Revista Brasileira de Climatologia
Brazilian Journal of Climatology

OPEN ACCESS

COMPARISON OF HOMOGENEOUS REGIONS OF PRECIPITATION FROM TWO DISTINCT DATA SOURCE FOR THE STATE OF PARÁ-BRAZIL

Comparação de regiões homogêneas de precipitação a partir de duas fontes de dados distintas para o estado do Pará-Brasil

Comparación de regiones homogéneas de precipitación a partir de dos fuentes de datos distintas para el estado de Pará-Brasil

David Figueiredo Ferreira Filho
Programa de Pós-Graduação em Engenharia Civil pela Universidade Federal do Pará – PPGEC/UFPA
davydferreira@gmail.com

Jéssica Ramos Abreu Ferreira
Programa de Pós-Graduação em Engenharia Civil pela Universidade Federal do Pará – PPGEC/UFPA
jessicaramosabreu@gmail.com

Francisco Carlos Lira Pessoa
Docente no Programa de Pós-Graduação em Engenharia Civil pela Universidade Federal do Pará – PPGEC/UFPA
fclpessoa@ufpa.br

Abstract: The aim of this study was to compare the regionalization of precipitation carried out using the Fuzzy C-Means grouping technique, with two distinct data sources, one provided by the National Water Agency (ANA) and the other obtained through the Global Precipitation Climatology Centre (GPCC) meteorological satellite provided by the German National Meteorological Service (DWD), for 30 years (1986 to 2015), with the aim of verifying through statistical techniques, what will be the representativeness and differences of the regions formed by traditional and satellite. The non-hierarchical technique of Fuzzy C-Means was applied to the formation of the regions for the two data, in order to group the stations, later owned by the groupings, validation techniques (Dunn, Silhouette and PBM) were applied, with the aim of forming the best cluster for data analysis. Performance analyses were also performed, using statistical methods. As results, 2 homogeneous rainfall regions were found after the calculations of the validation indices, in which they were specialized in GIS environment. The southwestern portion of the state was where there was the greatest divergence between the analyzed data, in such a way that, in the homogeneous rainfall region formed by GPCC,

there was a greater concentration of region 2, while in the analysis formed by ANA data, there were fragments of region 2. The results of the statistical tests showed that comparisons between the two regions are acceptable, with small differences, but of great value for hydrological studies in the region.
Keywords: GPCC. ANA. Fuzzy C-Means. Validation Indices. Regionalization.

Resumo: O objetivo desse estudo foi comparar a regionalização da precipitação realizada através da técnica de agrupamento Fuzzy C-Means, com duas fontes de dados distintas, uma fornecida pela Agência Nacional da Água e Saneamento Básico (ANA) e outra obtida através do satélite meteorológico Global Precipitation Climatology Centre (GPCC) fornecido pelo Serviço Meteorológico Nacional Alemão (DWD), durante 30 anos (1986 a 2015), com o objetivo de verificar, através de técnicas estatísticas, qual será a representatividade e as diferenças das regiões formadas por dados tradicionais e satélite. A técnica não-hierárquica de Fuzzy C-Means foi aplicada à formação das regiões para os dois dados, a fim de agrupar as estações, sendo aplicadas técnicas de validação (Dunn, Silhouette e PBM), com o objetivo de formar o melhor agrupamento para análise dos dados. Análises de desempenho também foram feitas, usando métodos estatísticos. Como resultados, 2 regiões homogêneas de precipitação foram encontradas após os cálculos dos índices de validação, sendo especializadas em ambiente SIG. A parte sudoeste do estado foi onde houve maior divergência entre os dados analisados, de tal maneira que, na região homogênea de chuvas formada pelo GPCC, houve uma maior concentração da região 2, enquanto na análise formada pelos dados da ANA, houve fragmentos da região 2. Os resultados dos testes estatísticos mostraram que as comparações entre as duas regiões são aceitáveis, com pequenas diferenças, mas de grande valor para estudos hidrológicos na região.
Palavras-chave: GPCC. ANA. Fuzzy C-Means. Índices de Validações. Regionalização.

Resumen: El objetivo de este estudio fue comparar la regionalización de la precipitación realizada a través de la técnica de agrupación Fuzzy C-Means, con dos fuentes de datos diferentes, una proporcionada por la Agencia Nacional de Agua y Saneamiento Básico (ANA) y otra obtenida a través del satélite meteorológico Global. Centro de Climatología de Precipitaciones (GPCC) proporcionado por el Servicio Meteorológico Nacional Alemán (DWD), durante 30 años (1986 a 2015), con el objetivo de verificar, a través de técnicas estadísticas, cuál será la representatividad y diferencias de las regiones formadas por tradicionales y datos satelitales. Se aplicó la técnica no jerárquica de Fuzzy C-Means a la formación de las regiones para los dos datos, con el fin de agrupar las estaciones, aplicándose técnicas de validación (Dunn, Silhouette y PBM), con el objetivo de formar la mejor agrupación. para el análisis de datos. También se realizaron análisis de rendimiento utilizando métodos estadísticos. Como resultado, se encontraron 2 regiones de precipitación homogéneas después de los cálculos del índice de validación, siendo especializadas en el entorno GIS. La zona suroeste del estado fue donde hubo mayor divergencia entre los datos analizados, de tal forma que, en la región de lluvia homogénea formada por el GPCC, hubo una mayor concentración de la región 2, mientras que en el análisis formado por la ANA datos, hubo fragmentos de la región 2. Los resultados de las pruebas estadísticas mostraron que las comparaciones entre las dos regiones son aceptables, con pequeñas diferencias, pero de gran valor para los estudios hidrológicos en la región.
Palabras clave: GPCC. ANA. Fuzzy C-Means. Índices de Validações. Regionalización.

Submetido em: 14/10/2022
Aceito para publicação em: 08/06/2023
Publicado em: 23/06/2023

1. INTRODUCTION

For hydrological studies, precipitation stands out as a climatic variable with great spatial-temporal variability, of great impact on hydrological cycles, providing information about the climate, becoming essential for the planning of human activities, promoting local development (DEZFULI, 2011; SHI et al., 2013).

However, recent studies have found that precipitation variability has been altered as a consequence of climate change (ÁVILA et al., 2014; SILVEIRA et al, 2016), being associated with these changes the difficulty of accurate data, low densities of stations, problems with measurements, errors presented in the data, difficulty of access to stations in some areas, as well as the maintenance of existing ones (FITZJARRALD et al., 2008; YAMANA; ELTAHIR, 2011; NOBRE et al., 2011), which makes hydrological studies difficult.

Such variabilities in the Amazon are also associated with increased deforestation, and the possible changes in regional and global climate have motivated a series of experiments and climate models, because there is great concern about changes in Amazonian land use by anthropic interference, whose trend points to changes and climatological and environmental consequences on a local and global scale (DAVIDSON et al., 2012).

According to Menezes and Fernandes (2016), there is a great spatial and temporal variability of precipitation in the state of Pará related to the action of climatic phenomena such as El Niño, and in the study, only 6 stations (Acará, Brasil Novo, Tracambeua, Cajueiro, Santo Antonio and Uruará) showed significant trends at a significance level of 5%, which may be related to various reasons, including failures in the records.

These problems have opened to the scientific community studies on the search for alternatives to traditional rainfall measurements, such as the use of meteorological satellites for hydrological studies (SARMADI; SHOKOOHI, 2014; SULOCHANA; CHANDRIKA; RAO, 2014; CAI et al., 2015).

The failures in measurements linked to advances in algorithms and data collection used by meteorological satellites open up studies on methods of obtaining hydrological information. Such complementary information can be obtained through the formation of homogeneous regions, whose main objective is to delimit areas with similar hydrological

behaviors, as explored by several works (ARELLANO; ESCALANTE, 2014; ASONG; KHALIQ; WHEATER, 2015; FAZEL et al., 2018).

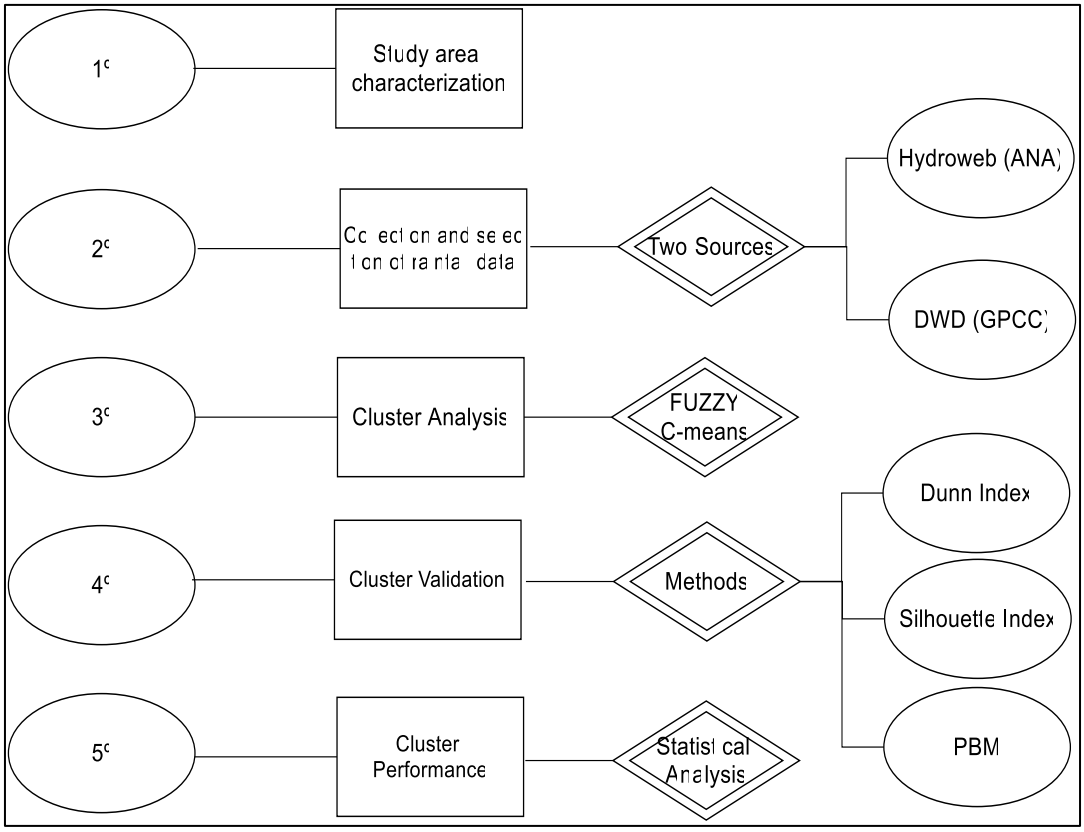
One of the techniques for clustering homogeneous regions is group formation by Fuzzy C-Means (AHANI; NADOUSHANI, 2016; NADOUSHANI; DEHGHANIAN; BAHRAM SAGHAFIAN, 2018), which is a well-known methodology and its importance is related to obtaining hydrological information in unmonitored locations. This technique is known as fuzzy (AGUADO; CANTANHEDE, 2010), and forms free choice groups and needs to be reapplied several times in order to avoid minimum sites and questions about the results, thus using validation indexes (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002), of which the most diverse indexes are found (PAKHIRA et al., 2004).

