New Approaches to Hydrological Prediction in Data-sparse Regions (Proc. of Symposium HS.2 at the Joint IAHS & IAH Convention, Hyderabad, India, September 2009). IAHS Publ. 333, 2009.

Regional frequency analysis of annual precipitation in data-sparse regions using large-scale atmospheric variables

P. SATYANARAYANA & V. V. SRINIVAS

Department of Civil Engineering, Indian Institute of Science, 560 012 Bangalore, India
vvs@civil.iisc.ernet.in

Abstract Regional precipitation frequency analysis (RPFA) is widely used for predicting precipitation quantiles at target sites in data-sparse areas. The RPFA involves fitting a frequency distribution to information pooled at target site from a region (group of similar sites). Conventional approaches to RPFA use precipitation statistics as attributes to form regions. Therefore, sufficient number of sites with contemporaneous data is required to form meaningful regions. This requirement cannot be met in data sparse areas. To address this issue, an approach is presented in this paper. Large-scale atmospheric variables affecting precipitation in the study area, location parameters (latitude, longitude and altitude) and seasonality of precipitation are suggested as attributes to form regions using fuzzy cluster analysis, and precipitation statistics are suggested for use in validating the delineated regions for homogeneity. Results from application to India indicate that the approach is effective for RPFA in data-sparse areas.

Key words regionalization; precipitation frequency analysis; fuzzy cluster analysis; large-scale atmospheric variables; India

INTRODUCTION

Effective prediction of the amount and frequency of precipitation is necessary for a wide range of applications that include design of irrigation projects, and investigating the frequency and spatial distribution of meteorological droughts. Regional precipitation frequency analysis (RPFA) is a potential alternative to at-site frequency analysis models for predicting precipitation quantiles at target sites in data-sparse areas. The RPFA involves fitting a frequency distribution to information pooled at the target site from a region (group of sites having similar precipitation characteristics). The process of identifying the region is called regionalization.

The past four decades have witnessed extensive research in regionalization of precipitation. The approaches that have been developed include elementary linkage analysis, spatial correlation analysis, common factor analysis, empirical orthogonal function analysis, principal component analysis (PCA), cluster analysis, and PCA in association with cluster analysis. Discussion on these approaches can be found in Satyanarayana & Srinivas (2008). In several past studies, the attributes that have often been used for regionalization of precipitation include statistics computed from precipitation data (e.g. mean, cross-correlation of annual/seasonal/monthly/daily precipitation; mean number of wet days). Therefore, a sufficient number of sites with contemporaneous data is required to form meaningful regions, and it may not be possible to arrive at effective regions in data-sparse areas. Moreover, validation of the identified regions for homogeneity is not possible, since the use of the same precipitation statistics to form regions and subsequently to test their homogeneity, is meaningless. Furthermore, it is not possible to identify regions for ungauged sites, because the attributes (precipitation statistics that are necessary to identify region) are unknown.

To address the aforementioned issues, an approach is proposed by Satyanarayana & Srinivas (2008), wherein large-scale atmospheric variables (LSAV) affecting precipitation in a region and location attributes (latitude, longitude and altitude) are suggested as features for regionalization by K-means cluster analysis, and the delineated regions are independently validated for homogeneity using the *L*-statistics of the observed precipitation. The present study extends the idea of the previous study (Satyanarayana & Srinivas, 2008) to frequency analysis of annual precipitation in data-sparse areas using fuzzy cluster analysis. The LSAV affecting precipitation at sites in the study area, location attributes and seasonality (i.e. average time of occurrence) of maximum monthly precipitation are suggested as features for regionalization using fuzzy c-means (FCM) cluster analysis. Effectiveness of the proposed approach is illustrated through application to India.

