# GCM Based Fuzzy Clustering to Identify Homogeneous Climatic Regions of North-East India

Arup K. Sarma, Jayshree Hazarika

***Abstract*—**The North-eastern part of India, which receives heavier rainfall than other parts of the subcontinent, is of great concern now-a-days with regard to climate change. High intensity rainfall for short duration and longer dry spell, occurring due to impact of climate change, affects river morphology too. In the present study, an attempt is made to delineate the North-eastern region of India into some homogeneous clusters based on the Fuzzy Clustering concept and to compare the resulting clusters obtained by using conventional methods and nonconventional methods of clustering. The concept of clustering is adapted in view of the fact that, impact of climate change can be studied in a homogeneous region without much variation, which can be helpful in studies related to water resources planning and management. 10 IMD (Indian Meteorological Department) stations, situated in various regions of the North-east, have been selected for making the clusters. The results of the Fuzzy C-Means (FCM) analysis show different clustering patterns for different conditions. From the analysis and comparison it can be concluded that nonconventional method of using GCM data is somehow giving better results than the others. However, further analysis can be done by taking daily data instead of monthly means to reduce the effect of standardization.

***Keywords*—**Climate change, conventional and nonconventional methods of clustering, FCM analysis, homogeneous regions.

## I. Introduction

In today’s era, the most important issue that humanity has ever faced is considered to be climate change. For the past few years, climate change is gaining attention of the climatologists as well as environmental and water-resources engineers because of its expected impacts on the environment and on human beings. The global impact of climate change and its potential effects on various fields of sustainable development has threatened the whole world. In vast countries like India, where climate conditions differ from place to place, climate change is of major concern, because its economic performance and social progress are dependent on rainfall. Especially the north-eastern part of India is of great concern now-a-days. This is the part on India which receives heavier rainfall than other parts of the subcontinent. High intensity rainfall for short duration and longer dry spells are the major problems due to impact of climate change. It also affects river morphology due to flood or drought. Since water is one of the major key area being affected by climate change, therefore planning and management of water resources related issues should definitely include the impacts of climate change.

Many researches and projects on climate change are already going on. Mehrotra & Mehrotra [1] have studied the climate change and hydrology with emphasis on the Indian subcontinent. They mentioned that, the major impacts of climate change in India would be on the hydrology, water resources and agriculture of the country. Wilby et al. [2] have investigated the empirical relation between climatic variables and local variables. Wilby et al. [3] made a comparison of different Statistical downscaling methods using GCM output. Zhang [4] developed a simple method for statistical downscaling of GCM output to predict soil erosion and crop production. Gosain et al. [5] have developed a hydrologic modelling for various river basins in India considering climate change effect. In IIT Guwahati also, many research projects are going on related to climate change. Studies have been done on simulating the impact of climate change with the use of downscaling methods, on the precipitation characteristics and stream flow behavior of Dhansiri River, a southern tributary of Brahmaputra basin.

After observing thoroughly the research works done by different people, it came into notice that more emphasis is given on various downscaling techniques which have been used to observe the impacts of climate change on water resources. Use of Global Climate model (GCM) is the most popular way of doing an analysis. However, there are a number of downscaling techniques which can be used. Furthermore, there are various types of climate models available for different emission scenarios, provided by various organizations and climate centers. The selection of appropriate model as well as downscaling technique is not very clear from any of those works.

From the previous studies, it came into notice that different downscaling models give different results for different regions. Hence it is not advisable and also is a tedious process to use different models for different regions, especially when the study area is large. Hence, if it becomes possible to delineate a region into some homogeneous clusters, in which same downscaling technique can be applied without having much variation, then it will be very useful in further studies related to water resources planning and management. Keeping this in mind, objective of the study is decided as, (i) to delineate the region into homogeneous clusters based on the Fuzzy Clustering concept, and (ii) to compare the resulting clusters obtained by using conventional methods and nonconventional methods.

Studies have also been done on research works related to regionalization or homogeneous clustering and on fuzzy

A.K.Sarma is with the Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India (phone: 91 361 2582409; fax: 91 258 2440; e-mail: aks@iitg.ernet.in).

J.Hazarika is with the Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India (e-mail: jayshree@iitg.ernet.in).

clustering techniques. Fill & Stedinger [6] performed homogeneity tests based upon Gumbel distribution and had done a critical appraisal of Dalrymple's test. Habib et al. [7] has discussed an innovative method for hydrological delineation of homogeneous regions in Tunisia, based on the shape of the empirical Cumulative Distribution Function (CDF) and similarities of physiographic and climatic characteristics. Satyanarayana & Srinivas [8] has done an experiment on regionalization of precipitation in data sparse areas based on fuzzy clustering approach.

