a big demand for water resources.

## Materials and methods

### Study area

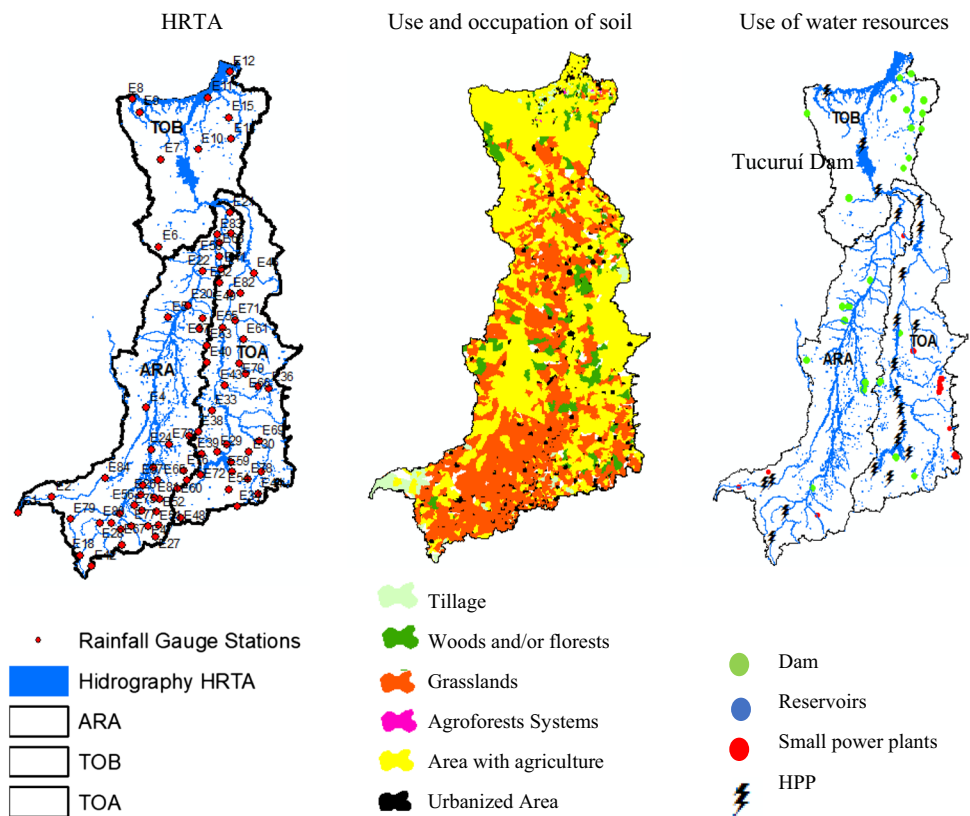
The region lies between the south parallel  $0^{\circ}30'$  and  $18^{\circ}05'$  and the longitude meridians  $45^{\circ}45'$  and  $56^{\circ}20'$ . Its

configuration is elongated, with a South–North direction, following the predominant direction of the main watercourses, the Tocantins and Araguaia rivers. The total drainage area of the HRTA is  $918.822 \text{ km}^2$  and covers part of the Midwest, North and North-east regions of Brazil. This region has a tropical climate, with an average annual temperature of  $26^{\circ}\text{C}$ , and two well-defined climatic periods: rainy, from October to April, with more than 90% precipitation, with the existence of some dry days between January and February, forming the so-called summer; and dry matter, from May to September, with low relative humidity. The water balance of the region estimates that the average annual precipitation is of the order of 1.837 mm and the flow is of  $13.624 \text{ m}^3/\text{s}$  and the actual evapotranspiration is 1.371 mm, which represents 75% of the precipitation, while the annual average real evapotranspiration of the country is 1.134 mm or 63% of the precipitation, and the mean coefficient of surface flow is 0.30 (National Water Agency (ANA) 2009). Figure 1 presents that HRTA is divided into three sub-basins: Alto Tocantins (ATO), Baixo Tocantins (BTO), and Araguaia (ARA). This figure also shows the use and occupation of soil, indicating strong anthropic action and the use of water resources mainly destined to hydroelectric production. Thus, the hydrographic basin is of great importance for Brazil, since its source has been exploited for the production of hydropower since the 1970s and has not yet been exhausted. Tucuruí Hydroelectric Power Plant, located in the state of Pará, is a large-scale hydroelectric power plant. In addition to hydroelectric potential, the region has excelled in mining, agroindustry, agriculture, and livestock, and especially in irrigation projects for corn, rice, and soybeans. According to the monitoring carried out by the Brazilian Agricultural Research Company—EMBRAPA, 109.5 thousand hectares of irrigable areas were registered in this region in 2014. The activities of land use and occupation are divided into urbanized areas of crops, of agroforestry systems, pastures, and agricultural establishments.

### Data sources

Historical series of rainfall amounts were adopted from 83 rainfall gauge station of the National Water Agency (ANA) database in HRTA (Fig. 1). The rainy seasons were chosen based on the historical series of data, opting for the stations with a larger series of data that were consistent and without observation failures. Of the 83 stations adopted, 70 had series with 30 years of data (1975–2004) and 13 stations had series ranging from 17 to 28 years (1977–2004). These series were organized in a database, which includes calculations for the average annual precipitation of each station. Information on altitude and geographic location was also extracted from the ANA database. The mean annual precipitation (MAP), altitude, latitude and longitude were used were used to apply the

**Fig. 1** Hydrographic region of Tocantins–Araguaia (HRTA)



Fuzzy c-means method and thus identify the homogeneous regions of precipitation. Table 1 shows the rainfall gauge stations and variables used in this study.

**Fuzzy c-means**

In the Fuzzy c-means clustering, the partitions were generated by minimizing a function, equated by an iterative algorithm (FCM), indicating the degree of membership of an element belonging to a particular cluster. Therefore, it is a technique in which each element belongs to a cluster with a certain degree of pertinence. The technique required pre-specifications of the number of clusters to be formed. The Fuzzy c-means cluster looks for the partition that minimizes the objective function, as represented by Eq. 1.

