

# Regionalization of Precipitation in Andhra Pradesh and Telangana State by Using PCA



Gore Vikas Sudam, Ashutosh Chaturvedi, and K. V. Jayakumar

**Abstract** Precipitation is one of the most important climatic variables for sustaining civilization. It is characterized by extremely high variability in space and time. The magnitude and frequency of precipitation are of great significance in hydrologic and hydraulic design and has wide application in various areas. However, precipitation data is available only in those regions where the rain gauges are installed. The magnitude and frequency of precipitation in ungauged sites can be assessed by grouping areas with similar characteristics. The procedure of grouping of areas having similar behaviour is termed as “regionalization”. Principal component analysis (PCA) which is based on consecutive sieving of stations was initiate for regionalization. PCA used for the reduction of the scale of the data set with minimum reducing of informations. Principal elements (PCs) that are orthogonal to every alternative are derived from inter-station correlation and/or variance matrix of precipitation within the study space. If the primary few leading PCs account for vital % of the full variance, their spatial patterns are analyzed to make undiversified precipitation regions. This approach may involve either plotting the unrotated and/or revolved computer loadings on the map of the study space or representing stations as points in two-dimensional house of leading PCs.

**Keywords** Precipitation · Regionalization · Principal component analysis · Normalization · Eigenvalue · ArcMap

---

G. V. Sudam (✉) · A. Chaturvedi  
Department of Civil Engineering, Alamuri Ratnamala Institute of Engineering and Technology,  
Thane, Maharashtra 421601, India  
e-mail: [vikassgore807@gmail.com](mailto:vikassgore807@gmail.com)

K. V. Jayakumar  
Water and Environmental Division, Department of Civil Engineering, National Institute of  
Technology Warangal, Hanamkonda, India  
e-mail: [kvj@nitw.ac.in](mailto:kvj@nitw.ac.in)

## 1 Introduction

The present study is taken up the following objectives: to identify, analyze, and describe hydrologically homogenous regions using regionalization approaches, and the magnitude and frequency of precipitation in ungauged sites can be assessed by grouping areas with similar characteristics. The meaning of region is a set of sites depicting similar hydrological behaviour, and the meaning of regionalization is the process of identifying regions. Regional frequency analysis is the frequency analysis based on regional information.

The conventional apply in regionalization was to delineate regions as geographically contiguous areas supported physiographic and/or political/administrative boundaries. As such, regions might not have any definite relationship to causal/explanatory variables influencing rain. Regions supported those factors need not be solid in rain. To delineate effective solid regions, a range of regionalization approaches are developed over the past six decades. They embody those supported correlation analysis (e.g. elementary linkage analysis and its variation) principal component analysis (PCA) additionally noted as eigenvector analysis or the connected, empirical orthogonal operate (EOF) analysis, divisor analysis, spectral analysis, cluster analysis, and PCA in association with cluster analysis, stratified approach, and region of influence.

Principal component analysis (PCA), approach of regionalization that is usually noted as eigenvector analysis or empirical orthogonal operate analysis uses the principal component analysis (PCA), PCs that area unit orthogonal to every different area unit derived from inter-station correlation and/or variance matrix of precipitation was enforced by Srinivas [1]. The author had used seasonally and monthly summer monsoon rain (SMR) of ninety stations. Six regions were known to be positively heterogeneous, another 3 were probably heterogeneous once homogeneity testing, and 10 afterwards coherent solid rain regions were known once applying the procedure to SMR. If the primary few leading PCs account for important PCs of the full variance, their spatial patterns area unit analyzed to make solid precipitation regions. This approach may involve either plotting the unrotated and/or turned computer loadings on the map of the study space, or representing stations as points in two-dimensional house of leading PCs. Within the context of regionalization of utmost rain, regional frequency (growth) curves of precipitation extremes area unit created for every of the described regions victimization pooled data from the region or by fitting regional regression relationship between PCs and parameters of the distribution. The developed curves or relationships area unit helpful to attain quantile estimates at sites that area unit ungauged or have inadequate knowledge.

Almazroui et al. [2] used the PCA for determination of the homogenous groups in Saudi Arabian climate were carried out by grouping similar long-term climatological features by using the two variables, precipitation, and air temperature. Five groups were identified: group A (Northern), B (Red Sea Coastal), C (Interior), D (Highland), and E (Southern). Each of the climatic groups exhibited a unique feature (with strong similarities among stations that belong to the same group), especially in terms of

inter-annual fluctuations and the degree of seasonality. This simplified overview of climatic characteristics was beneficial for assisting in economic planning, especially for any semi-arid or arid areas that are naturally vulnerable to climatic variability and extremes. Darand and Daneshvar [3] used the S-mode PCA for seasonal correlation matrix with eight derived variables which had been mapped using GIS. Based on the eight eigenvalues, three factors were extracted between the component and varimax rotation to enhance interpretability of retained PCA scores. The authors considered seasonal precipitation data with eight derived variables and used mean seasonal precipitation amount and its percentage had been interpolated. Using correlation matrix, as opposed to the covariance matrix, the dry grids in the deserts and wet grids in the mountains directly compared.

The recommendations of Miller et al. [4] were to apply the rotated orthogonal based on varimax scores to achieve eigenvector-based regionalization regime of seasonal precipitation. Author considered three rotated components, based on scree plot and eigenvalues over 1.0, the percentage of variance of the data explained as 88.95%. Darand and Daneshvar [3] considered as important of the loading PCs and showed good correlations between the variables and the PCs which had the correlation coefficient  $> 0.7$ . Regression method used to calculate the standardized rotated PC scores. The PCs loading values got PC1 which was explained 47.2% of the total variance, PC2 was explained 23.77% of the total variance, and PC3 was explained 17.96% of the total variance. Munoz-Díaz and Rodrigo [5] said that the purpose of PCA was to identify the most important correlation structures between numbers of variables in order to obtain a description of the major part of the overall variance with few linear combinations based on the original variables. Richman [6] summarized the multiple of techniques which were used for rotations in a PCA in a climatological context. The goal of all these rotational strategies was to obtain a clear pattern of loading which was clearly marked by high loading for others. Typical rotational strategies were VARIMAX and QUARTIMAX. VARIMAX allowed maximization of the variances of each individual record for the various factors.

