Importance of Regional Rainfall Data in Homogeneous Clustering of Data

Sparse Areas: A Study in the Upper Brahmaputra Valley Region

Jayshree Hazarika¹,², Arup Kumar Sarma¹

¹Department of Civil Engineering, Indian Institute of Technology Guwahati, Guwahati – 781039, Assam, India

²Department of Civil Engineering, Assam Engineering College, Guwahati – 781013, Assam, India

Corresponding Author: Dr. Jayshree Hazarika

Email ID: jayshree@iitg.ac.in; ORCID ID: 0000-0002-7155-8752

Abstract

Delineation of homogeneous regions has found its way into many hydrological applications as it helps in addressing the challenges in understanding the behavior of rainfall distribution and its variability at a local scale. In the present study, rainfall data recorded by 83 tea gardens in the upper Brahmaputra valley region of Assam have been used to identify homogeneous rainfall regions by using fuzzy clustering analysis. Further, seven different cluster validity indices (CVs) were utilized to find out the optimum clustering in the fuzzy c-means (FCM) algorithm. The clusters thus formed were assessed for statistical homogeneity by performing homogeneity tests based on L-moment. Three different combinations of feature vectors were employed in FCM algorithm and the outputs were compared for attaining best solutions to regionalization. The results were further compared with previous regionalization studies. The analysis and comparison conclude that if regionalization needs to be done at a local scale, further sub-clustering of a larger clustered region to smaller regions may be required. Local rainfall data can be used for the purpose provided a good dataset with large number of station points are available within the region. Along with rainfall data, geographical location
parameters (latitude, longitude and elevation) need to be taken into account for getting a

definite conclusion.



Keywords:

Regionalization, L-moments, Fuzzy clustering, Homogeneity test
1. Introduction

Rainfall is one of the most important hydrological parameters that requires to be studied scrupulously, both in spatial and temporal scale. However, challenge comes in understanding the behavior of rainfall distribution pattern when applied on a regional scale. Pattern of rainfall, its frequency and magnitude may drastically vary depending upon the orography (Venkatesh and Jose 2007), large-scale synoptic and convective precipitation types (Karaca et al. 2000, Unal et al. 2012; Baltaci et al. 2015; Efe et al. 2019) of the region. Scarcity of sufficient data at many sites of interest, may further complicate the investigation. To address this issue, the region may be classified into few homogeneous rainfall regions of similar rainfall distribution, also termed as regionalization. The Brahmaputra valley region is situated in the northeastern part of India, that has its own peculiar topography. The variations in the orographic arrangements and altitude differences in the region give rise to irregular and complex rainfall patterns at a local scale, which eventually amplifies the need of regionalization.

Regionalization has found its way into many applications in water resources planning, agricultural planning, drainage design, and estimating magnitude and frequency of extreme events like flood and drought. Literatures indicate that, in the past, political and geographical boundaries are used as a basis of forming homogeneous regions (Thomas and Benson 1970; NERC 1975; Beable and McKercher 1982; Chew et al. 1987). However, use of political and geographical boundaries is found to be not very convincing while forming hydrologically homogeneous regions (Bonell and Sumner 1992; Burn 1997; Rao and Srinivas 2006b; Satyanarayana and Srinivas 2011; Dikbas et al. 2012). A considerable amount of researches have been carried out in the recent years to identify homogeneous rainfall regions with various methods other than geographical divisions (Bedi and Bindra 1980; Bärring 1987; Sumner and Bonell 1988; Kulkarni et al. 1992; Gadgil et al. 1993; Burn 1997; Adelekan 1998). Regionalization with the use of principal component analysis (PCA) was found to be of use.
(Singh and Singh 1996; Wotling et al. 2000). When subjectivity involved with PCA came into
notice, the concept of cluster analysis started getting attention (Bonell and Sumner 1992;
Guttman 1993; Venkatesh and Jose 2007; Machiwal et al. 2019). Cluster analysis refers to a
varied group of statistical procedures used to classify a multivariate dataset into some clusters
or groups (Rao and Srinivas 2006a, b; Srinivas et al. 2008; Dikbas et al. 2012). Studies have
also been done where PCA was further associated with various cluster analysis techniques for
homogeneous clustering (Dinpashoh et al. 2004; Satyanarayana and Srinivas 2011; Darand and
Daneshvar 2014; Machiwal et al. 2019). Ward’s hierarchical cluster analysis is one of the
widely used methods, that is found to be suitable for homogeneous regionalization (Unal et al.
2003; Baltacı et al. 2017). Another vastly applied method is the k-means clustering (Rao and
Srinivas 2006a; Pelczer et al. 2007; Agarwal et al. 2016). K-means clustering splits a region
into hard clusters, i.e. with a degree of membership 1 or 0. This means that a site can at most
belong to only one cluster. However, this may not be valid in real world cases. To address this
matter, the concept of fuzzy clustering was introduced to regionalization, which allows a site
to fit into several clusters concurrently with a certain membership value. The fuzzy
membership value of a site signifies the extent to which it fits into a particular group of sites
(Rao and Srinivas 2006b). A lot of studies have successfully applied fuzzy clustering technique
for clustering of hydrologically homogeneous regions in the recent years (Hall and Minns 1999;
Owen et al. 2006; Plain et al. 2007; Sadri and Burn 2011; Satyanarayana and Srinivas 2011;
Chen et al. 2011; Dikbas et al. 2012; Mok et al. 2012; Chavoshi et al. 2013; Asong et al. 2015;
Bharath and Srinivas 2015; Goyal and Sharma 2016; Irwin et al. 2017; Wang et al. 2017).
On the basis of critical reviews of earlier studies on regionalization, the aim of the current study
is defined as to identify homogeneous rainfall regions in upper Brahmaputra Valley region of
northeast India by using fuzzy clustering analysis. To achieve the best possible partition from
the fuzzy c-means (FCM) algorithm, seven different cluster validity indices (CVs) were used.
Three different combinations of feature vectors were employed in FCM algorithm and the outputs were compared for attaining best solutions to regionalization. The homogeneous rainfall regions (fuzzy clusters) thus formed by the use of FCM algorithm and validated with CVs were then assessed for statistical homogeneity by performing homogeneity tests using L-moment approach (Hosking and Wallis 1997). The results were further compared with some previous regionalization studies and finally concluding remarks were presented.

2. Methodology

In the subsections below, the fuzzy c-means (FCM) algorithm for delineation of homogeneous clusters is described at first, which is followed by a brief description of the CVs, homogeneity test methods and process of adjustment of heterogeneous clusters. The methodology used is shown in Fig. 1, by means of a flow chart.

2.1 Fuzzy C-Means clustering (FCM)

The FCM approach is basically optimization of fuzzy c-means objective function. It was initially developed by Dunn (1973) and afterwards modified by Bezdek et al. (1984). The fuzzy c-means function to be minimized, is expressed as:

\[
J(Z; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m ||z_k - v_i||^2 A
\]  

\[
U = [\mu_{ik}] \in M_{fc} \text{ is the fuzzy partition matrix of } Z; \ V = [v_1, v_2, ..., v_c], \ v_i \in R^n \text{ is vector of cluster centers to be determined}; \ d^2(z_k, v_i)A = D_{ikA}^2 = ||z_k - v_i||^2 A = (z_k - v_i)^T A (z_k - v_i)
\]

is a squared distance norm; and \( m \) is the fuzziness parameter or fuzzifier, where \( m = [1, \infty) \).

Usually, its value falls between 1 and 2.5 (Pal and Bezdek 1995).

