Analysis of Precipitation Extremes for Indian Cities

Bharat Jambhulkar
Sharvil Ozarkar
Shrirang Pund

Department of Statistics,
Savitribai Phule Pune University

Project submitted to:
Indian Society for Probability and Statistics
M.Sc. Student Project Competition 2024

Abstract

High-impact weather events are intrinsic to the climate system, occurring at various spatial and temporal scales, with the potential to cause severe loss of life, property damage, and major disruptions to communication and transport systems. In the context of global climate change, understanding the patterns of extreme weather events has become increasingly critical. This study analyzes very heavy rainfall events across 50 Indian cities using daily rainfall data from 1951 to 2023 to identify long-term trends and groupings of cities that exhibit similar occurrences of extreme rainfall events.

Two extreme value modeling approaches—block maxima and peaks over threshold—were employed to analyze the intensity of these events. The Generalized Extreme Value (GEV) distribution and Generalized Pareto Distribution (GPD) were used to model the events and calculate return levels ([1] Coles, 2001), which are essential for estimating the recurrence intervals of extreme rainfall. Our findings highlight spatial and temporal variations in extreme rainfall events, offering insights critical for urban planning and disaster management.

Contents

1 Introduction

Very heavy rainfall events pose significant hazards, often leading to landslides and severe infrastructure damage. These events also contribute to widespread flooding, damaging properties, disrupting transportation, and overwhelming essential services. Urban areas, in particular, suffer from inadequate drainage systems, which exacerbate property damage and economic losses. Therefore, analyzing rainfall trends is crucial for improving disaster preparedness and mitigation strategies.

India's rainfall patterns are complex and diverse, driven largely by the Southwest Monsoon. While regions like the Western Ghats and northeastern states experience high rainfall, northwestern areas such as Rajasthan remain arid. This uneven distribution creates a dual challenge of managing both droughts and floods, impacting agriculture, water resources, and disaster management efforts. According to the India Meteorological Department (IMD), 'very heavy rainfall' refers to 115.6–204.4 mm of rain in 24 hours, with 'extremely heavy rainfall' exceeding 204.4 mm in 24 hours [2].

This study aims to explore rainfall patterns in highly populated Indian cities, focusing on trends in very heavy rainfall events. Cities like Delhi, Faridabad, and Meerut have shown decreasing trends, while cities such as Srinagar and Nagpur display an increase. Notably, Mumbai and Pune have been analyzed using both the block maxima and peaks-over-threshold methods to model extreme rainfall events, providing insight into their return levels. Additionally, decade-wise quartile clustering of very heavy rainfall events revealed shifting patterns in these events, with cities transitioning between clusters, reflecting dynamic rainfall changes over time. These findings highlight the evolving nature of urban rainfall distribution and its implications for disaster management strategies.

2 Theory

Statistical extreme value theory is a field of statistics dealing with extreme values, i.e., large deviations from the median of probability distributions. Extreme value distributions are the limiting distributions for the minimum or the maximum of large collections of independent random variables from the same arbitrary distribution. By definition extreme value theory focuses on limiting distributions (which are distinct from the normal distribution). Two approaches exist for practical extreme value applications. The first method relies on deriving block maxima (minima) series, the second method relies on extracting peak values above (below) a certain threshold from a continuous record. (The following sections are taken from ([1] Coles, 2001) & ([3] Tayybeh, Mahnoosh, Sedigheh and Rizwan, 2024).

2.1 Block Maxima

The block maxima(BM) approach discovers extreme values by taking the maximum values from observational data that have been structured into a specific block or period. The Fisher-Tippett theorem states the rescaled sample maxima converge in distribution to a variable having a generalized extreme value distribution.

Theorem 1: (Fisher-Tippett Theorem, 1928) Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random sequence with distribution function F and $M_n = \max(X_1, X_2, \ldots, X_n)$. The normalizing sequences $a_n > 0$, $b_n \in R$ such that $M_n - b_n/a_n$ converges in distribution, so that

$$\lim_{n\to\infty} P\left(\frac{\max(X_1, X_2, \ldots, X_n) - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) \to G(x) \tag{1}$$

where G can be described as a non-degenerate distribution function. The convergence in (1) occurs if and only if $n\{1 - F^n(a_n x + b_n)\} = -\log G(x)$. It is understood that G must belong to one of the three categories of limiting distributions, including Gumbel, Fréchet, and Weibull. These three categories can be combined to form a single generalized extreme value distribution(GEV). The cumulative distribution function of the three-parameter generalized extreme value distribution with location parameter μ, shape parameter ξ, and scale parameter σ is given by

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{2}$$

Define for

$$\left\{x : 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$$

where $\mu \in R$, $\sigma > 0$, and $\xi \in R$ may be any real numbers, with the case $\xi = 0$ being interpreted as the limit $\xi \to 0$,

$$G(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \tag{3}$$

which is widely called the Gumbel distribution. The shape parameter ξ (extreme value index) can decide the tail behavior of the distribution. The case $\xi > 0$ is that of the

polynomially decreasing tail function and therefore corresponds to a long-tailed parent distribution (Fréchet). The case $\xi < 0$ is the case of a finite upper endpoint and therefore short-tailed (Weibull) respectively. The log-likelihood function for GEV distribution is given by:

$$l(\mu, \sigma, \xi) = -n \log \sigma - \sum \log \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right] - \frac{1}{\xi} \sum \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]^{-1/\xi}, \quad (4)$$

where n is the number of block maxima $x_1, x_2, \ldots, x_n$. Based on the extreme value theory that derives the GEV distribution, we can fit a sample of extremes to the GEV distribution to obtain the parameters that best explain the probability distribution of the extremes. From the fitted distribution, we can estimate how often the extreme quantiles occur with a certain return level. Estimates of extreme quantiles of the annual maximum distribution are then obtained by inverting (1):

$$r_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - (-\log(1-p))^{-\xi}\right], & \text{for } \xi \neq 0, \\ \mu - \sigma \log(-\log(1-p)), & \text{for } \xi = 0, \end{cases} \quad (5)$$

where $G(x_p) = 1 - p$. In common terminology, r_p is the return level associated with the return period $1/p$, since to a reasonable degree of accuracy, the level r_p is expected to be exceeded on average once every $1/p$ years. More precisely, r_p is exceeded by the annual maximum in any particular year with probability p.

Since quantiles enable probability models to be expressed on the scale of data, the relationship of the GEV model to its parameters is most easily interpreted in terms of the quantile expressions in (5).

In particular, defining $y_p = -\log(1-p)$, so that

$$r_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{for } \xi \neq 0, \\ \mu - \sigma \log y_p, & \text{for } \xi = 0, \end{cases} \quad (6)$$

it follows that, if r_p is plotted against y_p on a logarithmic scale—or equivalently, if r_p is plotted against $\log y_p$—the plot is linear in the case $\xi = 0$. If $\xi < 0$ the plot is convex with an asymptotic limit as $p \to 0$ at $\mu - \frac{\sigma}{\xi}$; if $\xi > 0$ the plot is concave and has no finite bound. This graph is called the *Return Level* plot.

2.2 Peaks over Threshold

The peaks-over-threshold(POT) approach is also called the method of excess. The POT approach evaluates the distribution of exceedances above the specified threshold. Furthermore, it can be shown that for some sufficiently large threshold u, the distribution of the values exceeding the threshold is approximated to a Generalized Pareto(GP) distribution.