## 2 Study Area and Data

**Meteorological Data:** The Indian states like Telangana State (TS) and Andhra Pradesh (AP) were selected as the study area as shown in Fig. 1. Daily precipitation data sets were acquired for the period of 113 years ranging from 1900 to 2013 for the State of AP and TS. The data has been collected from Indian Meteorological Department (IMD) [7]. The precipitation data here was arranged in the grid patterns of  $0.25^\circ \times 0.25^\circ$  grid cells. It could be implied that each grid cells will have an area of  $625 \text{ km}^2$  ( $25 \text{ km} \times 25 \text{ km}$ ). The total number of grid cells taken in AP and TS is 433. The daily precipitation data for 433 rain stations, which is shown in Fig. 2, has been used for further analysis. If data is available which is not in required format and degree of resolution, then we have to interpolate the data in required format.

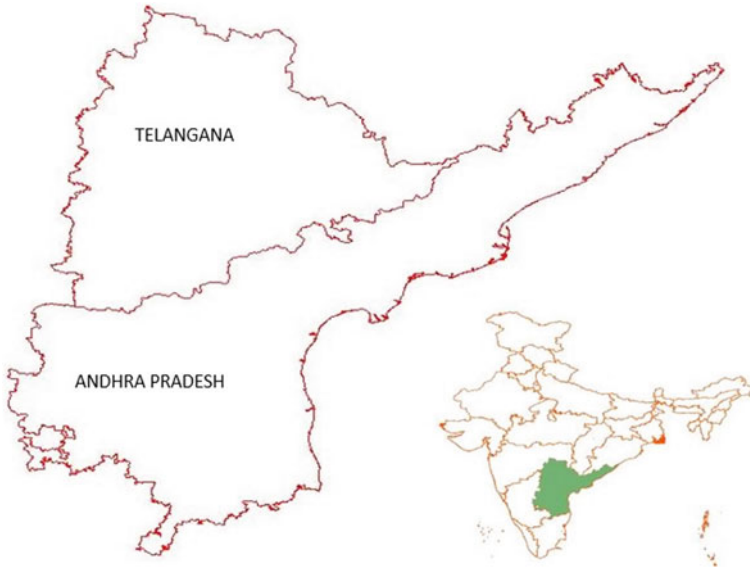


Fig. 1 Study area

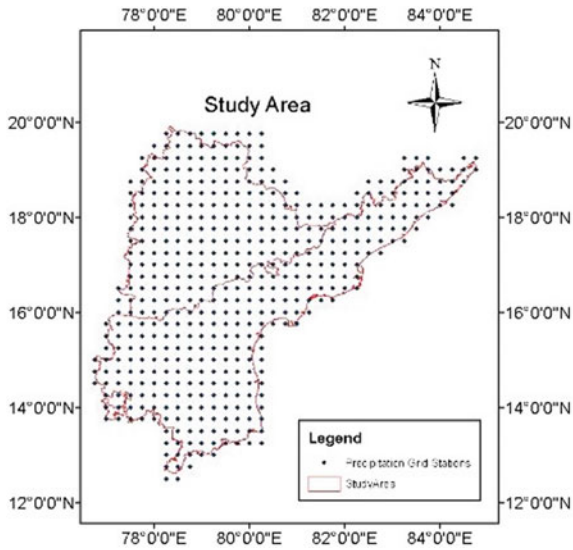


Fig. 2 Study area with  $0.25^\circ \times 0.25^\circ$  grid stations

### 3 Methodology

The methodology adopted in this research is based on the formulation of the research question, literature study, and concept built during the course period. To delineate effective homogeneous regions, a variety of regionalization technique have been developed over the six decades.

Principal component analysis (PCA) is a method that uses a corresponding degree conversion to change the position from a group of observations that may be related to a group of values of inconsistent variables called primary components. If there is a unit detection  $n$  variable  $p$ , then the number of distinct main components is  $\min(n - 1, p)$ . This change is somehow explained by the fact that the main object has the greatest possible variation (i.e. calculating the highest number of possible variations of information), and everything in sequence has the best variation that can be made under the orthogonal limit in the previous parts. The following vectors (each is a combination of linear variations corresponding to the degreed containing the visibility) are the unrelated orthogonal base element. PCA is sensitive to the related growth of initial variability.

PCA was fictional in 1901 by Karl Pearson, as associate degree analogue of the axis theorem in mechanics; it absolutely was later severally developed and named by Harold Hotellingin within the PCA is often used as a tool to search for information analysis and to build prophetic models. PCA is caused by the eigenvalues decay of information variability (or correction) of the matrix or the single price decay of the information matrix, usually as a first step in the formation of information. The results of the PCA area unit are generally defined in terms of object scores, commonly referred to as output scores (variable values used are the same point of information), and loading (weight that all relevant initial variables should be increased to make feature points). If the element scores area of a unit is matched by unit variations, the loading should have a variety of information in them (and that is the size of the eigenvalues). If element scores do not appear to be rated (and therefore contain variations in data), then loading should be measured in units (standard), and this weight area is called eigenvectors; they are cosines for the orthogonal conversion of variables into main components (PCs) or back. PCA is a widely used method of mathematics in climate and climate. When one is faced with a very large database, one attempt to reduce the size of the database, while minimizing any loss of information, the purpose of these mathematical methods is to better understand and interpret the structure of the data. A standard database can be viewed as a variable measurement of  $p$ . For example, if 113-year daily rain data for 433 different recording stations is available, in the study area, there are  $n = 41,273$  views (number of rainy days), of  $p = 433$  variable (rain channel number); there are also other types of data, for example, climate change is measured at  $n$  stations simultaneously. Typically, the  $p$  variables are highly concentrated (especially if the  $p$  channels are locally close). One must therefore choose a variable ( $m < p$ ) that displays all the information contained in the original matrix. This can be done by building new variables that are different

from the original but built on them. The simplicity of the PCA approach lies in this restriction of specific functions of the original variable.

## 4 Results

In the study, area over 75% of the annual precipitation is received during the five rainy months of June–October, only thereby leading to large variations in rainfall on temporal scale. Monthly normal precipitation in the study area is presented in Table 1 shows the monthly normal precipitation data (in centimetre) varying with time of a few selected station.

The first step in building a PCA structure is to build a matrix coefficients matrix using a data set of 433 channels. Thereafter calculated eigenvalues and eigenvectors of the coefficient's matrix for merging. The number of PCs with functional value is determined using eigenvalues. Eigenvectors are made up of a combination of linear symbols in the data set. The coefficients for each original attribute provide an indication of agreement or inconsistency in the actual signals to the new size (PC). The familiarity of each attribute consists of the focus of the statement, subtracting the amount of data from each of its variables so that its approximate value (value) is zero and adjusting each variation to make it equal to 1. The normalized set of data gives the mean as zero and standard deviation as 1. Figure 3 illustrates the variation of normalized monthly normal precipitation data of a few selected stations.

Principal components (PCs) are new flexible elements formed as line combinations or first-generation compounds. This combination is made in such a way that the new variables (i.e. principal components) are incompatible, and most of the details within the original variable are highlighted or pressed into the original elements.