METHODOLOGY

Suppose there are N sites in the study area. Identify n attributes affecting precipitation at each site, such as LSAV affecting precipitation, location attributes and seasonality of precipitation at the sites. Let $\boldsymbol{y}_i = \left[y_{1i}, \ldots, y_{ji}, \ldots, y_{ni} \right]' \in \Re^n$, denote ith feature vector depicting ith site in n-dimensional attribute space. y_{ji} is jth attribute. Rescale the feature vector as:

$$x_{ji} = \frac{\left(y_{ji} - \bar{y}_j \right)}{\sigma_j} \qquad \text{for } 1 \le i \le N;\ 1 \le j \le n \tag{1}$$

where x_{ji} denotes the rescaled value of y_{ji}; 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 to nullify the differences in their variance and relative magnitude. Use FCM algorithm to partition the matrix $\boldsymbol{X}$ containing rescaled feature vectors into c fuzzy clusters. The objective function and constraints of FCM algorithm are given by equations (2) and (3), respectively:

$$\text{Minimize } J(\boldsymbol{U},\boldsymbol{V}: \boldsymbol{X}) = \sum_{k=1}^{c} \sum_{i=1}^{N} (u_{ik})^{\mu} d^2(\boldsymbol{x}_i, \boldsymbol{v}_k) \tag{2}$$

$$\text{subject to: } \begin{cases} \sum_{k=1}^{c} u_{ik} = 1; & \forall i \in \{1, \ldots, N\} \\ 0 < \sum_{i=1}^{N} u_{ik} < N; & \forall k \in \{1, \ldots, c\} \end{cases} \tag{3}$$

where $u_{ik} \in [0, 1]$ denotes the membership of ith rescaled feature vector x_i in the kth fuzzy cluster; $\boldsymbol{U}$ is the fuzzy partition matrix which contains the membership of each rescaled feature vector in each fuzzy cluster; the parameter $\mu \in (1, \infty)$ refers to the weight exponent for each membership, and is known as fuzzifier; $\boldsymbol{V} = \left(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \ldots \boldsymbol{v}_c \right)$ represents a matrix of centroids of clusters where $\boldsymbol{v}_k$ denotes centroid of kth cluster; $d^2(\boldsymbol{x}_i, \boldsymbol{v}_k)$ is the distance from $\boldsymbol{x}_i$ to $\boldsymbol{v}_k$.

The iterative procedure of FCM algorithm (Bezdek, 1981) is summarized below:

(a) Initialize fuzzy partition matrix $\boldsymbol{U}$ using a random number generator. Let u_{ik}^{init} denote membership of $\boldsymbol{x}_i$ in cluster k.

(b) Adjust the initial memberships u_{ik}^{init} as:

$$u_{ik} = \frac{u_{ik}^{init}}{\sum_{k'=1}^{c} u_{ik}^{init}} \qquad \text{for } 1 \le k \le c, \quad 1 \le i \le N \tag{4}$$

(c) Compute the fuzzy cluster centroid $\boldsymbol{v}_k$ for $k = 1,2, \ldots, c$ as:

$$\boldsymbol{v}_k = \frac{\sum_{i=1}^{N} \left(u_{ik} \right)^{\mu} \boldsymbol{x}_i}{\sum_{i=1}^{N} \left(u_{ik} \right)^{\mu}} \tag{5}$$

(d) Update the fuzzy membership u_{ik} as:

$$u_{ik} = \frac{\left(\frac{1}{d^2\left(\boldsymbol{x}_i, \boldsymbol{v}_k \right)} \right)^{1/(\mu - 1)}}{\sum_{k'=1}^{c} \left(\frac{1}{d^2\left(\boldsymbol{x}_i, \boldsymbol{v}_k \right)} \right)^{1/(\mu - 1)}} \qquad \text{for } 1 \le k \le c, \quad 1 \le i \le N \tag{6}$$

Repeat steps (c) and (d) until change in the value of the memberships between two successive iterations becomes sufficiently small. Obtain different sets of clusters by varying c and μ. Subsequently, determine the optimal set of clusters using Xie-Beni fuzzy cluster validity index (Xie & Beni, 1991), computed as:

$$V_{XB,m}(\boldsymbol{U},\boldsymbol{V}:\boldsymbol{X}) = \frac{\sum_{k=1}^{c}\sum_{i=1}^{N}(u_{ik})^{\mu}\left\|\boldsymbol{v}_k - \boldsymbol{x}_i\right\|^2}{N \min\limits_{l,l\neq k}\left\|\boldsymbol{v}_l - \boldsymbol{v}_k\right\|^2} \tag{7}$$

To form fuzzy clusters, assign each site to cluster(s) in which it has membership greater than or equal to a threshold T_i, computed using equation (8), following Srinivas *et al*. (2008) and Rao & Srinivas (2008):

$$T_i = max\left\{\frac{1}{c}, \frac{1}{2}\left[\max_{1\leq k\leq c}(u_{ik})\right]\right\} \tag{8}$$