2.1.1 FCM algorithm for delineation of homogeneous rainfall regions

(i) The initial fuzzy partition matrix \( U \) is set.
(ii) Then, initial membership values $\mu_{ki}^{init}$ of $x_i$ that belongs to cluster $k$, is adjusted by using equation:

$$
\mu_{ki} = \frac{\mu_{ki}^{init}}{\sum_{j=1}^{c} \mu_{ji}^{init}}, \text{for } 1 \leq k \leq c, 1 \leq i \leq N \tag{2}
$$

(iii) Fuzzy cluster centroid $v_k$ is then calculated as

$$
v_k = \frac{\sum_{i=1}^{N} (\mu_{ki})^m x_i}{\sum_{i=1}^{N} (\mu_{ki})^m}, \text{for } 1 \leq k \leq c \tag{3}
$$

(iv) Fuzzy membership value $\mu_{ki}$ is updated as

$$
\mu_{ki} = \left( \frac{\frac{1}{d^2(x_i \cdot v_k)}}{\sum_{k=1}^{c} (\frac{1}{d^2(x_i \cdot v_k)})^{1/m}} \right)^{1-1/m}, \text{for } 1 \leq k \leq c, 1 \leq i \leq N \tag{4}
$$

(v) The objective function is then calculated as

$$
J(X; U, V) = \sum_{k=1}^{c} \sum_{i=1}^{N} (\mu_{ki})^m d^2(x_i, v_k) \tag{5}
$$

The above steps from (iii) to (v) are repeated till the difference in the objective function for two consecutive iterations becomes adequately small.

### 2.2 Cluster validity indices (CVs)

FCM algorithm divides the data into well separated and compact clusters, provided the optimal values of $c$ and $m$. Hence deciding the optimal values of these parameters is very crucial. Bezdek (1981) addressed this matter by stating the concept of validity indices. These indices
essentially measure the goodness of the partitioned clusters. In hydrological studies several indices are used (Hall and Minns 1999; Srinivas et al. 2008). In case of FCM algorithm, the following CVs are found to perform well:

1. Fuzzy partition coefficient ($V_{PC}$)

$$V_{PC}(U) = \frac{1}{N} \sum_{k=1}^{c} \sum_{i=1}^{N} (\mu_{ki})^2$$  \hspace{1cm} (6)

2. Fuzzy partition entropy ($V_{PE}$)

$$V_{PE}(U) = -\frac{1}{N} \left[ \sum_{k=1}^{c} \sum_{i=1}^{N} \mu_{ki} \log(\mu_{ki}) \right]$$  \hspace{1cm} (7)

3. Fuzziness performance index ($V_{FPI}$)

$$V_{FPI}(U) = 1 - \frac{cXV_{PC}(U) - 1}{c - 1}$$  \hspace{1cm} (8)

4. Normalized classification entropy ($V_{NCE}$)

$$V_{NCE}(U) = \frac{V_{PE}(U)}{\log(c)}$$  \hspace{1cm} (9)

Bezdek (1974a, b) formulated $V_{PC}$ and $V_{PE}$, whereas $V_{FPI}$ and $V_{NCE}$ were proposed by Roubens (1982). The range for $V_{PC}$ is $[1/c, 1]$; $V_{PC} = 1/c$ indicates equal sharing of clusters i.e. equal membership values of a data in all clusters (i.e. $\mu_{ki} = 1/c \ \forall i, k$) and $V_{PC} = 1$ indicates no sharing of membership among the clusters. Similarly, the range of $V_{PE}$ is $[0, \log (c)]$. $V_{PE} = 0$ implies no sharing of membership among clusters and $V_{PE} = \log(c)$ implies equal sharing of clusters (i.e. $\mu_{ki} = 1/c \ \forall i, k$). On the contrary, this range is $[0, 1]$ for $V_{FPI}$ and $V_{NCE}$; 0 implies no membership sharing between clusters and 1 implies equal sharing of clusters (i.e. $\mu_{ki} = 1/c \ \forall i, k$). As such, a maximum value for $V_{PC}$ indicates optimum partition (i.e. minimum value for $V_{PE}$, $V_{FPI}$ and $V_{NCE}$), that means least overlap among clusters.

Earlier studies have stated that these four CVs tend to display monotonous increasing or decreasing trend (Rao and Srinivas 2006b, Srinivas et al., 2008). Hence, they are not very effective in obtaining optimum partition to delineate rainfall regions. Xie and Beni (1991) found no direct correlation of $V_{PC}$ and $V_{PE}$ with any property of the data. Furthermore, they are
found to be very sensitive to the fuzzifier value, \( m \) (Halkidi et al. 2001). Here, these indices are used mainly to validate their performances in detecting optimum cluster number. To eliminate this drawback the other validity indices are introduced, as explained below:

5. Extended Xie and beni index (\( V_{XB} \))

\[
V_{XB,m}(U,V:X) = \frac{\sum_{k=1}^{c} \sum_{i=1}^{N} (\mu_{ki})^m ||x_i-v_k||^2}{N \min_{l \neq k} ||v_l-v_k||^2} \tag{10}
\]

Xie and Beni (1991) proposed the cluster validity index \( V_{XB,m} \). It quantifies the ratio of compactness within a fuzzy cluster to separation of clusters. Optimal partition of clusters should exhibit minimum value of \( V_{XB,m} \).

6. Fukuyama and Sugeno index (\( V_{FS} \))

\[
V_{FS}(U,V:X) = \sum_{i=1}^{N} \sum_{k=1}^{c} (\mu_{ki})^m ||x_i-v_k||^2 A - \sum_{i=1}^{N} \sum_{k=1}^{c} (\mu_{ki})^m ||v_i-v_k||^2 A \tag{11}
\]

Proposed by Fukuyama and Sugeno (1989), optimum partition is indicated by a minimum value of \( V_{FS} \).

7. Kwon index (\( V_{K} \))

\[
V_{K}(U,V:X) = \frac{\sum_{k=1}^{c} \sum_{i=1}^{N} (\mu_{ki})^m ||x_i-v_k||^2 + \frac{1}{2} \sum_{k=1}^{c} ||v_k-v_l||^2}{\min_{i \neq k} ||x_i-v_k||^2} \tag{12}
\]

The Extended Xie and beni index (\( V_{XB} \)) exhibits monotonous decreasing tendency as \( c \to N \). To address this problem Kwon (1998) proposed another index \( V_{K} \) that has an ad-hoc punishing function in numerator.

To determine the optimal values of \( c \) and \( m \), a range of values for the two parameters are selected and subsequent partitioning results show different sets of clusters, along with their validity indices. To decide the optimal set of values for \( c \) and \( m \) among those sets, first the optimum selection criteria of each of the validity indices are examined. Then, the sites having greater membership value in the clusters are identified, based on a threshold value of fuzzy membership (\( T_i \)). Thus, a fuzzy cluster is made by allocating those sites to the cluster, whose membership values are found to be higher than or equal to the threshold fuzzy membership.
value ($T_i$). The selection of this threshold value is subjective (Satyanarayana and Srinivas 2011). The most reasonable explanation would be to allocate the site to that group where its membership value is the highest. Yet, uncertainty arises when a site has low membership value in all the clusters or has equal memberships. To address this issue homogeneity test is done which is followed by adjustment of heterogeneous clusters. The methodologies for both of these are explained in the following subsection.

2.3 Homogeneity test and adjustment of heterogeneous clusters

The fuzzy clusters thus formed by using FCM algorithm and validated with CVs are then required to be assessed for statistical homogeneity by performing homogeneity tests. Heterogeneity measure (L-moment based) proposed by Hosking and Wallis performs better when skewness is low (average L-skew $< 0.23$) for a sample set of data, whilst for higher skewness bootstrap Anderson-Darling test is recommended (Viglione et al. 2007). Previous studies have shown that Hosking and Wallis’s homogeneity test is appropriate for delineation of homogeneous rainfall regions (Satyanarayana and Srinivas 2011), hence is considered in this study.

2.3.1 L-moment of data samples

L-moment is a method of explaining the probability distribution shape and evaluating the distribution parameters, especially for small sample sizes of environmental data, since it is unbiased and has a nearly normal distribution (Hosking 1990). Like usual moments, L-moments too determines the location, dispersion, peakedness, skewness and any other feature of shape of probability distribution. However, L-moments are derived from linear combination of data (Hosking 1990). These statistics are established by modifying “probability weighted moments” (Greenwood et al. 1979), which explains L-moments by means of linear combinations. Sample probability weighted moments as explained by Greenwood et al. (1979) is given below:
\[ b_0 = n^{-1} \sum_{j=1}^{n} x_j \]  
\[ b_r = n^{-1} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} x_j \]

The first few L-moments and L-moment ratios are defined as:

Location, mean \( (l_1) \): \( l_1 = b_0 \)

Scale, L-Cv \( (t_2) \): \( t_2 = \frac{t_2}{l_1} \) where \( l_2 = 2b_1 - b_0 \)

L-Skewness \( (t_3) \): \( t_3 = \frac{t_3}{l_2} \) where \( l_3 = 6b_2 - 6b_1 + b_0 \)

L-Kurtosis \( (t_4) \): \( t_4 = \frac{t_4}{l_2} \) where \( l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \)

2.3.2 Discordancy measures, heterogeneity measures and adjustment of heterogeneous clusters

1. Discordancy measures

Discordancy measure \( (D_i) \) detects those sites which are unacceptably discordant with the designated cluster (Hosking and Wallis 1993). This discordancy value for \( i^{th} \) site (Hosking and Wallis 1995) is given as,

\[ D_i = \frac{N}{3(N-1)} (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) \]

Here, \( S \) is the sample covariance matrix expressed as:

\[ S = (N - 1)^{-1} \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T \]

where \( u_i = [t^{(i)} t_3^{(i)} t_4^{(i)}]^T \) means a vector comprising of the values of \( t, t_3 \) and \( t_4 \) for \( i^{th} \) site.