**Table 1** Monthly normal precipitation data (in cm)

Month	Precipitation data (cm)	Percentage	Cumulative percentage
October	217	20	20
September	195	18	38
July	164	15	53
August	164	15	69
June	109	10	79
November	88	8	87
May	41	3	91
December	26	2	94
April	23	2	96
February	17	1	97
March	11	1	99
January	10	1	100

**Fig. 3** Monthly mean of normalized data of few selected stations

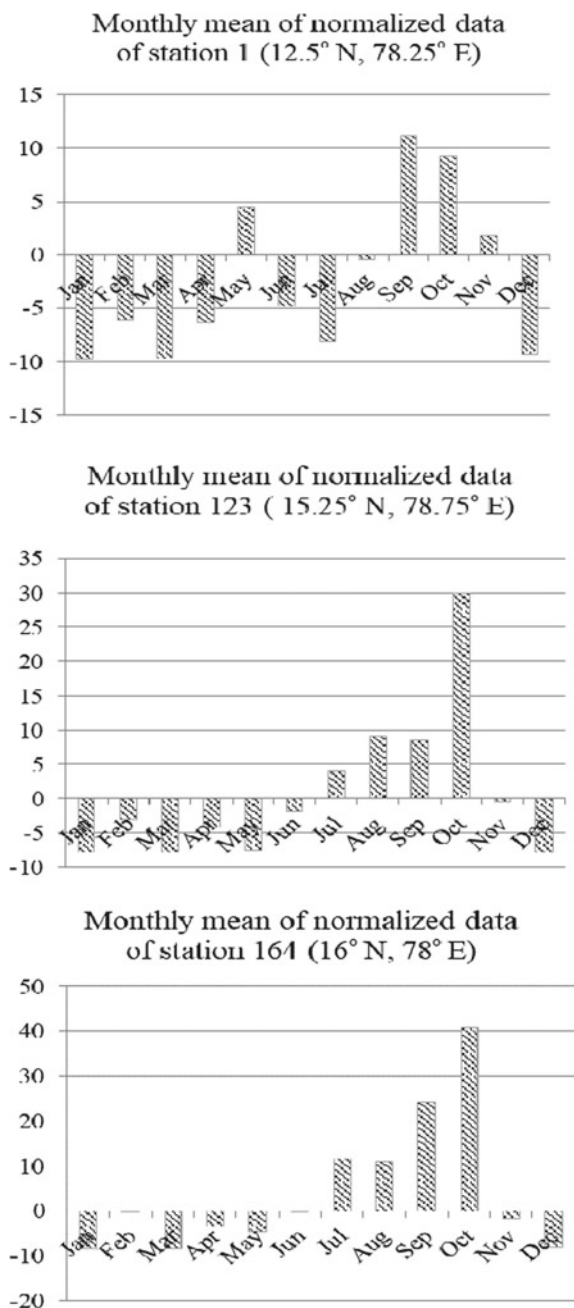
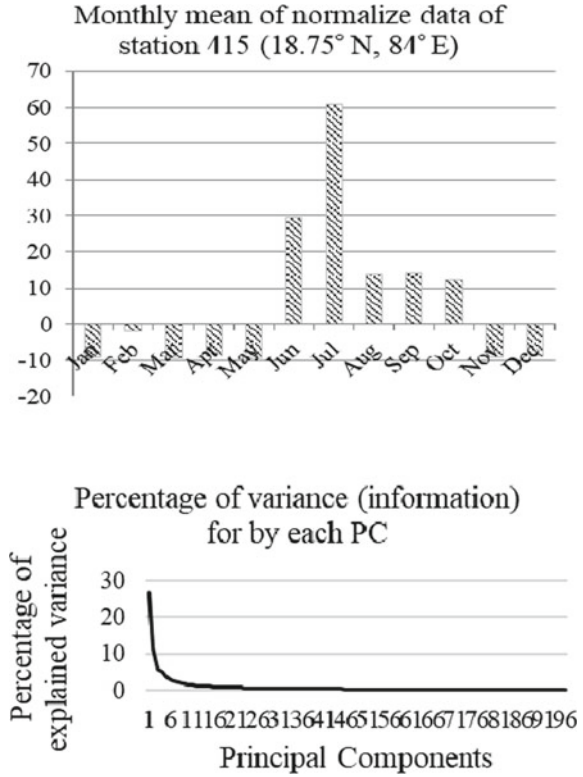


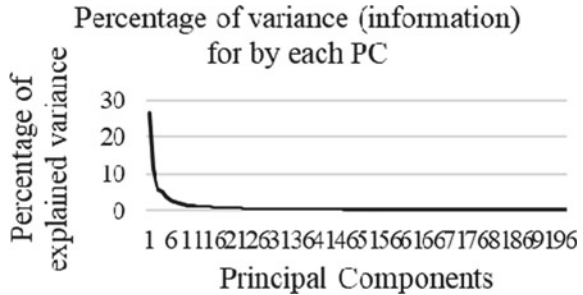
Fig. 3 (continued)



Thus, the view of the 433 data provided 433 key components, but the PCA tries to place the highest data in the first part, and then the remaining data in the second and so on, until something like screens in the scree structure of Fig. 4.

According to the scree plot shown in Fig. 4, it is not clear whether keeping fourteen or ninety-six PCs is better. However, if it is decided to choose to keep only fourteen data sets which have eigenvalue greater than one, then the total variability of the original data set is reduced to 66.67%. One of the aims of PCA is to reduce the dimensions

Fig. 4 Principal components





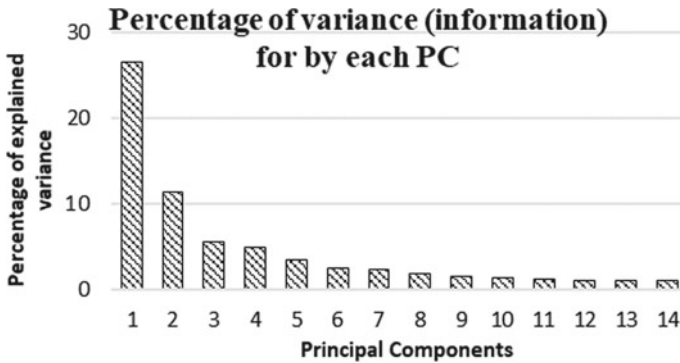


Fig. 5 First fourteen PCs

whiles keeping the maximum variability of the original data set, the number of PCs is too large if the ninety-six PCs are chosen, it is decided to take only first 14 PCs which have eigenvalues greater than one. Figure 5 shows the first 14 principal components, and the Table 2 shows the eigenvalues of them, ordered from the highest to the lowest variability of the data set. The first 14 PCs, which have the eigenvalue greater than one, which will covered the 66.67% of the total variance. For covering the 90% of the total variance, it gave the 97 eigenvalues or PCs. The first PC has the value of 114.95, which explains the 26.54% of the total variance. Similarly, the remaining 13 PCs, which have the values of 49.01, 24.11, 21.45, 15.46, 11.25, 10.47, 8.55, 7.17, 6.37, 5.41, 5.00, 4.80, and 4.58, explain the total variance of 11.32, 5.56, 4.95, 3.57, 2.59, 2.41, 1.97, 1.68, 1.47, 1.25, 1.15, 1.10, and 1.05%, respectively.

