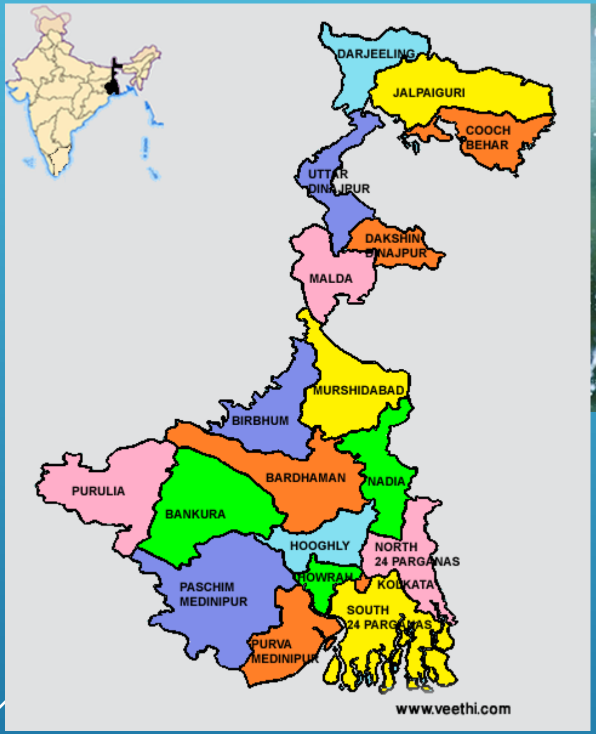
DOES THE VARIANCE OF DAILY RAINFALL DURING THE MONSOON IN WEST BENGAL CHANGE: AN EVIDENCE FROM A STATISTICAL HYPOTHESIS TESTING

Presented by:

Achyut Ghosh (ROLL 96/STS No. 190001)
Anirban Ghosh (ROLL 96/STS No. 190004)
Dipanta Mistry (ROLL 96/STS No. 190008)
Department of Statistics, University of Kalyani

Supervised by:

Dr. Raju Maiti
Economic Research Unit,
Indian Statistical Institute, Kolkata



DEPARTMENT OF STATISTICS
UNIVERSITY OF KALYANI

ABSTRACT

Objective: The aim of this study is to analyze the daily rainfall in six districts of West Bengal to study if there is any change in consistency of Monsoon rainfall over time. This will help to find any change in distribution of rainfall over years. Another purpose of this study is to forecast Monsoon rainfall of upcoming years based on past rainfall data.

Methods: We employ Regression analysis technique to detect trend in the Monsoon variance. Apart from Regression analysis, we also consider a non-parametric approach by performing Mann-Kendall Test to detect trend in Monsoon variance. Then we fit ARIMA model to forecast rainfall in Monsoon for next 10 years.

Results: The variance of Monsoon rainfall shows a significantly increasing trend for the districts Manbhum Purulia and South 24 Parganas. It shows a significantly decreasing trend for Darjeeling. For Nadia, Monsoon variance is stable over 120 years and no proper conclusion can be made for Coochbehar and Malda.

Conclusion: There is an indication that distribution of Monsoon rainfall has been changed in some certain regions of West Bengal. More detailed study can reveal more aspects of change in distribution of Monsoon rainfall in West Bengal.

Keywords: - Monsoon, Regression Analysis, Trend Detection, Mann-Kendall Test Statistic, Theil-Sen Slope Estimator, ARIMA Model

All the materials of this project can be found in the following GitHub repository: https://github.com/ganirban004/msc-project.git

DATA DESCRIPTION

Source: Multiple rain-gauge stations are situated across various districts of West Bengal. Station-wise daily rainfall data for 120 years (1901-2020) are collected from West Bengal Pollution Control Board.

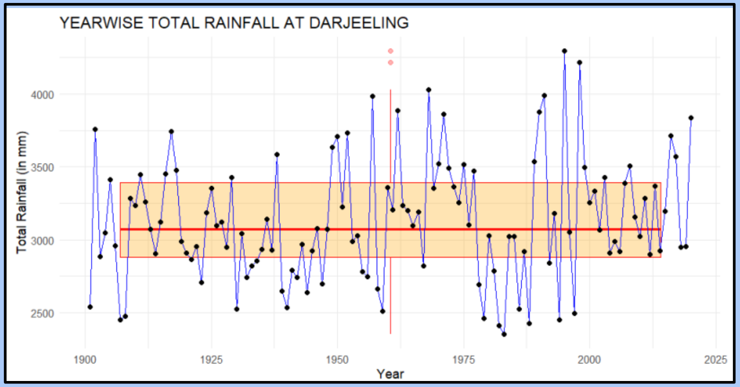
Preprocessing: The dataset contains two types of missing values to be considered here.

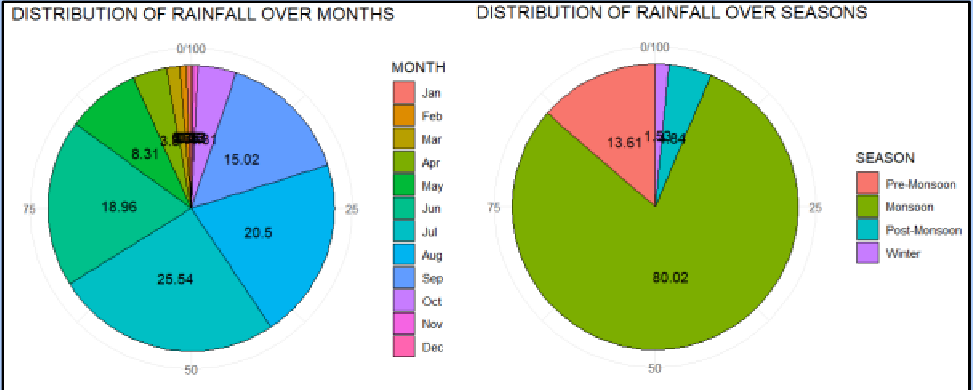
- For certain districts, some monthly data are missing for specific years. These missing values were imputed using the average of the corresponding months from the preceding and following ten years.
- For some districts, entire years of data were missing. These gaps were filled by averaging the data from the previous and next ten years for each respective date.

