International Conference "Water, Environment, Energy and Society" (WEES–2009)
New Delhi, 12–16 January 2009

Regional Frequency Analysis of Rainfall in India under Climate Change Scenarios

P. Satyanarayana and V.V. Srinivas[1]
Department of Civil Engineering, Indian Institute of Science
Bangalore - 560 012, INDIA
E-mail: [1]vvs@civil.iisc.ernet.in

ABSTRACT: Recently there is growth in scientific consensus that global climate is changing. In this scenario engineers are devoting their efforts to explore plausible implications of the climate change on water which is one of the most vulnerable resources of earth. In India, rainfall is the major source of water to river basins. Effective estimation of the magnitude and frequency of rainfall is necessary for hydrologic designs. However, often the available at-site information on rainfall is inadequate to arrive at reliable estimates. This necessitates the use of regional frequency analysis to pool adequate information from several locations in the region that are similar in terms of their hydro-meteorological characteristics. A few attempts have been made in the past three decades to identify homogeneous monsoon rainfall regions over India. However, the regions were not effectively validated. In this study, it is shown that the homogeneous monsoon regions that are in use by India Meteorological Department (IMD) are statistically heterogeneous. Subsequently, a novel regionalization procedure based on hydro-meteorological input is presented to form homogeneous rainfall regions in India. Following this, plausible implication of climate change on the delineated regions is assessed by using simulations from Canadian third Generation coupled General Circulation model. Results indicate that the proposed approach to regionalization is efficient in delineating homogeneous rainfall regions, and the future changes projected for the delineated homogeneous SMR regions are insignificant.

INTRODUCTION

Information pertaining to the amount and frequency of precipitation is necessary for a wide range of hydrologic applications that include planning of agriculture, design and operation of irrigation projects, and investigating frequency and spatial distribution of meteorological droughts. Traditionally, at-site frequency analysis methods were used to determine required rainfall quantile estimates. However, these methods are not suitable if target locations (sites) have inadequate data. Practicing hydro-meteorologists overcome this impediment by pooling information at target site with that from other locations depicting similar characteristics of precipitation to arrive at reliable quantile estimates. The procedure of identifying pooling group(s) or region(s), consisting of sites having similar characteristics, is known as regionalization. The frequency analysis based on the pooled information is called regional frequency analysis. Inherent in this analysis is the assumption that frequency distributions of data at all the sites in a pooling group are similar. Research has shown that even for a moderately heterogeneous region, quantile estimates based on regional frequency analysis can be considered sufficiently accurate for practical purposes (Hosking and Wallis, 1997).

Over the past four decades hydro-meteorologists have developed several approaches to regionalization of precipitation, which include elementary linkage analysis (e.g., Jackson, 1974; Sumner, 1983), spatial correlation analysis (e.g., Gadgil *et al.*, 1993), common factor analysis (e.g., Bärring, 1987), empirical orthogonal function analysis (e.g., Bedi and Bindra, 1980; Kulkarni *et al.*, 1992), Principal Component Analysis (PCA) (e.g., Singh and Singh, 1996; Iyengar and Basak, 1994), cluster analysis (e.g., Obregón and Nobre, 2006), and PCA in association with cluster analysis (e.g., Guttman, 1993; Dinpashoh *et al.*, 2004).

The traditional approaches to regionalization of precipitation are based on statistics computed from the observed precipitation, rather than attributes effecting hydro-meteorology in a region. Therefore independent validation of the delineated regions for homogeneity in precipitation was not possible. Herein, to address this issue, a new methodology is proposed for regionalization of precipitation. Large scale atmospheric variables effecting precipitation in a region and location attributes (latitude, longitude and altitude) are suggested for use as features for regionalization of precipitation by cluster analysis. This allows independent validation of the delineated regions for homogeneity, by using

[1]Conference speaker

statistics computed from the observed precipitation. The effectiveness of the proposed methodology is illustrated through application to Summer Monsoon Rainfall (SMR) data of India for delineating homogeneous regions.

The study region is selected because the knowledge of homogeneous rainfall regions is of great significance in India owing to its agro-based economy. The region receives significant amount of rainfall during summer monsoon season (June to September). The SMR regions that are currently in use by India Meteorological Department (IMD) are based on political boundaries, and are found to be statistically heterogeneous. Therefore there is a need to delineate new homogeneous SMR regions.

Furthermore, recently there is growth in interest on understanding plausible implications of climate change on spatial distribution of future rainfall. In this study, implication of climate change on regions delineated using proposed methodology was assessed using simulations from T63 version of Canadian third generation coupled General Circulation Model (CGCM3.1/T63). The General Circulation Models are numerical models representing physical processes in the atmosphere, ocean, cryosphere and land surface which are considered to depict earth's climate system. These models have been evolving steadily over the past few decades, and are considered as the most advanced tools currently available to simulate climatic conditions on earth several decades into the future.

METHODOLOGY

This section describes the proposed methodology to form homogeneous SMR regions in India. First, a brief note is provided on the selection of attributes for regionalization. Next, K-means clustering algorithm used to form plausible homogeneous rainfall regions based on the selected attributes is described. Following this, cluster validity indices considered for determining optimal number of clusters are presented. Subsequently, heterogeneity measures that were used to test homogeneity of the delineated regions, and the procedure adopted to adjust the regions are briefly described. Then details pertaining to estimation of precipitation quantiles are given. Finally, the procedure used to assess implication of climate change on the delineated regions is given.

Selection of Attributes

The selection of appropriate attributes is one of the most important steps in regionalization of rainfall. In general, meteorological variables (e.g., specific humidity, temperature, precipitable water, wind velocity and wind direction) and location parameters (e.g., latitude, longitude and elevation) which influence precipitation in a region can be considered as attributes.

Fifteen large scale atmospheric variables that were identified as predictors influencing precipitation in India were chosen as attributes for regionalization. The selected variables are air temperature at 4 pressure levels (925, 700, 500 and 200 milli bar (mb)), geopotential height at 3 pressure levels (925, 500, and 200mb), specific humidity at 2 pressure levels (925 and 850 mb), zonal and meridional wind velocities at 2 pressure levels (925 and 200 mb), precipitable water and surface pressure. The location attributes namely, latitude, longitude and elevation were also considered as attributes. Latitude and longitude were selected because geographically nearby sites could have similarities in precipitation events. Further, elevation is chosen as it influences precipitation.

K-Means Algorithm for Regionalization of Rainfall

The K-means algorithm (McQueen, 1967) is an iterative procedure that is commonly used to identify clusters in a given data set. Herein, the algorithm to arrive at homogeneous rainfall regions is described.

Let $\mathbf{Y} = \{\mathbf{y}_i / i = 1,\ldots,N\}$ denote a set of N feature vectors in n-dimensional attribute space {i.e., $\mathbf{y}_i = [y_{i1},\ldots,y_{ij},\ldots,y_{in}] \in \Re^n$}, where y_{ij} is the value of attribute j in i^{th} feature vector $\mathbf{y}_i$. Each feature vector represents one of the N sites (rain gauges) in the study region. It comprises of large scale atmospheric variables influencing precipitation at a site (or their principal components), and geographical location attributes.