The remainder of this paper is organized as follows: The proposed methodology, with special reference to Fuzzy-c-means (FCM) algorithm technique, for regionalization or homogeneous clustering is presented in Section II. Description of the study area and details concerning the data used for the study are provided in Section III. Results from the application of the proposed method in the north-eastern India are presented and discussed in Section IV. Results are compared for conventional and nonconventional data. Finally, the conclusion and the future study required are presented in Section V.

## II. Methodology

Clustering techniques are mostly unsupervised methods that can be used to organize data into groups based on similarities among the individual data items. Most clustering algorithms do not rely on assumptions common to conventional statistical methods, and therefore they are useful in situations where little prior knowledge exists. Clustering techniques are used in a wide variety of applications, including classification, image processing, pattern recognition, modelling and identification. This section summarizes the clustering and its types, along with FCM algorithm.

### *A. General Overview of Clustering Methods*

Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy (soft) or crisp (hard). Thus, the clustering methods are mainly of following types:

1. Hard Clustering

Methods, which are based on classical set theory, and require that an object either does or does not belong to a cluster, are referred to as hard clustering methods This means, partitioning the data into a specified number of mutually exclusive subsets. The discrete nature of the hard partitioning causes difficulties with algorithms based on analytic functionals, since these functionals are not differentiable.

The objective of clustering is to partition the data set $\boldsymbol{Z}$ into $c$ clusters (groups, classes). Generally, $c$ is assumed to be known, based on prior knowledge. Using classical sets, a hard partition of $\boldsymbol{Z}$ can be defined as a family of subsets $\{A_i | 1 \leq i \leq c\} \subset P(\boldsymbol{Z})1$ with the following properties:

$$\cup_{i=1}^{c} A_i = \boldsymbol{Z}, \tag{1}$$

$$A_i \cap A_j = \emptyset, 1 \leq i \neq j \leq c, \tag{2}$$

$$\emptyset \subset A_i \subset \boldsymbol{Z}, 1 \leq i \leq c. \tag{3}$$

In terms of membership (characteristic) functions, a partition can be conveniently represented by the partition matrix $\boldsymbol{U} = [\mu_{ik}]_{c \times N}$. The $i^{th}$ row of this matrix contains values of the membership function $\mu_i$ of the $i^{th}$ subset $A_i$ of $\boldsymbol{Z}$. The elements of $\boldsymbol{U}$ must satisfy the following conditions:

$$\mu_{ik} \in \{0,1\}, 1 \leq i \leq c, 1 \leq k \leq N, \tag{4}$$

$$\textstyle\sum_{i=1}^{c} \mu_{ik} = 1, 1 \leq k \leq N, \tag{5}$$

$$0 < \textstyle\sum_{k=1}^{N} \mu_{ik} < N, 1 \leq i \leq c. \tag{6}$$

2. Fuzzy Clustering

These methods allow the objects to belong to several clusters simultaneously, with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership.

Generalization of the hard partition to the fuzzy case follows directly by allowing $\mu_{ik}$ to attain real values in [0, 1]. Conditions for a fuzzy partition matrix are given by:

$$\mu_{ik} \in [0,1], 1 \leq i \leq c, 1 \leq k \leq N, \tag{7}$$

$$\textstyle\sum_{i=1}^{c} \mu_{ik} = 1, 1 \leq k \leq N, \tag{8}$$

$$0 < \textstyle\sum_{k=1}^{N} \mu_{ik} < N, 1 \leq i \leq c. \tag{9}$$

The $i^{th}$ row of the fuzzy partition matrix $\boldsymbol{U}$ contains values of the $i^{th}$ membership function of the fuzzy subset $A_i$ of $\boldsymbol{Z}$. Second equation constrains the sum of each column to 1, and thus the total membership of each $z_k$ in $\boldsymbol{Z}$ equals one.

3. Possibilistic Partition

A more general form of fuzzy partition, the possibilistic partition, can be obtained by relaxing the second constraint i.e. the total membership should be unity. This constraint, however, cannot be completely removed, in order to ensure that each point is assigned to at least one of the fuzzy subsets with a membership greater than zero. Thus the 2nd equation can be replaced by a less restrictive constraint $\forall k, \exists_i, \mu_{ik} > 0$. The conditions for a possibilistic fuzzy partition matrix are:

$$\mu_{ik} \in [0,1], 1 \leq i \leq c, 1 \leq k \leq N, \tag{10}$$

$$\exists_i, \mu_{ik} > 0, \forall k, \tag{11}$$

$$0 < \textstyle\sum_{k=1}^{N} \mu_{ik} < N, 1 \leq i \leq c. \tag{12}$$

### *B. Fuzzy C-Means (FCM) Clustering*

Most analytical fuzzy clustering algorithms are based on optimization of the basic c-means objective function, or some modification of it. The Fuzzy c-Means Functional, which is to be minimized, is formulated as:

$$J(\boldsymbol{Z};\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^{m}\|\boldsymbol{z}_k-\boldsymbol{v}_i\|^2\ A \tag{13}$$