Study Area:

Agro-Climatic Zone	Selected District
Northern Hill Zone	Darjeeling
Terai-Teesta Alluvial Zone	Coochbehar
Vindhyan Alluvial Zone	Maldah
Gangetic Alluvial Zone	Nadia
Undulating Red Zone	Purulia
Coastal Saline Zone	South 24 Parganas

EXPLORATORY DATA ANALYSIS: DARJEELING





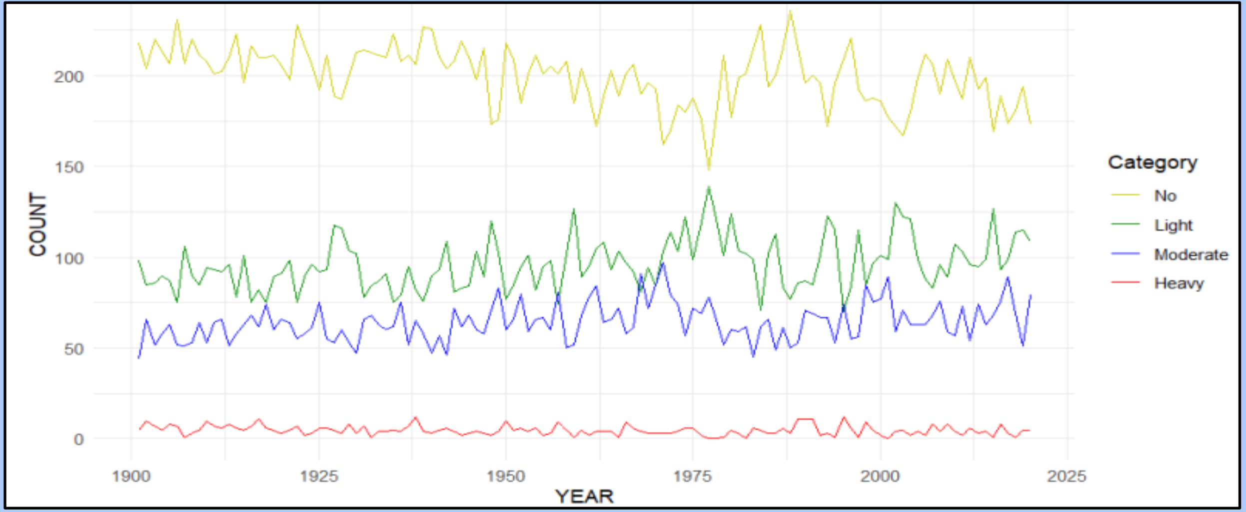
(a) Monthwise

(b) Seasonwise

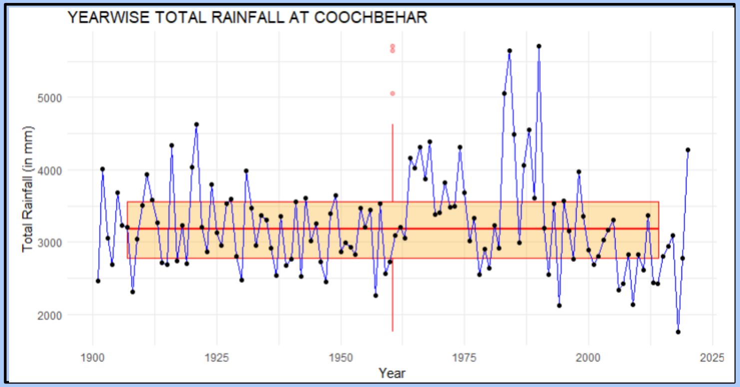
STATISTICS	VALUE (MM)
Mean	3128.725
SD	413.5391
Maximum (Year)	4294.108 (1995)
Minimum (Year)	2351.912 (1983)

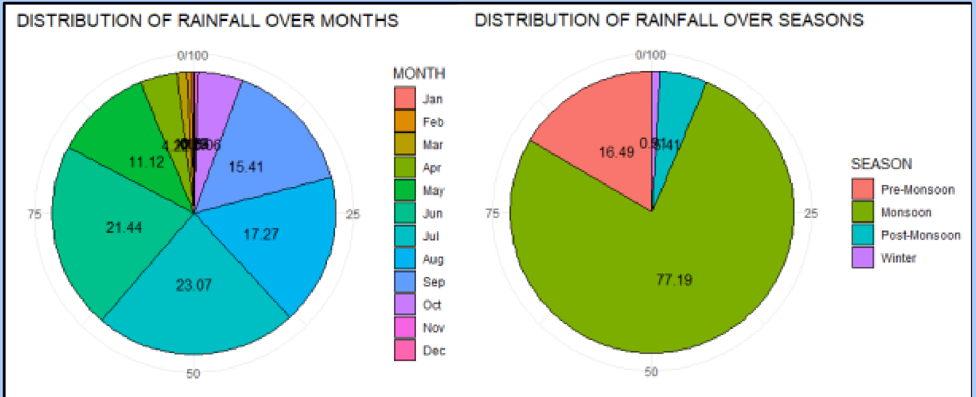
SEASON	STATISTICS	NO	LIGHT	MODERATE	HEAVY
Pre-Monsoon	Mean	51.9833	33.1333	6.825	0.0583
	SD	9.0829	8.7130	3.2266	0.2971
	Maximum	71	58	18	2
	Minimum	28	14	0	0
Monsoon	Mean	12.1750	50.2167	55.2333	4.375
	SD	6.4209	8.2615	9.4443	2.5757
	Maximum	30	73	80	11
	Minimum	0	28	38	0
Post-Monsoon	Mean	50.4333	8.3667	1.925	0.275
	SD	5.0707	4.5569	1.8581	0.7067
	Maximum	60	26	8	4
	Minimum	35	1	0	0
Winter	Mean	85.25	4.6	0.4	0
	SD	3.4767	3.2542	0.7	0
	Maximum	90	25	3	0
	Minimum	64	0	0	0

From seventh decade the rainfall started to decrease and keep decreasing till end of ninth decade of previous century. In the next decade, it deviates a lot and finally shows a regular pattern in current century.