Let $\mathbf{x}_i$ denote the i^{th} rescaled feature vector in the n-dimensional attribute space {i.e., $\mathbf{x}_i = [x_{i1},\ldots,x_{ij},\ldots, x_{in}] \in \Re^n$} obtained by rescaling $\mathbf{y}_i$ using Eqn. (1),

$$x_{ij} = \frac{(y_{ij} - \bar{y}_j)}{\sigma_j} \quad \text{for } 1 \le j \le n \qquad \ldots (1)$$

where x_{ij} denotes the rescaled value of y_{ij}, σ_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 because of the differences in their variance and relative magnitudes. If the attributes are not rescaled, those having greater magnitude and variance influence the

formation of clusters. Rescaling of principal components may not be necessary if they are considered as attributes.

In the K-means algorithm the feature vectors move from one cluster to another to minimize the objective function, F, defined as,

$$F = \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{i=1}^{N_k} d^2(x_{ij}^k - x_{\bullet j}^k) \quad \ldots (2)$$

where K denotes the number of clusters; N_k represents the number of feature vectors in cluster k; x_{ij}^k denotes the rescaled value of attribute j in the feature vector i assigned to cluster k; $x_{\bullet j}^k$ is the mean value of attribute j for cluster k, computed as,

$$x_{\bullet j}^k = \frac{\sum_{i=1}^{N_k} x_{ij}^k}{N_k} \quad \ldots (3)$$

By minimizing F in Eqn. (2), the distance of each feature vector from the center of the cluster to which it belongs is minimized. We have the option to incorporate the knowledge about the global shape or size of clusters by using an appropriate distance measure $d(\cdot)$, such as Euclidean or Mahalanobis. Euclidean distance measure is used in this study.

The steps in K-means algorithm to delineate clusters for a given value of K are as follows:

1. Set 'current iteration number' t to 0 and maximum number of iterations to t_max.
2. Initialize K cluster centers to random values in the n-dimensional feature vector space.
3. Initialize the 'current feature vector number' i to 1.
4. Determine Euclidean distance of i-th feature vector $\mathbf{x}_i$ from centers of each of the K clusters, and assign it to the cluster whose center is nearest to it.
5. If $i < N$ increment i to $i + 1$ and go to step (4), else continue with step (6).
6. Update the centroid of each cluster by computing average of the feature vectors assigned to it. Then compute F for the current iteration t using Eqn. (2).
   If $t = 0$, increase t to t+1 and go to step (3).
   If $t > 0$ compute the difference in the values of F for iterations t and t–1.Terminate the algorithm if change in the value of F between two successive iterations is insignificant, else continue with step (7).
7. If $t < t\_max$ update t to t+1 and go to step (3), else terminate the algorithm.

The optimal value attained by F depends on the assumed number of clusters (K) and initialized values of their centers. These *a priori* assumptions are necessary, however they do not guarantee optimal partition. In this study, for each value of K, cluster centers were randomly initialized 25 times to arrive at the optimal partition. Further to choose optimal number of clusters, cluster validity indices were considered.

Cluster Validity Indices

Cluster validity indices are useful to identify compact and well separated clusters (Halkidi *et al.*, 2001). In this study, three cluster validity indices, namely, Dunn's index (Dunn, 1973), Davies–Bouldin index (Davies and Bouldin, 1979) and Calinski-Harabasz index (Calinski and Harabasz, 1974) were used to determine optimal partition provided by the K-means clustering algorithm.

Dunn's Index (V_D) is computed as,

$$V_D = \min_{1 \le i \le K} \left\{ \min_{1 \le j \le K, j \ne i} \left[\frac{\delta(C_i, C_j)}{\max_{1 \le k \le K} \Delta(C_k)} \right] \right\} \quad \ldots (4)$$

where $\delta(C_i, C_j)$ denotes the distance between clusters C_i and C_j (inter-cluster distance) computed using Eqn. (5); $\Delta(C_k)$ represents the intra-cluster distance of cluster C_k defined by Eqn. (6). The value of K for which V_D is maximized is taken as the optimal number of clusters,

$$\delta(C_i, C_j) = \max_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} \left[d(\mathbf{x}_i, \mathbf{x}_j) \right] \quad \ldots (5)$$

$$\Delta(C_k) = \max_{\mathbf{x}_i, \mathbf{x}_j \in C_k} \left[d(\mathbf{x}_i, \mathbf{x}_j) \right] \quad \ldots (6)$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance between rescaled feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Davies-Bouldin Index (V_{DB}) is a function of the ratio of the sum of within-cluster scatters to between-cluster separation. The scatter within the k^{th} cluster, $S_{k,q}$ is computed using Eqn. (7) and $d_{kl,\lambda}$, the Minkowski distance of order λ between the centroids that characterize clusters C_l and C_k is defined by Eqn. (8),

$$S_{k,q} = \left(\frac{1}{N_k} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \mathbf{z}_k \right\|_2^q \right)^{1/q} \quad \ldots (7)$$

$$d_{kl,\lambda} = \left\| \mathbf{z}_k - \mathbf{z}_l \right\|_\lambda = \left(\sum_{j=1}^{n} \left| x_{\bullet j}^k - x_{\bullet j}^l \right|^\lambda \right)^{1/\lambda} \quad \ldots (8)$$

where $\mathbf{z}_k$ represents the centroid of cluster k and $S_{k,q}$ is the q^{th} root of the q^{th} moment of the Euclidean distance of feature vectors in cluster k with respect to its centroid. First moment (i.e. q = 1) and Minkowski

distance of order 2 (i.e. $\lambda = 2$) which are commonly adopted by practitioners (e.g., Pakhira *et al.*, 2004), were used in the present study. The Davies–Bouldin index is computed using Eqn. (9). A small value for V_{DB} indicates good partition, which corresponds to compact clusters with their centers far apart,

$$V_{DB} = \frac{1}{K}\sum_{k=1}^{K} \underset{k,1\le l\le K,k\ne l}{max}\left\{\frac{S_{k,q}+S_{l,q}}{d_{kl,\lambda}}\right\} \quad \ldots (9)$$

Calinski-Harabasz Index (V_{CH}) of a partition $G=\{C_1,\ldots,C_K\}$ comprising K clusters is computed as,

$$V_{CH} = \frac{[trace\ \mathbf{B}/(K-1)]}{[trace\ \mathbf{W}/(N-K)]} \quad \ldots (10)$$

where $\boldsymbol{B}$ is a matrix describing dispersion of cluster centroids and $\boldsymbol{W}$ is a matrix representing within-cluster dispersion. The traces of the matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ can be written as,

$$\text{trace } \mathbf{B} = \sum_{k=1}^{K} N_k \left\|\mathbf{z}_k - \bar{\mathbf{x}}\right\|^2 \quad \ldots (11)$$

$$\text{trace } \mathbf{W} = \sum_{k=1}^{K}\sum_{\mathbf{x}_i\in C_k} \left\|\mathbf{x}_i - \mathbf{z}_k\right\|^2 \quad \ldots (12)$$

where $\bar{\mathbf{x}}$ is centroid of the entire set of feature vectors $\{\mathbf{x}_i/i = 1,\ldots,N\}$. Maximum value of V_{CH} denotes optimal partition.