Therefore, the main objective of this work was to compare the regionalization performed through the Fuzzy C-Means clustering technique, but with two different data sources, one provided by the National Agency for Water and Basic Sanitation (ANA) and another obtained through the meteorological satellite Global Precipitation Climatology Centre (GPCC) provided by the German National Meteorological Service (DWD), for 30 years (1986 to 2015), with the aim of verifying through statistical techniques, that despite the error measurements in the traditional stations, what will be the representativeness and the differences of the regions formed by the traditional satellite and the satellite.

2. MATERIALS AND METHODS

For the development of this work, the methodology was subdivided according to Figure 1.

Figure 1 – Methodological Scheme.



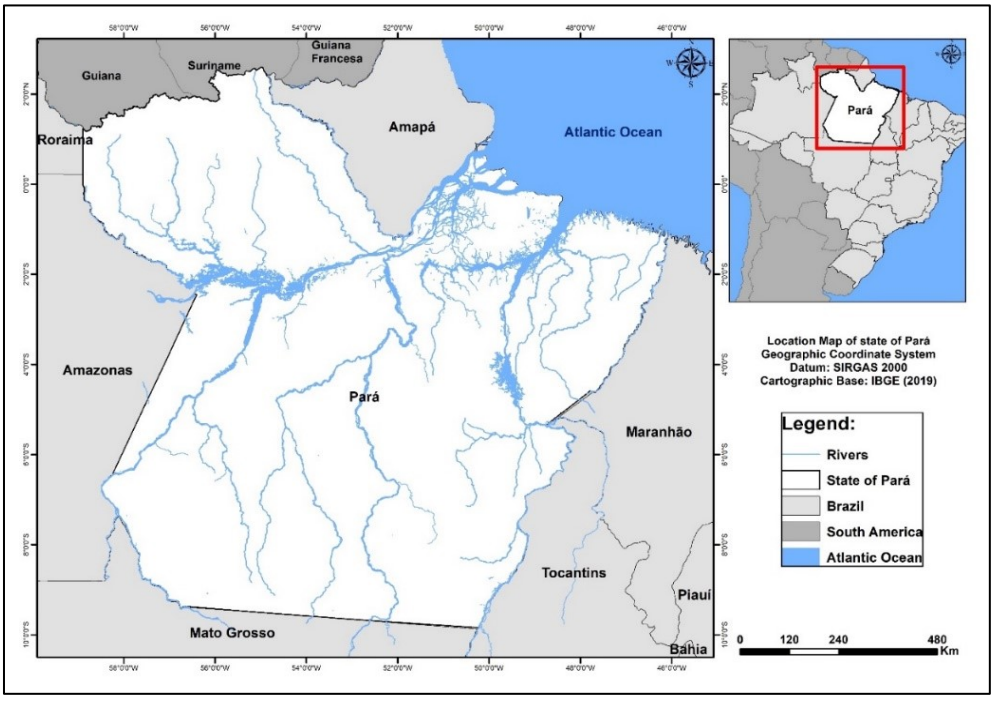
Source: Authors.

As a first step, the selection of the study area, Pará state, was made in order to later collect the average annual precipitation data for it. The area was chosen because the state has stations with poor records and/or absence of data in one part. The second step, data selection, consisted in obtaining data from two different sources, one provided by the National Water and Basic Sanitation Agency - ANA through the HidroWeb platform, and another provided by the Global Precipitation Climatology Center - GPCC. In possession of the data, the non-hierarchical method of Fuzzy C-Means was used, where the number of defined groups ranged from 2 to 10. In this phase, the variables Latitude, Longitude and Average annual precipitation of each station were grouped for each data source. Subsequently, with the clusters formed, validation indices, Dunn, Silhouette and PBM, were used to validate which group was best formed, for better analyses. Finally, statistical tests were applied to analyze the performance of the groups formed, in order to have a richer discussion and see the correlation between the different data sources.

2.1. Study Area (Área de Estudo)

The study area was the northern region of Brazil, more precisely the state of Pará (Figure 2), where, according to IBGE (2010), it has a population of approximately 8.5 million inhabitants. According to the Köppen-Geiger classification (1928), the region belongs to climate type "A", designated to tropical climates with high rainfall and average temperatures above 18 °C. In the equatorial subclass (Af), according to Menezes et al. (2015), the average annual temperatures are between 24 °C and 26 °C, with high rainfall, close to 2,000 mm per year in some analyses.

Figure 2 – Map of location of the study area.



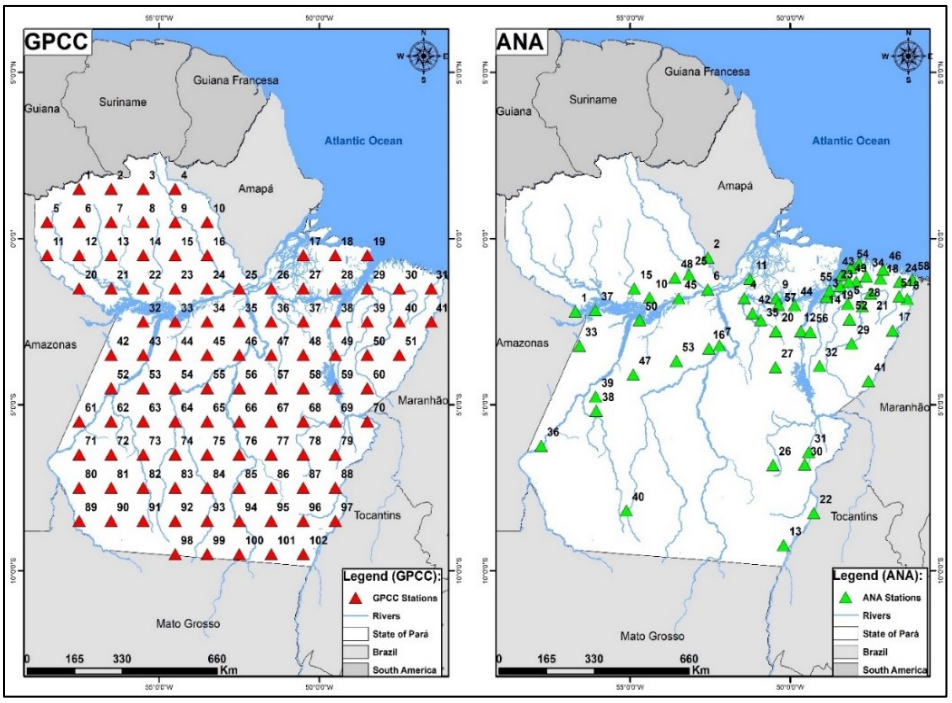
Source: Authors.

2.2. Data Selection (Seleção dos dados)

In order to compare the clustering, two different sources of precipitation data were selected, from the National Agency for Water and Basic Sanitation - ANA and the Global Precipitation Climatology Centre - GPCC, according to Figure 3. According to the World Meteorological Organization (WMO, 1984), a study period of at least three consecutive decades (30 years) is necessary for the calculation of climatological patterns. Therefore, the

selected study period was from January 1986 to December 2015 (30 years), which corresponds to the hydrological period of the region (January to December) for both data sources, using the average annual precipitation, to subsequently apply the methods.

Figure 3 – Location of ANA and GPCC stations.



Source: Authors.

2.2.1. Global Precipitation Climatology Centre – GPCC

In 1989 the WMO established the GPCC, operated by the German National Meteorological Service (DWD), to provide global precipitation analysis to help monitor and research the Earth's climate. The centre is a German contribution to the World Climate Research Programme (WCRP) and the Global Climate Observing System (GCOS), and is important in hydroclimatic and meteorological studies, especially in areas without reliable and accurate data sources (DINKU et al. 2008; RAZIEI et al. 2014).

The GPCC monitoring database is the result of the union of three data sources, the monthly precipitation totals derived from synoptic weather reports (SYNOP) of the German DWD, the US National Oceanic and Atmospheric Administration (NOOA) and the DWD climate

bulletins of the Japan Meteorological Agency (JMA) and the UK National Meteorological Service (Met. Office UK).

These are re-analysis data from climate models that fit field data and interpolate rainfall fields. According to Schneider et al. (2011), one of the goals of the GPCC is to provide the public with monthly and annual precipitation data, available at 1° x 1° and 2.5° x 2.5° latitude-by-longitude spatial resolution, from 1901 to the present day, interpolated and calculated by 13 statistical methods, available on the NOOA website.