Hence,

\[ \bar{u} = N^{-1} \sum_{i=1}^{N} u_i \]

Large values of \( D_i \) indicate probable errors in the site data. Hosking and Wallis (1993) explained that a particular site is not considered to be homogeneous with the region if \( D_i \) is more than a certain critical value, than that. \( D_i \geq 3 \) is suggested as the criterion for affirming a site to be discordant, for regions with 15 or more sites. However, Hosking and Wallis (1993,
1995) have advised to scrutinize the dataset for the largest $D_i$ values, irrespective of their magnitude.

2. Heterogeneity measures

Heterogeneity measures give the degree of heterogeneity existing within the region. It is estimated based on the extent of actual variability in L-moment ratios in relation to the expected variability in a homogeneous region. The heterogeneity measures to be estimated are $H_1$, $H_2$ and $H_3$. These measures are defined based on L-Cv, L-Skewness and L-Kurtosis. These three heterogeneity measures are given below.

i. Heterogeneity measures based on L-Cv

\[ H_1 = \frac{V - \mu_V}{\sigma_V} \]  

ii. Heterogeneity measures based on L-Cv and L-Skewness

\[ H_2 = \frac{V_2 - \mu_{V_2}}{\sigma_{V_2}} \]  

iii. Heterogeneity measures based on L-Cv and L-Kurtosis

\[ H_3 = \frac{V_3 - \mu_{V_3}}{\sigma_{V_3}} \]  

Here $V$ is the weighted standard deviation or dispersion of the sample coefficients of L-variation (L-Cv); $V_2$ and $V_3$ denote weighted average distance from the site to group weighted mean in a 2-D space of L-Cv/L-Skewness and L-Skewness/L-Kurtosis, respectively; $\mu_V$, $\mu_{V_2}$ and $\mu_{V_3}$ are the mean of $V$, $V_2$ and $V_3$ values, respectively, calculated from a large number of simulations ($N_{sim}$); $\sigma_V$, $\sigma_{V_2}$ and $\sigma_{V_3}$ are standard deviations of $V$, $V_2$ and $V_3$, respectively, calculated from a number of simulations. In this study $N_{sim}$ is taken as 1000, simulated from kappa distribution and fitted using regional average L-moment ratios.

\[ V = \left( \frac{\sum_{i=1}^{N} n_i (\ell_i - \ell(R))^2}{\sum_{i=1}^{N} n_i} \right)^{1/2} \]  

\[ V_2 = \frac{\sum_{i=1}^{N} n_i (\ell_i - \ell(R))^2 + \ell_3^2 - \ell_2^2}{\sum_{i=1}^{N} n_i} \]  

\[ V_3 = \frac{\sum_{i=1}^{N} n_i (\ell_i - \ell(R))^2 + \ell_3^2 - \ell_2^2}{\sum_{i=1}^{N} n_i} \]
\[ V_3 = \frac{\sum_{i=1}^{N} n_i \left( (t_3^{(i)}-t_3^{(R)})^2 + (t_4^{(i)}-t_4^{(R)})^2 \right) \right)^{1/2}}{\sum_{i=1}^{N} n_i} \]  

(27)

Here \( N \) indicates number of sites/stations in the clustered region; \( n_i \) indicates record length of site \( i \); \( t_3^{(i)} \) and \( t_4^{(i)} \) indicate sample L-moment ratios of site \( i \); \( t_3^{(R)} \) and \( t_4^{(R)} \) indicate regional average L-moment ratios (L-Cv, L-Skewness and L-Kurtosis respectively).

If \( H < 1 \) for a region, it is described as ‘acceptably homogeneous’, \( 1 \leq H < 2 \) implies ‘possibly heterogeneous’, and \( H \geq 2 \) implies ‘definitely heterogeneous’. For further details on heterogeneity measures, Hosking and Wallis (1997) can be referred.

3. Adjustments of heterogeneous clusters

Clusters formed by the use of clustering algorithms does not always exhibit statistical homogeneity. Even after performing homogeneity test, there is a need to adjust the possibly or definitely heterogeneous clusters to come to a definite conclusion. Furthermore, it is assumed that there is not any cross-correlation among the data. Nevertheless, in actual scenarios there is gradual variation of rainfall across space, which implies that there exists cross-correlation among geographically contiguous sites (Satyanarayana and Srinivas 2011). So further adjustments are essential to form homogeneous regions. Studies suggest that, to decrease heterogeneity measures values, the discordant sites of one region can be either removed or shifted to some other region, after confirming that the site has not exhibited high fuzzy membership value in that cluster and discordancy of that site is not because of sampling variability (Rao and Srinivas 2006b; Satyanarayana and Srinivas 2011). The heterogeneous regions can also be broken into two or more regions, and if required two or more small regions can be merged together.

3. Results and discussion

3.1 Geographic and climate properties of the study area

Fig. 2 shows the study area, that is situated in the upstream part of Brahmaputra River in the state of Assam, covering the central Brahmaputra valley and the eastern Brahmaputra valley.
region. It covers most of the upper and middle Assam districts. This region of the valley is surrounded by Eastern Himalayas towards the north, the Patkai Bum in the east, the Naga Hills in the southern side and Meghalaya plateau in the far south. The Brahmaputra valley has a great geographical as well as political significance. It is surrounded by Bhutan in the north, Arunachal Pradesh in the north and east, Nagaland and Karbi-Anglong hills in the south. The study area lies between 25.921° N and 27.619° N latitudes and 91.896° E and 95.768° E longitudes. Mean annual precipitation varies from 859 mm to 3412 mm and the elevation varies from 67 m to 427 m. Seasonal precipitation is found to be the highest in June-July-August (JJA) (406 mm – 1880 mm) and the lowest in December-January-February (DJF) (9 mm – 222 mm). The Brahmaputra valley has a subtropical climate which is influenced by northeast and southwest monsoon. The Meghalaya plateau, the Himalayas and the surrounding hills of Arunachal Pradesh, Manipur, Nagaland and Mizoram influences the climate. The monsoon winds coming from Bay of Bengal move towards the northeast and hits these mountains causing heavy precipitation on the valley.

3.2 Data used

A total of 83 raingauge stations with observed periods of varying number of years (from 5 years daily data to as long as 84 years and up to year 2016) were selected from various tea gardens of Assam. The location (longitude and latitude), seasonal precipitation (in mm), mean annual precipitation (in mm) and elevation (in m) of the raingauge stations are shown in the Table 1. Elevation of each station was determined from Shuttle Radar Topography Mission’s digital elevation model data (DEM) version 2.1 (SRTM-1) of 30m resolution. The DEM files come in tiles, that come as zipped SRTMHGT files at 1-arcsecond resolution (3601x3601 pixels) in a latitude/longitude projection (EPSG:4326).

3.3 Cluster analysis with Fuzzy C-Means (FCM) algorithm
Previous studies suggest that, the attributes to be considered in clustering of homogeneous rainfall regions may include large scale atmospheric variables (LSAVs) of Global Climate Models (GCMs) or else principal components (PCs) of GCMs, location parameters (latitude, longitude, elevation etc.), and seasonality measures (maximum, minimum, standard deviation of rainfall etc.). In the present study, total annual rainfall, total monthly rainfall, standard deviation of total annual rainfall and total monthly rainfall and all three location parameters i.e. latitude, longitude and elevation of all the 83 raingauge stations were included as attributes.

