Water Practice & Technology

IWA PUBLISHING



Water Practice & Technology Vol 16 No 4, 1446 doi: 10.2166/wpt.2021.086

# Comparision of performance of multi criteria decision making ensemble-clustering algorithms in rainfall frequency analysis

Nilotpal Debbarma, Parthasarathi Choudhury and Parthajit Roy

Civil Engineering Department, NIT Silchar, Assam, India
*Corresponding author. E-mail: nilotpal_rs@civil.nits.ac.in

ND, 0000-0003-0657-3742

## ABSTRACT

Non-availability of adequate extreme rainfall information at any place of interest are solved using regionalization where subjective grouping of similar attributes of nearby gauged stations is performed. K-Means and Fuzzy C-Means are commonly used methods in regionalization of rainfall, but application of genetic algorithms is very rarely explored. Genetic algorithms (GA) are highly efficient evolutionary algorithms, and through an appropriate objective function can effectively achieve the purpose of clustering. In the present study, Davies-Bouldin index is considered and validation is performed using a set of validation measures. Taking into account the varied output obtained in each validation measure, an ensemble approach involving multi criteria decision making is applied to obtain optimal ranked solutions, and the procedure is extended to K-Means and Fuzzy C-Means for comparison. From the results obtained, GA-based clustering is found to outperform the other two algorithms in formation of homogenous regions with better performance in leave-one-out cross validation (LOOCV) test and sensitivity analysis. Accuracy of regional growth curves of regions assessed using regional relative bias and RMSE suggest low uncertainty and accurate quantile estimates in GA regions. Further, information transfer index based on entropy evaluated among GA regions is found to be highest and K-Means lowest.

**Key words:** clustering, Fuzzy C-Means, genetic algorithm, information transfer index, sensitivity

## HIGHLIGHTS

- Comparison of genetic algorithm-based clustering to K-Means and Fuzzy C-Means with ensemble MCDM technique.
- Optimum cluster based on several cluster validation indices and MCDM.
- Sensitivity analysis of MCDM rankings.
- Regional growth curve comparision for regions delineated by all methods.
- Information transfer among stations in cluster regions with entropy based information transfer index.

## INTRODUCTION

Information related to extreme rainfall occurrence is a primary requirement for proceeding with planning and construction of any water resources-related projects. The interest in extreme rainfall quantiles in particular is more crucial when the negative consequences are devastating to a specific region, followed by non-availability of adequate rainfall information. Regionalization techniques in such situations have been mostly preferred, which provide reliable and accurate estimates, and an enormous amount of work can be found in the literature. Most of the regionalization methods involve application of clustering methods like K-Means, Fuzzy C-Means, hierarchical methods, and so on. But the application of evolutionary algorithms in regionalization studies of rainfall is scarcely available. Evolutionary algorithms have an efficient global search ability in generating optimal solutions and have therefore been successfully applied in many other domains relating to determination of number of clusters (Bandyopadhyay & Maulik 2002; Hruschka *et al.* 2004; Kuo *et al.* 2012; Ozturk *et al.* 2015). But, the application of the evolutionary algorithms in clustering rainfall stations is less studied and rare, and thus deserves more attention and exploration. The study therefore considers application of GA-based clustering in determination of annual extreme rain cluster regions utilizing seven characteristics of rainfall stations. To put the evolutionary algorithms to serve the purpose of clustering, selection of an appropriate objective function is

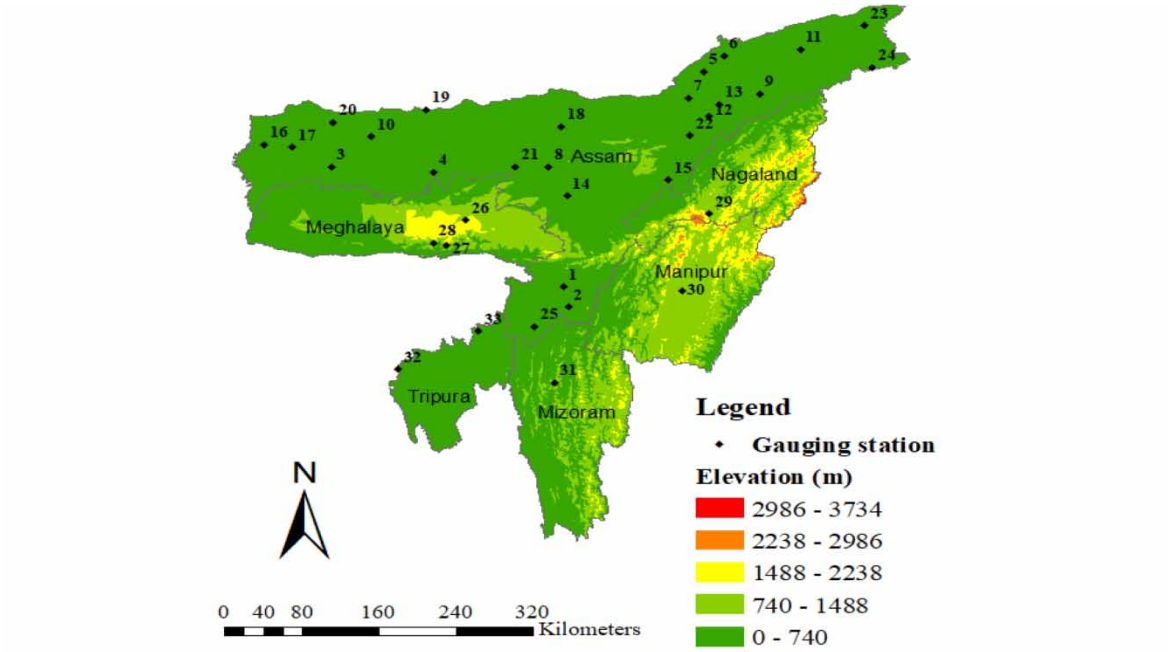
an important step that must satisfy the goal of achieving distinctly separate and compact clusters. Several problem-specific objective functions are available in the literature and only a few applications incorporating the Davies-Bouldin index in objective function of evolutionary algorithms to obtain clusters are studied (Maulik & Bandyopadhyay 2002; Lin *et al.* 2005; Liu *et al.* 2011; Agustín Blas *et al.* 2012). The index is found to be more preferable compared to Dunn, XB indices (Lin *et al.* 2005) and, hence, in the present study it is applied to obtain reliable compact rainfall regions that are distinctly different from other cluster regions based on station characteristics.

After the determination of cluster regions, a crucial step involves validating the clustering results to assess the quality and stability of determined clusters. External and internal type validity indexes are available in large amounts that address these issues, but the external type indices are mostly used to perform comparison of clustering outcomes from different algorithms or a single algorithm with different parameter settings. Whereas the internal indexes require no prior information and utilize the inherent information in a data. In today's results, however, prior information related to the data often remain concealed, posing difficulty to the users about the number of clusters. And therefore, internal indices are mostly preferred as they attempt to find both compactness within a cluster and separation between clusters. Some of the most well-known internal indices used are Calinski-Harabaz, Silhouette Coefficient, Dunn, Davies-Bouldin, CS and Xie-Beni (Gan *et al.* 2007; Peng *et al.* 2012; Zaki & Wagner 2014). In the present study, nine internal-type indices for GA-based clustering and K-Means are considered, and six fuzzy type indices for Fuzzy C-Means clustering. Previous studies on cluster validation in regionalization studies were limited to use of only a few types of cluster validation measures. As there exists a large amount of different cluster validation indices, selection of any particular type of index for the problem is preference based and not straightforward. There is no set rule for selecting the best cluster validation index and each index may provide different results that may prove suitable for a particular type of data and unsuitable for some. The selection of the optimum cluster will thus be affected by the type of indices selected and subjective choice of the decision maker. Further, for data sets with no prior knowledge on the number of clusters, selection of an inappropriate cluster validation measure may further lead to more ambiguity and uncertainty. So, in the present study, inclusion of a greater number of cluster validation indices will reduce the ambiguity and be more reliable in selecting the optimum cluster. But with the increase in number of selected validity indices, the ease in deciding the optimum cluster becomes complicated and thus emphasis has to be put again on the choice of decision-making. The required decision from the results of various indices can thus be designed as a multi criteria decision problem where one has to choose the best cluster from varying solution alternatives with different selection criteria. Multi criteria decision making methods like TOPSIS (Hwang & Yoon 1981), WASPAS (Zavadskas *et al.* 2012), VIKOR (Opricovic & Tzeng 2004), PROMETHEE (Mareschal *et al.* 1984), AHP (Saaty 1988), and so on, have efficiently solved multi criteria decision problems in various fields. The integrated approach of selection rainfall regions from several cluster validation measures with the application of multi-criteria decision-making techniques is rarely reported in literature. The study also extends the application to K-Means and Fuzzy C-Means and is compared to GA-based clustering, which is also rarely addressed. The comparative study will augment the performance metrics of MCDM methods in ranking rainfall cluster regions and their applicability in evaluating cluster validation measures in the area of water resources. Three MCDM techniques, namely WASPAS, VIKOR and TOPSIS, are applied and a similarity match in rankings of any two MCDM techniques will be taken as the final selected ranking.

Reports from previous studies (Basu & Srinivas 2014; Goyal & Gupta 2014) reveal that even after selection of best cluster based on the set of validity indices, the best cluster is not statistically homogeneous, necessitating further changes in cluster (Hosking & Wallis 1997) or to search for a potential solution in other cluster divisions. In such situations, the rankings provided by the application of the MCDM method will simplify both the lack of expertise in choosing suitable validity indices and the uncertainty of obtaining statistically homogeneous clusters. As studies suggest the selection of best cluster based on different types of cluster validation measures, there exists scope for more improvement in selection of the appropriate number of cluster validation measures. With the aim of inclusion of more appropriate cluster validation measures, and in combination with the multi-criteria decision-making technique, the study aims at achieving reliable and robust cluster. Keeping in view of the above scope of application, the paper mainly focuses on the performance of three different clustering algorithms guided by the application of the MCDM technique. Secondly, the appropriateness of the ranked clusters will be assessed through sensitivity analysis of the MCDM techniques to the cluster validation measures. Finally, with application of the L-Moments approach for the suitability of homogenous regions produced by the algorithms is reported through regional growth curves and entropy transfer index of regions.

