A PROBABILISTIC GRAPHICAL MODEL APPROACH TO IDENTIFYING SPATIAL CHANGES IN MONTHLY PRECIPITATION UNDER CLIMATE CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

A regionalization method based on Markov random field (MRF) model is proposed for grouping locations into spatially coherent homogeneous regions with respect to monthly rainfall. The model consists of discrete state variables for representing low and high rainfall at grid-scale, and also for temporal rainfall patterns. Chinese Restaurant Process (CRP) prior is used on the latent variables for computing their posterior distribution conditioned on observations of monthly rainfall across India. Inference of the latent variables is done by Gibbs sampling to estimate the final clustering of locations to obtain regions which are spatially coherent and homogeneous with respect to precipitation patterns. The regions are obtained for two time-intervals: 1950-1982 and 1983-2013 to assess the impact of potential climate change on precipitation patterns. We find that some locations have moved from high to low precipitation regions or vice versa due to large change in rainfall characteristics.

1 INTRODUCTION

The variability in the precipitation pattern over Indian region at multiple spatial and temporal scales affects agricultural productivity and economics in the society Paul et al. (2016). The changing trend of precipitation could increase the risk of failure of a conventional water management system and pose a threat to agricultural economics that involves food, agricultural and environmental policy Saha et al. (2014). Past studies have observed both increasing and decreasing trends in seasonal precipitation at different places across the world Zhai et al. (2005). Sayemuzzaman & Jha (2014) found an increasing trend in precipitation in the summer season and a decreasing trend in the winter season in North Carolina, United States. Spatio-temporal patterns could also change at regional scale due to global climate change. Hence spatio-temporal analysis of precipitation is necessary for sustainable hydrological and agricultural planning towards adaptation to potential climate change. Spatiotemporal analysis of precipitation at a regional scale aids in solving many hydrological problems, such as development of precipitation dataset in ungauged catchment, streamflow prediction in ungauged basin, identifying sites with similar precipitation pattern and construction of intensity-duration frequency curves at points lacking data for any hydraulic structure design Satyanarayana & Srinivas (2008). Regionalization of precipitation is a technique to identify groups of sites having similar temporal characteristics of rainfall, which can be called "homogeneous regions. It is essentially a clustering problem, and hybrid cluster analysis, kernel-based fuzzy cluster analysis, fuzzy clustering approach, K-means clustering, spectral clustering and hierarchical clustering are among the more commonly used techniques for regionalization Rao & Srinivas (2006); Basu & Srinivas (2014); Nasseri & Zahraie (2011); Türkeş & Tatlı (2011); Darand & Daneshvar (2014); Hsu & Li (2010). However, these techniques are not guaranteed to produce spatially coherent clusters, without which the purpose of regionalization is not fulfilled. Hence, further studies are required to develop techniques that can provide contiguous regions which are also homogenous. A variety of approaches have been tried for regionalization. Hsu & Li (2010) employed wavelet transform and self-organizing map neural network for the regionalization of annual precipitation over Taiwan. Other methods such as hybrid cluster analysis and K-means clustering falls into this category. The probabilistic approach in data mining uses statistical inferential methods for model-building to regionalize the precipitation Bock (1996). Cowpertwait (2011) performed regionalization using a

mixed multivariate Gaussian model to get spatially coherent regions defined by Voronoi tessellation. Their study selected the optimal cluster based on the Bayesian information criterion. This suggests that a probabilistic approach for spatial clustering of precipitation may be suitable for regionalization.

The objective of the present study is to provide a probabilistic framework that performs regionalization of precipitation over the Indian region, which would be used to investigate the regional patterns in two time periods. This study has used the Markov Random Field (MRF) model Cross & Jain (1983), to model the spatial characteristics of the process. The proposed framework first creates a discrete representation of precipitation by introducing latent random variables, whose values are estimated by probabilistic inference. Then this discrete representation is used for identifying spatially contiguous regions which are homogeneous with respect to temporal behaviour of the hydroclimatic variables under consideration. For example, we find that in case of India, the Thar desert in the west, and Tamil Nadu coast in the south-east, are distinct homogeneous regions with respect to monthly precipitation. The study further analyses the transition of the cluster patterns between two time-periods to investigate the variability that could be due to the effect of global climate change.