L-Moment Based Regional Homogeneity Test

The homogeneity of delineated rainfall regions was assessed using L-moment based heterogeneity measures defined by Hosking and Wallis (1997). In a homogeneous region all the sites are supposed to have the same population L-moment ratios. However, their sample L-moment ratios [LMRs: coefficient of L-variation (L-CV), L-skewness and L-kurtosis] may be different due to sampling variability. The regional homogeneity tests examine whether the between-site dispersion of the sample LMRs for the group of sites under consideration is larger than the dispersion expected in a homogeneous region.

Suppose that the region to be tested for homogeneity has N_R sites, with site i having record length of rainfall n_i. Further, let $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$ denote L-CV, L-skewness and L-kurtosis of rainfall at site i, respectively. The regional average L-CV, L-skewness and L-kurtosis of rainfall, represented by t^R, t_3^R and t_4^R respectively, are computed as,

$$\left.\begin{aligned} t^R &= \sum_{i=1}^{N_R} n_i t^{(i)} \Big/ \sum_{i=1}^{N_R} n_i \\ t_3^R &= \sum_{i=1}^{N_R} n_i t_3^{(i)} \Big/ \sum_{i=1}^{N_R} n_i \\ t_4^R &= \sum_{i=1}^{N_R} n_i t_4^{(i)} \Big/ \sum_{i=1}^{N_R} n_i \end{aligned}\right\} \quad \ldots (13)$$

where, $n_i \Big/ \sum_{i=1}^{N_R} n_i$ denotes the weight applied to sample LMRs at site i, which is proportional to the sites' record length. The regional average mean l_1^R is set to 1 by scaling monsoon rainfall totals at each site by its mean.

Three Heterogeneity Measures (*HMs*) are considered: (i) weighted standard deviation of the at-site sample L-CVs (V), (ii) weighted average distance from the site to the group weighted mean in the two dimensional space of L-CV and L-skewness (V_2), (iii) weighted average distance from the site to the group weighted mean in the two dimensional space of L-skewness and L-kurtosis (V_3),

$$\left.\begin{aligned} V &= \left\{\sum_{i=1}^{N_R} n_i (t^{(i)} - t^R)^2 \Big/ \sum_{i=1}^{N_R} n_i\right\}^{1/2} \\ V_2 &= \sum_{i=1}^{N_R} n_i \left\{(t^{(i)} - t^R)^2 + (t_3^{(i)} - t_3^R)^2\right\}^{1/2} \Big/ \sum_{i=1}^{N_R} n_i \\ V_3 &= \sum_{i=1}^{N_R} n_i \left\{(t_3^{(i)} - t_3^R)^2 + (t_4^{(i)} - t_4^R)^2\right\}^{1/2} \Big/ \sum_{i=1}^{N_R} n_i \end{aligned}\right\} \quad \ldots (14)$$

In these dispersion measures, distance of sample LMRs for site i from the regional average LMRs is weighted proportionally to the sites' record length, thus allowing greater variability of LMRs for sites having small sample size in a region.

A large number of realizations (N_{sim} = 500) of rainfall were simulated for each of the regions from kappa distribution fitted to regional average LMRs: l_1^R, t^R, t_3^R and t_4^R. Each realization constitutes a homogeneous region, with N_R sites having same record length as their real-world counterparts. Further, in each realization, the data simulated at any site in the region is serially independent and the data simulated at different sites in the region are not cross-correlated. For each simulated realization, V, V_2 and V_3 are computed. Let μ_V, μ_{V_2} and μ_{V3} denote the mean and

σ_V, σ_{V_2} and σ_{V3} the standard deviation of the N_{sim} values of V, V_2 and V_3 respectively. These statistics are used to estimate the following three *HMs*,

$$H_1 = \frac{(V - \mu_V)}{\sigma_V} \quad \ldots (15)$$

$$H_2 = \frac{(V_2 - \mu_{V_2})}{\sigma_{V_2}} \quad \ldots (16)$$

$$H_3 = \frac{(V_3 - \mu_{V_3})}{\sigma_{V_3}} \quad \ldots (17)$$

A region can be regarded as 'acceptably homogeneous' if $HM < 1$, 'possibly homogeneous' if $1 \leq HM < 2$, and 'definitely heterogeneous' if $HM \geq 2$. The values of H_2 and H_3 rarely exceed 2 even for grossly heterogeneous regions and hence lack power to discriminate between homogeneous and heterogeneous regions. Consequently, H_1 is considered to be superior to H_2 and H_3 (Hosking and Wallis, 1997).

Adjusting the Regions

The regions are adjusted to improve their homogeneity following the options suggested by Hosking and Wallis (1997), which include: (i) eliminating (or deleting) one or more sites from the data set; (ii) transferring one or more discordant sites from a region to other regions; (iii) dividing a region to form two or more new regions; (iv) allowing a site to be shared by two or more regions; (v) dissolving regions by transferring their sites to other regions; (vi) merging a region with another or others; (vii) merging two or more regions and redefining groups; (viii) obtaining more data and redefining regions. Among these, the first three options are useful in reducing the values of heterogeneity measures of a region, whereas the options (iv) to (vii) help in ensuring that each region is sufficiently large in terms of collective data length at all the sites in it.

The sites that are grossly discordant with respect to other sites in a region are identified using the discordancy measure of Hosking and Wallis (1997) given by Eqn. (18). The critical value of D_i for a region depends on its size,

$$D_i = \frac{1}{3} N_R (\mathbf{u}_i - \bar{\mathbf{u}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{u}_i - \bar{\mathbf{u}}) \quad \ldots (18)$$

where $\mathbf{u}_i = \left[t^{(i)} \; t_3^{(i)} \; t_4^{(i)} \right]^{\mathrm{T}}$ is a vector containing the t, t_3, and t_4 values for site i in the region, the superscript T denotes transpose, $\bar{\mathbf{u}}$ is the unweighted group average of the L-moment ratios computed using Eqn. (19) and $\boldsymbol{S}$ is a covariance matrix computed using Eqn. (20),

$$\bar{\mathbf{u}} = \frac{\sum_{i=1}^{N_R} \mathbf{u}_i}{N_R} \quad \ldots (19)$$

$$\mathbf{S} = \sum_{i=1}^{N_R} (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^{\mathrm{T}} \quad \ldots (20)$$

In this study, to adjust a region, firstly the sites that are flagged discordant by the discordancy measure were identified. Secondly, the heterogeneity measures (H_1, H_2 and H_3) of the region to be adjusted were examined as they changed with exclusion of each site from the region. In this context, one site is eliminated at a time with replacement. Thirdly, the discordant site, whose exclusion reduces the heterogeneity measures of a region by a significant amount, was identified and removed from the region after ensuring that the site discordancy is high. This procedure is followed in Rao and Srinivas (2008) and Srinivas *et al.* (2008).

