

# Spatial and Statistical Clustering Based Regionalization of Precipitation and Trend Identification in Pranhita Catchment, India

Rajashree Bothale<sup>1</sup>, Yashwant Katpatal<sup>2</sup>Scientist, National Remote Sensing Centre, Hyderabad, India<sup>1</sup>Professor, Department of Civil Engineering, Visvesvaraya National Institute of Technology (VNIT), Nagpur, India<sup>2</sup>

**Abstract:** Clustering based regionalisation and the trends in precipitation in the Pranhita catchment, India using the annual precipitation data pertaining to 131 rainfall stations during 1970 - 2011 is reported here. The rainfall data was subjected to various quality checks, homogeneity tests and discordancy measures which identify nonclimatic factors responsible for breaks in the data and the outliers in the dataset. The area was subdivided into homogeneous zones using statistical and spatial clustering. The heterogeneity measure based on coefficient of variations showed that the identified clusters are homogeneous in nature. The change in precipitation has been calculated by trend analysis on the Precipitation Index (PI). Though trend line of the entire study area did not show any variation, a positive trend was observed in Region #1. This clearly indicated that within large homogeneous areas, minor variations in climate are possible and mappable. Thus there is a need to divide large catchments into smaller regions to understand the subtle variations in the climate for further detailed analysis.

**Keywords:** Precipitation trends; Regionalization; ISODATA; Agglomerative Hierarchical Clustering; Homogeneity

## I. INTRODUCTION

The spatial and temporal distribution of precipitation affects the runoff process. A catchment which is a natural physical boundary may not exhibit homogeneous rainfall characteristics. Thus it is required to extract a dominant pattern of variability in the precipitation and identify the trend over a period of time in any catchment. Regionalization of precipitation is necessary for various applications that include meteorological drought analysis and agricultural planning to cope with water shortages that are likely due to low rainfall, forecasting and downscaling of precipitation and land-use planning and management. The regions are delineated for specific purpose and are based on annual, monthly, weekly or daily observations. Thus the regions delineated for different purpose need not be identical (Srinivas, 2013).

Trend analysis of rainfall is necessary for hydrology and climatology to detect the impact of climate variability over space and time. Trend detection can be used to help water managers recognize if the data upon which the design and operation of water resource systems were based are still valid in current conditions. Changes in the hydrological parameters such as precipitation, evaporation and streamflow affect the flow regimes. For microlevel water resources management, it is inevitable to understand the spatial and temporal variation pattern of rainfall.

Pranhita catchment, Godavari basin, India lies in one such large identified homogeneous rainfall zone in Central India. Before studying the precipitation trends, to capture the subtle variations due to climatic variability, the entire study area is regionalized into small homogeneous regions. Our objective was to identify sub regions within study area, study the performance of both statistical and spatial clustering technique on regionalisation of the data in addition to trend identification.

## II. LITERATURE SURVEY

Many researchers worked on the identification of homogeneous rainfall regime over different parts of the World using different approaches, viz., correlation analysis, Principal Component Analysis, Common Factor Analysis, Spectral

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

analysis, hierarchical approach etc. for regionalization of the data (Iyengar & Basak, 1994; Singh & Singh, 1996; Comrie & Glen, 1998 and Saikranthi et al., 2013). Clustering analysis is preferred by many researcher as it identifies the patterns in the data (Puvanewaran & Smithson, 1993; Kansakar et al., 2004; Kysely et al., 2007; Saris et al., 2010; Cretat et al., 2010). To check the homogeneity of the identified regions, L-moments (Hosking & Wallis, 1997) method is widely used by researchers all over the World (Lee & Maeng (2003), Modarres (2010), Ngongondo et al. (2011), Rahman et al.(2013)).

Sonali & Kumar (2013) indicated that parametric, non-parametric, Bayesian, time series and nonparametric methods with resampling approaches were mainly used in trend detections studies for different hydrologic and climatic variables. Slope based tests (Sen’s robust slope estimator, linear regression) and rank based tests(Mann – Kendall test, Spearman rank correlation) are used for trend detection. Trends were estimated for different durations of annual extreme rainfall using the regional average Mann–Kendall S trend test by Adamowski & Bougadis (2003). Yavuz & Erdogan(2012) analysed monthly and annual precipitation data over Turkey and worked out spatial precipitation trends.

Regionalisation of precipitation has been attempted by many researchers in India. Goswami et al. (2006) showed a significant decreasing trend in the frequency of moderate events over Central India using data from 1951 to 2000. Satyanarayana and Srinivas (2008) found that the summer monsoon rainfall regions that are currently in use by Indian Meteorological Department (IMD) are heterogeneous. They also divided entire country into 25 homogeneous annual rainfall regions using large-scale atmospheric variables (Satyanarayana and Srinivas, 2011). Azad et al. (2010) have divided India into 10 spectrally homogeneous regions which cut across or be subsets of homogeneous-rainfall zones defined earlier by various workers based on different criteria. The biggest spectrally homogeneous region is region 7 which encompasses 7 previously identified regions including Vidarbha, Telangana, Marathawara etc. Saikranthi et al. (2013) used correlation analysis techniques and identified 26 regions using annual, seasonal southwest monsoon and northeast monsoon. All these researchers worked with entire India data and divided the country into homogeneous zones.

