

Using clustering algorithms and GPM data to identify spatial precipitation patterns over southeastern Brazil

Bruno GUERREIRO MIRANDA¹, Rogério GALANTE NEGRI^{1,2} and Luana ALBERTANI PAMPUCH^{1,2*}

¹ *São Paulo State University (UNESP), Institute of Science and Technology, São José dos Campos, 12245000, São Paulo, Brazil.*

² *Graduate Program in Natural Disasters, São Paulo State University (UNESP), National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN), São José dos Campos, 12247004, São Paulo, Brazil.*

*Corresponding author: Luana A. Pampuch, luana.pampuch@unesp.br

Received: December 13, 2021; accepted: June 21, 2022

RESUMEN

El Sudeste de Brasil comprende una importante región geoeconómica y poblada de América del Sur. En consecuencia, es fundamental analizar y comprender los perfiles de precipitación en esta región. Entre las diferentes fuentes de datos y técnicas disponibles para realizar estos estudios, el uso de algoritmos de agrupamiento y la información del proyecto Global Precipitation Measurement (GPM) surge como una alternativa conveniente pero poco explotada. Precisamente, este estudio emplea los métodos K-Means, Hierarchical Ward y Self-Organizing Maps para agrupar los datos de precipitación en subregiones homogénea. Fueran utilizados los períodos anual y estacional registrados de 2001 a 2019 del proyecto GPM. Los métodos adoptados fueron comparados con el uso de medidas cuantitativas y el número de conglomerados definidos mediante una regla bien establecida. Los resultados demuestran que los períodos anuales y estacionales están organizados de acuerdo con diferentes números de conglomerados. Además, los resultados permiten: identificar la presencia de una distribución espacialmente heterogénea en el área de estudio; concluir que el algoritmo K-Means es un método de agrupamiento adecuado en el contexto de esta investigación en comparación con los métodos de Hierarchical Ward y Self-Organizing Maps en términos de las medidas Calinski-Harabasz y Davies-Bouldin; y que la precipitación espacial se distribuye sobre el sureste de Brasil está representada por 10 grupos en períodos anuales y de verano, 11 grupos en otoño y primavera y 9 grupos en período de invierno.

ABSTRACT

Southeastern Brazil comprises an important geoeconomic and populous region in South America. Consequently, it is essential to analyze and understand the precipitation profiles in this region. Among different data sources and techniques available to perform such study, the use of clustering algorithms and information from the Global Precipitation Measurement (GPM) project emerges as a convenient, yet less exploited alternative. This study employs the K-Means, the Hierarchical Ward, and the Self-Organizing Maps methods to cluster the annual and seasonal precipitation data from GPM project recorded from 2001 to 2019. The adopted methods are compared in terms of quantitative measures and the number of clusters defined through a well-established rule. The results demonstrate that the annual and seasonal periods are organized according to different number of clusters. Moreover, the results allow: identify the presence of a spatially heterogeneous distribution in the study area; to conclude that the K-Means algorithm is a suitable clustering method in the context of this investigation when compared to Ward's Hierarchical and Self-Organizing Maps methods in terms of the Calinski-Harabasz and Davies-Bouldin measures; and that the spatial precipitation distribution over Southeastern Brazil is represented by 10 clusters in annual and summer periods, 11 clusters in autumn and spring and 9 clusters in winter period.

Keywords: clustering algorithms, precipitation, southeastern Brazil, GPM.

1. Introduction

Each region of the globe has peculiar characteristics, such as latitude, altitude, distance from the oceans, and type of surface, which influence the weather and regional climate. South America presents different topography throughout its extensive territory and is surrounded by the Pacific and Atlantic Oceans. The combination of these two factors and the presence of several atmospheric systems lead to climate heterogeneity in the region, with eight different rainfall regimes (Reboita et al., 2010).

Precipitation southeastern Brazil has marked seasonality, with extensive rain during summer (rainy season) and scarcity over winter (dry season), and is modulated by the South American Monsoon System (SAMS) (Reboita et al., 2010; Marengo et al., 2010). The spatial distribution is also heterogeneous due to its location in a tropical/subtropical region (Nunes and Rampazo, 2017). The proximity to the South Atlantic Ocean favors the transport of moisture to southeastern Brazil (Gimeno et al., 2010; Drumond et al., 2008). In addition, Sea Surface Temperature (SST) anomalies in this Atlantic region may be responsible for precipitation variability and can be associated with extreme events (dry and wet) (Bombardi and Carvalho, 2009; Bombardi et al., 2015; Pampuch et al., 2016).

The South Atlantic Convergence Zone (SACZ) is the main system responsible for high rainfall accumulations during summer in the region (Nogués-Paegle and Mo, 1997; Carvalho et al., 2002; Seluchi and Chou, 2009). Cold fronts may advance over southeastern Brazil throughout the year and also play an essential role in accumulated precipitation (Seluchi and Chou, 2009; Pampuch et al., 2016; Ambrizzi et al., 2015). The position and persistence of the South Atlantic Subtropical High (SASH) may act as a blocking system that does not favor the occurrence of precipitation in the region (Souza and Reboita, 2021).

The topography of southeastern Brazil is also a relevant element in the climate of the region (Nunes and Rampazo, 2017). The presence of elevated surfaces in the central regions constitute dividers for large rivers that drain towards the coast and to the southwest, in addition to the broad coastal plains, which form vast coastal lowlands areas.

Southeastern Brazil is a highly urbanized region that according to the Brazilian Institute of Geography

and Statistics (IBGE) has a population of ~88 million, with businesses and technological development that contribute close to 53.1% to the national gross domestic product (IBGE, 2021). Some of the water basins that serve this population include the São Francisco basin, which covers 7.52% of the country's area; the Southeast Atlantic basin with 2.7%; the South Atlantic basin, with an extension of 2.18%, supplies a small part of the state of São Paulo; and the Paraná basin, with 10.33% of the total area, based on information from the National Water Agency (ANA, 2014).

Meteorological studies to better understand the regional precipitation distribution are very relevant in various environmental contexts, such as for city management and administration. Understanding such patterns may support several applications like preventing landslides and erosion, controlling water levels in reservoirs, and agriculture planning. Among several data of information available to support such studies, the products derived from remote sensing are a viable for extensive spatio-temporal analysis of precipitation. As an example, projects developed by the National Aeronautics and Space Administration (NASA) provide data for different types of studies and research, such the Tropical Rainfall Measuring Mission (TRMM) (Gonçalves et al., 2017) and the Global Precipitation Measurement (GPM) (NASA, 2015).

In particular, the GPM project aims to promote research and applications regarding the physical-meteorological processes of precipitation. Diverse studies based on data from the GPM project have shown its importance. Verma and Ghosh (2018) presented a study about the precipitation at the Gangotri glacier (Himalayas) and noted that, overall, for medium to heavy rainfall, the final run data products are close to the field data. According to Salles et al. (2019), considering a region of the central plateau of Brazil, satellite precipitation products based on GPM can guarantee the monitoring with precision, even better than the previous mission TRMM.

Gadelha et al. (2017) analyze data from 14 rainfall stations installed in the Gramame river basin (state of Paraíba, Brazil) and conclude that estimates from GPM are similar to in-situ measurements. Using clustering techniques, Freitas (2019) presents a comparison between the precipitation data collected for a large portion of the automatic rainfall network

area, maintained by the Brazilian National Center for Monitoring and Alerting of Natural Disasters (CE-MADEN), with the estimated precipitation provided by the Project GPM and concludes that the measurements from such project are reliable. Nonetheless, Freitas (2019) highlights that, for more complex studies, it is necessary to observe the occurrence of under/overestimations and analyze other climatic factors that may influence the final measurements, like the local topography, climate, atmospheric systems, temperatures, and airspeed.

Some studies have already used clustering techniques in Brazilian climate studies. Malfatti et al. (2018), employed clustering techniques to identify homogeneous regions of precipitation along the basin of the Paraná River. Uda et al. (2015) identified distinct precipitation profiles in the Iguaçú River basin. Comunello et al. (2013) outlined homogeneous rainfall environments in the state of Mato Grosso do Sul.

Clustering techniques may be distinguished according to their approaches. Search for hierarchical structures, minimizing a cost function, or modeling neural networks are three typical clustering approaches. Ward's hierarchical method (WH) (Murtagh and Legendre, 2014), K-Means (KM) algorithm (Kodinariya and Makwana, 2013) and the Kohonen's Self-Organizing Map (SOM) (Miljković, 2017) cover some of the approaches. The analysis of remote sensing data usually demands computational and statistical techniques. When the purpose is to identify the spatial precipitation patterns through remote sensing products, the clustering method becomes a convenient alternative. This kind of technique performs a partition (i.e., a subset configuration) on the dataset according to the similarity found among its elements. The spatial representation of these subsets may reveal relevant behaviors to understand and characterize precipitation regimes.

KM, WH, and SOM algorithms appear in different studies involving precipitation analysis in Brazil. Dourado et al. (2013) adopted the KM algorithm to analyze the homogeneous regions in the state of Bahia, Brazil, from 1981 to 2010. Lohmann et al. (2018) used the SOM method as a tool to identify rainfall patterns for the city of Curitiba, state of Paraná, Brazil, to analyze the causes of flooding. Moreover, analyses based on clustering methods and

precipitation data derived from the TRMM project have been carried for different regions of the world, such as Indonesia (Kuswanto et al., 2019) and Brazil (Santos et al., 2019). Pereira et al. (2013) used TRMM data to show consistency when analyzing the spatial distribution of precipitation in Brazil, with seasonal variation very similar to that in meteorological stations. On the other hand, investigations that contemplate clustering methods and GPM data are scarce compared to TRMM.