### Study area and rainfall data

The study region concentrates on the Southern bank of the Brahmaputra basin including the whole of Barak basin in northeast India. Thirty-three rain gauge stations covering spatially the Barak and Brahmaputrra basins are selected, and annual maximum daily rainfall data for 20 years are collected from the Regional Meteorological Centre, Guwahati. The stations located in the study area are presented in Figure 1. The Barak basin adjacent to the Brahmaputra basin has active floodplains with a large coverage of marshy lands, which are every year subjected to extreme flood inundation. The altitudinal pattern varies abruptly from place to place, followed with erratic rainfall occurrences. The heavy rainfall and cloud bursts have devastating effects in the region annually with widespread landslides and erosion. The extreme rain behaviour in the region is very uncertain and makes the future rainfall scenario highly vulnerable.




List of stations - Silchar(1), Dholai(2), Goalpara(3), Guwahati(4), North Lakhimpur(5), Choudhoaghat(6), Batadighat(7), Kampur(8), Sibsagar(9), Beki Rd. Bridge(10), Dibrugarh(11), Jorhat(12), Neamatighat(13), Kherunighat(14), Bokajan(15), Gossaigaon(16), Kokrajhar(17), Tezpur(18), Mellabazar(19), Aie NH.Xing(20), Dharamtul(21), Golaghat(22), Dhollabazar(23), Margherita(24), Gharmura(25), Shillong(26), Cherrapunjee(27), Mawsynram(28), Kohima(29), Imphal(30), Aizwal(31), Agartala(32) and Kailashahar(33).

**Figure 1** | Location of rain gauge stations in study area of northeast region of India.

## METHODOLOGY

This section provides a brief background on the three clustering algorithms used in the study for formation of homogenous rainfall regions. Figure 2 shows a schematic diagram of the steps for the formation of homogenous regions using the MCDM technique. The section also includes the sensitivity analysis of multi criteria decision-making technique rankings for each algorithm, information transfer index calculated based on entropy for each region. The assessment of regional growth curves using Monte Carlo simulation and comparison between regions are also discussed.

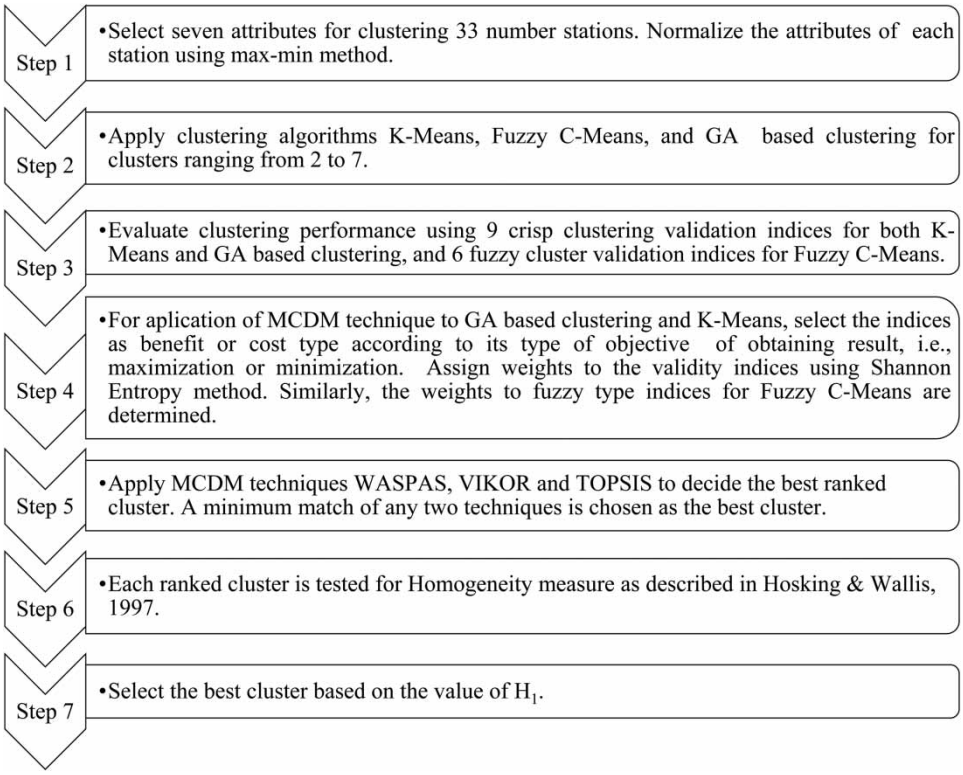



**Figure 2** | Steps for determination of homogenous cluster regions using MCDM technique.

### Genetic algorithm-based clustering

Genetic algorithms are evolutionary algorithms, which have provided precise optimal solutions for fitness functions of various complex optimization problems. Due to their superior capability in finding near optimal solutions, they have found applications in clustering problems with extremely good results in obtaining optimal clusters. In the present study, the chromosome representation is done with real number representation or floating-point representation.

### Initialization of population and fitness function

An initial random population is generated using chromosomes with a size of $N_p \times N_a$, where $N_p$ is the number of population size and $N_a$ is the number of attributes or the length of the chromosome. Here, $N_p$ is taken as 100 and $N_a$ is taken as 7. The range of clusters is studied between $K_{min}$ and $K_{max}$ chosen as two and $n^{0.5}$ respectively, where n represents the number of data points in the clustering data set. The chromosomes are generated using uniform distribution randomly in the range of 0 and 1. The population data set is normalized in the range 0–1 using the min-max method. The fitness function considered here in the study is the clustering problem to obtain desired clusters or solutions, where within-cluster similarity is maximized and between-cluster similarity is minimized. The Davies–Bouldin index is an appropriate index that determines the suitability of cluster solutions with determination of within-cluster scatter and between-cluster separation. Hence, it is incorporated in as the objective function and a minimum obtained value indicates a good clustering result. Davies-Bouldin (DB) (Davies & Bouldin 1979) is the ratio of within-cluster scatter and between-cluster separation and is given as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} (R_{ij}) \tag{1}$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \text{ where, } i \neq j, d_{ij} = \left\| c_i - c_j \right\|^2 \tag{2}$$

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \left\| x_j - c_i \right\|^2 \tag{3}$$

is the number of clusters, $d_{ij}$ is the distance (Euclidean distance in the present study) between centroids $c_i$ and $c_j$ of the $i^{th}$ and $j^{th}$ clusters respectively, $S_i$ represents the average Euclidean distance of all members in the $i^{th}$ cluster to its centroid($c_i$), $n_i$ is the number of members in the $i^{th}$ cluster.

### K-Means

K-Means clustering is the most popularly used clustering algorithm and is an iterative distance-based algorithm that aims to determine the clusters by minimization of the objective function given as:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{C} \|x_j - C_i\|^2 \tag{4}$$

Here, $\|x_j - C_i\|^2$ is the Euclidean distance between the $j^{th}$ data point and the $i^{th}$ cluster center. The number of clusters is decided beforehand, and cluster centers are allocated at random. Based on a set distance measure, each datum is individually assigned in any particular cluster. The data that are closest to a particular centroid are grouped together into a single cluster. All other data are similarly assigned to their respective clusters in accordance with their proximity to centroids. The centroids so formed are then recalculated and new centroids are created. All datasets and newly obtained centroids are further regrouped again repeatedly until there is no further change in the centroids formed, thus creating permanent clusters. As the method uses random numbers for initialization, the precision of the results very much depends on the initial selection of centroids.

### Fuzzy C-Means

K-Means in extended form with a fuzzy concept is the Fuzzy C-Means (Bezdek 1981) where data points are partitioned based on membership values ranging from 0 to 1. Here, a data point may belong to one or more clusters and is obtained by iterative optimization of minimizing an objective function. For a data set with N number of samples $X = \{x_1, x_2, x_3, \ldots x_N\}$ and each data having r dimensions, the objective function is

$$Z(U, C) = \sum_{i=1}^{K} \sum_{j=1}^{N} u_{ij}^{m} \|x_j - C_i\|^2 \tag{5}$$

where $x_j$ is the $j^{th}$ data of r-dimensional data, $C_i$ is the $i^{th}$ centroid with r-dimension, i = 1,2……. K and j = 1,2……., N. The method requires initialization of a membership matrix $U = [u_{ij}]$ using membership value $u_{ij}$ such that $\sum_{j=1}^{C} u_{ij} = 1.0$ for all data points. Center vectors are found at each step and the final U matrix is decided based on designed convergence criterion value $\varepsilon$. Here, C is the number of clusters, and m is the fuzzifier. The fuzzifier value depicts the dispersion of the data set within a cluster. The value is greater than 1. An important shortcoming to consider in Fuzzy C-Means clustering is its sensitivity to the parameters C and m.

### Cluster validity indices, multi criteria decision making methods and sensitivity

Nine cluster validation indices, viz. CS, Calinski-Harabasz, Davies-Bouldin, PBM, Xie-Beni, Kraznowski-Lai, Dunn, SD index, and Silhouette were selected to evaluate the clustering results from GA-based clustering and K-Means. Details of the methods can be found in (Calinsky & Harabasz 1974, Krzanowski & Lai 1988; Gan *et al.* 2007; Peng *et al.* 2012). And there are a total of eight cluster validation indices, namely Partition Coefficient, Classification Entropy, Xie-Beni index, Kwon, Tang, Fuzzy Silhouette index, Modified Partition Coefficient, Seperation index (Bensaid *et al.* 1996; Tang *et al.* 2005; Campello & Hruschka 2006; Wang & Zhang 2007) for fuzzy clustering. The decision matrix to be used for each MCDM is constructed from the performance of each cluster number for each cluster validation measure and is given by where xij is the performance of the ith alternative with the jth criterion, m is the number of alternatives (number of clusters) and n is the number of criteria (number of cluster validation indices). For determining the weights of the criteria (validity indices) Shannon Entropy method is used as in this method the weights are determined without consideration of the decision of the decision maker. A higher value of entropy weight indicates a higher importance of the criteria. All the elements of the decision matrix are normalized based on the type of indices;

that is, maximization or minimization.

$$y_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \text{ (Benefit or maximization type indices)} \tag{6}$$

$$y_{ij} = \frac{\max(x_{ij}) - x_{ij}}{\max(x_{ij}) - \min(x_{ij})} \text{ (Cost or minimization type indices)} \tag{7}$$

Three MCDM techniques, WASPAS, VIKOR and TOPSIS, are considered for the study in determining the rankings of the clusters for selection of homogenous regions. A larger value of VIKOR index indicates a better decision for an alternative. A detailed description on the TOPSIS and VIKOR methods can be found in Opricovic & Tzeng (2004) and WASPAS can be found in Zavadskas *et al.* (2012).