The fuzzy clusters identified using the foregoing procedure need to be evaluated for statistical homogeneity using homogeneity tests. In the current study, the Hosking & Wallis (1997) test is considered. The test is based on the idea that in a homogeneous region, all sites are supposed to have the same population L-moment ratios [LMRs: coefficient of L-variation (L-CV), L-skewness and L-kurtosis]. However, their sample L-moment ratios may be different due to sampling variability. In this test, a region k having N_k sites is defined as acceptably homogeneous or possibly heterogeneous or definitely heterogeneous, based on the error induced because of sampling variability. The error is computed using three heterogeneity measures (*HMs*), H_1, H_2 and H_3. A region can be regarded as "acceptably homogeneous" if $HM < 1$, "possibly heterogeneous" 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 & Wallis, 1997). Among the delineated fuzzy clusters, adjust heterogeneous clusters to improve their homogeneity by eliminating sites that are grossly discordant with respect to other sites. Identify the grossly discordant site(s) using the discordancy measure of Hosking & Wallis (1997) (not shown due to lack of space). Further, verify if the eliminated sites from a region can be considered as a new region. After adjustments, if a site i belongs to $R'_{(i)}$ of the R regions, the memberships of the site in each of the $R'_{(i)}$ regions has to be updated using equation (9):

$$u'_{ik} = \frac{u_{ik}}{\sum_{k'=1}^{R'_{(i)}} u_{ik}} \qquad \text{for } 1 \leq k \leq c, \quad 1 \leq i \leq N \tag{9}$$

where $u'_{ik} \in [0, 1]$ denotes the updated membership of the ith site in the kth fuzzy region. If a site belongs to only one fuzzy region, the membership of the site in that region has to be updated to 1, and its membership in all other regions has to be updated to zero.

Prediction of annual precipitation quantiles and performance assessment

Prediction of the annual precipitation quantile of T-years recurrence interval at any site in a region k is based on pooled information from all sites in the region. For this purpose, identify the regional frequency distribution to fit the pooled regional information using L-moment based goodness of fit (GOF) test of Hosking & Wallis (1997). Determine a dimensionless quantile function (known as regional growth curve) for each region. Subsequently, estimate regional precipitation quantile $\hat{P}_i(T)$ at site i for T-year recurrence interval using equation (10), based on the index flood method (Dalrymple, 1960; Srinivas *et al*., 2008).

$$\hat{P}_i(T) = \sum_{k=1}^{R'_{(i)}} u'_{ik}\ \overline{P}_i\ \hat{p}_k(T) \tag{10}$$

where $\overline{P}_i$ is the mean annual precipitation at site i; $\hat{p}_k(T)$ is the growth curve ordinate of region k for T-year recurrence interval. $R'_{(i)}$ is the number of regions in which site i has partial memberships; u'_{ik} is updated membership of ith site in kth fuzzy region. Readers are referred to Hosking & Wallis (1997) for further details on the GOF test and growth curve estimation.

Performance of the proposed method in predicting quantiles of annual precipitation is assessed using two performance measures, namely, the average relative bias (*Average R-bias*) and relative root mean square error (*R-RMSE*). The equations can be found in Satyanarayan & Srinivas (2008). Minimum values of these measures indicate better performance.

CASE STUDY

The study area India lies between latitude 8°4' and 37°6' north, and longitude 68°7′ and 97°25′ east, and has an area of 3 287 263 km^2. It receives average annual precipitation of 117 cm and more than 80% of the annual rainfall during June–September. 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 receives low rainfall.

For the study, high-resolution gridded daily rainfall data for the period 1951–2004 procured from the India Meteorological Department (IMD) (Rajeevan *et al.*, 2005) were considered. Parthasarathy *et al*. (1993) found no systematic trend in the all India precipitation in a study covering the period 1871–1990. The gridded re-analysis data of the monthly mean atmospheric variables (listed in Table 1), which influence precipitation in the study area (Anandhi *et al*., 2008), were extracted from database of the National Centers for Environmental Prediction (NCEP) (Kalnay *et al.*, 1996), for the period 1951 to 2004 from the website http://www.cdc.noaa.gov. 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°. The re-analysis data was re-gridded to 1° × 1° IMD grid resolution using Grid Analysis and Display System (GrADS; Doty & Kinter, 1993). For the analysis, the atmospheric variable at each pressure level was considered as a separate variable. Thus, there are a total of 15 atmospheric variables.

