Identification of homogeneous rainfall regions using a genetic algorithm involving multi-criteria decision making techniques

Nilotpal Debbarma, Parthasarathi Choudhury and Parthajit Roy

ABSTRACT

The detection of appropriate homogeneous regions is an important step in regional frequency analysis with the determination of homogeneity depending to a great extent on the type of method used in grouping. So, the study considers a genetic-algorithm-based clustering method to identify homogeneous precipitation regions for 39 gauge stations of the north-eastern region of India. The performance evaluation is done using six cluster validation measures. Further, considering all the six indices together, selection for the optimum cluster is modelled as a multi-criteria decision making (MCDM) problem. Three MCDM methods, namely TOPSIS, WASPAS and VIKOR, are applied to obtain ranked clusters which are then subjected to a heterogeneity test using the L-moments approach. The results suggested the stations to be grouped into three homogeneous regions. Comparison with the *k*-means method indicated relatively better performance for genetic-algorithm-based clustering. Finally, an L-moment ratio diagram and goodness-of-fit measures were conducted to select regional frequency distributions for the identified homogeneous regions.

Key words | genetic algorithm, homogeneous, *k*-means, multi-criteria decision making

Nilotpal Debbarma (corresponding author)
Parthasarathi Choudhury
Parthajit Roy
Civil Engineering Department,
NIT Silchar,
Assam,
India
E-mail: *dbnilotpal@gmail.com*

INTRODUCTION

Estimation of extreme rainfall events and their frequency of occurrence is a crucial requirement in the design and hydrological planning for many water resources projects. The estimation generally involves conducting a frequency analysis of the observed at-site data for the project site, followed by the fitting of a probability distribution. However, it has been found that the site of interest most often falls short in the desired length of records and becomes insufficient for the conducting of a reliable frequency analysis. Therefore, approaches like regionalization have been developed that can incorporate the information from nearby adjoining areas to provide reliable estimates of required extreme rainfall or flood events for the project site. This technique can also be extended to estimations for ungauged catchments or areas within a homogeneous region. Several methods of regionalization used in identification of homogeneous precipitation regions can be found from studies in the literature e.g. Bärring (1988), Iyengar & Basak 1994, Baeriswyl & Rebetez (1997), Comrie & Glenn (1998), Ngongondo *et al.* (2011), Gaál *et al.* (2008), Satyanarayana & Srinivas (2011), Goyal & Gupta (2014) and Asong *et al.* (2015). But, clustering analysis is found to be more popularly used in the regionalization studies of homogeneous precipitation regions.

In recent decades, the extreme precipitation scenario in the north-eastern region of India has caused a lot of catastrophic damage and havoc and human misery to be experienced, necessitating the requirement for more updated and reliable information on extreme events. The observed precipitation is seen to change both in magnitude and frequency over space and time in the region with numerous casualties and cases of damage reported every year caused by heavy rainfall and flood events during pre-monsoon and monsoon

doi: 10.2166/ws.2019.018

periods. So, their accurate and more reliable estimation will be essentially helpful in disaster mitigation works in the region and hydrological design and planning of any water resources projects. Some notable works on the study of extreme rainfall events in the region are available in the studies of Deka *et al.* (2011), Mahanta *et al.* (2013), Goyal & Gupta (2014), and Bora *et al.* 2016. For example, Bora *et al.* (2016) studied rainfall frequency analysis of maximum rainfall for 12 rain gauge stations using L-moments (linear combinations of order statistics for parameter estimation) and LQ-moments (quick estimators replacing expectation definitions in L-moments). Deka *et al.* (2011) applied higher order L-moments (LH-moments) considering nine ground gauge stations in the region for the maximum daily annual rainfall events. Goyal & Gupta (2014) formed homogeneous precipitation regions for the region using fuzzy *c*-means based on annual total precipitations of 68 station gauges. Other studies reporting on the scenario of maximum rainfall can be found in Mahanta *et al.* (2013). But, the rainfall estimation based on clustered homogeneous regions in the region is scarcely explored. Moreover, most of the studies have restricted to only a few stations except for Goyal & Gupta (2014). And to the best of our knowledge to date, there is no study available in the literature to apply the genetic algorithm in forming homogeneous precipitation station groups in the region. The main objectives of the paper, therefore, focus on (i) application of genetic-algorithm-based clustering and evaluation of performance using cluster validation measures, (ii) ranking of the clusters obtained involving the multi-criteria decision making (MCDM) technique to select appropriate homogeneous clusters using a heterogeneity test, and (iii) determining the best-fit regional probability distribution for the identified homogeneous regions.