### III. STUDY AREA

Pranhita catchment ( $75^{\circ}50'$  to  $80^{\circ}58'$  E,  $18^{\circ}10'$  to  $21^{\circ}55'$  N & elevation range 98 m to 1116 m) in Godavari basin of India lies in the Eastern Maharashtra and covers 1,08,352sq km(Figure 1). The rainfall characteristic of this region is governed by its location in the central part of the country. Annual precipitation in the area is 1071 mm and the mean monsoonal precipitation is 938 mm which is 88% of the total precipitation.

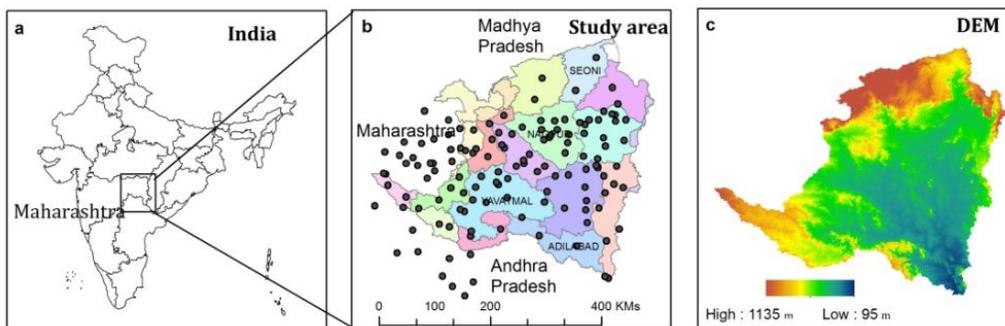


Figure 1: Study area with a) Location map b) Districts and rainfall stations c) Digital Elevation Model (DEM) data from Shuttle Radar Topographic Mission (SRTM).

### IV. DATA USED AND QUALITY CHECKS

#### 3.1 Data used

Yearly precipitation data (Precipitation total) pertaining to 131 stations (Figure 1) is obtained from Indian Meteorological Department (IMD). Data pertaining to recent years (2007 - 2011) was acquired from Agriculture

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

Department of Government of Maharashtra (<http://www.mahaagri.gov.in/rainfall/index.asp>). Maximum record length of data available for the study area is 42 years pertaining to the period 1970 – 2011.

### 3.2 Quality checks

Basic quality checks and visual graphical checks were carried out on the data. Any sudden jump in the rainfall was compared with the annual values of other nearby stations. Spurious values were rechecked from the website and compared with the normal values of the stations which ranged between 700 and 1600 mm in the study area. Visual graphical checks helped in identifying the outliers. Doubtful observations were checked by plotting the time series of rainfall from near by stations and stations with maximum correlation.

Homogeneity tests identify the break in the data. Non climatic factors often make the data unrepresentative of the actual climate variation and might bias the study’s conclusion. For example, a break in the data could result from instrument recalibration and a linear trend could result from a gradual but constant degradation of the sensor (Costa & Soares, 2009). A number of homogeneity tests are suggested by Peterson *et al.* (1998) to find the stability of the time series across time as opposed to a trend. Five different homogeneity tests were carried out at 5% significant level in the present analysis. The annual data series pertaining to 131 stations was subjected to Mann Kendall test (MKT) (Mann, 1945; Kendall, 1975), Pettitt test (PT) (Pettitt, 1979), the Standard Normal Homogeneity Test (SNHT), (Alexandersson, 1986), Buishand range test (BT) (Buishand, 1982) and Von Neumann ratio test (VNRT) (Von Neumann, 1941). MKT is a non parametric test where the null hypothesis ( $H_0$ ) suggests that there is no trend in the series. The PT allows identifying the time at which the shift occurs. SNHT detects the changes in series of rainfall data. The BT can be used on variables following any type of distribution, but its properties have been particularly studied for the normal case. VNRT is also used to test homogeneity. This test does not allow detecting the time at which the change takes place. Whenever the p value is less than the significance level  $\alpha = 0.05$ ,  $H_0$  is rejected. Rejecting  $H_0$  indicates that there is a trend or break in the data depending upon the type of test used. On rejecting the null hypothesis, the result is said to be statistically significant. MKT rejected 12 stations, SNHT rejected 11 stations, PT rejected 15 stations, VNRT rejected 7 stations and BT rejected 27 stations. Based on the combined analysis of all the five tests, 14 stations which failed the hypothesis for three or more tests were rejected. Table 1 shows the homogeneity tests results for 14 rejected stations.

**Table 1: Results of the homogeneity tests for annual rainfall data pertaining to the reject stations**

Stations	P value (two tailed test)					Rejected by number of tests
	MKT	PT	SNHT	BT	VNRT	
Arni	0.036	0.011	0.005	0.01	0.385	4
Biloli	0.016	0.011	0.222	0.048	0.027	4
Chamorshi	0.049	0.027	0.044	0.379	0.584	3
Chorkhamari	0.03	0.049	0.73	0.049	0.664	3
Dhanora	0.001	0.0001	0.0001	0.0001	0.245	4
Dharni	0.001	0.001	0.009	0.002	0.07	4
Etapalli	0.025	0.003	0.068	0.005	0.236	3
Hingna	0.025	0.017	0.033	0.044	0.092	4
Khairee	0.001	0.00	0.001	0.001	0.218	4
Nalsare	0.141	0.039	0.04	0.026	0.457	3
Pathri	0.16	0.025	0.083	0.04	0.019	3
Teosa	0.025	0.011	0.038	0.01	0.038	5
Umarkhed	0.109	0.014	0.037	0.015	0.784	3
Warud	0.008	0.012	0.011	0.006	0.088	4

Figure 2 shows the time series of annual rainfall plot for 6 stations showing the break in the trend of data detected through homogeneity tests.