Here three different combinations of feature vectors were used to form, consequently, three different sets of input data matrices (for FCM algorithm). This was done in order to observe the effect of a particular attribute in the cluster formation. The three combinations were as follows:

Case 1: Input data matrix with total monthly rainfall as attributes,
Case 2: Input data matrix with standard deviation of total monthly rainfall as attributes,
Case 3: Input data matrix with latitude, longitude, elevation, total annual rainfall and standard deviation of total annual rainfall as attributes.

To acquire reliable results from clustering, most crucial point is to assume the cluster number \((c)\), since the number of regions is unknown beforehand. The best way to do that is to choose a range of values for \(c\), and then to find out the most appropriate one. To achieve that, the cluster number \(c\) is changed from 2 to \(k\), \(k\) being a quantity lesser than total number of sites, \(N\).

The lower bound of \(c\) is taken as 2, because the dataset is apparently clustered into more than one group. The interesting point here is to define the upper bound of \(c\), i.e. the parameter \(k\). Varying the \(k\) value impacts the reliability of cluster number. From research, it was noticed that increasing the \(k\) value generates more consistent cluster number division (Mok et al 2012).

Many works can be found in literature on identification of optimal cluster numbers. In most of the cases, \(k \leq N^{1/2}\) is suggested as the upper bound (Xie and Beni 1991; Pal and Bezdek 1995;
Mok et al. (2012). In a similar way, optimum value of fuzzifier \((m)\) also needs to be found out. Pal and Bezdek (1995) presented that FCM algorithm works well when \(m\) varies from 1–2.5. It is hence suggested (Satyanarayana and Srinivas 2011) to attain a number of sets of clustered regions by selecting a range of values for \(c\) and \(m\), and then identify the final clustered regions based on the optimal values of \(c\) and \(m\) and by means of CVs.

In this study, FCM algorithm was executed for each case with cluster number varying from \(c_{\text{min}} = 2\) to \(c_{\text{max}} \approx 10\), with increment 1. The fuzzifier value \((m)\) is increased from 1.1 to 3.0, with increment 0.1. Change in the values of objective functions with respect to that of fuzzifier \(m\) were plotted for cluster number varying from 2 to 10, for all the three cases and are shown in Fig. 3a, 3b and 3c. It is observed that, as the fuzzifier value increases the optimum value of objective function declines, for a given cluster number. In a similar way, the optimum value decreases with rise in cluster number, for a given value of fuzzifier. The FCM algorithm was found to perform better for \(m\) in the range of 1.5–2.5.

Further, the CVs described in section 2.2, were calculated to achieve the optimal value of \(c\) and \(m\), for each case. The values of \(V_{PC}, V_{PE}, V_{FPI}, V_{NCE}, V_{XB}, V_{FS}\), and \(V_{K}\) are tabulated in supplementary material. Since the FCM algorithm performed better for \(m\) in the range of 1.5–2.5, the CVs for higher value of \(m\) are not considered and hence are not shown in the tables. In all the three cases, it is observed that, \(V_{PC}\) decreases monotonously whereas \(V_{PE}, V_{FPI}\) and \(V_{NCE}\) show increase monotonously with increase in the fuzzifier value, \(m\), for a given value of cluster number \(c\). Moreover, they show overall monotonous decreasing \((V_{PC})\) and increasing \((V_{PE}, V_{FPI}\) and \(V_{NCE}\) tendency with increase in the cluster number \(c\), hence giving always low value of \(c\) and \(m\) as the optimum set of values for homogeneous clustering. This indicates their ineffectiveness in determining optimum number of homogeneous rainfall regions. Extended Xie and beni index \((V_{XB})\) as well as the Kwon index \((V_{K})\) were found to be relatively effective in these cases. They didn’t show any monotonous increasing or decreasing tendency.
with change in the values of $c$ and $m$, and were in line with each other. In case of Fukuyama and Sugeno index ($V_{FS}$), although it didn’t have any monotonic tendency, the values were found to be in a conflict to the values of $V_{XB}$ and $V_{K}$ in some cases. Taking into account of the similar studies on homogeneous rainfall region identification, found in literature (Rao and Srinivas 2006b; Satyanarayana and Srinivas 2011; Farsadnia et al. 2014; etc.), $V_{XB}$ and $V_{K}$ were opted for clustering of rainfall regions. Henceforth, the clusters found from these two indices were considered for further analysis. The optimal value of clusters $c=3$ was identified for all the three cases, using these CVs with the corresponding value of $m=1.5$ for all three cases.

3.4 Homogeneity test and adjustment of heterogeneous clusters

The $H$ values of homogeneity test for all the three cases are given in Table 2. Initially, out of the nine clusters (three clusters for each case), four clusters were acceptably homogeneous, four clusters were possibly heterogeneous and one cluster was definitely heterogeneous. Hence, adjustments were done to make the possibly and definitely heterogeneous clusters as homogeneous, by either shifting discordant sites from one cluster to another or by removing those sites if necessary. For case 1, stations Chubwa and Dilli were found to discordant with all the clusters; hence they are removed during adjustment which lead to three acceptably homogeneous clusters with the rest 81 stations. Similarly, station Chubwa was found to be discordant for case 2; hence removed from the clusters and adjustments were done for 82 stations. However, in case 3, no stations were removed. The reason that the two stations did not show any discordancy in case 3 is because the geographic locations of the stations were considered during cluster analysis for case 3. The $H$ values of homogeneity test after adjustment and number of sites in each cluster are shown in Table 2. Final clusters formed after adjustment are shown in Fig. 4a, 4b and 4c.

3.5 Comparison with similar studies done previously
Although this is the first and foremost study to divide the Brahmaputra valley region of India into hydrologically homogeneous regions with the use of rainfall data record by the tea gardens by applying fuzzy clustering technique and seven cluster validity indices (CVs), few previous studies can be found on use of various clustering techniques for Indian subcontinent. Hence, comparisons have been made with those closely related studies to validate the performance of the present study. The homogeneous regions developed by Indian Meteorological Department (IMD) displays five large provinces, which are although delineated based on rainfall characteristics but are influenced by contiguity of area and administrative state boundaries. Iyengar and Basak (1994) used principal component analysis (PCA) for regionalization of Indian monsoon rainfall and recommended the PCA approach for further subdivision of the region. Ten homogeneous sequential regions were formed in India from their analysis, in which the stations of upper Brahmaputra valley regions were seemed to form similar kind of clusters as found in the present study, although few stations remained un-clustered. Singh and Singh (1996) have done regionalization of monthly as well as seasonal rainfall for sub-Himalayan areas and Gangetic plains, by using principal component analysis (PCA). They used rainfall data for a period of 114 years (1871-1984) from 90 well distributed stations which resulted into four distinct homogeneous rainfall areas for both monthly and seasonal scales. Srinivasa and Kumar (2007) utilized fuzzy cluster analysis (FCA) to classify 159 meteorological stations in India and concluded that FCA method performs well than the Kohonen Artificial Neural Networks (KANN) method in finding meteorologically homogeneous groups. They utilized location parameters (latitude, longitude and elevation) along with other meteorological parameters for clustering and the results exhibited 14 clusters over Indian region, the North-eastern region being in one cluster. Satyanarayana and Srinivas (2008) have done regional frequency analysis using LSAVs that affects the precipitation in a region instead of observed precipitation data and have used K-means clustering with adjustments and L-statistics for
regionalization. 17 homogeneous regions were formed after the analysis, two regions covering
the northeastern states. The upper Brahmaputra valley region came under the same
homogeneous cluster, hence producing similar results to those of the present study.
Satyanarayana and Srinivas (2011) have done regionalization of rainfall data, based on fuzzy
clustering method by utilizing GCM data, location parameters and seasonal precipitation data.
The stations of upper Brahmaputra valley regions were seemed to form similar kind of clusters
as found in the present study. Stations in middle Assam were found to form one cluster while
other stations on the upper Assam formed a different cluster. Saikranthi et al. (2013) used
correlation analysis for regionalization based on seasonal and annual rainfall data. They used
51 years (1951–2001) daily rainfall data collected for more than 1000 rain gauges across India
for the analysis, which produced 26 homogeneous rainfall zones. However, because of data
scarcity northeastern states were not included in the analysis. Bharath and Srinivas (2015) used
wavelet-based global FCM analysis, instead of PCA for determining homogeneous
hydrometeorological regions in India. The new approach proposed by them clustered the Indian
territory into 29 regions, northeastern region having 7 clusters. The clusters formed in the upper
Brahmaputra valley region were similar to the present study. Kulkarni (2017) has used
probability density function to divide the Indian subcontinent into homogeneous clusters, using
daily summer monsoon rainfall at 357 square grids of size 10000 sqkm. The study produced
five clusters, out of which one cluster covered adjoining regions and all other clusters were
scattered indicating irregular behaviour of daily rainfall pattern in India. The study was done
by using two time periods 1901–1975 and 1976–2010, and the resulting clusters were found to
be extremely different in the two time periods. The clusters formed in the northeastern region
is also different for the two periods which are not entirely in line with the present study.
However, they used gridded rainfall data instead of station data, thus the difference. Mannan
et al. (2018) have used climatic variables and self-organizing maps to regionalize India.
Artificial neural network is used along with four CVs for clustering and applied on gridded rainfall dataset ($0.25^\circ \times 0.25^\circ$) from IMD for 34 years (1980–2013) as well as climatic variables such as air temperature, surface pressure, geo-potential height, specific humidity, etc. 10 homogeneous regions were formed when only rainfall data was used, whereas incorporation of climatic variables divided the region into 15 regions. The region 2 in their study covered the northeast India with rainfall of 7.2 mm/day.