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

where  $n$  is the number of data;  $p$  is the number of clusters;  $u_{ij}$  is the degree of relevance of the sample  $X_i$  to the  $j$ -th cluster;  $m$  is the fuzziness parameter;  $d$  is the Euclidean distance between  $X_i$  and  $C_j$ ;  $X_i$  is the data vector, with  $i = 1, 2, \dots, n$ , representing a data attribute; and  $C_j$  is the centre of a fuzzy clustering. The objective function  $J$  is minimized, and the membership degrees  $u_{ij}$  are generated according to Eq. 2.

$$u_{ij} = \left[ \sum_{k=1}^c \left( \frac{d(X_i, C_j)}{d(X_i, C_k)} \right)^{2/(m-1)} \right]^{-1} \tag{2}$$

where  $C_j$  can be obtained by Eq. 3.

$$C_j = \frac{\sum_{i=1}^n (u_{ij})^m X_i}{\sum_{i=1}^n (u_{ij})^m} \tag{3}$$

The degrees of membership  $u_{ij}$ , representing the probabilities, are generated from a uniform distribution in the interval  $[0,1]$ . The clusterings are modified at each iteration following the algorithm (Fig. 2).

The fuzziness parameter ( $m$ ) is also known as the Fuzzy weight exponent and is a parameter that controls the level of diffusivity in the classification process. Thus, for  $m = 1$ , the clusters have strict limits equivalent to those of the k-means and, as the value increases, the boundaries become more diffuse. According to Cox (2005),  $m$  is usually in the range of 1.25–2.0. The cluster decision is defined by the greater degree of relevance presented for each element analysed. Thus, for a given  $X_i$ , its greater degree of pertinence will determine to which cluster this  $X_i$  belongs, which clusters all the data and avoids equivocations and rigidity in the formation of the clusterings.

**Table 1** Rainfall gauge stations and variables used in this study

ID	Stations	MAP (mm)	Altitude (m)	Latitude	Longitude
E1	São José da Serra	1613.78	797	-15.836	-55.322
E2	Rio das Mortes	1689.70	551	-15.315	-54.176
E3	Alô Brasil	1657.61	339	-12.164	-51.697
E4	Santo Antonio Leverger	1581.00	205	-12.292	-50.963
E5	Barreira do Campo	1417.30	195	-9.228	-50.211
E6	Fazenda Caiçara	1730.10	95	-6.815	-50.539
E7	Faz. Estrela Norte	1931.40	22	-3.854	-50.463
E8	Acampamento IBDF	2011.56	11	-1.792	-51.434
E9	Maracacuera Florestal	2619.68	20	-2.245	-51.179
E10	Cachoeira Tracambeua	2381.51	50	-3.510	-49.209
E11	Abaetetuba	2583.90	13	-1.738	-48.859
E12	Vigia	2843.30	15	-0.868	-48.110
E13	Faz. Maringa	1933.30	20	-3.161	-48.100
E14	Tomé-Açu	2552.80	45	-2.421	-48.149
E15	Abreulandia	2117.39	240	-9.624	-49.155
E16	Almas	1524.14	427	-11.579	-47.174
E17	Alto Araguaia	1681.33	659	-17.300	-53.219
E18	Ananas	1561.50	191	-6.364	-48.073
E19	Araguacema	2048.03	203	-8.810	-49.556
E20	Araguatins	1551.70	122	-5.648	-48.125
E21	Arapoema	1866.98	215	-7.659	-49.065
E22	Aruanã	1536.86	200	-14.927	-51.081
E23	Bandeirantes	1456.25	276	-13.698	-50.800
E24	Bom Jardim de Goiás	1649.99	402	-16.209	-52.169
E25	Britânia	1416.60	297	-15.246	-51.163
E26	Cachoeira GO	1513.54	766	-16.669	-50.649
E27	Caiaponia	1631.64	713	-16.949	-51.811
E28	Campinaçu	2441.46	683	-13.790	-48.567
E29	Cavalcante	1849.89	821	-13.797	-47.462
E30	Colinas do Sul	1573.37	530	-14.151	-48.078
E31	Colinas TO	1801.25	229	-8.053	-48.482
E32	Colonha	1419.70	264	-12.388	-48.711
E33	Contagem	1570.26	1242	-15.653	-47.878
E34	Córrego do ouro	1544.32	569	-16.298	-50.557
E35	Dianópolis	1449.21	679	-11.625	-46.811
E36	Dois Irmãos Tocantins	2028.89	264	-9.257	-49.064
E37	Entroncamento S M	1634.60	345	-13.103	-49.201
E38	Estrela do Norte	1750.67	467	-13.873	-49.071
E39	Fátima	1897.41	352	-10.764	-48.902
E40	Faz Primavera	1816.08	257	-7.559	-48.421
E41	Faz São Bernardo	1674.44	750	-17.680	-52.833
E42	Faz. Lobeira	1556.30	243	-11.531	-48.292
E43	Faz. Santa sé	1684.06	573	-15.216	-47.157
E44	Flores GO	1144.44	200	-14.450	-47.045
E45	Goiantins	1572.48	185	-7.711	-47.315
E46	Israelândia	1597.32	406	-16.300	-50.906
E47	Itaberáí	1827.62	726	-16.030	-49.800
E48	Itacaja	1844.53	250	-8.392	-47.763
E49	Itapirapua	1589.17	343	-15.816	-50.609
E50	Itapuranga	1644.69	646	-15.816	-50.608
E51	Jeroaquara	1779.80	400	-15.373	-50.500

Table 1 (continued)