Out of remaining 117 stations, 29 stations which cleared all the homogeneity tests and have 42 years data are selected as reference stations. Spatial distribution of stations with length of time series, reference stations and rejected stations are illustrated in Figure 3a. The reference stations are well distributed over the study area. Number of stations included in different analysis year is shown in Figure 3b. Valid data was available for minimum 81 stations in the year 1970 and maximum 115 stations in the year 1994.

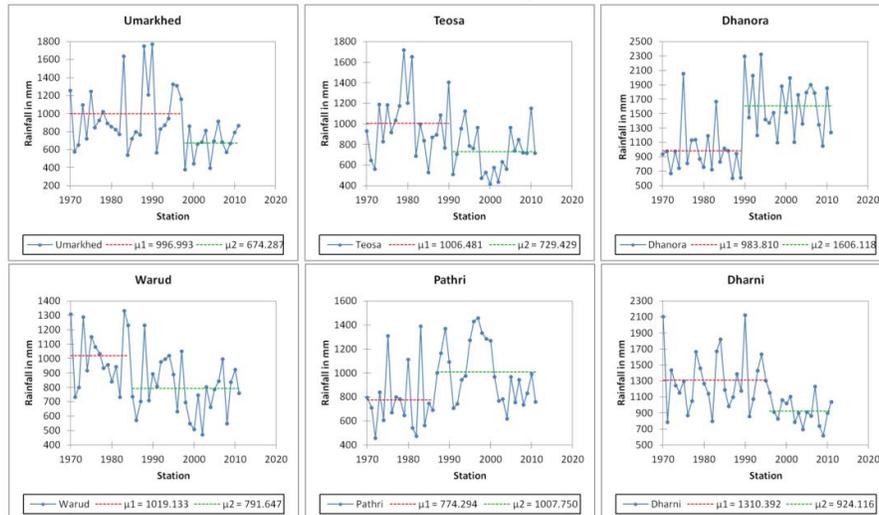


Figure 2: Plot of annual rainfall showing abrupt changes in the rainfall pattern

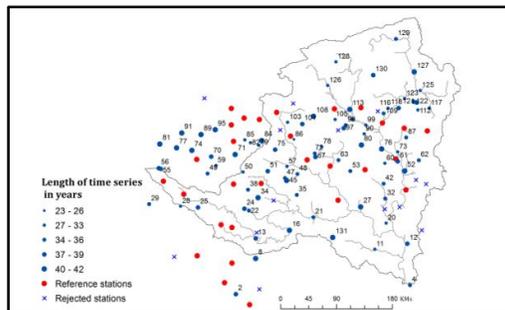


Figure 3a: Reference stations, rejected stations and remaining stations with dots scaled as per length of time series.

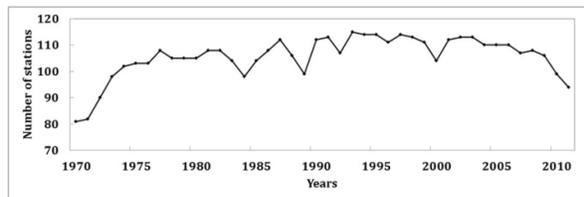


Figure 3b: Number of stations included in each analysis year

## V. METHODOLOGY

### 4.1 Analysis of missing data

Like any other data series, the present dataset has many missing values as visible in the Figure 3b. It is impractical to exclude the data altogether as this often leaves very few data points for analysis in the statistical domain. There are different methods available to calculate the missing data. Presti *et al.* (2010) applied four regressive methods to treat missing data in Candalaro river basin. Presently the correlation coefficient is calculated between target stations and 29 reference stations. Figure 4 shows correlation between two candidate stations with other reference stations. The dots are scaled as per magnitude of correlation. The nearest approach method is the easiest approach to fill the missing data but sometimes nearest station is not highly correlated with the candidate station (Figure 4). For example, Ashti station has maximum correlation with nearest Arvi station but Mukhed station has highest correlation with Hingoli station which is not nearest. Similar trends were seen for other stations also. Hence replacement by nearest method may not be best estimation technique always. No single method exists which can generate missing values with full satisfaction. After analyzing the sample results of missing data filled with nearest neighbor, correlation methods, methodology shown in Table 2 is adopted in the present study.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

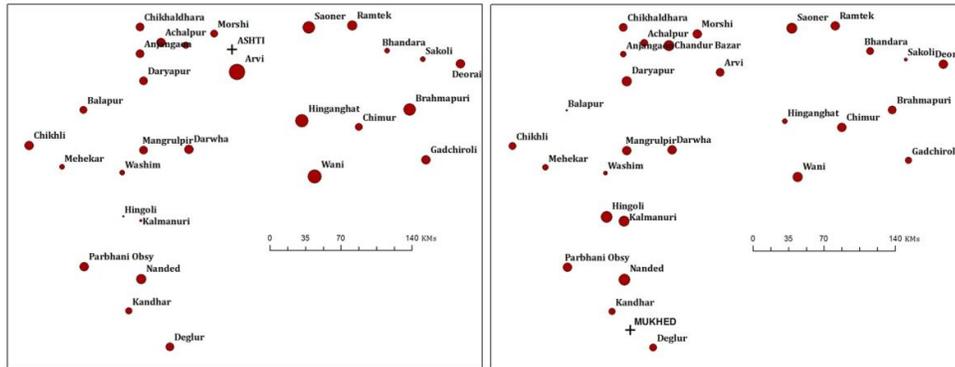


Figure 4: Spatial location of reference stations in relation to candidate station ‘Ashti’ (Left figure) shown by + mark and for station Mukhed (right figure) shown by + mark. The size of dots is scaled as per magnitude of correlation between candidate station and reference station.