Figure 5 shows the first 14 PCs, and the Table 2 shows the values of the PC, explained variance and the cumulative variance of the highest 14 PCs.

The first principal component vector, i.e. the first column of the final data set, is correlated with the standardized original data set and using the correlation data the homogeneous region that has the strong correlation among them is identified.

The loading value of PC1 explains 26.54% of total variance and indicates positive correlation with all precipitation stations. The normalized PC score shows high positive values in the Papikondalu forest region in catchment area of Godavari River, East Godavari district as well as West Godavari district of mouth of the Godavari River, Krishna district in Krishna river catchment, Vishakhapatnam district of AP, Laknavaram forest area eastern part of the TS, and some south-west station of TS. The first factor strongly correlated with that region which has highest precipitation region of AP and TS. The Nellore district of southern part of the coastal AP and Chittoor and Anantapur district of southern part of the Rayalseema region of AP have a lowest PC score. Figure 6 illustrates the regions based on correlation ranges of PC1.

The loading value of PC2 explains 11.32% of the total variance. The standardized PC score shows high efficacious values in Penna river catchment area and Tirumala forest region of AP. The northern and the eastern part of the TS and the northern

**Table 2** Eigenvalues, explained variance, and cumulative variance of PCs

Factor	Eigen values	Explained variance	Cumulative variance
1	114.95	26.54	26.54
2	49.01	11.32	37.86
3	24.11	5.56	43.4
4	21.45	4.95	48.39
5	15.46	3.57	51.96
6	11.25	2.59	54.56
7	10.47	2.41	56.98
8	8.55	1.97	58.95
9	7.17	1.65	60.61
10	6.37	1.47	62.08
11	5.41	1.25	63.33
12	5.00	1.15	64.49
13	4.80	1.10	65.60
14	4.58	1.05	66.66

part of the AP have negative correlation with second factor. Figure 7 illustrates the regions based on correlation ranges of PC2.

The loading value of PC3 explains 5.56% of the total variance. The standardized PC score shows high efficacious values in Hyderabad city area and Nallamala forest area of TS. Third factor negatively correlated with larger area of the northern part and the eastern part of the TS and the Rayalseema region, the Guntur, Prakasam, and Nellore district of the coastal AP. Figure 8 illustrates the regions based on correlation ranges of PC3.

The loading value PC4 explains 4.95% of the total variance. The standardized PC score shows high efficacious values in Tirumala forest region and coastal region near the Pulicat Lake south east part of the AP. The southern part and the central part of the TS and the Rayalseema region, the Guntur, Prakasam, Krishna, East Godavari, Vizianagaram, and Nellore district of the coastal AP have negative correlation with fourth factor. Figure 9 illustrates the regions based on correlation ranges of PC4.

Figure 10 shows the regions of PC5–PC14. The loading value of PC5 explains 3.57% the total variance. The standardized PC score shows zenith positive values in the Nallamalla forest region in the TS and Tirumala mountain ranges, and negative values in the Rayalaseema region, and Northern parts of the TS and AP. The PC6 loading value explains 2.59% of the total variance. The standardized PC score shows zenith positive values in the Palkonda forest ranges and has positive PC score Tirumala mountain ranges and Nallamala forest regions, the northern parts of the TS and AP.

The loading value PC7 explains 2.41% of total variance. The standardized PC score shows high positive values in the southern part Rayalaseema region, the Kondapalli forest region of the AP. It has positive values in the catchment area of the Musi

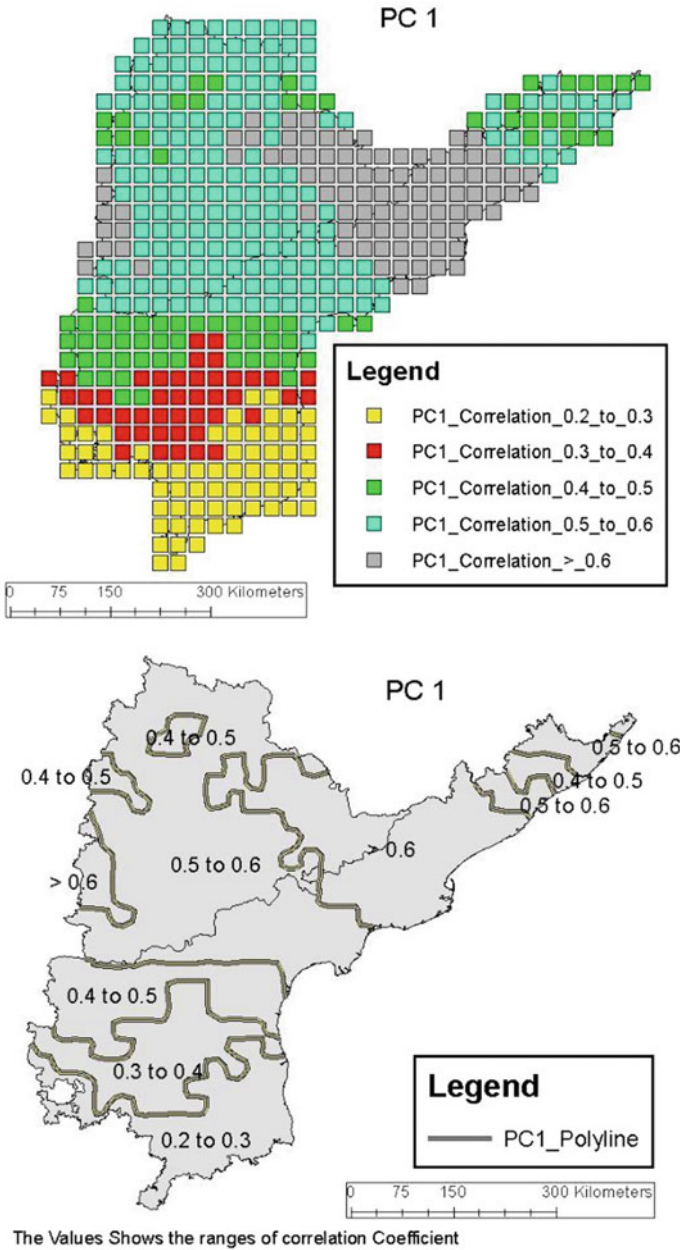
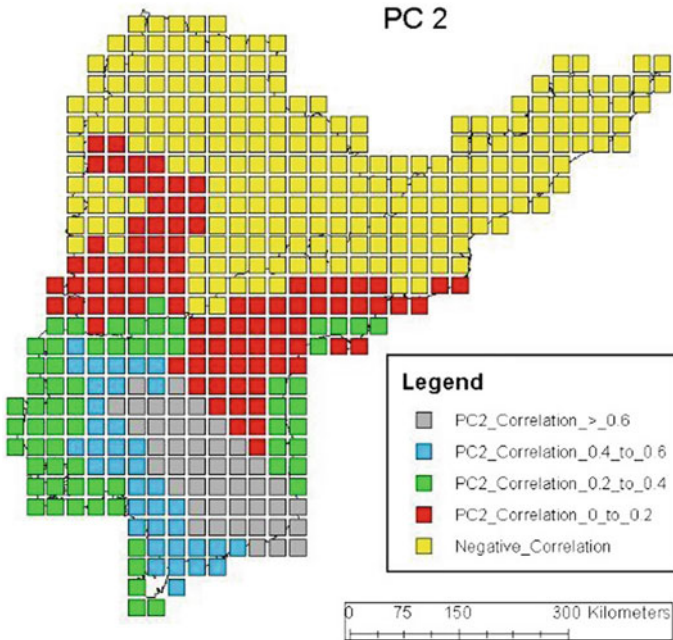