In the present study, the sensitivity of the three MCDM methods to the weights of validity indices in selection of optimum cluster is investigated. Four perturbation changes are applied to every weight where any weight $w_p$ ($p = 1,2, \ldots$. n) reduces to $w_p^* = \gamma_p w_p$ using initial variation ratio $\gamma_p$ (Li *et al.* 2013). As all weights must sum to 1 the other weights change accordingly and are given as:

$$\left.\begin{aligned} w_1^{/} &= \frac{w_1}{w_1 + w_2 + \ldots w_p^* + \ldots w_n} = w_1/\{1 + (\gamma_p - 1)w_p\} \\ w_2^{/} &= \frac{w_2}{w_1 + w_2 + \ldots w_p^* + \ldots w_n} = w_2/\{1 + (\gamma_p - 1)w_p\} \\ &\vdots \\ w_p^{/} &= \frac{w_p^*}{w_1 + w_2 + \ldots w_p^* + \ldots w_n} = \gamma_p w_p/\{1 + (\gamma_p - 1)w_p\} \\ &\vdots \\ w_n^{/} &= \frac{w_n}{w_1 + w_2 + \ldots w_p^* + \ldots w_n} = w_n/\{1 + (\gamma_p - 1)w_p\} \end{aligned}\right\} \tag{8}$$

where, $\gamma_p = \beta_p - \beta_p w_p/1 - \beta_p w_p$ and $\beta_p = w_p^{/}/w_p$. $\beta_p$ is the unitary variation ratio of $w_p$ for disturbance values 0.01, 0.1, 0.2 and 0.5. Note when $\beta_p = 1.0$, the weights get reduced to their original weights. Thus, the lowest value of $\beta_p$ produces the highest variation in a particular weight and vice versa.

## Heterogeneity measure

Hosking & Wallis (1993, 1997) proposed a heterogeneity statistic for computing the degree of heterogeneity in a group of sites. The test compares the variability of the L-Statistics of at-site L-moment ratios to the expected variability within representative synthetic homogeneous regions.

$$H_i = \frac{V_i - \mu_{vi}}{\sigma_{vi}}, \text{ i} = 1, 2, 3 \tag{9}$$

Three heterogeneity measurements $H_1$, $H_2$ and $H_3$ are calculated using V-statistic, mean ($\mu_{vi}$) and standard deviation $\sigma_{vi}$ of $N_{sim}$ values of V-statistic. V-Statistic V1, V2 and V3 correspond to the weighted standard deviation of the at-site sample L-coefficient of variation, the weighted measure based on deviation of at-site sample L-CV and L-Skewness, and the weighted measure based on at-site sample L-Skewness and L-Kurtosis, respectively. On the basis of homogeneity measurements, as suggested by Hosking & Wallis (1997), a region or group of sites is considered 'acceptably homogenous' if $H_1 < 1$, possibly heterogeneous if $1 \leq H_1 < 2$ and definitely heterogeneous if $H_1 \geq 2$. Further, $H_2$ and $H_3$ measures lack discriminating power compared to the $H_1$ measure (Hosking & Wallis 1997; Rao & Hamed 2000) and hence the $H_1$ statistic is accounted for as the main heterogeneity measure in the study.

### Entropy based information flow in regions using information transfer index

For determining the information transfer between one station to an adjacent station in produced cluster regions, the information transfer index (ITI) as expressed by Ridolfi *et al.* (2016) is applied and is expressed as:

$$ITI(A, B) = \frac{T(A, B)}{H(A, B)} \tag{10}$$

$$T(A, B) = H(A) + H(B) - H(A, B) \tag{11}$$

Here, $H(A)$ and $H(B)$ are marginal entropies of stations A and B, while $H(A,B)$ is their joint entropy. *ITI* is a symmetric index and represents the mutual information transfer between the two stations. The value lies between 0 and 1 and a higher value indicates a better information communication between the two stations.

### Assessment of accuracy of estimated regional growth curve

The regional growth curve in each homogenous region was determined using Hosking & Wallis (1997) and the assessment of accuracy of the curve was expressed through regional relative bias and regional relative RMSE. The procedure involves generation of regional average L-moments using a Monte Carlo simulation and in the process of simulation, quantile estimates for various return periods are calculated. At the $m^{th}$ repetition, the estimated quantiles for non-exceedance probability F is $\hat{Q}_i^{[m]}(F)$ and compared with the true values $Q_i(F)$. The relative error of this estimate at site $i$ for non-exceedance probability $F$ is:

$$\frac{\hat{Q}_i^{[m]}(F) - \hat{Q}_i(F)}{\hat{Q}_i(F)} \tag{12}$$

This quantity is squared and averaged over the M repetitions to obtain relative bias and mean relative quadratic error:

$$B_i(F) = M^{-1} \sum_{m=1}^{M} \frac{\hat{Q}_i^{[m]}(F) - \hat{Q}_i(F)}{\hat{Q}_i(F)} \tag{13}$$

$$R_i(F) = \left[ M^{-1} \sum_{m=1}^{M} \left\{ \frac{\hat{Q}_i^{[m]}(F) - \hat{Q}_i(F)}{\hat{Q}_i(F)} \right\}^2 \right]^{1/2} \tag{14}$$

Finally, the regional relative bias and relative root mean square error of the estimated quantile are:

$$B^R(F) = N^{-1} \sum_{i=1}^{N} B_i(F) \tag{15}$$

$$R^R(F) = N^{-1} \sum_{i=1}^{N} R_i(F) \tag{16}$$

## RESULTS AND DISCUSSION

The rain gauge stations were verified for trend and randomness of data series using the Mann-Kendall test and Ljung Box test. The analysis results suggest no trend and the data were serially independent and hence are suitable for statistical frequency analysis and fitting of probability distributions.

Nine cluster validity indices were computed to evaluate the performance of GA-based clustering. The performance of each cluster validity measure for each cluster number are reported in Table 1. From the table it can be seen that the CH index suggests cluster 5 as optimum cluster, while indices DB XB, CS and SD, Dunn, PBM, KL and Sil indices suggest optimum cluster as 2. Cluster division 2 is not statistically homogenous and ineffective as 31 stations are held in one group and two stations in the other cluster. The heterogeneity value $H_1$ of the group of 31 stations is 1.06 with many stations as discordant stations. Neglecting the results of cluster 2, the search for

**Table 1** | Performance of GA-based clustering for various cluster validation indices

| Cluster number | CS | CH | DB | Dunn | PBM | XB | KL | SD | Sil |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.3869 | 19.0408 | 0.3276 | 1.0017 | 1.1204 | 0.1291 | 10.8527 | 1.9716 | 0.6164 |
| 3 | 0.6242 | 17.8914 | 0.6844 | 0.2252 | 0.8862 | 2.0257 | 8.9706 | 4.4135 | 0.2790 |
| 4 | 0.6483 | 16.3139 | 0.6788 | 0.2288 | 0.5036 | 1.4869 | 7.9457 | 4.5594 | 0.3271 |
| 5 | 0.6728 | 22.1092 | 0.6583 | 0.1643 | 0.5554 | 3.4521 | 5.4735 | 4.8160 | 0.3989 |
| 6 | 0.6093 | 20.6454 | 0.5836 | 0.1643 | 0.4541 | 2.9763 | 4.9714 | 4.4910 | 0.4045 |
| 7 | 0.7746 | 17.1320 | 0.5752 | 0.1532 | 0.4252 | 4.1657 | 5.0587 | 5.5300 | 0.4079 |

the next best cluster in other cluster divisions points to different clusters with no single solution. For example, indices DB, Dunn, Sil give cluster 7 as optimum, CS as 2, CH gives 5, PBM, KL, SD give 3 as optimum cluster and XB gives 4 as optimum cluster. So, obtaining a single optimum cluster by normal judgement is not possible and requires a lot of subjective decision making. The MCDM approach is therefore applied in selecting the optimum cluster by ranking of clusters, which will simplify both the relation of results of each criterion (cluster validity measure) and the verification of statistically homogeneous clusters. Herein, the range of clusters determined in cluster analysis are ranked, utilizing a multi-criteria decision method with number of clusters as alternatives and cluster validity indices as criteria. Before proceeding with the application of multi-criteria decision making, the weights of each cluster validity index to act as criteria are evaluated using the Shannon Entropy method and is reported in Table 2. Indices Sil, Dunn, PBM, KL and CH are maximization type and considered as benefit-type criteria, while CS, XB, DB and SD type indices are minimization type or cost criteria. In the study, a total of three MCDM techniques of TOPSIS, VIKOR and WASPAS were considered for the ranking of clusters. As the rankings provided by any single MCDM method may generally differ from those of another MCDM method, so only the best similarly matched rankings generated by all three MCDM methods are considered. All the MCDM methods were executed using the R software package 'MCDM' (Blanca & Ceballos 2016). The results obtained in Table 3 show a similarity in rankings for GA-based clustering for TOPSIS and

**Table 2** | MCDM ranking for clusters in GA-based clustering

| | | TOPSIS | | WASPAS | | VIKOR | |
|---|---|---|---|---|---|---|---|
| Algorithm | Cluster number | Value | Rank | Value | Rank | Value | Rank |
| GA-based clustering | 2 | 0.9766 | 1 | 0.9891 | 1 | 0 | 1 |
| | 3 | 0.2231 | 2 | 0.4573 | 2 | 0.8294 | 2 |
| | 4 | 0.1755 | 3 | 0.418 | 3 | 0.8818 | 3 |
| | 5 | 0.089 | 5 | 0.3872 | 4 | 0.9287 | 5 |
| | 6 | 0.1005 | 4 | 0.3841 | 5 | 0.9233 | 4 |
| | 7 | 0.0651 | 6 | 0.3534 | 6 | 1 | 6 |