**Table 2: Amalgamated methodology to replace missing data**

% data missing	Number of stations	The method followed for filling missing records
No data missing	35	
< 5%	20	Replace with data of the nearest station
5 – 10%	36	Replace with the data of the station with which the candidate station has maximum correlation
10 - 25%	26	Replace with the values calculated using equation of reference station having a maximum correlation with the candidate station which was obtained by a linear equation developed by plotting the reference station rainfall values on the X axis and candidate station rainfall values on the Y axis. For Ashti station maximum correlation is with Arvi station. The equation is $Y = 0.836X + 73.30$ with $r^2$ value of 0.75.
> 25%		Reject the series

As per Johnson (2003) for < 5% of the missing data any method would work well. We replaced the values with simple mean. For 5 – 10% of missing data, data is replaced with the data of the reference station with maximum correlation. When more dataset is missing (>10%), use of a linear equation obtained from station with maximum correlation is likely to give more reliable results.

## 4.2 Regionalization of data

4.2.1 Statistical clustering –To divide the data based on statistical similarity, clustering in the statistical domain was carried out using annual precipitation. The AHC technique has been used in the analysis which is a bottom-up clustering method where clusters have sub-clusters. It creates a nested tree structure from a dissimilarity matrix and goes on combining all the objects (117 stations here) into sub clusters. In the present study there are 42 (years) observations and 117 variables (Stations). The proximity between two objects is measured by measuring similarity or dissimilarity on dendrogram plot. Euclidian distance dissimilarity is used for clustering. Hierarchical clustering returns a more meaningful and subjective division of clusters. For agglomeration, Ward’s method (Ward, 1963) which aggregates two groups based on variance so that within-group inertia increases as little as possible to keep the clusters homogeneous was used. This technique has been successfully employed with rainfall data (Ramos, 2001; Tennant & Hewitson, 2002; Cretat *et al.*, 2010). In Ward’s method of clustering, the distance between two clusters is calculated as the sum of squares between two clusters, added up over all variables. At each cluster generation, the sum of squares is minimized.

4.2.2 Spatial domain clustering –This approach involves the analysis of rainfall stations in the spatial domain. After converting point data into raster format clustering is performed. For any hydrologic analysis and to understand the spatial variability in the data, spatial maps are generated by data interpolation. Interpolation methods are ways to estimate the value at any given point by using values from different nearby points. Australian National University’s

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

Digital Elevation Model Algorithm (ANUDEM) algorithm is used for spatial interpolation of the data. It essentially uses a spline technique which is a deterministic method that creates surfaces from samples based on the extent of similarity or degree of smoothing. The spline surface passes exactly through each sample point. For deriving clusters, ISODATA (Iterative Self Organizing Data Analysis Technique) clustering algorithm is used to classify the dataset into homogenous groups. It uses partitioning clustering algorithm to generate various partitions which are optimized for intra-cluster similarity and inter-cluster dissimilarity. Partitioning clusters result in exact number of desired clusters. The partitions are independent of each other. Elevation of rainfall stations extracted from the SRTM DEM data is also used as input for deriving clusters.

Since the clustering in both the domains was carried out after filling the missing values in the rainfall time series, an attempt was also made to derive these zones without replacing the missing values. 42 rainfall station maps were generated with valid data for each year and then were used to generate grid surfaces as per the method given above.

### 4.2.3 Test for homogeneity of regions

Comparing the regions derived through statistical and spatial clustering was one approach followed in the study. Apart from this approach, L-moment approach was followed to test the homogeneity of the delineated regions. Within the identified regions, to identify the outliers in the rainfall stations, Discordancy measure ( $D_i$ ) for each site was calculated. The discordancy measure assists in identifying those sites whose L-moment ratios are discordant or different relative to L-moment ratios for the collection of sites (Hosking & Wallis, 1997).

L-moments are improvement over conventional product moment statistics for characterizing the shape of a probability distribution and estimating the distribution parameters, particularly for environmental data where sample sizes are commonly small. L-moment ratio measures of scale, skewness and kurtosis are calculated based on the formula given by Hosking and Wallis (1997, pp18-27).