Theorem 2: Let $X_1, X_2, \ldots$ be a sequence of independent random variables with common distribution function F, and let

$$M_n = \max(X_1, \ldots, X_n).$$

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies Theorem 1, so that for large n,

$$\Pr(M_n \leq z) \approx G(z),$$

where

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{7}$$

for some $\mu, \sigma > 0$ and ξ. Then, for large enough u, the distribution function of $(X-u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1+\xi\frac{y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{8}$$

defined on $\{y : y > 0 \text{ and } (1+\xi\frac{y}{\tilde{\sigma}}) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u-\mu). \tag{9}$$

The family of distributions defined by equation(8) is called the *generalized Pareto family*. The parameters of the generalized Pareto distribution of threshold excesses are uniquely determined by those of the associated GEV distribution of block maxima. In particular, the parameter ξ in (8) is equal to that of the corresponding GEV distribution.

2.3 Non-Stationary Extremes

In the context of environmental processes, it is common to observe non-stationarity– for example, due to different seasons having different climate patterns, or perhaps due to more long-term trends owing to climate change. But what about extremes for which we cannot assume stationarity? To date, no general theory for non-stationary extremes has been established. In practice, it is common to adopt pragmatic 'workarounds' based on the type of non-stationarity observed.

One way of capturing this trend is by allowing the GEV location parameter to vary across time.

$$X_t \sim \text{GEV}(\mu(t), \sigma, \xi),$$

where

$$\mu(t) = \beta_0 + \beta_1 t \tag{10}$$

and t is an indicator of year. In this way, variations over time are modelled as a linear trend in the location parameter of the GEV. As in a standard simple linear regression, β_1 represents the slope – in this case, the annual rate of change in annual maxima of rainfall values. The time-homogeneous model is a special case of this time-dependent model, with $\beta_1 = 0$; since this is nested within the model which allows for a time dependence, the deviance statistic can be used to formally compare models. Not just linear, but we can essentially model any functional relationship of a parameter with any environmental variables like *Wind Speed*, *Temperature*.

Parameter estimation

In the log-likelihood equation(4), we simply replace μ in the above expression with equation(10), giving

$$l(\beta_0, \beta_1, \sigma, \xi) = -n \log \sigma - \sum \log \left[1 + \xi \left(\frac{x_i - (\beta_0 + \beta_1 t)}{\sigma} \right) \right] - \frac{1}{\xi} \sum \left[1 + \xi \left(\frac{x_i - (\beta_0 + \beta_1 t)}{\sigma} \right) \right]^{-1/\xi}, \quad (11)$$

We could then maximise this log-likelihood.

2.4 Threshold Selection

The raw data consist of a sequence of independent and identically distributed measurements $x_1, \ldots, x_n$. Extreme events are defined by defining a high threshold u, for which the exceedances are $\{x_i : x_i > u\}$. Label these exceedances by $x(1), \ldots, x(k)$, and define threshold excesses by $y_j = x(j) - u$, for $j = 1, \ldots, k$. y_j are regarded as independent realizations of a random variable whose distribution can be approximated by a member of the generalized Pareto family. Inference consists of fitting the generalized Pareto family to the observed threshold exceedances, followed by model verification and extrapolation.

This approach contrasts with the block maxima approach through the characterization of an observation as extreme if it exceeds a high threshold.The issue of threshold choice is analogous to the choice of block size in the block maxima approach, implying a balance between bias and variance. In this case, too low a threshold is likely to violate the asymptotic basis of the model, leading to bias; too high a threshold will generate few excesses with which the model can be estimated, leading to high variance. The standard practice is to adopt as low a threshold as possible, subject to the limit model providing a reasonable approximation. Two methods are available for this purpose: one is an exploratory technique carried out prior to model estimation; the other is an assessment of the stability of parameter estimates, based on the fitting of models across a range of different thresholds.

2.4.1 Mean Residual Life plot

The *threshold stability property* of the GPD means that if the GPD is a valid model for excesses over some threshold u_0, then it is valid for excesses over all thresholds $u > u_0$. Denoting by σ_{u_0} the GPD scale parameter for excesses over threshold u_0, the expected value of our threshold excesses, conditional on being greater than the threshold, is

$$E[(X - u) \mid X > u] = \frac{\sigma_{u_0} + \xi u}{1 - \xi}. \quad (12)$$

Thus, for all $u > u_0$, $E[(X - u) \mid X > u]$, is a linear function of u. Furthermore, $E[(X - u) \mid X > u]$ is simply the mean of the excesses of the threshold u, for which the sample mean of the threshold excesses of u provides an estimate. This leads to the *mean residual life plot*, a graphical procedure for identifying a suitably high threshold for modelling extremes via the GPD. In this plot, for a range of candidate values for u, we identify the corresponding mean threshold excess; we then plot this mean threshold excess against u, and look for the value u_0 above which we can see linearity in the plot. (See Figure 10)

2.4.2 Shape parameter stability

A complementary technique is to fit the generalized Pareto distribution at a range of thresholds, and to look for stability of parameter estimates. The argument is as follows. If a generalized Pareto distribution is a reasonable model for excesses of a threshold u_0, then excesses of a higher threshold u should also follow a generalized Pareto distribution. The shape parameters of the two distributions are identical. (See Figure 11)

3 Study Area

India's geographical diversity plays a significant role in its varied rainfall patterns. The country's vast landscape, which includes the Himalayas in the north, the Thar Desert in the west, coastal plains, and the Deccan Plateau, influences the distribution and intensity of rainfall. The study focuses on India's top 50 most populous cities (Figure 1). This includes cities such as Mumbai, Delhi, Chennai, Kolkata, Bangalore, etc.[4] A high population increases the risk of significant damage to infrastructure, loss of life, and economic disruption during heavy rainfall events. Urbanization often leads to reduced natural drainage and increased surface runoff, exacerbating the effects of flooding. Additionally, understanding rainfall patterns in these cities helps in better urban planning, disaster preparedness, and mitigation strategies.

The IMD gridded daily rainfall data is a reliable source for studying rainfall patterns across India. This dataset includes daily rainfall values from 1951 to 2023 and allows for a detailed analysis of long-term trends and extreme weather events. The data consists of high spatial resolution daily gridded rainfall measurements (0.25 x 0.25 degrees), with rainfall amounts recorded in millimetres (mm).[5] The rainfall value recorded at the nearest weather station is used for a city's given latitude and longitude.

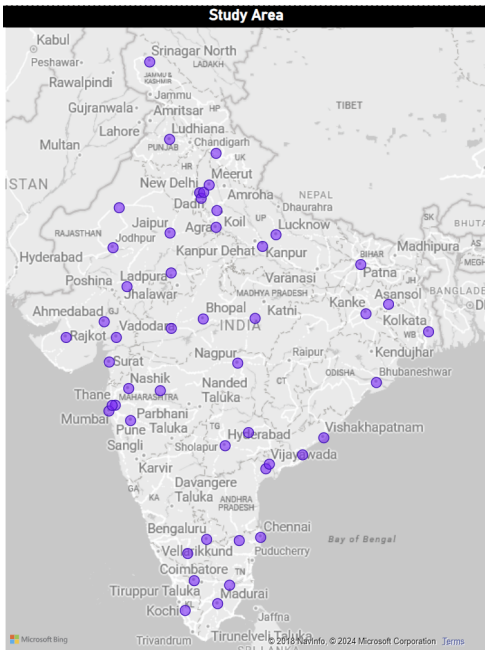


Figure 1:Location of Indian Cities Under Study Area

4 Exploratory Analysis

Daily rainfall is classified into several categories based on the amount of rainfall recorded. Light rain occurs when the daily rainfall is between 2.5 and 15.5 mm, while moderate rain is defined as 15.6 to 64.4 mm of daily rainfall. If the rainfall is between 64.5 to 115.5 mm, it is categorized as heavy rain. Very Heavy rain occurs when daily rainfall ranges from 115.6 to 204.4 mm. Finally, extremely heavy rain is classified when the rainfall amount in a day is 244.5 mm or more.[2]

In our study, we focused on the analysis and modelling of very heavy rainfall events. The data were segmented into decades, covering the period from 1953 to 2023. The initial decade spanned from 1954 to 1963, while the most recent covered 2014 to 2023. We set a threshold of 115.6 mm and counted the frequency of instances where the daily rainfall value exceeded this threshold within each decade.