2.2.2. National Water and Sanitation Agency

On the HidroWeb portal of the National Water and Basic Sanitation Agency (ANA), stations with historical series data were selected within the study period. The result was 57 rainfall stations located in the state of Pará, with the codes and geographic coordinates described in Table 1.

Table 1 - Pluviometric stations selected in the State of Pará.

Cod	Lat (°)	Long (°)	Cod	Lat (°)	Long (°)	Cod	Lat (°)	Long (°)
148010	-1,75	-48,87	**849000**	-8,26	-49,26	**855000**	-8,19	-55,12
151001	-1,79	-51,43	**47003**	-0,74	-47,85	**447001**	-4,29	-47,56
148009	-1,96	-48,21	**146010**	-1,29	-46,58	**251000**	-2,25	-51,18
152005	-1,53	-52,58	**152006**	-1,08	-53,16	**148012**	-1,09	-48,40
352001	-3,21	-52,21	**650001**	-6,82	-50,54	**149003**	-2,00	-49,86
146008	-1,82	-46,34	**350000**	-3,87	-50,46	**153000**	-1,80	-53,48
150003	-1,79	-50,43	**247003**	-2,04	-47,75	**47004**	-0,93	-47,10
154000	-1,77	-54,40	**348001**	-3,15	-48,09	**455004**	-4,09	-54,90
151002	-1,21	-51,26	**649001**	-6,79	-49,55	**253000**	-1,18	-53,60
249003	-2,79	-49,67	**649000**	-6,43	-49,42	**148003**	-1,30	-48,17
950001	-9,23	-50,21	**349002**	-3,83	-49,09	**254000**	-2,44	-54,71
148002	-1,44	-48,44	**356002**	-3,23	-56,59	**146005**	-1,73	-46,60
154001	-1,50	-54,87	**147010**	-1,13	-47,63	**248003**	-2,42	-48,15
352005	-3,31	-52,54	**250002**	-2,45	-50,92	**353000**	-3,68	-53,55
247005	-2,77	-46,80	**657000**	-6,24	-57,78	**48006**	-0,87	-48,11
147002	-1,20	-47,18	**151000**	-1,12	-52,00	**148011**	-1,57	-48,77
147007	-1,30	-47,94	**256001**	-2,15	-56,09	**249002**	-2,80	-49,38
250000	-2,79	-50,45	**555000**	-5,18	-56,06	**250001**	-1,99	-50,37
147011	-1,66	-47,49	**455003**	-4,75	-56,08	**146009**	-1,23	-46,19

Source: Authors.

However, the series presented gaps, so that to make the research feasible, the gaps were filled using the regional weighting method. According to Ishihara et al. (2014), the method is simplified, but its results are quite representative when applied to homogeneous climatic regions.

2.3. Fuzzy C-Means Clustering Analysis - FCM

The identification of homogeneous rainfall regions through the application of clustering techniques contributes to the understanding of the spatial-temporal variation of precipitation (AMANAJÁS and BRAGA, 2012). In this research, the non-hierarchical Fuzzy C-Means method proposed by Dunn (1973) and then generalized by Bezdek (1981) was used.

Fuzzy Clustering, as it is known, is characterized by the basic idea that a dataset X = {x1, x2, ..., xn} is divided into p groups, and the clustering result is expressed by degrees of relevance, such that each element can belong to a single group or more.

The algorithm is formed with an assumption, such that a database X = {x1,x2,x3,...,xn) is required, in which each point xk, k=(1,2,3,...,n) is a vector in $\Re p$, n is the total data in the database X and $\Re p$ represents a p-dimensional space of the real numbers. Thus, the partition matrix for the domain X is arranged in Equation 1:

$$Mfnc = \left\{ U \epsilon Ucn : Uik \epsilon [0,1], \sum_{i=1}^{c} Uik = 1, 0 < \sum_{k=1}^{n} Uik < n \right\} \quad \text{Equation (1)}$$

Where, Ucn is the true matrix group c x n, c is the number of groups that will be found, arranged 2≤c≤n, U is the Fuzzy partition matrix for the domain X and Uik is the degree of relevance of xk in a group i.

Thus, the task of generating an indicator to help verify convergence is assigned to the objective function (J), defined through Equation 2 below:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{p} (Uik)^m * d(Xk, Cj)^2 \quad \text{Equation (2)}$$

Where n is the number of data, p is the number of groups, m is the fuzzy parameter, d is the Euclidean distance between Xk,Cj, Xk is the data vector, where i=1, 2,..., n, represents a data attribute and Cj is the centre of a fuzzy cluster.

Then the objective function J is minimized and the relevance degrees Uik are generated according to Equation 3:

$$\mathrm{Uik} = \left[\sum_{\mathrm{k=1}}^{\mathrm{c}} \left(\frac{\mathrm{d(XkCj)}}{\mathrm{d(Xk, Cj}}\right)^{2/(\mathrm{m}-1)}\right] \quad \text{Equation (3)}$$

And Cj is a vector called centroid or prototypes (PEDRYCZ; VUKOVICH, 2004), which can be obtained through Equation 4:

$$Cj = \frac{\Sigma(\mathrm{Uik})^{\mathrm{m}}\ \mathrm{Xk}}{\Sigma(\mathrm{Uik})^{\mathrm{m}}} \quad \text{Equation (4)}$$

2.4. Groups Validation (Validação dos grupos)

All clustering processes produce a solution, even when the original data has no substructures (TAN; STEINBACH;) KUMAR, 2005), especially in the Fuzzy C-Means method, where it can generate several solutions because it is a free choice method, and for this, in this study, 3 validation indexes were used to check which is the best arrangement of clusters for the state of Pará, i.e., the one that best represents the behavior according to the input data from the two data sources, in order to avoid bad clusters for the analyses, and among those chosen to evaluate them are the indexes of Dunn (PAKHIRA et al. 2004), Silhouette (ROUSSEEUW, 1987) and PBM (PAKHIRA et al, 2004) indexes.

2.4.1. Dunn

The Dunn method (PAKHIRA et al., 2004) is defined as S and T two non-empty subsets in RN. Then, the diameter Δ(S) is defined as the distance δ between S and T:

$$\Delta(\mathrm{S}) = \max_{x,y\ E\ S}\{d(x,y)\} \quad \text{Equation (5)}$$

$$\delta(\mathrm{S},\mathrm{T}) = \min_{xES,yET}\{d(x,y)\} \quad \text{Equation (6)}$$

Where d (x; y) is the distance between points x and y. For any Dunn partition, you have defined the following index:

$$Vd = \min_{1\le s\le K}\{\min_{1\le t\le K, t\neq s}\{\frac{\delta i\{Cs, Ci)}{\max_{1\le k\le K}\Delta j(Ck)}\}\}$$ Equation (7)

Thus, the higher Vd, the corresponding is the number of clusters that maximises Vd is considered the optimal number of clusters (PAKHIRA et al., 2004).

2.4.2. Silhouette

The silhouette method, developed by Rousseeuw (1987), states that the silhouette width evaluates the quality of a clustering solution, considering both compactness and separation, the distance between data points in two neighboring groups. This method allows determining the appropriate number of groups, so that the value of k is chosen so as to provide the best average value of the Silhouette (GIL et al., 2015), thus according to the following equation:

$$s(i) = \frac{b_i - w_i}{max(b_i, w_i)}$$ Equation (8)

For b_i= 〖min〗 _k (B_(i,k)). Where Wi is the average distance of the i-th point to other points in the same group and B_(i,k) is the average distance of the i-th point to points in another group k. Thus, by presenting a positive unit value, the points are correctly arranged, on the other hand, by being characterized by the value zero, it becomes impossible to identify to which group they belong and, finally, by presenting a negative unit value, the points are probably allocated to the wrong groups.

Thus, the average silhouette width of a group $\bar{}$(k) for all i in a given group is defined as the average of all individual silhouettes, where n is the number of objects in the data set, according to the following equation:

$$\bar{s}(k) = \frac{1}{n}\sum_{i=1}^{n} s(i)$$ Equation (9)

2.4.3. PBM

The PBM index is defined as the product of three factors (Equation 10), where the objective is to maximise it to obtain the actual number of clusters, in other words, the maximum value for the best partition (PAKHIRA et al., 2004).

$$PBM(k) = (\frac{1}{k} * \frac{E1}{Ek} * Dk)^2$$ Equation (10)

Where, k is the number of clusters.

The first factor decreases as the value of K increases, reducing the value of the index. The second factor consists of the ratio E1 (center point of the data set), which is constant for a given data set, and EK, which decreases as the value of K increases. So, the value of the PBM index increases as the value of EK decreases. This, on the other hand, indicates that the formation of more clusters, which are naturally compressed, should be encouraged. Finally, the third factor (DK), which measures the maximum separation between a pair of clusters, increases with the value of K.

Thus, the factor E1 is the sum of the distances of sample units from the geometric centre of all samples (Equation 11), this factor does not depend on the number of initial clusters.

$$E1 = \sum_{t-1}^{n} d(x(t), Wo)$$ Equation (11)

Where, d is the Euclidean Distance, x(t) is the data vector, Wo is the centre of a cluster. With E1, the value of Ek (Equation 12) is calculated, determined by the distance between K clusters and weighted by the pertinence value corresponding to each sample for the cluster.

$$Ek = \sum_{t=1}^{n} \sum_{i=1}^{k} Ui(t) d[x(t) x\ Wi]^2$$ Equation (12)

Where Ui(t) = Degrees of relevance, and Dk represents the maximum separation of each pair of clusters (Equation 13).

$$Dk = \max_{i,j=1...k} (d\ (Wi, Wj))$$ Equation (13)

To select the best cluster, the maximum value of the PBM index (Equation 14), i.e. the best partition (argmax(PBM(k)) must be achieved).

$$K = argmax\ (PBM(k))$$ Equation (14)

2.5. Group Performance (Performance dos agrupamentos)

To analyze the performance of the estimates of homogeneous regions formed between ANA stations and the GPCC satellite, a mesh of one thousand points generated in the interpolation maps was used for the correlation. In addition, some statistical criteria were applied, namely: the Correlation Coefficient (R) and Determination (R^2) used to measure the overall agreement; the Mean Square Error (MSE) and Root Mean Square Error (RMSE) used to verify error and data bias; and the Nash-Sutcliffe Coefficient (NASH) used to evaluate the efficiency among the estimates (Table 2).