Heterogeneity measure ( $H_1$ ) which is primary indicator for accepting or rejecting an identified region was calculated for all the identified regions. Heterogeneity of a proposed region is computed based on the magnitude of the site-to-site variability in the L-moment ratios relative to the level of variability expected in a homogeneous region. The heterogeneity measures  $H_1$ ,  $H_2$  and  $H_3$  are for the L-moment ratios L-Cv, L-Skewness and L-Kurtosis, respectively. In general it has been found that  $H_1$  measure is better in comparison to  $H_2$  and  $H_3$ . To calculate the value of  $H_1$ , a weighted-average regional value of L-Cv ( $L-Cv^R$ ) is computed from the sample values of L-Cv for the proposed sites in the region, where the weights are based on record length. A weighted-average standard deviation is then computed for the at-site L-Cv values for the collection of sites ( $V_{LCV}$ ). A four-parameter Kappa distribution is fitted using the weighted-average regional values of the L-moment ratios for the sites. 500 computer simulations are then conducted using the fitted 4-parameter Kappa distribution, where each simulation has the same number of sites and record lengths as that for the proposed region. The mean ( $\mu_v$ ) and standard deviation ( $\sigma_v$ ) are computed of the 500 samples of the standard deviation of the at-site samples of L-Cv.  $H_1$  is then computed as:

$$H_1 = \frac{(V_{LCV} - \mu_v)}{\sigma_v} \tag{1}$$

### 4.3 Trend analysis

Efforts are made to study the existence of trends and other temporal patterns in precipitation over the entire study area and also within identified regions. To understand the time variability of precipitation in the present study, Precipitation Index (PI) given by Kraus (1977) is calculated which provides a synthesis of the area coverage behavior of precipitation at time t (Delitala *et al.*, 2000; Raziqi, 2008). The PI is defined as

$$PI_t = \frac{1}{N_t} \sum_{j=1}^{N_t} \frac{(P_{jt} - \mu_j)}{\sigma_j} \tag{2}$$

Where  $P_{jt}$  is the precipitation (mm) at the station j at time t (year),  $\mu_j$  and  $\sigma_j$  are the mean and standard deviation for the precipitation at the same station j, and  $N_t$  is the number of stations available at time t.

This method can be used for any time duration 't' like month, season or year. It considers the number of stations used in the analysis along with mean and standard deviation of precipitation within each region. To examine the trends and

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

long term fluctuations in the PI for individual region, a three year, five year and seven years moving average filter was used. The years were selected looking to return interval of ENSO events in any area with average return interval of four and half years, but can recur as early as two years and as much as ten years apart. MKT was carried out for individual regions and study area as a whole to detect trend.

## VI. RESULTS AND DISCUSSIONS

### 5.1 Data regionalization results

The result of AHC method of statistical clustering is to ultimately combine all the stations into one single cluster as shown in the dendrogram (Figure5). It shows the attributes distances between each pair of merged classes sequence. To avoid crossing lines, the diagram is graphically arranged so that members of each pair of classes to be merged are neighbors in the diagram. The height of the vertical lines indicates the degree of difference or dissimilarity between branches or groups. The distance between two clusters is obtained by substituting the correlation between clusters from one. For highly correlated clusters (Correlation ~1), the intercluster distance will approach zero. So longer the line on dendrogram, greater is the difference between clusters. Based on the separability of clusters shown in the Figure 5, three clusters are retained here. 59 stations are assigned to cluster 1, 33 stations to cluster 2 and 25 stations are assigned to cluster 3.

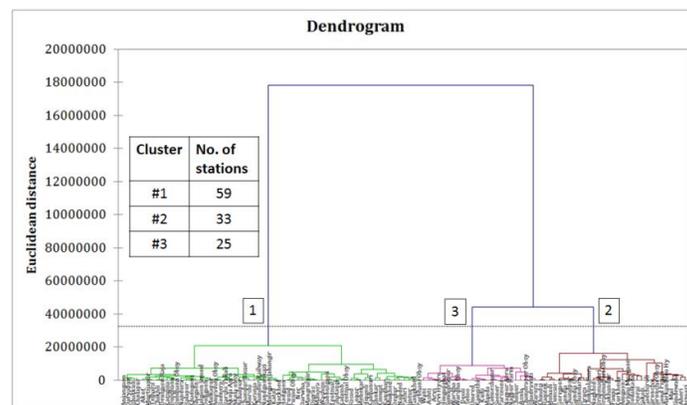


Figure 5: Dendrogram (Cluster tree) for the 117 stations. Number in square on the tree branches represents the cluster. Number of stations in each cluster is shown in the table.

To get the spatial zone of clusters, the Thiessen polygon map was generated for each rainfall station for spatial comparison. The grouped Thiessen polygon represents the rainfall region map for the study.

Figure 6 shows the region map generated using statistical clustering (Figure 6A) and spatial clustering (Figure 6B). The Figure 6C is the result of spatial clustering using original values.

Cluster analysis of rainfall has resulted in distinct spatial patterns. Although only rainfall and elevation values were used in statistical clustering, the resulting zones are continuous in nature similar to the zones obtained by spatial clustering where location attributes were also included apart from elevation and rainfall. The distribution of rainfall regions is similar in all the three cases except in North and South of the study area. This may be attributed to non-uniform distribution of the rain gauges, because a uniform distribution of the rain gauge could stabilize the boundaries. There is a variation in extent of Region #1 (Figure 6B, 6C) caused probably due to the behavior of one station in the South classified into Region #1. Characteristics of the regions obtained through different clustering techniques are given in Table 3.