METHODOLOGY

Genetic-algorithm-based clustering

Genetic algorithms proposed by Holland (1975) are evolutionary algorithms which use the principle of natural genetics and evolution, and have been widely studied for clustering problems. In the present study, a real-life dataset consisting of 39 rain gauge stations located on the south bank of Brahmaputra valley and Barak valley are taken up for clustering. The chromosome representation is done with real numbers considering station characteristics as attributes. Here, each chromosome represents the cluster centres with a length of $K \times N_a$, where K is the number of clusters and N_a is the number of attributes. Each chromosome is assigned a probability of selection on the basis of its cost in the sorted population and then selected based on roulette wheel selection. Heuristic crossover is considered with crossover probability set as 0.8 for generation of off-spring. Mutation percentage is taken as 0.3, mutation rate of 0.01, and the gene values are modified at randomly selected locations. The fitness function in the study considered is the Davies–Bouldin index, which is an appropriate cluster validity measure to obtain a desired set of centres or solutions with maximum possible within-cluster similarity and minimum between-cluster similarity. Results of the genetic-algorithm-based clustering are also compared in the study with the *k*-means algorithm. To measure the performance of the clustering results, seven cluster validation measures are selected and defined briefly in Table 1.

Multi-criteria decision methods

Assessment of performance using the cluster validation measures and selection for best optimum clusters is decided using three MCDM methods, namely WASPAS, TOPSIS and VIKOR. Many MCDM techniques are available in the literature for choosing the number of clusters, but there is no particularly efficient technique. So, in the study all three MCDM techniques are chosen so that a more agreeable ranking can be decided in choosing the number of clusters. The study considers Shannon information entropy theory in determining the weights in all the three MCDM methods. The best ranked clusters given by the MCDM methods are then subjected to a heterogeneity test using L-moments theory to select the final homogeneous region.

L-moments and heterogeneity measure

L-moments (Hosking 1990; Hosking & Wallis 1997) are alternative ways of determining the parameters of a probability distribution from a data sample and are the modified forms of probability weighted moments (PWMs)

Table 1 | Cluster validation measures used for the study

Cluster validation measure	Definition	Optimal value
CS index	$CS = \frac{\sum_{i=1}^{k}\left\{\frac{1}{n_i}\sum_{x_j\in C_i}\max_{x_k\in C_i}\{d(x_j, x_k)\}\right\}}{\sum_{i=1}^{k}\{\min_{j\in\{1,2,\ldots,K\},j\neq i}\{d(c_i, c_j)\}\}}, c_i = \frac{1}{n_i}\sum_{x_j\in C_i} x_j$	Min value
SD index	$SD = \alpha S_a + S_t\ ;\ S_a = \frac{1}{k}\sum_{i=1}^{k}\frac{\|\sigma(v_i)\|}{\|\sigma(X)\|};$ $S_t = \frac{\max_{1\le i,j\le k}\|v_i - v_j\|}{\min_{1\le i,j\le k}\|v_i - v_j\|}\left\{\sum_{i=1}^{k}\left(\sum_{i=1}^{k}\|v_i - v_j\|\right)^{-1}\right\}$	Min value
Dunn index	$Dunn = \min_{i=1}^{k}\left\{\min_{j=i+1}^{k}\left(\min_{x\in c_i, y\in c_j}\|x - y\|^2\right)\right\}/\max_{p=1}^{k}\left\{\max_{x,y\in c_p}\|x - y\|^2\right\}$	Max value
Calinski–Harabasz index	$CH(k) = \frac{(n-k)\{\sum_{i=1}^{k}\|C_i\|(Z_i - Z)^T(Z_i - Z)}{(k-1)\sum_{i=1}^{k}\sum_{x\in C_i}(x - z_i)^T(x - z_i)}$	Max value
Kraznowski and Lai index	$KL = \left\|\frac{DIFF(k)}{DIFF(k+1)}\right\|;$ $DIFF(k) = (k-1)^{2/p}W_{k-1} - k^{2/p}W_k,\ p$ is number of variables	Max value
Davies–Bouldin index	$DB = \frac{1}{k}\sum_{i=1}^{N}\max_{i,j\neq i}\left\{\left[\frac{1}{n_i}\sum_{x\in C_i}\|x - c_i\|^2 + \frac{1}{n_j}\sum_{x\in C_j}\|x - c_j\|^2\right]/\|c_i - c_j\|^2\right\}$	Min value
Silhouette index	$Silh = \frac{1}{k}\sum_{i}\left\{\frac{1}{n_i}\sum_{x\in C_i}\left(\frac{b(x) - a(x)}{\max[b(x), a(x)]}\right)\right\}$	Max value