Table 2 - Statistics to estimate performance.

Method	Equation	Interval	Good Value	Nº Equação
R	$R = \frac{\sum(Pi - \bar{P}\imath)(Piy - \overline{P\imath y})^2}{\sqrt{\sum(Pi - \bar{P})^2}\sqrt{\sum(Piy - \overline{P\imath y})^2}}$	-1 to 1	1	(15)
R^2	$R^2 = \frac{\sum(Piy - \bar{P})^2}{\sum(Pi - \bar{P})^2}$	0 to 1	1	(16)
MSE	$MSE = \frac{\mathbf{1}}{\mathbf{N}}\sum(Pi - \boldsymbol{Piy})^2$	0 to ∞	0	(17)
RMSE	$RMSE = \left[\frac{\mathbf{1}}{\mathbf{N}}\sum(Pi - \boldsymbol{Piy})^2\right]^{1/2}$	0 to ∞	0	(18)
NASH	$NASH = 1 - \frac{\sum_{i=1}^{n}(Pi - Piy)^2}{\sum_{i=1}^{n}(Pi - \bar{P})^2}$	- ∞ to 1	1	(19)

Note: Where Pi is ANA precipitation, Piy is GPCC precipitation.

Source: Authors.

Pearson's R coefficient was determined by (Eq. 1) and can range from -1 (negative perfect correlation), +1 (positive perfect correlation) to 0 (no correlation). Thus, as suggested by Rencher and Christensen (2012), R correlations can be: 0.00 to 0.19 (very weak); 0.20 to 0.39 (weak); 0.40 to 0.69 (moderate); 0.70 to 0.89 (strong); 0.90 to 1.0 (very strong). The coefficient of determination (R^2) is a statistical criterion that measures the proportion of change in Y (dependent variable) that can be explained by the X variable (independent variable). The value of R^2 can range from 0 to 1, therefore $0 \leq R^2 \leq 1$ (NASH; SUTCLIFFE, 1970).

Another criterion used for the adjustment of the results was the Nash-Sutcliffe (NASH) coefficient (NASH; SUTCLIFFE, 1970). Thus, a classification suggested by Pereira et al. (2016) and Costa et al. (2019) was used: good values (NASH > 0.75); acceptable (0.36 ≤ NASH ≤ 0.75); inadequate (NASH < 0.36). The RMSE was applied to portray the difference between precipitation values (COSTA et al., 2019).

3. RESULTS AND DISCUSSION

From a simulation comparison using the mean annual precipitation as dependent variable generated from two data sources, ANA and GPCC, for the state of Pará, with the application of non-hierarchical Fuzzy C-Means tests from 2 to 10 groups, and taking as input data the independent variables - latitudes, longitudes and precipitation, for each station, we obtained the homogeneous regions (or homogeneous groups) of precipitation for the state.

The formation of these groups was validated with validation indices, since the non-hierarchical technique applied can generate results without substructures for analysis (Tan; Steinbach; Kumar, 2005), in which coefficients were generated as those presented in Table 3, also applied by Parchure and Gerdam (2019) in Mumbai, India, and different from those applied by Menezes, Fernandes and Rocha (2015) for the state of Pará.

Table 3 - Result of the validation indexes of GPCC and ANA data.

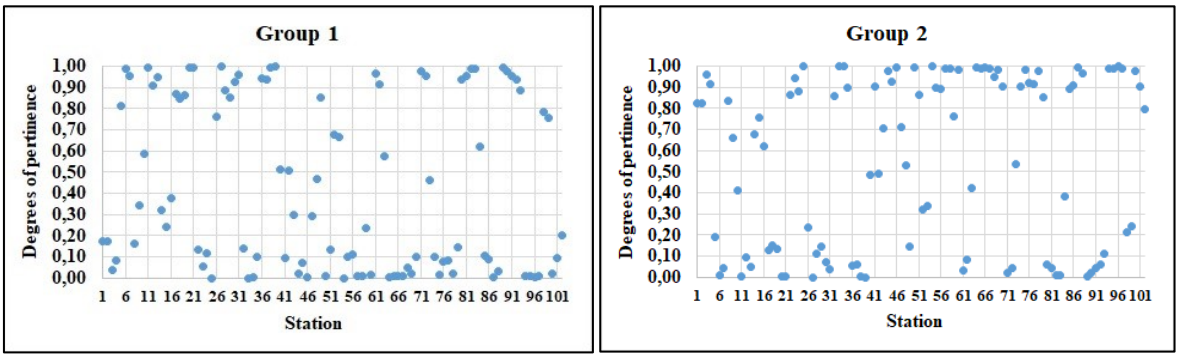
Nº of Groups	GPCC			ANA		
	Dunn	Silhouette	PBM	Dunn	Silhouette	PBM
2	**0.13**	0.33	**2.98**	0.17	**0.42**	**3.46**
3	0.11	0.33	2.41	**0.19**	0.37	2.51
4	0.11	**0.40**	2.14	0.09	0.27	1.83
5	0.09	0.35	2.16	0.09	0.36	2.90
6	0.11	0.33	2.2	0.14	0.35	2.57
7	0.11	0.32	2.02	0.15	0.41	2.55
8	0.11	0.31	1.71	0.15	0.39	2.28
9	0.11	31	1.71	0.15	0.39	2.28
10	0.11	0.26	1.28	0.17	0.36	1.96

Source: Authors.

The choice of the best cluster was defined by the highest values of the coefficients, since all tend to maximize the best result, thus, for the GPCC data, the Dunn and PBM indexes generated an ideal cluster for 2 clusters, with respective values of 0.13 and 2.98, and in the Silhouette index, 0.40, with ideal formation for 4 clusters. For the data provided by ANA, the Dunn index resulted in a coefficient value of 0.19, Silhouette and PBM showed a better cluster for 2 clusters, with indices of 0.42 and 3.46 respectively. Therefore, although the Dunn and Silhouette index presented different formations for their data sources, the PBM index pointed to 2 clusters, however, as in the tests the best formation regardless of the method was for 2 clusters, it is soon concluded that the ideal formation was for 2 arrangements.

In possession of the ideal number for the formation of the groups, for the GPCC data, they were organized according to the greatest degree of relevance of each station, that is, the stations were subdivided into groups with greater affinity among themselves. Figure 4 illustrates all GPCC stations with their respective degrees of relevance for the two groups, in which one can see that for group 1 some stations have greater affinity with group 2 and vice-versa, therefore the closer to 1, the greater their relationship with the group formed.

Figure 4 - Degrees of pertinence of the groups formed with the GPCC data source.



Source: Authors.

With the formation of the respective groups and, consequently, with the subdivision of the stations according to their degree of relevance (Table 4), two distinct groups were formed among themselves, but with similar internal individuals (stations), to later spatialize them in the GIS environment with the application of the Kriging technique.

For group 1 of GPCC, 44 stations were grouped, with an average annual precipitation of 2321.00 mm and for group 2, 58 stations with an average annual precipitation of 1920.98 mm.

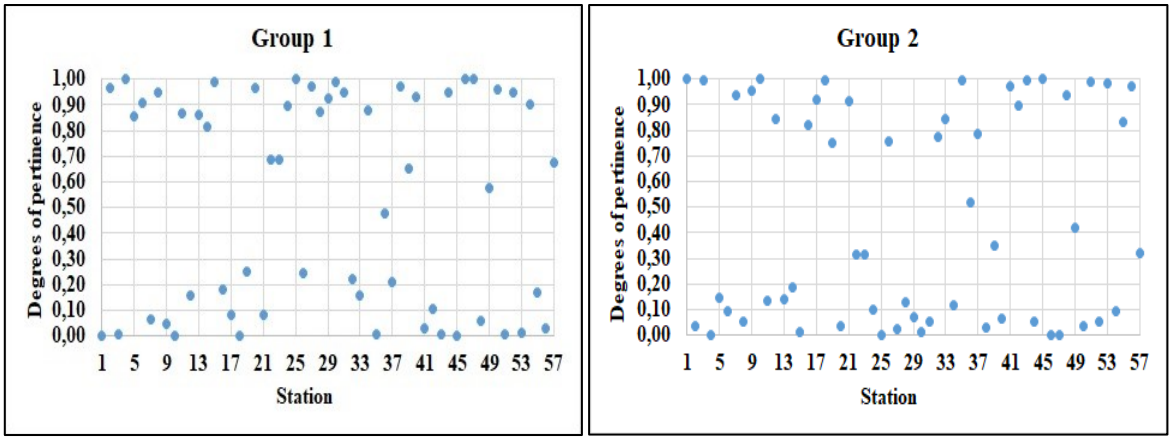
Table 4 - Degrees of relevance of each GPCC station with the formation of its respective group.