ID	Stations	MAP (mm)	Altitude (m)	Latitude	Longitude
E52	Lagoa da Flexa	1436.41	200	-14.333	-50.730
E53	Mimoso	1307.59	687	-15.058	-48.159
E54	Miracema Tocantins	1706.89	210	-9.564	-48.388
E55	Monte Carlos GO	1542.97	400	-15.606	-51.357
E56	Mozarlandia	1654.07	400	-14.741	-50.577
E57	Muricilandia	1671.40	393	-7.154	-48.470
E58	Niquelândia	1703.61	568	-14.475	-48.041
E59	Nova América	1605.70	800	-15.021	-49.892
E60	Novo Acordo	1598.19	300	-9.961	-47.675
E61	Novo Planalto	1588.39	286	-13.245	-49.502
E62	Paraíso do TO	2280.65	390	-10.165	-48.891
E63	Perez	1499.26	299	-15.890	-51.853
E64	Pilar de Goiás	1948.00	765	-14.761	-49.580
E65	Pindorama do Tocantins	1614.57	444	-11.140	-47.576
E66	Piranhas	1582.92	356	-16.423	-51.823
E67	Piraquê	1760.80	184	-6.672	-48.470
E68	Ponte Paranã	1245.44	363	-13.425	-47.139
E69	Porto Gilandia	1656.37	220	-10.786	-47.800
E70	Porto Real	1599.03	200	-9.307	-47.929
E71	Porto Uruaçu	1468.39	572	-14.555	-49.139
E72	Rio Pintado	1443.73	200	-13.529	-50.188
E73	Sama	1410.71	375	-13.533	-48.227
E74	Santa fé	1614.51	400	-15.767	-51.104
E75	Santa Terezinha GO	1504.88	400	-14.433	-49.706
E76	São Ferreira	1673.12	361	-16.306	-51.471
E77	São João Aliança	1498.55	1009	-14.707	-47.524
E78	Tesouro	1714.94	389	-16.078	-53.549
E79	Torixoreu	1405.73	307	-16.201	-52.550
E80	Travessão	1516.55	450	-15.369	-50.705
E81	Tupiratins	1740.09	192	-8.398	-48.130
E82	Xambioá	1694.99	148	-6.413	-48.533
E83	Xavantina	1526.45	263	-14.672	-52.355

#### Algorithm of the Fuzzy c-means method

- Determine the value for  $p$  (number of groups),  $m$  (fuzziness parameter) and  $\mathcal{E}$  (error);
- Initialize the centroids according to Equation 3;
- Initialize the iteration counter  $t$  as  $t = 0$ ;
- Calculate the objective function  $J$  by means of Equation 1;
- Calculate the degrees of membership according to Equation 2;
- Check the stop condition:  $d(V_t, V_{t+1}) \leq \epsilon$ ;
- If stop condition = false then repeat the previous steps; otherwise, finalize the algorithm.

Fig. 2 Structure of the Fuzzy c-means algorithm

#### PBM validation index

One of the questions in a clustering analysis is the validation of the formed clusters. To achieve a good result, it is necessary to evaluate which partition is most suitable for the data and whether the partition generated by the algorithm is of good quality. To answer these questions, there are several validation indices in the literature, such as the  $V_{PC}$  and  $V_{PE}$  index (Bezdek 1981); the  $V_{WPE}$  index (Windhan 1981); the  $V_{MPC}$  index (Fukuyama and Sugeno 1989) and the  $PBM$  index (Pakhira et al. 2004).

In this study, the PBM index was used to validate the clusters and assess both the distances between the clusters formed and those between the elements and the centres of the formed clusters, which makes the validation safer. According to Pakhira et al. (2004), the PBM index serves to validate the number of clusters or subsets formed from a

dataset. This index is defined as the product of three factors (Eq. 4), of which maximization ensures that the partition has a small number of compact clusters with large separations between at least two.

$$PBM(k) = \left( \frac{1}{k} \cdot \frac{E_1}{E_k} \cdot D_k \right)^2 \tag{4}$$

where  $k$  is the number of clusters.

The factor  $E_1$  (Eq. 5) is the sum of the distances of each sample to the geometric centre of all samples  $w_0$ . This factor does not depend on the number of clusters.

$$E_1 = \sum_{t=1 \dots n} d(x(t), W_0) \tag{5}$$

The factor  $E_k$  (Eq. 6) is the sum of the distances between the clusters of  $K$  clusterings and is weighted by the corresponding relevance value of each sample to the cluster.

$$E_k = \sum_{t=1 \dots n} \sum_{i=1 \dots k} ui(t)d(x(t), W_i)^2 \tag{6}$$

$D_k$  (Eq. 7) represents the maximum separation of each pair of clusterings.

$$D_k = \max_{i,j=1 \dots k} (d(w_i, w_j)) \tag{7}$$

The procedure to calculate the PBM index can be described as follows:

1. Select the maximum number of clusters  $M$ ;
2. Calculate the factor  $E_1$ ;
3. For  $K=2$  to  $K=M$ , do:
  - a. Run the FCM algorithm;
  - b. Calculate the factors  $E_k$  and  $D_k$ ;
  - c. Calculate the PBM index (k).
4. Determine the best number of clusters  $K$  (Eq. 8).

$$K = \text{maxarg}(PBM(k)) \tag{8}$$

The PBM index is an optimization index, so to obtain the best partition, one must process the algorithm for several  $K$  values and choose the one that results in the highest index value because the higher the PBM index, the better the partition (Pakhira et al. 2004).

### L-moments

The L-moments make up a system of more reliable statistical measures for describing the characteristics of probability distributions and are derived from the probability-weighted moments (PWM) as generalized by Hosking and Wallis (1993). These moments are considered measures of the position, scale and shape of the probability distributions and are

similar to conventional moments, but estimated by linear combinations (Eq. 9), asymmetry, kurtosis and the coefficient of variation.

$$\beta_r = E\{X[Fx(x)]\} \tag{9}$$

where  $\beta_r$  is the probability-weighted moment (PWM);  $E$  is the probability of occurrence of the variable; and  $Fx(x)$  is the cumulative distribution function of  $X$ . According to Naghettini and Pintpo (2007), the estimation of  $\beta_r$  from a finite sample of size  $n$ , begins with the ordering of its constituent elements in ascending order, that is,  $X_1: n \leq X_2: n \leq \dots X_n: n$  and the values of the observed variable. Thus, the sample L-moments are calculated (Eqs. 10–13).