Estimation of Precipitation Quantiles

Precipitation quantile at site i for T-year recurrence interval was estimated by index flood method (Dalrymple, 1960) as,

$$\hat{P}_i^R(T) = \bar{P}_i \, \hat{p}_k^R(T) \qquad i = 1, \ldots, N \quad \ldots (21)$$

where $\bar{P}_i$ is the sample mean of the SMR at site i in region k; and $\hat{p}_k^R(T)$ is the growth curve ordinate of region k for T-year recurrence interval.

Assessment of Implication of Climate Change on Delineated Regions

Plausible implications of climate change on the delineated regions was assessed by using simulations from CGCM3.1/T63 for the period 2001–2030 for four scenarios namely A1B, A2, B1 and COMMIT. For this purpose, feature vectors were prepared using location attributes and the climate data extracted from CGCM3.1/T63 and clusters were formed using K-means algorithm. Further details of this analysis are provided in the following section.

CASE STUDY

Description of the Study Region

The study region India lies between 8°4′ and 37°6′ north latitude and 68°7′ and 97°25′ east longitude, and has an area of 32,87,263 km^2. The climate of the

region can be classified into four seasons: winter (January and February), summer (March to May), summer monsoon (June to September), and post-monsoon (October to December). The region receives more than 80% of the annual rainfall during summer monsoon. Heavy rainfall is confined largely to the Western Ghats and the northeastern parts of the country. The central region and Gangetic plain receive moderate rainfall, while the northwestern part of the country receives low rainfall towards the end of monsoon season (Sharma *et al.*, 2003). Parthasarathy *et al.* (1993) found no systematic trend in the all India rainfall in a study covering the period 1871–1990. However, they reported noting large interannual and decadal variations.

Data Used

For the study high resolution gridded daily rainfall data for the period 1951–2004 procured from IMD (Rajeevan *et al.*, 2005, 2006) were considered. Further, gridded reanalysis data of the monthly mean atmospheric variables for the study region were extracted from the database of National Centers for Environmental Prediction (NCEP) and National Center for Atmospheric Research (NCAR) (Kalnay *et al.*, 1996), for the period 1951 to 2004 from the web site http://www.cdc.noaa.gov. The reanalysis data is prepared based on historical (past) data assimilated from 1948 to the present. The spatial domain of the extracted data ranges from 47.5° N to 0° latitude, and 57.5° E to 110° E longitude at a spatial resolution of 2.5°.

Average elevation of terrain in each of the NCEP grid boxes was computed from Shuttle Radar Topography Mission (SRTM) data processed by Consortium for Spatial Information of the Consultative Group for International Agricultural Research (CGIAR-CSI), available at the web site http://srtm.csi.cgiar.org.

The climate data simulated by CGCM3.1/T63 were collated at monthly time scale for the period January 2001 to December 2030 for four scenarios namely A1B, A2, B1 and COMMIT. These scenarios are prescribed in the fourth Assessment Report (AR4) of Intergovernmental Panel on Climate Change (IPCC). They are widely known as SRES scenarios, indicating Emission Scenarios (ES) prescribed in Special Report (SR) of IPCC. The CGCM3.1/T63 has a surface grid whose spatial resolution is roughly 2.81 degrees along both latitude and longitude, and 31 levels in the vertical. The spatial domain of the extracted data ranges from 46.04° N to 1.41° N latitude, and 53.44° E to 106.88° E longitude.

RESULTS AND DISCUSSION

The IMD currently uses five homogeneous SMR regions. The statistical homogeneity of each of these regions were tested with SMR data at all the 1° × 1° grid points in it using L-moment based homogeneity test. The location of grid points in the regions is shown in Figure 1(a). The values of heterogeneity measures computed for the regions are presented in Table 1, which indicate that all the regions are statistically heterogeneous. In particular, the heterogeneity statistics for Peninsular, West central and Northwest regions are found to be very high. Herein, a region is declared as homogeneous or heterogeneous based on H_1 index because it has much better discriminatory power than H_2 and H_3 (Hosking and Wallis, 1997, p. 68).

The IMD regions were adjusted to improve their homogeneity following the procedure described earlier. The results of this analysis presented in Table 2 and Figure 1(b) show that excessive number of sites (grid points) had to be eliminated from the regions to make them acceptably homogenous. This indicates that the IMD regions are not useful as the precursors to derive homogeneous SMR regions, and new regions need to be formed.

To delineate new homogeneous SMR regions in the study area, 52 out of 60 NCEP grid boxes covering India were considered. The discarded eight grid boxes are in Himalayan mountainous region, and are shown as crossed boxes in Figure 2. They were not considered for the analysis because the density of rain gauges in the Himalayan region is very low (Rajeevan *et al.*, 2005, 2006), and some of the pressure levels considered in this study (e.g., 925 mb) are not defined for several locations in the region.

The spatial domain of 15 atmospheric variables (mentioned in the subsection "Selection of Attributes"), which influence precipitation in each NCEP grid box, was chosen as 16 NCEP grid points surrounding it. For example, Figure 2 shows spatial domain of the predictor variables for the hashed NCEP grid box enclosed by points 6, 7, 11 and 10.

At each of the sixteen NCEP grid points, the mean monthly values of each of the fifteen atmospheric variables were computed for each of the four summer monsoon months (June–September). Thus 960 values (16 grid points × 15 variables × 4 months) were obtained for each of the 52 NCEP grid boxes. Subsequently, from the 960 values, five principal components (PCs) that preserve more than 97% of the variance and the corresponding principal directions (PDs) were extracted. The standardized location attributes (latitude, longitude, and average elevation of

