A Detailed Analysis of Rainfall Data of West Bengal from 1901 to 2021

Swapnaneel Bhattacharyya (BS2105) [1],
Srijan Chattopadhyay (BS2126) [1],
Sevantee Basu (BS2129) [1],
Koustav Mallick (BS2107) [1],
Soumava Mondal (BS2014) [1]

[1] *Indian Statistical Institute, 203 Barrackpore Trunk Rd, Kolkata, WB 700108, India.*

Abstract: In this project, we study the West Bengal rainfall data over the years 1901-2020 considering different regions. We estimate the onset and departure of monsoon for each of the years and check whether there is any change in the monsoon onset, Departure, monsoon duration, monsoon rainfall, and total rainfall over the years. We also introduce a new clustering method for the years based on the similarity of rainfall patterns. We incorporate this method through simulation studies. We also interpret the results and try to find possible explanations for them.

Keywords and phrases: Onset, Departure, Changepoint, Clustering, Spectral Density.

Contents

1 Introduction . . . 3
2 West Bengal Rainfall Data . . . 3
3 Formulation of the problem . . . 3
4 Methods . . . 4
4.1 Onset, Departure Detection and Change Point of Onset, Departure . . . 5
4.2 Changepoint detection for Monsoon onset-Departure and total Rainfall over the years . . . 7
4.3 Change in the Rainfall Pattern . . . 8
4.4 Clustering based on Mean . . . 8
4.5 Subclustering based on Covariance . . . 9
5 Simulation Studies . . . 9
6 Results . . . 10
6.1 Change point detection of Onset, Departure, Monsoon Rainfall, and Total Rainfall . . . 10
6.2 Change in the Rainfall Pattern . . . 13
7 Conclusion . . . 13
References . . . 15
8 Appendix . . . 15
A Nonparametrtic Spectral Density Estimate of a Second-order Stationary Time Series . . . 15

1. Introduction

Rainfall patterns are crucial not only for agricultural productivity and water resource management but also for shaping socio-economic dynamics. In West Bengal, where agriculture is a cornerstone of the economy, understanding the temporal evolution of rainfall is essential for informed decision-making and sustainable development.

Our project embarks on a thorough analysis of over a century's worth of rainfall data, from 1901 to 2020, across West Bengal's diverse districts. This presentation will illuminate the methodologies we employed, the significant findings we uncovered, and the profound implications of our research.

In this project, we conduct a region-wise analysis of rainfall data. We estimate the onset and Departure of the monsoon for each year in each region. Based on these estimations, we examine whether the start and end of the monsoon, its duration, the total rainfall during the monsoon, and the annual total rainfall have changed over the past 120 years. Additionally, we perform region-wise clustering of the years based on the similarity of rainfall patterns. We also look into the possible implications of the results found.

2. West Bengal Rainfall Data

We use the West Bengal rainfall data over the years 1901-2020 for different stations. A part of the West Bengal rainfall dataset is shown in the following table

LAT	LON	DISTRICT	STATION	YEAR	J1	J2	J3	$\cdots$
22.72	88.48	24 PGS N	BARASAT	1901	0	0	0	$\cdots$
22.72	88.48	24 PGS N	BARASAT	1902	0	0	0	$\cdots$
22.72	88.48	24 PGS N	BARASAT	1903	0	0	0	$\cdots$
22.72	88.48	24 PGS N	BARASAT	1904	0	0	0	$\cdots$
22.72	88.48	24 PGS N	BARASAT	1905	0	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

In the data, each row consists of rainfall of a particular year of a station in West Bengal (e.g. J1 refers to the first day of January). We downloaded the shape file of West Bengal from the internet and used the latitude and longitude of that data along with that of our data the corresponding rainfall to plot the following maps. This was done mainly using *terra, sf, dplyr, spatstat* packages in R. We plot the district-wise total rainfall of West Bengal for several years in Figure 1. This total means averaged over all the stations in the district and then summed over all the days. In the figure, darker the shade of blue, higher the total rainfall in that year. For missing data, we used a 5 day (backward) local averaging (for first 5 day if missing, then 0) to remove the missingness.

3. Formulation of the problem

We have data on rainfall over various time points in different locations of West Bengal for 120 (121 in some locations) years. So the data has the form $\{R^i_{s,t} : t \in \Gamma_T, s \in \Gamma_S, i = 1, \cdots, 120(or, 121)\}$, where s denotes the spatial point of an observation. t denotes the temporal point of an observation. Γ_T denotes the set of temporal sampling regimes. For our case, $\Gamma_T = \{1, 2, \cdots, 366\}$. Γ_S denotes the set of spatial sampling regimes. Assume $\Gamma_S = \{s_1, \cdots, s_m\}$. $R^i_{t,s}$ denotes the observed rainfall in ith year in location s during the tth time point of that year.

Define $\mu^i_{s,t} = \mathbb{E}(R^i_{s,t})$. So, observe that for each **year** i, we have a matrix of datapoints. In each of the matrices, each row consists of observations $\{R^i_{s,t} : t \in \Gamma_T\}$ for a spatial location s. In each of the matrices, each column consists of observations $\{R^i_{s,t} : s \in \Gamma_S\}$ for a timepoint t. For i th year the data matrix is,

$$R^i = \begin{bmatrix} * & 1 & 2 & \cdots & 366 \\ s_1 & R^i_{s_1,1} & R^i_{s_1,2} & \cdots & R^i_{s_1,366} \\ s_2 & R^i_{s_2,1} & R^i_{s_2,2} & \cdots & R^i_{s_2,366} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_m & R^i_{s_m,1} & R^i_{s_m,2} & \cdots & R^i_{s_m,366} \end{bmatrix}$$

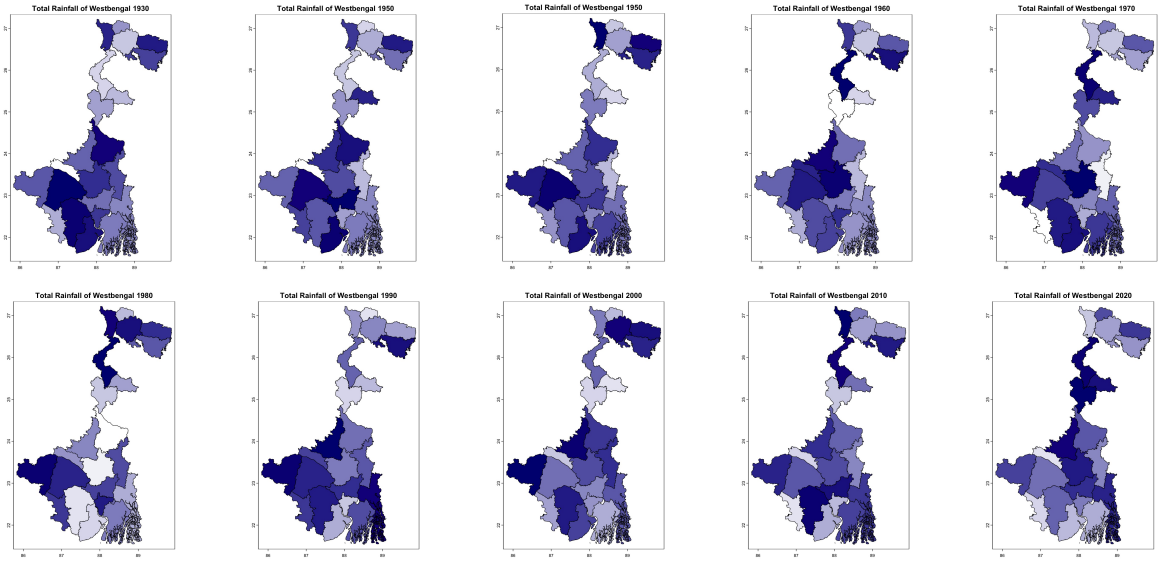


Fig 1. Total Rainfall of West Bengal over the years (less rainfall (light blue) to more rainfall (deep blue))

Note that, some of the entries of this may not be available since the original dataset has some missing data. However, we assume the rainfall data over different years are independent.