EXPLORATORY DATA ANALYSIS: COOCHBEHAR



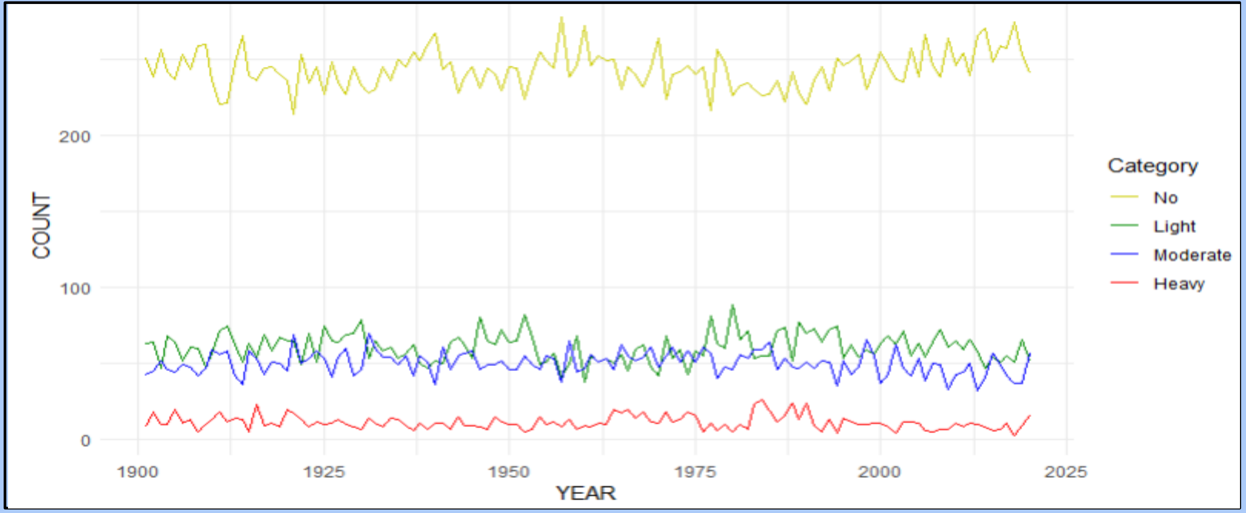


(a) Monthwise (b) Seasonwise

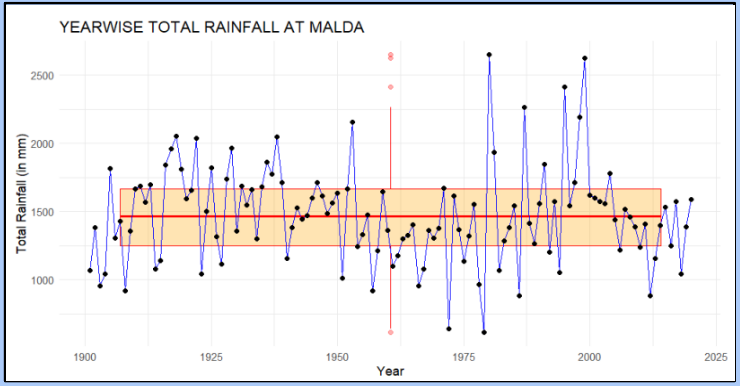
STATISTICS	VALUE (MM)
Mean	3243.816
SD	673.3498
Maximum (Year)	5709.813 (1990)
Minimum (Year)	1761.043 (2018)

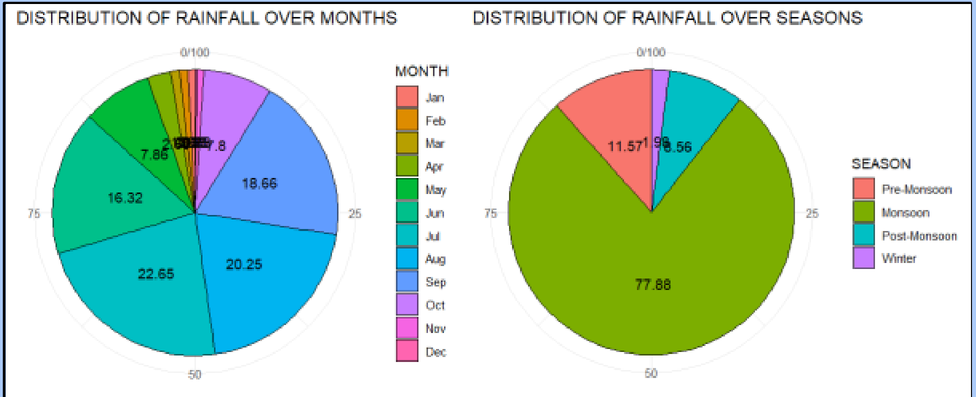
SEASON	STATISTICS	NO	LIGHT	MODERATE	HEAVY
Pre-Monsoon	Mean	63.1000	16.6333	11.5500	0.7167
	SD	6.7890	4.9597	4.1107	1.0423
	Maximum	82	34	25	5
	Minimum	41	5	3	0
Monsoon	Mean	39.3333	37.0167	35.5500	10.1000
	SD	7.6598	6.0635	5.7443	4.3749
	Maximum	59	55	51	24
	Minimum	24	25	22	2
Post-Monsoon	Mean	53.1583	4.6333	2.6750	0.5333
	SD	3.8794	2.6769	2.1531	0.8158
	Maximum	60	14	11	4
	Minimum	38	0	0	0
Winter	Mean	87.5583	2.2083	0.4750	0.0083
	SD	2.0404	1.6528	0.7299	0.0909
	Maximum	91	8	3	1
	Minimum	82	0	0	0

From middle of seventh decade to middle of eight decade more rainfall occurred. Then from middle of ninth decade, it shows a decreasing trend till present, except for the year 2020.



EXPLORATORY DATA ANALYSIS: MALDA



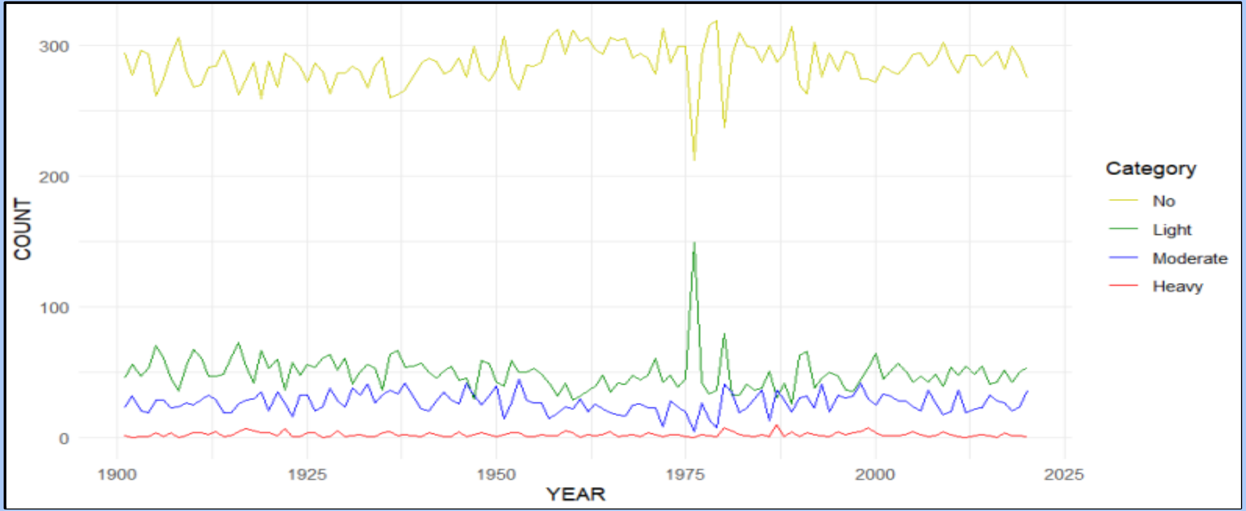


(a) Monthwise (b) Seasonwise

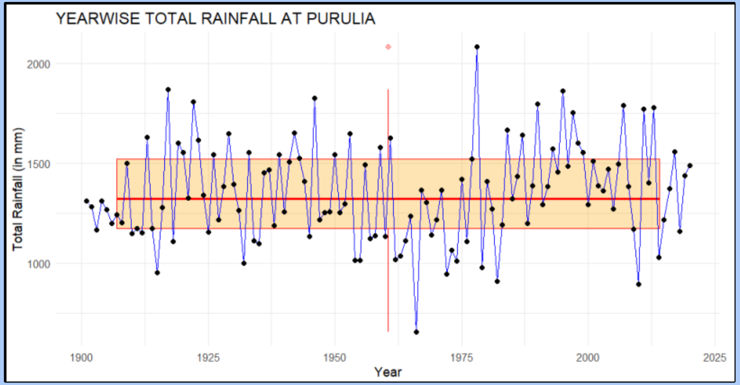
STATISTICS	VALUE (MM)
Mean	1475.512
SD	357.8816
Maximum (Year)	2646.826 (1980)
Minimum (Year)	615.800 (1979)