The average elevation of terrain in each of the IMD grid boxes was computed from Shuttle Radar Topography Mission (SRTM) data processed by the Consortium for Spatial Information of the Consultative Group for International Agricultural Research (CGIAR-CSI; http://srtm.csi.cgiar.org).

Table 1 The list of atmospheric variables considered for regionalization.

Site no.	Variable name	Pressure levels (in kPa)
1	Air temperature	92.5, 70, 50, 20
2	Geopotential height	92.5, 50, 20
3	Specific humidity	92.5, 85
4	Zonal wind	92.5, 20
5	Meridional wind	92.5, 20
6	Surface pressure	-
7	Precipitable water	-

RESULTS AND DISCUSSION

To delineate homogeneous annual precipitation regions in India, 294 out of 357 (1° × 1°) IMD grid boxes covering the study area were considered. The discarded 63 grid boxes are in the Himalayan mountain region where some of the pressure levels (e.g. 92.5 kPa) considered for atmospheric variables are not defined at several locations. The spatial domain of the 15 atmospheric variables (listed in Table 1), which influence precipitation in each IMD grid box, was chosen as 25

IMD grid points surrounding it. Further, to reduce the problem of high dimensionality, the mean monthly values of each of the 15 atmospheric variables were computed at each IMD grid point. Thus 4500 values (15 variables × 25 grid points × 12 months) were obtained for each of the 294 IMD grid boxes. Several of the atmospheric variables are correlated and convey similar information. Therefore to avoid redundancy, four principal components (PCs), which are orthogonal and preserve more than 97% of the variance, were extracted from the 4500 values. Seasonality of maximum precipitation in each of the IMD grid boxes was computed using the middle day of a 30-day maximum precipitation in each year. The procedure was based on seasonality measures defined in Burn (1997).

A feature vector representing each IMD grid box was prepared using its four PCs, seasonality of maximum precipitation and location attributes (latitude, longitude, and average elevation of terrain). The 294 feature vectors, thus formed, were partitioned into fuzzy clusters using the FCM algorithm. As the exact number of regions is not known *a priori*, the algorithm was executed by varying the number of clusters c from 2 to 25, with an increment of 1, and the value of fuzzifier μ from 1.1 to 3.0 with an increment of 0.1. The resulting clusters are plotted on the map of India for visual interpretation. Further, the Xie-Beni cluster validity index (equation (7)) was computed to determine optimal partition consisting of plausible homogeneous precipitation regions. The foregoing analysis suggested $c = 20$ and $\mu = 1.9$ as optimal partition.

The statistical homogeneity of each of the 20 clusters in optimal partition was tested by computing *HMs* using annual precipitation data at 1° × 1° grid points in it (Fig. 1(a)). The results showed that six clusters were acceptably homogeneous, one cluster was possibly heterogeneous, and the remaining 13 clusters were definitely heterogeneous. All the heterogeneous clusters were adjusted to improve their homogeneity following the procedure described in the methodology section. Twenty-two regions and 13 unallocated sites were obtained after adjustments (Fig. 1(b)). Among the regions, those numbered 21 and 22 were formed by using sites eliminated from clusters 1 and 20, respectively. The characteristics of the 22 regions are given in Table 2. It can be noted that regions, 8, 13, 18, 21 and 22 are possibly heterogeneous and the remaining 17 regions are acceptably homogeneous.