Assumption 1. *We assume that for each location $s \in \Gamma_S$, for different years $\{i_1, \cdots, i_k\}$ the time processes $\{R^{i_1}_{s,t}\}_{t\in\Gamma_t}, \cdots, \{R^{i_k}_{s,t}\}_{t\in\Gamma_t}$ are independent.*

In practice, this is a very mild assumption since rainfall in one year is rarely affected by the previous year's natural factors. The central focus of this project is to explore the rainfall patterns in the different regions of West Bengal over different years and to inspect whether there is any change in the pattern or not. To be precise, We focus on the following questions :

- ☞ **Onset and Departure of Rainfall:** Has the timing of rainfall initiation and termination changed over time?
- ☞ **Duration of Monsoon:** Is there a difference in the length of the monsoon season over the years?
- ☞ **Rainfall Quantity:** Are there any fluctuations in the total amount of rainfall and monsoon rainfall over the years?
- ☞ **Monsoon Pattern:** Have there been any alterations in the typical pattern of rainfall across the years?

Figure 2 shows the daily rainfall averaged over the years. Each of these questions is addressed to different geographical regions of West Bengal.

4. Methods

All of the questions in our goals typically address changepoint detection problems. We address these questions region-wise for each of the regions : **Sub-Himalayan Zone, Himalayan Zone, Barind Zone, Gangetic Alluvial Plane, Rarh Plane, Saline Coastal Zone, and Western Plateau Zone**. The divisions referred to were acquired by utilizing the officially designated geographical divisions of West Bengal. For every year, for each of these regions $\mathbf{S}$, we find the average of rainfall in ith year at $t-$th time, over all locations inside the region $\mathbf{S}$ ($U^i_{\mathbf{S},t}$) i.e.

$$U^i_{\mathbf{S},t} = \frac{1}{|\mathbf{S}|}\sum_{s\in\mathbf{S}} R^i_{s,t}$$

So now for each region $\mathbf{S}$ at each year i, we have a time series data $\{U^i_{\mathbf{S},t}\}_{t\in\{1,2,\cdots,366\}}$. By Assumption 1, for each region $\mathbf{S}$ for different years $i_1, \cdots, i_k$, the time processes $\{U^{i_1}_{\mathbf{S},t}\}, \{U^{i_2}_{\mathbf{S},t}\}, \cdots, \{U^{i_k}_{\mathbf{S},t}\}$ are independent.

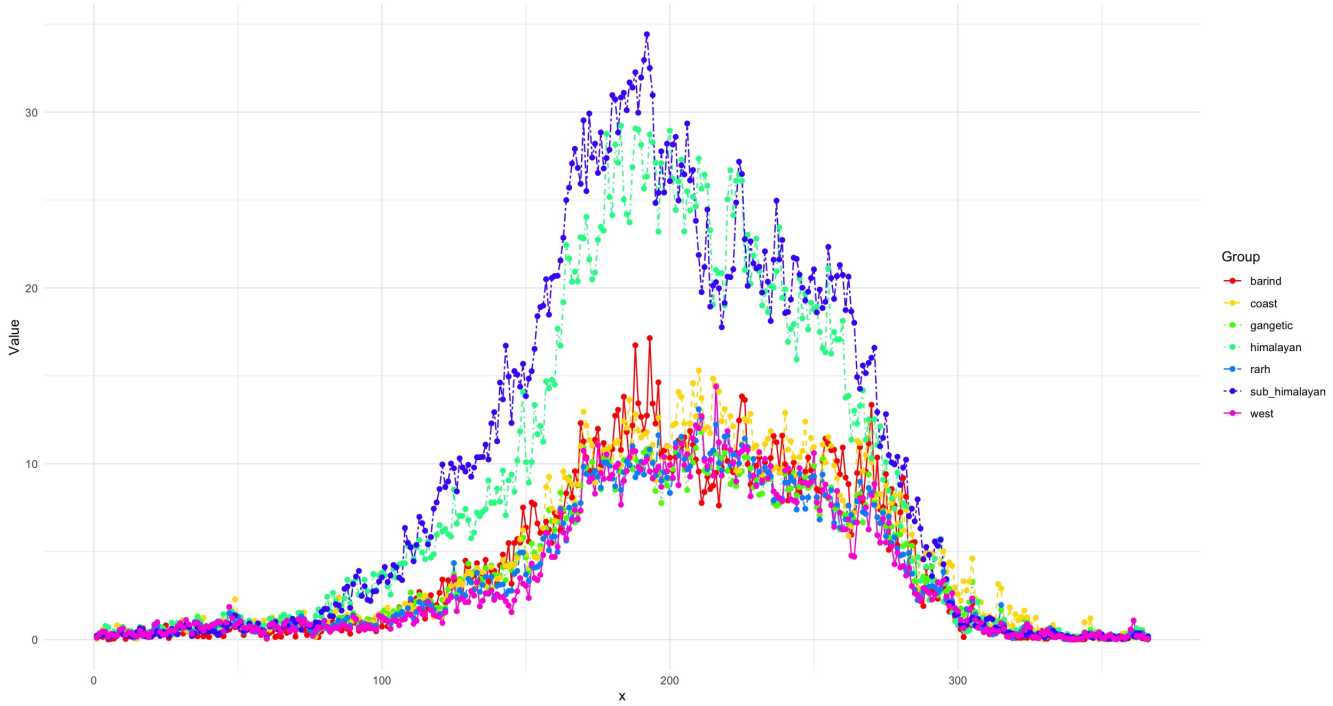
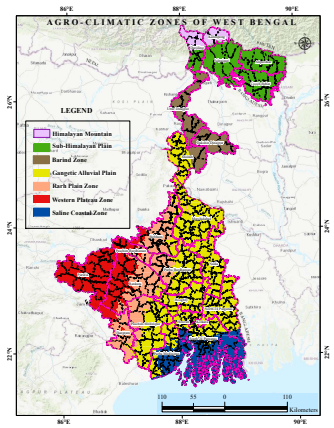