This study aims to assess the application of clustering methods on data from the GPM project to identify homogeneous regions in southeastern Brazil concerning the annual and seasonal precipitation from 2001 to 2019. The WH, KM, and SOM methods are analyzed, and the quality of results are submitted to quantitative measures for further comparison and discussions. This paper is organized as follows: Section 2 presents relevant details about the study area and the GPM data as well as a formal discussion about clustering methods, accuracy measures for clustering assessment, and the study design; Section 3 presents the results and their respective discussions. Section 4 summarizes the findings of this paper.

2. Materials and Methods

2.1 Study Area and Data

The study focuses on southeastern Brazil, located at latitude (25° 18' 35"S, 14° 13' 58"S) and longitude (53° 05' 15"W, 39° 41' 18"W). The area covers the states of São Paulo (SP), Rio de Janeiro (RJ), Minas Gerais (MG) and Espírito Santo (ES). Figure 1 shows the location and topography of the study area.

Precipitation data for this area were provided by the GPM project from 2001 to 2019. This product estimates daily precipitation at a spatial resolution of 0.1° × 0.1° (Gadelha, 2018). Evaluation of the GPM product over Brazil was done by Rozante et al. (2018) in comparison with rain gauge data. They show that the behavior of maximum precipitation in summer and minimum in winter is well estimated although there is the usual overestimation in summer.

We calculate annual and seasonal accumulated precipitation. The seasons are considered as: summer (December-February; DJF); autumn (March-May; MAM); winter (June-August; JJA); and spring (September-November; SON).

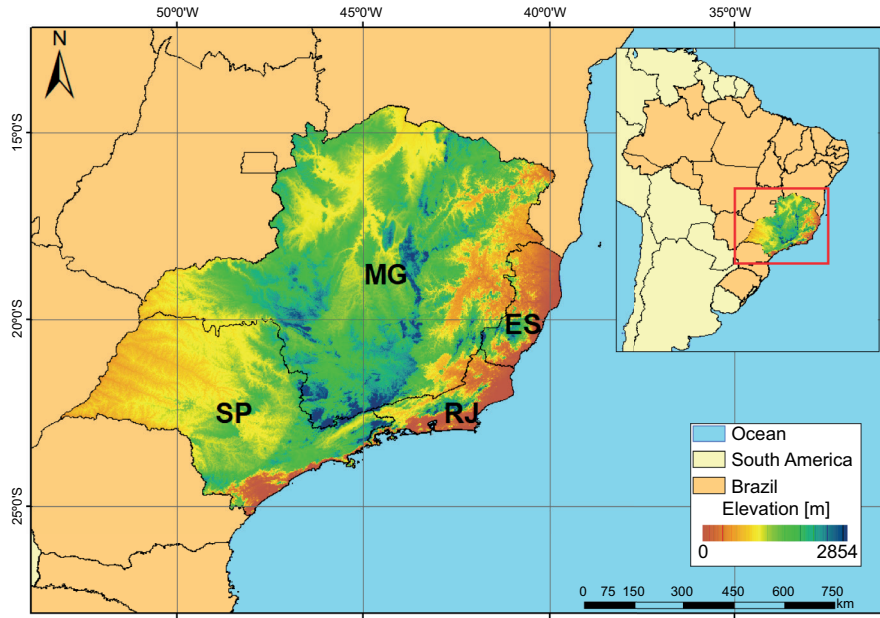


Fig. 1. Study area - southeastern region of Brazil.

2.2 Clustering Methods

Clustering methods play an important role in exploratory data analysis, especially in cases where there is little if any knowledge about the data (Jain and Dubes, 1988). Formally, a clustering comprehends the modelling and application of a function $F: X \rightarrow Y$ on a set of observations $I \subset X$. Through such function each observation $\mathbf{x} \in I$ is assigned to a specific cluster G_j , by means of an indicator $y \in Y = \{1, 2, \dots, k\}$. An observation $\mathbf{x} = [x_1, x_2, \dots, x_n]$, as element of the attribute space X , is usually referred to as attribute vector. The components $x_i, i = 1, \dots, n$, are measures regarding a specific feature. Once $G_j \subseteq I, j = 1, \dots, k$, composes a partition of I , then $\cup_{j=1}^k G_j = I$ and $\cap_{j=1}^h G_j = \emptyset$.

The clustering methods in the literature comprise different approaches to modeling F . Identifying hierarchical relationships, minimizing cluster variability, and performing modeling based on neural networks are some examples of approaches.

Hierarchical methods exploit the relationship structure drawn by the dissimilarity values found between the elements in a dataset. The Ward's Hierarchical (WH) method (Murtagh and Legendre, 2014) is one of the hierarchical methods, that is defined by the following dissimilarity measure between clusters:

$$\begin{aligned}
 D(G_j \cup G_k, G_l) = & \\
 & \frac{\#G_j + \#G_l}{\#G_j + \#G_k + \#G_l} D(G_j, G_l) \\
 & + \frac{\#G_k + \#G_l}{\#G_j + \#G_k + \#G_l} D(G_k, G_l) \\
 & - \frac{\#G_l}{\#G_j + \#G_k + \#G_l} D(G_j, G_k)
 \end{aligned} \tag{1}$$

where $D(\cdot, \cdot)$ represents a recursively computed dissimilarity measure over the clusters. This measure corresponds to the Euclidean distance when applied to compare clusters composed by a single element each.

After computing the dissimilarities between clusters throughout $D(\cdot, \cdot)$, a hierarchical organization naturally arises. In turn, it is possible to establish a threshold τ that determines k clusters such that $D(G_j, G_l) > \tau$ for $j \neq l$ and $j = 1, \dots, k$.

Methods based on minimizing objective cost function define a partition on the dataset such that the variabilities within and between clusters are both minimized and maximized, respectively. Among different methods proposed in the literature following

this approach, the K -Means (KM) (Kodinariya and Makwana, 2013) algorithm is highlighted for its simplicity, effectiveness, and robustness. According to this algorithm, the within/between cluster variabilities minimization/maximization are achieved by solving (Webb and Copsey, 2011):

$$\min_{\mu_j} \frac{1}{m} \sum_{j=1, \dots, k} \sum_{x_i \in G_j} \|\mathbf{x}_i - \mu_j\|^2 \quad (2)$$

where μ_j is the centroid of the cluster G_j , for $j = 1, \dots, k$.

The optimization problem expressed by Equation 2 is solved iteratively through the following steps: (i) each observation $x_i \in I$, for $i = 1, \dots, m$, is assigned to the cluster G_j , with $j = 1, \dots, k$, based on the smallest dissimilarity to the centroid μ_j , determined in terms of the smallest Euclidean distance between \mathbf{x}_i and μ_j ; (ii) after assigning each element to a cluster, the respective centroid is updated according to the mean vector computed concerning the elements allocated to this cluster. This process is repeated until a convergence in the centroids updating is achieved.

Unlike the previous methods, the Self-Organizing Maps (SOM) comprise a data clustering model based on neural network concepts. Formally, a matrix $L_1 \times L_2$ of weight vectors $\mathbf{w}_{ij} = [w_{ij1}, w_{ij2}, \dots, w_{ijn}]$ denotes a map of neurons. Under these conditions, an observation $\mathbf{x} \in I$ is submitted to the neuron map, which promotes an adjustment on the weights according to:

$$\mathbf{w}_{ij} := \mathbf{w}_{ij} + \eta V((i, j), (u, v); \sigma)(\mathbf{x} - \mathbf{w}_{ij}) \quad (3)$$

where $\eta \in \mathbb{R}_+$ represents a learning rate in the neural network. Timely, $V(\mathbf{a}, \mathbf{b}; \sigma) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2 / (2\sigma^2))$ model the neighborhood relations on the map of neurons, where $\mathbf{a}, \mathbf{b} \in \mathbb{N}^2$ represents a spatial coordinate and $\sigma \in \mathbb{R}_+$ is a parameter that controls the neighborhood range. Further details and discussions about the SOM method are found in Haykin (2009).

2.2.1 Clustering Assessment

The use of measures for the assessment of the clustering results is essential to provide a quantitative notion about the quality of the partition obtained and allow comparisons between different results and methods. The Calinski-Harabasz (CH) and Davies-Bouldin

(DB) (Debbarma et al., 2019) indices are examples of measures helpful for this purpose.

The quantification performed by the CH measure is based on the variability values from the clusters that define the dataset partition. Assuming that a dataset I is partitioned according to the clusters G_1, \dots, G_k , the measure CH is expressed by:

$$\text{CH} = \frac{\text{Tr}(\mathbf{V}_E) \#I - k}{\text{Tr}(\mathbf{V}_I) k - 1} \quad (4)$$

given:

$$\mathbf{V}_I = \frac{1}{m} \sum_{j=1}^c \sum_{i=1}^m \delta(\mathbf{x}_i, G_j) (\mathbf{x}_i - \mu_j)^T (\mathbf{x}_i - \mu_j) \quad (5)$$

$$\mathbf{V}_E = \sum_{j=1}^c \frac{\#G_j}{m} (\mu - \mu_j)^T (\mu - \mu_j) \quad (6)$$

where $\mu = \frac{1}{m} \sum_{x_i \in I} \mathbf{x}_i$ is the mean vector observed over all the data in I ; $\mu_j = \frac{1}{\#G_j} \sum_{x_i \in G_j} \mathbf{x}_i$ is the mean vector observed on the cluster G_j ; and $\delta(\mathbf{x}_i, G_j) = \begin{cases} 1; & \text{if } \mathbf{x}_i \in G_j \\ 0; & \text{if } \mathbf{x}_i \notin G_j \end{cases}$ is a membership function.