**Table 3** | Heterogeneity measure for first three similar rankings given by MCDM

| Clustering algorithms | Number of clusters | Number of rain gauge stations in each cluster region | Heterogeneity $H_1$ | Heterogeneity $H_2$ | Heterogeneity $H_3$ | Region type |
|---|---|---|---|---|---|---|
| All 33 stations | 1 | All 33 stations | 1.11 | 1.12 | 0.41 | Heterogenous |
| GA-based clustering | 2 | Region I–2 stations | −0.54 | −0.83 | −0.15 | Homogenous |
| | | Region II–31 stations | 1.04 | 1.03 | 0.38 | Heterogenous |
| | 3 | Region I–9 stations | 0.33 | 0.92 | 0.37 | Homogenous |
| | | Region II–2 stations | −0.54 | −0.83 | −0.15 | Homogenous |
| | | Region III–22 stations | −0.06 | 0.34 | 0.19 | Homogenous |
| | | Region III*–20 stations | 0.06 | −1.11 | −1.83 | Homogenous |
| | 4 | Region I–7 stations | −0.55 | −1.11 | −0.99 | Homogenous |
| | | Region I–2 stations | −0.54 | −0.83 | −0.15 | Homogenous |
| | | Region I–2 stations | −0.87 | 1.89 | 1.20 | Homogenous |
| | | Region I–22 stations | 1.02 | 1.59 | 0.96 | Heterogenous |

*After removal of discordant stations Goalpara and Golaghat.

VIKOR while WASPAS is similar for the first three rankings. Hence, the first three similar rankings of all three MCDM methods are considered for selection of optimum cluster and homogeneity analysis. Cluster 2 is the first in rank, 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. Initially, with the region considered as one whole homogeneous region, the heterogeneity test conducted gave $H_1$, $H_2$ and $H_3$ values as 1.11, 1.12 and 0.41 respectively. After evaluating cluster groups 2, 3 and 4, the homogeneity of region 3 was found to be better with $H_1$, $H_2$ and $H_3$ values as 0.33, −0.54 and 0.06 respectively.

### Comparision to heterogeneity measure of regions based on IMD, Pune sub-divisions

As per IMD Pune, the north eastern region of India has been divided into three homogenous meteorological sub-divisions (2a – Arunachal Pradesh, 2b – Assam and Meghalaya, 2c – Nagaland, Manipur, Mizoram and Tripura (NMMT). And the study area comprises two meteorological subdivisions 2b and 2c. 2b comprises the states Assam and Meghalaya, while meteorological subdivision 2c comprises Nagaland, Manipur, Mizoram and Tripura. Thirty-three stations were considered in the study region of 2b and 2c. From heterogeneity analysis, the considered stations in Assam and Meghalaya do not form homogenous regions with values of heterogeneity measure $H_1$ as 1.11. But the rest of the stations considered in the NMMT region form a homogenous region with $H_1$ value as 0.63. Thus, the analysis of the stations is in agreement with region 2c subdivision, but sub-division 2b has stations Shillong, North Lakhimpur, Goalpara and Golaghat with discordances greater than 3. So, the stations were removed to reduce the heterogeneity value to 0.08 for formation of a homogenous rainfall region. The other heterogeneity statistics $H_2$ and $H_3$ also gave negative values, suggesting intercorrelation between the sites in the considered sub-division. As important representative stations from both the states of Assam and Meghalaya, they had to be removed to form a homogenous region, a necessary relook at the homogeneity needs to be explored due to change in scenario of climatic changes, socio-economic development and land use changes.

### Comparison of GA-based clustering, K-Means and Fuzzy C-Means clustering

GA-based clustering gave three homogenous rainfall regions, K-Means gave seven homogenous rainfall regions and Fuzzy C-Means gave six homogenous rainfall regions. From the table it can be seen that the homogeneity of all three regions of GA-based clustering is better in $H_1$ values than the other two clustering algorithms. Negative values of $H_1$ indicates inter correlation among sites in the group.

Comparatively, GA based clustering has only one group with negative value of $H_1$, whereas K-Means and Fuzzy C-Means have 4 and 3 groups with relatively higher negative values of $H_1$. This indicates the grouping by GA based clustering to be superior to the other two algorithms. Also, region III in GA based clustering comprises of 20 stations with a value of $H_1$ as 0.06, while the other algorithms regions are lesser in number with negative values of $H_1$ (Tables 4–8).

### Sensitivity analysis of MCDM methods to change in weights

From Table 9, the rankings of clusters utilizing the cluster validity indexes give the rankings for GA clustering in the order as $1 > 2 > 3 > 5 > 4 > 6$ for both VIKOR and TOPSIS while for WASPAS as $1 > 2 > 3 > 4 > 5 > 6$. To see the effect of the new weights in the order of ranking or if the results change, four changes in the weights of criteria are applied. According to Shannon Entropy method the sum of weights must equal 1 and for a change in any one weight the weights of other criterias also changes. By applying the changes in weights and performing the MCDM rankings for GA based clustering again, it can be seen that, the decision for optimum cluster is unchanged for TOPSIS in all criteria. Sil, XB and SD indices are not affected to any change in weights, and there occurs only three changes in WASPAS and VIKOR with no alteration in the first three rankings. Thus, the first three rankings are not affected to any criteria change in all three methods. Further, there were always two MCDM methods not affected by a change in any criteria and decision could be taken on the similar rankings in the two methods. Thus, the cluster rankings given by the methods are significantly insensitive to weight variation in criteria, and justifies as a robust and accurate methods in decision making of optimum cluster. Among the indices, WASPAS was highly sensitive to PBM, and VIKOR to Dunn and DB indices. Overall, it was found that the TOPSIS approach was stable and the best MCDM, as no weight changes of any indices had an influence on it.

Similarly, the sensitivity of the MCDM analysis of K-Means and Fuzzy C-Means are reported. There were seventeen changes in rankings for K-Means due to change in criteria weight. Only SD index was found to be

**Table 4** | Performance of K-Means clustering for various cluster validation indices

| Cluster number | CS | CH | DB | Dunn | PBM | XB | KL | SD | Sil |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.3869 | 19.0408 | 0.3234 | 1.0017 | 1.1204 | 0.1291 | 10.8527 | 1.9716 | 0.7666 |
| 3 | 0.7505 | 18.6312 | 0.9972 | 0.1045 | 0.8778 | 9.2036 | 8.7734 | 4.8568 | 0.5036 |
| 4 | 0.7034 | 23.1511 | 0.8625 | 0.1370 | 0.6895 | 6.0782 | 6.2905 | 4.8224 | 0.4819 |
| 5 | 0.6808 | 23.4982 | 0.7601 | 0.1643 | 0.5928 | 3.2949 | 5.2242 | 5.0546 | 0.4882 |
| 6 | 0.6137 | 24.4491 | 0.6924 | 0.2101 | 0.4951 | 2.5970 | 4.3379 | 5.0721 | 0.4455 |
| 7 | 0.6552 | 27.4552 | 0.7088 | 0.2520 | 0.5187 | 1.3605 | 3.4158 | 6.5063 | 0.4135 |

**Table 5** | Performance of Fuzzy C-Means clustering for various cluster validation indices

| Cluster number | PC | MPC | XB | Sep | Kwon | Tang |
|---|---|---|---|---|---|---|
| 2 | 0.599 | 0.198 | 1.006 | 1.006 | 34.214 | 34.214 |
| 3 | 0.613 | 0.42 | 0.505 | 0.505 | 32.935 | 24.804 |
| 4 | 0.557 | 0.41 | 0.402 | 0.402 | 38.402 | 21.646 |
| 5 | 0.556 | 0.445 | 0.321 | 0.321 | 42.862 | 18.661 |
| 6 | 0.581 | 0.498 | 0.198 | 0.198 | 41.88 | 13.605 |
| 7 | 0.554 | 0.48 | 0.387 | 0.387 | 108.196 | 28.688 |

**Table 6** | Entropy weights determined using Shannon Entropy method

| Crisp indices | *CH* | *Dunn* | *PBM* | *KL* | *Sil* | *CS* | *DB* | *XB* | *SD* |
|---|---|---|---|---|---|---|---|---|---|
| GA based clustering | 0.076 | 0.241 | 0.128 | 0.120 | 0.076 | 0.065 | 0.144 | 0.069 | 0.081 |
| K-Means | 0.094 | 0.212 | 0.132 | 0.099 | 0.134 | 0.113 | 0.080 | 0.058 | 0.077 |
| Fuzzy indices | ***PC*** | ***MPC*** | ***XB*** | ***Sep*** | ***Kwon*** | ***Tang*** | | | |
| Fuzzy C-Means | 0.159 | 0.161 | 0.169 | 0.171 | 0.170 | 0.170 | | | |

**Table 7** | MCDM ranking for clusters in K-Means and Fuzzy C-Means clustering

| | | TOPSIS | | WASPAS | | VIKOR | |
|---|---|---|---|---|---|---|---|
| Algorithm | Cluster number | Value | Rank | Value | Rank | Value | Rank |
| K-Means | 2 | 0.9343 | 1 | 0.9686 | 1 | 0 | 1 |
| | 3 | 0.1873 | 4 | 0.4022 | 5 | 1 | 6 |
| | 4 | 0.1485 | 6 | 0.402 | 6 | 0.9496 | 5 |
| | 5 | 0.1777 | 5 | 0.407 | 4 | 0.9152 | 4 |
| | 6 | 0.2049 | 3 | 0.4117 | 3 | 0.8669 | 3 |
| | 7 | 0.2378 | 2 | 0.417 | 2 | 0.8325 | 2 |
| Fuzzy C-Means | 2 | 0.355 | 6 | 0.4687 | 6 | 1 | 6 |
| | 3 | 0.6771 | 4 | 0.6652 | 4 | 0.1659 | 2 |
| | 4 | 0.7602 | 3 | 0.6863 | 3 | 0.5932 | 4 |
| | 5 | 0.8371 | 2 | 0.7479 | 2 | 0.5556 | 3 |
| | 6 | 0.941 | 1 | 0.9536 | 1 | 0 | 1 |
| | 7 | 0.5318 | 5 | 0.5829 | 5 | 0.8545 | 5 |