$$\beta_r = \frac{1}{n} \sum_{j=r+1}^n \frac{(j-1)(j-2) \dots (j-r)}{(n-1)(n-2) \dots (n-r)} x_{j:n} \tag{10}$$

$$\beta_0 = \frac{1}{n} \sum_{j=1}^n x_{j:n} \tag{11}$$

$$\beta_1 = \frac{1}{n} \sum_{j=2}^n \frac{(j-1)}{(n-1)} x_{j:n} \tag{12}$$

$$\beta_2 = \frac{1}{n} \sum_{j=1}^n \frac{(j-1)(j-2)}{(n-1)(n-2)} x_{j:n} \tag{13}$$

where  $X_j$  represents the samples, and  $n$  is the number of samples. These estimators serve to calculate the first four moments:  $\lambda_1, \lambda_2, \lambda_3$ , and  $\lambda_4$ , which are obtained using Eqs. 14, 15, 16 and 17, respectively.

$$\lambda_1 = \beta_0 \tag{14}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{15}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{16}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{17}$$

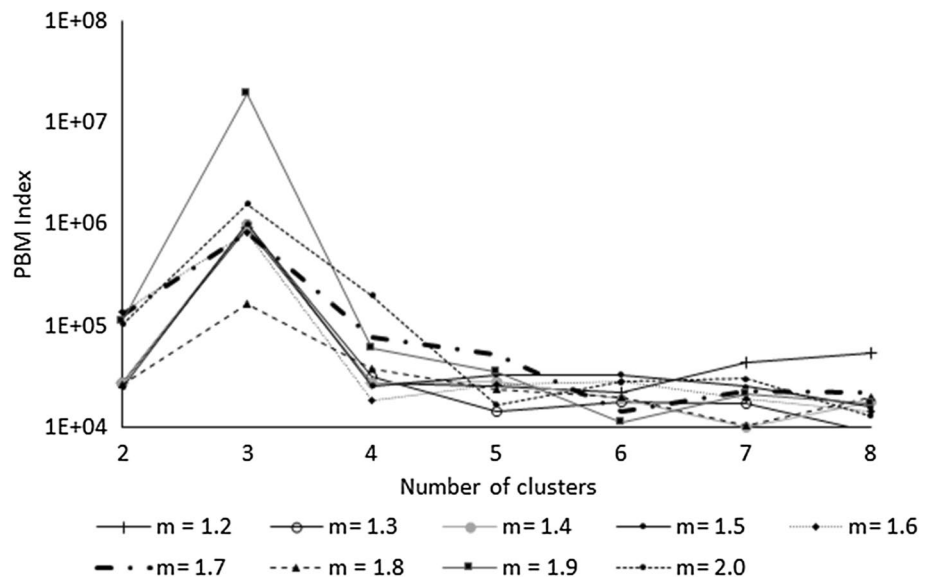
Regarding shape measurements of distributions, it becomes more convenient for the L-moments to be expressed in dimensionless quantities. These quotients serve to determine the standard deviation of the homogeneous regions and are obtained using Eqs. 18–20.

$$Cv - L = \tau_2 = \frac{\lambda_2}{\lambda_1} \tag{18}$$

$$\gamma - L = \tau_3 = \frac{\lambda_3}{\lambda_2} \tag{19}$$



**Fig. 3** PBM indices of the clusters



**Table 2** Results of the application of the PBM index to the clusterings of the algorithm FCM

<i>p</i>	<i>m</i> =1.2	<i>m</i> =1.3	<i>m</i> =1.4	<i>m</i> =1.5	<i>m</i> =1.6	<i>m</i> =1.7	<i>m</i> =1.8	<i>m</i> =1.9	<i>m</i> =2.0
2	3E+04	3E+04	3E+04	2E+04	1E+05	1E+05	3E+04	1E+05	1E+05
3	9E+05	1E+06	1E+06	1E+06	8E+05	8E+05	2E+05	<b>2E+07</b>	2E+06
4	3E+04	3E+04	3E+04	3E+04	2E+04	8E+04	4E+04	6E+04	2E+05
5	3E+04	1E+04	3E+04	3E+04	3E+04	5E+04	2E+04	3E+04	2E+04
6	2E+04	2E+04	2E+04	3E+04	3E+04	1E+04	2E+04	1E+04	3E+04
7	4E+04	2E+04	1E+04	3E+04	2E+04	2E+04	1E+04	2E+04	3E+04
8	5E+04	9E+03	2E+04	2E+04	1E+04	2E+04	2E+04	2E+04	1E+04

$$K - L = \tau_4 = \frac{\lambda_4}{\lambda_2} \tag{20}$$

The determination of the L-moments (MML) and L-moment quotients in hydrological studies of a given region can help in the treatment of data consistency, regional analysis and the identification of homogeneous regions. The advantage of this method is that it requires less computational effort to solve systems of equations (Naghettini and Pintpo 2007). The use of this methodology allows the use of the H test, which uses the L-moment quotients to test the homogeneity of regions classified as homogeneous.

**Heterogeneity Test H**

The measure of heterogeneity *H* (Eq. 21), 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 and expected variability of a homogeneous region based on L-statistics. This measure assists in determining the homogeneity of the regions formed in the cluster.

$$H = \frac{(V - \mu_v)}{\sigma_v} \tag{21}$$

where *V* is the sample-weighted standard deviation for *CV-L*,  $\mu_v$  is the arithmetic mean of the statistics *V<sub>j</sub>* obtained by simulation, and  $\sigma_v$  is the standard deviation between the values of the dispersion measure of the simulated samples (*n<sub>sim</sub>*), which are obtained using Eqs. 22, 23 and 24, respectively.

$$V = \left[ \frac{\sum_{i=1}^n n_i (t^i - t^R)^2}{\sum_{i=1}^n n_i} \right] \tag{22}$$

$$\mu_v = \frac{\sum_{j=1}^{n_{sim}} V_j}{n_{sim}} \tag{23}$$

$$\sigma_v = \sqrt{\frac{\sum_{j=1}^{n_{sim}} (V_j - \mu_v)^2}{N_{sim}^{-1}}} \tag{24}$$