The values of CH are in $[0, \infty]$, where higher values imply better clustering results. On the other hand, the CH values tend to decrease as the number of clusters increases. Also, it is worth noting that there is no ‘‘acceptable’’ cut-off value for such measure. Nevertheless, the comparisons should involve results with an equivalent number of clusters and prioritize results with higher CH.

In contrast, the DB index provides an assessment that does not depends on the number of clusters. This measure has been useful in studies related to the analysis of meteorological data (Raju and Kumar, 2007; Pansera et al., 2013) and it is given by:

$$\text{DB} = \frac{1}{k} \sum_{i=1}^k \max_{\substack{j=1, \dots, k \\ j \neq i}} \left\{ \frac{S_i + S_j}{\|\mu_i - \mu_j\|} \right\} \quad (7)$$

where $S_i = \frac{1}{\#G_i} \sum_{x_i \in G_i} \|\mathbf{x}_i - \mu_i\|$ expresses the mean distance between each element of the cluster to its centroid. The DB values are in $[0, \infty]$ and smaller values imply better clustering results.

2.2.2 Optimum Number of Clusters

Defining the suitable number of subsets to perform the clustering process may not be a simple task.

Distinct approaches in the literature may support this decision. Some examples include the use of Information Criteria concept (Akogul and Erisoglu, 2016), minimizing the error based on Information - Theoretic Standards (Sugar and James, 2003) or even using a heuristic approach based on the explained variance as a function of the number of clusters, also known as Elbow's Method (Ketchen and Shook, 1996).

Elbow's Method assumes that the increase in the number of clusters implies a reduction of the accumulated variance inside the clusters. On the other hand, the indiscriminate increase in the number of clusters may not imply significant gains in reducing such variance (Han et al., 2012). This approach has been used in studies related to the determination of homogeneous precipitation regions in different regions of the world, such as Malaysia (Ahmad et al., 2013), India (Akhisha et al., 2018) and Ethiopia (Zhang et al., 2016).

Formally, for a given dataset I partitioned in clusters G_j , for $j = 1, \dots, k$, the summation of the deviations from the members of each cluster to its centroid may be expressed by:

$$Q(k) = \sum_{j=1}^k \frac{1}{\#G_j} \sum_{x_i \in G_j} \|x_i - \mu_j\| \quad (8)$$

Consequently, when admitting different values for $k \in \{k_1, k_2, \dots, k_h\}$, with $k_1 < k_2 < k_h$, it is observed that the value of k that simultaneously provides the smaller within-clusters variability and larger separability between clusters is determined by the greatest distance between the point $(k, Q(k))$ and the line-segment with extremes at the points $(k_1, Q(k_1))$ and $(k_h, Q(k_h))$. Such distance is expressed by:

$$L(k; k_1, k_h) = \frac{\left| (Q(k_h) - (Q(k_1))k - (k_h - k_1)Q(k) + k_h Q(k_1) - k_1 Q(k_h)) \right|}{\left\| (k_h, Q(k_h)) - (k_1, Q(k_1)) \right\|} \quad (9)$$

Finally, $k = \arg \max_{k_1, \dots, k_h} L(k; k_1, k_h)$ stands for the optimal number of clusters.

2.3 Experiment Design

Figure 2 depicts the experiment design of this study. Firstly, data from the GPM project, expressed in terms of monthly accumulated precipitation, were obtained and limited to the period of interest (2001 to 2019) and study area (Figure 1). Considering this precipitation time series and admitting a significance of 1%, the Mann-Kendall (MK) (Kendall and Gibbons, 1990) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (Kwiatkowski et al., 1992) tests were applied to reveal the existence of temporal trend in the data. The precipitation values of each instant in the time series stand for the average value observed in the study area in each season.

Subsequently, the precipitation values were organized into five distinct periods of analysis: annual, which comprises the accumulated precipitation in each year; and seasonal, containing the accumulated precipitation in the seasons of each year in the study period. Specifically, the seasons are: summer (December-February; DJF); autumn (March-May; MAM); winter (June-August; JJA); and spring (September-November; SON).

The KM, WH, and SOM methods (Section 2.2) where applied to the annual and seasonal periods

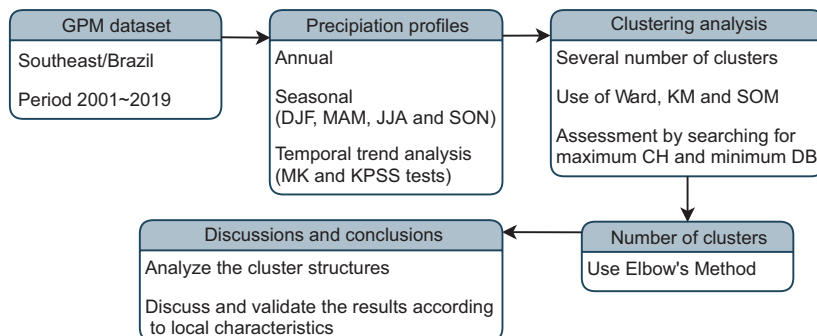


Fig. 2. Experiment design workflow.

considering the number of clusters varying from 2 to 50. The clustering results (a total of $3 \times 49 = 147$) are submitted to the CH and DB measures (Section 2.2.1), to determine the most suitable method to identify the precipitation patterns over the study area. An analysis of the optimal number of clusters is then performed using Elbow's Method (Section 2.2.2). Once the most appropriate method is chosen as well as the optimum number of clusters, the results are discussed considering the topographical and climatic characteristics of the study area.

All the data processing was implemented using the Python 3.8 programming language (van Rossum and Drake, 2011) and the Numpy (Van Der Walt et al., 2011), ScikitLearn (Pedregosa et al., 2011) and GDAL (GDAL/OGR, 2021) libraries.

3. Results and Discussion

Figure 3 shows the time series of the spatial average precipitation in the study area over the analyzed period, where it is possible to observe a stationary profile. Moreover, both MK (p -value = 0.556) and KPSS (p -value = 0.1) confirm that the precipitation profile has not increased or decreased over the time.

The results demonstrate large spatial variability for the region, as shown in Figure 4. Considering the annual period (Fig. 4(a)), the GPM data shows accumulated precipitation from 800 mm (north of MG state) to 2000 mm (coast of SP state). This spatial pattern agrees with results by Neto (2011), Pampuch et al. (2016), Vasconcellos and Reboita (2021) and Silva et al. (2021).

Considering the seasonal periods (Figs. 4(b) to 4(e)), the patterns agree with precipitation patterns already documented in previous studies (Nunes and Rampazo, 2017; Huffman et al., 2007; Vasconcellos and Reboita, 2021); in summary the highest accumulated precipitation is in summer and it is lowest in winter, with autumn and winter being transition seasons.

In the summer, the highest rainfall values occur on the coast of SP and in the central portion of the southeast sector (about 900 mm) due to the SACZ activity and breeze effects (Vasconcellos and Reboita, 2021). The smallest values are recorded in northern MG state (with 250 mm). The wet summer in the region is characteristic of the SAMS, expressed at upper levels by an anticyclonic circulation over Bolivia and a trough near the coast of northeastern Brazil. High pressure systems and an anticyclonic circulations over the sub-tropical Pacific and Atlantic oceans, a low pressure over northern Argentina (Chaco), SACZ, and the South American Low-Level Jet east of the Andes are observed at low levels (Marengo et al., 2010). All these circulation patterns are responsible for more than 60% of the precipitation that occurs in this season (Vasconcellos and Reboita, 2021).

The transition between summer and winter (Fig. 4(c)) has characteristics of both seasons and shows precipitation values from 200 to 600 mm. This season is characterized by a reduction in solar heating and convection, weakening of the trade winds and actuation of the SASH (Vasconcellos and Reboita, 2021). Thus, it is possible to notice a reduction in rainfall compared to the previous season (summer), but not intense as in the winter.

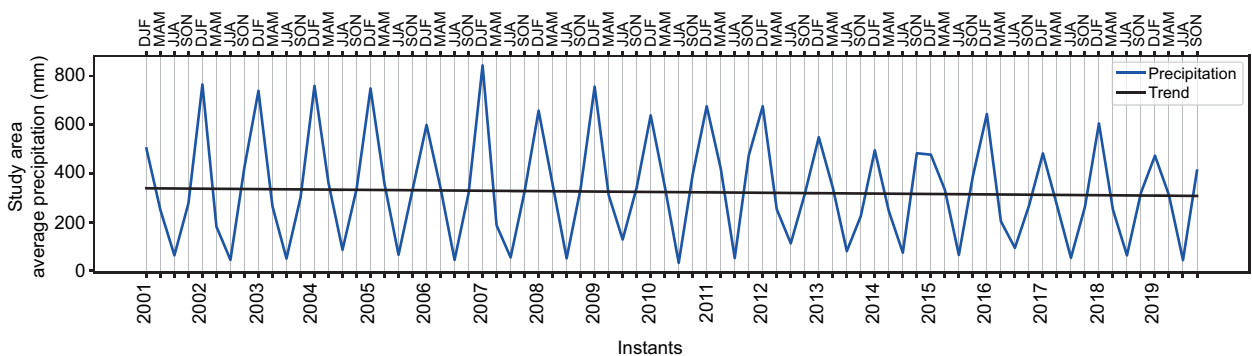


Fig. 3. The precipitation time series observed in the study area over analyzed period.