K – number of clusters; N – number of objects of the dataset; n_i – number of objects in cluster C_i; α – weighting factor; S_j – silhouette of object j; z – the mean of the entire dataset; z_i – the mean of cluster C_i.

as defined by Greenwood *et al.* (1979). A heterogeneity statistic, H was proposed by Hosking & Wallis (1997) to identify the homogeneity in a group of regions based on a discordancy measure and L-moment statistics. For the heterogeneity test of a group, a four-parameter kappa distribution is fitted to the regional dataset generated from series of 500 equivalent region data using Monte-Carlo simulation. For each region, the regional L-moment ratios and V-statistic are computed and based on the vector of the V-statistic, mean (μ_{vi}) and standard deviation σ_{vi}, and the heterogeneity measure can be calculated as:

$$H_i = \frac{V_i - \mu_{vi}}{\sigma_{vi}},\ i\ =\ 1,\ 2,\ 3 \tag{1}$$

On the basis of homogeneity measurements as suggested by Hosking & Wallis (1997), a region or group of sites is considered 'acceptably homogeneous' if $H_1 < 1$, possibly heterogeneous if $1 \le H_1 < 2$ and definitely heterogeneous if $H_1 \ge 2$. Finally in the study, the identified homogeneous regions are examined with an L-moment ratio diagram and goodness-of-fit test in selection of the best fitted distributions of extreme rainfall in the region.

STUDY AREA

The study area chosen is located in the north-eastern region of India covering the south bank of Brahmaputra valley in Assam and Barak valley. It lies between 22° N to 28° N and 89° E to 96° E and covers the north-eastern states Assam, Manipur, Nagaland, Meghalyaya, Mizoram and Tripura. The areas on the south bank of Brahmaputra valley are plain lands and the areas covering Barak valley

Table 2 | Summary of attributes of the 39 rain gauge stations

Attribute	Range
Mean of annual daily maximum rainfall (mm)	80.97 to 585.98
Latitude	23.73 °N to 27.75 °N
Longitude	89.98 °E to 95.68 °E
Altitude (m)	16 to 1,598

are hilly in nature. The areas are poorly gauged and the availability of data especially in the Barak valley is sparse. Thirty-nine rain gauge stations were selected from the two valleys with a study period covering years from 2001 to 2010. The annual daily maximum rainfall data were obtained from the Regional Meteorological Centre, Guwahati. Four attributes, namely longitude, latitude, altitude and average annual maximum daily precipitation, were selected as feature attributes for use in regionalization. The statistical summary of the considered attributes is presented in Table 2.

RESULTS AND DISCUSSION

Identification of cluster groups using genetic algorithm

Choosing the minimum number of clusters as two and the maximum number of clusters as seven, a clustering study using the genetic algorithm was conducted. The upper bound of maximum clusters, i.e. seven, was fixed on the basis of the formula $n^{1/2} = 7$ (Urcid & Ritter 2012), where n is the number of rain gauge stations. The mutation rate was set at 0.01 to avoid dilution of the best results, the crossover percentage was found to be best at 0.8, and population was tried for 500 iterations. The data were normalized using the min–max normalization method before clustering. From the results in Table 3, the Dunn and KL indices gave values in a decreasing trend with cluster 2 as the optimum cluster. Of the other indices, Calinski–Harbasz showed a reverse increasing trend with cluster 6 having the maximum index value and cluster 5 as second best, while the Silhouette index was found varying with the optimum value at cluster 2. In case of the CS index, the best minimal values were observed for cluster 2 followed by cluster number 5. As the formed clustered regions are to be further checked for statistical homogeneity (Hosking & Wallis 1997), the selection for optimum cluster will depend on the other ranked optimum values. The SD index gave an increasing trend with cluster number 2 having the minimum value. Overall, no single clusters with optimum value in all six validation measures were found. So, the judgement for choosing an optimum cluster based on the performance of any single type of chosen validation measure will be less reliable and incomplete. To decide on an optimum cluster, the problem is considered as a MCDM problem that will give the ranking for optimum clusters utilizing all the six validation measures.