2 Data and Methodology

2.1 Data and Study Area

India, our study region, lies in between $8\circ04$'N, $37\circ06$' N latitude and $68\circ07$'E, $97\circ25$' E longitude. India is highly diverse with respect to the climate variabilities exist in the country. Major climatic divisions of India are tropical wet, subtropical humid, arid and mountain regions. According to the Indian Meteorological Department (IMD), India has four climatological seasons: Winter, Pre-Monsoon, Monsoon, Post-Monsoon. The diversified climatic conditions lead to highly variable precipitation patterns at both spatial and temporal scales across the country. In the current study, gridded daily rainfall obtained from the Indian Meteorological Department (IMD) at $0.25\circ$ spatial resolution for the period 1950-2013, was used as the observed rainfall dataset. Pai et al. (2014) prepared the gridded IMD daily rainfall data using the 6955 rain gauge stations data in India with different availability period from 1901 to 2010.

2.2 Change-point Detection

Many studies have been done in time series analysis to analyse the trend exists in the hydroclimatic variables Chandniha et al. (2017); Chattopadhyay & Edwards (2016); Minaei & Irannezhad (2018). However, a few have focused on the research of irregularities that is present within the long-term time series data Mallakpour & Villarini (2016); Conte et al. (2019). Regional spatio-temporal analysis have been carried out for the time periods before and after any change that has taken place. Here, we try to find an abrupt change in the time series of all-India annual precipitation using the Pettitts test Pettitt (1979) for assessing temporal variability.. The Pettitts test for change detection is a non-parametric and most commonly used test that detects the abrupt changes in climatic records. A non-parametric test statistics U_t is defined below:

$$U_t = \sum_{i=1}^{t} \sum_{j=t+1}^{n} sign(x_i - x_j) \tag{1}$$

Where $\{x_i\}_{i=1}^{n}$ is a series of observed data that has a change point at t. The next step is to define the statistical change point test (SCP) using the test statistic K (where $K = max(\|U_t\|)$) and the associated confidence level ρ (where $\rho = exp(\frac{-K}{(n^2+n^3)})$ for the sample length (n). When ρ is smaller than the predefined confidence level, the null hypothesis is rejected. The use of Pettitts test on a time series data creates two subseries data between which a significant change point exists.

2.3 Markov Random Field Model

Markov Random Field (MRF) is a statistical model which involves sets of random variables, some of which may be observed, i.e. directly measurable, and others which are latent i.e. not directly measurable. Let X be the $S \times T$ observation matrix, where element (s, t) stores the observed

value (rainfall, temperature, etc) at location s for month t. The S locations of the study area are ordered sequentially first according to longitude and then according to latitude, though any other ordering is also acceptable. For each locations s, we identify a set of neighbouring locations $\Omega(s)$, which are within a range of 0.5 degree from s along either latitudes or longitudes. We model the observations X as random variables, using location-specific mixture distributions. In other words, at each location we fit two distributions on the rainfall volume one for low values of X and another for high values. These distributions can be chosen according to the nature of the observations. In this work, we choose Gamma distribution for rainfall.

In the proposed model there are two sets of latent variables: 1) discrete state (Z) at each spatio-temporal location (s, t) which can take binary values indicating which of the two distributions fit the corresponding observation $X(s, t)$ as mentioned above, and 2) spatial clustering variables V for each grid point. The variable $V(s)$ takes integer values specifying the temporal pattern of precipitation at location s. These temporal patterns are time-series of binary values, indicating high or low values of rainfall at different months.

We assume that these Z-variables are spatio-temporally coherent, i.e. spatio-temporally adjacent locations should have same values of Z. This will help us to carry out regionalization more effectively. We construct a spatio-temporal graphical model, using every Z and X variable as a vertex. Each $Z(s, t)$ variable vertex is connected to the corresponding $X(s, t)$ vertex using a data edge. If two locations s and s‘ are geographical neighbors, we connect the vertices corresponding to $Z(s, t)$ and $Z(s‘, t)$ using spatial edges, for all values of t. Similarly, we connect the vertices corresponding to $Z(s, t)$ and $Z(s, t - 1)$ with temporal edges.