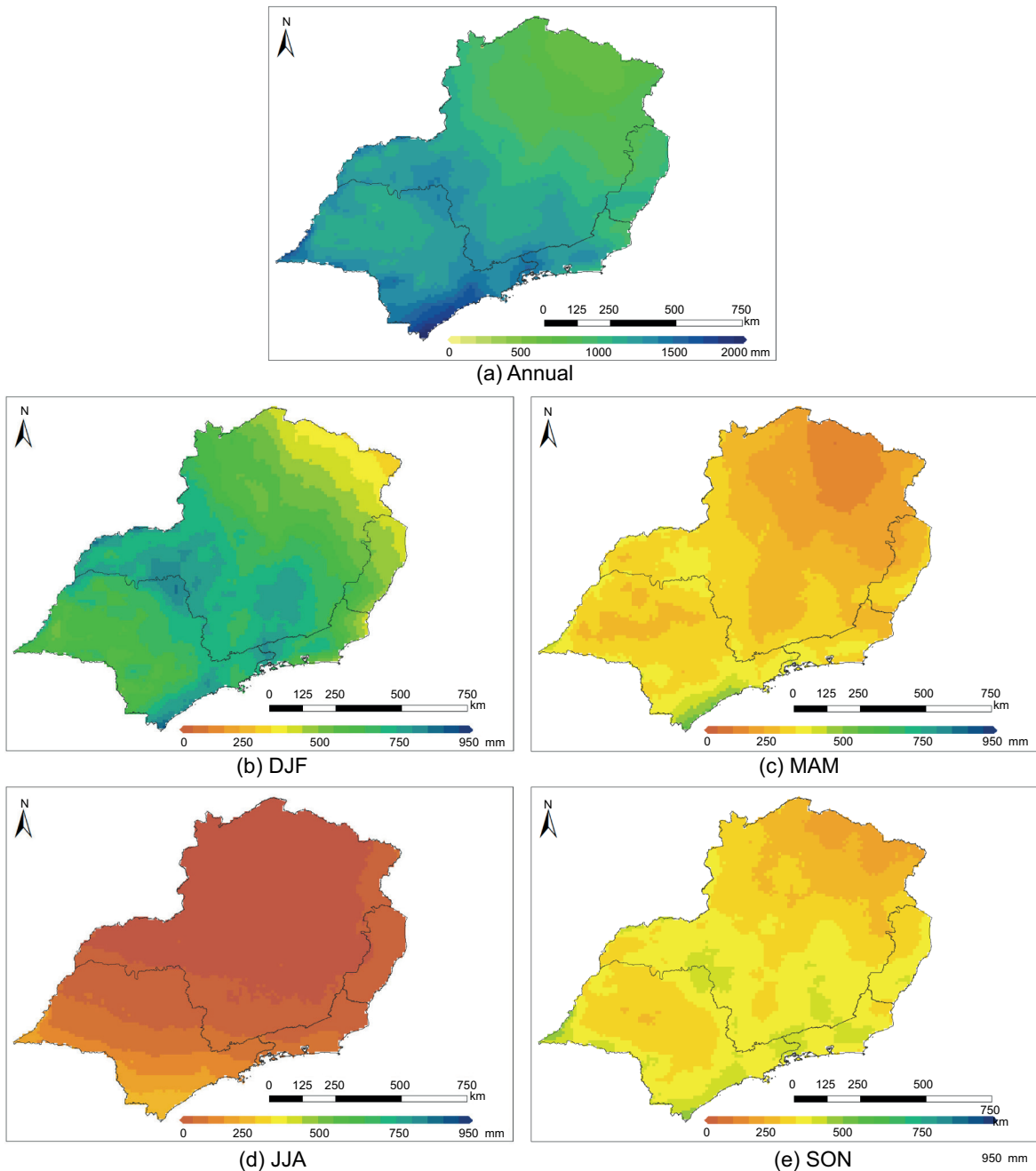


Fig. 4. Temporal average of precipitation values for the (a) Annual (b) DJF (c) MAM (d) JJA (e) SON regarding the period 2001-2019 using the GPM data.

It is evident from the GPM data that the lowest precipitation values (about 0 and 300 mm) are observed in June, July and August. Neto (2011) showed that in the period 1961–2011 less than 5% of the accumulated annual precipitation occurred in winter. In this situation, the SASH is displaced westward,

affecting southeastern Brazil and preventing the equatorward advance of frontal systems (Vasconcellos and Reboita, 2021; Silva et al., 2014).

The spring season shares similarities with autumn and shows a rainfall increase compared to winter. This increase in precipitation occurs due to the

increase in heating and the establishment of the SAMS atmospheric systems (Marengo et al., 2010; Vasconcellos and Reboita, 2021). The average precipitation ranges from 200 to 400 mm.

The performance of the KM, WH, and SOM methods for clustering precipitation data with a

different number of clusters (2 to 50) was then assessed through the CH and DB measures (see experiment design in Fig. 2) and distinct periods of analysis (i.e., annual and seasonal). Figure 5 depicts this assessment, where the KM method shows better results compared to the WH and SOM methods, as it often

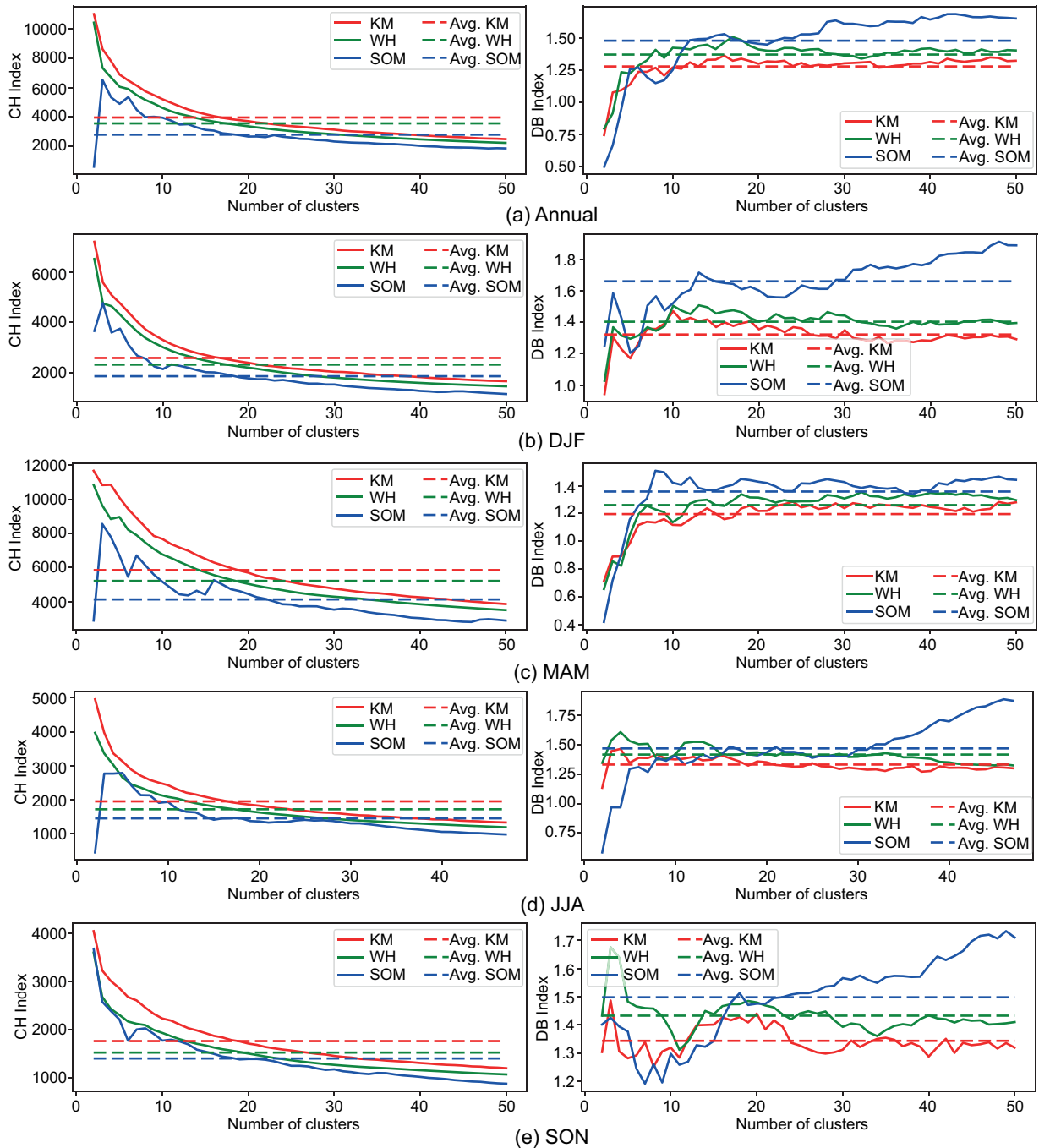


Fig. 5. Clustering assessment based on CH and DB indexes for distinct periods of analysis and number of clusters.

displays high CH and low DB values independently of the number of clusters. The dashed lines indicate the central tendency in which each method, regarding the CH and DB measures, behaves as a function of the number of clusters. Based on this tendency profile, the KM method always provides higher CH values and lower DB values, implying a better performance than the WH and SOM methods regardless of the period considered (annual or seasonal).