City	1954-1963	1964-1973	1974-1983	1984-1993	1994-2004	2004-2013	2014-2023
Mumbai	62	34	41	41	29	46	54
Delhi	3	2	2	1	0	0	0
Bengaluru	0	0	0	0	3	0	2
Kolkata	5	3	4	12	3	1	6
Chennai	4	5	9	15	4	12	10
Hyderabad	1	0	1	2	1	1	1
Ahmedabad	4	2	9	4	8	6	4
Pune	3	1	0	1	0	27	1
Jaipur	1	1	3	4	1	1	1
Surat	7	10	19	13	11	22	6
Lucknow	3	3	6	1	2	9	2
Kanpur	2	2	2	1	0	3	1
Nagpur	1	1	2	2	4	5	5
Visakhapatnam	6	1	6	6	1	8	6
Vadodara	3	3	9	2	2	11	2
Bhopal	5	8	6	14	5	5	9
Indore	6	8	8	5	8	9	5
Coimbatore	0	2	3	0	2	0	0
Nashik	0	5	1	0	0	14	3
Patna	2	1	1	2	2	3	2
Ghaziabad	4	3	2	4	6	3	0
Ludhiana	2	2	1	6	2	1	2
Agra	3	2	2	1	3	0	2
Madurai	1	1	0	1	0	1	0
Rajkot	10	3	5	3	5	10	13
Faridabad	4	3	3	2	0	0	1
Meerut	7	8	9	7	6	1	1
Kalyan-Dombivli	30	21	14	28	38	29	26
Vijayawada	1	4	2	5	4	5	2
Aurangabad	0	0	0	3	1	0	1
Thane	51	32	33	24	27	28	50
Dhanbad	3	3	4	1	2	2	5
Jodhpur	0	0	1	2	0	0	3
Kota	2	5	2	0	2	1	4
Guntur	1	4	1	3	4	4	2
Mysuru	1	0	0	2	1	0	2
Vellore	0	2	0	1	1	3	1
Ranchi	4	2	1	0	2	1	0
Jabalpur	1	5	5	10	4	8	2

City	1954-1963	1964-1973	1974-1983	1984-1993	1994-2004	2004-2013	2014-2023
Bhubaneswar	6	9	1	3	11	10	11
Salem	0	2	0	2	1	1	1
Warangal	1	2	2	3	1	1	2
Srinagar	0	0	0	0	0	1	1
Tiruchirappalli	1	1	2	2	1	7	0
Udaipur	0	1	1	0	0	4	1
Dehradun	12	13	1	5	4	9	8
Bikaner	0	0	1	0	0	1	0
Aligarh	5	4	2	2	1	0	2
Kottayam	1	3	2	1	4	4	7
Kakinada	1	2	3	6	5	6	2

Table 1: Very Heavy Rainfall Events Across Indian Cities Arranged In Decreasing Order of Population

4.1 Quartile Based clustering of Decade Wise Exceedances

To enhance our understanding of the exceedances and to facilitate cross-city comparisons, we clustered the cities based on the quartiles of their decade-wise exceedances. The clustering was done as follows:

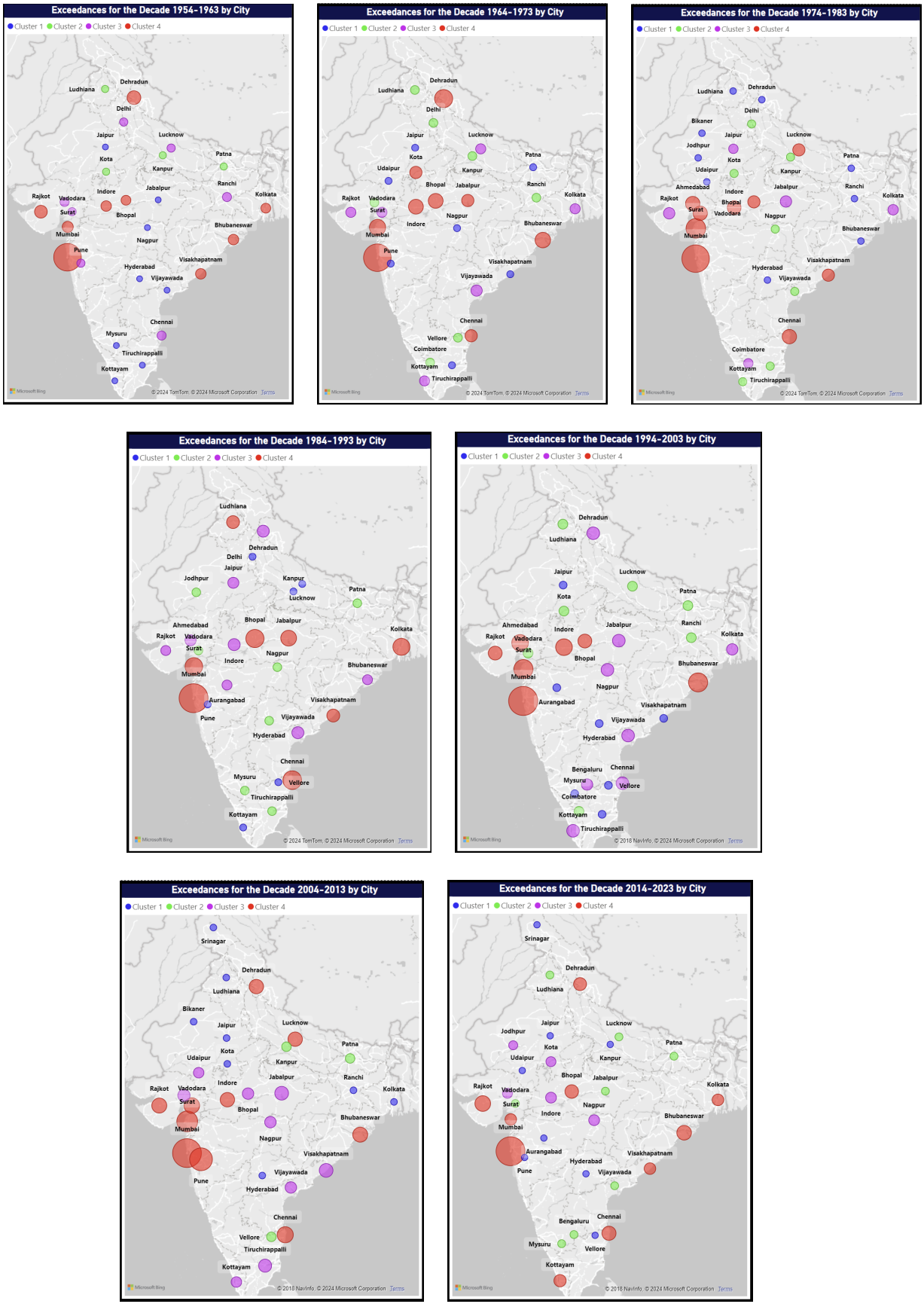
- **Cluster 1:** Exceedances $\leq$ first quartile value (Q1)
- **Cluster 2:** Q1 $<$ Exceedances $\leq$ second quartile value (Q2)
- **Cluster 3:** Q2 $<$ Exceedances $\leq$ third quartile value (Q3)
- **Cluster 4:** Q3 $<$ Exceedances

The quartile values for the 7 decades are given in the table below based on which the clustering of the cities has been done.