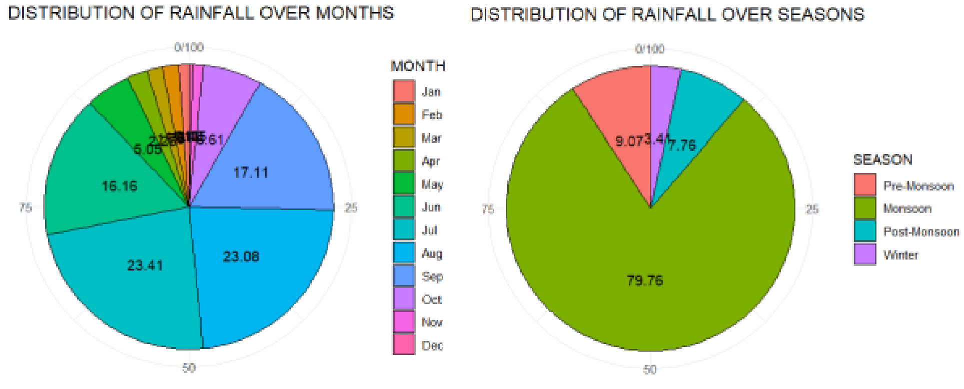
SEASON	STATISTICS	NO	LIGHT	MODERATE	HEAVY
Pre-Monsoon	Mean	81.1	7.2983	3.5333	0.1583
	SD	4.6177	3.7481	2.32	0.4471
	Maximum	90	24	11	3
	Minimum	68	1	0	0
Monsoon	Mean	62.025	36.5	21.2333	2.2417
	SD	11.5546	10.0631	6.4765	1.8028
	Maximum	89	113	38	10
	Minimum	6	19	3	0
Post-Monsoon	Mean	54.9417	3.8917	1.8	0.3667
	SD	3.9397	3.4322	1.5578	0.7063
	Maximum	61	31	7	4
	Minimum	30	0	0	0
Winter	Mean	87.775	1.9667	0.5	0.0083
	SD	1.9934	1.6224	0.7416	0.0909
	Maximum	91	7	3	1
	Minimum	82	0	0	0

In the last two decade of previous century, rainfall there was some inconsistency in rainfall. Otherwise, the rainfall is more or less stable.



EXPLORATORY DATA ANALYSIS: MANBHUM PURULIA



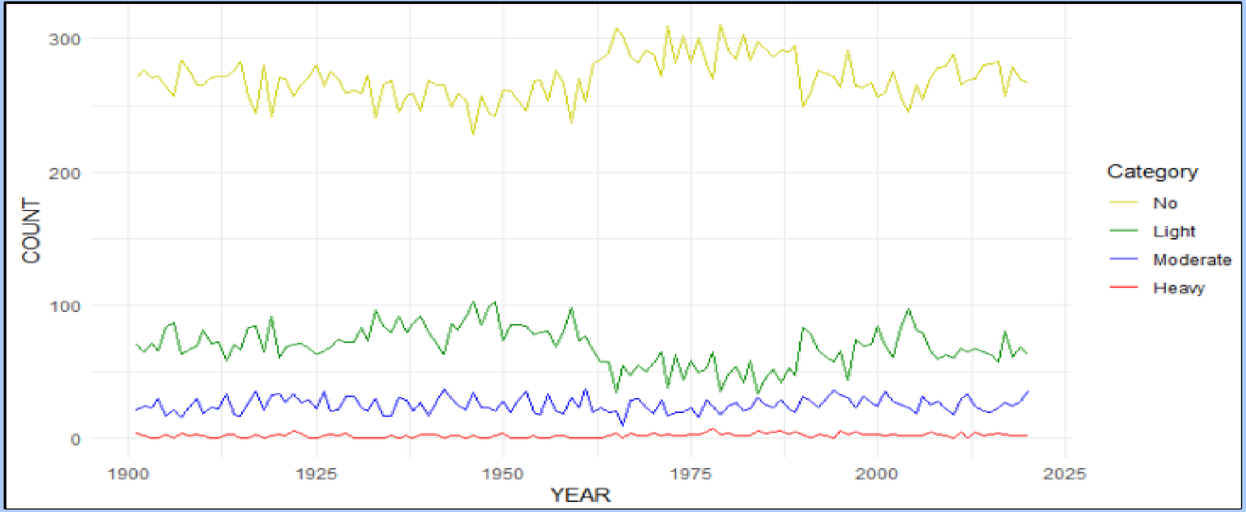


(a) Monthwise (b) Seasonwise

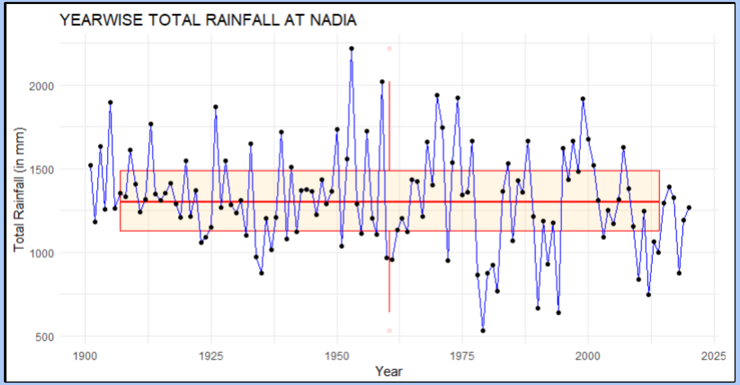
STATISTICS	VALUE (MM)
Mean	1350.746
SD	245.49
Maximum (Year)	2081.5 (1978)
Minimum (Year)	655.17 (1966)