Fig. 6 Regions of first principal component



**Fig. 7** Regions of second principal component

River, the mouth of the Godavari River, southern parts of the AP and TS. It has negative values in the northern part of the Rayalaseema region, catchment area of the Penna River, and the northern parts of the AP and TS.

The loading value PC8 explains 1.97% of total variance. The standardized PC score shows high positive values in the North Godavari district and Cumbun forest range in AP. It has the positive values in Tirumala mountain ranges and Nallamala forest regions and has negative values in southern part of the Rayalaseema region, north part of AP, and most part of the central Telangana. The loading value PC9 explains 1.65% of total variance. The standardized PC score shows high positive values in the Krishna River basin in southern part of the TS. The loading value PC10 explains 1.47% of total variance. The standardized PC score shows high positive values in the northern parts of the AP and TS and the southern part of the Rayalaseema region. The loading value PC11 explains 1.25% of total variance. The standardized PC score shows high positive values in the southern part of the coastal AP. The loading value PC12 explains 1.15% of total variance. The standardized PC score shows high positive values in the Nallamala forest region, southern parts of the Rayalaseema region, and the northern parts of the TS. The loading value PC13 explains 1.10% of total variance. The standardized PC score shows high positive values in the Koundinya forest region in southern part of the Rayalaseema. The loading value PC14 explains 1.05% of total variance. The standardized PC score shows high positive values in the Koundinya forest region in southern part of the

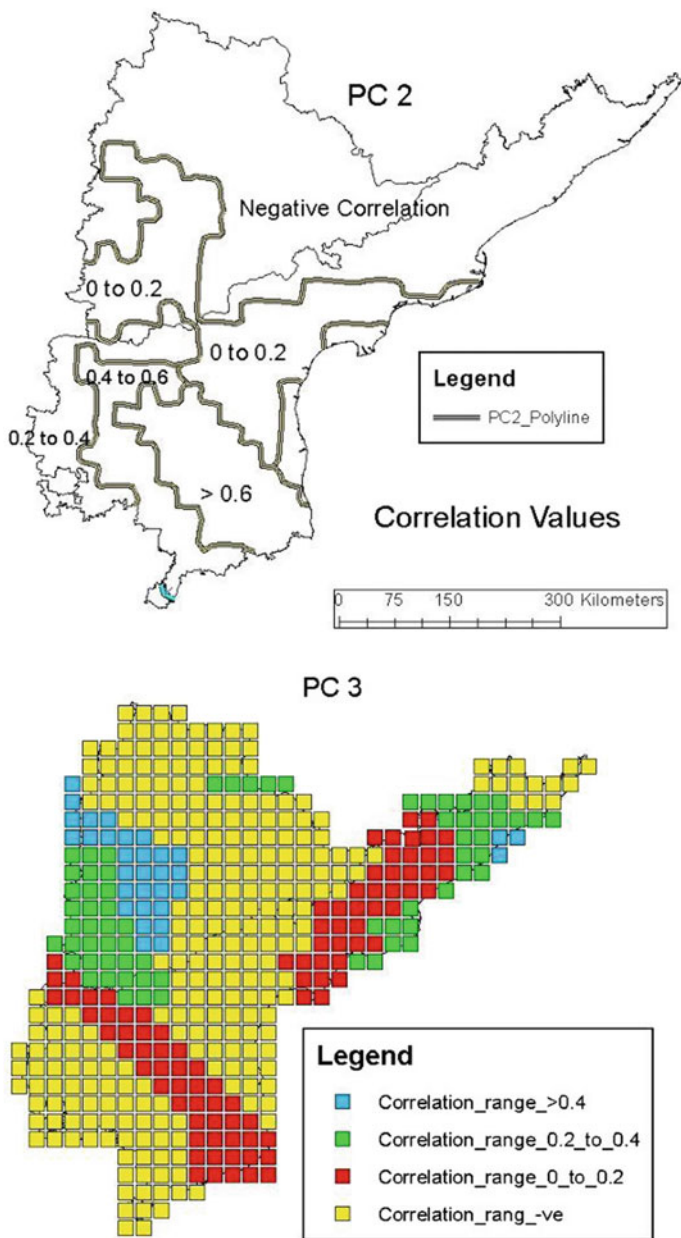


Fig. 8 Regions of third principal component

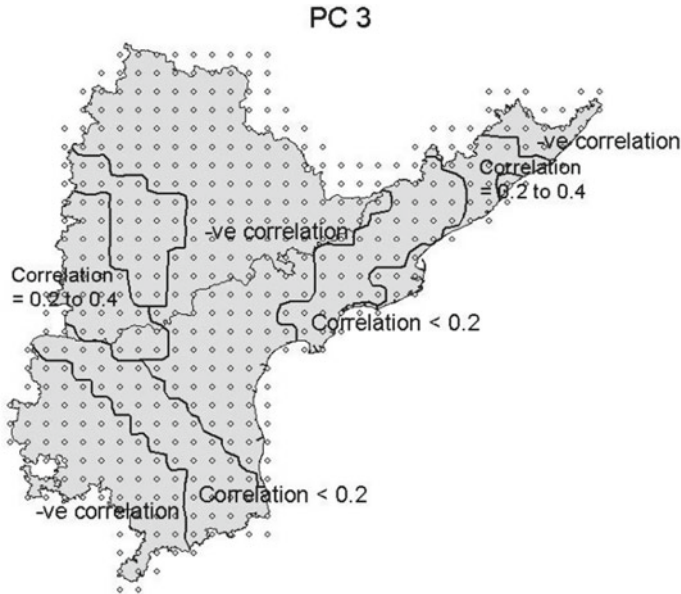


Fig. 8 (continued)

Rayalaseema and has positive PC score in the Nallamala Forest region, Laknavaram forest regions.