Ranking of clusters using MCDM analysis

Three MCDM methods, namely TOPSIS, WASPAS and VIKOR, are considered for the purpose of ranking the clusters. And as the rankings provided by a single MCDM method may generally differ from those of another MCDM method, so only the best similarly matched rankings generated in all three MCDM methods are considered. Here for each MCDM method, cluster validation measures were taken as the criteria and the number of clusters as the

Table 3 | Results of cluster validation indices for genetic-algorithm-based clustering

		Cluster number					
Clustering algorithm	Validation measures	2	3	4	5	6	7
Genetic algorithm	Dunn	0.5832	0.3208	0.2404	0.1231	0.1087	0.1980
	Silh	0.7145	0.6568	0.5340	0.5982	0.6412	0.6144
	CH	12.4079	16.0868	21.1853	32.2067	33.4479	24.5129
	KL	10.4990	9.0672	7.0411	4.6288	4.0019	4.6869
	CS	0.5334	0.6721	0.6934	0.6412	0.6460	0.7031
	SD	2.6463	3.6152	4.5500	4.7072	4.5834	5.0079

alternatives. And the weights of each cluster validation measure were determined using Shannon information entropy. Then, cluster validation indices with the objective of attainment of maximum value (i.e. Dunn, Silh, KL and CH indices) were considered as benefit criteria, whereas indices requiring minimal value (i.e. CS and SD indices) were taken as cost criteria. All the MCDM methods were executed using the R software package 'MCDM' (Blanca & Ceballos 2016). The results obtained in Table 4 show a similarity in rankings generated by all the three MCDM methods with the first rank for optimum cluster being 2, followed by cluster number 3 and cluster number 4. So, the first three best rankings obtained are chosen to be further subjected to heterogeneity analysis for homogeneous regions.

To compare the performance, *k*-means clustering was applied for the same range of clusters and is presented in Table 5. Both Dunn and KL indices gave cluster 2 as the optimum cluster. The Silhouette index gave the optimum for cluster number 2 and the CH index gave cluster 6 as the optimum. The DB and CS indices showed a decreasing trend giving cluster number 7 as the optimum cluster. Table 6 produces the similarity in rankings generated for *k*-means in all three MCDM methods for all cluster numbers, suggesting cluster number 2 as the optimum choice.

Table 4 | MCDM ranking by TOPSIS, WASPAS and VIKOR for genetic-algorithm-based clustering

		TOPSIS		WAPSAS		VIKOR	
Clustering algorithm	**Cluster number**	**Value**	**Rank**	**Value**	**Rank**	**Value**	**Rank**
Genetic algorithm	2	0.7487	1	0.8852	1	0.1507	1
	3	0.4874	2	0.7112	2	0.2861	2
	4	0.3235	3	0.6180	3	0.7108	3
	5	0.2565	5	0.5737	5	0.9018	5
	6	0.2612	4	0.5652	6	0.9174	6
	7	0.2351	6	0.5760	4	0.7721	4

Table 5 | Results of cluster validation indices for *k*-means

		Cluster number					
Clustering algorithm	**Validation measures**	**2**	**3**	**4**	**5**	**6**	**7**
k-means	Dunn	0.3144	0.2332	0.1466	0.1851	0.1720	0.1722
	Silh	0.6945	0.5384	0.5979	0.6885	0.6939	0.6432
	CH	19.9935	19.5681	28.3695	42.0843	51.6409	44.6178
	KL	9.1016	8.2270	5.7777	3.7249	2.7518	2.8005
	DB	0.9521	0.9439	0.6177	0.5389	0.5234	0.4881
	CS	1.0855	1.2556	0.6550	0.6274	0.6096	0.5693