Group 1 - GPCC				Group 2 - GPCC			
Id	**Graus**	**Id**	**Graus**	**Id**	**Graus**	**Id**	**Graus**
5	0.8115	63	0.5759	1	0.8270	57	0.9919
6	0.9883	71	0.9770	2	0.8276	58	0.9905
7	0.9545	72	0.9554	3	0.9592	59	0.7656
10	0.5869	80	0.9393	4	0.9154	60	0.9825
11	0.9957	81	0.9537	8	0.8392	64	0.9949
12	0.9080	82	0.9919	9	0.6588	65	0.9919
13	0.9490	83	0.9919	14	0.6784	66	0.9924
17	0.8723	84	0.6188	15	0.7596	67	0.9883
18	0.8486	89	0.9934	16	0.6225	68	0.9511
19	0.8658	90	0.9775	22	0.8664	69	0.9809
20	0.9938	91	0.9558	23	0.9463	70	0.9021
21	0.9946	92	0.9379	24	0.8822	73	0.5350
26	0.7613	93	0.8859	25	0.9993	74	0.9022
27	0.9999	98	0.7871	32	0.8614	75	0.9848
28	0.8893	99	0.7578	33	0.9989	76	0.9232
29	0.8519	-	-	34	0.9978	77	0.9146
30	0.9270	-	-	35	0.8975	78	0.9769
31	0.9621	-	-	41	0.9041	79	0.8548
36	0.9450	-	-	43	0.7036	85	0.8944
37	0.9393	-	-	44	0.9793	86	0.9120
38	0.9932	-	-	45	0.9270	87	0.9953
39	0.9998	-	-	46	0.9963	88	0.9691
40	0.5152	-	-	47	0.7094	94	0.9915
42	0.5089	-	-	48	0.5315	95	0.9876
49	0.8526	-	-	50	0.9925	96	0.9985
52	0.6770	-	-	51	0.8647	97	0.9909
53	0.6643	-	-	54	0.9997	100	0.9784
61	0.9690	-	-	55	0.8980	101	0.9040
62	0.9138	-	-	56	0.8909	102	0.7988

Source: Authors.

The same procedure was applied to the data from the ANA stations, Figure 5, where they were shown with their respective degrees of relevance to the two groups formed, and the closer to 1, the greater the relationship between this and the group.

Figure 5 - Degrees of relevance of the groups formed with the data source ANA.



Source: Authors.

Thus, two distinct groups were also formed, but with similar internal individuals (stations), to later spatialize them, in which group 1 had 31 stations, with average annual precipitation of 2169.00 mm, and in group 2, 26 stations were grouped, with average annual precipitation of 2640.16 mm.

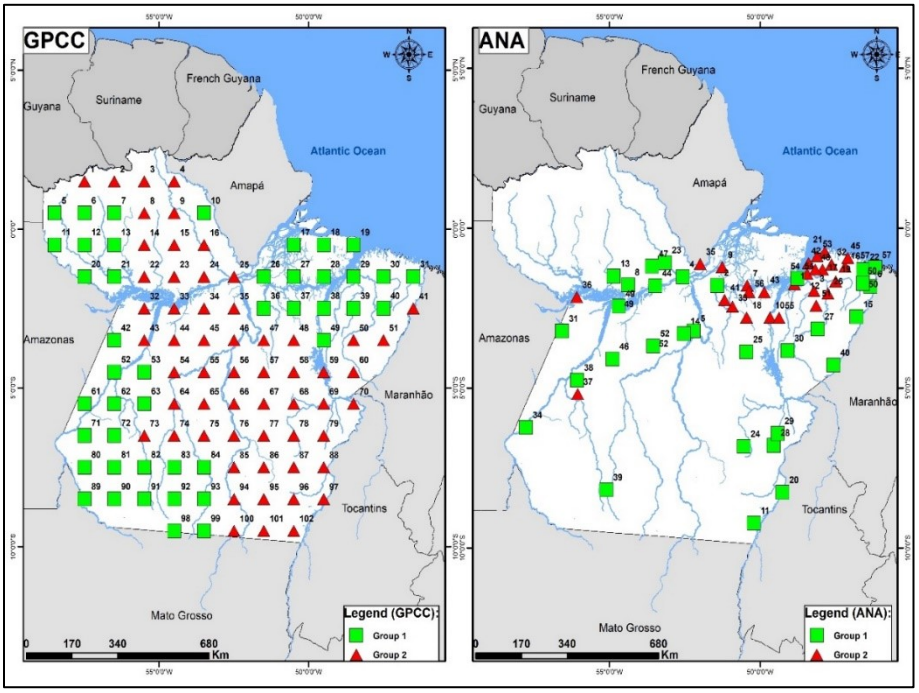
Table 5 - Degrees of relevance of each ANA station with the formation of its respective group.

Group 1 - ANA				Group 2- ANA			
Id	**Graus**	**Id**	**Graus**	**Id**	**Graus**	**Id**	**Graus**
2	0.9662	29	0.9276	1	0.9997	37	0.7872
4	0.9999	30	0.9862	3	0.9927	41	0.9706
5	0.8548	31	0.9475	7	0.9368	42	0.8958
6	0.9059	34	0.8806	9	0.9512	43	0.9943
8	0.9462	38	0.9697	10	0.9997	45	1.00
11	0.8657	39	0.6511	12	0.8426	48	0.9384
13	0.8617	40	0.9322	16	0.8204	51	0.9909
14	0.8144	44	0.9464	17	0.9166	53	0.9845
15	0.9896	46	0.9980	18	0.9965	55	0.8330
20	0.9645	47	0.9998	19	0.7508	56	0.9725
22	0.6869	49	0.5778	21	0.9150	-	-
23	0.6870	50	0.9616	26	0.7549	-	-
24	0.8981	52	0.9462	32	0.7758	-	-
25	0.9977	54	0.9040	33	0.8420	-	-
27	0.9735	57	0.6774	35	0.9955	-	-
28	0.8714	-	-	36	0.5204	-	-

Source: Authors.

Plotting these results in a GIS environment, without spatializing them, resulted in the formation of two groups, for two distinct data sources, for the state of Pará (Figure 6). Clearly the GPCC station grid proposes a more central homogeneous region, dividing in the western portion and part of the northeast another homogeneous region. For the ANA data, a portion of the northeast region is highlighted as homogeneous, and the remainder is considered another region.

Figure 6 - Formation of clusters by the highest degree of relevance of each station.



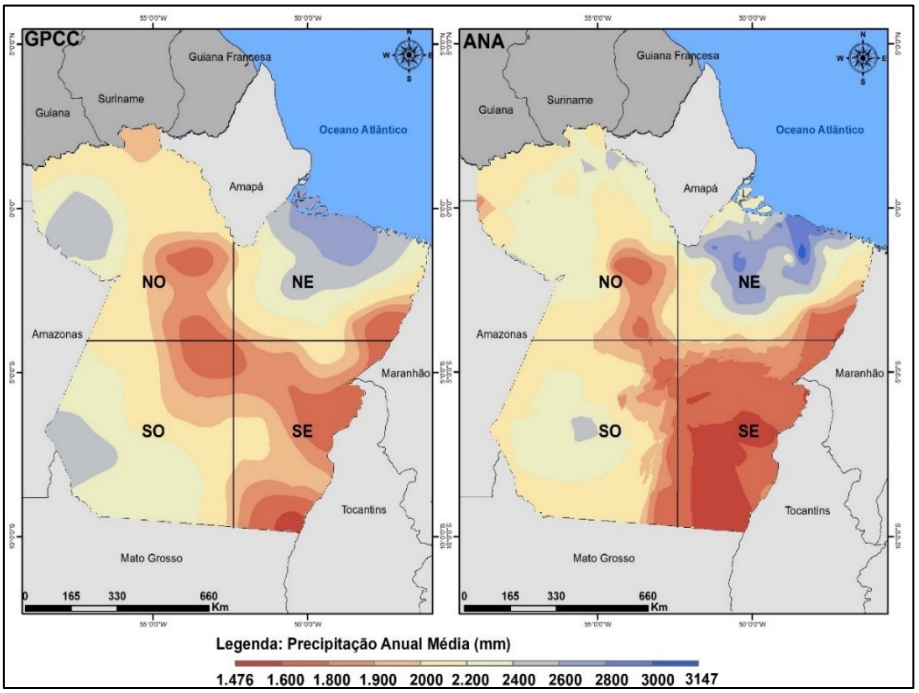
Source: Authors.

When these data were spatialized for the formation of homogeneous regions of precipitation, it was observed that both are well distributed throughout the study area (Figure 7), with greater similarity in the northeastern part of the state, where, according to Albuquerque et al. (2010), this fact is due to the performance of climatological phenomena, such as the Intertropical Convergence Zone (ITCZ), which operates in this region.

Ferreira Filho et al. (2019) performed comparative analyses of rainfall variability (Figure 7) for the same two data sources and found similar results to those obtained by Amaral et al. (2016) and Albuquerque et al. (2010), who found that the highest average annual rainfall was

in the northeastern part of the state, covering areas within these three mesoregions, as well as a portion in the lower Amazon.

Figure 7 - Long-term average annual precipitation for Pará (1986 to 2015) for both data sources.



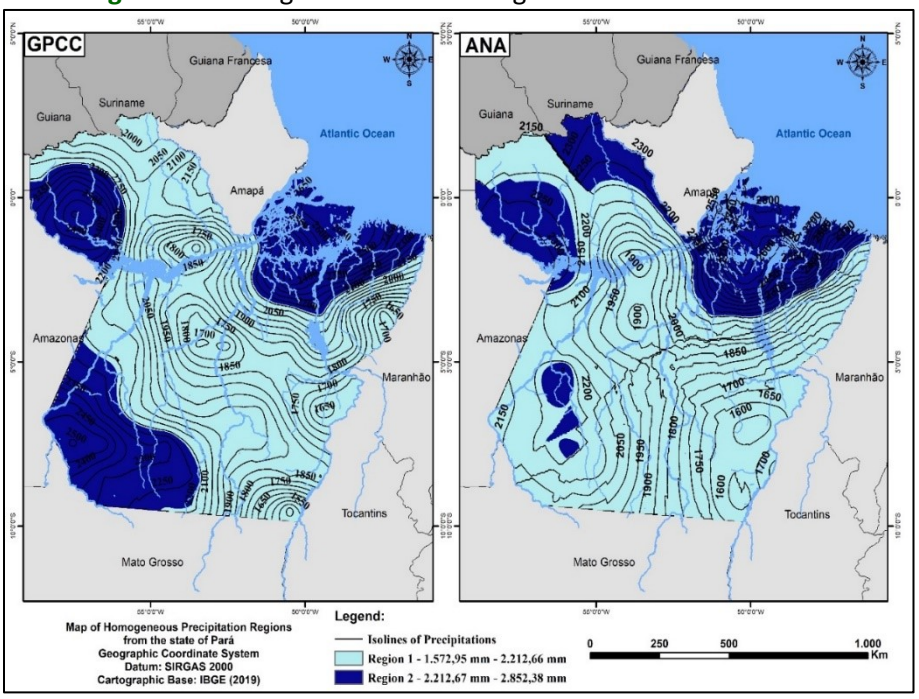
Source: Authors.