**Summary:** In this study, regionalization is used to hydrologic and hydraulic design, meteorological drought analysis as well as planning of agricultural area to cope water shortages that are likely due to low rains.

In the present study, precipitation regions are regionalized for Andhra Pradesh and Telangana States based on meteorological daily precipitation spatial resolution data of  $0.25^\circ \times 0.25^\circ$  of 433 grid stations of 113 years during 1901 to 2013, provided by IMD by using ArcMap. Using principal component analysis, the 14 principal components extracted based on the standardized precipitation data which have eigenvalue  $> 1$ . The 14 PCs explain 66.67% of the total variance. The standardized PC score of the first PC explained 26.54% of the total variance, and highest PC score was obtained in the area of where the highest precipitation occurs in the study area.

## 5 Conclusions

The main objective of the study was to identify the homogeneous regions those that has similar characteristics of precipitation using regionalization approaches with the help of ArcMap. The states in study area are too largely extended to be considered as a similar characteristics (homogeneous) region.



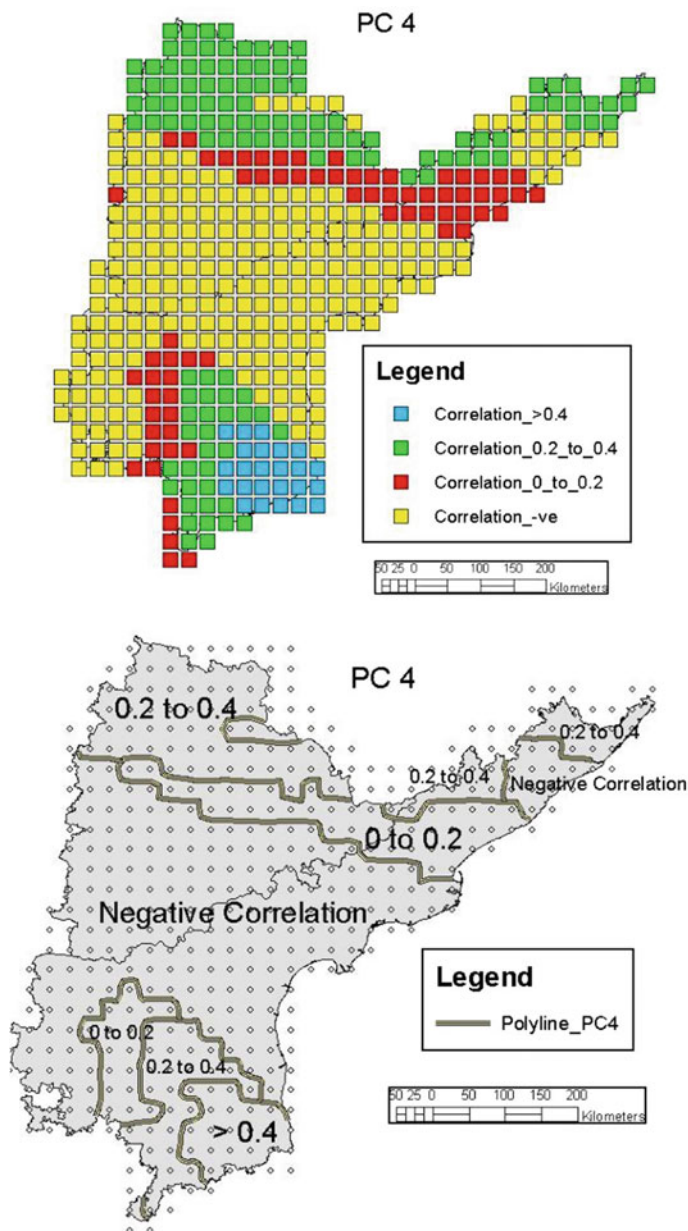
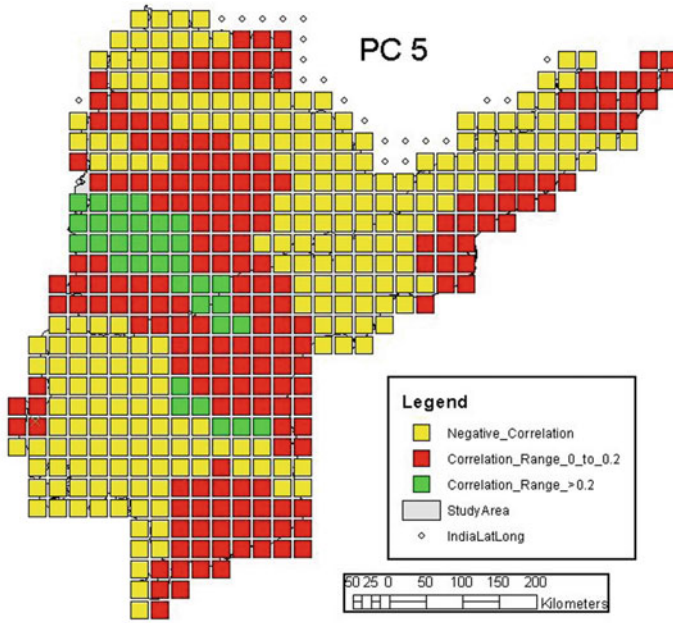
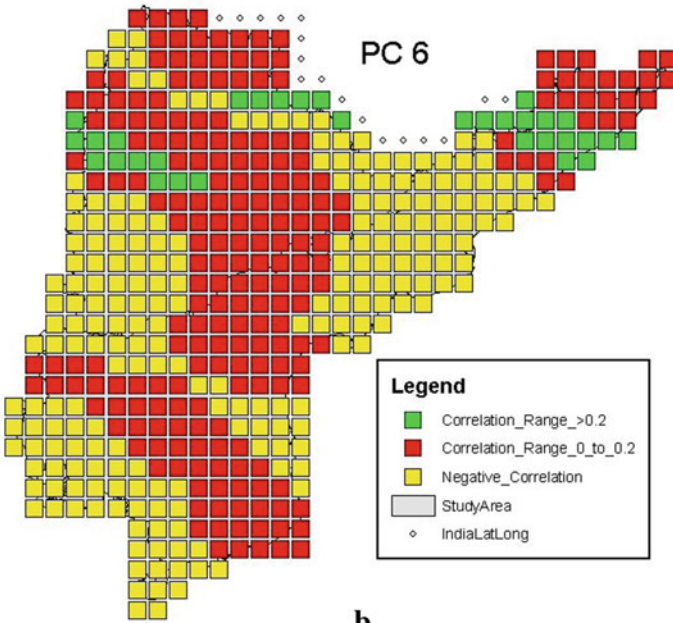