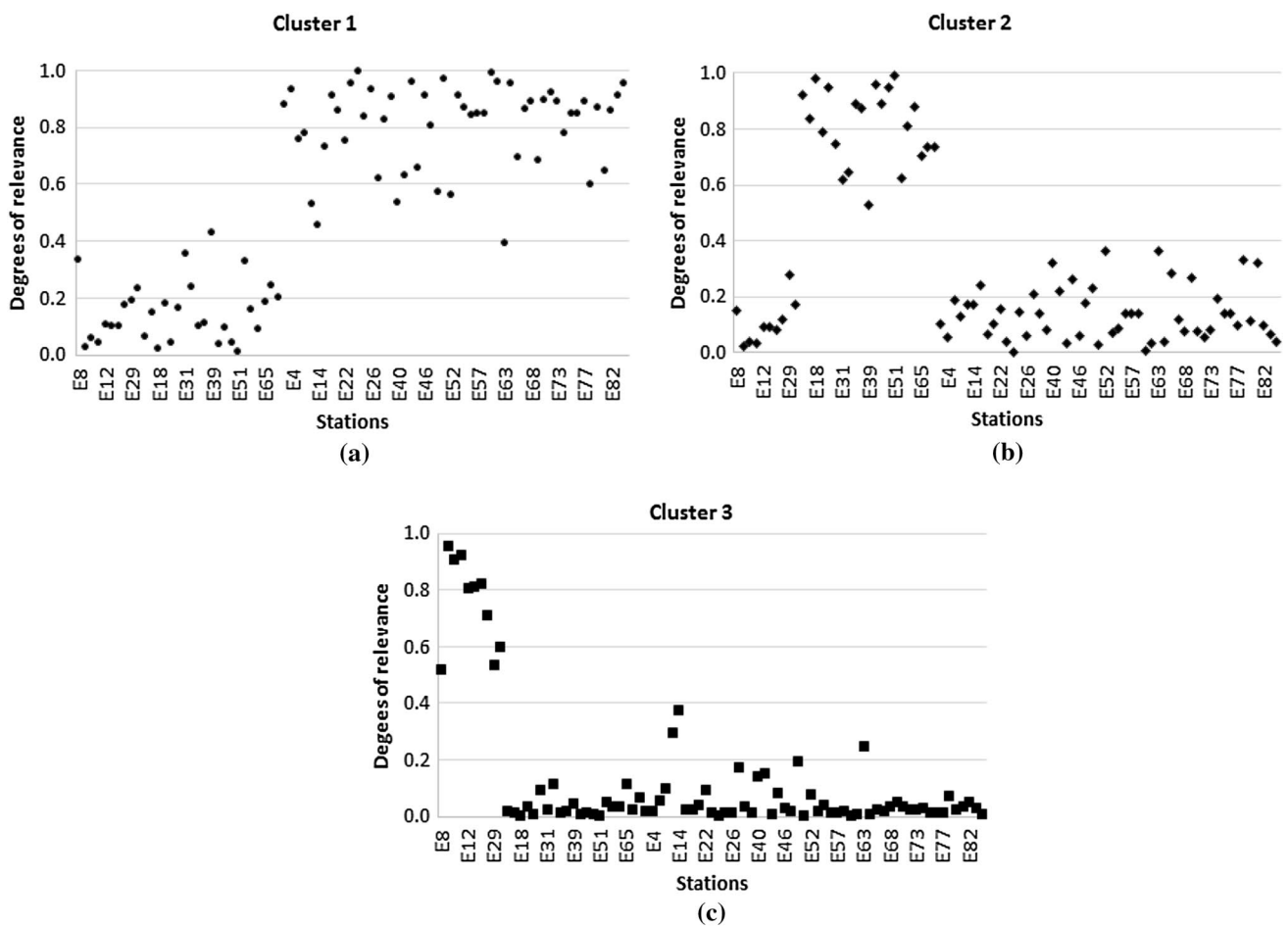


Fig. 4 Pertinence degrees of the stations by cluster

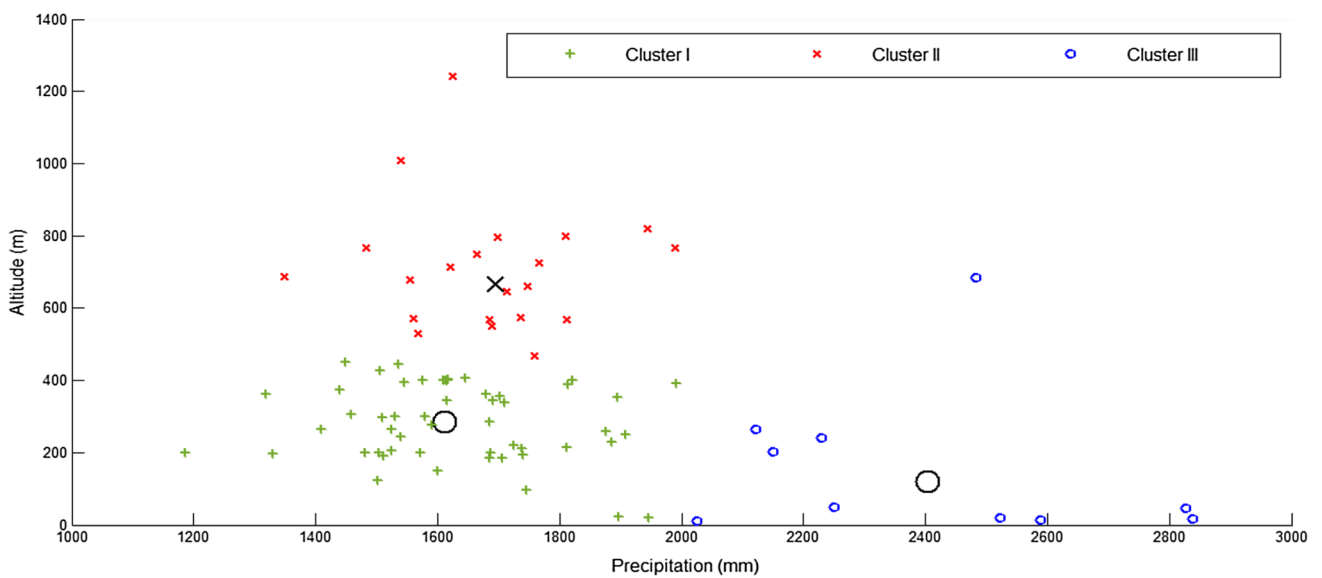


Fig. 5 Clusterings according to the precipitation and altitude characteristics



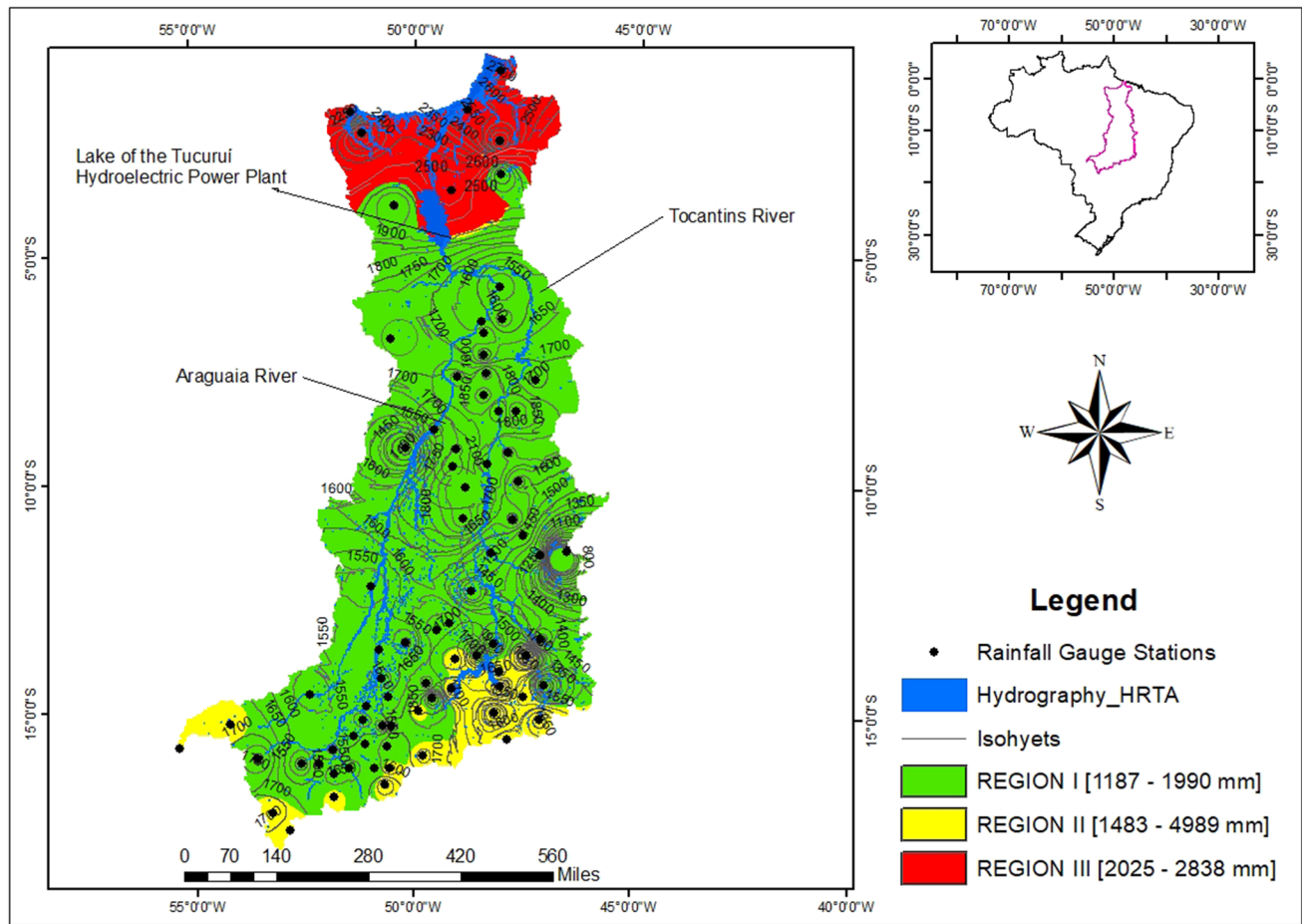


Fig. 6 Homogeneous regions of HRTA precipitation

The determination of  $H$  starts with the calculation of the weighted standard deviation  $V$  of  $CV$ - $L$ s of the observed samples. Then, the simulation of the homogeneous region of precipitation is simulated from the adjusted Kappa distribution (Eq. 25), which obtains the quotients of regional  $L$ -moments. Next, the statistics  $V_j$  ( $j = 1, 2, \dots, N_{sim}$ ) are calculated (Eq. 23) for all homogeneous regions.

$$F(x) = \left\{ 1 - h \left[ 1 - \frac{k(x - \xi)}{\alpha} \right]^{\frac{1}{k}} \right\}^{1/h} \quad (25)$$

Table 3 Average of L-moment quotients and the H value of the homogeneous regions

Region	NS	Average CV-L Obs.	Average CV-L Sim.	H
I	52	0.1087	0.1075	0.0047
II	21	0.1134	0.1133	-0.0049
III	10	0.1120	0.1125	-0.7874

where  $x$  is the studied variable,  $\xi$  is the position parameter,  $\alpha$  is the scale parameter, and  $k$  and  $h$  are the shape parameters. According to the test of significance, which was proposed by Hosking and Wallis (1997), if  $H < 1$ , the region is considered “acceptably homogeneous”, if  $1 \leq H < 2$ , the region is “possibly homogeneous,” and finally, if  $H \geq 2$ , the region should be classified as “definitely heterogeneous”.