Decade	1954-1963	1964-1973	1974-1983	1984-1993	1994-2004	2004-2013	2014-2023
Q1	1	1	1	1	1	1	1
Q2	2	2	2	2	2	3	2
Q3	4.75	4.75	5	5	4	8.75	5

Table 2: Decade wise values of Quartiles

The plots in Figure 2 illustrate the results for this clustering. Cities that remain within the same clusters across decades suggest stability in the occurrence of very heavy rainfall events over the years. Conversely, cities that shift between clusters across decades may suggest a potential shift in the pattern of decade-wise exceedances, warranting further investigation to identify underlying contributing factors.



*Bubble size of a city is proportional to count of decade wise exceedances.
*Cities which are very close to each other have not been displayed in the plots

Figure 2: Quartile Based clustering of Decade Wise Exceedances

To understand the behavior of very heavy rainfall events across Indian cities over the decades, regression line is fitted using the number of very heavy rainfall events in each decade as the response variable and decades as the regressor(*1954-1963:1,. . . ,2013-2023:7*). This analysis was conducted for all 50 cities under study. Cities like Delhi, Faridabad, Meerut, and Aligarh exhibited a significant negative slope, indicating a decrease in the occurrence of very heavy rainfall events over the decades. In contrast, Nagpur and Kottayam showed a significant positive trend, reflecting an increase in these extreme events over decades(Figure 3). Table 3 shows the R^2 value of the fitted lines and the p-values of the slope parameter.

	Delhi	Faridabad	Meerut	Aligarh	Nagpur	Kottayam
R^2	0.9073	0.7788	0.6867	0.6639	0.9167	0.6436
p-value	0.0009	0.0085	0.0212	0.0255	0.0007	0.0299

Table 3: R^2 and p-value

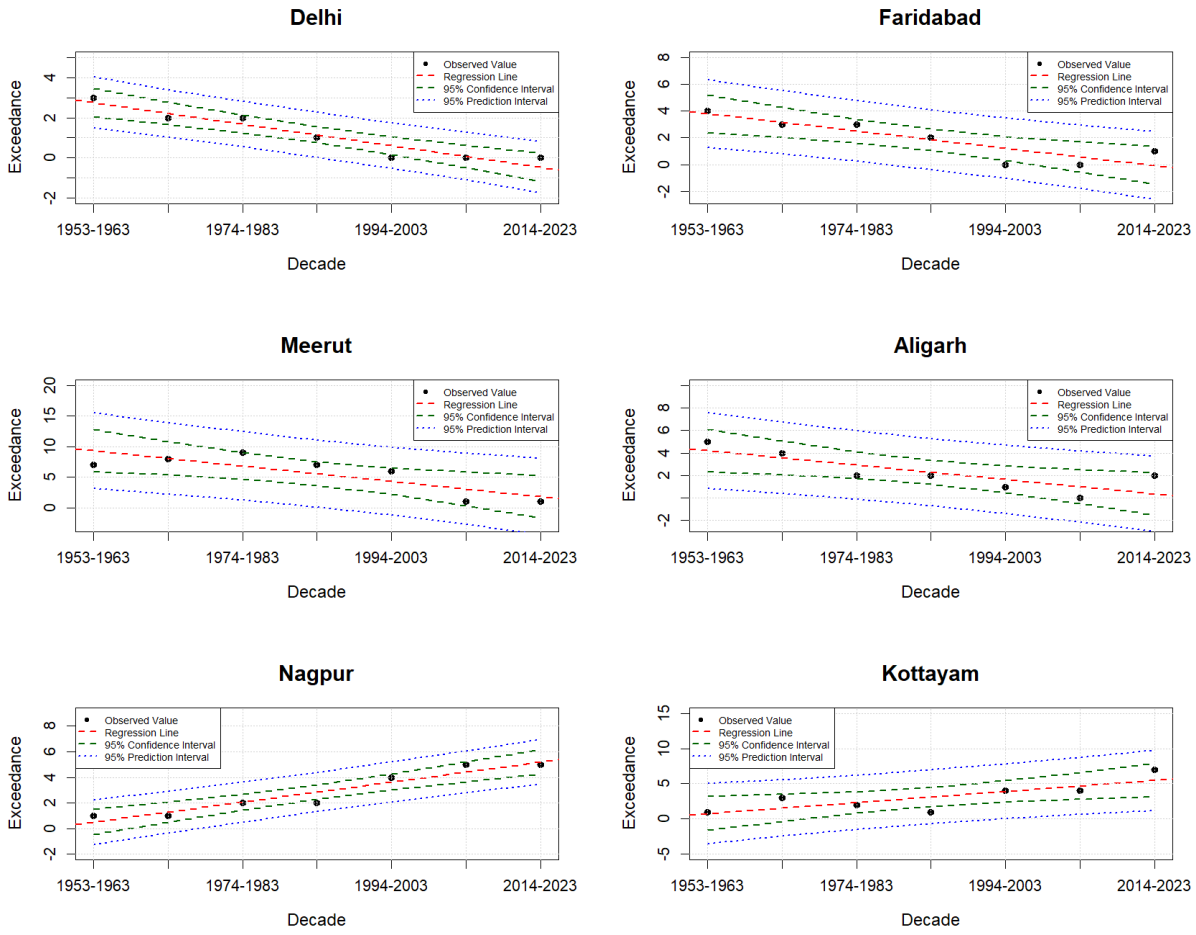


Figure 3: Regression Plots For Decadewise Exceedances

4.2 Insights

- The decade from 2004 to 2013 is notably unusual, with cities such as Pune, Nashik, Surat, Lucknow, Vadodara, Rajkot, Tiruchirappalli, and Udaipur experiencing a significant increase in the frequency of heavy rainfall events.
- The decade from 1984 to 1993 also showed an unusual number of exceedances in cities like Kolkata, Chennai, Bhopal, Ludhiana, Aurangabad, and Jabalpur.
- Bhubaneswar has exhibited a marked rise in heavy rainfall occurrences over the last three decades, suggesting a shift in its climatological patterns.
- Tiruchirappalli has shown a general upward trend in the frequency of heavy rainfall events from 1953 to 2013, with a slight deviation in the most recent decade.
- In Srinagar, no instances of rainfall exceeding the 115.6 mm threshold were observed during the first five decades. However, exceedances were noted in the last two decades, indicating a deviation from historical patterns and suggesting potential influences of emerging climatic or environmental factors.
- A significant decrease was observed in Delhi, Faridabad, Meerut, and Aligarh, while Srinagar, Nagpur, and Kottayam experienced a notable increase. These findings highlight the need for a serious investigation into the environmental changes impacting these cities.(Figure 3)
- Cities such as Ahmedabad, Jaipur, Kanpur, Visakhapatnam, Indore, Coimbatore, Patna, Ghaziabad, Ludhiana, Agra, Madurai, Kalyan-Dombivli, Vijayawada, Aurangabad, Dhanbad, Jodhpur, Kota, Guntur, Mysuru, Vellore, Salem, Warangal, and Kakinada exhibit minimal variation in heavy rainfall events over the decades, indicating relative stability in their rainfall patterns.

This analysis underscores the importance of proactive measures for cities exhibiting an increasing trend in very heavy rainfall events, as they may face heightened risks in the future. Conversely, cities with stable occurrences over the decades may not need to allocate excessive resources for such events. Overall, the findings reveal significant regional variations in climate patterns across India, highlighting the potential hazards of escalating heavy rainfall events and the broader implications of ongoing climatic changes. Understanding these shifts is crucial for developing adaptive strategies to mitigate potential climate-related risks.

5 Modelling of Extremes

5.1 Block Maxima approach for Mumbai

Mumbai, a densely populated city, is highly vulnerable to extreme weather events, particularly heavy rainfall. These downpours frequently overwhelm the city's drainage system, leading to waterlogging and flooding that disrupt daily life, halt transportation, and pose health and safety risks. The resulting delays and infrastructure damage severely impact the city's economy. Given Mumbai's vulnerability, there is an increasing need to better understand and model these extreme rainfall events, both to anticipate their occurrence and to mitigate their impact.