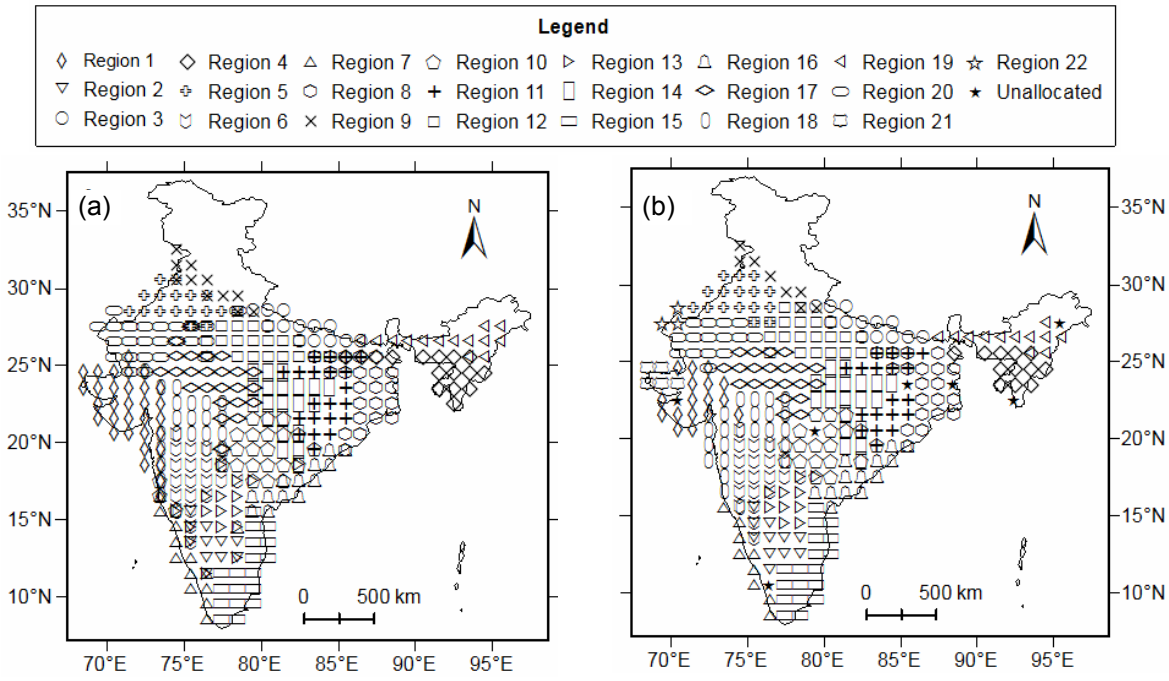

Fig. 1 (a) Annual rainfall clusters in optimal partition; the 1° × 1° IMD grid points in each cluster are shown by a different symbol. (b) Homogeneous annual rainfall regions formed by adjusting the clusters shown in (a).

To predict annual precipitation quantiles at any of the sites in a region, a frequency distribution suitable to fit the pooled regional information was identified using the *L*-moment based regional goodness-of-fit (GOF) test of Hosking & Wallis (1997). For this purpose, the frequency distributions considered were Generalized logistic (GLO), Generalized extreme value (GEV), Generalized Pareto (GPA), Generalized normal (GNO), Pearson type III (PE3), and Wakeby. Among the distributions accepted at the 90% confidence level for each region, the distribution for which the GOF measure is sufficiently close to zero was selected for estimation of growth curve ordinates for use in equation (10). It is found that the Wakeby distribution is suitable to fit pooled information of regions in northeast and central northeast India. The GNO distribution is suitable to fit pooled data in Tamil Nadu and north Uttar Pradesh. The GEV distribution (followed by GNO and PE3 distributions) are found suitable to fit data in Gujarat and Rajasthan. For the rest of India, the GLO distribution is found suitable. The frequency distributions of rainfall vary across India, owing to variation of climate and geographical conditions influencing rainfall.

To assess the potential of the proposed method in predicting quantiles of annual precipitation, *Average R-bias* and *R-RMSE* were computed. For determining at-site estimates of *T*-year precipitation quantiles, GOF of various hydrological distributions to at-site annual precipitation data in the study region was performed by Kolmogorov-Smirnov and Chi-squared tests considering the 90% confidence level. The distributions considered for this purpose were Normal (N), 2-parameter Gamma (G2), GLO, GEV, GPA, GNO and PE3. The method of probability weighted moments was used for parameter estimation. The *Average R-bias* and *R-RMSE* computed are presented in Table 3 for recurrence intervals: 2, 5, 10, 25, 50, 100 and 250-years, for brevity. A negative value of *Average R-bias* indicates that regional estimates of precipitation quantiles are greater than at-site estimates. It can be noted from Table 3 that *Average R-bias* is generally negative, and the values of *Average R-bias* and *R-RMSE* increase with return period. The value of *Average R-bias* varies from +0.13% to –4.46%, whereas the value of *R-RMSE* varies from 2.19% to 13.28%. Small values of *Average R-bias* and *R-RMSE* indicate that the approach is effective in predicting precipitation quantiles.