Once the KM method is identified as the most suitable to cluster precipitation data, the optimal number of clusters was analyzed using the Elbow's Method. Figure 6 illustrates the relation between $\log Q(k)$ to k from 2 to 50. The values for k that maximize $L(k; 2, 50)$ for each period are highlighted as symbols in Figure 6, indicating 10 clusters for the annual and DJF periods, 11 clusters for the MAM and SON periods and 9 clusters for JJA.

Based on the optimal number of clusters for each period, Figure 7 maps the regions with distinct precipitation patterns over the study area. Figures 8 to 12 shows the annual/seasonal averages extracted from each region, allowing to infer the rainfall regimes of the considered periods and respective partitioning.

The annual period, clustered into 10 regions, reinforces the consistency of the results, which agree with homogeneous precipitation regions compared to climatology (Figure 4). The highest accumulated precipitation (1600 mm) is found on the coast of SP (region 9), decreasing to the north towards region 1, where the smallest precipitation amount is observed (740 mm). The annual variability is similar in all

regions. Note that a decrease in precipitation in all regions occurred in 2014, associated with the well-known drought event observed that year in the study region (Coelho et al., 2016; Otto et al., 2015; Nobre et al., 2016).

Precipitation during summer (DJF) clustered into 10 regions, indicates the highest rainfall, about 750 mm (region 6) in the northern portions of SP and the southwest of MG states. Precipitation decreases in the center of SP and reaches the lowest values in the north of MG, about 400 mm (region 1). Summer corresponds to the largest seasonal precipitation accumulated in all regions, due to the SAMS as mentioned above. Nevertheless, summer precipitation indicates a remarkable interannual variability between clusters.

Precipitation during autumn (MAM), clustered into 11 regions, clearly separate regions with high (450 mm) and low (190 mm) accumulated precipitation, corresponding to the coast of SP (region 10) and north of MG (region 1), respectively. There is a reduction in accumulated precipitation compared to summer. Considering the annual variability in MAM, the pattern is similar for all regions, except in 2011 in region 7, where the largest accumulated precipitation of the entire series was recorded in this season.

Precipitation during winter (JJA), clustered into 9 regions, shows that the north of MG state comprises a vast region, while the SP state is divided into several regions. The largest precipitation is observed in region 9 (255 mm) and the smallest in region 1 (20 mm). The large amounts of precipitation in the south and southeast portions are associated with cold fronts

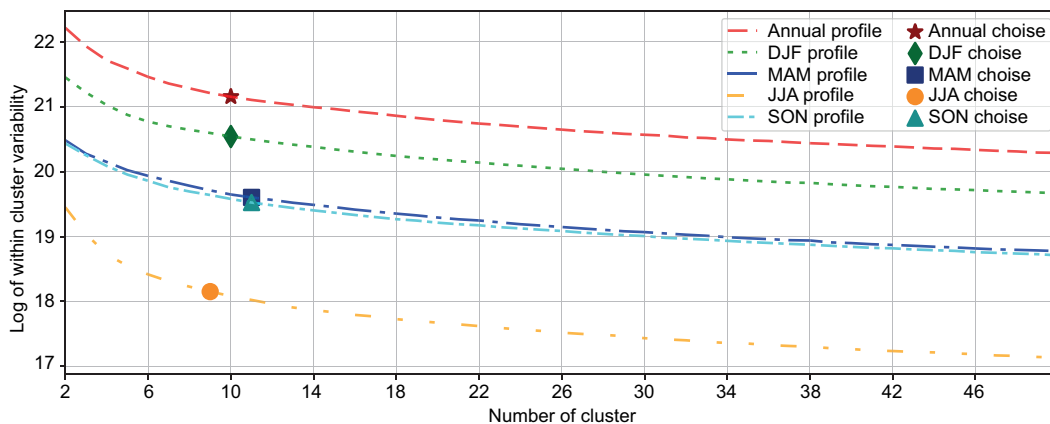


Fig. 6. Number of clusters according to Elbow's Method for KM method.

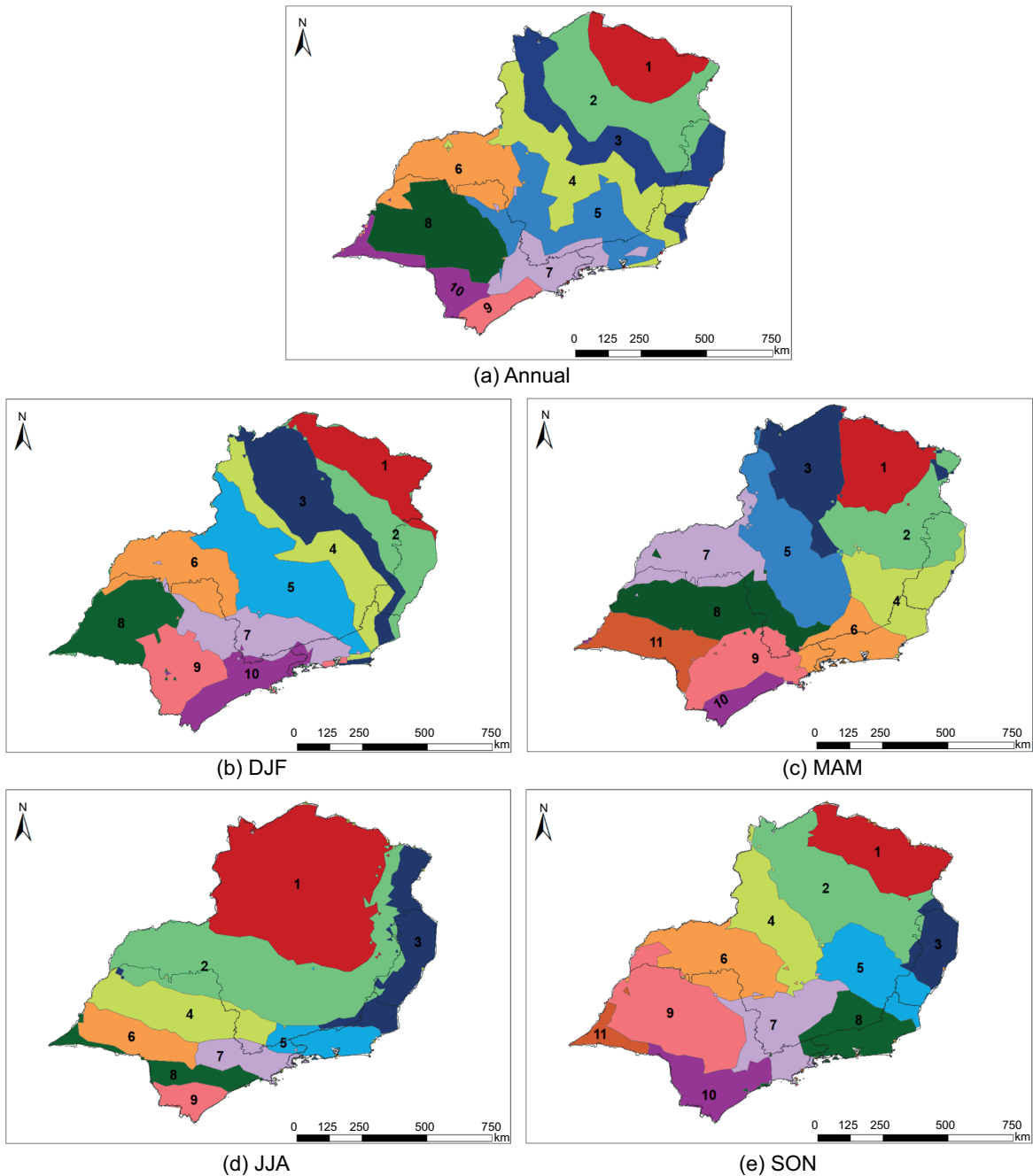


Fig. 7. Clustering results for each analyzed periods using the KM method and optimal number of clusters.

that may reach the region and the SASH area. Winter is the season with both less accumulated precipitation and similar annual variability in all clusters.