**Table 8** | Comparision of homogenous regions of GA based clustering, K-Means and Fuzzy C-Means

| Algorithm | Regions | No of stations | Stations with discordancy value | Heterogeneity measure | | |
|---|---|---|---|---|---|---|
| | | | | $H_1$ | $H_2$ | $H_3$ |
| GA based clustering | I | 9 | North Lakhimpur(2.09), Choudhowghat(0.99), Bokajan(0.33), Sibsagar(0.69), Margherita(0.98), Dibrugarh(0.25), Jorhat(1.67), Dhollabazar(1.66), Neamatighat(0.34) | 0.33 | 0.92 | 0.37 |
| | II | 2 | Cherrapunjee(1.0), Mawsynram(1.0) | −0.54 | −0.83 | −0.15 |
| | III | 20 | Silchar(0.06), Dholai(1.12), Guwahati(0.48), Batadighat(0.72), Kampur(0.17), Kherunighat(0.26), Imphal(0.63), Gossaigaon(0.83), Kokrajhar(0.98), Tezpur(1.50), Mellabazar(1.44), Aie N. H. Xing(0.11), Dharamtul(0.34), Gharmura(0.98), Beki Road Bridge(0.69), Shillong (2.70), Kohima(2.39), Aizwal(1.14), Agartala(0.60), Kailashahar(2.83) | 0.06 | −1.11 | −1.83 |
| K-Means | I | 8 | North Lakhimpur (1.84), Choudhowghat(0.85), Sibsagar(0.70), Margherita (1.05), Dibrugarh(0.24), Jorhat(1.57), Dhollabazar(1.44), Neamatighat (0.31) | 0.46 | 0.98 | 0.57 |
| | II | 7 | Guwahati(0.69), Batadighat(1.40), Kampur(0.20), Kherunighat(0.62), Bokajan(1.64), Tezpur(1.79), Dharamtul(0.57) | −0.4 | −0.92 | −1.23 |
| | III | 5 | Gharmura(1.30), Shillong(1.02), Kohima(1.31), Imphal(0.27), Aizwal(1.09) | 0.88 | 1.47 | 0.52 |
| | IV | 2 | Cherrapunjee(1.0), Mawsynram(1.0) | −0.54 | −0.83 | −0.15 |
| | V | 2 | Goalpara(1.0), Golaghat(1.0) | −0.54 | −1.1 | −0.45 |
| | VI | 5 | Beki Rd Bridge(0.65), Gossaigaon (0.45), Kokrajhar(1.29), Mellabazar(1.29), Aie NH Xing(1.32) | 0.72 | −0.28 | −1.12 |
| | VII | 4 | Silchar(1.0), Dholai(1.0), Agartala(1.0), Kailashahar(1.0) | −1.87 | −1.42 | −0.76 |
| Fuzzy C-Means | I | 7 | Guwahati(0.44), Batadighat(1.95), Kampur(0.23), Dharamtul(0.06), Kherunighat(1.81), Tezpur(1.90) Golaghat(0.61) | −1.84 | −0.24 | 0.36 |
| | II | 9 | North Lakhimpur(2.09), Choudhowghat(0.99), Bokajan(0.33), Sibsagar(0.69), Margherita(0.98), Dibrugarh(0.25), Jorhat(1.67), Dhollabazar(1.66), Neamatighat(0.34) | 0.33 | 0.92 | 0.37 |
| | III | 5 | Silchar(0.46), Dholai(0.89), Gharmura(1.32), Agartala(1.08), Kailashahar(1.24) | −0.78 | −1.43 | −1 |
| | IV | 2 | Cherrapunjee(1.0), Mawsynram(1.0) | −0.54 | −0.83 | −0.15 |
| | V | 5 | Goalpara (1.32), Shillong (1.05), Kohima(1.31), Imphal(0.19), Aizwal(1.12) | 0.83 | 3.34 | 2.33 |
| | VI | 5 | Beki Rd Bridge(0.65), Gossaigaon(0.45), Kokrajhar(1.29), Mellabazar(1.29), Aie NH Xing(1.32) | 0.8 | −0.22 | −1.1 |

insensitive to the method and did not affect the rankings in the two methods. Dunn, PBM, CH, KL and XB were found to be highly sensitive in K-Means and thus indicates that K-Means clustering was less successful in delineating regions having distinctly separated clusters and within cluster variation. The ranking orders changed with increase in B from $B = 0.5$ to $B = 0.1$ with production of seventeen types of changes. The decision requirement of similarity in rankings of minimum two MCDM methods was satisfied only for Sil, CS, DB and SD indices. With the increase in variation in weights the MCDM ranking for K-Means was found to vary. Thus, the formed regions in K-Means are comparatively weaker from the point of cluster compactness and separation.

For partition of Fuzzy C-Means clustering, there were only three types of ranking change with variation of weights from lowest to highest perturbation. The rankings are better than K-Means and except for Kwon index, the minimum requirement of similarity of any two MCDM methods are satisfied. At 100 percent perturbation the weight changes to its original weights and hence the rankings observed will be the same as the original rankings. The results indicate that the weights determined by the Shannon Entropy method were accurate and suitable for application to all three MCDM methods. The determination of best cluster was satisfactorily achieved by the clustering algorithms with the utilization of MCDM methods. K-Means clustering result is associated with some uncertainty due to the ranking sensitivity for change in weights, and this may be due to non-generation of optimal centroids or solution converging at some local optimum. Whereas, the GA based clustering

outperformed the other two algorithms and the sensitivity analysis did not alter the decision of optimum cluster from rankings produced by the MCDM methods (Tables 10–12).

### Stability of homogenous regions formed using leave one out cross validation test

To study the effect of stability of the clusters formed by clustering algorithm the leave one out cross validation (LOOCV) test was conducted on the final regions produced be the three clustering algorithms. The three regions of genetic algorithm were identified and the stations in each region were dropped each time to see the effect on the heterogeneity measure $H_1$. Region 1 comprised 9 stations and the $H_1$ value obtained each time after removal of one station is plotted in Figure 3. Region 3 comprises 20 stations and the $H_1$ values are plotted in the same figure. All values of $H_1$ obtained in the leave-one-out test gave $H_1$ values less less than 1, suggesting the region to be adequately homogenous as there was no occurrence of any $H_1$ value greater than 1. Region 2 comprises of only two stations and hence was strictly homogenous, and the heterogeneity value in the plot was the same value as that of the homogeneity of the region. The mean was also taken as the same value.

While for K-Means, the LOOCV procedure also gave no values of regions greater than 1 or was not heterogenous. Hence, the 7 cluster regions formed were adequate homogenous regions. In two regions where the number of stations in the region were two each, the LOOCV could not be carried out and the heterogeneity value for the regions is plotted in the plot with the mean. There were occurrences three times for region 3 and two times for region 6 exceeding the value 1 but was less than 2; that is, it was not definitely heterogenous. Also, four regions means were found to be having negative $H_1$ values in the test suggesting inter-site dependence among the sites in those regions.

For Fuzzy C-Means, the regions 2 and 3 were seen to have $H_1$ values greater than 1 with the removal of 3 stations in region 2 and two stations in region 3. Three regions were found to have negative $H_1$ values suggesting inter-dependence of sites in the region. Overall, the regions produced by GA-based clustering gave better results with the only one region with inter-site station dependence, and the results in K-Means and Fuzzy C-Means was possibly heterogenous for 5 stations but was not fully heterogenous.

From the table, the maximum and minimum of annual extreme rain in each region were obtained. Region 2, region 4 of K-Means and region 4 of Fuzzy C-Means were having similar values because the region comprises the similar stations Cherrapunjee and Mawsynram. Region 1 of GA-based clustering and region 3 of Fuzzy C-Means are similar. Region 5 of K-Means gave the highest bias among all regions and has the highest standard deviation with a value of 54.95. It comprises two stations, Goalpara and Golaghat, and are having high values of skewness and kurtosis among all stations. Comparatively, the standard deviation of the annual extreme of stations for the clustering algorithms, the GA-based clustering gave relatively better regions. From the values reported for average skewness of each group for each clustering algorithm, the skewness and kurtosis are better for GA-based clustering followed by Fuzzy C-Means and K-Means. Hence, the grouping of stations by genetic algorithm were more appropriate and better among the three algorithms. The information transfer index of entropy of each station in the groups is calculated and the average is reported. The information transfer in the regions among the stations is highest among GA cluster regions. Lowest information transfer is found in region 5 in K-Means regions, with a value of 0.074. For the fuzzy C-Means regions, the information transfer between stations is slightly lower than GA-based regions, while it is relatively better than K-Means regions. Also the region 3 of GA based clustering has the highest number of stations but the stations are well grouped and have comparatively higher average information transfer compared to some regions of K-Means and Fuzzy C-Means.

### Comparison of accuracy of regional growth curves using Monte Carlo simulation

Regional relative bias explains the tendency for quantile estimates to be uniformly too high or too low across the whole region. The probability distributions identified for each region are more appropriate to true distribution if the bias is lesser (Atiem & Harmancioglu 2006). From Figure 4, GA-based clustering is seen to have the minimum bias. Bias obtained by regions of IMD meteorological sub-divisions are seen to have more bias than GA clustering. Also, for higher return periods, the bias of all regions by GA-based clustering seems to disperse least among the algorithms, thus signifying that the regions formed are relatively more homogenous. Thus, the GA-based homogenous regions have less bias in quantile estimates in all regions compared to IMD based sub-divisions in the region. Region 5 in K-Means is sensitive with the increase in return period giving negative values of bias, indicating that the region seems to deviate from the true quantile estimates with lower estimates, and the uncertainty of estimates in the region is more.