Fig 2. Yearwise average daily rainfall for the regions



4.1. Onset, Departure Detection and Change Point of Onset, Departure

IMD officially defines onset of monsoon as "If after 10th May, 60% of the available 14 stations enlisted, viz. Minicoy, Amini, Thiruvananthapuram, Punalur, Kollam, Allapuzha, Kottayam, Kochi, Thrissur, Kozhikode, Thalassery, Kannur, Kudulu and Mangalore report rainfall of 2.5 mm or more for two consecutive days, the onset over Kerala be declared on the 2nd day, provided the following criteria are also in concurrence. Depth of westerlies should be maintained upto 600 hPa, in the box equator to Lat. 10^{o}N and Long. 55^{o}E to 80^{o}E. The zonal wind speed over the area bounded by Lat. 5-10^{o}N, Long. 70-80^{o}E should be of the order of 15 – 20 Kts. at 925 hPa. The source of data can be RSMC wind analysis/satellite derived winds. INSAT derived OLR value should be below 200 wm-2 in the box confined by Lat. 5-10^{o}N and Long. 70-75^{o}E. Southwest monsoon normally sets in over Kerala around 1st June. It advances northwards, usually in surges, and covers the entire country around 15th July. The NLM is the northern most limit of monsoon upto which it has advanced on any given day."
But first of all, our data is confined in West Bengal only. Monsoon generally comes a bit later in West Bengal.

Also, the above definition not only uses rainfall data, but information from many other natural factors as well. So, we can't use this as our definition of Onset. So, based on the rainfall pattern of West Bengal, we define the Onset in the following way:

Definition 1. *We define* ***the onset of monsoon*** *as a specific time when there is a discernible overall increasing trend, significant variability in rainfall, and a sudden surge in average rainfall.*

We will first describe the method (Algortithm 1) of detecting onset from rainfall data and then use changepoint detection method described in Section 4.2. First we restrict our data to 19 May and 31 October. Our intuition is, at the start of rainfall there should be a higher trend, jump in the mean and variance of per day rainfall. We check the trend by spearman rank correlation between 9 days rainfall and it's sorted version. As it is a very small number, to create a larger impact, first we divide by it's modulus if it is non zero and then make a transform $f(x) = (-1 + \frac{1}{1-|x|})\text{sgn}(x)$. The graph of $f(x)$ is increasing in $[-1, 1]$. We will do the whole work independently for each region **S**. From Algorithm 1, if we input the region along with a particular θ, we get the input. Now, the question is how to choose that θ? Now, we had the official onset data for North Bengal and South Bengal separately. So, we took that θ which minimizes the sum of squared errors of the actual and predicted over all the regions, i.e. we define

$$L^N(\theta) = (\hat{B(\theta)} - N)^2 + (\hat{H(\theta)} - N)^2 + (\hat{SH(\theta)} - N)^2$$

$$L^S(\theta) = (\hat{G(\theta)} - S)^2 + (\hat{R(\theta)} - S)^2 + (\hat{W(\theta)} - S)^2 + (\hat{C(\theta)} - S)^2$$

where B, H, SH, G, R, W, C represents Barind, Himalayan, Sub Himalayan, Gangetic, Rarh, West Plateau, Coast regions respectively and $\hat{B(\theta)}$ means the onset of barind calculated at θ. Also, N and S means North and South official onset respectively. So, we take $\hat{\theta}^N = \underset{\theta>0}{argmin} L^N(\theta)$ and $\hat{\theta}^S = \underset{\theta>0}{argmin} L^S(\theta)$. Then we use $\hat{\theta}^N$ and $\hat{\theta}^S$ in our onset detection algorithm to use as the final form. So, our final onset detection algorithm will be Onset Detection$(X, \hat{\theta}^N)$ for northern regions and Onset Detection$(X, \hat{\theta}^S)$ for south regions. For North, $\hat{\theta} = 20$ and for south $\hat{\theta} = 1$. We then calculate the onset.

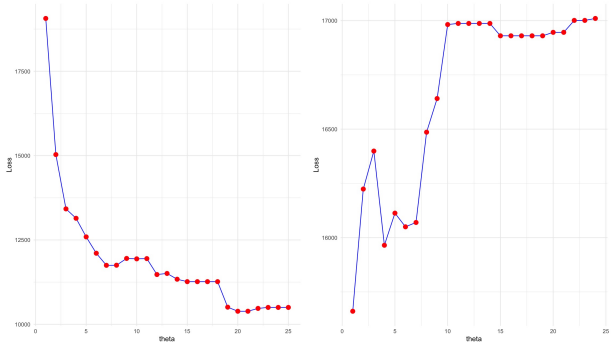


Fig 3. Plot of Loss with θ for North (left) and South (right)

We plotted the official onset data and the onset predicted by our algorithm in Figure 4. As we can see, in most of the cases, the difference between the official and predicted ones are at most 2-5 days. Now, the official onset is not only decided by rainfall, rather it takes into account various factors like air condition, sea temperature and other weather factors. But here we can see, only using rainfall data can give us a good estimate of the onset. Now the official onset data is available from 1983 only. So we have only used that part of the data to check accuracy.

Definition 2. *We define* ***the Departure of monsoon*** *as a specific time starting from around 1.5 months after the onset when there is a consecutive average rainfall less than 2 units.*

The algorithm for Departure is in Algorithm 2. Using estimated Departure and onset, we find the duration of monsoon, estimated monsoon rainfall and actual total rainfall for each region **S**. Then, we apply the same change point detection technique for detecting changes in the rainfall amount.

Algorithm 1: Onset Detection(X, θ)

Input: Rainfall data of a particular year of a region **S**, i.e., a vector X of length 366
Output: Estimated Onset for X
$X_1 = X[140:300]$
for $i \leftarrow 2$ ***to*** $length(X_1) - 10$ **do**
  $kt = \rho_{spearman}(X_1[i:i+9], sort(X_1[i:i+9]))$;
  $kt = \mathbb{1}(kt \neq 0)\frac{kt}{\sqrt{|kt|}}$;
  $spread = Standard\ Deviation(X_1[i:i+9])$;
  $avg = Mean(X_1[i:i+9])$;
  $X_2[i-1] = (avg + \theta sgn(kt) * (-1 + \frac{1}{1-|kt|})) * spread$
for $j \leftarrow 10$ ***to*** $length(X_2) - 10$ **do**
  if(max(X_2[(i-9):(i+9)]) = $X_2[i]$) **return** *i+139*
  else **return** *140*