The northern region of the state, located in the northwestern portion, is considered a homogeneous rainfall region (PESSOA; BLANCO; MARTINS, 2011; FERREIRA FILHO et al., 2019a), for this reason, the behaviors in both data sources occur similarly. Pessoa, Blanco and Martins (2011) also stated that this behavior is directly related to the dense vegetation of the area, since it is considered a region little explored, contributing to high rates of humidity and, consequently, high rates of rainfall.

The southwestern portion of the state was where there was greater divergence between the analyzed data, so that in the homogeneous rainfall region formed by the GPCC, there was a greater concentration of region 2, while in the analysis formed by the ANA data, there were fragments of region 2. This fact can be justified in a study by Ferreira Filho et al. (2020), who stated that this portion of the state has areas of low density of ANA stations, while the GPCC has a uniform distribution of the data network in the region. Despite the differences

in the behaviour of the homogeneous regions in this area, they have quite similar rainfall indices, with differences of only 34 mm between them. With the lack of data for this portion, the GPCC data grid becomes more complete for obtaining information in this portion of the state.

Figure 8 - Homogeneous Rainfall Regions for the State of Pará.



Source: Authors.

In a study developed by Silva et al. (2018) in the Tapajós basin in the state of Pará, the formation of 3 homogeneous rainfall regions was observed, although they used a different method from the present study, but it is close to the result obtained using the GPCC data (2 homogeneous regions).

Visually the central portion of the state is better represented in both cases, since the station densities in the two data sources are distributed in this region, thus being able to affirm that it is the region with the greatest precipitation information. Despite these distinctions, both show similar behaviour.

A similar result was obtained by Limberger; Silva (2018) in which they used GPCC data and observed that they followed the monthly precipitation variability of ANA data, being more evident in the Northern and Central sub-regions of the Amazon region. In general, in this

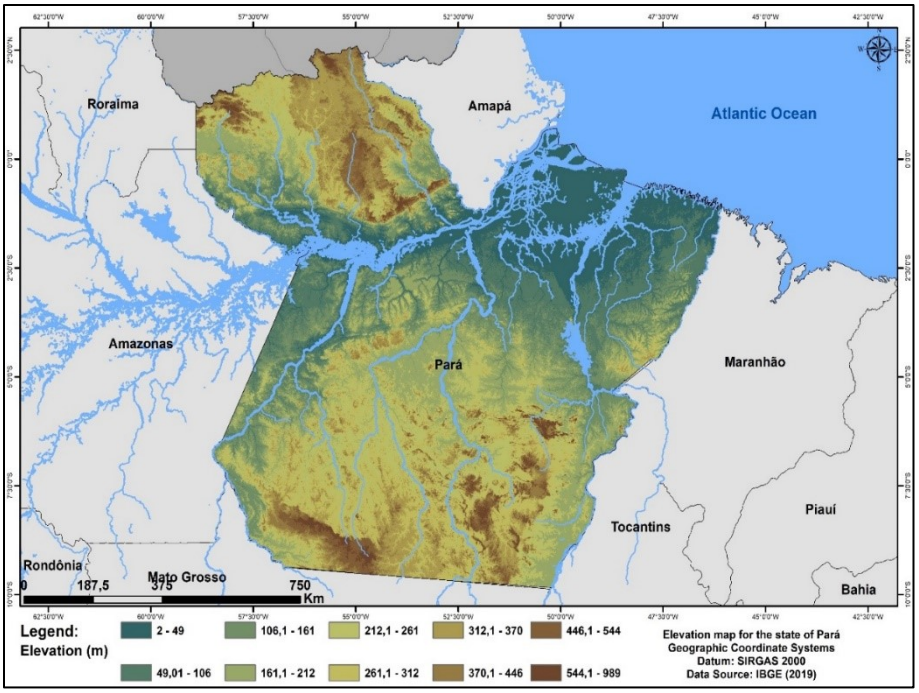
studied area, the GPCC products show comparable hydrological performance to the ANA product after real time, presenting the great potential for real time application.

When relating the regions formed with the vegetation present in the study area, it can be observed that in region 1, there is a diversity in the vegetation, where in the southeast portion, the lowest precipitation rates are concentrated along with the Arc of Deforestation, with the presence of sparse, sparse vegetation, and in some points, without vegetation, a fact due to the increase of cattle ranching, agriculture related to the expansion of soybean and timber production (COSTA; PIRES, 2010; LEMOS;SILVA, 2011), another factor that can associate this index is due to the presence of climatic phenomena, El Niño and La Niña (GONÇALVES et al. , 2016). When the central portion of the state is analyzed, an intermediate and dense vegetation is noted, going to the northwest region, another intrinsic characteristic of this group formed, which should be noted, that despite the variability of precipitation in this group, the variability of vegetation is also noted.

In relation to homogeneous group 2, the presence of intermediate and dense vegetation, in its territorial majority, can justify the formation of rainfall with higher precipitation rates, the main highlight being the Calha Norte region, as a previous attack. It is observed the concentration of ANA stations in the southern portion of the state, due to areas of difficult access, presence of dense vegetation and high altitudes (Figure 10), thus the deployment of monitoring stations in this area becomes difficult, another fact that justifies the use of GPCC grid data for hydrological studies.

When comparing regions with elevation (Figure 9), this was an input variable in the clustering model, therefore, it also influenced the results; therefore, region 1 was the one with the highest altitudes, another fact that justifies it having lower precipitation rates, and in view of homogeneous region 2, it can be observed that due to the presence of low altitudes, with the exception of the southwestern portion, the clustering occurs in areas closer to sea level.

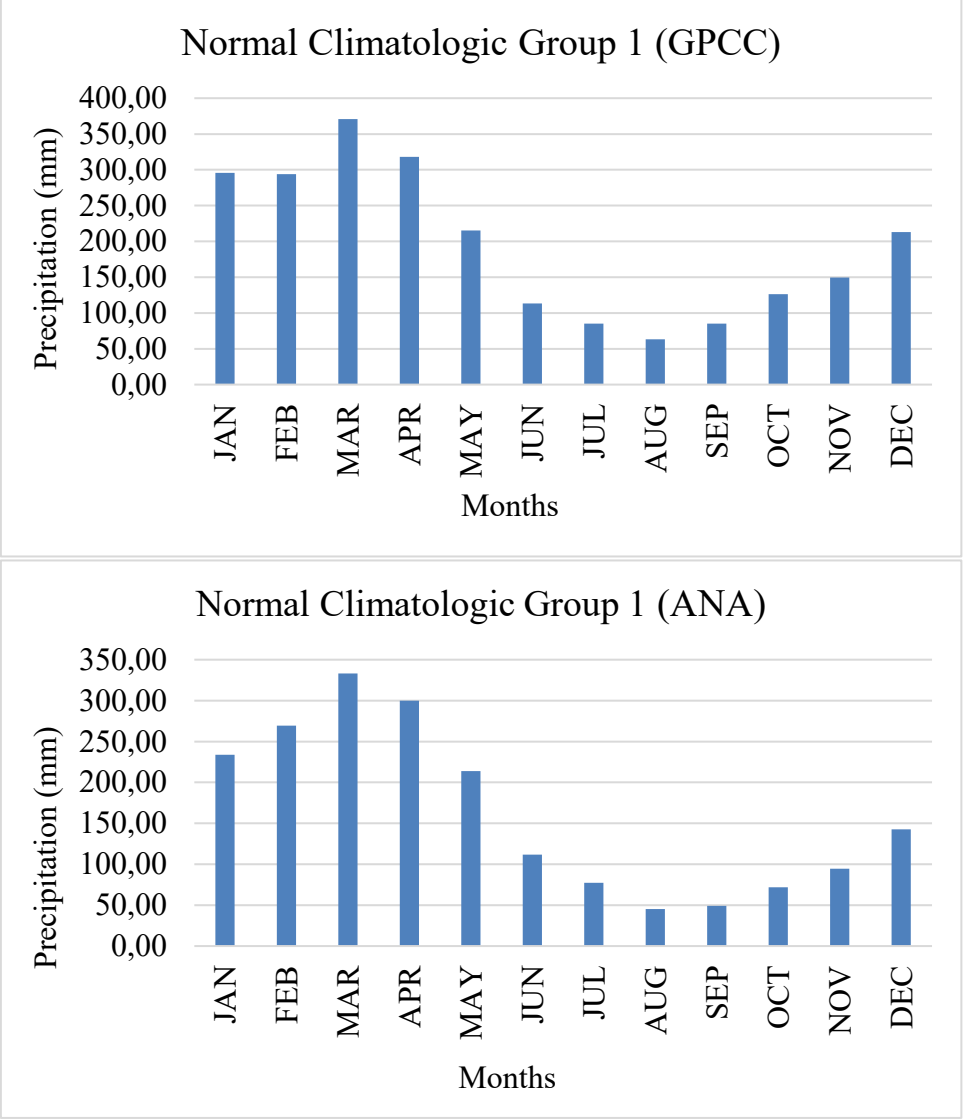
Figure 9 – Elevation map in metres for the state of Pará.



Source: Authors.