Now, we define edge potential functions on each of these edges. The graphical model is now a Markov Random Field (Kindermann and Laurie, 1980), which specifies the joint probability distribution of all the variables as the product of all the edge potential functions. We choose the edge potential functions in such a way that (i) spatio-temporal coherence of the Z variables is maximized, (ii) the Z-variables fit the observations X well through the mixture distributions. For details of these functions, the reader is referred to Mitra & Seshadri (2019). The inference problem now is to find the optimal values of all the Z-variables such that the joint probability is maximized, which is achieved through Gibbs Sampling Casella & George (1992). The details of the approach is provided in Mitra & Seshadri (2019). However, here we introduce new discrete variables into our model, $V = \{V(1), V(2), \ldots, V(S)\}$ which indicate a clustering of the spatial locations. This clustering is done based on binary vectors of Z-variables: each location is represented with a T-dimensional time-series of Z-variables. For any location s, $V(s)$ denotes its cluster index. We do not fix the number of clusters to be formed, rather this number is inferred by the model from the data. Each cluster$k$ is associated with a canonical T-dimensional binary time-series, and if $V(s) = k$ for any location, then the time-series of Z at location s is a noisy version of the k-th canonical time-series. These canonical time-series too need to be estimated from the data. The average amount of noise, i.e. number of time-points where the time-series of Z mismatches with the canonical time-series, is a user-tuneable parameter called μ. Small value of μ necessitates the formation of pure clusters with less noise, hence each cluster is small in size and number of clusters formed is high, while the reverse happens for large values of μ.

To obtain spatially contiguous clusters, we put a prior distribution on the V-variables based on Chinese Restaurant Process Pitman (1995), such that each location s joins the same cluster as at least one of its neighbouring locations in $\Omega(s)$, and is more likely to join clusters which already contains more locations. These two constraints on the clustering ensures that a reasonably small number of spatial clusters are formed which have reasonable size and are spatially contiguous. To obtain the optimal values of V, we again define a joint distribution of X, Z, V variables, and use Gibbs Sampling.

3 Results and Discussions

3.1 Pettitts test

In the present study, Pettitts test was applied to the annual precipitation time-series (1950-2013) over India. The test was carried out for 95% confidence interval. For the sake of brevity, the change point for one randomly selected grid location grid location $(8.3 \circ N, 77 \circ E)$ is shown in Figure 1. It was

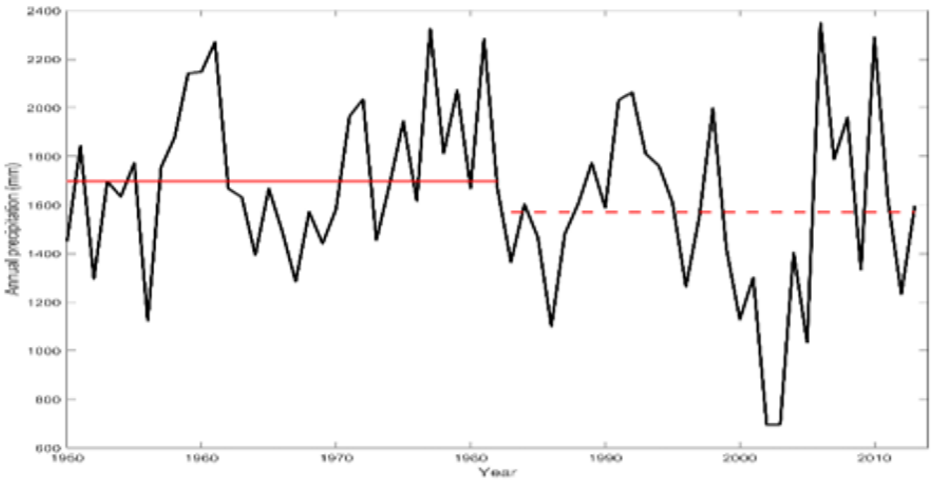

Figure 1: The annual rainfall over India in each year (1950-2013) is plotted. The red lines show the mean annual rainfall before and since 1983. The obvious shift justifies using 1983 as the change-point.

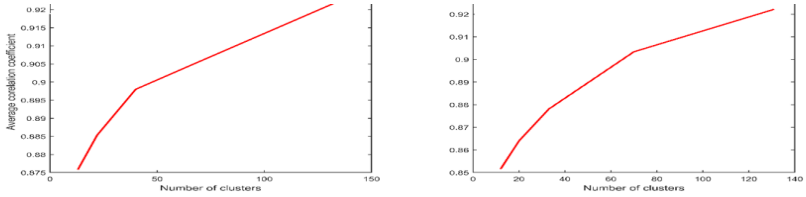

Figure 2: Plot between the average correlation coefficient for the derived homogeneous regions and the corresponding cluster number for regionalization based on monthly precipitation for H1(1950-1982) and H2 (1983-2013) periods

found that a change in mean of the precipitation occurred in 1983 for the same grid location. The change point analysis of precipitation across India indicated different change points from the year 1975 to 1986, with maximum change points in and around 1981 and 1982. Based on the outcomes of change point analysis regionalisation was performed for two different time periods that is 1950-1982 and 1983-2013. For the rest of the paper, these two periods will be referred to as Historical Period 1 (H1) and Historical Period 2 (H2) respectively.