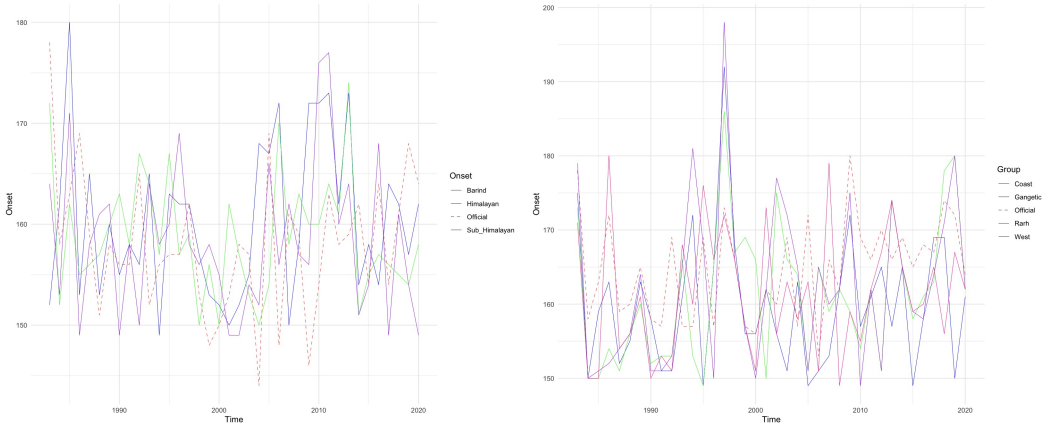


Fig 4. Plot of official data with our predicted onset

Algorithm 2: Departure Detection(X)

Input: Rainfall data of a particular year of a region **S**, i.e., a vector X of length 366
Output: Estimated Departure for X
for $i \leftarrow 2$ ***to*** $length(X) - 5$ **do**
  initialize vector u;
  $u[j] = \text{Mean}(X[j:j+5])$
initialize vector s;
$s = \text{which}(u[229:300] \leq 2)$;
If s is empty Departure = 300
else Departure = $\min(s + 228, 300)$
Return Departure.

4.2. Changepoint detection for Monsoon onset-Departure and total Rainfall over the years

We have the estimates of the onset and Departure for monsoon over the years. We now detect whether there has been any significant change in the onset or Departure time over the years for every region. We use the changepoint detection method introduced in Fearnhead and Rigaill (2019). For the data $\mathbf{y} = (y_1, \cdots, y_n)$, for $1 \leq s < t \leq n$, by $\mathbf{y_{s:t}}$ we denote the subvector $(y_s, \cdots, y_t)$. If we assume that there are k changepoints in the data, this will correspond to the data being split into $k+1$ distinct segments. We let the location of the jth changepoint be τ_j for $j = 1, \ldots, k$, and set $\tau_0 = 0$ and $\tau_{k+1} = n$. The jth segment will consist of data points $y_{\tau_{j-1}+1}, \ldots, y_{\tau_j}$. We let $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_{k+1})$ be the set of changepoints.The cost associated with a

segment of data, $y_{s:t}$, is taken to be:

$$\mathcal{C}(y_{s:t}) = \min_{\theta} \sum_{i=s}^{t} \gamma(y_i; \theta)$$

where $\gamma(y;\theta)$ is a loss function taken to be,

$$\gamma(y;\theta) = \begin{cases} (y-\theta)^2 & \text{if } |y-\theta| < K, \\ K^2 & \text{otherwise,} \end{cases}$$

In literature, this is known as *biweight loss function*. The penalized cost for a segmentation is then

$$Q(y_{1:n}; \tau_{1:k}) = \sum_{i=0}^{k} \{\mathcal{C}(y_{\tau_i+1:\tau_{i+1}}) + \beta\} :$$

where $\beta > 0$ is a chosen constant that penalizes the introduction of changepoints. We take the changepoints $(\tau_1, \cdots, \tau_k)$ to be the minimizer of cost function $Q(y_{1:n}; \tau_{1:k})$.

For our case, for a particular region $\mathbf{S}$ let $L_1^{\mathbf{S}}, \cdots, L_n^{\mathbf{S}}$ and $R_1^{\mathbf{S}}, \cdots, R_n^{\mathbf{S}}$ be the estimated onsets and Departures respectively for the years $1901, \cdots, 2020$ at region $\mathbf{S}$. So $n = 121$ for our case. We use the above method to determine whether there is any changepoint for the monsoon onset and Departure, the monsoon span, the total amount of rainfall in the year, and the highest amount of rainfall over the years.

4.3. Change in the Rainfall Pattern

Let$\mu_{s,t}^i = \mathbb{E}(R_{s,t}^i)$ be the mean of the rainfall data in $i-$th year, $t-$th time, $s-$th location. We determine, for each region of West Bengal, if there has been a change in the rainfall pattern over the years. In case there is any change, we will find the optimal number of clusters based on the similarity of the rainfall pattern. To be precise, we capture the similarity of the mean and covariance pattern of the rainfall data, i.e. for a region $\mathbf{S}$, we consider two years i_1 and i_2 in the same cluster if

$$\mathbb{E}U_{\mathbf{S},t}^{i_1} = \mathbb{E}U_{\mathbf{S},t}^{i_2} \quad \forall t \in \{1, 2, \cdots, 366\} \tag{4.1}$$

and

$$\mathrm{Cov}(U_{\mathbf{S},t+h}^{i_1}, U_{\mathbf{S},t}^{i_1}) = \mathrm{Cov}(U_{\mathbf{S},t+h}^{i_2}, U_{\mathbf{S},t}^{i_2}) \quad \forall h, t \tag{4.2}$$

4.4. Clustering based on Mean

We at first obtain the clusters based on the mean of the process i.e. only satisfying condition (4.1). For a region, let $\tilde{y}_1', \cdots, \tilde{y}_{120}'$ be the rainfall data vectors (i.e. each vector is of size 366). The key idea to obtain the clusters based on mean is

- For each possible K, obtain K clusters of the whole data.
- Now obtain the most suitable K.

We first make a transformation the rainfall data vectors$(\tilde{y}_i)$ to the cumulative rainfall data vector $(\tilde{y}_i')$ as

$$\tilde{y}_i[j] = \sum_{k=1}^{j} \tilde{y}_i'[k]$$

where $\tilde{y}_i[k]$ denotes the $k-$th entry of the vector $\tilde{y}_i$.

For a fixed k, we repeat the following large number of times (say, 10,000): for $l-$th iteration ($1 \le l \le 10,000$), we choose k distinct years uniformly from $\{1, 2, \cdots, 120\}$. Obtain the cluster $C_{k,l}$ by the *K-means algorithm* considering the chosen k years to be the initial centers for the *K-means algorithm*. Let $C_{k,l}$ have

the partition $\bigcup_{d=1}^{k} P_{d,l}$ of the year indices. We define the **Aggregate Intra-Cluster Variance (AICV)** corresponding to the clusters $C_{k,l}$ to be

$$\text{AICV}_{k,l} = \sum_{d=1}^{k} \sum_{j \in P_{d,l}} \left| y_j - \frac{1}{|P_{d,l}|} \sum_{j \in P_{d,l}} y_j \right|^2$$

where $|\ .\ |$ denotes the L_2-norm of a vector defined as

$$|(x_1, \cdots, x_{366})| = \sqrt{x_1^2 + \cdots + x_{366}^2}$$

Take the cluster $C_{k,l}$ to be the one with the lowest $\text{AICV}_{k,l}$. So we $\hat{C}_k = C_{k,\hat{l}}$ where $\hat{l} = \underset{1 \leqslant l \leqslant 10{,}000}{\text{argmin}}\ \text{AICV}_{k,l}$. For this cluster, we call the AICV to be AICV_k i.e. $\text{AICV}_k = \text{AICV}_{k,\hat{l}}$. So using this method for each given k, we can obtain a clustering with k many clusters. Now we have to choose the most optimal k. We describe two methods for obtaining the most suitable k: Elbow Method, Using an $F-$type statistic

The first one is a graphical method. In practice, any method of the two above can be used. We also show the efficiencies of the methods by simulation.