Fig. 9 Regions of fourth principal component



**a**



**b**

**Fig. 10** Regions of 5th–14th PCs



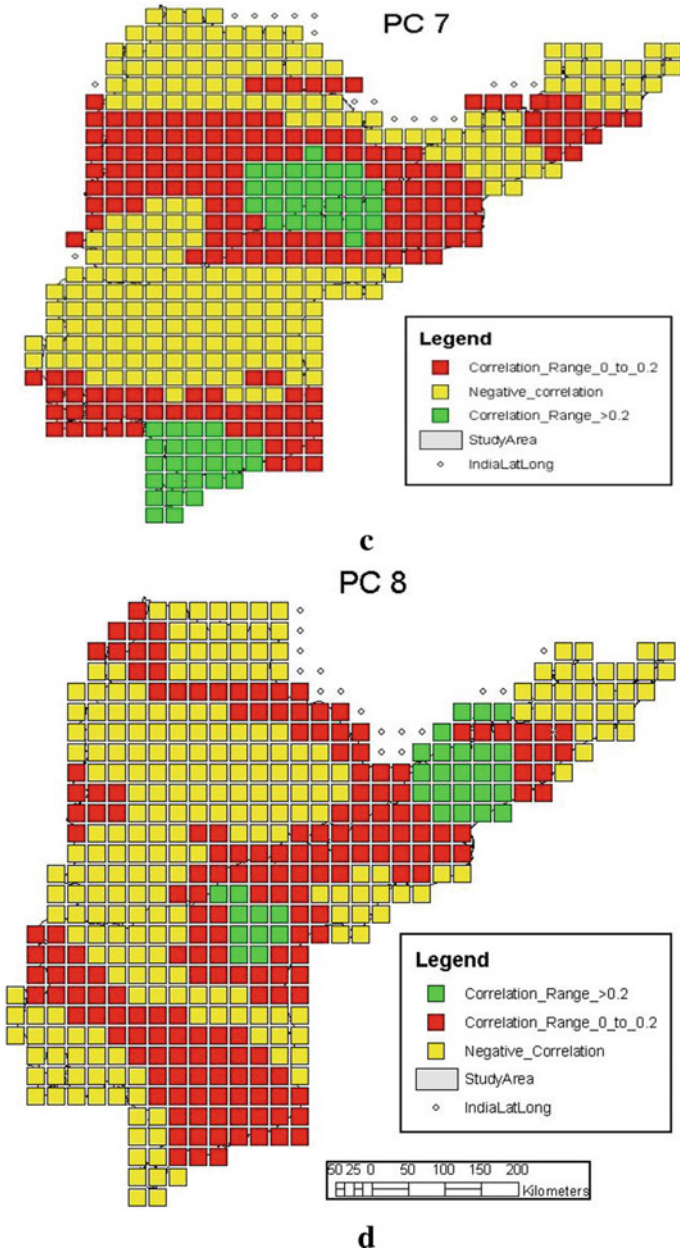
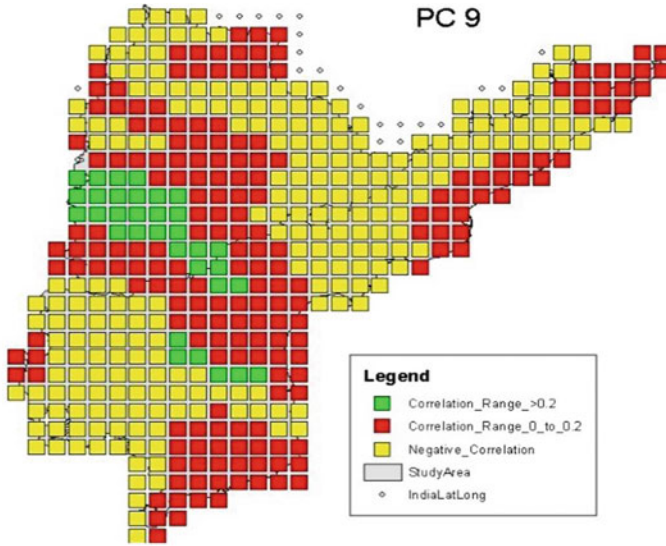
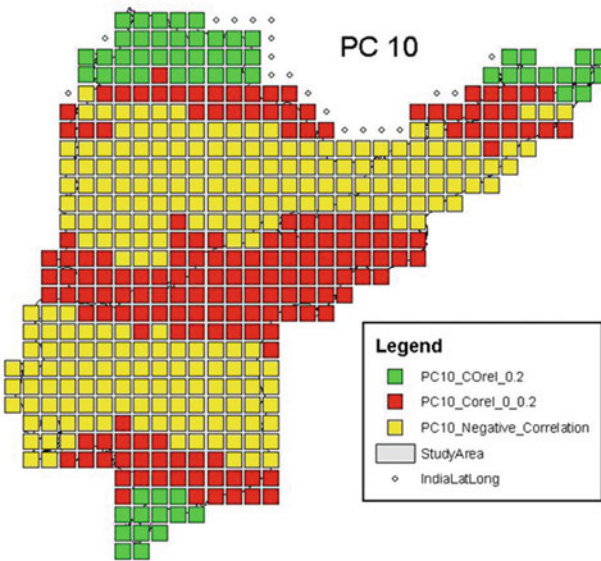


Fig. 10 (continued)



e



f

Fig. 10 (continued)