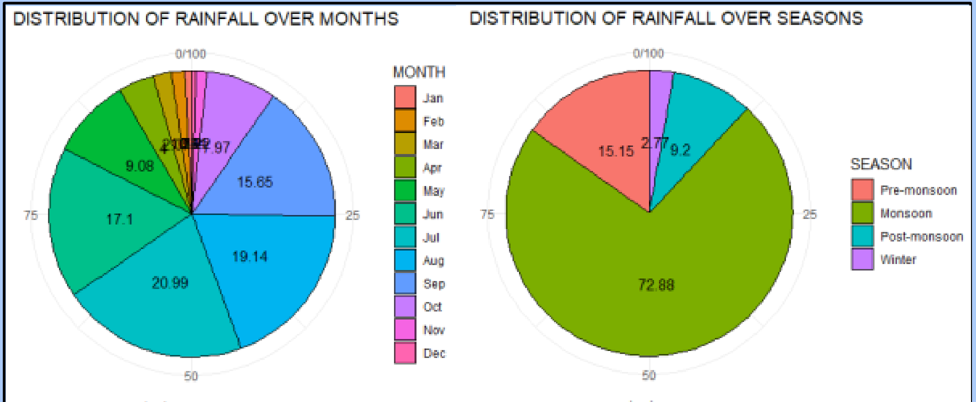
SEASON	STATISTICS	NO	LIGHT	MODERATE	HEAVY
Pre-Monsoon	Mean	79.6583	10.7000	1.6000	0.0417
	SD	6.0463	6.0424	1.6093	0.1998
	Maximum	91	31	6	1
	Minimum	61	1	0	0
Monsoon	Mean	51.1833	48.8750	20.6667	1.2750
	SD	11.6919	10.2839	5.3281	1.3842
	Maximum	81	69	37	7
	Minimum	26	24	8	0
Post-Monsoon	Mean	53.6500	5.4583	1.7583	0.1333
	SD	4.3944	3.5116	1.6882	0.3859
	Maximum	61	16	10	2
	Minimum	40	0	0	0
Winter	Mean	86.0417	3.6000	0.6083	0.0000
	SD	3.4166	2.8763	0.9156	0.0000
	Maximum	91	14	4	0
	Minimum	75	0	0	0

Rainfall started to increase from middle of seventh decade and keep the flow till middle of last decade of previous century. Then it shows a more or less regular pattern in the current century.



EXPLORATORY DATA ANALYSIS: NADIA



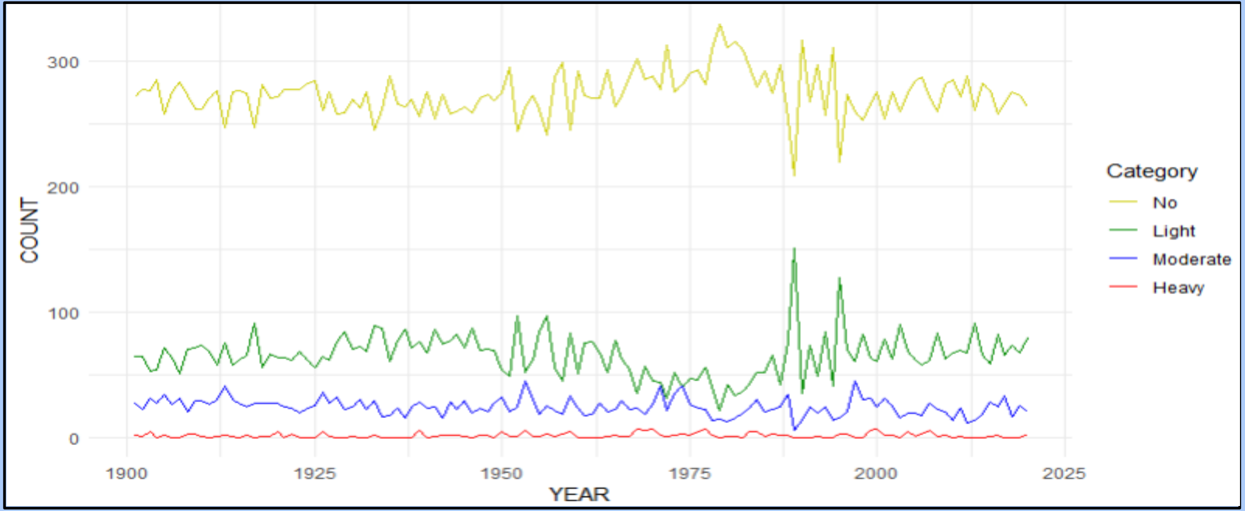


(a) Monthwise (b) Seasonwise

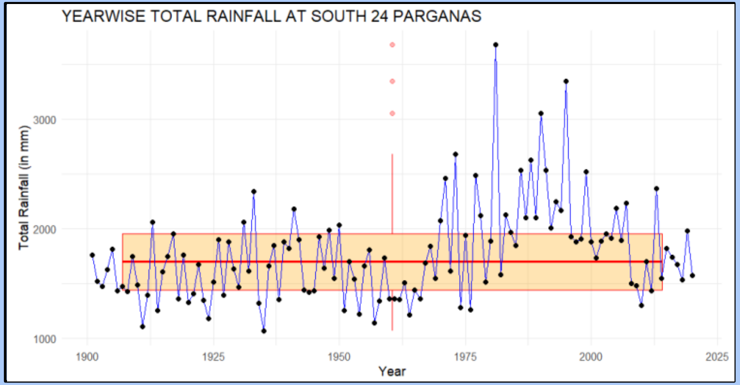
STATISTICS	VALUE (MM)
Mean	1313.007
SD	298.6616
Maximum (Year)	2219.967 (1953)
Minimum (Year)	532.200 (1979)

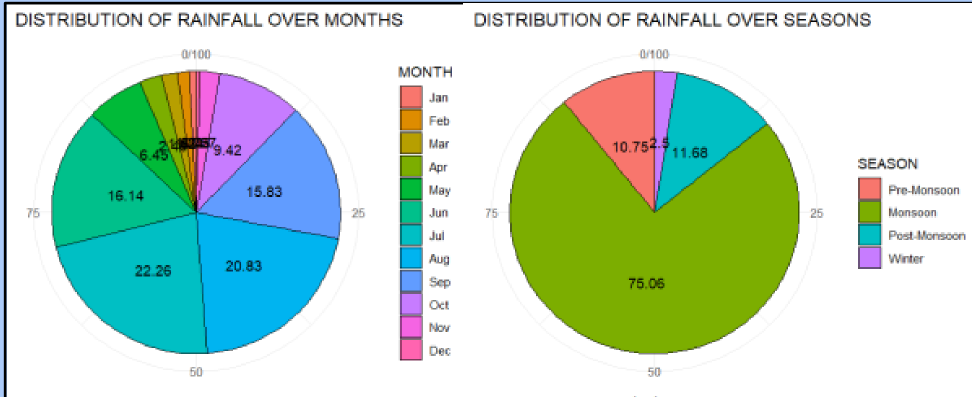
SEASON	STATISTICS	NO	LIGHT	MODERATE	HEAVY
Pre-Monsoon	Mean	76.4667	11.5583	3.8667	0.1083
	SD	6.5038	6.2474	2.6924	0.3365
	Maximum	92	34	10	2
	Minimum	57	0	0	0
Monsoon	Mean	57.7750	45.5333	17.4250	1.2667
	SD	12.2130	12.0892	5.6342	1.4761
	Maximum	96	113	40	5
	Minimum	4	16	5	0
Post-Monsoon	Mean	52.9667	5.7000	2.1583	0.1750
	SD	4.7468	4.1324	1.8303	0.4409
	Maximum	61	31	9	2
	Minimum	30	0	0	0
Winter	Mean	87.1000	2.6000	0.5417	0.0083
	SD	2.5146	2.0632	0.8841	0.0909
	Maximum	91	9	5	1
	Minimum	81	0	0	0

There is a decreasing trend of rainfall in last two decades. In last 60 years, the rainfall showed more variability than that of first 60 years.