Elbow Method: For each k, we plot the AICV_k in a graph. Consider the k to be most optimal for which there is a maximum sharp drop in the graph.

Using an $F-$type statistic: For a given k, we have a clustering with k clusters. Let $\hat{\mu}_i$ be the vector which is the average of all of the observations in the ith cluster ($1 \leqslant i \leqslant k$). Let $\hat{\mu} = \frac{1}{120} \sum_{i=1}^{120} \tilde{y}_i$. Let F_k be defined as

$$F_k = \left(\frac{\sum_{i=1}^{k} |\hat{\mu}_i - \hat{\mu}|^2/(k-1)}{\text{AICV}_k/(120-k)} \right) \Big/ F_{k-1,120-k}^{1-\alpha}$$

We take the optimal number of clusters to be

$$\hat{K} = \underset{k}{\text{argmax}}\ F_k$$

4.5. Subclustering based on Covariance

We have obtained a clustering based on the mean of the rainfall data. Now we obtain the further finer clusters based on covariance pattern i.e. consider condition (4.2) now. Suppose the obtained cluster sets using the above method are $C_1, \cdots, C_K$. We discuss how we obtain further subclusters of each of the clusters C_i, $1 \leqslant i \leqslant k$. We discuss only in the context of C_1. The same method will be applied for all of the other clusters. Suppose cluster set C_1 consists of years $j_1, \cdots, j_r$ where $r = |C_i|$ is the size of C_1 i.e. $C_1 = \{\tilde{y}_{j_1}, \cdots, \tilde{y}_{j_r}\}$. For each year, the rainfall data of that year is a time series data. Let $\tilde{z}_{j_i} = \tilde{y}_{j_i} - \frac{1}{r} \sum_{i=1}^{r} \tilde{y}_{j_i}$. Let $\hat{f}_{j_i}$ be the spectral density estimate of $\tilde{z}_{j_i}$. So $\hat{f}_i$ is obtained from the data $\tilde{z}_{j_i}$. Since spectral density is a periodic function with period 2π, we work with a set of arguments in $[0, 2\pi]$ and call them $\theta_1, \cdots, \theta_m$. Consider the r (size of cluster C_1) vectors each consisting of the spectral density estimates of a year from the cluster C_1 at $\theta_1, \cdots, \theta_m$ i.e., consider $\left(\hat{f}_{j_1}(\theta_1), \cdots, \hat{f}_{j_1}(\theta_m)\right), \left(\hat{f}_{j_2}(\theta_1), \cdots, \hat{f}_{j_2}(\theta_m)\right), \cdots, \left(\hat{f}_{j_r}(\theta_1), \cdots, \hat{f}_{j_r}(\theta_m)\right)$. Now considering these r vectors as input, apply the first clustering method (Mean Based Clustering with Elbow Method) to obtain the subclusters of C_1.

5. Simulation Studies

To check the performance of our clustering method, we perform simulation studies. For that, we consider a collection of time-series vectors of equal size from different processes and run our clustering algorithm to find the number of clusters. We include our results of simulation studies in the following table

From the summary, it is pretty evident that in most cases, our method finds the exact number of clusters.

No of Vectors	Processes	Decomposition	Original No of Clusters	Detected No of Clusters
170	AR(1), MA(2), AR(2), MA(1), MA(3), AR(5)	30, 20, 40, 30, 20, 30	6	6
180	AR(2), MA(3), AR(3), MA(4), MA(5), AR(6)	40, 30, 20, 30, 30, 30	6	6
160	AR(1), MA(2), AR(2), MA(1), MA(3), AR(4)	20, 30, 30, 20, 30, 30	6	5
150	AR(2), MA(1), AR(3), MA(2), MA(4), AR(5)	30, 20, 20, 30, 20, 30	6	6
140	AR(1), MA(3), AR(4), MA(5), MA(2), AR(3)	20, 30, 30, 20, 20, 20	6	6
130	AR(3), MA(2), AR(5), MA(4), MA(1), AR(2)	30, 20, 20, 20, 20, 20	6	5
120	AR(2), MA(4), AR(1), MA(3), MA(5), AR(6)	20, 20, 20, 20, 20, 20	6	6
110	AR(3), MA(5), AR(2), MA(1), MA(4), AR(3)	10, 20, 20, 20, 20, 20	6	5
100	AR(4), MA(1), AR(5), MA(2), MA(3), AR(6)	20, 20, 20, 10, 10, 20	6	6
90	AR(1), MA(4), AR(3), MA(5), MA(2), AR(4)	10, 10, 10, 20, 20, 20	6	5

TABLE 1
Summary of Clustering Results

6. Results

6.1. Change point detection of Onset, Departure, Monsoon Rainfall, and Total Rainfall

We use the described changepoint detection method to investigate whether there has been any change over the years in monsoon onset and departure, duration of monsoon, and monsoon rainfall. We interpret the results we got one by one.

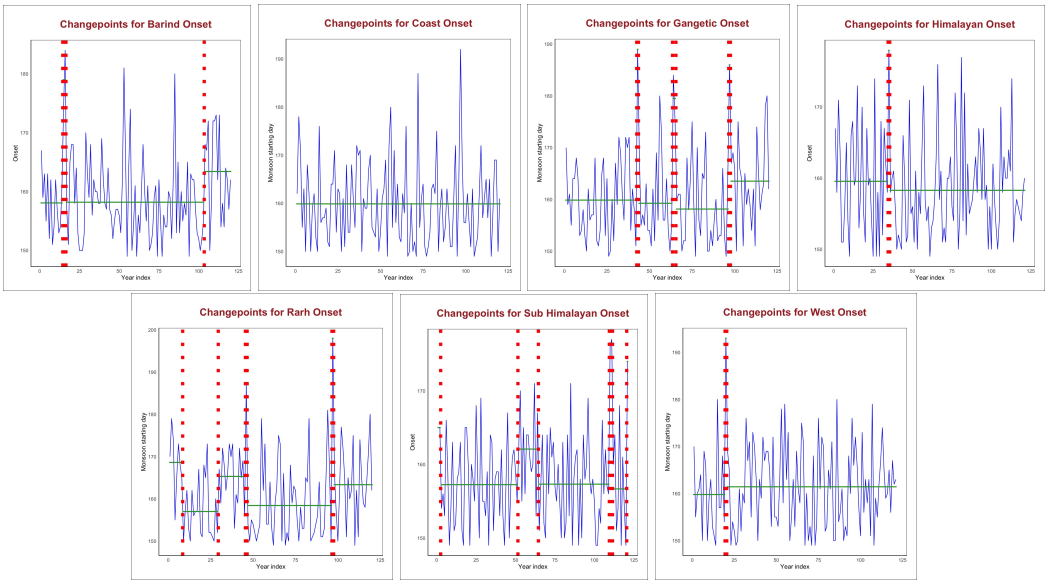


FIG 5. *Change Point Detection of Onset for Various Regions*

From Figure 5, we see For **Coastal Region** there is no change in the Onset time of Monsoon, possibly because they receive rainfall from the winds which originate around the same time every year, i.e. the Bay of Bengal branch. There is also on an average not much change (slight advancement recently) in onset for **Sub Himalayan Region** in the year 2020 which receive torrential rain. For all of the other regions except

Himalayan region (which are facing a glacier melting which could be a possible explanation) in the present years, the monsoon gets delayed in comparison with a century ago. The amount of the delay happens to be increasing as the distance from a region to the sea increases.