**Table 9** | Sensitivity of rankings for GA based clustering to change in criteria weight for various values of unitary variation ratio $\beta$

| | Rankings for criterias with $\beta = 0.5$ (lowest variation) | | | | | | | | | Rankings for criterias with $\beta = 0.2$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MCDM method** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | **DB** | **XB** | **SD** | **CH** | **Dunn** | **PBM** | **KL** | **Sil** | **CS** | **DB** | **XB** | **SD** |
| WASPAS | NW | NW | C1 | NW | NW | NW | NW | NW | NW | NW | NW | C1 | C1 | NW | NW | NW | NW | NW |
| VIKOR | NV | C2 | NV | NV | NV | NV | C3 | NV | NV | NV | C2 | NV | NV | NV | NV | C3 | NV | NV |
| TOPSIS | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT |
| | **Rankings for criterias with $\beta = 0.1$** | | | | | | | | | **Rankings for criterias with $\beta = 0.01$ (highest variation)** | | | | | | | | |
| **MCDM method** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | ***DB*** | ***XB*** | ***SD*** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | ***DB*** | ***XB*** | ***SD*** |
| WASPAS | C1 | NW | C1 | C1 | NW | NW | NW | NW | NW | C1 | NW | C1 | C1 | NW | NW | NW | NW | NW |
| VIKOR | NV | C2 | NV | NV | NV | C3 | C3 | NV | NV | NV | C2 | NV | NV | NV | C3 | C3 | NV | NV |
| TOPSIS | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT |

*Rankings*: Original WASPAS (NW): 1-2-3-4-5-6, Original VIKOR (NV) : 1-2-3-5-4-6, Original TOPSIS (NT) : 1-2-3-5-4-6, Change 1 (C1) : 1-2-3-5-4-6, Change 2 (C2) : 1-3-5-4-2-6, Change 3 (C3) : 1-2-3-4-5-6.

**Table 10** | Sensitivity of rankings for K-Means to change in criteria weight for various values of unitary variation ratio $\beta_p$

| | Rankings for criterias with $\beta = 0.5$ (lowest variation) | | | | | | | | | Rankings for criterias with $\beta = 0.2$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MCDM method** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | ***DB*** | ***XB*** | ***SD*** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | ***DB*** | ***XB*** | ***SD*** |
| WASPAS | C1 | C2 | C3 | C3 | C3 | NW | NW | NW | NW | C8 | C9 | C3 | C3 | C3 | C1 | C1 | C10 | NW |
| VIKOR | NV | C4 | NV | NV | NV | NV | NV | NV | NV | NV | C4 | NV | NV | NV | NV | NV | NV | NV |
| TOPSIS | NT | C5 | C6 | C6 | NT | NT | NT | C7 | NT | NT | C11 | C6 | C6 | NT | NT | NT | C12 | NT |
| | **Rankings for criterias with $\beta = 0.1$** | | | | | | | | | **Rankings for criterias with $\beta = 0.01$ (highest variation)** | | | | | | | | |
| **MCDM method** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | ***DB*** | ***XB*** | ***SD*** | ***CH*** | ***Dunn*** | ***PBM*** | ***KL*** | ***Sil*** | ***CS*** | ***DB*** | ***XB*** | ***SD*** |
| WASPAS | C13 | C9 | C3 | C3 | C3 | C5 | C1 | C14 | NW | C13 | C9 | C3 | C3 | C3 | C4 | C17 | C17 | NW |
| VIKOR | C15 | C4 | NV | NV | NV | NV | NV | NV | NV | C15 | C4 | C6 | NV | NV | NV | NV | NV | NV |
| TOPSIS | NT | C16 | C6 | C6 | NT | NT | NT | C12 | NT | NT | C16 | C3 | C6 | NT | NT | NT | C12 | NT |

*Rankings*: Original WASPAS (NW) : 1-5-6-4-3-2, Original VIKOR (NV) : 1-6-5-4-3-2, Original TOPSIS (NT) : 1-4-6-5-3-2, Change 1 (C1) : 1-4-6-5-3-2, Change 2 (C2) : 1-2-4-3-5-6, Change 3 (C3) : 1-6-5-4-3-2, Change 4 (C4) : 1- 4-2-3-5-6, Change 5 (C5) : 1-3-6-4-5-2, Change 6 (C6) : 1-5-6-4-3-2, Change 7 (C7) : 1-3-6-5-4-2, Change 8 (C8) : 1-2-6-5-3-4, Change 9 (C9) : 1-2-3-4-5-6, Change 10 (C10) : 1-2-5-6-3-4, Change 11 (C11) : 1-2-6-5-4-3, Change 12 (C12) : 1-2-5-6-4-3, Change 13 (C13) : 1-2-6-4-3-5, Change 14 (C14) : 1-2-4-6-3-5, Change 15 (C15) : 1-6-5-4-2-3, Change 16 (C16) : 1-2-6-4-5-3, Change 17 (C17) : 1-2-4-5-3-6.

**Table 11** | Sensitivity of rankings for Fuzzy C-Means to change in criteria weight for various values of unitary variation ratio $\beta_p$

| | Rankings for criterias with β = 0.5 (lowest variation) | | | | | | Rankings for criterias with β = 0.2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MCDM method** | ***PC*** | ***MPC*** | ***XB*** | ***Sep*** | **Kwon** | ***Tang*** | ***PC*** | ***MPC*** | ***XB*** | ***Sep*** | ***Kwon*** | ***Tang*** |
| WASPAS | NW | NW | NW | NW | NW | NW | NW | NW | NW | NW | C2 | NW |
| VIKOR | C1 | NV | NV | NV | NV | NV | C1 | NV | NV | NV | NV | C3 |
| TOPSIS | NT | NT | NT | NT | NT | NT | NT | NT | NT | NT | C2 | NT |
| | **Rankings for criterias with β = 0.1** | | | | | | **Rankings for criterias with β = 0.01 (highest variation)** | | | | | |
| **MCDM method** | ***PC*** | ***MPC*** | ***XB*** | ***Sep*** | ***Kwon*** | ***Tang*** | ***PC*** | ***MPC*** | ***XB*** | ***Sep*** | **Kwon** | **Tang** |
| WASPAS | NW | NW | NW | NW | C2 | NW | NW | NW | NW | NW | C2 | NW |
| VIKOR | C1 | NV | NV | NV | NV | C3 | C1 | NV | NV | NV | NV | C3 |
| TOPSIS | NT | NT | NT | NT | C2 | NT | NT | NT | NT | NT | C2 | NT |

*Rankings*: Original WASPAS (NW): 6-4-3-2-1-5, Original VIKOR (NV): 6-2-4-3-1-5, Original TOPSIS (NT) : 6-4-3-2-1-5, Change 1 (C1) : 6-4-3-2-1-5, Change 2 (C2) : 6-5-3-2-1-4, Change 3 (C3) : 6-1-4-3-2-5.

**Table 12** | Comparision of statistical parameters and average information transfer index of regions

| Algorithm | Regions | Annual extreme rain average | Standard deviation | Average Skewness | Average Kurtosis | Average Information Transfer Index |
|---|---|---|---|---|---|---|
| GA based clustering | Region 1 | 114.69 | 23.9 | 0.08 | 0.09 | 0.185 |
| | Region 2 | 561.63 | 13.77 | 0.03 | 0.10 | 0.255 |
| | Region 3 | 130.38 | 43.54 | 0.16 | 0.14 | 0.188 |
| K-Means | Region 1 | 117.48 | 23.93 | 0.09 | 0.09 | 0.188 |
| | Region 2 | 95.18 | 8.94 | 0.13 | 0.11 | 0.194 |
| | Region 3 | 106.40 | 29.14 | 0.14 | 0.12 | 0.172 |
| | Region 4 | 561.63 | 13.77 | 0.03 | 0.10 | 0.255 |
| | Region 5 | 134.15 | 54.95 | 0.47 | 0.41 | 0.074 |
| | Region 6 | 189.64 | 34.66 | 0.11 | 0.08 | 0.177 |
| | Region 7 | 136.46 | 9.69 | 0.13 | 0.16 | 0.172 |
| Fuzzy C-Means | Region 1 | 95.60 | 8.86 | 0.19 | 0.17 | 0.164 |
| | Region 2 | 114.69 | 23.90 | 0.08 | 0.09 | 0.185 |
| | Region 3 | 134.66 | 9.32 | 0.15 | 0.16 | 0.168 |
| | Region 4 | 561.63 | 13.77 | 0.03 | 0.10 | 0.255 |
| | Region 5 | 115.52 | 41.76 | 0.19 | 0.15 | 0.185 |
| | Region 6 | 189.64 | 34.66 | 0.11 | 0.08 | 0.177 |

Overall quantile deviation from true quantiles for sites in a region, measured by regional average relative RMSE, has the most weight to determine whether an estimation procedure is better than another (Atiem & Harmancioglu 2006) and is presented in Figure 5. Region 1 comprises Assam and Meghalaya with four stations removed while Region 2 in IMD comprises five stations of the NMMT subdivision. Region 2 was less accurate in terms of regional average RMSE. Except for region 2 of GA-based clustering, comprising Cherrapunjee and Mawsynram, the RMSE values were close to zero and were relatively low, thus the distributions could appropriately define the extreme rainfall behaviour of the stations and quantiles. Whereas the estimated quantiles in K-Means and Fuzzy C-Means clustered regions were more dispersed thereby suggesting the quantiles estimated for the regions were relatively more uncertain.