EXPLORATORY DATA ANALYSIS: SOUTH 24 PARGANAS



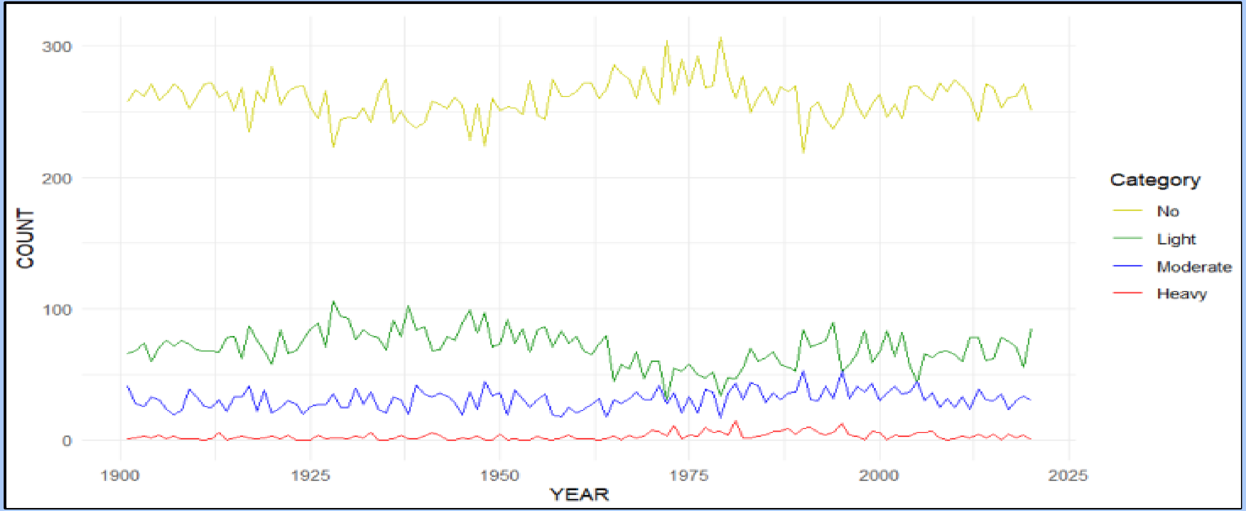


(a) Monthwise (b) Seasonwise

STATISTICS	VALUE (MM)
Mean	1768.882
SD	435.6788
Maximum (Year)	3677.583 (1981)
Minimum (Year)	1069.62 (1935)

SEASON	STATISTICS	NO	LIGHT	MODERATE	HEAVY
Pre-Monsoon	Mean	77.6167	10.875	3.35	0.1583
	SD	5.9008	4.609	2.5841	0.5627
	Maximum	90	24	15	4
	Minimum	55	1	0	0
Monsoon	Mean	46.2083	49.5667	23.8417	2.3833
	SD	9.0092	9.3137	5.6701	2.4839
	Maximum	73	72	37	14
	Minimum	22	24	11	0
Post-Monsoon	Mean	49.8333	7.2417	3.425	0.5
	SD	5.3764	3.9094	2.3933	0.7853
	Maximum	60	19	10	3
	Minimum	34	0	0	0
Winter	Mean	86.725	2.7583	0.7333	0.0333
	SD	2.5947	2.1016	0.9463	0.1795
	Maximum	91	8	4	1
	Minimum	79	0	0	0

There is slight deficiency in rainfall in sixth to seventh decade. After that rainfall started to increase from middle of eighth decade and keep increasing till middle of last decade of previous century, then decreased till present.



METHODOLODY: TESTING LINEAR REGRESSION COEFFICIENT

For a single year, we compute the marginal variance using all Y_t where t belongs to the Monsoon of that year and the formula to compute it is

$$MV_i = \frac{1}{T}\sum_{t\in S_i}(Y_t - \overline{Y})^2$$

where S_i denotes the time interval during Monsoon for year i and MV_i refers to Monsoon variance of the year i. Now plot MV_i against years i.

Now once we have the rainy season variance for all the years then we want to fit a regression line of MV_i on i, i = 1, 2, …, 120, given by

$$\boxed{MV_i = \alpha + \beta i + \varepsilon_i}$$

Then our testing would be $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$.

Let $SE(\hat{\beta})$ be the standard error for estimating β. Then $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$ follows a t-distribution with d.f. n − 2, i.e.,

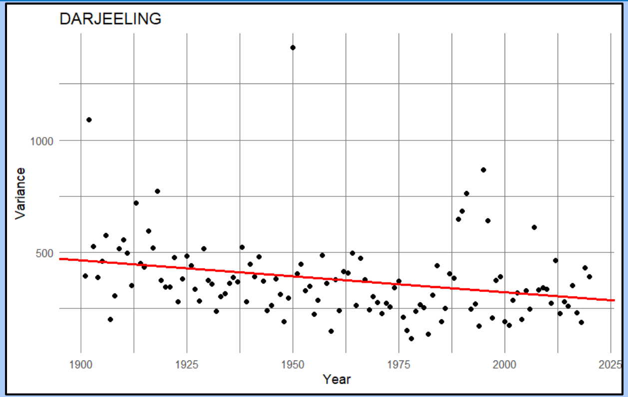
$$\boxed{t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim t_{n-2}}$$