4. Conclusion

In this paper, fuzzy clustering approach has been used to classify regions with homogeneous rainfall in upper Brahmaputra Valley region of northeast India. Three different combinations of feature vectors were employed in FCM algorithm to attain the best solutions to regionalization. Seven different CVs were used to determine the optimal partition in the fuzzy c-means (FCM) algorithm, out of which Extended Xie and beni index ($V_{XB}$) and Kwon index ($V_K$) were opted for clustering of rainfall regions owing to their satisfactory performance. The optimal value of cluster number for all three cases was identified as $c=3$, with corresponding value of $m=1.5$. The clustered regions were then assessed for statistical homogeneity by performing homogeneity tests using L-moment approach. Four clusters were found to be acceptably homogeneous. Other possibly heterogeneous and definitely heterogeneous clusters were made homogenous by adjusting the discordant sites. It was found from the results that the clustering pattern was improved in case 3, where geographical location parameters (latitude, longitude and elevation) were included along with local rainfall data of tea gardens. It indicates that if regionalization needs to be done at a local scale such as an average sized watershed, further sub-clustering of a clustered region may be required. Local rainfall data, along with geographical location parameter details, can be used for the purpose since GCM data will not be of much use in this aspect because of their coarse resolution. However, a good rainfall dataset with large number of station points is required to be available within the region.
Acknowledgements

Authors acknowledge the teagarden authorities for making the data available. Our deepest gratitude is extended to the editors and anonymous reviewers for their valuable comments and suggestions.

Author’s Contribution

Conceptualization: Jayshree Hazarika, Arup Kumar Sarma; methodology: Jayshree Hazarika, Arup Kumar Sarma; formal analysis and investigation: Jayshree Hazarika; writing - original draft preparation: Jayshree Hazarika; writing - review and editing: Arup Kumar Sarma; resources: Arup Kumar Sarma; supervision: Arup Kumar Sarma.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and material

All relevant data are provided in both main manuscript and supplementary material.

Code availability

Not applicable

Declaration

Ethics approval All procedures performed in this study were in accordance with the ethical standards of the institution or any comparable ethical standards.

Consent to participate We (authors) are agreed that Jayshree Hazarika planned, performed the analysis, and wrote the paper, while Arup Kumar Sarma provided the resources and added his expertise in the analysed results.

Consent for publication We give consent for our paper to be published in your Theoretical and Applied Climatology journal.
Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Adelekan IO (1998) Spatio-temporal variations in thunderstorm rainfall over Nigeria. International Journal of Climatology 18:1273–1284. https://doi.org/10.1002/(SICI)1097-0088(199809)18:11<1273::AID-JOC298>3.0.CO;2-4

Agarwal A, Maheswaran R, Sehgal V, Khosa R, Sivakumar B, Bernhofer C (2016) Hydrologic regionalization using wavelet-based multiscale entropy method. Journal of Hydrology 538:22–32. https://doi.org/10.1016/j.jhydrol.2016.03.023

Asong ZE, Khaliq MN, Wheater HS (2015) Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical attributes. Stochastic Environmental Research and Risk Assessment 29:875–892. https://doi.org/10.1007/s00477-014-0918-z

Baltacı H, Göktürk OM, Kındap T, Ünal A, Karaca M (2015) Atmospheric circulation types in Marmara Region (NW Turkey) and their influence on precipitation. International Journal of Climatology 35:1810–1820. https://doi.org/10.1002/joc.4122

Baltacı H, Kındap T, Ünal A, Karaca M (2017) The influence of atmospheric circulation types on regional patterns of precipitation in Marmara (NW Turkey). Theoretical and Applied Climatology 127:563–572. https://doi.org/10.1007/s00704-015-1653-1

Bärring L (1987) Spatial patterns of daily rainfall in central Kenya: application of principal component analysis and spatial correlation. Journal of Climatology 7(3):267–290. https://doi.org/10.1002/joc.3370070306

Beable ME, McKercher AI (1982) Regional flood estimation in New Zealand. Water and Soil Technical Publication 20, Ministry of works and development, Wellington. N.Z
Bedi HS, Bindra MMS (1980) Principal components of monsoon rainfall. Tellus 32(3):296–298. https://doi.org/10.3402/tellusa.v32i3.10584

Bezdek JC (1974a) Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology 1(1):57–71. https://doi.org/10.1007/BF02339490

Bezdek JC (1974b) Cluster validity with fuzzy sets. Journal of Cybernetics 3(3):58–73. https://doi.org/10.1080/01969727308546047

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York. https://doi.org/10.1007/978-1-4757-0450-1

Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Computers & Geosciences 10:191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bharath R, Srinivas VV (2015) Delineation of homogeneous hydrometeorological regions using wavelet-based global fuzzy cluster analysis. International Journal of Climatology 35:4707–4727. https://doi.org/10.1002/joc.4318

Bonell M, Sumner G (1992) Atmospheric circulation and daily precipitation in Wales. Theoretical and Applied Climatology 46:3–25. https://doi.org/10.1007/BF00866443

Burn DH (1997) Catchment similarity for regional flood frequency analysis using seasonality measures. Journal of Hydrology 202:212–230. https://doi.org/10.1016/S0022-1694(97)00068-1

Chavoshi S, Azmin Sulaiman WN, Saghafian B, bin Sulaiman MN, Manaf LA (2013) Regionalization by fuzzy expert system based approach optimized by genetic algorithm. Journal of Hydrology 486:271–280. https://doi.org/10.1016/j.jhydrol.2013.01.033

Chen J, Zhao S, Wang H (2011) Risk Analysis of Flood Disaster Based on Fuzzy Clustering Method. Energy Procedia 5:1915–1919. https://doi.org/10.1016/j.egypro.2011.03.329
Chew HH, Heiler M, David T (1987) Magnitude and frequency of floods in peninsular Malaysia (revised and updated) 1987, 1987th edn. Ministry of Agriculture and Fisheries, Malaysia

Darand M, Daneshvar MRM (2014) Regionalization of precipitation regimes in Iran using principal component analysis and hierarchical clustering analysis. Environmental Processes 1:517–532. https://doi.org/10.1007/s40710-014-0039-1

Dikbas F, Firtat M, Koc AC, Gungor M (2012) Classification of precipitation series using fuzzy cluster method. International Journal of Climatology 32:1596–1603. https://doi.org/10.1002/joc.2350

Dinpashoh Y, Fakheri-Fard A, Moghaddam M, Jahanbakhsh S, Mirnia M (2004) Selection of variables for the purpose of regionalization of Iran’s precipitation climate using multivariate methods. Journal of Hydrology 297:109–123. https://doi.org/10.1016/j.jhydrol.2004.04.009

Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. Journal of Cybernetics 3(3):32–57. https://doi.org/10.1080/01969727308546046

Efe B, Lupo AR, Deniz A (2019) The relationship between atmospheric blocking and precipitation changes in Turkey between 1977 and 2016. Theoretical and Applied Climatology 138:1573–1590. https://doi.org/10.1007/s00704-019-02902-z

Farsadnia F, Kamrood MR, Nia AM, Modarres R, Bray MT, Han D, Sadatinejad J (2014) Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. Journal of Hydrology 509:387–397. https://doi.org/10.1016/j.jhydrol.2013.11.050

Fukuyama Y, Sugeno M (1989) A new method of choosing the number of clusters for the fuzzy c-means method. Proceedings of Fifth Fuzzy Systems Symposium, pp. 247–250 (in Japanese).
Gadgil S, Yadumani, Joshi NV (1993) Coherent rainfall zones of the Indian region. International Journal of Climatology 13(5):547–566. https://doi.org/10.1002/joc.3370130506

Goyal MK, Sharma A (2016) A fuzzy c-means approach regionalization for analysis of meteorological drought homogeneous regions in western India. Natural Hazards 84(3):1831–1847. https://doi.org/10.1007/s11069-016-2520-9

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. Water Resources Research 15(5):1049–1054. https://doi.org/10.1029/WR015i005p01049

Guttman NB (1993) The Use of L-Moments in the Determination of Regional Precipitation Climates. Journal of Climate 6:2309–2325. https://doi.org/10.1175/1520-0442(1993)006<2309:TUOLMI>2.0.CO;2

Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. Journal of Intelligent Information Systems 17:107–145. https://doi.org/10.1023/A:1012801612483

Hall MJ, Minns AW (1999) The classification of hydrologically homogeneous regions. Hydrological Sciences Journal 44(5):693–704. https://doi.org/10.1080/02626669909492268

Hosking JRM (1990) L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society: Series B (Methodological) banner 52(1):105–124. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x

Hosking JRM, Wallis JR (1993) Some Statistics useful in regional frequency analysis. Water Resources Research 29(2):271–281. https://doi.org/10.1029/92WR01980

Hosking JRM, Wallis JR (1995) Correction to “Some statistics useful in regional frequency analysis”. Water Resources Research 31(1):251. https://doi.org/10.1029/94WR02510

Hosking JRM, Wallis JR (1997) Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511529443
Irwin S, Srivastav RK, Simonovic SP, Burn DH (2017) Delineation of precipitation regions using location and atmospheric variables in two Canadian climate regions: the role of attribute selection. Hydrological Sciences Journal 62:191–204. https://doi.org/10.1080/02626667.2016.1183776

Iyengar RN, Basak P (1994) Regionalization of Indian monsoon rainfall and long-term variability signals. International Journal of Climatology 14:1095–1114. https://doi.org/10.1002/joc.3370141003

Karaca M, Deniz A, Tayanc M (2000) Cyclone Track Variability Over Turkey in Association with Regional Climate. International Journal of Climatology 20:1225–1236. https://doi.org/10.1002/1097-0088(200008)20:10%3C1225::AID-JOC535%3E3.0.CO;2-1

Kulkarni A (2017) Homogeneous clusters over India using probability density function of daily rainfall. Theoretical and Applied Climatology 129:633–643. https://doi.org/10.1007/s00704-016-1808-8

Kulkarni A, Kripalani RH, Singh SV (1992) Classification of summer monsoon rainfall patterns over India. International Journal of Climatology 12(3):269–280. https://doi.org/10.1002/joc.3370120304

Kwon SH (1998) Cluster Validity Index for Fuzzy Clustering. Electronics Letters 34(22):2176–2177. https://doi.org/10.1049/el:19981523

Machiwal D, Kumar S, Meena HM, Santra P, Singh RK, Singh DV (2019) Clustering of rainfall stations and distinguishing influential factors using PCA and HCA techniques over the western dry region of India. Meteorological Applications 26:300–311. https://doi.org/10.1002/met.1763

Mannan A, Chaudhary S, Dhanya CT, Swamy AK (2018) Regionalization of rainfall characteristics in India incorporating climatic variables and using self-organizing maps. ISH
Mok PY, Huang HQ, Kwok YL, Au JS (2012) A robust adaptive clustering analysis method for automatic identification of clusters. Pattern Recognition 45:3017–3033. https://doi.org/10.1016/j.patcog.2012.02.003

NERC (1975) Flood Studies Report, five volumes. Natural Environmental Research Council (NERC), Department of the Environment, London

Owen SM, MacKenzie AR, Bunce RGH, Stewart HE, Donovan RG, Stark G, Hewitt CN (2006) Urban land classification and its uncertainties using principal component and cluster analyses: A case study for the UK West Midlands. Landscape and Urban Planning 78:311–321. https://doi.org/10.1016/j.landurbplan.2005.11.002

Pal NR, Bezdek JC (1995) On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy Systems 3(3):370–379. https://doi.org/10.1109/91.413225

Pelczer I, Ramos J, Domínguez R, González F (2007) Establishment of regional homogeneous zones in a watershed using clustering algorithms. In: 32 Congress of IAHR Harmonizing the Demands of Art and Nature in Hydraulics, IAHR, Venice, Italy, File. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.3470

Plain MB, Minasny B, McBratney AB, Vervoort RW (2008) Spatially explicit seasonal forecasting using fuzzy spatiotemporal clustering of long-term daily rainfall and temperature data. Hydrology and Earth System Sciences Discussions 5:1159–1189. https://doi.org/10.5194/hessd-5-1159-2008

Rao AR, Srinivas VV (2006a) Regionalization of watersheds by hybrid-cluster analysis. Journal of Hydrology 318:37–56. https://doi.org/10.1016/j.jhydrol.2005.06.003

Rao AR, Srinivas VV (2006b) Regionalization of watersheds by fuzzy cluster analysis. Journal of Hydrology 318:57–79. https://doi.org/10.1016/j.jhydrol.2005.06.004
Roubens M (1982) Fuzzy clustering algorithms and their cluster validity. European Journal of Operational Research 10:294–301. https://doi.org/10.1016/0377-2217(82)90228-4

Sadri S, Burn DH (2011) A Fuzzy C-Means approach for regionalization using a bivariate homogeneity and discordancy approach. Journal of Hydrology 401:231–239. https://doi.org/10.1016/j.jhydrol.2011.02.027

Saikranthi K, Rao TN, Rajeevan M, Bhaskara Rao SV (2013) Identification and Validation of Homogeneous Rainfall Zones in India Using Correlation Analysis. Journal of Hydrometeorology 14:304–317. https://doi.org/10.1175/JHM-D-12-071.1

Satyanarayana P, Srinivas VV (2008) Regional frequency analysis of precipitation using large-scale atmospheric variables. Journal of Geophysical Research 113:1–16. https://doi.org/10.1029/2008JD010412

Satyanarayana P, Srinivas VV (2011) Regionalization of precipitation in data sparse areas using large scale atmospheric variables – A fuzzy clustering approach. Journal of Hydrology 405:462–473. https://doi.org/10.1016/j.jhydrol.2011.05.044

Singh KK, Singh SV (1996) Space-time variation and regionalization of seasonal and monthly summer monsoon rainfall of the sub-Himalayan region and Gangetic plains of India. Climate Research 6:251–262. https://doi.org/10.3354/cr006251

Srinivas VV, Tripathi S, Rao AR, Govindaraju RS (2008) Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348:148–166. https://doi.org/10.1016/j.jhydrol.2007.09.046

Srinivasa RK, Nagesh KD (2007) Classification of Indian meteorological stations using cluster and fuzzy cluster analysis, and Kohonen artificial neural networks. Nordic Hydrology 38(3):303–314. https://doi.org/10.2166/nh.2007.013
Sumner G, Bonell M (1988) Variation in the spatial organisation of daily rainfall during the
north Queensland wet seasons, 1979–82. Theoretical and Applied Climatology 39:59–72.
https://doi.org/10.1007/BF00866390

Thomas DM, Benson Ma (1970) Generalization of streamflow characteristics from drainage-
basin characteristics. https://doi.org/10.3133/wsp1975

Unal Y, Kindap T, Karaca M (2003) Redefining the climate zones of Turkey using cluster
analysis. International Journal of Climatology 23:1045–1055. https://doi.org/10.1002/joc.910