Augmented Dicky Fuller (ADF) and Pettitt's tests were used to check the assumptions of stationarity and change in the location parameter of the distribution on yearly maximum rainfall values for Mumbai city. Pettitt's test is a non-parametric test that detects changes in the location parameter of a distribution. The ADF test resulted in a p-value of 0.0486, which is less than 0.05, indicating rejection of the null hypothesis (H_0) that a unit root is present in the data. For Pettitt's test, p=0.2971 indicates no evidence to reject the null hypothesis (H_0) of no change in the location parameter of the distribution. Thus, a stationary GEV model was fitted to the yearly maximum rainfall values of Mumbai city.

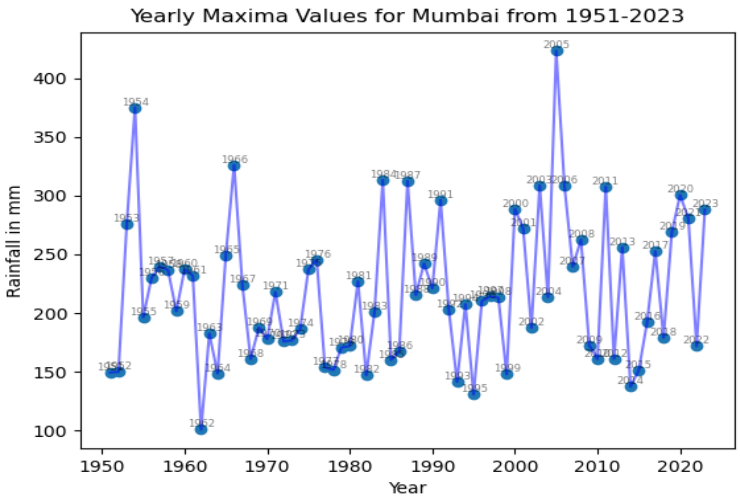


Figure 4

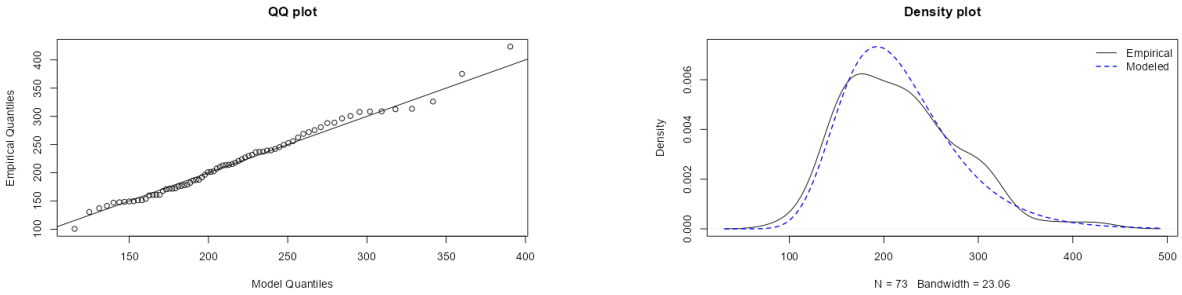
We used the `fevd()` function in the `extRemes`[6] package to fit the GEV distribution. Maximum likelihood estimates for the fitted distribution are:

ξ	μ	σ
-0.0361	190.7472	50.1745
(0.0901)	(6.649)	(4.828)

Table 4: Parameter Estimates for GEV [1]

[1]Standard Errors are reported in brackets

To assess the adequacy of the model, QQ plots and density plots were examined. Kolmogorov-Smirnov(KS) test resulted into a p-value=0.968 which fails to reject H_0 that there is no difference between the two distributions.



Return levels (in mm) for 2-year and 5-year period are 209mm and 264mm, respectively. The interpretation for which is that the annual maxima exceeding 209mm is a 2-year event, meaning that the expected number of times annual maxima exceeds 209mm in a 2-year period is one.

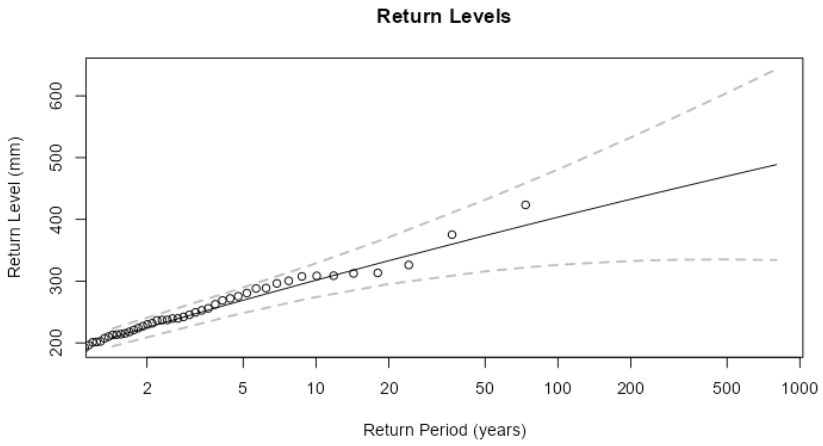


Figure 5: Return Level plot for Mumbai(GEV)

Return periods are incorrectly interpreted as "100-year event is an event which happens only once in 100 years," which may lead to inaccurate assessment of risks. A more nuanced way of looking at this is to consider a time period within which a risk is evaluated. For example, the return level plot above suggests that the annual rainfall maxima exceeding approximately 332 mm is a 20-year event, but the same event will have a probability of approximately 40.1% to be exceeded at least once within 10 years. This is calculated using the following formula:

$$1 - (1 - p)^n$$

where n is the number of return period blocks within a time period (10 for 10 years with return period block of size 1 year), and p is 0.05 (for a 20-year event).[7]

5.2 Block Maxima approach for Pune

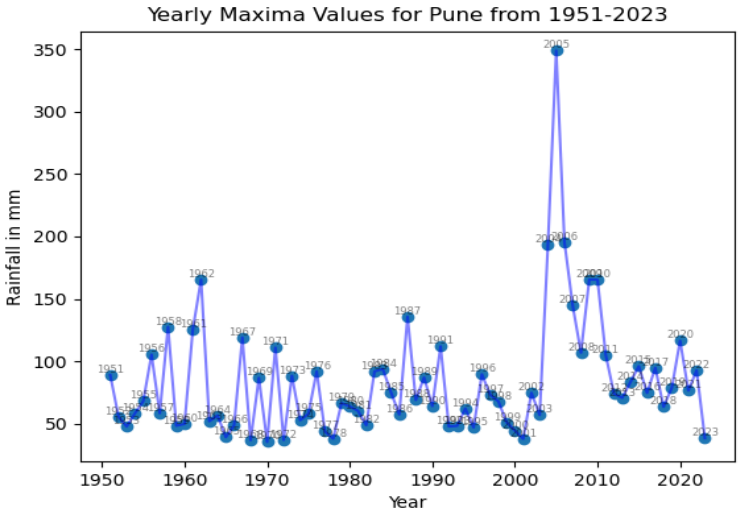


Figure 7

Now consider annual maximum rainfall values for Pune city. ADF test gives a p-value=0.2868 which fails to provide evidence to reject the null hypothesis, which means there exists a unit root indicating non-stationarity in the data. Pettitt's test also provides evidence for a probable change in the location parameter (p=0.0067). We consider multiple models where the distribution parameters are functions of time. These include linear as well as quadratic relations [8].