In the analysis of the rainfall distribution of the homogeneous region 1 (Figure 10), along the months, it can be observed that in both months the month with the highest rainfall index is registered in March, similar to that found by Amanajás and Braga (2012); Menezes, Fernandes and Rocha (2015) and Ferreira Filho et al. (2020), with the beginning of a rainy season from November to April, and a less rainy season from May to October. One of the few differences between the data formed by the two sources was for the month of January, where the GPCC obtained a higher rainfall index than the ANA data, a fact that is justified because the GPCC overestimated between January and August (FERREIRA FILHO et al., 2020).

Figure 10 - GPCC and ANA climate normals respectively for homogeneous region 1.



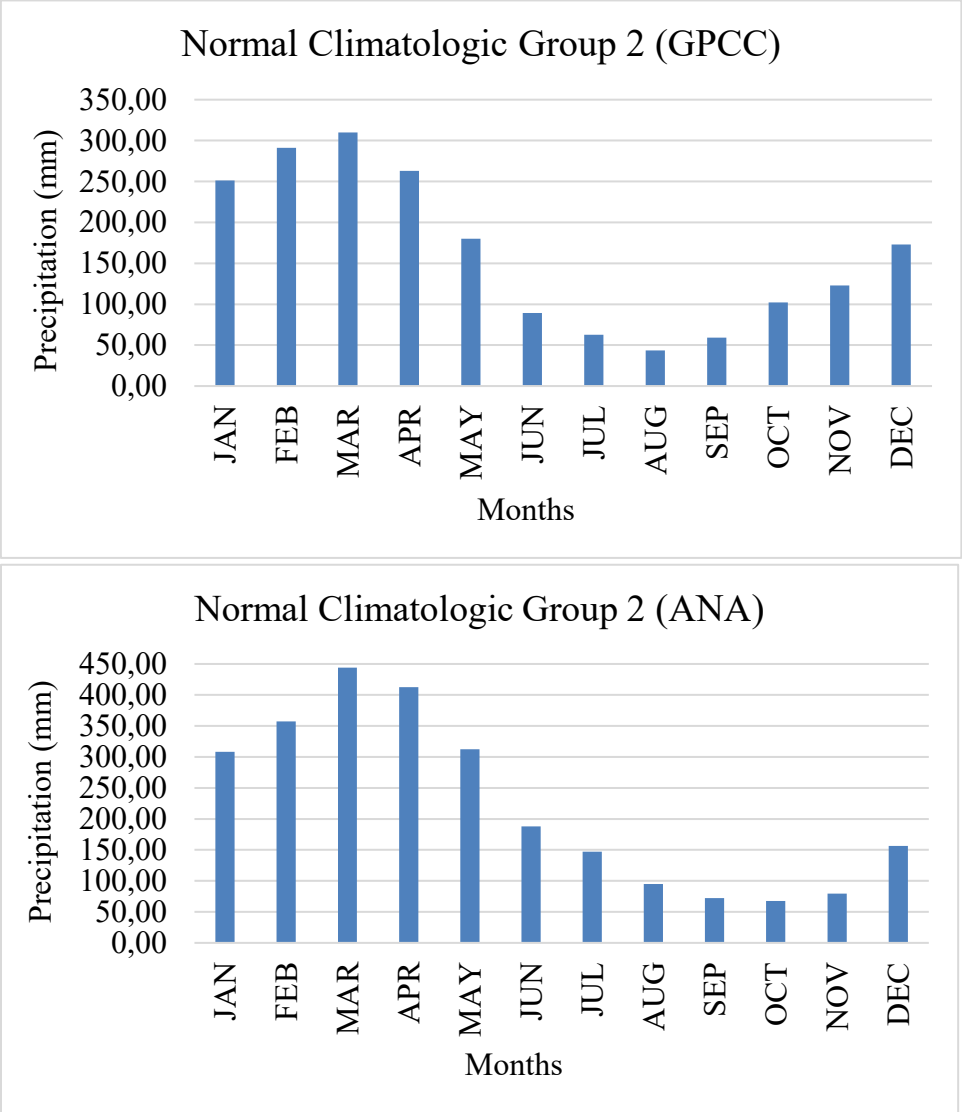
Source: Authors.

In a study applied by Menezes, Fernandes and Rocha (2015), the behaviour of the normal climatological distributions showed similarities with that of the research, thus confirming that the study was valid.

The same analysis was made for homogeneous region 2, represented by Figure 11, where March was also recorded as the rainiest month. In this region, the divergence between the normals occurred at the top of the least rainy month, for the GPCC, it was the month of August, identical to the study of Ferreira Filho et al., (2019), and for the ANA data, the month of October. It is worth noting that this region diverged more in the southwestern portion of the state, due to the low density of information and stations from ANA in relation to the GPCC

grid, so this may have been the greatest influence on the behavior of the region formed, along with the ZCIT (RIBEIRO et al., 1996).

Figure 11– GPCC and ANA climate normals respectively for homogeneous region 2.



Source: Authors.

Thus, to verify the performance of the estimates of the regions formed by ANA and GPCC, from the spatialization, a mesh of 1,000 random points was generated and the values of the precipitation points were extracted, in which the statistical tests were analyzed and applied (Table 6).

Table 6 - Results of the statistical tests on the performance of the regions formed.

Methods	Results	Interval	Good Value
R	0,82	-1 to 1	1
R^2	0,67	0 to 1	1
MSE	24936,40	0 to ∞	0
RMSE	157,91	0 to ∞	0
NASH	0,66	- ∞ to 1	1
Standard Error	143,10	-	-

Source: Authors.

Several widely used statistical indexes were used to quantitatively evaluate the accuracy of the values of the data generated by ANA and GPCC with respect to precipitation, showing that the performance results presented a good correlation between the data, in which the R coefficient presented a value considered positive and good, R^2 as a moderate value, with good results between the data. As for the errors, a value of 143.10 mm was obtained as standard error, which is acceptable, considering that it is annual average data. The SBM and RMSE methods presented values of 24,936.0 and 157.91. The MSE is sensitive to large errors, because it squares the individual differences, but it is always positive, and the closer to zero, the better it indicates a perfect simulation. In this case, as we worked with annual data, the MSE presented a value considered good in statistical terms, evidenced in the RMSE statistical test, with a root value of 157.91.

Finally, the NASH coefficient, 0.66, according to the classification of Pereira et al. (2016) and Costa et al. (2019), presented an acceptable value, for statistical performance analysis. Therefore, the coefficients showed that the regions formed have a good correlation between the analyzed data, finally, illustrated in the normal probability plot of the data.

4. CONCLUSION

The study made important contributions to the formation of homogeneous rainfall regions for the state of Pará. It was found that hydro-meteorological monitoring in the region presents problems such as small data coverage, low density of stations and others with

incomplete series, which may be one of the main influences in the distinction of the regions presented.

As the GPCC data present a more complete grid of information, it is suggested, as researched and seen in other works, the use of these to obtain precipitation data in areas with little data available from the National Water and Basic Sanitation Agency - ANA.

This factor was evidenced through statistical techniques (R, R^2, MSE, RMSE and NASH), whose values presented a good correlation between the regions formed, with the data used for reanalysis (GPCC) and ANA, analyzing mainly the southwestern portion of the state, with low information from monitoring stations.

For both regions 1 and 2 formed, in both data sources, the climatological norms obey their variability as verified in other studies presented, demonstrating that despite the state being considered large, comparisons between the two data sources have become important tools on hydrological studies.

It is worth noting that any data sets, observed or simulated, whether of reanalysis or not, may present problems of data consistency due to failures in obtaining or recording, as well as the statistical tools employed, among other factors. For this reason, the comparison between the two sources becomes of great value.

In general, the results for Pará state indicated that the GPCC data performed well in relation to the observed data, leading it to indicate the use of GPCC data as an alternative source of precipitation data in locations without rainfall stations or long data series.

REFERENCES

AGUADO, A. G.; CANTANHEDE, Marco André. **Lógica Fuzzy**. 2010. Available at: < http://www.sysrad.com.br/redmine/attachments/1843/Artigo_logicaFuzzi.pdf >.

AHANI, A.; NADOUSHANI, S. M. Assessment of some combinations of hard and fuzzy clustering techniques for regionalization of catchments in Sefidroud basin. **Journal of Hydroinformatics,** v. 18, n. 6, p. 1033–1054. 2016.

de ALBUQUERQUE, M. F.; DE SOUZA, E. B.; DE OLIVEIRA, M. D. C. F.; DE SOUZA JÚNIOR, J. A. Precipitação nas mesorregiões do estado do Pará: climatologia, variabilidade e tendências nas últimas décadas (1978-2008). **Revista Brasileira de Climatologia,** n. *6*. 2010.

AMANAJÁS, J. C.; BRAGA, C. C. Spatio-temporal rainfall patterns in eastern Amazônia using multivariate analysis. **Brazilian Journal of Meteorology,** v. 27, n. 4, p.423 – 434. 2012.

ARELLANO-LARA, F.; ESCALANTE-SANDOVAL, C. A. Multivariate delineation of rainfall homogeneous regions for estimating quantiles of maximum daily rainfall: a case study of northwestern Mexico. **Atmosphere,** v. 27, n. 1, p. 47-60. 2014.

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 Env. Res. and Risk Asses.,** v. 29, n. 3, p. 875-892. 2015.

ÁVILA, P. L. R.; DE SOUZA, E. B.; PINHEIRO, A. N.; FIGUEIRA, W. S. Analysis of simulated seasonal precipitation using regcm4 over the state of Pará in years of climatic extremes. **Brazilian Journal of Climatology,** n. 14. 2014.

BEZDEK, James C. Pattern recognition with fuzzy objective function algorithms. **Plenum Press,** New York. 1981.

CAI, Y.; JIN, C.; WANG, A.; GUAN, D.; WU, J.; YUAN, F.; XU, L. Spatiotemporal analysis of tropical multisatellite precipitation analysis accuracy 3B42 precipitation data at high mid-latitudes in China. **PloS One,** v. 10, n. 4, e0120026. 2015.