where, $\boldsymbol{U} = [\mu_{ik}] \in M_{fc}$ is a fuzzy partition matrix of $\boldsymbol{Z}$. $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_c], \boldsymbol{v}_i \in R^n$ is a vector of cluster prototypes (centers), which have to be determined. $d^2(\boldsymbol{z}_k, \boldsymbol{v}_i)A = D_{ikA}^2 = \|\boldsymbol{z}_k - \boldsymbol{v}_i\|^2 A = (\boldsymbol{z}_k - \boldsymbol{v}_i)^T A(\boldsymbol{z}_k - \boldsymbol{v}_i)$ is a squared inner-product distance norm, and $m = [1, \infty)$ is a parameter which determines the fuzziness of the resulting clusters.

### *C.Fuzzy C-Means (FCM) Algorithm to Delineate Homogeneous Precipitation Regions*

Suppose there are *N* sites in a study area. The '*n*' attributes, influencing precipitation at each site, have to be identified. The attributes may include large scale atmospheric variables (LSAVs) or their principal components, location parameters (latitude, longitude and altitude), and seasonality measures. Subsequently, a feature vector is formed for each site using the identified attributes for the site.

The $i^{th}$ site is denoted in n-dimensional attribute space by the feature vector

$$\boldsymbol{y}_i = [y_{1i}, \dots, y_{ji}, \dots, y_{ni}]^T \in R^n, \tag{14}$$

where, $y_{ji}$ is the value of $j^{th}$ attribute in $y_i$. The attributes of $y_i$ are rescaled using

$$x_{ji} = \frac{(y_{ji} - \bar{y_j})}{\sigma_j}, for\ 1 \le j \le n, 1 \le i \le n \tag{15}$$

where, $x_{ji}$ denotes the rescaled value of $y_{ji}$, $\sigma_j$ represents the standard deviation of attribute *j*, and $\bar{y_j}$ is the mean value of attribute *j* over all the *N* feature vectors.

Rescaling the attributes is necessary to nullify the differences in their variance, relative magnitude and importance. Otherwise, attributes having greater magnitude and variance influence the formation of clusters, which is undesirable. If certain attributes are known to be more important than others in influencing precipitation in the study area, the rescaling should be such that the variances of rescaled values of those attributes are greater than those of the less important attributes.

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_N)$ denote matrix containing rescaled feature vectors, where xi is rescaled feature vector for the $i^{th}$ site. Next task is to partition $\boldsymbol{X}$ into *c* soft clusters using Fuzzy c-means (FCM) algorithm, to arrive at optimum value of the following objective function:

Minimize,

$$J(\boldsymbol{X};\boldsymbol{U},\boldsymbol{V}) = \sum_{k=1}^{c}\sum_{i=1}^{N}(\mu_{ki})^{m}\|\boldsymbol{x}_i-\boldsymbol{v}_k\|^2 \tag{16}$$

or can be written as,

Minimize,

$$J(\boldsymbol{X};\boldsymbol{U},\boldsymbol{V}) = \sum_{k=1}^{c}\sum_{i=1}^{N}(\mu_{ki})^{m}\ d^2(\boldsymbol{x}_i, \boldsymbol{v}_k) \tag{17}$$

Subject to the following constraints,

$$\sum_{k=1}^{c}\mu_{ki} = 1\ \forall i\ \in \{1, \dots., N\} \tag{18}$$

$$0 < \sum_{i=1}^{N}\mu_{ki} < N\ \forall k \in\ \{1, \dots, c\} \tag{19}$$

where $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, \dots, \boldsymbol{v}_c)$ represents a matrix containing cluster centroids and $\boldsymbol{v}_k = [v_{1k}, \dots, v_{jk}, \dots, v_{nk}] \in R^n$ denotes centroid of $k^{th}$ soft cluster, $\mu_{ki} \in [0,1]$ denotes the membership of $x_i$ in the $k^{th}$ soft cluster,

$\boldsymbol{U}$ is the fuzzy partition matrix which contains the membership of each rescaled feature vector in each soft cluster,

The parameter $m \in [1, \infty)$ refers to the weight exponent for each fuzzy membership, and is known as fuzzifier;

d($\boldsymbol{x}_i$, $\boldsymbol{v}_k$) is the distance from $\boldsymbol{x}_i$ to $\boldsymbol{v}_k$.

The iterative procedure of FCM algorithm used to arrive at homogeneous precipitation regions is summarized below:

(i) Initialize fuzzy partition matrix $\boldsymbol{U}$ using a random number generator.