Table 6 | MCDM ranking for *k*-means

		TOPSIS		WAPSAS		VIKOR	
Clustering algorithm	**Cluster number**	**Value**	**Rank**	**Value**	**Rank**	**Value**	**Rank**
k-means	2	0.9958	1	0.9972	1	0	1
	3	0.8118	2	0.8462	2	0.2654	2
	4	0.4408	3	0.6060	3	0.6261	3
	5	0.1627	4	0.4974	4	0.8449	4
	6	0.0503	5	0.4107	5	0.9876	5
	7	0.0472	6	0.4105	6	0.9961	6

Table 7 | Homogeneity measure of the clustering algorithms

Clustering algorithm	Number of clusters	Number of rain gauge stations in each cluster region	Heterogeneity			Region type
			H1	H2	H3	
Genetic algorithm	2	Region I : 2 stations	0.12	−0.83	−0.98	Homogeneous
		Region II* : 36 stations	0.36	0.05	0.21	Homogeneous
	3	Region I : 2 stations	0.12	−0.83	−0.98	Homogeneous
		Region II : 4 stations	−0.15	0.43	−0.22	Homogeneous
		Region III* : 32 stations	0.4	0.05	0.47	Homogeneous
	4	Region I : 7 stations	1.77	−0.21	0.09	Possibly heterogeneous
		Region II : 2 stations	0.12	−0.83	−0.98	Homogeneous
		Region III : 4 stations	−0.15	0.43	−0.22	Homogeneous
		Region IV* : 25 stations	0.08	−0.08	−0.19	Homogeneous
k-means	2	Region I : 5 stations	0.19	−0.64	−1.28	Homogeneous
		Region II* : 33 stations	0.61	0.07	0.4	Homogeneous
	3	Region I : 3 stations	0.4	−1.3	−1.81	Homogeneous
		Region II : 10 stations	1.79	0.6	0.73	Possibly heterogeneous
		Region III* : 25 stations	0.15	−0.03	−0.12	Homogeneous
	4	Region I : 2 stations	0.12	−0.83	−0.98	Homogeneous
		Region II : 4 stations	−0.15	0.43	−0.22	Homogeneous
		Region III : 10 stations	0.51	0.63	0.42	Homogeneous
		Region IV* : 22 stations	0.61	−0.08	0.35	Homogeneous

*Indicates adjusted region after removal of discordant Goalpara station.

Heterogeneity test

Initially, with the region considered as one whole homogeneous region the heterogeneity test conducted gave H1, H2 and H3 values as 0.80, −0.03 and −0.01 respectively. For genetic-algorithm-based clustering in Table 7, both cluster numbers 2 and 3 gave homogeneous regions, whereas cluster number 4 formed one heterogeneous region. In the analysis, each cluster had one discordant station, Goalpara, which was removed from the particular region to increase its homogeneity measure. The H1 values obtained for cluster number 2 are found to be better than the region as a single homogeneous region. Cluster 3 also gave three homogeneous regions with Region III giving an H1 value of 0.40, suggesting good homogeneity among the stations. And region II comprising of Shillong, Aizawl, Kohima and Imphal gave H1 with a negative value of −0.15 thereby indicating slightly positive correlation to exist among the frequency distribution of the sites. This may be reasonable, as all the four stations alongside Cherrapunjee and Mawsynram are situated at approximately equal elevations and are comparatively much higher than the remaining stations. Also, the region II stations are located in the Barak valley region and receive more heavy rainfall compared with stations located on the south bank of Brahmaputra valley during the pre-monsoon periods. The H2 and H3 values of regions I, II and III are below the value of −2 (Hosking & Wallis 1993) and are within the acceptable range. But, the H1 statistic will be mainly considered in the study for determination of the homogeneity of a region as the H2 and H3 statistics lack discriminatory power between homogeneous and heterogeneous regions (Hosking & Wallis 1997). So, we can conclude that cluster 3 is more preferable, producing more acceptably homogeneous regions without losing its heterogeneity in comparison with cluster 2 and cluster 4. So, the genetic-algorithm-based clustering suggests the study region to be divisible into three acceptably homogeneous regions.