3.2 Parameter Selection of MRF-based Regionalization

We attempted to regionalize of the Indian landmass in order to obtain regions of similar precipitation pattern using the model discussed above. In the adopted framework, the number of clusters is not predefined by user but identified by the model, though it can be influenced by the user-specified parameter μ. Smaller the μ, higher is the number of clusters and vice versa. However, the homogeneity of these clusters should also be high, which we measure using correlation coefficient between the precipitation time-series at the locations belonging to each homogeneous region. Figure 2 shows the average correlation in the different regions that are obtained when we use $\mu = 5$ and find the values to be around 0.85-0.95, indicating high homogeneity. We also vary μ to 3, 5, 7, 9 and 11, and it appears that the average correlation coefficient improves with the number of clusters until a particular value, after which it does not show any significant improvement. Lower values of are associated with a high number of regions and high correlation, i.e. we get small and very homogeneous regions. For the rest of the analysis we use $\mu = 5$.

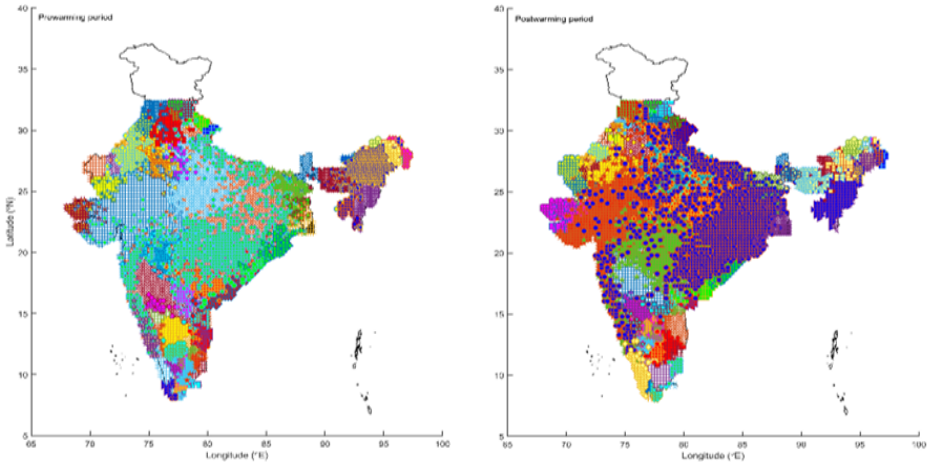

Figure 3: Homogeneous regions of monthly precipitation in India identified using the MRF model with $\mu = 5$. The colours represent the regions identified for the periods 1950-1982 (left panel) and 1983-2013 (right panel).

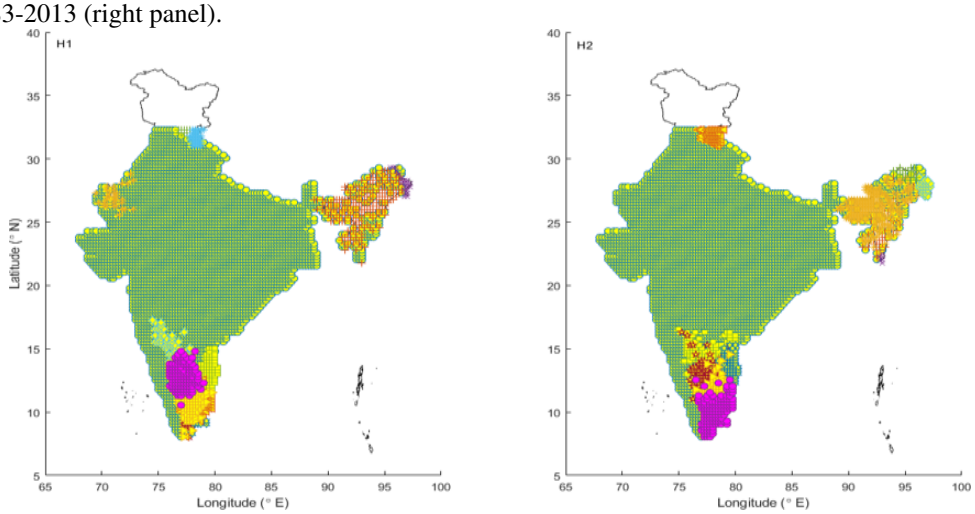

Figure 4: Homogeneous regions of monthly precipitation in India identified using the MRF model with =11. The colours represent the regions identified for the periods 1950-1982 (left panel) and 1983-2013 (right panel).