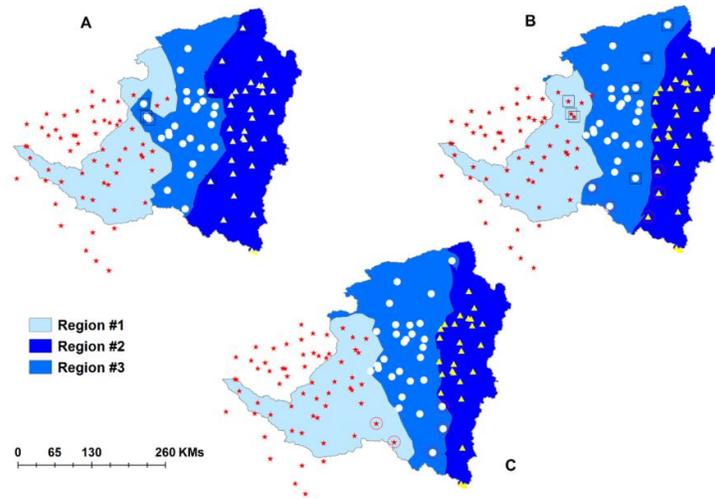


Figure 6: Regions in the study area based on A) Statistical AHC clustering and Thiessen map of cluster points B) ANUDEM interpolation and ISODATA clustering in spatial domain with interpolated rainfall values. C) Spatial clustering with original rainfall values. Black squares represent result variation between statistical and spatial clustering. Red circles represent variation between original and filled values.

All the stations were compared for the region in which they have been classified. The comparison between statistical and spatial technique showed variation in 7 stations (Black squares in Figure 6). This number reduced to five when the comparison was made between original data and filled data (Red circles in Figure 6). Thus the process of data filling before the process of regionalization does not seem to have much impact on the clustering of stations in the present study. As far as actual area under each region is concerned, there is small variation in spatial boundary delineated through each method.

**Table 3: Region characteristics with different methods of clustering**

Parameter	Statistical			Spatial (Filled values)			Spatial (Original values)		
	Region #1	Region #2	Region #3	Region #1	Region #2	Region #3	Region #1	Region #2	Region #3
No of stations	59	33	25	62	26	29	61	28	28
% rainfall station	50%	28%	22%	53%	22%	25%	52%	24%	24%
% Area	29%	42%	29%	29%	30%	41%	34%	27%	39%
Ave rainfall	838 mm	1286 mm	1040 mm	862 mm	1267 mm	1037 mm	898 mm	1295 mm	1030 mm
Location	West	East	Central	West	East	Central	West	East	Central
Rainfall	Moderate	Very high	High	Moderate	Very high	High	Moderate	Very high	High

**5.2 Homogeneity and discordancy tests**

L-moment statistics was calculated for all the groups identified by cluster analysis (Table 4). The results of discordancy test for three groups indicate that there are no discordant stations within the identified regions ( $D_i < 3$ ). From the homogeneity measure ( $H_1$ ) for the groups (Table 4) it is clear that all the three groups are homogeneous with value of  $H_1$  being less than 1.

**5.3 Trend analysis results**

Table 5 shows the values slope for linear trend observed in the zone wise annual PI series with intercept set as zero. From the Table 5 it is clear that a positive trend is observed in Region #1 which is prominent for 5 years moving average.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

**Table 4: L-moments for the identified regions and homogeneity measure**

Regions	Mean of (L-Cv)	Mean of (L-Cs)	Mean of (L-Ck)	Mean of (L-Di)	Homogeneity measure $H_1$
#1	0.134	0.064	0.102	0.17	-0.845
#2	0.116	0.112	0.092	0.96	-0.357
#3	0.119	0.102	0.160	.27	-0.262

**Table 5: Values of slope of the linear trend detected in the identified regions for annual PI series**

Region	Moving average for 3 Years	Moving average for 5 Years	Moving average for 7 years
#1	0.29	0.47	0.27
#2	-.05	-.04	-.02
#3	-.08	-.04	-.03

Figure 7 shows the time variability of PI series for entire study area as a whole and for three identified regions with solid line representing the 5 year moving average time series. Considering the time series it is clear that entire study area is characterised by two dry periods and three wet periods. The year 2005 with PI value -1.2 is the driest in the entire study area. Maximum rainfall was recorded at 1990 with PI value reaching around 1.2. PI series of Region #1 shows that the region is characterized by two wet periods and two deficit periods. It differs from other two regions with major wet period in the years 1988 - 1991 and 1996 - 2001. Years 1982-1987 and 2002 - 2007 were dry periods for Region #1. Region #2 fluctuated around the long term average with alternate wet and dry periods with one major wet period from 1990 - 1996. Region #3 has two major surplus zones and one major deficit zone between 1982 to 1987.