For *k*-means clustering, the heterogeneity test confirmed homogeneous regions only for cluster numbers 2 and 4. The H1 and H3 values of the regions of cluster 4 were better than those of cluster 2, indicating less deviation between at-site estimates and observed data. So, cluster number 4 can be considered more preferable with more acceptable homogeneous sub-regions. Here, region I (Cherrapunjee and Mawsynram) and region II (Shillong, Aizawl, Imphal and Kohima) are commonly identified in both the algorithms; but the remaining 32 stations were considered as

Table 8 | Goodness-of-fit test for candidate distributions

Algorithm	Region	Acceptable distribution	Best fit Z^{dist}	Best distribution
Genetic algorithm	A	GLO(0.02), GEV(−0.36), GNO(−0.38), PE3(−0.48), GPA(−1.17)	0.02	GLO
	B	GLO(0.39), GEV(−0.12), GNO(−0.15), PE3(−0.28), GPA(−1.22)	0.12	GEV
	C	GEV(0.91), GNO(0.94), PE3(0.67)	0.67	PE3
k-means	A	GLO(0.02), GEV(−0.36), GNO(−0.38), PE3(−0.48), GPA(−1.17)	0.02	GLO
	B	GLO(0.39), GEV(−0.12), GNO(−0.15), PE3(−0.28), GPA(−1.22)	0.12	GEV
	C	GLO(0.19), GEV(−0.70), GNO(−0.66), PE3(−0.76)	0.19	GLO
	D	GEV(1.32), GNO(1.25), PE3(0.92)	0.92	PE3

one region in genetic-algorithm-based clustering and two sub-regions in *k*-means. Also, the region III and IV values of *k*-means for H1 are 0.51 and 0.61 respectively, which is more than the value of 0.40 as generated by genetic-algorithm-based clustering. Thus, dividing the 32 stations into two further sub-regions increases the heterogeneity measure. Region III in *k*-means has a larger H2 value compared with region III of genetic-algorithm-based clustering, indicating relatively more deviation between regional and at-site estimates. The H2 value of region IV of *k*-means is negative, indicating correlation between the frequency distribution of the sites. Thus, the comparative performances of heterogeneity measures of genetic-algorithm-based clustering to *k*-means clustering confirmed the former to be far superior.

Identification of regional distributions based on L-moment ratio diagrams and goodness-of-fit statistics

The suitability of fitting of each of the distributions to each homogeneous region is assessed through L-moment ratio diagrams (Hosking & Wallis 1997) and goodness-of-fit measure, the Z-statistic (Hosking & Wallis (1993, 1997)). The Z-statistic for a distribution is accepted at a significance level of 10% and the best distribution is selected based on satisfying the condition, $\min Z^{\text{Dist}}_{\text{crit}} \leq |1.64|$. The results of the goodness-of-fit test for 500 simulations for the regions identified by the clustering algorithms are presented in Table 8. The L-moment ratio diagrams of genetic-algorithm-based clustering for identifying regional frequency distributions are presented in Figure 1. Five candidate distributions chosen for the study are Generalized Logistic (GLO), Generalized Extreme Value (GEV), General Normal (GNO), Pearson Type-III (PE3) and Generalized Pareto (GPA). In Table 8, both genetic-algorithm-based clustering and *k*-means identified region A (Cherrapunjee and Mawsynram) as fitting the generalized logistic distribution. And from Figure 1 of the L-moments ratio diagram, the regional weighted L-moments for region A are the GLO distribution. Region II consisting of Shillong, Kohima, Imphal and Aizawl was found to be close to GEV distribution in the moment ratio diagram and the results were also confirmed by the goodness-of-fit test for both the algorithms. The rest of the stations that are lying in low-lying areas on the south bank of Brahmaputra valley and Barak valley are found to have Pearson Type-III distribution by genetic-algorithm-based clustering, and GLO and PE3 distributions by *k*-means clustering. Figure 2 shows the distribution of the homogeneous stations identified by genetic-algorithm-based clustering.