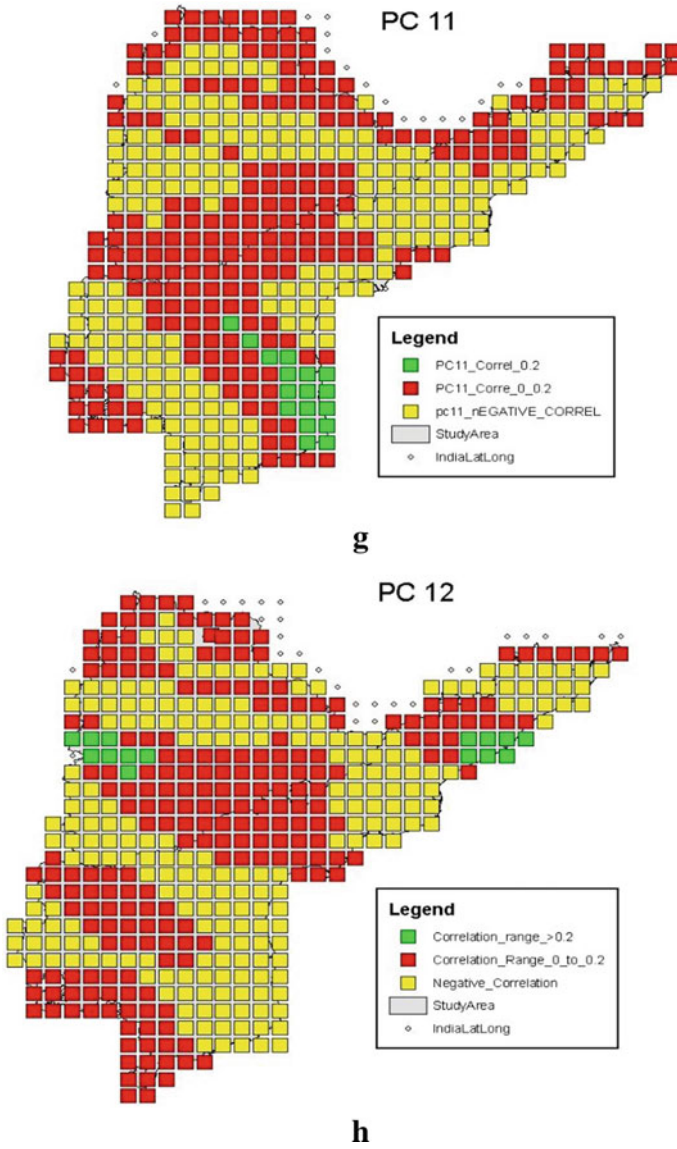
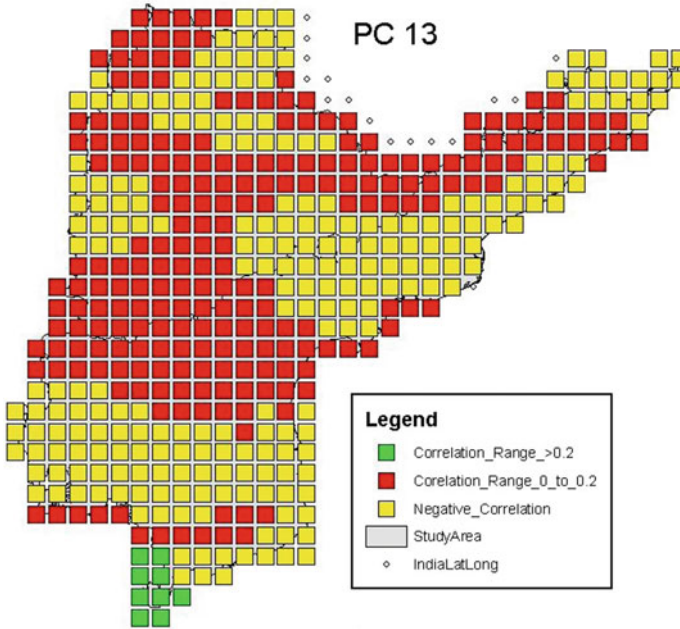
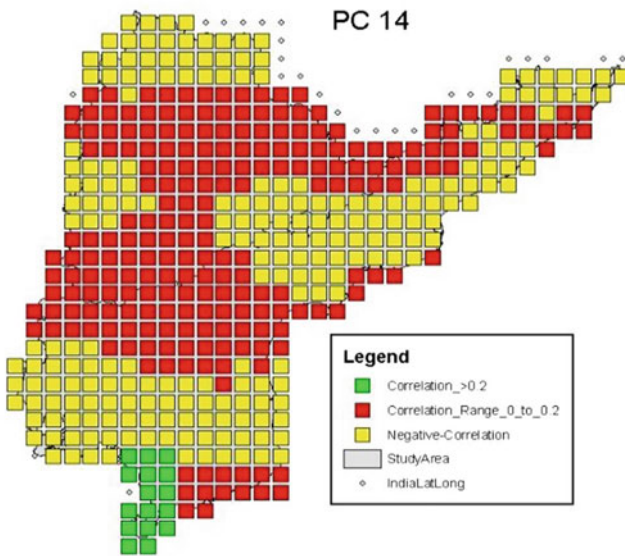


Fig. 10 (continued)



i



j

Fig. 10 (continued)

The 14 PCs have the eigenvalue  $> 1$  which is explaining only 66.67% of the total variance of the daily precipitation data. For covering 90% of the total variance, it is taking 97 principal components. The main purpose of the PCA is to reduce the variables with minimum loss of information, so that with a few reduced variables, further analysis can be carried out. But here, the number of reduced variables still large. If only first 14 PCs which have eigenvalues more than 1 are considered, it covered only 66.67% of the information. So the PCA advise is not the best possible choice to descried homogeneous for daily precipitation data, because the first few leading PCs (derived from inter-station correlation and/or covariance matrix) do not cumulative account for a consequential per cent of the total variance shared by stations.

## References

1. Srinivas VV (2013) Regionalization of precipitation in India: a review. *J Indian Instit Sci* 93(2):153–162
2. Almazroui M, Dambul R, Islam MN, Jones PD (2015) Principal components-based regionalization of the Saudi Arabian climate. *Int J Climatol* 35(9):2555–2573
3. Darand M, Daneshvar MRM (2014) Regionalization of precipitation regimes in Iran using principal component analysis and hierarchical clustering analysis. *Environ Process* 1(4):517–532
4. Miller SN, Semmens DJ, Goodrich DC, Hernandez M, Miller RC, Kepner WG, Guertin DP (2007) The automated geospatial watershed assessment tool. *Environ Modell Softw* 22(3):365–377
5. Munoz-Diaz D, Rodrigo FS (2004) Spatio-temporal patterns of seasonal rainfall in Spain (1912–2000) using cluster and principal component analysis: comparison. *Ann Geophys* 22(5):1435–1448
6. Richman MB (1986) Rotation of principal components. *J Climatol* 6(3):293–335
7. Indian Meteorology Department. <http://www.imd.gov.in/>