3.3 Results of MRF-based Regionalization

The MRF-based regionalization results on monthly precipitation for both H1 and H2 periods using $\mu = 5$ and $\mu = 11$ are shown in Figure 3 and 4 respectively, where each region is represented by a color. The total number of grid-points used for regionalization is 4551. For $\mu = 5$, the H1 and H2 periods were then found to have 67 and 70 homogeneous regions, respectively, while for $\mu = 11$ these numbers are 13. The homogeneity is somewhat higher (around 0.9) in the former case, compared to the latter (0.85), as discussed earlier. The regions were ranked and labelled in ascending order of the regional average precipitation.

The results of the above analysis could be explained in the context of Indian Summer Monsoon Rainfall patterns. The regions that receive the highest rainfall in the two periods were found to be in north-eastern region and Kerala coast in the South-west. Regions that represents very low rainfall regions were found to be in the north-western part of India in both H1 and H2 period. Parts of Rajasthan and Gujarat in the west were assigned to separate clusters as the factors such as the presence of desert and proximity to sea affect the rainfall characteristics. Under all the settings, the

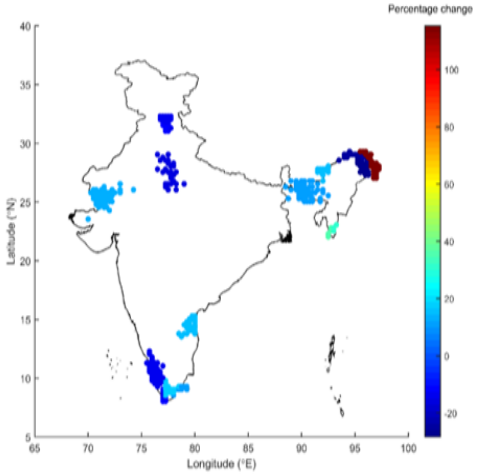

Figure 5: Grid points showing significant changes in the H2 period compared to H1

southern and the north-eastern parts of India were found to contain several regions, as may be seen in Figures 3 and 4. It reflects the heterogeneous nature of precipitation pattern that exist in those regions due to the complex geographical features (eastern, western ghats mountain ranges and seas on three sides of South India, and a complex structure of orography in North-eastern India). Owing to climate variability possibly due to climate change, the precipitation patterns at a regional scale undergo a significant change which can be seen in the regionalization pattern in both periods. A few grid points in drier/wet regions of H1 period shifted to the regions associated with heavy/less rainfall in H2. To quantify the changes in regionalization between the two periods, the Rand index was computed using the two clustering in two periods and found to be 0.88 respectively that indicates significant changes in the regionalization. It was also found that using the same regionalization in both periods (H1 and H2) result in reduced homogeneity, verifying further that the spatial patterns of precipitation have indeed changed between the two periods. The percentage change in the regional average was calculated at each grid point to identify those grid points. The grid points that are more than one standard deviation from the mean of percentage change was identified as grid points with severe changes and it is plotted in Figure 5. Parts of North-east India, South-western (Kerala) and northern India were found to be associated with low rainfall regions in the H2 period, while parts of Thar desert in Rajasthan moved to higher rainfall regions in H2. Increase in rainfall was also observed in some parts of eastern India (close to the Eastern Ghat hills) and North-east India.

4 Conclusion

The present study investigated the changes in regionalization of precipitation between two historical periods obtained using change point analysis. The MRF model was used to obtain regions using monthly precipitation time-series. The regions identified were both homogeneous and spatially contiguous. The homogeneous regions were identified and labelled to understand the changes in characteristics of regional patterns in the future scenario. The grid points associated with higher percent changes in the H2 period were identified. The results of the study showed that there are appreciable changes in regionalization as well as the characteristics of the homogenous regions regarding precipitation patterns in the H2 period. The major conclusions of the study is that parts of North-eastern, South-eastern and Northern India experienced significant changes during the post-warming period. Increase in rainfall was observed in some parts of Eastern India and Thar desert in Rajasthan. This framework could similarly be carried out for future climate under different climate scenarios. It is essential that changed precipitation patterns be considered in regional water resources planning and water resource management.