COSTA, J. C.; PEREIRA, G.; SIQUEIRA, M. E.; DA SILVA CARDOZO, F.; DA SILVA, V. V. Validation of rainfall data estimated by CHIRPS for Brazil. **Brazilian Journal of Climatology,** v. 24, p. 228-243. 2019.

COSTA, M. H.; PIRES, G. F. Effects of Amazon and Central Brazil deforestation scenarios on the duration of the dry season in the arc of deforestation. **International Journal of Climatology,** v. 30, n. 13, p. 1970-1979. 2010.

DAVIDSON, E. A.; DE ARAUJO, A. C.; ARTAXO, P.; BATCH, J. K.; BROWN, I. F.; BUSTAMANTE, M. M.; et al. The Amazon basin in transition. **Nature,** v. 481, p. 321-328. 2012.

DEZFULI, A. K. Spatio-temporal variability of seasonal rainfall in western equatorial Africa. **Theoretical and applied climatology,** v. 104, n. 1-2, p. 57-69. 2011.

DINKU, T.; CONNOR, S. J.; CECCATO, P.; ROPELEWSKI, C.F. Comparison of global gridded precipitation products over a mountainous region of Africa. **International Climatology,** v. 11, p. 2960–2979. 2008.

Do AMARAL, M. A. C. M.; JOSÉ, J. V.; FOLEGATTI, M. V.; COELHO, R. D.; BARROS, T. H. S. Spatial distribution of rainfall in relation to the topography in the state of Pará. **Irriga,** v.1, n. 1, p.1-10. 2016.

DUNN, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. **Cybernetics and Systems,** v. 3, p. 32-5. 1973.

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.

FERREIRA FILHO, D. F.; BEZERRA, P. E. S..; SILVA, M. de N. A. da; RODRIGUES, R. S. S. .; de FIGUEIREDO, N. M. Application of interpolation techniques for spatialization of rainfall in the

hydrographic region of Calha Norte, Pará. **Brazilian Journal of Climatology,** v. 24, p. 341-363. 2019.

FERREIRA FILHO, D. F.; LIRA, B. R. P. .; CRISPIM, D. L.; PESSOA, F. C. L. .; FERNANDES, L. L. Rainfall analysis in the state of Pará: comparison between data obtained from rainfall stations and the GPCC satellite. **Brazilian Journal of Climatology,** v. 26. 2020.

FITZJARRALD, D. R; SAKAI, R. K.; MORAES, O. L.; COSME DE OLIVEIRA, R.; ACEVEDO, O. C.; CZIKOWSKY, M. J.; BELDINI, T. Spatial and temporal rainfall variability near the Amazon-Tapajós confluence. **Journal of Geophys. Res.,** v. 113. 2008.

De OLIVEIRA GIL, V.; FERRARI, F.; EMMENDORFER, L. Research on the application of clustering algorithms for the astrophysical problem of classification galaxies. **Brazilian Journal of Applied Computing**, v. 7, n. 2, p. 52-61. 2015.

GONÇALVES, M. F.; BLANCO, C. J. C.; DOS SANTOS, V. C.; DOS SANTOS OLIVEIRA, L. L.; PESSOA, F. C. L. Identification of Rainfall Homogenous Regions taking into account El Niño and La Niña and Rainfall Decrease in the state of Pará, Brazilian Amazon. **Acta Scientiarum. Technology,** v. 38, n. 2, p. 209-216. 2016.

HALKIDI, M.; BATISTAKIS, Y.; VARGIANNIS, M. **Cluster validity methods: Part. I.** ACM SIGMOD Record, v. 31, n. 2. 2002.

IBGE (2010). **Census 2010.** https://cidades.ibge.gov.br. Access on 10 August 2022.

ISHIHARA, J. H.; FERNANDES, L. L.; DUARTE, A. A. A. M.; DUARTE, A. R. C. L. M.; PONTE, M. X.; LOUREIRO, G. E. Quantitative and Spatial Assessment of Precipitation in the Brazilian Amazon (Legal Amazon) - (1978 to 2007). **Brazilian Journal of Water Resources.,** Porto Alegre, v. 19, p. 29-39. 2014.

KÖPPEN, W.; GEIGER, R. Klimate der Erde. Gotha: Verlagcondicionadas. **Justus Perthes,** 1928.

LEMOS, A. L. F.; SILVA, J. de A. Deforestation in the Legal Amazon: evolution, causes, monitoring and mitigation possibilities through the Amazon Fund. **Forest and Environment**, v. 18, n. 1, p. 98-108. 2012.

LIMBERGER, L.; SILVA, M. E. S. Observed precipitation in Brazilian Amazonia: conventional networks and data from Reanalysis I of NCEP/NCAR, CRU and GPCC. **Brazilian Journal of Climatology,** v. 22, ed. Jan/Jun. 2018.

MENEZES, F. P.; FERNANDES, L. L.; DA ROCHA, E. J. P.. The Use of Statistics for Precipitation Regionalization in the State of Pará, Brazil. **Brazilian Journal of Climatology**, v. 16, p. 64-71. 2015.

MENEZES, F.; FERNANDES, L. Analysis of trend and variability of precipitation in the State of Pará. **Biosphere Encyclopedia,** v. 13, no. 24, 2016.

NADOUSHANI, S. S. M.; DEHGHANIAN, N.; SAGHAFIAN, B. A fuzzy hybrid clustering method for identifying hydrologic homogeneous regions. **Journal of Hydroinformatics,** v. 20, n.6. 2018.

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.

NOBRE, C.; YOUNG, A. F.; SALDIVA, P. H. N.; ORSINI, J. A. M.; NOBRE, A. D.; OGURA, A. T.; et al. **Vulnerability of Brazilian Megacities to Climate Change: the São Paulo Metropolitan Region (RMSP).** Climate Change in Brazil: economic, social and regulatory aspects. Brasilia: IPEA., v., p. 197-219. 2011.

PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U. Validity index for crisp and fuzzy clusters, **Pattern Recognition,** n. 37, p.481-501.2004.

PARCHURE, A. S.; GEDAM, S. K. Homogeneous regionalization via L-moments for Mumbai City, India. **Meteorology Hydrology and Water Management,** v. 7, n. 2, p. 73 – 83. 2019.

PEDRYCZ, W.; VUKOVICH, G. Fuzzy clustering with supervision. Pattern Recognition. The **Journal of the Pattern Recognition Society,** v.37, p. 1339-1349. 2004.

PEREIRA, D. D. R.; ULIANA, E. M.; MARTINEZ, M. A; DA SILVA, D. D. Performance of a concentrated hydrologic model and a semidistributed in the prediction of daily flows. **Irriga, Botucatu,** v. 21, n. 2, p.409-424. 2016.

PESSOA, F. C. L.; BLANCO, C. J. C.; MARTINS, J. R. Regionalization of flow permanence curves in the region of Calha Norte in the State of Pará. **Brazilian Journal of Water Resources,** v. 16, p. 65-74. 2011.

RAZIEI, T.; DARYABARI, J.; BORDI, I.; PEREIRA, L. S. Spatial patterns and temporal trends of precipitation in Iran. **Theor Appl Climatol,** v. 115, p. 531–540. 2014.

RENCHER, A. C.; CHRISTENSEN, W. F. **Methods of multivariate analysis.** New Jersey: John Wiley and Sons, 2012. 768 p.

RIBEIRO, A.; VICTORIA, R. L.; PEREIRA, A. R.; VILLA NOVA, N. A.; MARTINELLI, L. A.; MORTATTI, J. Analysis of the Rainfall Regime of the Amazon Region from Data from Eleven Locations. **Brazilian Journal of Meteorology,** v. 11, p. 25 – 35. 1996.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics,** v. 20, p. 53–65. 1987.

SARMADI, F.; SHOKOOHI, A. Regionalizing precipitation in Iran using GPCC gridded data via multivariate analysis and L-moment methods. **Theor Appl Climatol,** p. 122:121–128. 2014.

SCHNEIDER, U.; FUCHS, T.; MEYER-CHRISTOFFER, A.; RUDOLF, B. **Global Precipitation Analysis Products of the GPCC.** Global Precipitation Climatology Centre (GPCC) Deutscher Wetterdienst, Offenbach a. M., 2011, Germany.

SHI, W.; YU, X.; LIAO, W.; WANG, Y.; JIA, B. Spatial and temporal variability of daily precipitation concentration in the Lancang River basin, China. **Journal of Hydrology,** v. 495, p. 197–207. 2013.

SILVA, M. D. N. A. D.; PESSOA, F. C. L.; SILVEIRA, R. N. P. D. O.; ROCHA, G. S.; MESQUITA, D. A. Determination of Homogeneity and Trend of Precipitations in the Tapajós River Basin. **Brazilian Journal of Meteorology,** v. 33, v. 4, p. 665 675. 2018.

SILVEIRA, C. D. S.; SOUZA FILHO, F. D. A. D.; MARTINS, E. S. P. R.; OLIVEIRA, J. L.; COSTA, A. C.; NÓBREGA, M. T.; et al. Climatic changes in the basin of the São Francisco River: An analysis for precipitation and temperature. **Brazilian Journal of Water Resources,** v. 21, 416-428. 2016.

SULOCHANA, Y.; CHANDRIKA, P.; BHASKARA RAO, S. V. Rainrate and rain attenuation statistics for different homogeneous regions of India. **Indian Journal of Radio & Space Physics,** v. 43, p. 301-314. 2014.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining.** Addison Wesley. 2005.

WWO. World Meteorological Organization. **Commission on Instruments and Methods of Observation**. International Organizing Committee for the WMO Solid Precipitation Measurement Intercomparison, final report of the first session, 31 pp., Geneva, 1985.

YAMANA, T. K.; ELTAHIR, E. A. B. On the use of satellite-based estimates of rainfall temporal distribution to simulate the potential for malaria transmission in rural Africa. **Water Resour. Res.,** v. 47. 2011.