## Results and discussion

### Formation of homogeneous regions

In total, 63 clusterings were performed by varying the fuzziness parameter from 1.2 to 2.0 and the number of clusters from 2 to 15. However, it was verified that the larger the number of clusters, the smaller the value of the  $PBM$  index. In this way, tests of up to 8 clusters were performed, ensuring the objectivity of the research, since the  $PBM$  index would tend to decrease with clusters greater than 8. The choice for the best cluster was determined by the  $PBM$  index, which

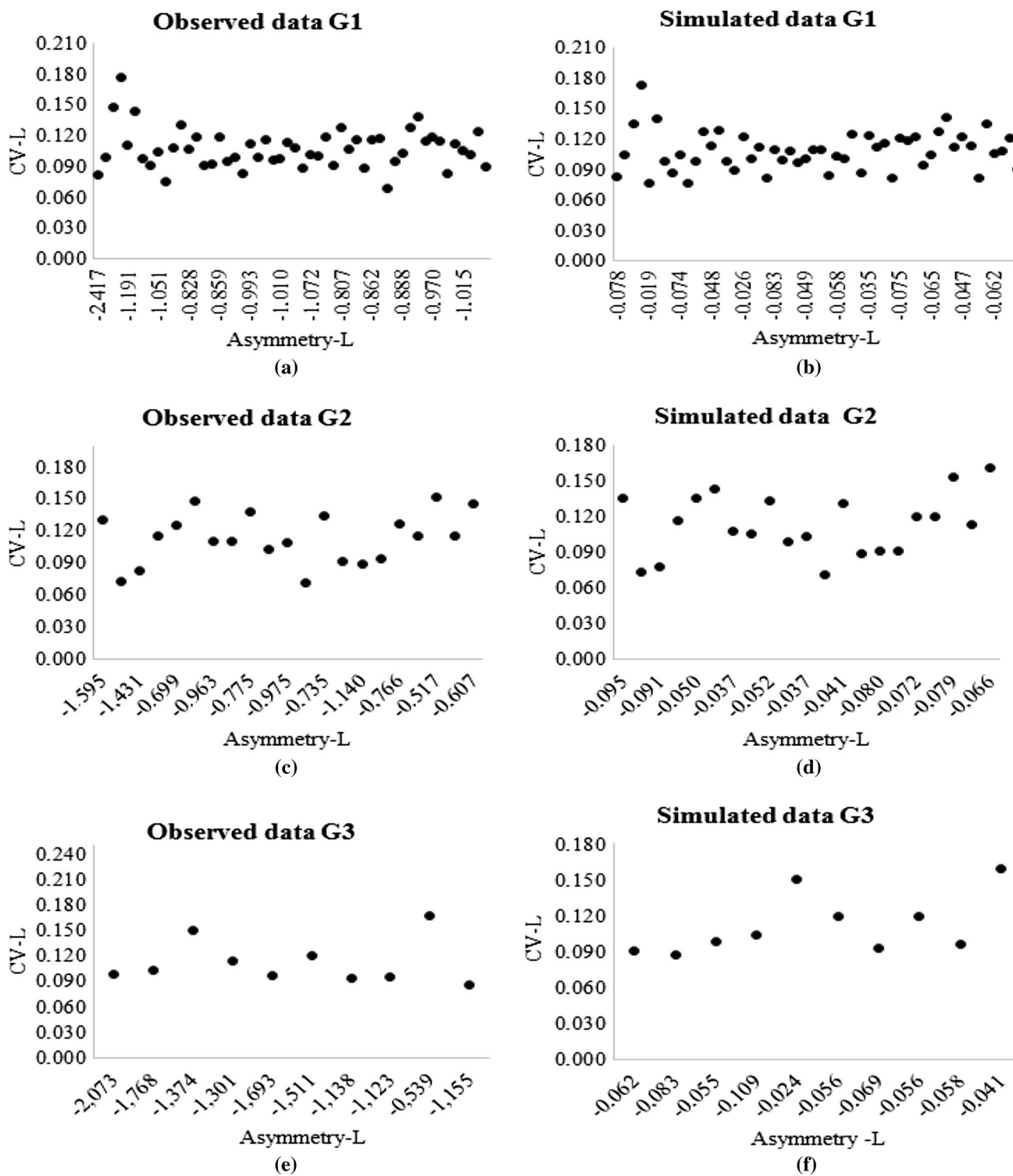


Fig. 7 Scatter plot of CV-L and L-Asymmetry quotients

presented a higher index (Fig. 3) in the formation of three clusters with a fuzziness parameter equal to 1.9 (Table 2).

One of the results from the FCM algorithm is the degree of the pertinence of the clustered elements. This degree of

pertinence refers to the probability that an element belongs to a particular cluster. Thus, all of the rainfall gauge stations, which are represented by their characteristics of mean annual precipitation, altitude, and location, are presented a

pertinence degree for each cluster. For example, the Acampamento IBDF Station (E8) has for Clusters 1, 2, and 3 a degree of pertinence of 0.34, 0.15, and 0.52. According to Mingoti (2005), this station has an approximate 52% probability of belonging to Cluster 3. Thus, the decision to allocate the station to a given cluster is due to its degree of pertinence (Fig. 4).

The clusters formed represent the homogeneous regions of precipitation. Region I is formed by 52 stations, Region II is formed by 21 stations, and Region III is formed by 10 stations according to their pertinence degrees. In the formation of the clusters by the Fuzzy c-means, the rainfall stations are clustered considering the similarity between the elements of the cluster, according to the characteristics involved in the clustering analysis, in which each cluster has its clustering centre (Fig. 5).