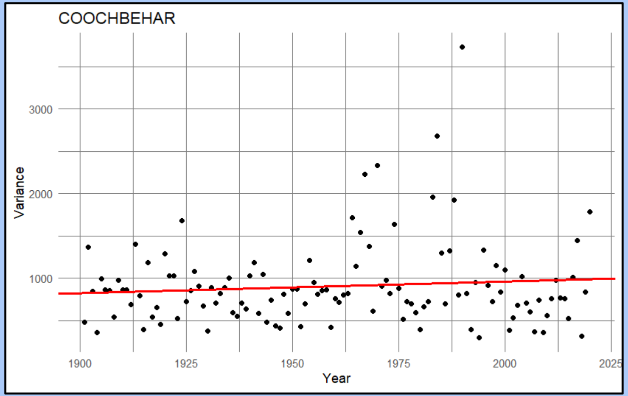
So

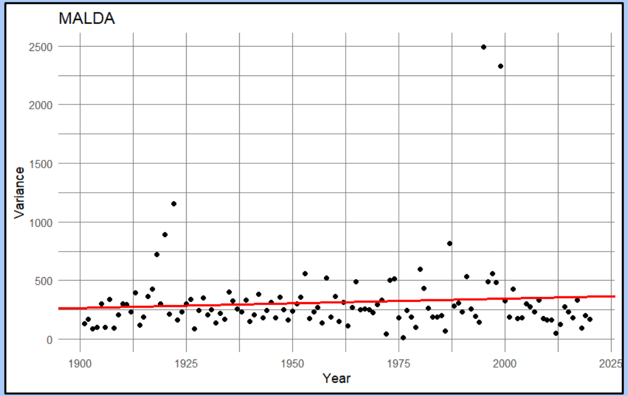
$$\boxed{\text{p-value} = P(|t_n{-}2| > t)}$$

for testing H_0 against H_1.

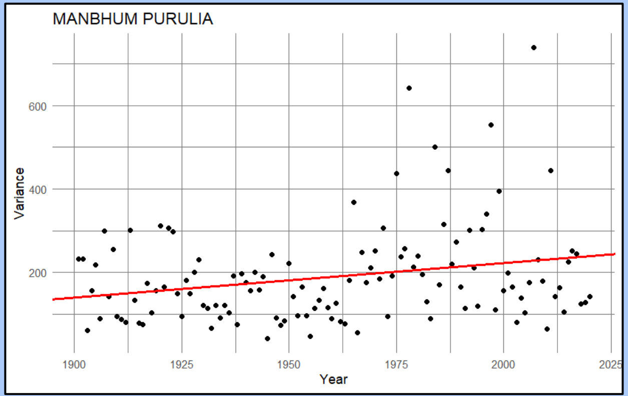
RESULTS: TESTING LINEAR REGRESSION COEFFICIENT

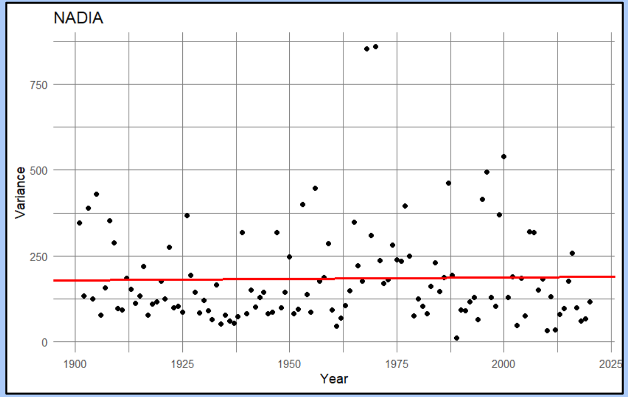


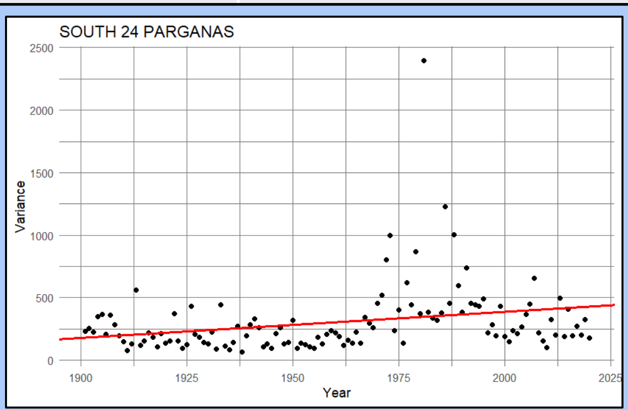




District	Trend of Variance
Darjeeling	Significantly Decreasing
Coochbehar	Non-significantly Increasing
Maldah	Non-significantly Increasing
Manbhum Purulia	Significantly Increasing
Nadia	Non-significantly Increasing
South 24 Parganas	Significantly Increasing







METHODOLOGY: MANN-KENDALL TEST AND THEIL-SEN ESTIMATOR

Mann-Kendall Test: The Mann-Kendall test uses relative magnitudes of the data to calculate trend. It is calculated from the sum of the sign of the slopes. The statistic S is $S = \sum_{k=1}^{n-1}\sum_{j=k+1}^{n} \mathrm{sgn}(x_j - x_k)$, where n = number of data point and x_i = the i^{th} observation and the $\mathrm{sgn}(x_j - x_k)$ is an indicator function which takes on values 1, 0 or -1 according to the sign of $(x_j - x_k)$:

$$\mathrm{sgn}(x_j - x_k) = \begin{cases} 1 & if \quad (x_j - x_k) > 0 \\ 0 & if \quad (x_j - x_k) = 0 \\ -1 & if \quad (x_j - x_k) < 0 \end{cases}$$

When n is large enough, under null hypothesis of no trend, S is normally distributed with

$$E(S) = 0 \text{ and } Var(S) = \frac{n(n-1)(2n+5) - \sum_{j=1}^{p} t_j (t_j - 1)(2t_j + 5)}{18}$$

where p = number of tied groups in the dataset and t_j is the number of data points in the j^{th} tied group. Then S and Var(S) are used to compute the test statistic Z, which is computed as:

$$Z = \begin{cases} \frac{S-1}{\{Var(S)\}^{1/2}} & if \quad S < 0 \\ 0 & if \quad S = 0 \\ \frac{S+1}{\{Var(S)\}^{1/2}} & if \quad S > 0 \end{cases}$$

METHODOLOGY: MANN-KENDALL TEST AND THEIL-SEN ESTIMATOR

Under the null hypothesis of no trend, Z has a Standard Normal distribution. The trend is said to be decreasing if Z is negative and increasing if Z is positive. H_0, the null hypothesis of no trend, is rejected if the absolute value of Z is greater than, upper $(1-\frac{\alpha}{2})^{th}$ percentile obtained from the Standard Normal cumulative distribution tables.

Theil-Sen Slope Estimator: Theil-Sen slope estimator (TSSE) is not greatly affected by gross data errors or outliers and even it can be computed when data are missing. The slope estimates of all data pairs is computed as:

$$Q_i = \frac{x_j - x_k}{j - k} \text{ for } i = 1, 2, \ldots, N; \ (j > k)$$

where Q_i = slope between data points x_j and x_k, x_j = data measurement at time j, x_k = data measurement at time k, N = number of data pairs. The median of these N values of Q_i is Sen's estimator of slope.

$$Q_{med} = \begin{cases} Q\left[\frac{N+1}{2}\right] & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(Q[N/2] + Q[N/2+1]\right) & \text{if } N \text{ is even} \end{cases}$$

The sign of Q_{med} reflects data trend reflection, while its value indicates the steepness of the trend.