Precipitation during the spring (SON) season, clustered into 11 regions, mark separations at the north of MG, the central portion of SP, and the coastal

zone of RJ. The difference in accumulated rainfall between regions is not as evident as in other seasons, and the values range from 280 mm and 400 mm. The accumulated precipitation increases during spring as the atmospheric systems characteristic of the SAMS become established. The annual variability in spring

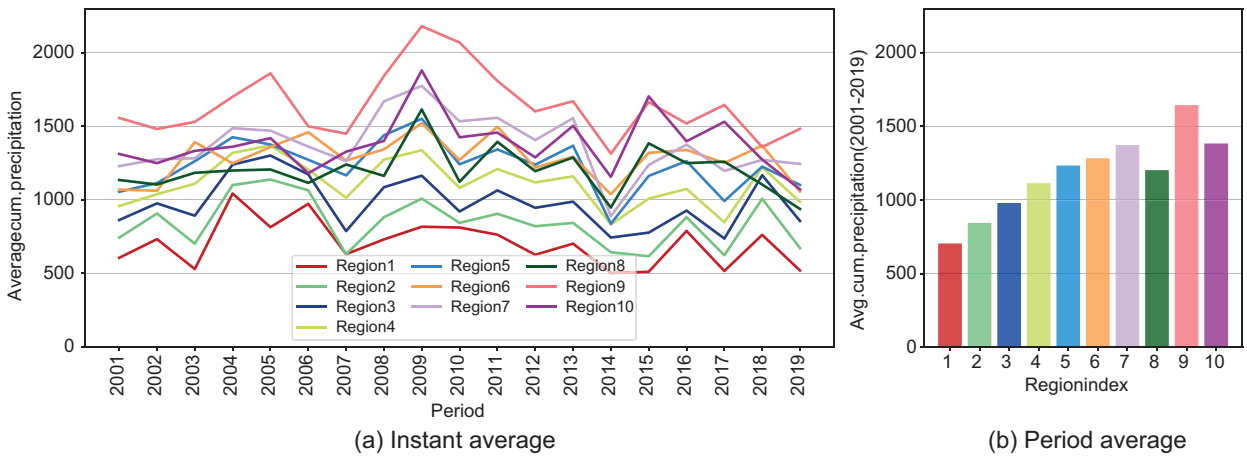


Fig. 8. Precipitation profiles for the identified regions regarding the Annual period

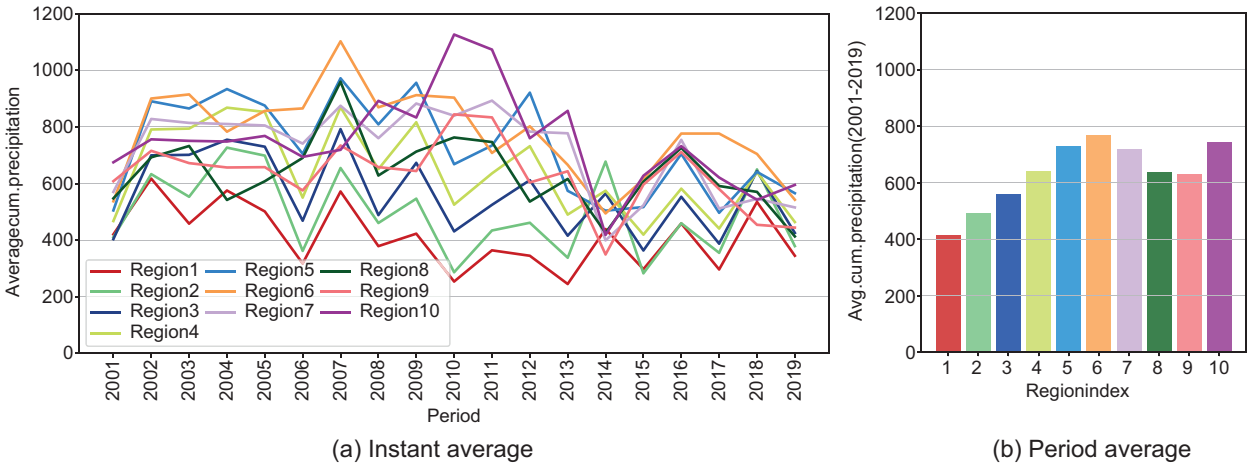


Fig. 9. Precipitation profiles for the identified regions regarding the DJF period.

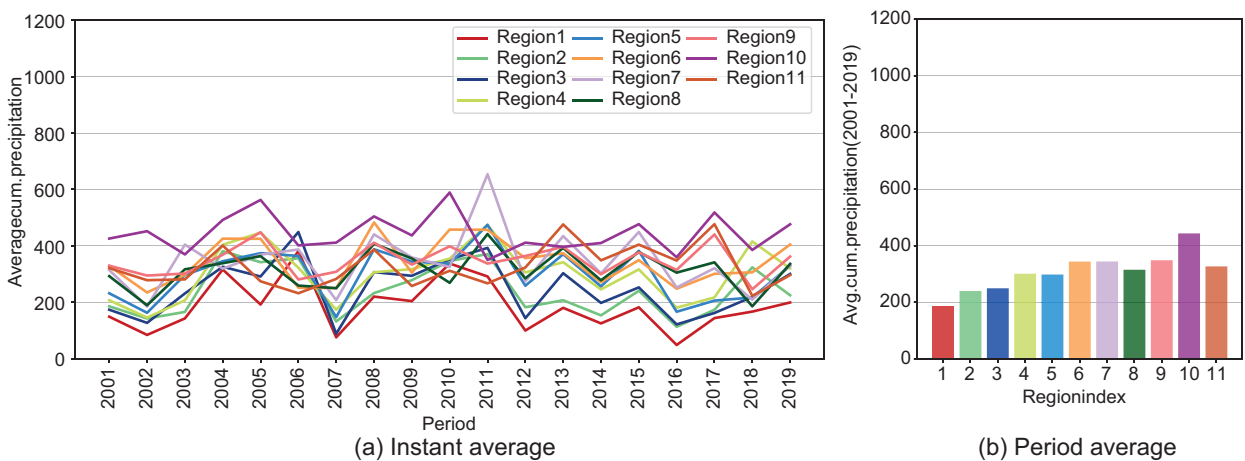


Fig. 10. Precipitation profiles for the identified regions regarding the MAM period

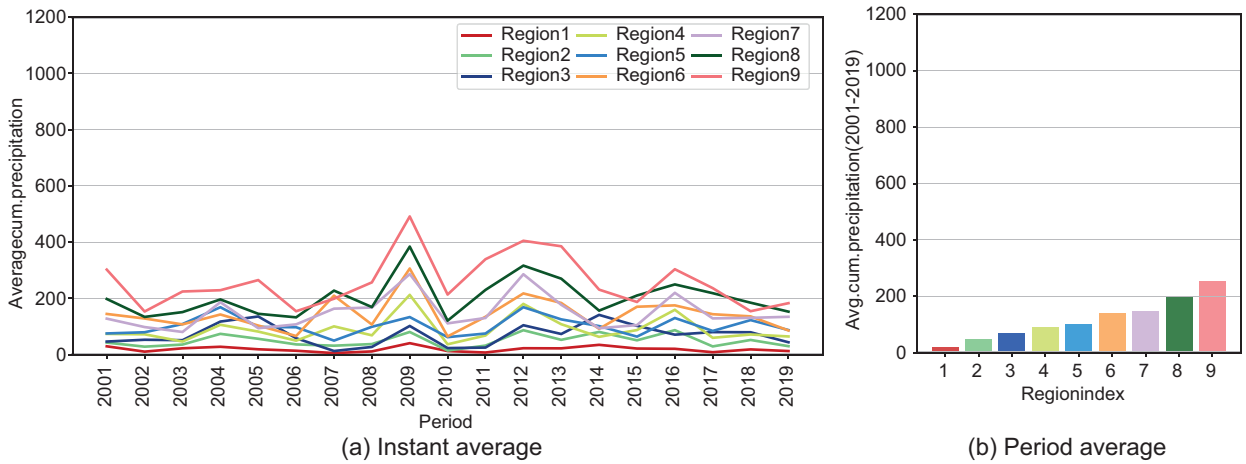


Fig. 11. Precipitation profiles for the identified regions regarding the JJA period.

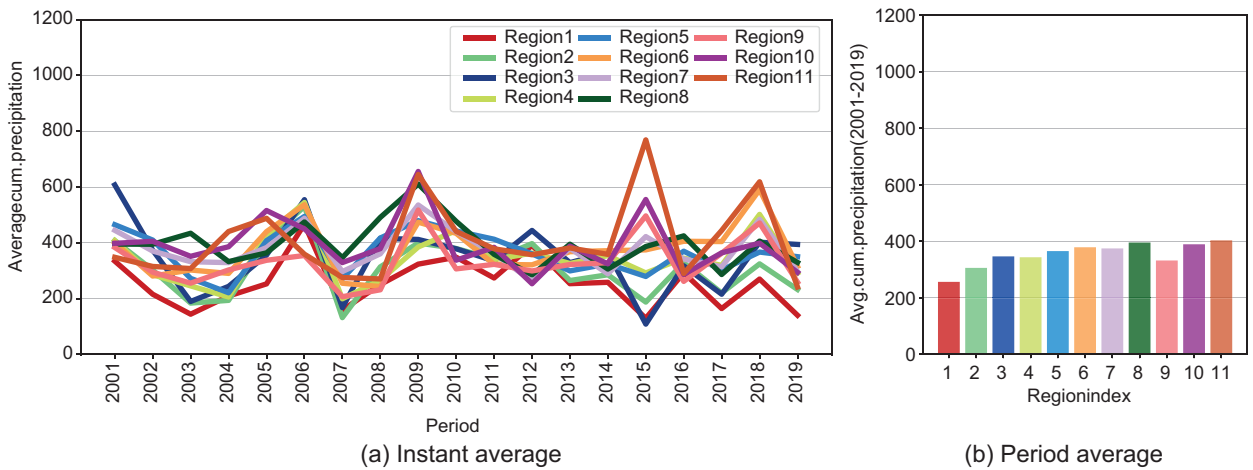


Fig. 12. Precipitation profiles for the identified regions regarding the SON period.

is similar for all clusters except in 2015, where the dry period does not occur with the same intensity in all regions.