Models	**Parameter forms**
BM00	$\mu = \mu_0$ (constant), σ (constant), ξ (constant)
BM10	$\mu = \mu_0 + \mu_1 T$, σ (constant), ξ (constant)
BM01	$\mu = \mu_0$ (constant), $log\sigma = \sigma_0 + \sigma_1 T$, ξ (constant)
BM11	$\mu = \mu_0 + \mu_1 T$, $log\sigma = \sigma_0 + \sigma_1 T$, ξ (constant)
BM20	$\mu = \mu_0 + \mu_1 T + \mu_2 T^2$, σ (constant), ξ (constant)
BM02	$\mu = \mu_0$ (constant), $log\sigma = \sigma_0 + \sigma_1 T + \sigma_2 T^2$, ξ (constant)
BM12	$\mu = \mu_0 + \mu_1 T$, $log\sigma = \sigma_0 + \sigma_1 T + \sigma_2 T^2$, ξ (constant)
BM21	$\mu = \mu_0 + \mu_1 T + \mu_2 T^2$, $log\sigma = \sigma_0 + \sigma_1 T$, ξ (constant)
BM22	$\mu = \mu_0 + \mu_1 T + \mu_2 T^2$, $log\sigma = \sigma_0 + \sigma_1 T + \sigma_2 T^2$, ξ (constant)

Table 5: Stationary and Non-stationary BM Models

Since we only model the annual maxima for the years 1951–2023, we consider a vector T which has values 1 to 73 sequentially representing the years. Then we fit model parameters linearly against T and T^2. (Refer to Table 5)

Observe Table 6, the AIC(Akaike Information Criterion) and BIC(Bayesian Information Criterion) values for different models stay almost the same. Is the non–stationary model worthwhile? In other words, does the non–stationary model provide an improvement in fit over the simpler models? With models $M_0 \subset M_1$, we define the deviance statistic as:

$$D = 2\left\{\ell_1(M_1) - \ell_0(M_0)\right\},$$

where $\ell_1(M_1)$ and $\ell_0(M_0)$ are the maximised log–likelihood under models M_1 and M_0 respectively. The asymptotic distribution of D is given by the χ^2_k distribution with k degrees of freedom, where k is the difference in dimensionality of M_1 and M_0; thus, calculated values of D can be compared to critical values from χ^2_k, where large values of D suggest that model M_1 explains substantially more of the variation in the data than M_0.[1]

	BM00	**BM10**	**BM01**	**BM11**	**BM20**	**BM02**	**BM12**	**BM21**	**BM22**
ξ	0.3416	0.3053	0.3281	0.3187	0.2987	0.3927	0.3183	0.3325	0.3624
	(0.1251)	(0.1196)	(0.1254)	(0.1180)	(0.1283)	-	-	(0.1192)	-
μ	60.2100	-	59.6347	-	-	59.6740	-	-	-
	(3.1940)	-	(3.2942)	-	-	-	-	-	-
σ	23.0569	23.1455	-	-	23.2087	-	-	-	-
	(2.7691)	(2.7047)	-	-	(2.7838)	-	-	-	-
μ_0	-	54.9829	-	50.0608	57.4368	-	50.0447	56.9200	59.1997
	-	(4.7530)	-	(5.0800)	(8.5104)	-	-	(5.8962)	-
μ_1	-	0.1642	-	0.2949	0.0051	-	0.2953	-0.3601	-0.6309
	-	(0.1171)	-	(0.1383)	(0.4596)	-	-	(0.4361)	-
μ_2	-	-	-	-	0.0020	-	-	0.0098	0.0141
	-	-	-	-	(0.0055)	-	-	(0.0067)	-
σ_0	-	-	20.5919	16.5238	-	19.0107	16.5933	13.4249	16.6999
	-	-	(1.2548)	(1.2486)	-	-	-	(1.3312)	-
σ_1	-	-	1.0029	1.0084	-	1.0086	1.0082	1.0134	0.9950
	-	-	(1.0051)	(1.0049)	-	-	-	(1.0065)	-
σ_2	-	-	-	-	-	0.9999	1.0000	-	1.0002
	-	-	-	-	-	-	-	-	-
NNLH	358.3590	357.3078	358.1839	355.8636	357.2437	358.1494	355.8639	354.8018	354.3833
AIC	722.7180	722.6155	724.3677	721.7272	724.4874	726.2988	723.7277	721.6037	722.7666
BIC	729.5894	731.7774	733.5295	733.1795	735.9397	737.7511	737.4705	735.3464	738.7998

Table 6: Parameter Estimates for BM Models[2]

[2]NNLH: Maximized Non-Negative Likelihood

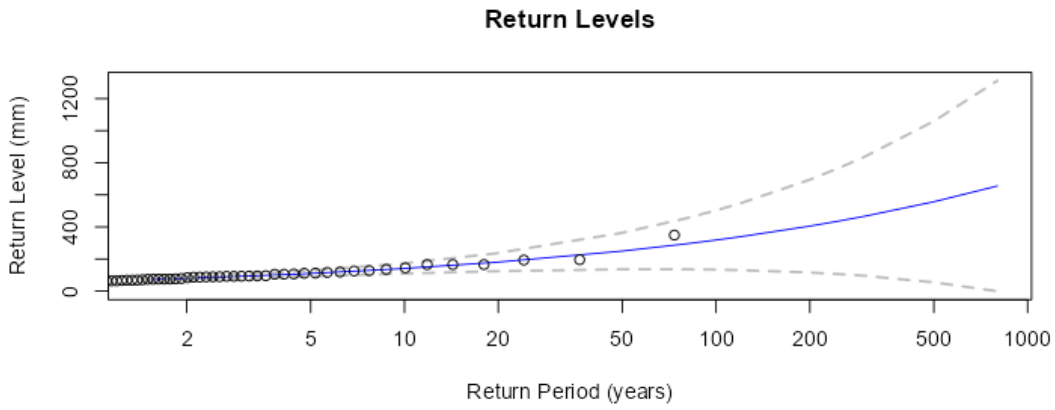


Figure 8: Return Level plot for Pune(BM00)

Return Levels for 2, 5 and 10 year periods are 69mm, 105mm and 138mm respectively(Figure 8). `extRemes` package[6] provides a function `lr.test()` which calculates the deviance statistic and outputs the result in a format presented below. Calculating the deviance statistic for BM00 and BM22, we find that the BM22 with its parameters being quadratic functions of time do not improve upon the variation explained. Thus, we stick to simpler models for inference.(Refer to Figure 8)

D	Chi-square critical value	Alpha	DF	*p*-value
7.9514	9.4877	0.0500	4	0.09337

Table 7: Output for `lr.test(BM00, BM22)`

5.3 Peaks Over Threshold Approach for Mumbai

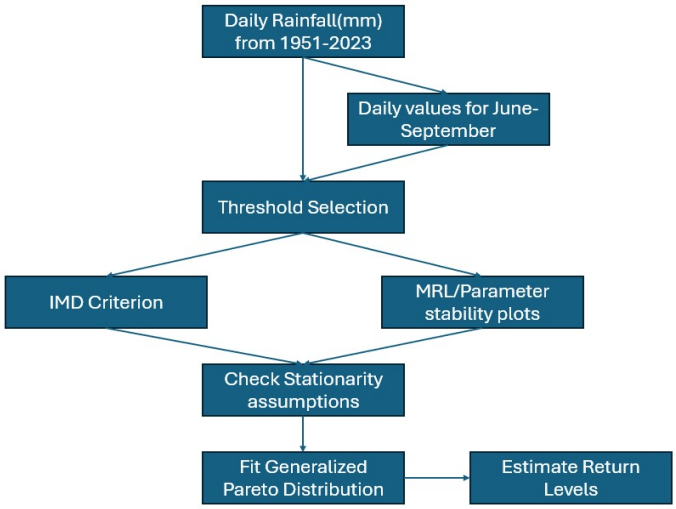


Figure 9: Flowchart for POT approach

5.3.1 Using daily rainfall data for years 1951-2023

The steps involved in fitting GPD distribution is provided in Figure 9. We model the exceedances over threshold defined by IMD for very heavy rainfall (115.6mm) by fitting a Generalised Pareto Distribution (GPD). Once again, we run ADF and Pettitt's test on values above the threshold to check for stationarity and constant threshold parameter assumption. ADF test yields a p-value of 0.01 which is less than 5% level of significance implying rejection of H_0. Pettitt's test yields a p-value of 0.3452 which leads to failure of rejection of H_0 implying constant threshold parameter. Thus, we fit a stationary GPD distribution with threshold parameter of 115.6. But, KS test yields a p-value$< 2.2\times10^{-11}$ which means the theoretical distribution does not fit the empirical distribution well for $u = 115.6$.