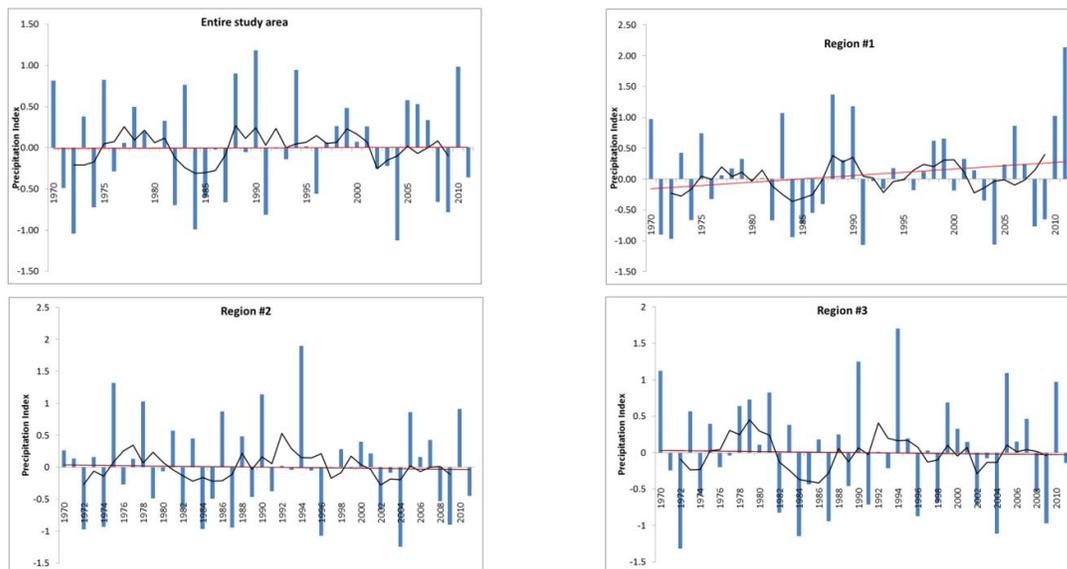


Figure 7: Regional PI series for entire study area, region #1, region #2, region #3 and Region #4. . The continuous thick black line shows the five year moving average. Red line is the trend line generated with linear method.

No trend is observed in the time series of PI for the entire study area. It highlights a positive trend for Region #1. The increasing trend of Region #1 is significant at  $\alpha=0.05$  and  $r^2$  as 0.18. There is no significant trend in Region #2 and Region #3. To confirm the zone wise trend in the rainfall data, MKT was carried out within each zone (Table 6). Results confirm the positive trend in the Region #1 with positive value of S which is significant. There is a non significant negative trend in Region #2 and Region #3. MKT also confirms no trend for entire study area. Thus trend analysis results also indicate that within large homogeneous areas, it is possible to map subtle variations in climate by regionalising the climatic variable.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

**Table 6: Values of Mann Kendall test for different regions**

Region	Kendall's tau	S	p-value (Two-tailed)
Region #1	0.080	69.000	0.464
Region #2	-0.029	-25.000	0.880
Region #3	-0.017	-15.000	0.796
Entire study area	0.032	10.00	0.670

## VII. CONCLUSIONS

This paper regionalises the annual precipitation using statistical and spatial clustering techniques and identifies the trend over Pranhita catchment. After performing quality checks and homogeneity checks (Figure 2), out of 131 stations, 117 stations data was further used in the analysis for regionalisation. Reference stations, rejected stations and remaining stations along with number of stations included every year are given in Figure 3a and Figure 3b. Statistical clustering done through AHC method of clustering divided the stations into three distinct groups (Figure 5). Both the methods grouped the rainfall stations into identical regions and only 7 stations are found to be categorised in different regions (Figure 6). The boundaries thus identified make physical sense. The regions provided a physical boundary to study the impact of climate variability. The regions also represent magnitude regime and three different magnitudes based regions were identified, viz., Moderate rainfall zone, High and very high rainfall zone. Rainfall regions show the clear spatial distribution of rainfall from east to west. However, distribution of rain-gauges is not uniform enough in the study to finalize the spatial regions, particularly in the North and South of the study area.

The procedure used to fill the missing values did not produce any significant difference for a majority of stations (96%). Out of 117 stations, only 5 stations fell in different regions when original and filled datasets were compared. Since most of the time it is difficult to get continuous error free long term quality data set, our study shows that a combined approach used to fill the missing data can also yield satisfactory results. The changing nature of precipitation has been captured by performing trend analysis on the data. Though trend line of the entire study area does not show any variation, a positive trend is observed in Region #1 (Figure 7). This indicates that within large homogeneous areas, minor variations in climate are possible and mappable. Thus there is a need to divide large catchments into small regions to capture the subtle variations in the climate. The study can further extend by considering daily rainfall data and other climate indicators. An effort will be made to understand the impact of El Niño/La Niña events in the identified regions at watershed scale.

## REFERENCES

1. Adamowski, K and Bougadis, J. 2003. Detection of trends in annual extreme rainfall. *Hydrol. Process.* 17, 3547–3560.
2. Alexandersson H. 1986. A homogeneity test applied to precipitation data. *Journal of Climatology*6(6): 661-675.
3. Iyengar R N and Basak P, 1994. Regionalization of Indian monsoon rainfall and long-term variability signals, *International Journal of Climatology*, 14(10), 1095–1114.
4. Azad, S, Vignesh, T. S., & Narasimha, R. 2010. Periodicities in Indian monsoon rainfall over spectrally homogeneous regions. *International Journal of Climatology*, 30(15), 2289-2298.
5. Buishand TA. 1982. Some methods for testing the homogeneity of rainfall data. *Journal of Hydrology*58: 11-27.
6. Comrie AC & Glenn EC. 1998. Principal components-based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability. *Climate research*. Vol. 10: 201–215, 1998
7. Costa AC, & Soares A. 2009. Trends in extreme precipitation indices derived from a daily rainfall database for the South of Portugal. *Int. J. Climatol.*29: 1956–1975. DOI: 10.1002/joc.1834
8. Cretat J, Richard Y, Pohl B Rouault M, Reason C & Fauchereau N. 2010. Recurrent daily rainfall patterns over South Africa and associated dynamics during the core of the austral summer. *Int. J. Climatol.*32(2): 261–273, DOI: 10.1002/joc.2266
9. Delitala AMS, Cesari D, Chessa PA, & Ward MN. 2000. Precipitation over Sardinia Italy during the 1946–1993 rainy seasons and associated large-scale climate variations, *Int. J. Climatol.*20(5): 519–541, 2000.
10. Goswami BN, Venugopal V, Sengupta D, Madhusoodanan MS, Xavier PK, 2006. Increasing trend of extreme rain events over India in a warming environment. *Science*314 (5840): 1442–1445.
11. Hosking J R M and Wallis J R. 1997. *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, New York, USA.
12. Iyengar RN and Basak P. 1994. Regionalization of Indian monsoon rainfall and long term variability signals. *International Journal of Climatology*14(10):1095-1114
13. Johnson, ML. 2003. Lose something? Ways to find your missing data. Houston Center for Quality of Care and Utilization Studies Professional Development Series. 17-09-2003.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