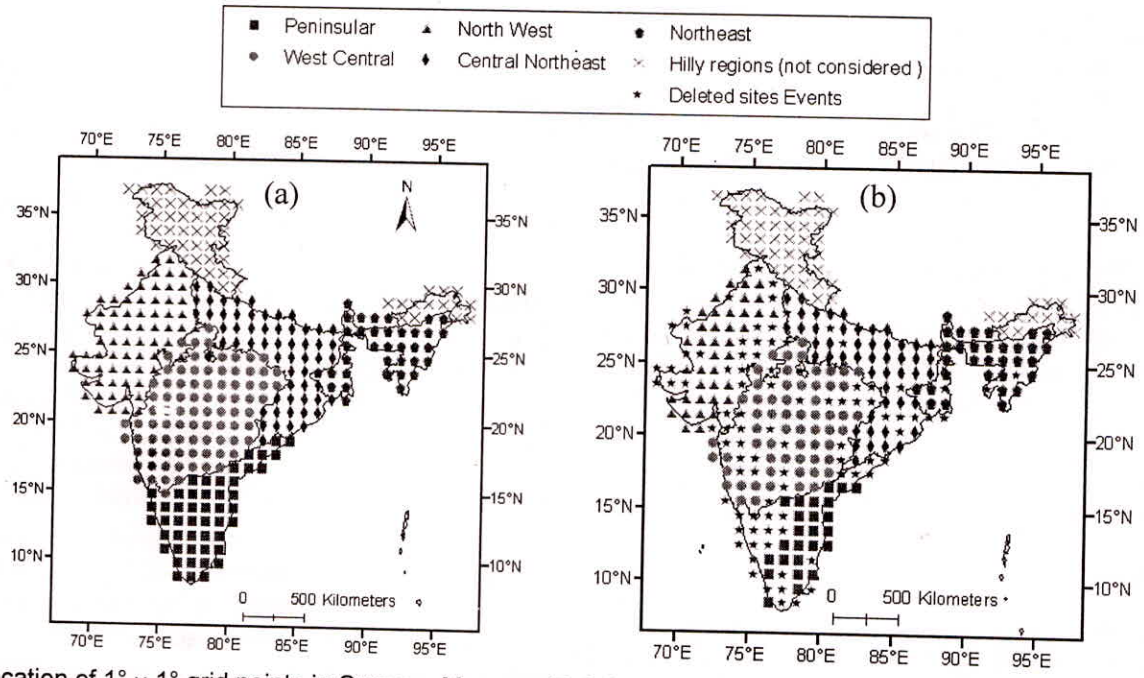

Fig. 1: Location of 1° × 1° grid points in Summer Monsoon Rainfall (SMR) regions that are considered homogeneous by IMD are shown in (a). Homogeneous regions formed by adjusting the SMR regions are shown in (b)

Table 1: Characteristics of the IMD Summer Monsoon Rainfall Regions Determined Using Heterogeneity Measures Given in Eqns. (15) to (17)

Sl. No.	*Region Name*	*Number of Grid Points*	H_1	H_2	H_3	*Region Type*
1.	Peninsular	49	23.28	5.93	0.26	Definitely Heterogeneous
2.	West Central	86	10.89	0.64	–1.33	Definitely Heterogeneous
3.	Northwest	69	20.96	5.87	–1.08	Definitely Heterogeneous
4.	Central northeast	59	4.32	–0.73	–1.90	Definitely Heterogeneous
5.	Northeast	36	4.44	–0.91	1.06	Definitely Heterogeneous

Table 2: Characteristics of the Homogeneous Regions Formed by Adjusting the IMD Summer Monsoon Rainfall Regions

Sl. No.	*Region Name*	*Number of Grid Points*	*Heterogeneity Measures*			*Number of Grid Points Eliminated*
			H_1	H_2	H_3	
1.	Peninsular	27	0.75	–0.34	1.35	22
2.	West Central	62	0.80	–1.17	–2.03	24
3.	Northwest	40	0.84	–0.86	–1.90	29
4.	Central northeast	45	0.74	–0.86	–1.47	14
5.	Northeast	32	0.45	–1.30	–1.06	04

the terrain in each of the NCEP grid boxes) and PCs were considered as attributes to form 52 feature vectors for K-means cluster analysis.

As the exact number of regions is not known *a priori*, the K-Means algorithm was executed by varying the number of clusters from 2 to 25. The resulting clusters were plotted on India map for visual interpretation. Further, cluster validity indices were computed to determine optimal number of clusters. Davies-Bouldin and Calinski-Harabasz indices suggested $K = 20$ as optimal partition, whereas Dunn's index suggested $K = 22$ as optimal partition. The difference in the value of Dunn's index for $K = 20$ and $K = 22$ was found to be insignificant. Consequently, the clusters obtained for $K = 20$ were selected as optimal partition. Figure 3 shows the location of these clusters on India map. The statistical homogeneity of each of these clusters (plausible homogeneous rainfall regions)

was tested using SMR data at 1° × 1° grid points. The values of L-moment based heterogeneity measures computed for each of the clusters are shown in Table 3. It can be seen from the Table that clusters 13 and 17 are acceptably homogeneous, clusters 4, 9 and 14 are possibly homogeneous, whereas the remaining clusters are heterogeneous. Herein it is worth mentioning that the values of heterogeneity measures for several of the delineated clusters are significantly less than that of IMD SMR regions (See Tables 1 and 3).

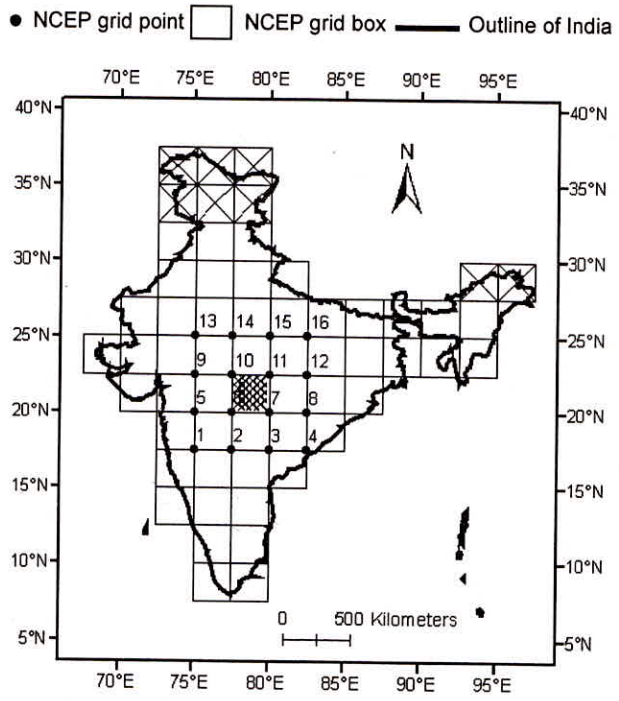

Fig. 2: NCEP grid boxes covering India. The uncrossed boxes were considered for regionalization of summer monsoon rainfall. Atmospheric variables influencing rainfall in the hashed box were considered at 16 NCEP grid points shown as black dots surrounding the box.

The heterogeneous clusters were adjusted to improve their homogeneity following the procedure described earlier. Finally 21 regions were obtained. The adjustment of clusters is necessary because the set of attributes considered for cluster analysis is not exhaustive (i.e. the attributes do not comprise the entire set of causal variables which effect rainfall in the study region). In practice, the current state of knowledge and paucity of data are some of the factors that make it impossible to collate information on exhaustive set of attributes to perform regionalization. Nevertheless, the adjustments should not be substantial if the attributes used for cluster analysis include a reasonable number of causal variables effecting rainfall and if a good approach is used for clustering the data.