We discard the previous model with IMD defined threshold and move on to choosing threshold using MRL(Mean Residual Life) and parameter stability plots. R Library `evmix`[9] provides functions such as `mrlplot()` and `tshapeplot()` which provide the following plots as outputs:

Figure 10

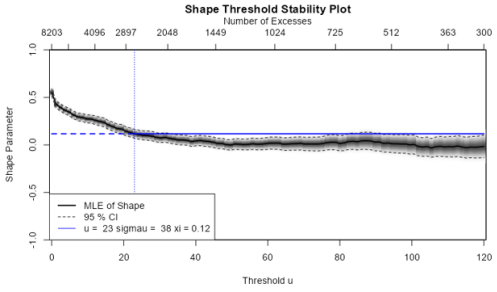


Figure 11

The above plots suggest $u = 23$ to be an optimal threshold. We then perform ADF test which resulted into a p=0.01 implying rejection of H_0. Pettitt's test, with p=0.3935 failed to detect change in threshold parameter(In GPD, the threshold parameter can be treated as the location parameter). Model adequacy is checked by calculating the KS test statistic which gives a p-value=0.1267 indicating a good fit. Figure 12 shows the QQ and density plots for the GPD.

Parameter Estimates for GPD

ξ	Threshold	σ
0.1157	23	37.6988
(0.0239)		(1.158)

Table 8: Daily rainfall for 1951-2023 for Mumbai

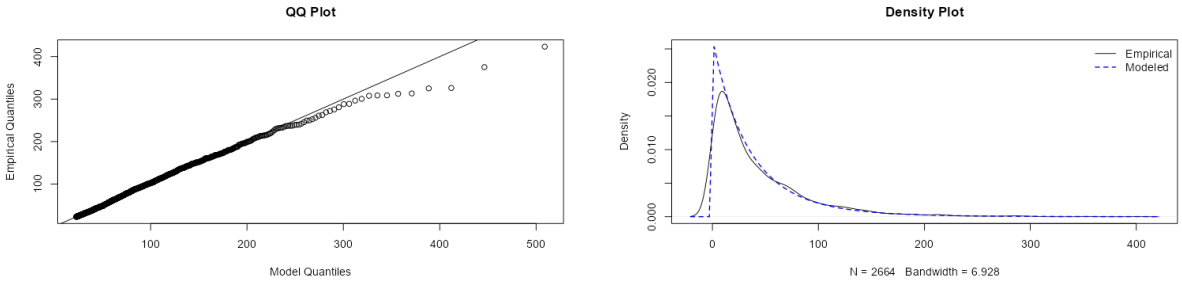


Figure 12

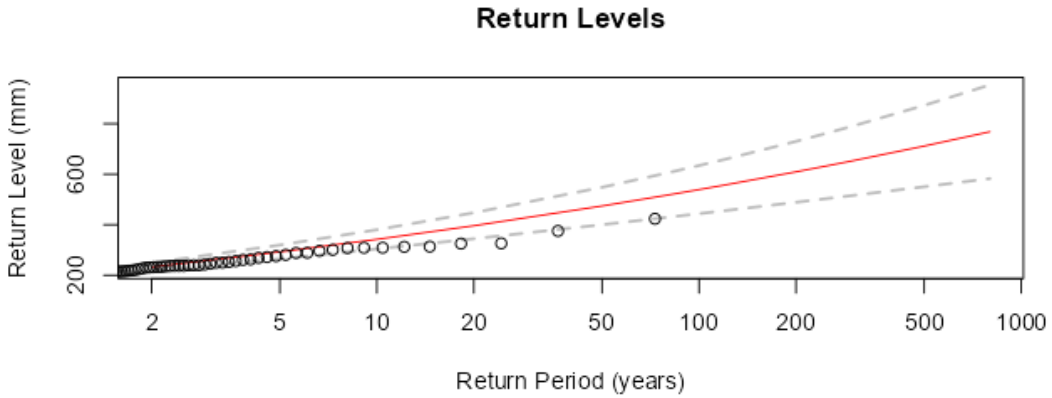


Figure 13: Return Level plot for Mumbai (1951-2023)

The 2, 5 and 10 year return levels for Mumbai city are 233 mm, 292 mm and 342 mm respectively. But one might question, how does a threshold of 23mm of rainfall make sense when we are trying to model extreme rainfall cases? One reason for getting such a low threshold might be because here we modelled daily rainfall for 73 years (26663 values in total) and since it rains only in the monsoon season in India, approximately 30% of 26663 values are zeroes. Thus, due to large number of zeroes $u = 23$ becomes the 90^{th} percentile of the data. In the next section, we only consider values for months June-September(Monsoon months for India) for years 1951-2023 to see whether the return levels differ from the one obtained from fitting a model with daily values for 73 years.

5.3.2 Using data only from months June-September for years 1951-2023

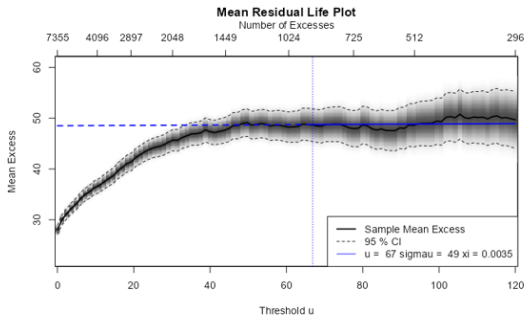


Figure 14

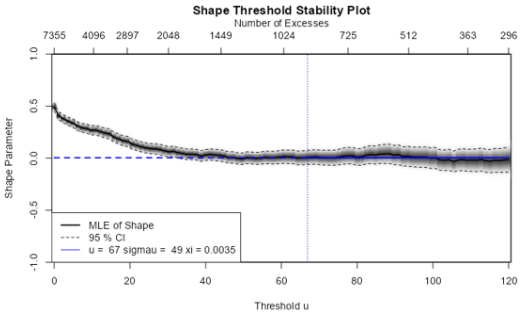


Figure 15

MRL and parameter stability plots suggest an optimal threshold $u = 67$. This value of threshold is the 90^{th} percentile of the data and 115.6(IMD criterion) is approximately the 96^{th} percentile. KS test(p=0.203) indicates a good fit for the observed data. Refer to Table 9 for parameter estimates for this model. We also fit a GPD model with the IMD defined criterion($u = 115.6$) and found that the return levels for periods less than 10 is approximately equal for both the models. The 2, 5 and 10 year return levels are 276 mm, 322 mm and 356 mm respectively which differ significantly from the return levels of the previous model which used entire 26663 values.