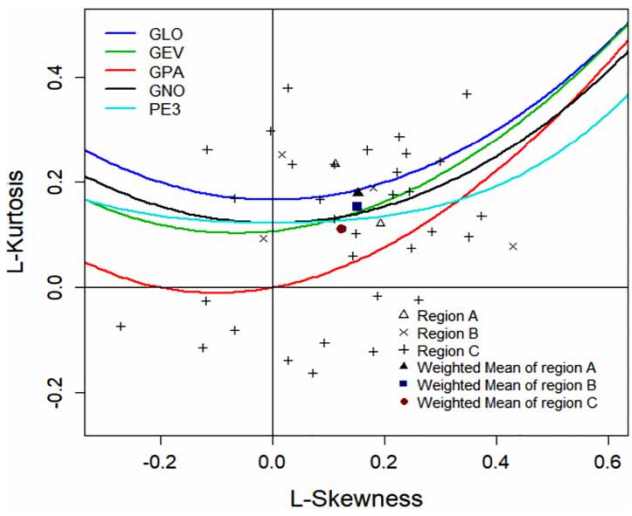

Figure 1 | L-moments ratio diagram of regions generated by the genetic algorithm method.

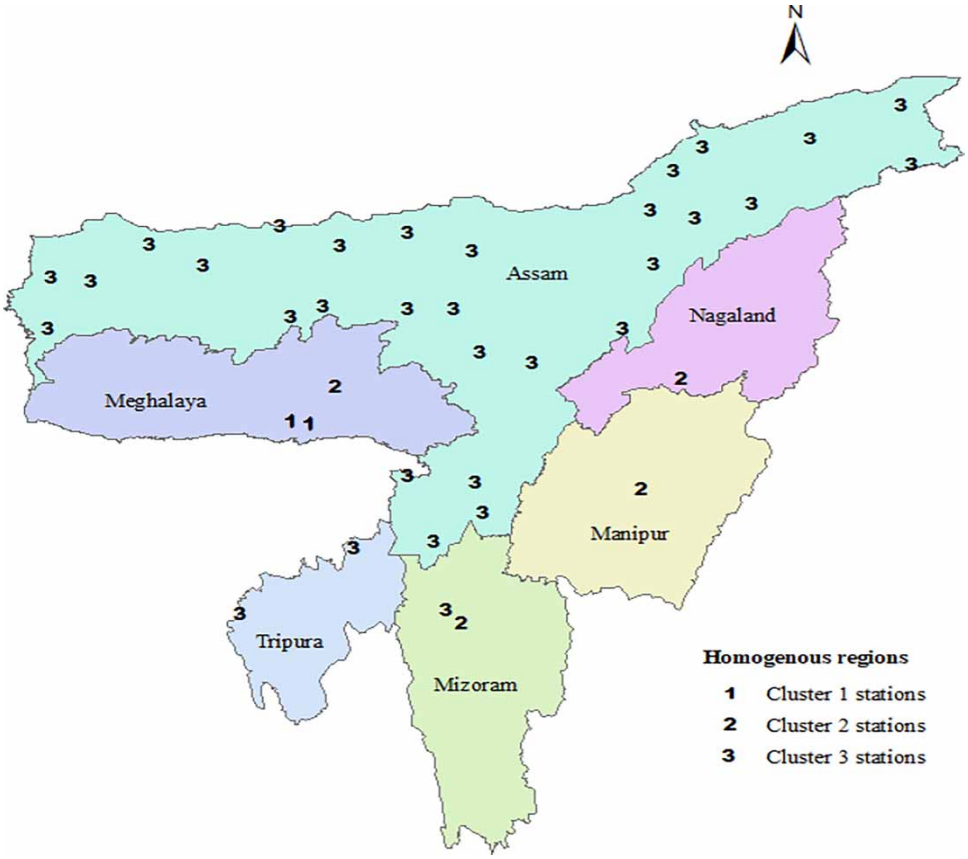

Figure 2 | Location of stations in regions identified by the genetic algorithm.