In some regions like (**Barind, Rarh**) before 2000, the Monsoon used to appear before the usual time but after 2000, there is a significant delay in the Monsoon Onset for each of the regions.

Global warming is causing an increase in the average sea temperature, resulting in a reduced temperature gradient between the sea and the land in West Bengal. Consequently, the initial winds that bring the monsoon to various regions of West Bengal are weaker than before. This weaker wind flow takes more time to gather sufficient moisture from the sea, delaying its progression beyond the coastal regions. As a result, the monsoon onset is now delayed in the interior regions of West Bengal, like the **Western Plateau**.

To look into the change of the departure of monsoons over the years, we refer to Figure 6.

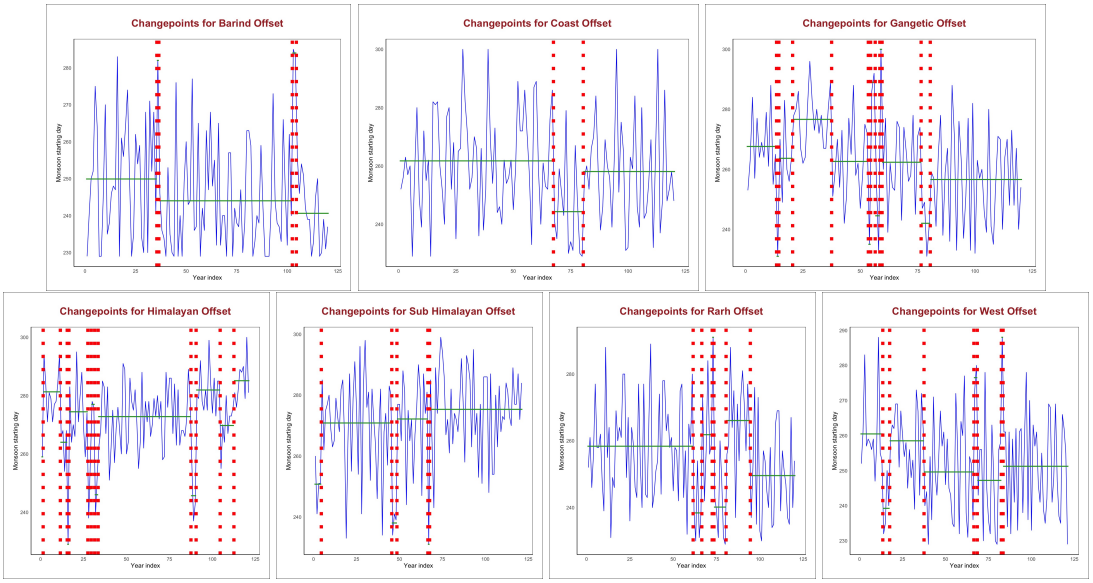


Fig 6. Change Point Detection of Departure for Various Regions

Table 2
Region-wise Onset and Departure Status

Region	Onset Status	Departure Status
Barind	Delayed	Advanced
Coast	No Change	Advanced
Himalayan	Advanced	Delayed
Sub Himalayan	Slightly Advanced	Slightly Delayed
Rarh	Delayed	Advanced
West	Delayed	Advanced
Gangetic	Delayed	Advanced

From the table 2, it turns out for **Barind, Rarh, West Plateau, Gangetic Region**, the Onset has delayed and Departure has appeared early, so **the monsoon length has been shorter than before.** To understand the change in the monsoon duration over the years for each region, we refer to Figure 7.

We observe For all of the regions except the **Himalayan and Sub-Himalayan Regions**, the monsoon duration has been shorter than before. This observation is in accordance with the previous result.This might possibly be due to glacier melting which increases the water content which can then easily be precipitated. Now, to comment on the change of the Monsoon Rainfall over the years, we refer to Figure 8.

We observe that for most of the regions except **Gangetic Region**, there are quite fluctuations in the monsoon rainfall. In General, possibly due to Global Warming, the Monsoon rainfall tends to decrease for most of the regions. However, in the **Gangetic Region** it is pretty stable with a slight decrease rainfall in the last 20 years, which is in accordance with the observations of Yaduvanshi and Ranade (2017). For