Table 2 Characteristics of the annual rainfall regions.

CN	*CS*	H_1	*CN*	*CS*	H_1	*CN*	*CS*	H_1	*CN*	*CS*	H_1
1	15	0.98	7	8	–0.05	13	10	1.08	19	12	0.99
2	10	0.42	8	20	1.20	14	19	0.59	20	15	0.68
3	13	0.51	9	8	0.70	15	18	0.02	21	5	1.25
4	10	0.96	10	17	–0.21	16	13	0.19	22	3	–1.85
5	15	0.75	11	19	0.89	17	21	0.85			
6	20	–0.62	12	23	0.85	18	24	1.58			

CN: cluster number, and *CS*: cluster size (in number of IMD grid points).

Table 3 *Average R-bias* and *R-RMSE* computed to assess the performance of the proposed method in predicting annual precipitation quantiles at 1° × 1° IMD grid points.

Performance measure	Return period (years): 2	5	10	25	50	100	250
Average R-bias (%)	0.13	–0.38	–1.03	–1.94	–2.65	–3.39	–4.46
R-RMSE (%)	2.19	5.73	7.35	8.33	9.21	10.65	13.48

SUMMARY AND CONCLUSIONS

The conventional approaches to RPFA use statistics of precipitation as attributes to form regions for pooling information. Consequently, they may not be useful to form meaningful regions in data-sparse areas. Besides this, regions delineated using precipitation statistics cannot be independently validated for homogeneity in precipitation. To alleviate these problems, a fuzzy RPFA approach is proposed. The LSAV, location parameters and seasonality of rainfall are suggested as features for

regionalization using FCM cluster analysis. The seasonality of precipitation in an area can be reliably obtained from local inhabitants, even at ungauged sites. The proposed approach allows independent validation of the identified regions for homogeneity by using statistics computed from the observed precipitation, and it has the ability to form regions even in areas where the raingauge density is sparse. The effectiveness of the proposed approach is illustrated through application to India. Overall, 22 regions are obtained, of which 5 are possibly heterogeneous and the remaining 17 are acceptably homogeneous. Through *L*-moment based analysis it is shown that the approach is effective in predicting precipitation quantiles. The proposed method can be extended to prediction of precipitation quantiles at the daily time scale. This, however, requires identification of attributes influencing daily rainfall for delineation of regions that are homogeneous in frequency distribution of daily rainfall. Research in this direction is under way. Further, a simulated annealing-based FCM algorithm can be used instead of a conventional FCM algorithm to explore the possibility of arriving at global optimal set of regions.

REFERENCES

Anandhi, A., Srinivas, V. V., Nanjundiah, R. S. & Kumar, D. N. (2008) Downscaling precipitation to river basin in India for IPCC SRES scenarios using Support Vector Machine. *Int. J. Climatol*. **28**(3), 401–420.

Bezdek, J. C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York. USA.

Burn, D. H. (1997) Catchment similarity for regional flood frequency analysis using seasonality measures. *J. Hydrol*. **202**, 212–230.

Dalrymple, T. (1960) Flood frequency analysis. *US Geol. Survey Water Supply Paper 1543-A.*

Doty, B. & Kinter, J. L. (1993) The Grid Analysis and Display System (GrADS): a desktop tool for earth science visualization. American Geophysical Union 1993 Fall Meeting, San Francisco, USA.

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

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. & Joseph, D. (1996) The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Met. Soc*. **77**(3), 437–471.

Parthasarathy, B., Kumar, R. K. & Munot, A. A. (1993) Homogeneous Indian monsoon rainfall: variability and prediction. *J. Earth Sys. Sci*. **102**, 121–155.

Rajeevan, M., Bhate, J., Kale, J. D. & 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.

Rao, A. R. & Srinivas, V. V. (2008) *Regionalization of Watersheds – An Approach Based on Cluster Analysis*. Springer Publishers, Germany.

Satyanarayana, P. & Srinivas, V. V. (2008) Regional frequency analysis of precipitation using large-scale atmospheric variables. *J. Geophys. Res*. **113**, D24110, doi:10.1029/2008JD010412.

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

Xie, X. L. & Beni, G. (1991) A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence* **13**(8), 841–847.