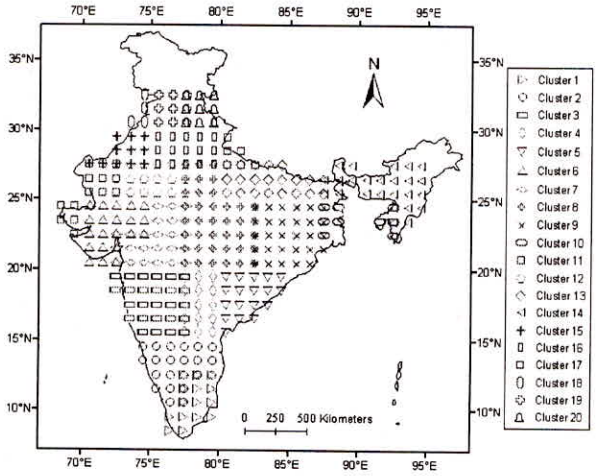

Fig. 3: Clusters in optimal partition obtained using K-means algorithm

Table 3: Characteristics of the Clusters in Optimal Partition Obtained Using K-means Algorithm

Cluster Number	*Cluster Size (in number of IMD grid points)*	*Heterogeneity Measures*		
		H_1	H_2	H_3
1.	15	10.06	3.14	0.80
2.	22	13.36	4.41	1.03
3.	26	4.50	–0.08	–0.84
4.	15	1.30	–1.17	–2.16
5.	18	4.89	0.23	–1.23
6.	21	7.13	0.51	–1.38
7.	24	4.74	–1.19	–1.92
8.	39	2.23	–1.02	–1.34
9.	29	1.73	–2.39	–2.87
10.	9	4.47	–0.80	–1.31
11.	14	6.04	3.16	0.53
12.	18	8.04	2.13	–0.48
13.	21	–0.53	–1.39	–1.28
14.	26	1.62	–1.26	–1.46
15.	11	2.46	–0.05	–1.06
16.	15	9.32	2.31	0.35
17.	6	–0.43	–1.85	–1.55
18.	4	2.45	0.86	–0.14
19.	9	4.27	0.36	–1.07
20.	9	8.51	6.18	4.55

The adjusted regions are shown in Figure 4 and their characteristics are presented in Table 4. The results show that each of the 21 regions are either acceptably homogeneous or possibly homogeneous. Overall, 8 out of the 297 IMD grid points considered for regionalization were unallocated, as they were eliminated from different regions to improve statistical homogeneity.

Table 4: Characteristics of the Regions Formed by Adjusting Clusters Obtained Using K-means Algorithm

Region Number	Number of IMD Grid Points	Heterogeneity Measures			
		H_1	H_2	H_3	Region Type
1.	14	1.32	0.38	–0.18	Possibly Homogeneous
2.	14	0.45	2.22	1.69	Acceptably Homogeneous
3.	23	1.94	–0.70	–1.05	Possibly Homogeneous
4.	21	0.65	–1.21	–1.87	Acceptably Homogeneous
5.	13	1.54	–0.82	–2.00	Possibly Homogeneous
6.	14	0.96	–0.54	–0.71	Acceptably Homogeneous
7.	22	0.58	–2.11	–2.70	Acceptably Homogeneous
8.	36	0.79	–1.08	–1.36	Acceptably Homogeneous s
9.	30	1.82	–2.58	–2.87	Possibly Homogeneous
10.	7	0.26	–0.91	–0.79	Acceptably Homogeneous
11.	16	1.55	–1.49	–2.44	Possibly Homogeneous
12.	10	–0.09	–1.52	–2.01	Acceptably Homogeneous
13.	21	–0.53	–1.39	–1.28	Acceptably Homogeneous
14.	26	1.62	–1.26	–1.46	Possibly Homogeneous
15.	16	0.89	0.25	–0.36	Acceptably Homogeneous
16.	11	1.84	0.91	0.16	Possibly Homogeneous
17.	7	–0.43	–1.85	–1.55	Acceptably Homogeneous
18.	8	0.61	0.04	–0.06	Acceptably Homogeneous
19.	7	0.91	2.00	2.21	Acceptably Homogeneous
20.	8	1.07	–0.19	–0.45	Possibly Homogeneous
21.	4	0.24	–1.06	–1.62	Acceptably Homogeneous

The L-moment based regional GOF test (Hosking and Wallis, 1997) was used to identify distributions that are suitable to fit rainfall data in each region. Among the distributions accepted at 90% confidence level for each region, the distribution for which the GOF measure is sufficiently close to zero was selected for estimation of regional rainfall quantiles using index-method (Dalrymple, 1960). The growth curve ordinates computed for each of the 21 homogeneous SMR regions delineated in this study are shown in Table 5.

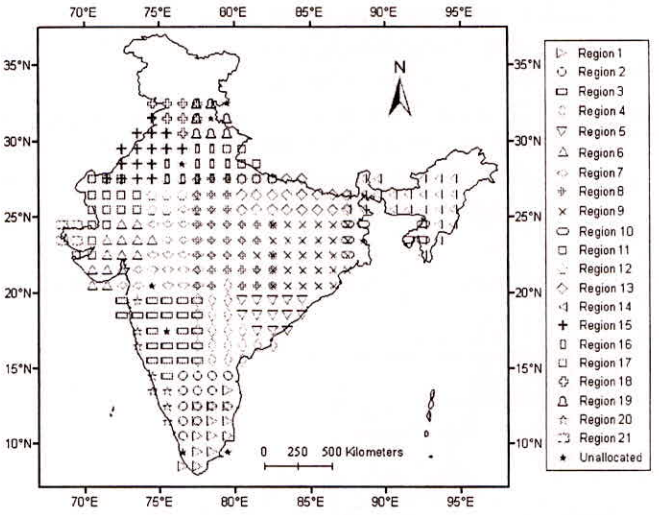

Fig. 4: Homogeneous rainfall regions obtained by adjusting the clusters shown in Figure 3

The foregoing results suggest that the proposed approach to regionalization is efficient in forming homogeneous rainfall regions. Rainfall quantiles estimated for the newly formed SMR regions can be considered as more reliable than those estimated for SMR regions currently in use by IMD.

Plausible implication of climate change on the regions delineated using the proposed approach was assessed by using simulations from CGCM3.1/T63 for the period 2001–2030 for four scenarios namely A1B, A2, B1 and COMMIT. For each of the scenarios, the CGCM3.1/T63 data on fifteen atmospheric variables (mentioned in the subsection "Selection of Attributes"), were re-gridded from 2.81 degree resolution to the NCEP/NCAR grid resolution (2.5 degree) using Grid Analysis and Display System (GrADS; Doty and Kinter, 1993). The spatial domain of the re-gridded CGCM3.1/T63 data influencing future precipitation in any NCEP grid box was considered as 16 NCEP grid points surrounding it, as described earlier. At each of the grid points, the mean monthly values of each of the fifteen re-gridded CGCM3.1/T63 variables were computed for each of the four summer monsoon months (June–September). Thus 960 values (16 grid points × 15 variables × 4 months) were obtained for each of the 52 NCEP grid boxes considered for regionalization. Subsequently, five PCs were extracted from the 960 values along the PDs obtained from NCEP data. The standardized location attributes (latitude, longitude, and average elevation of the terrain in each of the NCEP grid boxes)