Unal YS, Deniz A, Toros H, Incecik S (2012) Temporal and spatial patterns of precipitation
variability for annual, wet, and dry seasons in Turkey. International Journal of Climatology
32(3):392–405. https://doi.org/10.1002/joc.2274

Venkatesh B, Jose MK (2007) Identification of homogeneous rainfall regimes in parts of
Western Ghats region of Karnataka. Journal of Earth System Science 116(4):321–329.
https://doi.org/10.1007/s12040-007-0029-z

Viglione A, Laio F, Claps P (2007) A comparison of homogeneity tests for regional frequency
analysis. Water Resources Research 43(W03428):1–10. https://doi.org/10.1029/2006WR005095

Wang Z, Zeng Z, Lai C, Lin W, Wu X, Chen X (2017) A regional frequency analysis of
precipitation extremes in Mainland China with fuzzy c-means and L-moments approaches.
International Journal of Climatology 37: 429–444. https://doi.org/10.1002/joc.5013

Wotling G, Bouvier C, Danloux J, Fritsch J-M (2000) Regionalization of extreme precipitation
distribution using the principal components of the topographical environment. Journal of
Hydrology 233:86–101. https://doi.org/10.1016/S0022-1694(00)00232-8

Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Transactions on Pattern
Analysis and Machine Intelligence 13(8):841–847. https://doi.org/10.1109/34.85677

28
Table 1 Location details of raingauge stations located in various tea gardens of the upper Brahmaputra valley region with mean annual and seasonal precipitation

Table 2 Results of homogeneity test
| Sl. No. | Name of the station | Longitude (E) | Latitude (N) | Seasonal rainfall in mm | Total annual rainfall in mm | Elevation in m |
|--------|---------------------|--------------|--------------|-------------------------|----------------------------|----------------|
| 1      | Abhoijan            | 93.8986111   | 26.4225      | 70 610 1127 510         | 2317                       | 121            |
| 2      | Achabam             | 95.2644444   | 27.24        | 82 579 1249 401         | 2312                       | 123            |
| 3      | Amsoi               | 92.4102778   | 26.143333    | 31 326 843 346          | 1546                       | 67             |
| 4      | Anand               | 94.2300797   | 27.4619764   | 101 525 1880 719         | 3225                       | 112            |
| 5      | Arin                | 93.9752778   | 26.487778    | 49 478 831 321          | 1679                       | 100            |
| 6      | Arun                | 92.4402778   | 26.669444    | 32 451 823 257          | 1563                       | 80             |
| 7      | Athabari            | 94.0347222   | 26.413889    | 9 301 525 162          | 996                        | 104            |
| 8      | Azizibaghi          | 95.1302778   | 27.206389    | 111 698 1354 479        | 2642                       | 113            |
| 9      | Bahani              | 94.2025      | 26.755556    | 79 552 905 449          | 1985                       | 93             |
| 10     | Basmatia            | 95.0671167   | 27.360556    | 81 637 1126 409         | 2253                       | 113            |
| 11     | Bateli              | 92.2537422   | 26.775958    | 35 430 907 340          | 1712                       | 112            |
| 12     | Bhelaguri           | 94.3846566   | 26.704831    | 57 489 830 279          | 1654                       | 116            |
| 13     | Bokajan             | 93.7751667   | 26.022       | 22 260 548 254          | 1084                       | 136            |
| 14     | Bokakhat            | 93.6438889   | 26.637777    | 58 553 975 382          | 1968                       | 87             |
| 15     | Borahi              | 94.9902778   | 27.043333    | 59 684 1291 454         | 2488                       | 104            |
| 16     | Borchaporhi         | 93.6899      | 26.6381972   | 75 550 1143 406         | 2174                       | 92             |
| 17     | Borhat              | 95.2888889   | 27.138888    | 29 238 406 187          | 859                        | 122            |
| 18     | Borjan              | 94.0577778   | 26.5619444   | 39 393 745 271          | 1448                       | 104            |
| 19     | Borpathar           | 93.8488889   | 26.272222    | 44 404 708 274          | 1430                       | 127            |
| 20     | Chubwa              | 95.1730556   | 27.466111    | 99 835 1447 601         | 2981                       | 118            |
| 21     | Cinnatollah         | 94.0857124   | 27.342037    | 110 653 1709 589        | 3061                       | 123            |
| 22     | Dalowjan            | 93.9747222   | 26.433888    | 54 425 729 347          | 1555                       | 105            |
| 23     | Deamoolie           | 95.5388889   | 27.590555    | 91 670 1275 442         | 2478                       | 134            |
| 24     | Dejoo               | 94.0031433   | 27.280721    | 106 661 1690 544        | 3002                       | 123            |
| 25     | Dekorai             | 92.9636778   | 26.81        | 56 608 1238 445         | 2347                       | 83             |
| 26     | Deohall             | 95.2890372   | 27.4223095   | 103 710 1152 379        | 2344                       | 125            |
| 27     | Dhekiajuli          | 92.4615703   | 26.6904332   | 57 572 1090 391         | 2110                       | 77             |
| 28     | Dholaguri           | 93.84161     | 26.5128231   | 49 479 822 272          | 1622                       | 99             |
|   | City       | Latitude  | Longitude | Population | Area 1 | Area 2 | Area 3 | Area 4 | Area 5 |
|---|------------|-----------|-----------|------------|--------|--------|--------|--------|--------|
| 29| Digulturrung | 95.4119444 | 27.6104528 | 58         | 631    | 905    | 309    | 1903   | 126    |
| 30| Dilli       | 95.3672222 | 27.1638889 | 118        | 732    | 1507   | 519    | 2877   | 132    |
| 31| Diphloo     | 93.5685556 | 26.6394444 | 31         | 578    | 910    | 303    | 1822   | 84     |
| 32| Dooria      | 93.9191667 | 26.6316667 | 46         | 460    | 927    | 242    | 1675   | 97     |
| 33| Duklingia   | 94.2852151 | 26.6861544 | 74         | 685    | 1329   | 428    | 2515   | 112    |
| 34| Durrung     | 92.7297222 | 26.7227778 | 50         | 445    | 768    | 271    | 1534   | 76     |
| 35| Furkating   | 94.0137903 | 26.4617283 | 55         | 395    | 768    | 318    | 1535   | 106    |
| 36| Halem       | 93.4523611 | 26.8696667 | 44         | 449    | 1022   | 341    | 1856   | 90     |
| 37| Halmira     | 93.9466667 | 26.5241667 | 47         | 468    | 832    | 340    | 1688   | 101    |
| 38| Harchurah   | 92.7537139 | 26.775488  | 61         | 561    | 1226   | 435    | 2282   | 94     |
| 39| Hatigarh    | 94.0433333 | 26.3888889 | 44         | 452    | 785    | 309    | 1590   | 102    |
| 40| Hatikhuli   | 93.370899  | 26.5848439 | 51         | 540    | 906    | 373    | 1870   | 87     |
| 41| Kakojan     | 94.3861111 | 26.735     | 46         | 440    | 786    | 247    | 1518   | 107    |
| 42| Kellyden    | 92.948746  | 26.483845  | 41         | 440    | 927    | 292    | 1700   | 86     |
| 43| Keyhung     | 95.3279024 | 27.4345426 | 101        | 658    | 1238   | 461    | 2457   | 127    |
| 44| Khowang     | 94.8938766 | 27.2431657 | 80         | 627    | 1052   | 337    | 2096   | 103    |
| 45| Koomsong    | 95.6537533 | 27.6188283 | 109        | 925    | 1182   | 476    | 2693   | 144    |
| 46| Kopati      | 92.25073   | 26.59      | 45         | 504    | 839    | 297    | 1685   | 75     |
| 47| Lakwa       | 94.8722222 | 27.0244444 | 93         | 607    | 1286   | 429    | 2416   | 100    |
| 48| Lamabari    | 92.277784  | 26.842633  | 35         | 504    | 1218   | 408    | 2166   | 134    |
| 49| Ledo        | 95.7683333 | 27.3020278 | 77         | 392    | 577    | 288    | 1333   | 153    |
| 50| Lengeree    | 93.7166667 | 25.8911111 | 41         | 299    | 711    | 302    | 1354   | 154    |
| 51| Lepektata   | 94.8635316 | 27.3783221 | 222        | 649    | 1113   | 417    | 2402   | 104    |
| 52| Madhuting   | 95.3667005 | 27.3581907 | 27         | 262    | 451    | 157    | 897    | 129    |
| 53| Mahalakshmi | 93.0608333 | 26.8555556 | 28         | 432    | 733    | 280    | 1473   | 88     |
| 54| Maijan      | 94.9783246 | 27.5079422 | 118        | 623    | 1359   | 529    | 2629   | 111    |
| 55| Mancotta    | 94.9175806 | 27.4385028 | 130        | 865    | 1764   | 653    | 3412   | 108    |
| 56| Mazbat      | 92.2515106 | 26.7810162 | 37         | 559    | 1185   | 379    | 2160   | 113    |
| 57| Moran       | 94.8846352 | 27.1563477 | 72         | 609    | 1060   | 405    | 2146   | 103    |
| 58| Murphuloni  | 93.9247879 | 26.457985  | 36         | 495    | 868    | 295    | 1694   | 106    |
| 59| Nahorjan    | 93.6065063 | 26.6075556 | 62         | 553    | 895    | 302    | 1812   | 98     |
| 60| Namburnodi  | 93.83188   | 26.29877   | 50         | 399    | 731    | 326    | 1506   | 134    |
| 61| Namdang     | 95.7220833 | 27.2655556 | 111        | 730    | 1555   | 534    | 2930   | 427    |
|   | City       | Latitude   | Longitude  | 1  | 2  | 3  | 4  | 5  | 6  |
|---|------------|------------|------------|----|----|----|----|----|----|
| 62| Namrup     | 95.3286111 | 27.1947222 |  50|  429| 731| 256| 1465| 125|
| 63| Nitinnagar | 94.052032  | 26.345432  |  35|  294| 561| 223| 1113| 107|
| 64| Nonoi      | 92.921667  | 26.4041667 |  35|  487|1090| 383| 1994|  79|
| 65| Ouphulia   | 95.0147222 | 27.2177778 |  92|  704|1350| 493| 2639| 109|
| 66| Paneery    | 91.8962681 | 26.7463385 |  64|  680|1057| 453| 2254| 122|
| 67| Panitola   | 95.2573121 | 27.4941394 |  95|  470|1087| 463| 2114| 124|
| 68| Pavoijan   | 93.9055538 | 26.298396  |  44|  371| 654 | 275| 1344| 116|
| 69| Powai      | 95.6482333 | 27.3478431 |  77|  613|1262| 377| 2329| 158|
| 70| Rungamatty | 93.8986111 | 26.6702778 |  53|  543| 894 | 335| 1825|  90|
| 71| Rupai      | 95.4952778 | 27.6123056 | 111|  726|1313| 458| 2608| 132|
| 72| Rupajuli   | 92.7217448 | 26.7260395 |  53|  584|1225| 440| 2301|  78|
| 73| Sagmootea  | 93.004862  | 26.544426  |  49|  480|1013| 309| 1851|   92|
| 74| Santi      | 95.4125    | 27.3038889 | 102|  965|1368| 447| 2882| 127|
| 75| Sepoi      | 92.4102034 | 26.7837818 |  17|  704|1077| 327| 2125| 110|
| 76| Sepon      | 94.845947  | 27.1150673 |  77|  584|1141| 394| 2195| 103|
| 77| Scketting  | 94.087428  | 26.546912  |  57|  422| 900 | 314| 1693|   98|
| 78| Sonabheel  | 92.7849483 | 26.7369584 |  48|  465|1045| 364| 1921|   76|
| 79| Sundarpur  | 95.1924133 | 27.1453118 |  53|  725|1612| 656| 3046| 113|
| 80| Teloijan   | 94.9437618 | 27.2513485 |  55|  355| 798 | 329| 1537| 103|
| 81| Tezpore &  | 92.7395976 | 26.7232366 |  61|  570|1030| 359| 2020|   79|
gogra                  |
| 82| Thanai     | 95.0936111 | 27.5357694 |  73|  576|1071| 396| 2116| 116|
| 83| Uday jyoti | 93.9638889 | 26.5       |  19|  383| 678 | 264| 1344|   99|