(ii) Adjust the initial memberships $\mu_{ki}^{init}$ of $x_i$ belonging to cluster *k* using the following equation:

$$\mu_{ki} = \frac{\mu_{ki}^{init}}{\sum_{j=1}^{c}\mu_{ji}^{init}}, for\ 1 \le k \le c, 1 \le i \le N \tag{20}$$

(iii) Compute the fuzzy cluster centroid $\boldsymbol{v}_k$ as

$$\mathbf{v}_{\mathrm{k}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}}(\mu_{\mathrm{ki}})^{\mathrm{m}}\mathbf{x}_{\mathrm{i}}}{\sum_{\mathrm{i=1}}^{\mathrm{N}}(\mu_{\mathrm{ki}})^{\mathrm{m}}}, \mathrm{for}\ 1 \le \mathrm{k} \le \mathrm{c} \tag{21}$$

(iv) Update the fuzzy membership $\mu_{ki}$ as

$$\mu_{ki} = \frac{\left(\frac{1}{d^2(x_i v_k)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d^2(x_i v_k)}\right)^{\frac{1}{m-1}}}, for\ 1 \le k \le c, 1 \le i \le N \tag{22}$$

(v) Compute the value of objective function as

$$J(\boldsymbol{X};\boldsymbol{U},\boldsymbol{V}) = \sum_{k=1}^{c}\sum_{i=1}^{N}(\mu_{ki})^{m}\ d^2(\boldsymbol{x}_i, \boldsymbol{v}_k) \tag{23}$$

Repeat the steps (iii) to (v) until change in the value of objective function between two successive iterations becomes sufficiently small.

### *D.Parameters of the FCM Algorithm*

Before using the FCM algorithm, the following parameters must be specified:

1. The Number of Clusters, *c*

The number of clusters *c* is the most important parameter, in the sense that the remaining parameters have less influence on the resulting partition. When clustering real data without any a priori information about the structures in the data, one usually has to make assumptions about the number of underlying clusters. The chosen clustering algorithm then searches for *c* clusters, regardless of whether they are really present in the data or not.

2. The Fuzziness Exponent or Fuzzifier, $m$

The weighting exponent $m$ is an important parameter because it significantly influences the fuzziness of the resulting partition. As $m$ approaches $1 (m \rightarrow 1)$, the partition becomes hard ($\mu_{ik} \in \{0,1\}$). As $m \rightarrow \infty$, the partition becomes completely fuzzy ($\mu_{ik} = 1/c$). Usually, $m = 2$ is initially chosen.

3. The Termination Tolerance (Absolute Difference between Two Successive Iterations)

The FCM algorithm stops iterating when the norm of the difference between $\boldsymbol{U}$ in two successive iterations is smaller than the termination tolerance. Usually 0.001 is taken, even though 0.01 works well in most cases, while drastically reducing the computing times.

4. The Fuzzy Partition Matrix, $\boldsymbol{U}$

The fuzzy partition matrix must be initialized at the beginning. However, taking random value for $\boldsymbol{U}$ is also acceptable as the algorithm is not affected by the initial value of $\boldsymbol{U}$.

## III. Application

This section provides description of the study area and details concerning the data used for the study.

### *A. Description of the Study Area*

In this study, for making the clusters, 10 IMD (Indian Meteorological Department) stations have been selected, situated in various regions of the North-east. On the Basis of IMD data available for those stations, homogeneous clustering has been done using large scale atmospheric variables (GCM data) with the use of fuzzy clustering approach.

The IMD stations selected, along with their latitude-longitude-elevation, are given in Table I. The locations of the stations are shown in the Fig. 1.

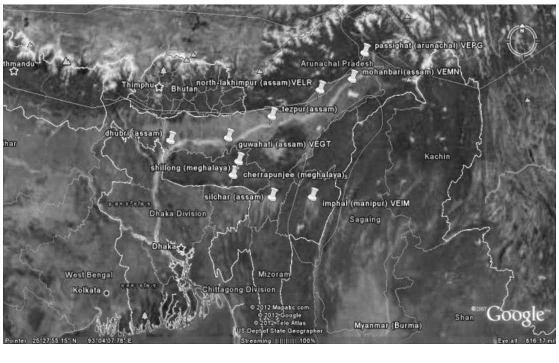

Fig. 1 Location of the IMD stations (courtesy: Google Earth)