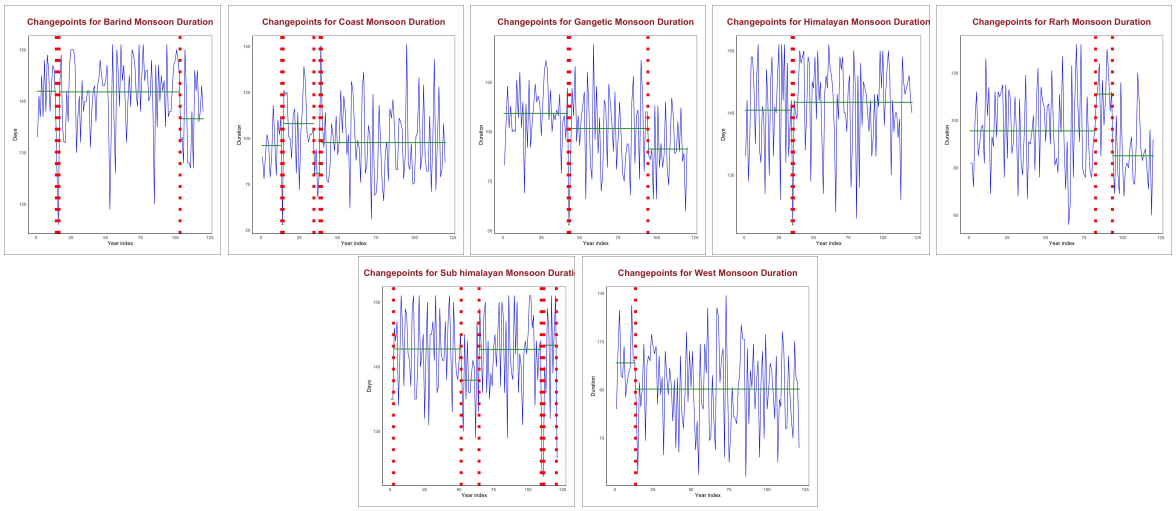


Fig 7. Change Point Detection of Duration of Monsoon for Various Regions

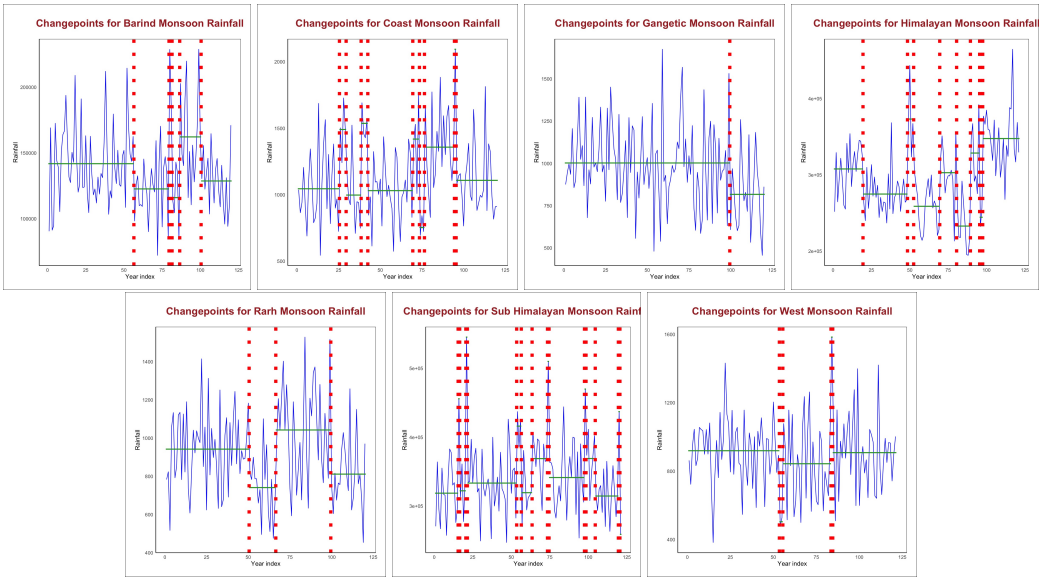


Fig 8. Change Point Detection of Monsoon Rainfall for Various Regions

each of the regions there are several peaks that signify a rise in the monsoon rainfall in that period. For the regions **Barind, Gangetic, Rarh, and Sub Himalayan** the amount of rainfall in Monsoon has decreased recently. For the **Himalayan Region** it has increased a bit. We can explain the observations as follows. The intense heat develops a low-pressure system that attracts rain-bearing winds from the sea. Therefore, rainfall is greatly affected by the flow of these winds and also the occurrence of cyclones. Since the effect of cyclones is the most in the **Coastal region**, that is why in the coastal region the total amount of rainfall in monsoons changes often over several years based on the occurrence of cyclones. e.g. in the years 1966,1925,1926, 1928, 2013-14 in the Coastal Region severe cyclones were experienced along with a significant rise in rainfall (as evident from the plot). These observations are in accordance with the studies in Sarkar and Chakraborty (2021).

The **Himalayan region** in the northern part of West Bengal experiences varying rainfall patterns, influenced by its proximity to the Himalayan mountain range. Areas closer to the foothills receive higher rainfall due to orographic lifting, where moist air is forced to rise over the mountains, leading to enhanced

precipitation. Similar to the Himalayan region, the **Sub-Himalayan region** receives higher rainfall due to its proximity to the mountains. Higher elevation areas might experience lower temperatures and increased precipitation, contributing to the overall water balance in the region. The **Western Plateau** experiences almost constant and low rainfall due to the geographical location away from coasts or water bodies.

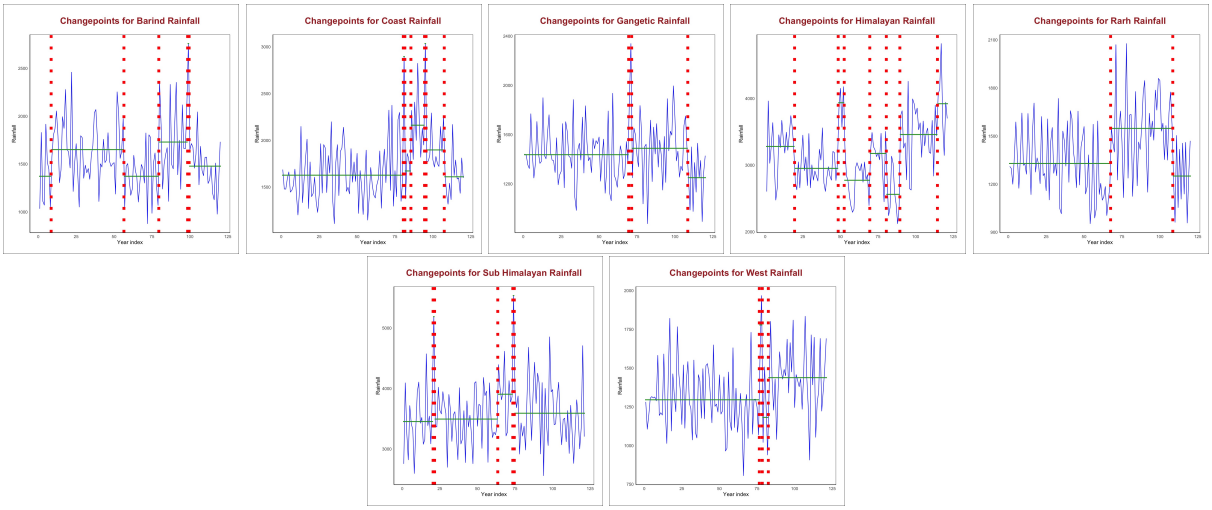


Fig 9. Change Point Detection of Total Rainfall for Various Regions

From figure Figure 8 and Figure 9, we see that for almost all regions except **Barind**, the fluctuations have reduced. However for Coastal,Gangetic,Himalayan, Rarh and Sub-Himalayan, the pattern is quite similar as that of monsoon rainfall, the explanations being quite the same. However, for **Western Plateau** region, we notice an increase in the total rainfall in recent years. This could be due to the combination of global climatic changes, altered atmospheric circulation, increase in convective activity outside the monsoon season, local factors like deforestation and topographical effects, and climate change which can lead to a decrease in monsoon-specific rainfall but an increase in overall annual rainfall due to higher rainfall in pre-monsoon or post-monsoon periods.

6.2. Change in the Rainfall Pattern

For each region, we obtain the clusters based on the similarity of rainfall patterns. We plot the rainfall data for each region-wise cluster. Now, as the rainfall data itself is very varied and sparse, hence there will be added noise when the clustering method will be applied directly to it. That's why instead of applying directly to the daily rainfall data, we instead use the cumulative rainfall data. That is not only stable, it also contains all the information till date in a summed version. So, in that sense more informative and stable. It also gives a more stable clustering than that of the actual daily rainfall data. In table 3 we note the number of clusters for each region. It is to be noted that the number of clusters for regions of South Bengal is higher than the number of clusters of regions of North Bengal. Since Southern Bengal is more affected by cyclones and sea winds which cause variation in rainfall patterns, the southern Bengal region has more clusters. In fig. 10, we show cluster-wise cumulative rainfall plots for the regions.