REFERENCES

Bidroha Basu and VV Srinivas. Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resources Research*, 50(4):3295–3316, 2014.

Hans H Bock. Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23(1):5–28, 1996.

George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46 (3):167–174, 1992.

Surendra Kumar Chandniha, Sarita Gajbhiye Meshram, Jan Franklin Adamowski, and Chandrashekhar Meshram. Trend analysis of precipitation in jharkhand state, india. *Theoretical and Applied Climatology*, 130(1-2):261–274, 2017.

Somsubhra Chattopadhyay and Dwayne R Edwards. Long-term trend analysis of precipitation and air temperature for kentucky, united states. *Climate*, 4(1):10, 2016.

Luiza Chiarelli Conte, Débora Missio Bayer, and Fábio Mariano Bayer. Bootstrap pettitt test for detecting change points in hydroclimatological data: case study of itaipu hydroelectric plant, brazil. *Hydrological Sciences Journal*, 64(11):1312–1326, 2019.

Paul SP Cowpertwait. A regionalization method based on a cluster probability model. *Water Resources Research*, 47(11), 2011.

George R Cross and Anil K Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):25–39, 1983.

Mohammad Darand and Mohammad Reza Mansouri Daneshvar. Regionalization of precipitation regimes in iran using principal component analysis and hierarchical clustering analysis. *Environmental Processes*, 1(4):517–532, 2014.

Kuo-Chin Hsu and Sheng-Tun Li. Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources*, 33(2):190–200, 2010.

Iman Mallakpour and Gabriele Villarini. A simulation study to examine the sensitivity of the pettitt test to detect abrupt changes in mean. *Hydrological Sciences Journal*, 61(2):245–254, 2016.

Masoud Minaei and Masoud Irannezhad. Spatio-temporal trend analysis of precipitation, temperature, and river discharge in the northeast of iran in recent decades. *Theoretical and applied climatology*, 131(1-2):167–179, 2018.

Adway Mitra and Ashwin K Seshadri. Detection of spatiotemporally coherent rainfall anomalies using markov random fields. *Computers & geosciences*, 122:45–53, 2019.

Mohsen Nasseri and Banafsheh Zahraie. Application of simple clustering on space-time mapping of mean monthly rainfall pattern. *International Journal of Climatology*, 31(5):732–741, 2011.

DS Pai, Latha Sridhar, M Rajeevan, OP Sreejith, NS Satbhai, and B Mukhopadhyay. Development of a new high spatial resolution ($0.25\times$ 0.25) long period (1901–2010) daily gridded rainfall data set over india and its comparison with existing data sets over the region. *Mausam*, 65(1):1–18, 2014.

Supantha Paul, Subimal Ghosh, Robert Oglesby, Amey Pathak, Anita Chandrasekharan, and RAAJ Ramsankaran. Weakening of indian summer monsoon rainfall due to changes in land use land cover. *Scientific reports*, 6(1):1–10, 2016.

AN Pettitt. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):126–135, 1979.

Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

A Ramachandra Rao and VV Srinivas. Regionalization of watersheds by fuzzy cluster analysis. *Journal of Hydrology*, 318(1-4):57–79, 2006.

Anamitra Saha, Subimal Ghosh, AS Sahana, and EP Rao. Failure of cmip5 climate models in simulating post-1950 decreasing trend of indian monsoon. *Geophysical Research Letters*, 41(20): 7323–7330, 2014.

P Satyanarayana and VV Srinivas. Regional frequency analysis of precipitation using large-scale atmospheric variables. *Journal of Geophysical Research: Atmospheres*, 113(D24), 2008.

Mohammad Sayemuzzaman and Manoj K Jha. Seasonal and annual precipitation time series trend analysis in north carolina, united states. *Atmospheric Research*, 137:183–194, 2014.

Murat Türkeş and Hasan Tatlı. Use of the spectral clustering to determine coherent precipitation regions in turkey for the period 1929–2007. *International Journal of Climatology*, 31(14):2055–2067, 2011.

Panmao Zhai, Xuebin Zhang, Hui Wan, and Xiaohua Pan. Trends in total precipitation and frequency of daily precipitation extremes over china. *Journal of climate*, 18(7):1096–1108, 2005.