Table 5: Growth Curve Ordinates for New Homogeneous SMR Regions Formed Over India. *R* Denotes Region Number, *Z* Represents Goodness-Of-Fit (GOF) Statistic. A Distribution is not Rejected by the GOF Test at 90% Confidence Level if $|Z| \leq 1.64$. GLO: Generalized Logistic; GEV: Generalized Extreme Value; GPA: Generalized Pareto; GNO: Generalized Normal (also known as three-parameter log-normal, LN3); PE3: Pearson Type 3

R	*Z*	*Distribution*	*Nonexceedence Probability*						
			0.50	*0.80*	*0.90*	*0.98*	*0.99*	*0.998*	*0.999*
1.	0.16	GNO	0.344	1.364	1.644	2.241	2.489	3.062	3.310
2.	–0.86	GEV	0.359	1.354	1.639	2.266	2.530	3.137	3.396
3.	–0.28	GNO	0.479	1.287	1.494	1.922	2.095	2.487	2.654
4.	–0.42	GNO	0.490	1.279	1.475	1.873	2.032	2.388	2.538
5.	1.24	GLO	0.598	1.193	1.322	1.610	1.738	2.052	2.196
6.	–0.44	GEV	0.354	1.357	1.627	2.193	2.420	2.922	3.128
7.	1.17	GLO	0.542	1.224	1.402	1.843	2.057	2.636	2.925
8.	–	WAKEBY	0.546	1.216	1.382	1.723	1.853	2.121	2.223
9.	0.35	GLO	0.628	1.181	1.315	1.635	1.785	2.174	2.363
10.	–0.94	GLO	0.640	1.175	1.311	1.640	1.798	2.216	2.423
11.	0.41	GEV	0.212	1.434	1.796	2.611	2.963	3.790	4.153
12.	–0.02	GNO	0.472	1.285	1.469	1.828	1.967	2.268	2.392
13.	–0.24	GLO	0.547	1.221	1.400	1.846	2.064	2.658	2.957
14.	0.06	GLO	0.657	1.167	1.298	1.619	1.775	2.188	2.394
15.	0.41	GNO	0.281	1.399	1.709	2.370	2.645	3.283	3.560
16.	–1.00	GNO	0.490	1.278	1.466	1.843	1.992	2.320	2.457
17.	–0.59	GLO	0.540	1.224	1.389	1.781	1.965	2.441	2.671
18.	–0.30	GNO	0.497	1.272	1.450	1.799	1.935	2.232	2.354
19.	0.17	GLO	0.416	1.284	1.490	1.977	2.203	2.785	3.063
20.	0.71	GLO	0.623	1.184	1.323	1.656	1.814	2.229	2.431
21.	0.51	GPO	0.094	1.584	2.242	3.760	4.409	5.905	6.545

Note: Wakeby is selected if none of the other distributions considered for GOF test are accepted. 'Z' is not given for Wakeby distribution

Table 6: Optimum Number of Clusters Suggested by Validity Indices for Different Climate Change Scenarios. The Selected Optimal Number of Clusters is shown in Bold Font

Scenario	*Davis-Bouldin Index*	*Dunn's Index*	*Calinski-Harabasz Index*
COMMIT	**22**	**22**	**22**
A1B	25	**21**	**21**
A2	25	**22**	**22**
B1	25	**24**	**24**

and PCs obtained using regridded data were considered as attributes to form 52 feature vectors for K-means cluster analysis. Table 6 shows optimal number of clusters determined for each of the scenarios using the three cluster validity indices, and Figure 5 shows the location of the same on India map. Comparison of Figures 4 and 5 indicate that the future changes projected for the homogeneous SMR regions are insignificant.

SUMMARY AND CONCLUDING REMARKS

The traditional approaches to regionalization of precipitation are based on statistics computed from the observed precipitation, rather than attributes effecting hydro-meteorology in a region. Therefore independent validation of the delineated regions for homogeneity in precipitation is not possible. To circumvent this it is proposed to form regions using large scale atmospheric variables and location attributes, so that independent validation of regions is possible with precipitation data. The effectiveness of the proposed approach is demonstrated through application to SMR data of India for delineating homogeneous regions.

The SMR regions that are in use by IMD were found to be statistically heterogeneous. Subsequently, the regions were adjusted to improve their homogeneity. The number of sites that had to be eliminated from the regions for improving their statistical homogeneity was found to be excessive, indicating that the IMD SMR regions are not useful as precursors to derive homogeneous SMR regions. Following this, new SMR regions were delineated using the proposed methodology.

Large scale atmospheric variables influencing precipitation in the study region were identified. Feature vectors prepared using standardized location attributes and principal components extracted from the selected atmospheric variables were clustered using K-means algorithm to arrive at clusters (plausible homogeneous regions). The optimal number of clusters in the data was identified as 20 using cluster validity indices. These clusters were subsequently adjusted to arrive at 21 acceptably/possibly homogeneous SMR regions. The results suggested that the proposed approach to regionalization is efficient in forming homogeneous rainfall regions.