4. Conclusions

This study presents an analysis of the distribution of precipitation patterns in southeastern Brazil using different clustering methods applied to data from the GPM project.

The results obtained indicate that the GPM data reveal a precipitation distribution consistent with the seasonal characteristics already discussed by Neto (2011), Pampuch et al. (2016), Vasconcellos and

Reboita (2021) and Silva et al. (2021), as Figure 4 demonstrates: (i) an increase in precipitation in DJF; (ii) a gradual decrease in precipitation in MAM; (iii) the lowest precipitation values occur in JJA; and (iv) in SON a gradual increase in SON.

The annual analysis indicates that the coastal regions of São Paulo and Rio de Janeiro concentrate the highest precipitation in the study area. The precipitation volume decreases significantly towards the north and northeast regions.

According to the CH and DB measures applied to analyze the performance of the different clustering methods, the KM algorithm was the most suitable for identifying the regions of distinct precipitation

patterns. The Elbow's Method was then applied to determine the optimum number of regions that partitions the study area in each period (i.e., annual and seasonals).

Based on these analyses the following can be inferred:

- *Seasonality* – seasonal and annual distributions have different impacts on the analysis of rainfall patterns in the study area, as they can be influenced by various factors such as convergence zones, droughts, longer duration of the rainy season, fires and others. The highest precipitation is observed in summer and spring and the lowest in winter.
- *Clustering method* – the KM algorithm is identified as a suitable clustering method in the domain considered, becoming more refined through the Elbow's Method. A remarkable similarity is found in the precipitation regime, both for the annual and the seasonal periods.

Finally, no prior studies have used this combined methodology for the southeastern region of Brazil, since Pampuch et al. (2016) used Ward's Hierarchical clustering and Singular Value Decomposition, and Machado (2014) used KMeans. Consequently, analyses with similar methodology should be carried out in future studies to identify subgroups within the presented clustering and obtain more information about the precipitation regime in southeastern Brazil.

Acknowledgments

The authors thank FAPESP (grants 2018/010333, 2021/013056) and CNPq (grant 426530/20187) for their financial support of this research.

References

Ahmad NH, Othman IR, and Deni SM. 2013. Hierarchical cluster approach for regionalization of peninsular Malaysia based on the precipitation amount. *Journal of Physics: Conference Series*, 423. <https://doi.org/10.1088/1742-6596/423/1/012018>

Akhisha PA, Rao VS, Umar SKN, Devi KU. 2018. Determination of rainfall regions among the districts of Kerala state. *MAUSAM: Quaterly Journal of Meteorology, Hydrology and Geophysics*, 69(3): 433-436. <https://doi.org/10.54302/mausam.v69i3.334>

Akogul S, Erisoglu M. 2016. A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications*, 21(3). <https://doi.org/10.3390/mca21030034>

Ambrizzi T, Jacobi PR, Dutra LMM. 2015. *Ciência das Mudanças Climáticas e sua Interdisciplinaridade*. Anna Blume Editora. 272 pp.

ANA. 2014. ANA divulga região hidrográfica Atlântico Sul nas redes sociais. Available at: http://www2.ana.gov.br/Paginas/imprensa/noticia.aspx?id_noticia=12422 (accessed 2021 December 15).

Bombardi RJ, Carvalho LM. 2009. The South Atlantic dipole and variations in the characteristics of the South American monsoon in the WCRP-CMIP3 multi-model simulations. *Climate Dynamics*, 36(3): 2091–2102. <http://doi.org/10.1007/s00382-010-0836-9>

Bombardi RJ, Zhu J, Marx L, Huang B, Chen L, Lu J, Krishnamurthy L, Krishnamurthy V, Colfescu I, Kinter III JL, Kumar A, Hu ZZ, Moorthi S, Tripp P, Wu X, Schneider EK. 2015. Evaluation of the CFSv2 CMIP5 decadal predictions. *Climate Dynamics*, 44(1-2): 543–557. <https://doi.org/10.1007/s00382-014-2360-9>

Carvalho LMV, Jones C, Liebmann B. 2002. Extreme precipitation events in South eastern South America and largescale convective patterns in the South Atlantic Con vergence Zone. *Journal of Climate*, 15: 2377–2394. [https://doi.org/10.1175/1520-0442\(2002\)015%3C2377:EPEISS%3E2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015%3C2377:EPEISS%3E2.0.CO;2)

Coelho CA, Oliveira CP, Ambrizzi, T, Reboita MS, Carpenedo CB, Campos JLPS, Tomazzielo ACN, Pampuch LA, Custódio MS, Dutra LMM, Rocha RP, Rehbein A. 2016. The 2014 southeast Brazil austral summer drought: regional scale mechanisms and teleconnections. *Climate Dynamics*, 46: 3737-3752. <https://doi.org/10.1007/s0038201528001>

Comunello E, Araújo LB, Sentelhas PC, Araújo MFC, Dias CTS, Fietz CR. 2013. O uso da análise de cluster no estudo de características pluviométricas. *Sig-mae*, 2(3): 29–37. ISSN 23170840. Available at: <https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/314> (accessed 2021 December 15).

Debbarma N, Choudhury P, Roy P. 2019. Identification of homogeneous rainfall regions using a genetic algorithm involving multi-criteria decision making techniques. *Water Supply*, 19(5): 1491-1499. <https://doi.org/10.1080/14747065.2019.1644444>

- doi.org/10.2166/ws.2019.018
- Dourado CS, Oliveira SRM, Avila AMH. 2013. Analysis of homogeneous zones in time series of precipitation in the state of Bahia. *Brazilian Journal of Meteorology*. <https://doi.org/10.1590/S000687052013000200012>
- Drumond A, Nieto R, Gimeno L, Ambrizzi T. 2008. A Lagrangian identification of major sources of moisture over central Brazil and La Plata Basin. *Journal of Geophysical Research Atmospheres*, 113. <https://doi.org/10.1029/2007JD009547>
- Freitas EdS. 2019. Avaliação do uso do IMERG (Integrated Multisatellite Retrievals for GPM) para determinação de eventos chuvosos e suas propriedades no Brasil: uma análise na escala subdiária. Master's thesis, Universidade Federal da Paraíba. Available at: https://repositorio.ufpb.br/jspui/bitstream/123456789/18947/1/EmersonDaSilvaFreitas_Dissert.pdf (accessed 2021 December 15).
- Gadelha AN. 2018. Análise da missão GPM (Global Precipitation Measurement) na estimativa da precipitação sobre território brasileiro. Master's thesis, Universidade Federal da Paraíba. Available at: <https://repositorio.ufpb.br/jspui/bitstream/123456789/13132/1/Arquivototal.pdf> (accessed 2021 December 15).
- Gadelha AN, Almeida CN, Freitas ES, Coelho VHR, Barbosa LR. 2017. Comparison of the estimated precipitation from GPM with data from rain gauge in the coast of the Paraíba state – Brazil. XX Simpósio Brasileiro de Recursos Hídricos. Available at: <https://files.abrhidro.org.br/Eventos/Trabalhos/60/PAP022159.pdf> (accessed 2021 December 15).
- GDAL/OGR contributors. GDAL/OGR Geospatial Data Abstraction software Library. Open Source Geospatial Foundation, 2021. Available at: <https://gdal.org> (accessed 2021 December 15).
- Gimeno L, Drumond A, Nieto R, Trigo RM, Stohl A. 2010. On the origin of continental precipitation. *Geophysical Research Letter*, 37. <https://doi.org/10.1029/2010GL043712>
- Gonçalves S, Brasil Neto R, Santos C, Silva R. 2017. Análise da variabilidade espaçotemporal da precipitação no Cariri Paraibano utilizando dados do satélite TRMM. In *Simpósio Brasileiro de Recursos Hídricos*. Available at: <https://files.abrhidro.org.br/Eventos/Trabalhos/60/PAP023279.pdf> (accessed 2021 December 15).
- Han J, Kamber M, Pei J. 2012. *Data Mining: Concepts and Techniques*. Elsevier Inc., 3 edition. Available at: <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf> (accessed 2021 December 15).
- Haykin S. 2009. *Neural Networks and Learning Machines*. Number v. 10 in *Neural networks and learning machines*. Prentice Hall. ISBN 9780131471399. Available at: <https://lps.ufrj.br/caloba/Livros/Haykin2009.pdf> (accessed 2021 December 15).
- Huffman GJ, Bolvin DT, Nelkin EJ, Wolff DB, Adler RF, Gu G, Hong Y, Bowman KP, Stocker EF. 2007. The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, 8(1): 38-55. <https://doi.org/10.1175/JHM560.1>
- IBGE. Instituto Brasileiro de Geografia e Estatística produto interno bruto - pib, 2021. Available at: <https://www.ibge.gov.br/explica/pib.php> (accessed 2021 December 15).
- Jain AK, Dubes RC. 1988. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. ISBN 0-13-022278-X.
- Kendall M, Gibbons JD. 1990. *Rank Correlation Methods*. A Charles Griffin Title, 5th edition.
- Ketchen DJ, Shook CL. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17(6):441-458. Available at: <http://www.jstor.org/stable/2486927> (accessed 2021 December 15).
- Kodinariya T, Makwana P. 2013. Review on determining of cluster in Kmeans clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1: 90-95. Available at: <http://www.ijarcsms.com/docs/paper/volume1/issue6/V1I60015.pdf> (accessed 2021 December 15).
- Kuswanto H, Setiawan D, Sopaheluwakan A. 2019. Clustering of precipitation pattern in Indonesia using TRMM satellite data. *Engineering, Technology and Applied Science Research*. <https://doi.org/10.48084/etasr.2950>
- Kwiatkowski D, Phillips PC, Schmidt P, Shin Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1): 159-178. [https://doi.org/10.1016/0304-4076\(92\)90104Y](https://doi.org/10.1016/0304-4076(92)90104Y)