7. Conclusion

Our analysis of rainfall patterns in West Bengal reveals significant regional variations influenced by global warming and local geographical factors. The onset of the monsoon remains unchanged in the Coastal Region due to consistent wind patterns from the Bay of Bengal, while the Sub-Himalayan Region shows a slight advancement in onset. Other regions, including the Barind and Rarh, have experienced delayed monsoon onset since 2000, likely due to a reduced temperature gradient between the sea and land weakening initial monsoon winds. The delay increases with distance from the sea, particularly affecting the Western Plateau.

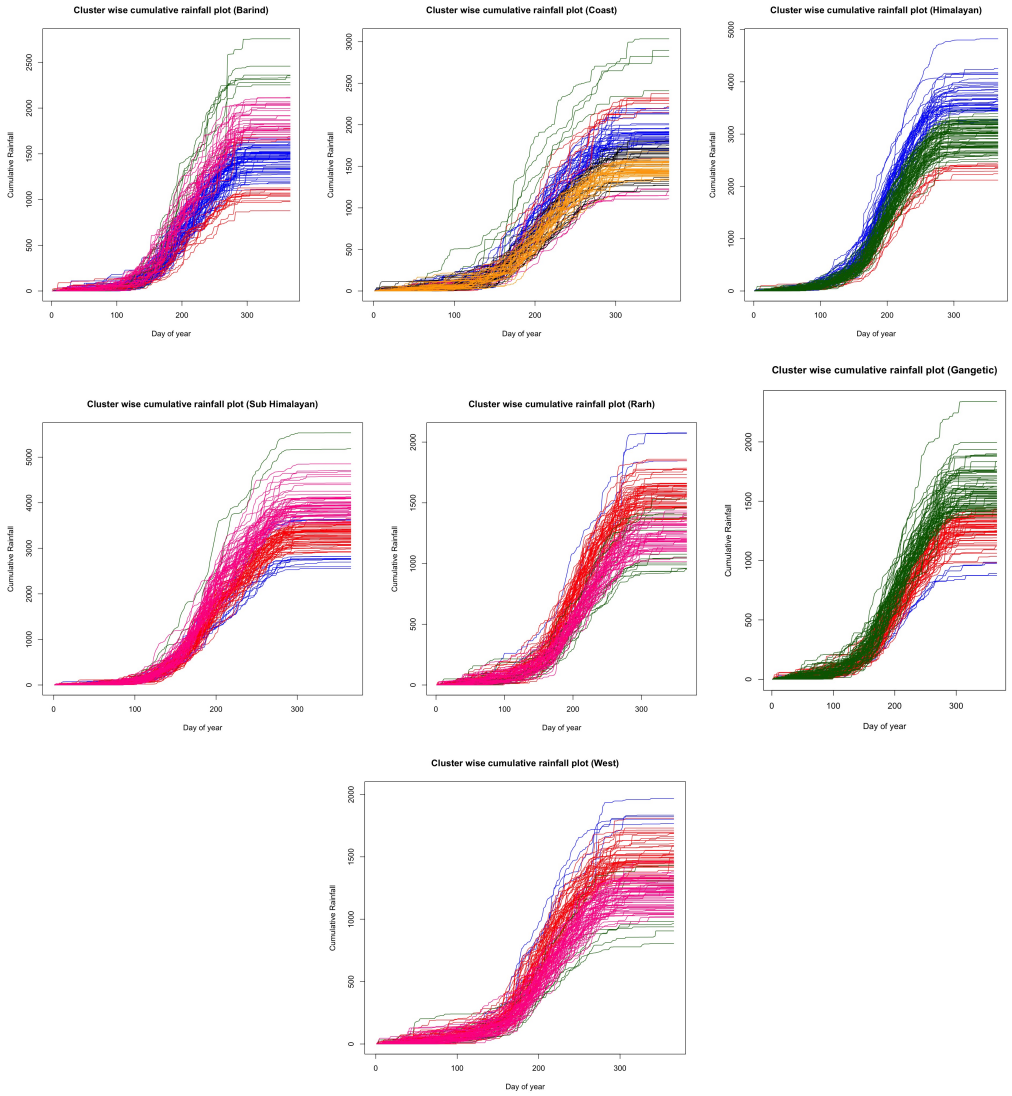


Fig 10. Cluster wise cumulative rainfall plot for the regions

TABLE 3
Region-wise Cluster Number

Region	Number of Clusters
Barind	4
Coast	6
Himalayan	3
Sub Himalayan	4
Rarh	4
West	4
Gangetic	3

Offsets have generally advanced, shortening the monsoon duration in most regions except the Himalayan and Sub-Himalayan areas, where glacier melt may contribute to extended rainfall. Monsoon rainfall amounts fluctuate significantly, with a general decrease attributed to global warming, except in the relatively stable Gangetic Region and a slight increase in the Himalayan Region. Total annual rainfall shows reduced fluctuations in most regions, with the Western Plateau experiencing an increase due to enhanced pre-monsoon and post-monsoon rainfall. Clustering analysis indicates more varied rainfall patterns in South Bengal, driven by cyclones and sea winds, compared to North Bengal. This comprehensive study underscores the complex interplay of global and local factors in shaping West Bengal's rainfall dynamics.

References

FEARNHEAD, P. and RIGAILL, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association* **114** 169–183.

SARKAR, C. S. and CHAKRABORTY, A. (2021). Historical Trend of Cyclone over the Bay of Bengal and People's Perception about Coping with Cyclone in Kakdwip Block of South 24 Parganas, West Bengal, India. *Turkish Online Journal of Qualitative Inquiry* **12**.

YADUVANSHI, A. and RANADE, A. (2017). Long-term rainfall variability in the eastern gangetic plain in relation to global temperature change. *Atmosphere-Ocean* **55** 94–109.

8. Appendix

Appendix A: Nonparametrtic Spectral Density Estimate of a Second-order Stationary Time Series

For a stationary time series $\{X_t\}_{t\in\mathbb{Z}}$, the autocovariance function $\gamma(h)$ is defined as

$$\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t)$$

If for a stationary time series $\{X_t\}_{t\in\mathbb{Z}}$, the autocovariance is summable i.e. $\sum_{h\in\mathbb{Z}} |\gamma(h)| < \infty$, then the spectral density of the process is defined as

$$f(\theta) = \frac{1}{2\pi} \sum_{u\in\mathbb{Z}_{\geqslant 0}} \gamma(u) e^{-iu\theta} \qquad \theta \in [-\pi, \pi]$$

The spectral densities of a time series can be estimated using Kernel Density Estimates. The periodogram $(I(\theta))$ based on samples $X_1, \cdots, X_n$ from a stationary time series is defined as

$$I(\theta) = |J(\theta)|^2$$

where

$$J(\theta) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t e^{-it\theta}$$

Then the nonparametric kernel density estimator of $f(\theta)$ is defined as

$$\hat{f}(\theta) = \frac{1}{n} \sum_{k=-\lfloor \frac{n-1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} K_h \left(\theta - 2\pi \frac{k}{n} \right) I \left(2\pi \frac{k}{n} \right)$$

where

$$K_h = \frac{1}{h} K \left(\frac{\cdot}{h} \right)$$