TABLE I
Latitude-Longitude-Elevation of the IMD Stations

| Index No. | Name of the stations | Region | Latitude | Longitude | Elevation |
|---|---|---|---|---|---|
| 42220 | Passighat (VEPG) | Arunachal | 28 °6' 0" N (28.1°) | 95 °23' 0" E (95.3833°) | 157 m (515 ft) |
| 42309 | North-lakhimpur (VELR) | Assam | 27 °14' 0" N (27.233°) | 94°7' 0" E (94.1167 °) | 101 m (331 ft) |
| 42314 | Mohanbari (VEMN) | Assam | 27 °29' 2" N (27.4839 °) | 95°1' 1" E (95.0169 °) | 111 m (364 ft) |
| 42404 | Dhubri | Assam | 26 °1' 0" N (26.0167 °) | 89 °59' 0" E (89.9833°) | 35 m (115 ft) |
| 42410 | Guwahati (VEGT) | Assam | 26 °6' 22" N (26.1061 °) | 91°35' 9" E (91.5859°) | 54 m (177 ft) |
| 42415 | Tezpur | Assam | 26 °37' 0" N (26.6167 °) | 92 °47' 0" E (92.7833 °) | 91 m (299 ft) |
| 42515 | Cherrapunjee | Meghalaya | 25 °15' 0" N (25.25°) | 91 °44' 0" E (91.7333 °) | 1300 m (4265 ft) |
| 42516 | Shillong | Meghalaya | 25 °34' 0" N (25.5667 °) | 91 °53' 0" E (91.8833°) | 1600 m (5249 ft) |
| 42619 | Silchar | Assam | 24 °45' 0" N (24.75°) | 92 °48' 0" E (92.8°) | 21 m (70 ft) |
| 42623 | Imphal (VEIM) | Manipur | 24 °45' 36" N (24.7599 °) | 93 °53' 48" E (93.8967°) | 774 m (2539 ft) |

### *B. Data Used*

Two types of data have been used in this study till now as described below:

1. Observed Precipitation Data

Observed daily precipitation data has been collected from Indian Meteorological Department (IMD) under an MOU between IIT Guwahati and IMD. The time period of data collection is from 01-01-1969 to 31-01-2012. However, no data are available for any station from 2001 to 2005.

2. GCM Data

The GCM data are downloaded from Intergovernmental Panel on Climate Change (IPCC) from Fourth Assessment Report (AR4). The time period of fourth assessment is from 2001 to 2100. In this study, GCM model, HadCM3, with A2 simulation run has been used. A2 scenario considers a very heterogeneous world with continuously increasing global population and regionally oriented economic growth. It considers the forcing effect of greenhouse gases and sulphate aerosol direct effect, which are based on IPCC SRES-A2 (Special Report on Emission Scenario A2).

HadCM3 stands for Hadley centre Coupled Model version 3. It is a coupled atmosphere-ocean general circulation model (AOGCM) developed at the Hadley Centre in the United Kingdom Met Office (UKMO). Unlike earlier AOGCMs at the Hadley Centre and elsewhere, this model does not require flux adjustment. HadCM3 is composed of two components: the atmospheric model HadAM3 and the ocean model (which includes a sea ice model). Simulations often use a 360-day calendar, where each month is 30 days. The atmospheric model, HadAM3, is a grid point model and has a horizontal resolution of 3.75° in longitude × 2.5° in latitude, which gives a global grid of 96 × 73 grid points. This gives a resolution of approximately 300 km. There are 19 levels in the vertical. The ocean model has a resolution of 1.25° × 1.25° with 20 vertical levels. This model has been used in lot of projects involving climate change and its prediction. It was one of the major models used in IPCC third assessment report. This model has higher resolution compared to other models.

### *C. The Fuzzy Clustering Approach*

In the present work, precipitation data for all the stations are available from IMD. Hence, attempt is made to regionalize the stations using both GCM data as well as observed precipitation data and to make a comparison between the results. For doing that, the observed rainfall data of the 10 IMD stations were taken. Since the GCM data are available only at the grid points, hence for a particular IMD station, data of the nearest grid point was taken.

#### 1. Regionalization Using GCM Data

##### a) Selection of Attributes

The first work is to find out the LSAVs, which influence the precipitation of a particular station. For doing this, a relation has to be established between the observed precipitation data and the LSAVs. In the present work, Pearson Correlation is considered to be an effective way to find out the relation. Hence, for each station, Pearson co-efficients were calculated using the data available. It has been observed from the correlation results (not shown for brevity) that precipitation data of different stations have different correlation with different LSAVs. Hence those LSAVs, which have good relations (more than 55%) with precipitation data of most of the stations, are selected as attributes. The 13 selected LSAVs are: hur200, hur500, hur850, mrso, psl, ta200, ta500, ta850, tas, ts, ua200, ua500 and zg200.

##### b) Principal Component Analysis

To reduce the curse of high dimensionality, mean monthly values of each of the 13 LSAVs were computed at each of the 10 stations. The mean monthly value of a variable denotes the average value of the variable computed for the month, overall years of the historical record. Thus there are 12 mean monthly values for each LSAV at each IMD station. Consequently, total 1560 values (13 LSAVs × 12 months × 10 stations) were obtained. Since several of the atmospheric variables are correlated to each other, hence to avoid redundancy, Principal Component Analysis (PCA) has been done.

Principal Component Analysis (PCA) is a method of dimensionality reduction without much sacrificing the accuracy. PCA aims to summarize data with many independent variables to a smaller set of derived variables, in such a way, that the first component has maximum variance, followed by the second component and so on. In this case, the Principal Components (PCs) which preserved more than 98% of the variance were extracted from the 1560 values.