## SUMMARY AND CONCLUSIONS

The regional frequency study involves a wide variety of grouping techniques that utilize similarity in characteristic attributes of stations. Numerous grouping techniques and algorithms are available, but the application with genetic algorithms is less documented. Also, it is often found that the clustered findings are not always conclusive, and require further subjective adjustments to make them statistically homogeneous. Further, when the number of clusters is unknown and the selection criteria are numerous, the choice is more difficult, and the solutions

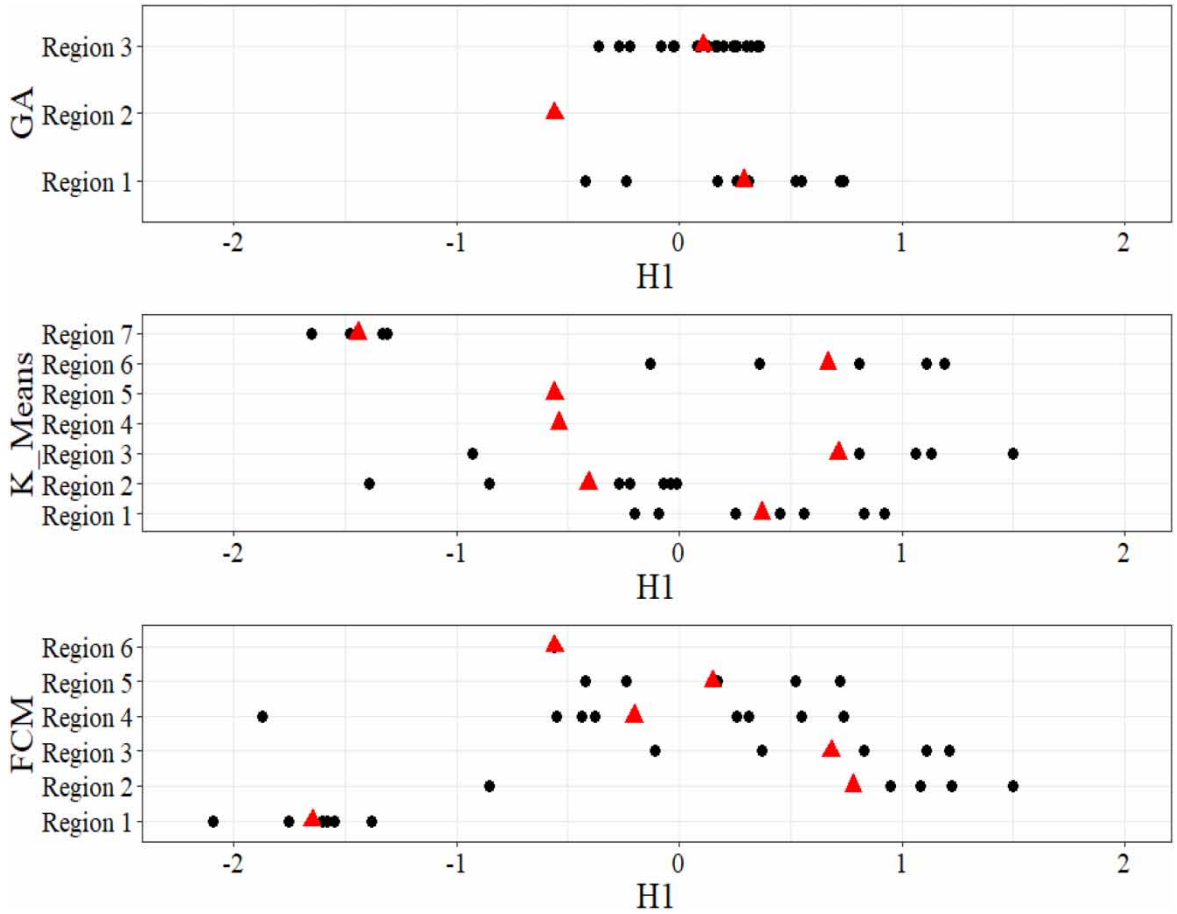



**Figure 3** | LOOCV test performance for regions of all three algorithms.

provided by MCDM rankings are more convenient. With the application of three separate integrated MCDM, results indicated GA-based clustering was seen to outperform the other two algorithms and provided stable and better results. The homogeneity measure, $H_1$, of all GA-based cluster regions was found to be comparatively better, while K-Means and Fuzzy C-Means provided several regions with significantly larger negative $H_1$ values, indicating the presence of inter-site dependency in regions.

Sensitivity analysis was conducted to study the robustness of the MCDM rankings to selected cluster validation measures. GA-based clustering TOPSIS rankings were found to be stable as they were not affected by change in any criteria weight. WASPAS and VIKOR rankings were seen to have slight change but there remained always two MCDM methods that were not affected by the change in criteria, and decision could be taken on the similar rankings of the methods. The K-Means rankings were found to vary considerably with increase in variation in weights, thereby indicating K-Means formed regions were comparatively weaker in cluster compactness and separation. Fuzzy C-Means rankings were better than K-Means, and there was very little change in rankings with the requirement of similarity of any two MCDM methods fulfilled. Shannon entropy weights were found to be accurate for application to determine weights in all three MCDM methods, and determination of best cluster was satisfactorily achieved with the utilization of MCDM methods. Overall, the sensitivity analysis on criteria weights suggested GA-based clustering outperformed the other two algorithms, indicating robust grouping of regions. The TOPSIS method was found to be the best MCDM method for all three algorithms with lesser sensitivity to weight changes of validation indices.

Stability of the clusters formed by the clustering algorithms was evaluated using the LOOCV test to see the effect on the change in heterogeneity measure $H_1$. Results obtained found two GA-based cluster regions with $H_1$ values less than 1 each time for all regions, suggesting the regions were adequately homogenous. One common region in GA-based clustering and Fuzzy C-Means, two in K-Means, comprised only two stations and hence was a strictly formed group and the LOOCV test could not be conducted for the regions. In K-Means, there were few occurrences exceeding the value 1 for region 3 and region 6 but was less than 2; that is, it was

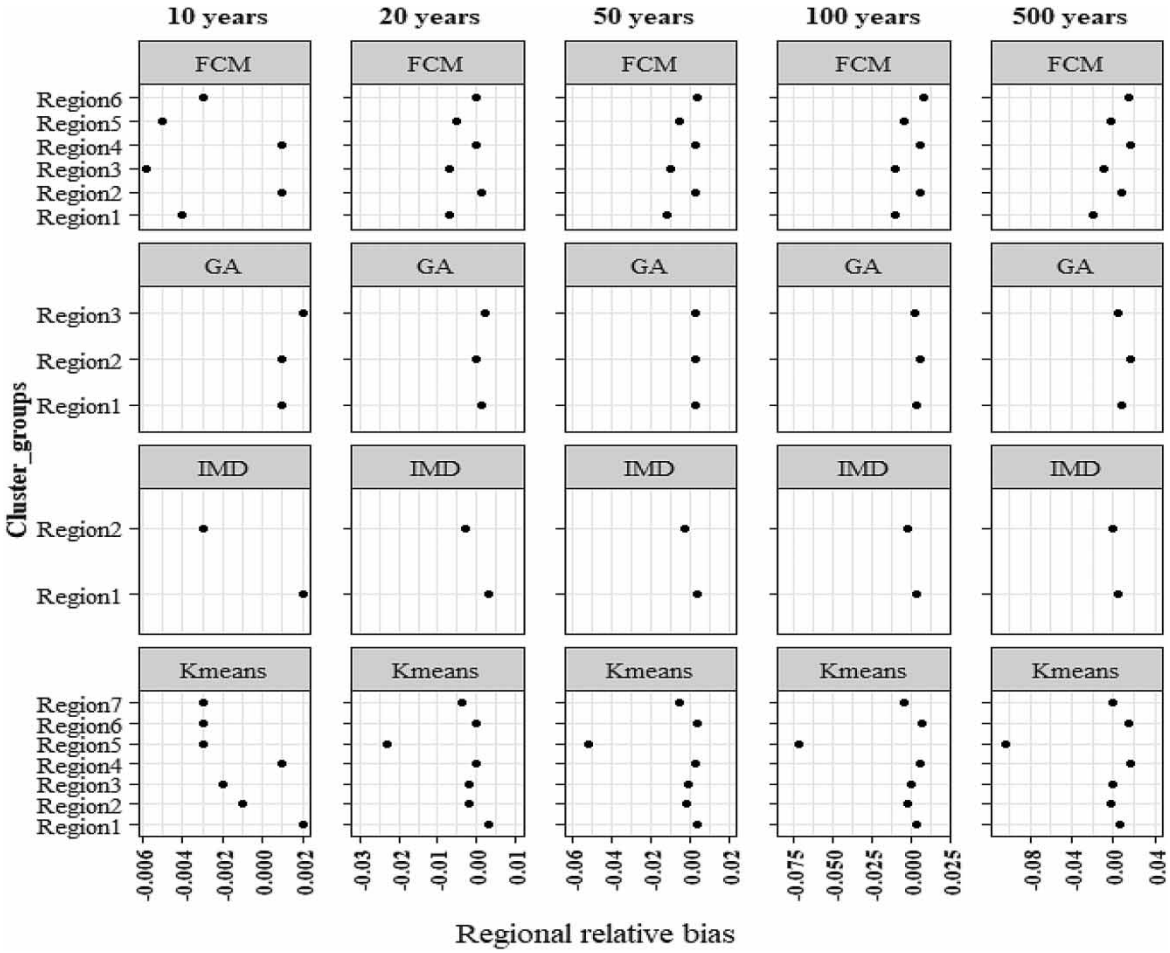



**Figure 4** | Regional relative bias of regional growth curve in regions obtained by GA-based clustering, K-Means, Fuzzy C-Means and IMD Pune.

not 'definitely heterogenous'. Whereas for Fuzzy C-Means, regions 2 and 3 gave few occurrences of $H_1$ values greater than 1, but were not 'definitely heterogenous'. Overall, it can be deduced that the regions produced by GA-based clustering produced better robustness and homogeneity.

Region 5 of K-Means gave the highest bias among all regions and had the highest standard deviation with a value of 54.95. Comparing the standard deviation of the annual extreme of stations for the clustering algorithms, the GA-based clustering gave relatively better regions. The average skewness of each group for each clustering algorithm and kurtosis are better for GA-based clustering followed by Fuzzy C-Means and K-Means. Hence, again the grouping of stations by genetic algorithm was more appropriate and better among the three algorithms. The information transfer index of entropy of each station in the groups was calculated and the region average reported. The information transfer among the stations produced by GA cluster regions was comparatively higher than the other two algorithms. Lowest information transfer is found in region 5 in K-Means regions with a value of 0.074. Information transfer index among Fuzzy C-Means regions are relatively better than K-Means, but is lower than GA-based regions.