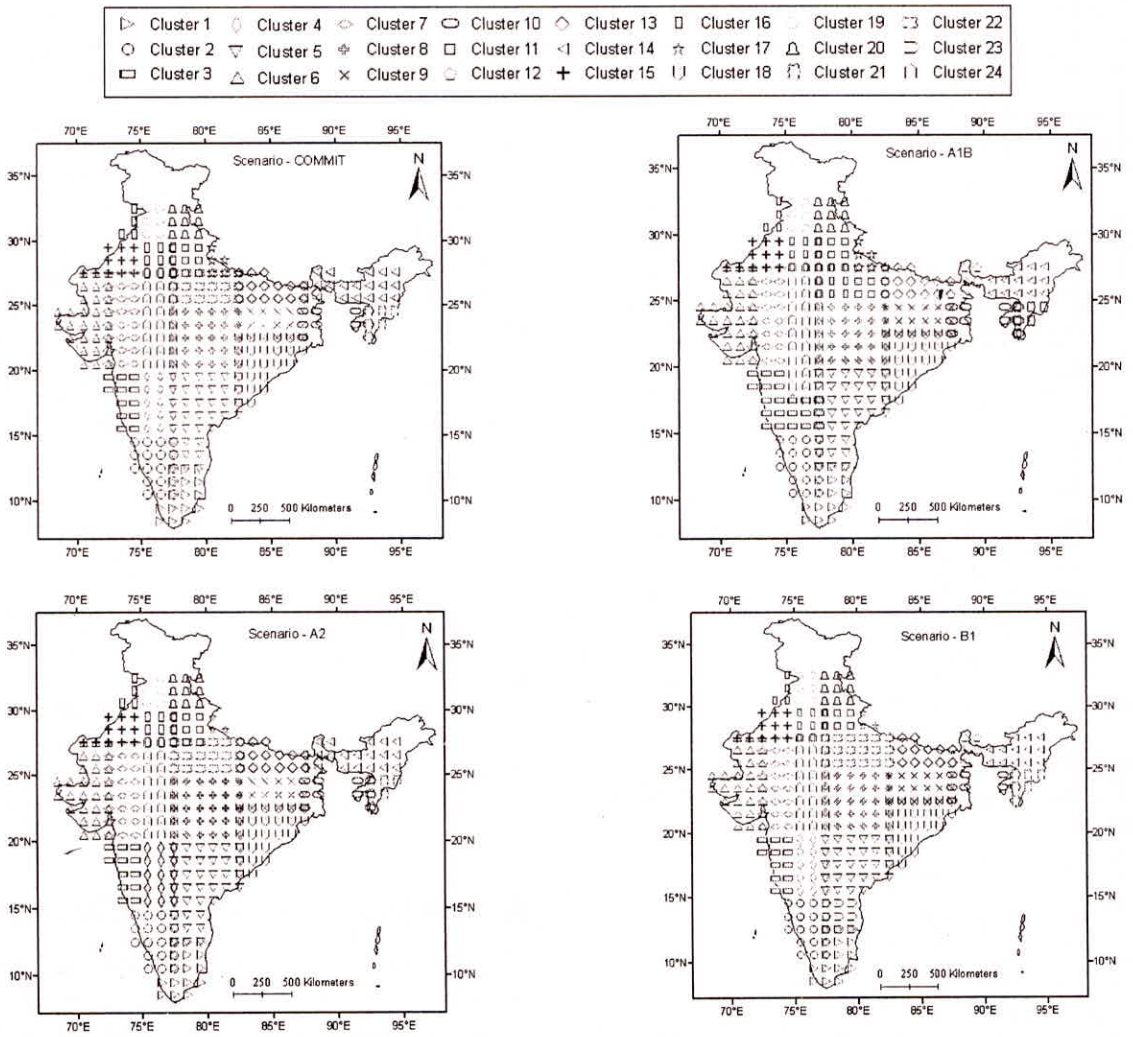

Fig. 5: Plausible homogeneous SMR regions projected using CGCM3.1/T63 data for the period 2001–2030 for four scenarios namely A1B, A2, B1 and COMMIT

Plausible implication of climate change on the regions delineated using the proposed methodology was assessed by using simulations from CGCM3.1/T63 for the period 2001–2030 for four scenarios namely A1B, A2, B1 and COMMIT. The future changes projected for the homogeneous SMR regions are found to be insignificant.

REFERENCES

Barring, L. (1987). "Spatial patterns of daily rainfall in central Kenya: application of principal component analysis and spatial correlation", *Journal of Climatology*, 7(3), 267–290.

Bedi, H.S. and Bindra, M.M.S. (1980). "Principal components of monsoon rainfall", *Tellus*, 32, 296–298.

Calinski, R.B. and Harabasz, J. (1974). "A dendrite method for cluster analysis", *Communications in Statistics*, 3, 1–27.

Dalrymple, T. (1960). "Flood frequency analysis", *Water Supply Paper* 1543-A, US Geological survey, Reston, VA.

Davies, D.L. and Bouldin, D.W. (1979). "A cluster separation measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224–227.

Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S. and Mirnia, M. (2004). "Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods", *Journal of Hydrology*, 297, 109–123.

Doty, B. and Kinter, J.L. III. (1993). "The Grid Analysis and Display System (GrADS)": a desktop tool for earth science visualization. American Geophysical Union 1993 Fall Meeting, San Fransico, CA, 6–10 December.

Dunn, J.C. (1973). "Fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters", *Journal of Cybernetics*, 3(3), 32–57.

Gadgil, S., Yadumani and Joshi, N.V. (1993). "Coherent rainfall zones of the Indian region", *International Journal of Climatology*, 13, 547–566.

Guttman, N.B. (1993). "The use of L-Moments in the determination of regional precipitation climates", *Journal of Climate*, 6, 2309–2325.

Halkidi, M., Batistakis, Y. and Vazirgiannis M. (2001). "On clustering validation techniques", *Journal of Intelligent Information systems*, 17(2/3). 107–145.

Hosking, J.R.M. Wallis, J.R. (1997). "Regional frequency analysis: an approach based on L-moments", Cambridge University Press, New York, USA.

Iyengar, R.N. and Basak, P. (1994). "Regionalization of Indian monsoon rainfall and long-term variability signals", Royal Meteorological Society, 0899-8418/94/101095-1114.

Jackson, I.J. (1974). "Inter-station rainfall correlation under tropical conditions", *Catena*, 1, 235–256.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y. , Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. and Joseph, D. (1996). *The NCEP/NCAR 40 year reanalysis project*, Bulletin of the American Meteorological Society, 77(3), 437–471.

Kulkarni, A., Kripalani, R.H., Singh, S.V. (1992). "Classification of summer monsoon rainfall patterns over India", *International Journal of Climatology*, 12, 269–280.

McQueen, J. (1967). "Some methods for classification and analysis of multivariate observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, Berkeley, CA, 1, 281–297.

Obregón, G.O. and Nobre, C.A. (2006). "Rainfall regionalization on the Amazon basin", *Proceedings of 8 ICSHMO*, Foz do Iguaçu, Brazil, INPE, 1149–1152.

Pakhira, M.K., Bandyopadhyay, S. and Maulik, U. (2004). "Validity index for crisp and fuzzy clusters", *Pattern Recognition*, 37, 487–501.

Parthasarathy, B., Rupa Kumar, K. and Munot, A.A. (1993). "Homogeneous Indian Monsoon rainfall: Variability and prediction", *Journal of Earth System Science*, 102, 121–155.

Rajeevan, M., Bhate, J., Kale, J.D. and Lal, B. (2005). "Development of a high resolution daily gridded rainfall data for the Indian Region (Version 2)", India Meteorological Department, India, *Met. Monograph Climatology No. 22/2005.*

Rajeevan, M., Bhate, J., Kale, J.D. and Lal, B. (2006). "High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells", *Current Science*, 91(3), 296–306.

Rao, A.R. and Srinivas, V.V. (2008). "Regionalization of Watersheds—an Approach Based on Cluster Analysis", Springer Publishers.

Sharma, C., Roy, J., Kumar, R.K., Chadha, D.K., Singh, R.N. , Saheb, S.P. and Mitra, A.P. (2003). "Impacts of climate change on water resources in India", *Proceedings of workshop on climate change and water resources in South Asia,* Kathmandu, Nepal, Asianics Agro Dev International, Islamabad, 61–90.

Singh, K.K. and Singh, S.V. (1996). "Space time variation and regionalization of seasonal and monthly summer monsoon rainfall of the sub-Himalayan region and Gangetic plains of India", *Climate Research*, 6, 251–262.

Srinivas, V.V., Tripathi, S., Rao, A.R. and Govindaraju, R.S. (2008). "Regional flood frequency analysis by combined self-organizing feature map and fuzzy clustering", *Journal of hydrology*, 348, 148–166.

Sumner, G.N. (1983). "Seasonal changes in the distribution of rainfall over the great dividing range: general trends", *Austr. Met.* Mag, 31, 121–130.