In the present work, PCA has been done with the help of MATLAB program. A table is made of the 1560 values, in MSExcel, putting the monthly data of all the 13 LSAVs along a row for each of the 10 stations. Putting the table as an input in MATLAB, the PCs were found out. In this case, 2 PCs (PC1 and PC2) were found preserving 100% of the variance, thus producing a 2 × 10 matrix of 20 values. This matrix has been used as the input data matrix X in the FCM algorithm used for the analysis.

##### c) Selection of Initial Partition Matrix

To proceed with the FCM algorithm, partition matrix U has to be initialized at first. The initial partition matrix Uinit was determined with the help of geographical location of the stations and cross-correlation among the stations.

Cross-correlations for all the stations have been determined for the followings:

- Cross-correlation for mean monthly precipitation,
- Cross-correlation for monthly maximum precipitation,
- Cross-correlation for monthly minimum precipitation,
- Cross-correlation for monthly dry period.

The cross-correlation tables are not shown here for brevity. With a close observation of the cross-correlations and the locations of the stations, the partition matrix U is initialized. Initially the cluster number was considered to be three, based on the reason that the total number of IMD stations was small and their locations are comparatively close to each other with respect to the grid size of GCM data.

##### d) Regionalization Using FCM Algorithm

Using the initial partition matrix $U_{init}$ and the 10 × 2 data matrix X (output matrix of PCA), regionalization i.e. homogeneous clustering has been done with the help of FCM algorithm. A MATLAB program is made for the FCM algorithm where $U_{init}$ and X were used as inputs. The algorithm gives the final partition matrix U as output.

#### 2. Regionalization Using Observed Precipitation Data

The conventional method of using directly the precipitation data for regionalization has also been included in the present study. Same procedure for PCA, Uinit and FCM analysis was followed. The only difference is that, input matrix of PCA was composed of monthly precipitation data instead of LSAV data.

#### 3. Regionalization using GCM Data Excluding Dry Period

By giving a close observation at the daily precipitation data of the 10 station points, it came into notice that, during the months January, February, November and December no precipitation occurs. Sometimes for a day or two, precipitation of small magnitude occurs which doesn’t have any significance over monthly mean value. Since this case is similar for all the stations, hence it gives good correlation among the stations when monthly mean values are being considered. Thus a possibility arises that the results of homogeneous clustering may get affected.

Therefore, a further analysis has been done without considering the data of the dry period (i.e. from November to February). Pearson Coefficients were again calculated for all the stations to determine the most influencing LSAVs. Those LSAVs, which have good relations (more than 50%) with precipitation data of most of the stations, are selected as attributes. The 4 selected LSAVs are: ta500, ua200, ua500 and zg200. Other variables have very poor correlation with precipitation.

Principal component analysis was done by using the GCM data of the 4 selected LSAVs excluding the dry period. Regionalization was done in the same way with the help of FCM algorithm.

4. Regionalization Using Observed Precipitation Data Excluding Dry Period

Regionalization has also been done by the conventional method, using directly the precipitation data and excluding dry period. Same procedure was followed. The only difference is that, input matrix of PCA was composed of monthly precipitation data instead of LSAV data.

5. Regionalization Using Interpolated GCM Data

In this study it was noticed that due to the use of nearby grid point data instead of exact data, giving definite statement on some of the stations has become difficult. Moreover in some cases the results are showing equal distribution of the stations among the clusters, further making the results unclear (detailed figures are shown in the next section). Hence, further study has been done by considering interpolated GCM data for each station. Studies have been done for both yearly mean data and for monsoon (excluding dry period) data. The 16 LSAVs selected in the former case are hur200, hur500, hur850, mrso, pr, prc, psl, ta200, ta500, ta850, tas, ts, ua200, ua500, ua850 and zg200. The 4 LSAVs selected in the later case (excluding dry period) are: ta500, ua200, ua500 and zg200.

## IV. Results and Discussion

From the results of the FCM analysis it came into notice that, different clustering has been found for different conditions. In the resulting clusters, obtained for both monthly means of the whole year and for monsoon only (excluding dry period), shown in Fig. 2, it is seen that all the stations are coming under single cluster (except Cherrapunjee) when observed precipitation data are used. It indicates similarity in the rainfall pattern in all the stations both in monsoon and non-monsoon period. The only conclusion found here is that conventional method is not very useful for clustering.