CONCLUSIONS

In this study, we have applied genetic-algorithm-based clustering to identify homogeneous precipitation regions involving MCDM. From the results obtained, the following conclusions can be made:

(i) The heterogeneity test conducted on genetic-algorithm-based clustering confirmed the formation of three acceptably homogeneous regions while *k*-means clustering suggested four homogeneous regions.
(ii) Comparative performances between the algorithms confirmed the superiority of genetic-algorithm-based clustering, resulting in partitioning of the south bank of Brahmaputra and Barak valley regions into three homogeneous precipitation regions.
(iii) The L-moments ratio diagram and goodness-of-fit measure suggested Generalized Logistic (GLO) distribution for homogeneous region I (Cherrapunjee and Mawsynram), Generalized Extreme Value (GEV) distribution for homogeneous region II (Shillong, Kohima, Imphal and Aizawl) and Pearson Type-III (PE3) distribution for region III (the remaining 32 stations) for the extreme rainfall values in the region.

Thus, the study has successfully identified the homogeneous regions utilizing the genetic algorithm as a clustering tool and this study can prove to augment the effectiveness of the genetic algorithm in the area of regionalization studies for hydro-meteorological data.

ACKNOWLEDGMENTS

The authors are thankful to the anonymous reviewers for their helpful suggestions in improving the quality and representation of this paper.

REFERENCES

Asong, Z. E., Khaliq, M. N. & Wheater, H. S. 2015 Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical attributes. *Stoch. Environ. Res. Risk Assess.* **29**, 875–892.

Baeriswyl, P.-A. & Rebetez, M. 1997 Regionalization of precipitation in Switzerland by means of principal component analysis. *Theoret. Appl. Climatol.* **58**, 31–41.

Bärring, L. 1988 Regionalisation of daily rainfall in Kenya by means of common factor analysis. *Journal of Climatology* **8**, 371–389.

Blanca, A. & Ceballos, M. 2016 *MCDM: Multi-Criteria Decision Making Methods for Crisp Data*. R Software Package, Version 1.2.

Bora, D. J., Borah, M. & Bhuyan, A. 2016 Regional analysis of maximum rainfall using L-moment and LQ-moment: a comparative case study for the North East India. *International Journal of Applied Mathematics & Statistical Sciences* **5** (6), 79–90.

Comrie, A. C. & Glenn, E. C. 1998 Principal components-based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability. *Clim. Res.* **10** (3), 201–215.

Deka, S., Borah, M. & Kakaty, S. C. 2011 Statistical analysis of annual maximum rainfall in North-East India: an application of LH-moments. *Theor. Appl. Climatol.* **104** (1), 111–122.

Gaál, L., Kyselý, J. & Szolgay, J. 2008 Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia. *Hydrol. Earth Syst. Sci.* **12** (3), 825–839.

Goyal, M. K. & Gupta, V. 2014 Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis. *Water Resources Management* **28** (13), 4491–4511.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. & Wallis, J. R. 1979 Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research* **15** (5), 1049–1054.

Holland, J. H. 1975 *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA.

Hosking, J. R. M. 1990 L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc. Ser. B* **52**, 105–124.

Hosking, J. R. M. & Wallis, J. R. 1993 Some statistics useful in regional frequency analysis. *Water Resources Research* **29** (2), 271–281.

Hosking, J. R. M. & Wallis, J. R. 1997 *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, Cambridge, UK.

Iyengar, R. N. & Basak, P. 1994 Regionalization of Indian monsoon rainfall and long term variability signals. *International Journal of Climatology* **14**, 1095–1114.

Mahanta, R., Sarma, D. & Choudhury, A. 2013 Heavy rainfall occurrences in northeast India. *International Journal of Climatology* **33**, 1456–1469.

Ngongondo, C. S., Xu, C. Y., Tallaksen, L. M., Alemaw, B. & Chirwa, T. 2011 Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches. *Stoch. Env. Res. Risk Assess.* **25**, 939–955.

Satyanarayana, P. & Srinivas, V. V. 2011 Regionalization of precipitation in data sparse areas using large scale atmospheric variables: a fuzzy clustering approach. *Journal of Hydrol.* **405**, 462–473.

Urcid, G. & Ritter, G. X. 2012 C-means clustering of lattice auto-associative memories for endmember approximation. In: *Advances in Knowledge-Based and Intelligent Information and Engineering Systems* (M. Graña, C. Toro, J. Posada, R. J. Howlett & L. C. Jain, eds). IOS Press, Amsterdam, The Netherlands, pp. 2140–2149.

First received 12 September 2018; accepted in revised form 8 January 2019. Available online 28 January 2019