- Lohmann M, Cunico C, Maganhotto RF. 2018. Neural network of the type SOM (Self Organizing Map) as a tool for identifying rain patterns. In I National Symposium on Geography and Territorial Management and XXXIV Geography Week at the State University of Londrina. Available at: <http://anais.uel.br/portal/index.php/sinagget/article/view/369/323> (accessed 2021 December 15).
- Machado LA. 2014. Classificação climática para Minas Gerais por meio do método de agrupamento não hierárquico de KMeans. *Cadernos do Leste - Artigos Científicos*, 14(14). <https://doi.org/10.29327/249218.14.143>
- Malfatti MGL, Cardoso AO, Hamburger DS. 2018. Identificação de regiões pluviométricas homogêneas na bacia hidrográfica do Rio Paraná. *Geociências UNESP*, 37(2):409 – 421. <https://doi.org/10.5016/geociencias.v37i2.11564>
- Marengo JA, Liebmann B, Grimm AM, Misra V, Dias PLS, Cavalcanti IFA, Carvalho LMV, Berbery EH, Ambrizzi T, Vera CS, Saulo AC, Nogués-Paegle J, Zipser E, Seth A, Alves LM. 2010. Recent developments on the South American monsoon system. *International Journal of Climatology*, 32(1):1–21. <https://doi.org/10.1002/joc.2254>
- Miljković D. 2017. Brief review of selforganizing maps. In 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1061-1066. <https://doi.org/10.23919/MIPRO.2017.7973581>
- Murtagh F, Legendre P. 2014. Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. *Journal of Classification*, 31: 274–295. <https://doi.org/10.1007/s00357-014-9161-z>
- NASA. 2015. Integrated multi-satellite retrievals for GPM (IMERG) version 4.4. NASA's Precipitation Processing Center. https://arthurhou.pps.eosdis.nasa.gov/Documents/Master_List_of_PPS_Data_Products.html (accessed 2021 December 15).
- Neto SAJ. 2011. Decálogo da climatologia do sudeste brasileiro. *Revista Brasileira de Climatologia*, 1:43–60. <https://doi.org/10.5380/abclima.v1i1.25232>
- Nobre CA, Marengo JA, Seluchi ME, Cuartas LA, Alves LM. 2016. Some characteristics and impacts of the drought and water crisis in southeastern Brazil during 2014 and 2015. *Journal of Water Resource and Protection*, 8: 252–262. <https://doi.org/10.4236/jwarp.2016.82022>
- NoguésPaegle J, Mo KC. 1997. Alternating wet and dry conditions over South America during summer. *Monthly Weather Review*, 125: 279–291. [https://doi.org/10.1175/1520-0493\(1997\)125<0279:AWADCO>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<0279:AWADCO>2.0.CO;2)
- Nunes LH, Rampazo NAM. 2017. Tendências da precipitação diária no estado de São Paulo a partir do Índice de Concentração (IC). <https://doi.org/10.20396/sbgfa.v1i2017.2312>
- Otto FEL, King CAS, Perez EC, Wada Y, van Oldenborgh GJ, Haarsma R, Haustein K, Uhe P, van Aalst M, Aravequia JA, Almeida W, Cullen H. 2015. Factors other than climate change, main drivers of 2014/15 water shortage in Southeast Brazil. *Bulletin of the American Meteorological Society*, 96(12): 51-56. <https://doi.org/10.1175/BAMS-D-15-00120.1>
- Pampuch LA, Drumond A, Gimeno L, Ambrizzi T. 2016. Anomalous patterns of SST and moisture sources in the South Atlantic Ocean associated with dry events in southeastern Brazil. *International Journal of Climatology*, 36: 4913–4928. <https://doi.org/10.1002/joc.4679>
- Pansera WA, Gomes BM, Vilas Boas MA, Mello EL. 2013. Clustering rainfall stations aiming regional frequency analysis. *Journal of Food, Agriculture and Environment*, 11 (2):877-885. <https://doi.org/10.1002/joc.4679>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blon del M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830. <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>. (accessed 2021 December 15).
- Pereira G, Silva MES, Moraes EC, Cardozo FS. 2013. Avaliação dos dados de precipitação estimados pelo satélite TRMM para o Brasil. *Revista Brasileira de Recursos Hídricos*, 18(3): 139-148. <https://doi.org/10.21168/rbrh.v18n3.p139-148>
- Raju KS, Kumar DN. 2007. Classification of Indian meteorological stations using cluster and fuzzy cluster analysis, and Kohonen artificial neural networks. *Hidrology Research*, 38 (3): 303-314. <https://doi.org/10.2166/nh.2007.013>
- Reboita MS, Gan MA, Rocha RP, Ambrizzi T. 2010. Regimes de precipitação na América do Sul: uma revisão bibliográfica. *Revista Brasileira de Meteorologia*, 25(2): 185-204. <https://doi.org/10.1590/S0102-77862010000200004>
- Rozante JR, Vila DA, Barboza Chiquetto J, Fernandes ADA, Souza Alvim D. 2018. Evaluation of TRMM/

- GPM blended daily products over Brazil. *Remote Sensing*, 10(6): 882. <https://doi.org/10.3390/rs10060882>
- Salles L, Satgé F, Roig H, Almeida T, Olivetti D, Ferreira W. 2019. Seasonal effect on spatial and temporal consistency of the new GPM-based IMERG-v5 and GSMaP-v7 satellite precipitation estimates in Brazil's central plateau region. *Water*. <https://doi.org/10.3390/w11040668>
- Santos CAG, Moura Brasil Neto R, Silva RM, Costa SGF. 2019. Cluster analysis applied to spatiotemporal variability of monthly precipitation over Paraíba state using Tropical Rainfall Measuring Mission (TRMM) data. *Remote Sensing*, 11(6): 637. <https://doi.org/10.3390/rs11060637>
- Seluchi ME, Chou SC. 2009. Synoptic patterns associated with landslide events in the Serra do Mar, Brazil. *Theoretical and Applied Climatology*, 98: 66–67. <https://doi.org/10.1007/s00704-008-0101-x>
- Silva LJ, Reboita MS, daRocha RP. 2014. Relação da passagem de frentes frias na região sul de Minas Gerais (RSMG) com a precipitação e eventos de geada. *Revista Brasileira de Climatologia*, 14. <https://doi.org/10.5380/abclima.v14i1.36314>
- Silva WL, Oscar Júnior AC, Cavalcanti IFA, Treistman F. 2021. An overview of precipitation climatology in Brazil: Spacetime variability of frequency and intensity as associated with atmospheric systems. *Hydrological Sciences Journal*, 66(2): 289–308. <https://doi.org/10.1080/02626667.2020.1863969>
- Souza CA, Reboita MS. 2021. Ferramenta para o monitoramento dos padrões de teleconexão na América do Sul. *TerraE Didática*, 17: e02109. <https://doi.org/10.20396/td.v17i00.8663474> (accessed 2021 December 15).
- Sugar CA, James GM. 2003. Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, 98(463): 750–763. <http://www.jstor.org/stable/30045303> (accessed 2021 December 15).
- Uda PK, Franco ACL, Queen G, Bonumá NB, Kobiyama M. 2015. Análise de cluster da precipitação na bacia do rio Iguaçu, região sul do Brasil. XXI Simpósio Brasileiro de Recursos Hídricos. ISSN 2318-0358. <https://www.ufrgs.br/gpden/wordpress/wp-content/uploads/2016/11/UDA.CLUSTER.pdf> (accessed 2021 December 15).
- Van Der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2): 22–30. <https://doi.org/10.1109/MCSE.2011.37>
- van Rossum G, Drake FL. 2011. *The Python Language Reference Manual*. Network Theory Ltd. ISBN 1906966141, 9781906966140. <https://docs.python.org/3/reference/> (accessed 2021 December 15).
- Vasconcellos FC, Reboita MS. 2021. Clima das regiões brasileiras e variabilidade climática. In: Iracema F. A. Cavalcanti, Nelson J. Ferreira. (Eds.), volume 1. *Oficina de Textos*, São Paulo. ISBN: 978-65-86235-24-1.
- Verma P, Ghosh SK. 2018. Study of GPM/IMERG rainfall data product for Gangtouri glacier. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-5. <https://doi.org/10.5194/isprs-archives-XLII-5-383-2018>
- Webb AR, Copsey KD. 2011. *Statistical Pattern Recognition*. John Wiley & Sons, Ltd, 3rd edition. ISBN 9780470682289. <https://doi.org/10.1002/9781119952954>
- Zhang Y, Moges S, Block P. 2016. Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatiotemporal variability: Application to western Ethiopia. *Journal of Climate*, 29(10): 3697–3717. <https://doi.org/10.1175/JCLI-D-15-0582.1>