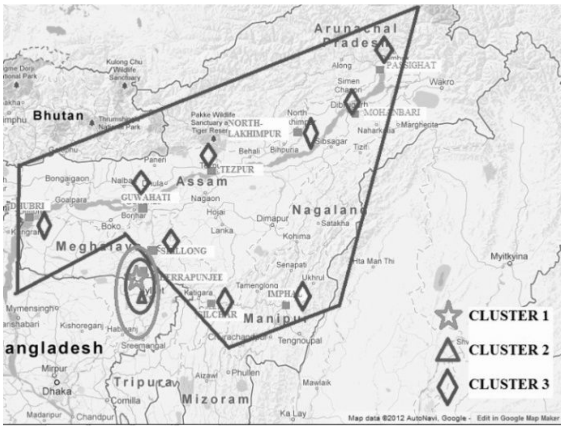

Fig. 2 Clustering done for precipitation data (both yearly and monsoon data)

The clusters obtained from non-conventional regionalization (using GCM data) are shown in Figs. 3 and 4.

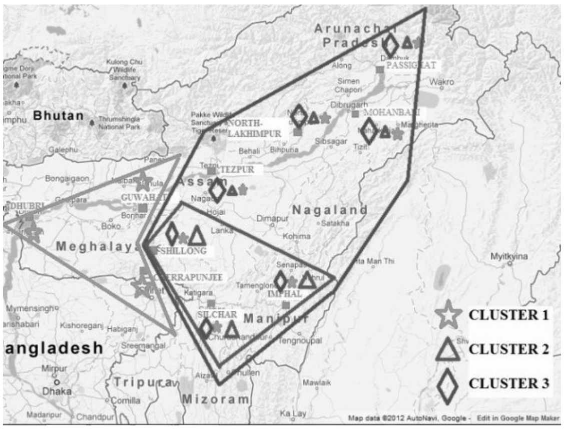

Fig. 3 Clustering done for GCM data

Here it is found that, when GCM data were considered, the analysis gave good results. However there is a little difference in the results of yearly means and monsoon period, but can be adjustable. The resulting pattern of clusters shows similarity with the pattern of GCM data used in the analysis. This can be attributed to the reason that GCM data were considered at the nearby grid points, instead of actual interpolated GCM data. Hence, further study has been done by considering interpolated GCM data for each station to get more reliable.

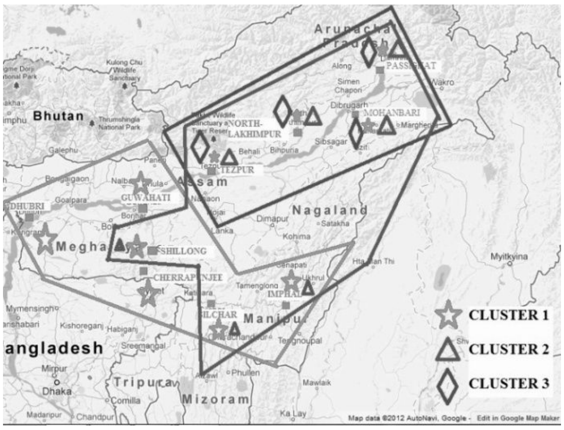

Fig. 4 Clustering done for GCM data (excluding dry period)

Studies have been done for both yearly mean data and for monsoon (excluding dry period) data. Resulting clusters are shown in Figs. 5 and 6.

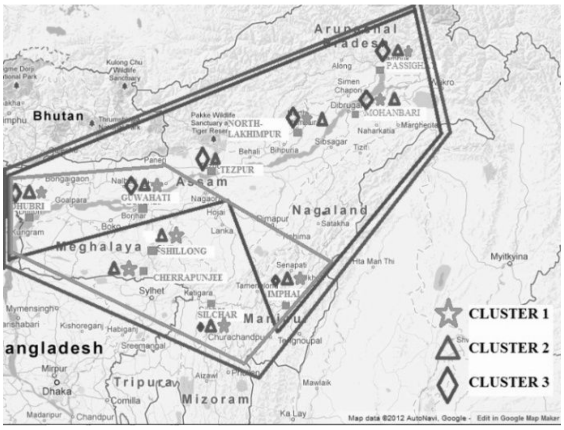

Fig. 5 Clustering done for interpolated GCM data

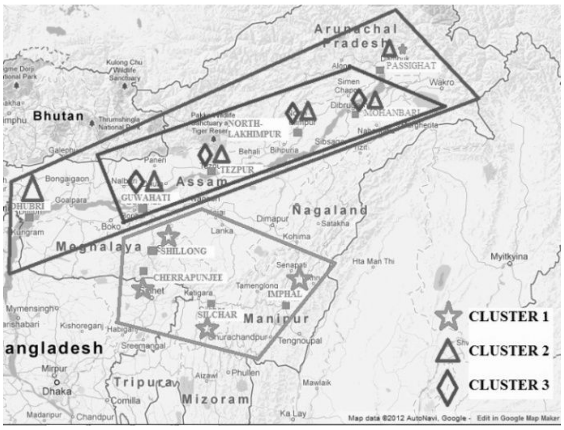

Fig. 6 Clustering done for interpolated GCM data (excluding dry period)

The clusters obtained for interpolated GCM data shows a different kind of pattern. Yearly means give a mixed clustering, whereas monsoon data shows that the stations on the Brahmaputra valley region are coming in the same cluster and the remaining stations which are nearer to Barak valley region are coming under a different cluster.