Regional growth curve in GA-based regions is seen to have the minimum relative bias compared to IMD meteorological sub-divisions, K-Means and Fuzzy C-Means. At higher return periods, the bias of all regions by GA-based clustering seems to disperse the least among the algorithms, thus signifying the regions to be relatively more homogenous. Most of the regions of K-Means were sensitive with increase in return period giving negative values of bias, thus indicating the regions deviate from the true quantile estimates, increasing the uncertainty in quantile estimates. Though, IMD region 1 was less deviating, region 2 was considerably inaccurate in terms of regional average RMSE. Except for region 2 of GA based clustering comprising of Cherrapunjee and Mawsynram, the regional average RMSE values were close to zero and were relatively low, thus the distributions

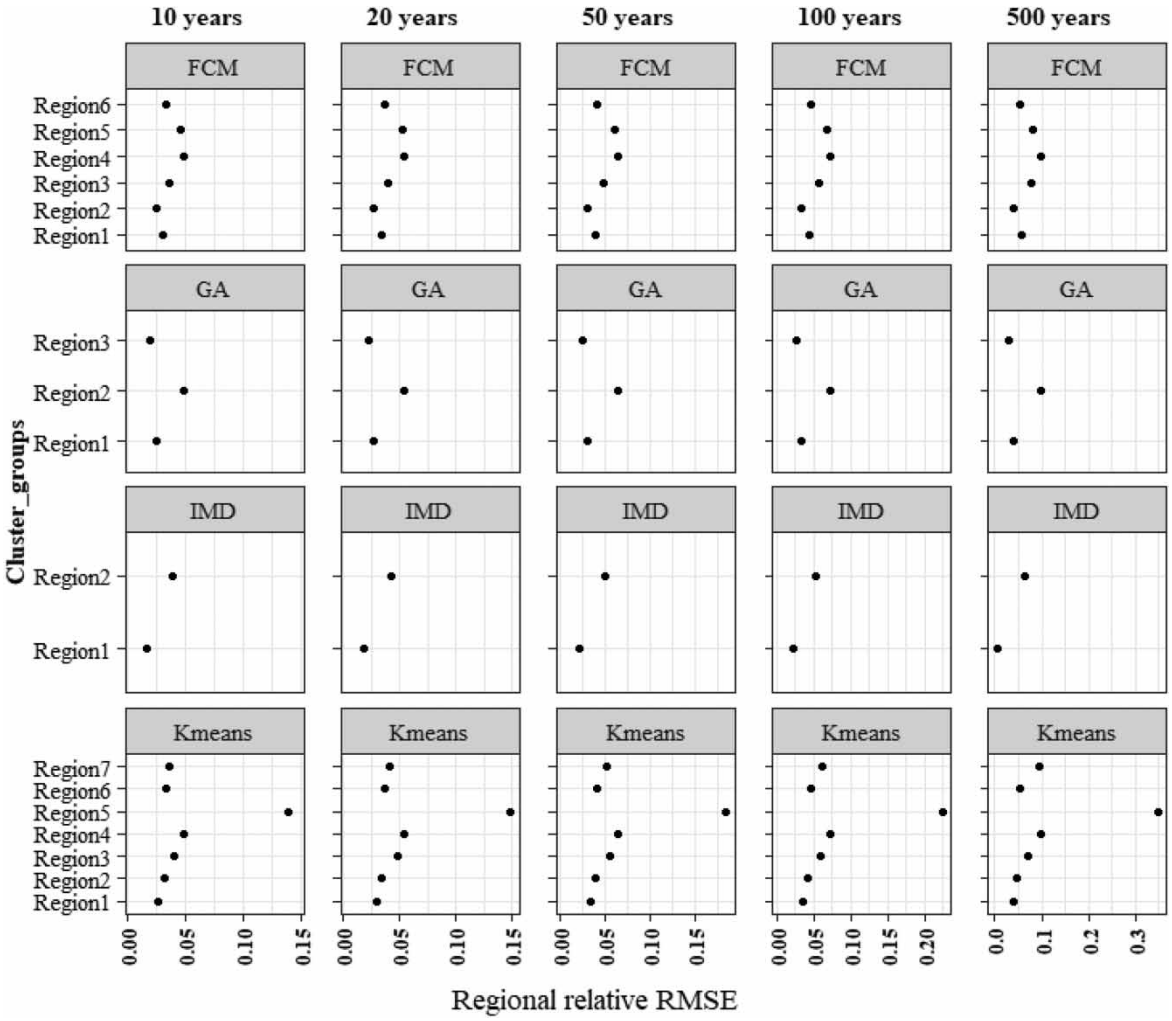



**Figure 5** | Regional relative RMSE of regional growth curve in regions obtained by GA-based clustering, K-Means, Fuzzy C-Means and IMD Pune.

appropriately define the extreme rainfall behaviour of the stations and quantiles. Whereas, the estimated quantiles in K-Means and Fuzzy C-Means clustered regions were more dispersed, thereby suggesting the quantiles estimated were relatively more uncertain. The results obtained from this study thus shows the efficiency of the MCDM ensemble approach for clustering algorithms in choosing appropriate homogenous cluster regions, and can be applied to various other algorithms. The uncertainty in achieving homogenous regions in frequency analysis can be simplified by the integrated approach and the lack of expertise in choosing cluster validation measure can be satisfactorily overcome.

## ACKNOWLEDGEMENTS

We are thankful to Regional Meteorological Centre Guwahati, India for the supply of data for the study area, and the study would not have been possible without the availability of free software package 'lmomRFA' and 'MCDM' in R environment.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

## REFERENCES


Agustín-Blas, L.-E., Salcedo-Sanz, S., Jiménez-Fernández, S., Carro-Calvo, L., DelSer, J. & Portilla-Figueras, J. A. 2012 A new grouping genetic algorithm for clustering problems. *Expert Systems with Applications* **39**, 9695–9703.

Atiem, A. & Harmancioglu, N.-B. 2006 Assessment of regional floods using L-moments approach: the case of the River Nile. *Water Resources Management* **20**, 723–747.

Bandyopadhyay, S. & Maulik, U. 2002 Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition* **35**, 1197–1208.

Basu, B. & Srinivas, V.-V. 2014 Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resources Research* **50**(4), 3295–3316.

Bensaid, A.-M., Hall, L.-O., Bezdek, J.-C., Clarke, L.-P., Silbiger, M.-L., Arrington, J.-A. & Murtagh, R. F. 1996 Validity-guided (Re) clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems* **4**, 112–123.

Bezdek, J.-C. 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, NY.

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

Calinsky, R. & Harabasz, J. 1974 A dendrite method for cluster analysis. *Communications in Statistics* 1–27.

Campello, R.-J.-G.-B. & Hruschka, E.-R. 2006 A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems* **157**(21), 2858–2875.

Davies, D.-L. & Bouldin, D.-W. 1979 A clustering separation measure. *IEEE Transactions on PAMI* **1**(1979), 224–227.

Gan, G., Ma, C. & Wu, J. 2007 *Data Clustering: Theory, Algorithms, and Applications (ASA-SIAM Series on Statistics and Applied Probability) (Society for Industrial and Applied Mathematics, Philadelphia, USA)*.

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.

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.

Hruschka, E.-R., deCastro, L.-N. & Campello, R.-J.-G.-B. 2004 Evolutionary algorithms for clustering gene-expression data. In: *Fourth IEEE International Conference on Data Mining (ICDM '04)*, 2004. p. 403–406.

Hwang, C.-L. & Yoon, K. 1981 *Multiple Attributes Decision Making Methods and Applications*. Springer, Berlin Heidelberg. 1981.

Krzanowski, W. & Lai, Y. 1988 A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* **44**, 23–34.

Kuo, R. J., Syu, Y. J., Chen, Z.-Y. & Tien, F. C. 2012 Integration of particle swarm optimization and genetic algorithm for dynamic clustering. *Information Science* **195**(2012), 124–140.

Li, P., Qian, H., Wu, J. & Chen, J. 2013 Sensitivity analysis of TOPSIS method in water quality assessment: I. Sensitivity to the parameter weights. *Environmental Monitoring and Assessment* **185**, 2453–2461.

Lin, H.-J., Yang, F.-W. & Kao, Y.-T. 2005 An efficient GA-based clustering technique. *Tamkang Journal of Scientific and Engineering Research* **8**, 113–122.

Liu, Y., Wu, X. & Shen, Y. 2011 Automatic clustering using genetic algorithms. *Applied Mathematics and Computation* **218**, 1267–1279.

Mareschal, B., Brans, J.-P. & Vincke, P. 1984 PROMETHEE: a new family of outranking methods in multicriteria analysis. In: *ULB Institutional Repository*. ULB–Universite Libre de Bruxelles, Brussels

Maulik, U. & Bandyopadhyay, S. 2002 Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(12), 1650–1654.

Opricovic, S. & Tzeng, G.-H. 2004 The compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research* **156**(2), 445–455.

Ozturk, C., Hancer, E. & Karaboga, D. 2015 Dynamic clustering with improved binary artificial bee colony algorithm. *Applied Soft Computing* **28**(2015), 69–80.

Peng, Y., Zhang, Y., Kou, G. & Shi, Y. 2012 A multicriteria decision making approach for estimating the number of clusters in a data set. *PLoS ONE* **7**(7), e41713. doi:10.1371/journal.pone.0041713.

Rao, A.-R. & Hamed, K.-H. 2000 *Flood Frequency Analysis*. CRC Press, Boca Raton, FL, USA.

Ridolfi, F., Rianna, E., Trani, G., Alfonso, L., Baldassarre, G.-D., Napolitano, G. & Russo, F. 2016 A new methodology to define homogeneous regions through an entropy based clustering method. *Advances in Water Resources* **96**, 237–250.

Saaty, T. 1988 What is the analytic hierarchy process? In: *Mathematical Models for Decision Support*, Vol. 48. (Mitra, G., Greenberg, H., Lootsma, F., Rijkaert, M. & Zimmermann, H., eds). Springer, Berlin, pp. 109–121.

Tang, Y., Sun, F. & Sun, Z. 2005 Improved validation index for fuzzy clustering. In: *Proceedings of the 2005 American Control Conference*. IEEE Computer Society, pp. 1120–1125.

Wang, W. & Zhang, Y. 2007 On fuzzy cluster validity indices. *Fuzzy Sets and Systems* **158**, 2095–2117.

Zaki, M.-J. & Wagner Meira, J. 2014 *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, Cambridge, UK, 2014.

Zavadskas, E.-K., Turskis, Z., Antucheviciene, J. & Zakarevicius, A. 2012 Optimization of weighted aggregated sum product assessment. *Elektronika ir elektrotechnika* **122**(6), 3–6.


First received 25 May 2021; accepted in revised form 23 August 2021. Available online 2 September 2021