14. Kansakar SR, Hannah DM, Gerrard AJ & Rees G. 2004. Spatial pattern in the precipitation regime of Nepal. *Int. J. Climatol.* **24**(13): 1645–1659.
15. Kendall M. 1975. *Multivariate Analysis*. Charles Griffin & Company, London.
16. Kraus EB. 1977. Subtropical droughts and cross-equatorial energy transports, *Mon. Wea. Rev.*, **105**, 1009-1018.
17. Kysely J, Picek J & Huth R. 2007. Formation of homogeneous regions for regional frequency analysis of extreme precipitation events in the Czech republic. *Stud. Geophys. Geod.* **51**(2): 327–344
18. Lee S and Maeng, SJ. 2003. Frequency analysis of extreme rainfall using L-moment. *Irrigation and Drainage*. Volume 52, Issue 3, pages 219–230, September 2003
19. Mann HB. 1945. Nonparametric tests against trend. *Econometrica*, **13**(3): 245-259.
20. Modarres, Reza and Sarhidi, Ali. 2010. Statistically-based regionalization of rainfall climates of Iran. *Global and Planetary Change* **75** (2011) 67–75.
21. Ngongondo CS, Xu Chong-Yu and Tallaksen LM. 2011. Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches. *Stoch Environ Res Risk Assess* (2011) 25:939–955. DOI 10.1007/s00477-011-0480-x
22. Peterson TC, Easterling DR, Karl TR, et al. 1998. Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol.* **18**(13): 1493–1517
23. Pettitt AN. 1979. A non-parametric approach to the change-point problem. *Appl. Statist.* **28**(2): 126-135.
24. Presti RL, Barca, E & Passarella, G. 2010. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin Italy. *Environ Monit Assess* **160**(1-4):1–22. DOI 10.1007/s10661-008-0653-3
25. Puvaneswaran KM & Smithson PA. 1993. Objective Classification of Homogeneous Rainfall Regimes in Sri Lanka. *Theor. Appl. Climatol.* **48** (2-3): 133-145
26. Rahman, MM, Sarkar, S, ;Najafi, MN and Rai, RK . 2013. Regional Extreme Rainfall Mapping for Bangladesh Using L-Moment Technique. *Journal of hydrologic engineering*. Volume 18, Issue 5(May 2013). /doi/abs/10.1061/(ASCE)HE.1943-5584.0000663
27. Ramos MC. 2001. Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region. *Atmospheric Research* **57** (2): 123–138
28. Raziqi Tayeb. 2008. Investigation of annual precipitation trends in homogeneous precipitation sub-divisions of Western Iran. *BALWOIS 2008 – Ohrid, Republic of Macedonia – 27, 31 May 2008*
29. Saikranthi K, Rao T N, Rajeevan M, and Rao S V B. 2013. Identification and validation of homogeneous rainfall zones in India using correlation analysis, *Journal of Hydrometeorology*, 14(1), 304–317, DOI: 10.1175/JHM-D-12-071.1.
30. SarisFaize, Hannah DM & Eastwood WJ. 2010. Spatial variability of precipitation regimes over Turkey. *Hydrological Sciences Journal* **55** (2): 234-249.
31. Satyanarayana P and Srinivas VV. 2008. Regional frequency analysis of precipitation using large-scale atmospheric variables, *Journal of Geophysical Research—Atmospheres*, 113, (2008) D24110, doi:10.1029/2008JD010412.
32. Satyanarayana P, and Srinivas VV. 2011. Regionalization of precipitation in data sparse areas using large scale atmospheric variables - A fuzzy clustering approach. *Journal of Hydrology*. **405** (3–4): 462–473
33. 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(3), (1996) 251–262.
34. Sonali P, and Nagesh Kumar D. 2013. Review of trend detection methods and their application to detect temperature changes in India. *Journal of Hydrology* 476 (2013) 212–227
35. Srinivas, VV. 2013. Regionalization of Precipitation in India—A Review. *Journal of the Indian Institute of Science*, 93(2), 153-162.
36. Tennant W J & Hewitson BC. 2002. Intra-seasonal rainfall characteristics and their importance to the seasonal prediction problem. *International Journal of Climatology* 22 (9): 1033–1048.
37. Von Neumann J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.* **12** (4): 367-395.
38. Ward JH. 1963. Hierarchical Grouping to optimize an objective function. *Journal of American Statistical Association* **58**(301): 236-244.
39. Yavuz H and Erdogan S. 2012. Spatial analysis of monthly and annual precipitation trends in Turkey. *Water resources management* **26** (3): 609-621. DOI 10.1007/s11269-011-9935-6