## V. Summary and Conclusion

From all the results, it was being noticed that Passighat, North-lakhimpur, Mohanbari and Tezpur were always coming under the same cluster. Hence it can be concluded that these four stations will always come under the same cluster. Therefore further analysis on the downscaling can be done at any two stations among them. Similarly, Shillong, Silchar, cherrapunjee and imphal were also coming under the same cluster for most of the analyses. Guwahati is showing similarity with both the clusters. Dhubri is showing different results; hence clear statement cannot be given and further analysis has to be done.

From the above analysis and comparison it can be concluded that non-conventional method of using GCM data is somehow giving better results than the others.

It is also a matter of concern that many of the input variables of the FCM algorithm were taken arbitrarily, such as the cluster number and the fuzziness parameter m. Further analysis can be done by changing the number of clusters and the value of fuzziness parameter to get more reliable results. A homogeneity test can be done to test the acceptability of the resulting clusters by any one of the various homogeneity test methods that is found suitable.

In the present study, monthly means (averaged over all the years of record) for each station have been used for the analysis to reduce the curse of high dimensionality. However, this consideration eliminated the effect of maximum and minimum values. Hence there can also be a future analysis with the use of daily data.

## Acknowledgment

The authors express their sincere thanks to Indian Meteorological Department (IMD) and Intergovernmental Panel on Climate Change (IPCC) for providing the data required for doing all the analyses. Sincere gratitude is extended towards Prof. Swaroop Nandan Bora, Department of Mathematics, IIT Guwahati, for his invaluable suggestions and co-operation at different stages of the work. We are also thankful to friends and well-wishers for their help and support.

## References

[1] D. Mehrotra and R. Mehrotra, “Climate change and hydrology with emphasis on the Indian subcontinent,” *Hydrological Sciences Journal*, vol. 40:2, pp. 231-242, April 1995.

[2] R. L. Wilby, H. Hassan, and K. Hanaki, “Statistical downscaling of hydrometeorological variables using general circulation model output,” *Journal of Hydrology*, vol. 205, pp. 1-19, 1998.

[3] R. L. Wilby, T. M. L. Wigley, D. Conway, P. D. Jones, B.C. Hewitson, J. Main, and D. S. Wilks, “Statistical downscaling of general circulation model output: A comparison of methods,” *Water Resources Research*, vol. 34, no. 11, pp. 2995-3008, Nov. 1998.

[4] X.-C. Zhang, “Spatial downscaling of global climate model output for site-specific assessment of crop production and soil erosion,” *Agricultural and Forest Meteorology*, vol. 135, pp. 215-229, 2005.

[5] A. K. Gosain, S. Rao, and D. Basuray, “Climate change impact assessment on hydrology of Indian river basins,” *Current Science*, vol. 90, no. 3, pp. 346-353, Feb. 2006.

[6] H. D. Fill, and J. R. Stedinger, “Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple’s test,” *Journal of Hydrology*, vol. 166, pp. 81-105, 1995.

[7] A. Habib, and M. Ellouze, “Hydrological delineation of homogeneous regions in Tunisia,” *Water Resources Management*, vol. 20, pp. 961–977, 2006.

[8] P. Satyanarayana, and V.V. Srinivas, “Regionalization of precipitation in data sparse areas using large scale atmospheric variables – A fuzzy clustering approach,” *Journal of Hydrology*, vol. 405, pp. 462–473, 2011.

**Arup K. Sarma** (ISTE’91–LM’01–M’03) became Member of Indian Society for Technical Education (ISTE) in 1991, a Life Member of ICE (LM) in 2001 and Member (M) of ASCE in 2003. Dr. Sarma was born in India on 24th January 1963. He received BE (civil engg.) from Jorhat Engineering College, M.E. (watershed management and flood control) from Assam Engineering College and PhD degree (Hydraulic and Water Resources Engineering) from Gauhati University of India, in the year 1999.

He is currently working as professor and head, Department of Civil Engineering, IIT Guwahati in India. He has to his credit more than 90 publications in reputed journals, books and conferences. His video course on “Hydraulics” is receiving wide appreciation from all over the world. Numerical simulation of free surface flow, water resources system optimization, climate change impact on water resources and ecological management practices for flood and erosion control are some of his research field. He has so far produced 6 PhDs, 49 M.Tech and has completed more than 50 consultancy and research projects from India and abroad.

**Dr. Sarma** was awarded with prestigious *B.P.Chaliha* Chair for Water Resources, a chair given by ministry of Water Resources, Govt. of India, He is reviewer of several reputed journals, which include journal of Hydrology, ASCE, Journals of Water Management and Maritime Engg of ICE(London), Journal of Water resources management (WARM) and Journal of Hydro Env Research (JHER).