Note: DJF = December, January, February
MAM = March, April, May
JJA = June, July, August
SON = September, October, November
Table 2 Results of homogeneity test

| Case | Cluster no. | H test | Homogeneity result | After adjustment | Homogeneity result | Final no of sites |
|------|-------------|--------|--------------------|-----------------|-------------------|------------------|
|      |             | H1     | H2     | H3     |                   | H1     | H2     | H3     |                   |                   |                  |
| 1    | C1          | -0.08  | -1.25  | -0.99  | Acceptably homogeneous | 0.68   | -0.97  | -1.06  | Acceptably homogeneous | 31               |
|      | C2          | 1.08   | -0.69  | 1.1    | Possibly heterogeneous | 0.49   | -1.04  | 0.9    | Acceptably homogeneous | 34               |
|      | C3          | 2.96   | 1.11   | 0.46   | Definitely heterogeneous | -0.84  | 0.56   | -0.05  | Acceptably homogeneous | 16               |
| 2    | C1          | -0.98  | -0.73  | 0.04   | Acceptably homogeneous | -0.91  | -0.62  | 0.1    | Acceptably homogeneous | 31               |
|      | C2          | 0.54   | -1.25  | 0.09   | Acceptably homogeneous | 0.41   | -0.71  | 0.39   | Acceptably homogeneous | 44               |
|      | C3          | 1.52   | 0.86   | 0.69   | Possibly heterogeneous | 0.41   | -1.23  | -1.28  | Acceptably homogeneous | 7                |
| 3    | C1          | 1.86   | 0.91   | 0.15   | Possibly heterogeneous | 0.71   | 0.85   | -0.11  | Acceptably homogeneous | 13               |
|      | C2          | -0.88  | -2.24  | -0.25  | Acceptably homogeneous | 0.73   | -1.95  | 0.16   | Acceptably homogeneous | 49               |
|      | C3          | 1.62   | -0.05  | -0.16  | Possibly heterogeneous | 0.69   | -0.86  | -0.93  | Acceptably homogeneous | 21               |
LIST OF FIGURES

Fig. 1 Methodology used to identify homogeneous rainfall regions

Fig. 2 Raingauge stations located in various tea gardens of the upper Brahmaputra valley region

Fig. 3 Variation in the optimum value of objective function of FCM algorithm with variation of fuzzifier $m$ and cluster number $c$, for (a) Case 1: with total monthly rainfall as attributes; (b) Case 2: with standard deviation of total monthly rainfall as attributes; and (c) Case 3: with latitude, longitude, elevation, total annual rainfall and standard deviation of total annual rainfall as attributes

Fig. 4 Clusters formed by the FCM algorithm after adjustment for (a) Case 1: with total monthly rainfall as attributes; (b) Case 2: with standard deviation of total monthly rainfall as attributes; and (c) Case 3: with latitude, longitude, elevation, total annual rainfall and standard deviation of total annual rainfall as attributes
Form fuzzy clusters applying FCM algorithm

Identify optimum cluster number by using cluster validity indices

Test whether the clusters are homogeneous

Yes

Homogeneous rainfall regions

No

Adjust the heterogeneous clusters

Fig. 1 Methodology used to identify homogeneous rainfall regions
Fig. 2 Raingauge stations located in various tea gardens of the upper Brahmaputra valley region
(a) Case 1

(b) Case 2
Fig. 3 Variation in the optimum value of objective function of FCM algorithm with variation of fuzzifier $m$ and cluster number $c$, for (a) Case 1: with total monthly rainfall as attributes; (b) Case 2: with standard deviation of total monthly rainfall as attributes; and (c) Case 3: with latitude, longitude, elevation, total annual rainfall and standard deviation of total annual rainfall as attributes.
(a) Case 1
Fig. 4 Clusters formed by the FCM algorithm after adjustment for (a) Case 1: with total monthly rainfall as attributes; (b) Case 2: with standard deviation of total monthly rainfall as attributes; and (c) Case 3: with latitude, longitude, elevation, total annual rainfall and standard deviation of total annual rainfall as attributes.