Parameter Estimates for GPD

ξ	Threshold	σ
0.0029	67	48.612
(0.0343)		(2.336)

Table 9: Jun-Sept for years 1951-2023 for Mumbai

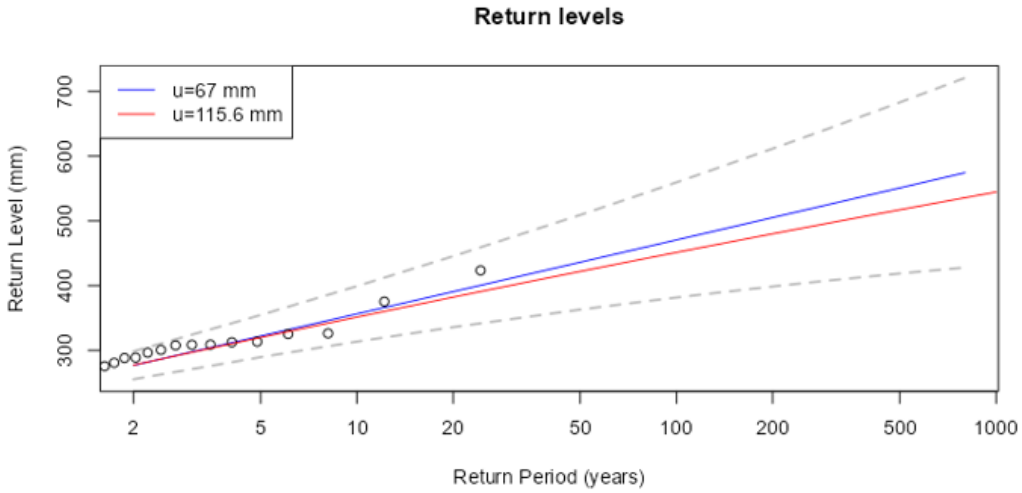


Figure 14: Return Level plot for Mumbai (1951-2023)

6 Conclusions and Discussion

6.1 Conclusions

The present study offers a comprehensive analysis of very heavy rainfall events across 50 Indian cities. The findings reveal trends in the occurrence of very heavy rainfall events, indicating climatic changes in some of the cities. For the sake of brevity, the computations of return levels are restricted to two cities only, but the same procedure can be employed for any other city.

6.1.1 Findings from the Exploratory & Regression Analysis

- Cities such as Pune, Nashik, Surat, Lucknow, Vadodara, Rajkot, Tiruchirappalli, and Udaipur experienced a significant increase in the frequency of heavy rainfall events during the decade from 2004 to 2013.
- Similarly, the decade from 1984 to 1993 also showed an unusual number of exceedances in cities like Kolkata, Chennai, Bhopal, Ludhiana, Aurangabad, and Jabalpur.
- A keen observation of climatological changes in cities such as Srinagar, Bhubaneswar, and Tiruchirappalli is necessary due to the increase in very heavy rainfall events.
- A significant decrease in very heavy rainfall events was observed in Delhi, Faridabad, Meerut, and Aligarh, indicating a notable change in the climate over the years.
- A significant increase in very heavy rainfall events was observed in Nagpur and Kottayam. These cities are more vulnerable to such hazardous events, suggesting that precautionary actions need to be taken.
- The remaining cities under study exhibit relative stability in the occurrence of very heavy rainfall events.

6.1.2 Prediction of Return Levels

Extreme event modeling has been performed for all 50 cities under study. The block maxima approach for analyzing rainfall in Mumbai and Pune has been discussed earlier to understand the process.

- Through the block maxima approach for Mumbai, the 2-year and 5-year return levels were estimated to be 209 mm and 264 mm, respectively. This means that the expected number of times daily rainfall exceeds 209 mm in a 2-year period is one. Similarly for Pune, return Levels for 2, 5 and 10 year periods are 69 mm, 105 mm and 138 mm respectively.
- The Peak-Over-Threshold (POT) approach for Mumbai was performed using two thresholds: one defined by the IMD and another estimated using MRL and parameter stability plots.
- Using the Kolmogorov-Smirnov (KS) test, it was concluded that the theoretical distribution does not fit the empirical distribution when using the IMD-defined threshold.

- June to September is the monsoon season in India. Considering daily rainfall only during these months from 1951 to 2023, MLR and parameter stability plots suggested a threshold of 67 mm, corresponding to the 90^{th} percentile of the data.
- The IMD-defined threshold for a very heavy rainfall event is 115.6 mm, which is approximately the 96^{th} percentile of the data. Using this threshold, a Generalized Pareto Distribution (GPD) model was fitted, and the 2-year, 5-year, and 10-year return levels were predicted to be 276 mm, 322 mm, and 356 mm, respectively.

6.2 Discussion

This study did not consider the influence of other environmental factors such as temperature, humidity, city elevation or coastal proximity on the frequency of very heavy rainfall events but it can certainly be adopted into models in future studies. Inclusion of these variables may provide more accurate estimations of return levels and thus provide a better assessment of environmental risks. Urban centers like Mumbai, which have a long history of infrastructure challenges related to heavy rainfall, could benefit from long-term planning adjustments. Implementing sustainable drainage solutions, such as green infrastructure or rainwater harvesting systems, could be crucial to reducing flood risks in the future. Cities that have shown increasing rainfall trends over the decades, such as Nagpur and Kottayam, should adopt similar measures. The results of this study highlight the importance of dynamic government interventions. Policies related to climate resilience, urban development, and emergency preparedness need to be adjusted based on the specific vulnerabilities of cities that are prone to very heavy rainfall events. The differences observed between cities (e.g., decreasing events in Delhi vs. increasing events in Kottayam) suggest that a one-size-fits-all policy would not be effective. Educating local populations about the risks posed by heavy rainfall events and the importance of preparedness can reduce casualties and property damage during such events. Public outreach programs that explain return periods and the probability of extreme weather could foster greater community involvement in implementing risk reduction measures.

7 Acknowledgment

We would like to thank the Head of our Department, Dr. Vinayak Gedam and express our sincere gratitude to Dr. Akanksha Kashikar for her invaluable guidance and support throughout this project. Our heartfelt thanks extend to Dr. Rohini Bhawar and researcher Ashwin Jadhav from the Department of Atmospheric and Space Sciences, Savitribai Phule Pune University for providing critical domain knowledge and necessary resources that significantly contributed to our work. We also deeply appreciate the Department of Statistics, Savitribai Phule Pune University, for offering excellent support and resources that facilitated this project. Finally, we thank the Indian Society of Probability and Statistics for giving us an opportunity to work on such exciting topic.

References

[1] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer London, 2001.

[2] Indian Meteorological Department. *IMD Data Portal*. `https://www.imdpune.gov.in/index.php`.

[3] Tayybeh Mohammadi et al. "Estimation of non-stationary return levels of extreme temperature by CMIP6 models". In: *Water Practice and Technology* 19 (2024), pp. 594–610.

[4] United Nations. *UN World Urbanization Prospects 2018*. `https://population.un.org/wup/Download/`. 2018.

[5] D. Pai et al. "Development of a new high spatial resolution ($0.25^\circ \times 0.25^\circ$) long period (1901-2010) daily gridded rainfall data set over India". In: *Mausam* 65 (2014), pp. 1–18. DOI: `10.54302/mausam.v65i1.851`.

[6] Eric Gilleland and Richard W. Katz. "extRemes 2.0: An Extreme Value Analysis Package in R". In: *Journal of Statistical Software* 72.8 (2016), pp. 1–39. DOI: `10.18637/jss.v072.i08`.

[7] George Bocharov. *pyextremes (Version 2.3.3)*. `https://github.com/georgebv/pyextremes`. 2023.

[8] K. M. Sakthivel and V. Nandhini. "Modeling extreme values of non-stationary precipitation data with effects of covariates". In: *Indian Journal of Science and Technology* 17.22 (2024), pp. 2283–2295.

[9] Yanan Hu and Carl Scarrott. "evmix: An R package for Extreme Value Mixture Modeling, Threshold Estimation and Boundary Corrected Kernel Density Estimation". In: *Journal of Statistical Software* 84.5 (2018), pp. 1–27.