Region I is formed by the rainfall stations with a mean of 1625 mm, a minimum of 1187 mm, and a maximum of 1990 mm. These stations are concentrated in the central and south-western portion of the HRTA, specifically in the sub-basins of Alto Tocantins and Araguaia, where the Cerrado biome dominates the tropical climate with a low rainfall index. Region II is formed by stations with average annual precipitations of approximately 1700 mm, a minimum of 1349 mm and a maximum of 1989 mm. Most of the stations in this cluster are distributed in the south and south-east portions of the HRTA. The predominant biome in this region is also the Cerrado. Region III is formed by stations that present higher volumes of precipitation, with an average of 2400 mm, a minimum of 2025 mm, and a maximum of 2843 mm. The stations of this cluster are concentrated in the northern portion of the HRTA and in the region Baixo Tocantins, where the Amazonian biome predominates with a hot and humid climate and a high rainfall index (Fig. 6).

### Heterogeneity Test $H$

The calculation of the heterogeneity measure of the homogeneous regions was made by comparing the variances between the observed and simulated  $CV-L$ . In this way, the heterogeneity measure is calculated according to Eq. 21. In the verification of the Heterogeneity Test  $H$ , a value of 0.047,  $-0.0049$ , and  $-0.7874$  was obtained for Region I, Region II, and Region III, respectively (Table 3), which confers acceptably homogeneous regions, since all  $H < 1$ .

The significance of the measure of heterogeneity can be visualized using the L-moment quotient diagrams (Fig. 7). In diagrams such as these, a possibly homogeneous region would have  $CV-L$  samples less dispersed than those obtained by simulation. In quantitative terms, this idea can be translated by the difference centred between the observed and simulated dispersions. The dispersion in the simulated regions, for the L-moment quotients, shows that there was

no dispersion of the data, and, therefore, there are no stations with mean values much greater or less than the expected values. Thus, the simulated and observed dispersions are similar and form an acceptably homogeneous region.

### Conclusion

The combined use of the Fuzzy c-means method, the  $PBM$  index, and the  $H$  Heterogeneity Test was satisfactory for the formation and validation of homogeneous regions of precipitation. The satisfactory results of the application of the methodology were indicated by the formation of distinct clusters, with well-defined homogeneous regions, showing the spatial variability of annual rainfall totals in the region. In addition to contributing to the understanding of the hydrological behaviour of the region, the formation of these homogeneous regions of precipitation will aid in regionalization studies and support the management and planning of water resources in the Hydrographic Region of Tocantins—Araguaia—HRTA that is of great importance for Amazon and Brazil.

**Acknowledgements** The authors thank the National Council of Scientific and Technological Development (CNPq) for granting the Master's scholarship.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

### References

- Bezdek JC (1981) Modified objective function algorithms in pattern recognition with fuzzy objective function algorithms. Kluwer, Norwell
- Brazilian Agricultural Research Corporation (EMBRAPA) (1994) National Center for Research of Cerrados—CPAC. Rainfall in the Cerrados. Brazilian Agricultural Research Corporation and National Center for Research of Cerrados, Brasília, Brazil (in Portuguese)
- Brazilian Agricultural Research Corporation (EMBRAPA) (2014) Levantamento da agricultura irrigada por pivôs centrais no Brasil. Brazilian Agricultural Research Corporation, Brasília (in Portuguese)
- Cox E (2005) Fuzzy modeling and genetic algorithms for data mining and exploration, 1st ed. Elsevier/Morgan Kaufmann. Hardcover (Morgan Kaufmann series in data management systems)
- Dikbas F, Firat M, Cok CA, Gungor M (2011) Classification of precipitation series using fuzzy cluster method. *J Climatol* 32:1596–1603
- Farsadnia R, Kamrood RM, Nia MA, Rodarres R, Bray TM, Hand D, Sadatnejad J (2014) Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing features maps. *J Hydrol* 509:387–397

- Fukuyama Y, Sugeno M (1989) A New method of choosing the number of clusters for the Fuzzy c-means method. In: Proceedings of fifth fuzzy systems symposium, pp 247–250
- Goyal MK, Gupta V (2014) Identification of homogeneous rainfall regimes in northeast region of India using fuzzy cluster analysis. *Water Resour Manag* 28:4491–4511
- Hosking J, Wallis J (1993) Some statistic useful in regional frequency analysis. *Water Resour Res* 29(2):271–281
- Hosking J, Wallis J (1997) *Regional frequency analysis: an approach based on L-moments*, 1st edn. Cambridge University Press, New York
- Loureiro GE, Fernandes LL, Ishihara JH (2015) Spatial and temporal variability of rainfall in the Tocantins–Araguaia hydrographic region. *Acta sci Technol* 37(1):89–98
- Mingoti SA (2005) *Data analysis using multivariate statistical methods* (in Portuguese). Editora UFMG, Belo Horizonte (**in portuguese**)
- Naghattini M, Pintpo EJA (2007) *Hydrology Statistics*, Ed. CPRM, Belo Horizonte, Brazil (**in Portuguese**)
- National Water Agency (ANA) (2009) *National water resources plan of the hydrographic region of Tocantins–Araguaia*. National Water Agency, Brasília (**in Portuguese**)
- Pakhira MK, Bandyopadhyay S, Maulik K (2004) Validity index for crisp and fuzzy clusters. *Pattern Recognit* 37:481–501
- Patil S, Stieglitz M (2011) Hydrologic similarity among catchments under variable flow conditions. *Hydrol Earth Syst Sci* 15:989–997
- Sadri S, Burn DH (2011) A fuzzy c-means approach for regionalization using a bivariate homogeneity and discordancy approach. *J Hydrol* 401:231–239
- Satyanarayana P, Sirvinas VV (2011) Regionalization of precipitation in data sparse areas using large scale atmospheric variables—a fuzzy clustering approach. *J Hydrol* 405:462–473
- Swain JB, Sahoo MM, Patra KC (2016) Homogeneous region determination using linear and nonlinear techniques. *Phys Geogr* 37(5):361–384
- Wazneh H, Chebana F, Ouarda TBMJ (2013) Depth-based regional index-flood model. *Water Resour Res* 49:7957–7972
- Windhan MP (1981) Cluster validity for fuzzy clustering algorithms. *Fuzzy Sets Syst* 5:177–185

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.