RESULTS: MANN-KENDALL TEST AND THEIL-SEN ESTIMATOR

Results obtained from Mann-Kendall test are tabulated below. The MK test statistic S, p-value, Theil-Sen Slope estimate and resulting trend of Monsoon variance are shown.

District	MK Test Statistics	P-Value (MK Test)	Sen's Slope Estimator	Trend of Variance
Darjeeling	-1810	0.00002	−1.409	Significantly Decreasing
Coochbehar	56	0.4504	0.1109	Non-significantly Increasing
Malda	2	0.4991	0.0018	Non-significantly Increasing
Manbhum Purulia	1012	0.0109	0.5006	Significantly Increasing
Nadia	-88	0.4218	-0.0411	Non-significantly Decreasing
South 24 Parganas	1594	0.0002	1.2252	Significantly Increasing

METHODOLOGY: FORECASTING WITH ARIMA MODEL

AutoRegressive Integrated Moving Average (ARIMA) Model effectively combines three components. The **Autoregressive (AR)** component involves regressing the variable on its own lagged (past) values:

$$X_t = \alpha + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

where X_t is the time series at time t, α is a constant, ϕ_i are the autoregressive parameters, and ε_t is white noise.

The **Integration (I)** component is used to transform a non-stationary time series into a stationary one by differencing the observations. A series is differenced by subtracting the previous observation from the current observation. If d differences are required to achieve stationarity, the process is repeated d times:

$$Y_t = X_t - X_{t-1}$$

where Y_t is the differenced series.

The **Moving Average (MA)** model suggests that the current value of the series is influenced by the errors from the previous q periods:

$$X_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

where μ is the mean of the series, θ_j are the moving average parameters, and ε_t is white noise.

Combining these components, the comprehensive ARIMA(p, d, q) model equation is:

$$\boxed{(1 - \sum_{i=1}^{p} \phi_i L^i)\,(1 - L)^d X_t = (1 + \sum_{j=1}^{q} \theta_j L^j)\,\varepsilon_t}$$

where p is the number of lag observations (autoregressive terms), d is the number of times the observations are differenced to achieve stationarity and q is the size of the moving average window (moving average terms) and L is the lag operator.

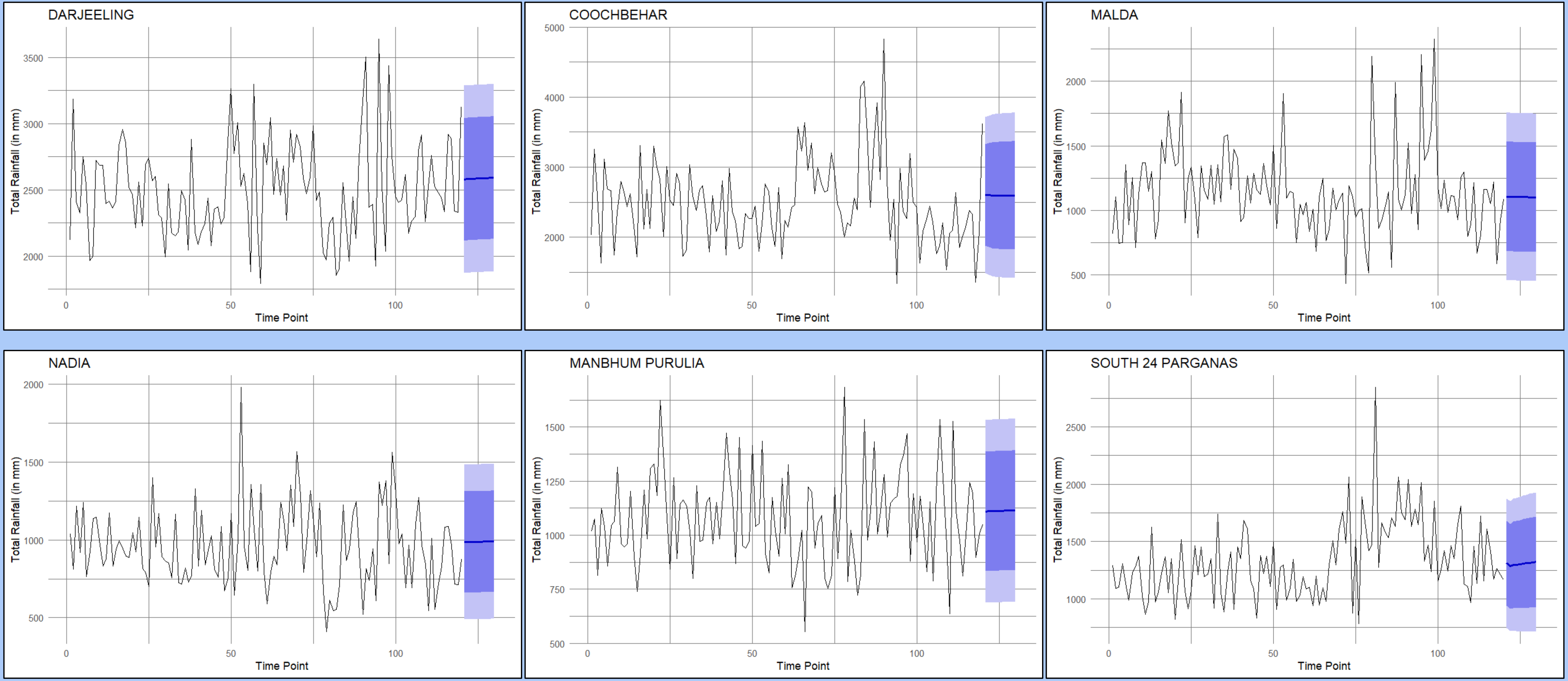
RESULTS: FORECASTING WITH ARIMA MODEL

Monsoon rainfalls (in mm) for next 10 years (2021-2030) are forecasted for each district using the ARIMA Model.

District	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Darjeeling	2580.69	2581.97	2583.25	2584.52	2585.80	2587.08	2588.36	2589.63	2590.91	2592.19
Coochbehar	2601.90	2599.51	2597.80	2596.63	2595.91	2595.56	2595.50	2595.69	2596.08	2596.63
Malda	1106.88	1106.18	1105.48	1104.78	1104.09	1103.39	1102.69	1101.99	1101.29	1100.59
Manbhum Purulia	1110.32	1110.87	1111.42	1111.97	1112.52	1113.07	1113.62	1114.17	1114.72	1115.27
Nadia	985.92	986.41	986.90	987.39	987.87	988.36	988.85	989.34	989.82	990.31
South 24 Parganas	1310.55	1285.10	1296.32	1299.25	1303.70	1307.54	1311.26	1314.77	1318.10	1321.25

RESULTS: FORECASTING WITH ARIMA MODEL

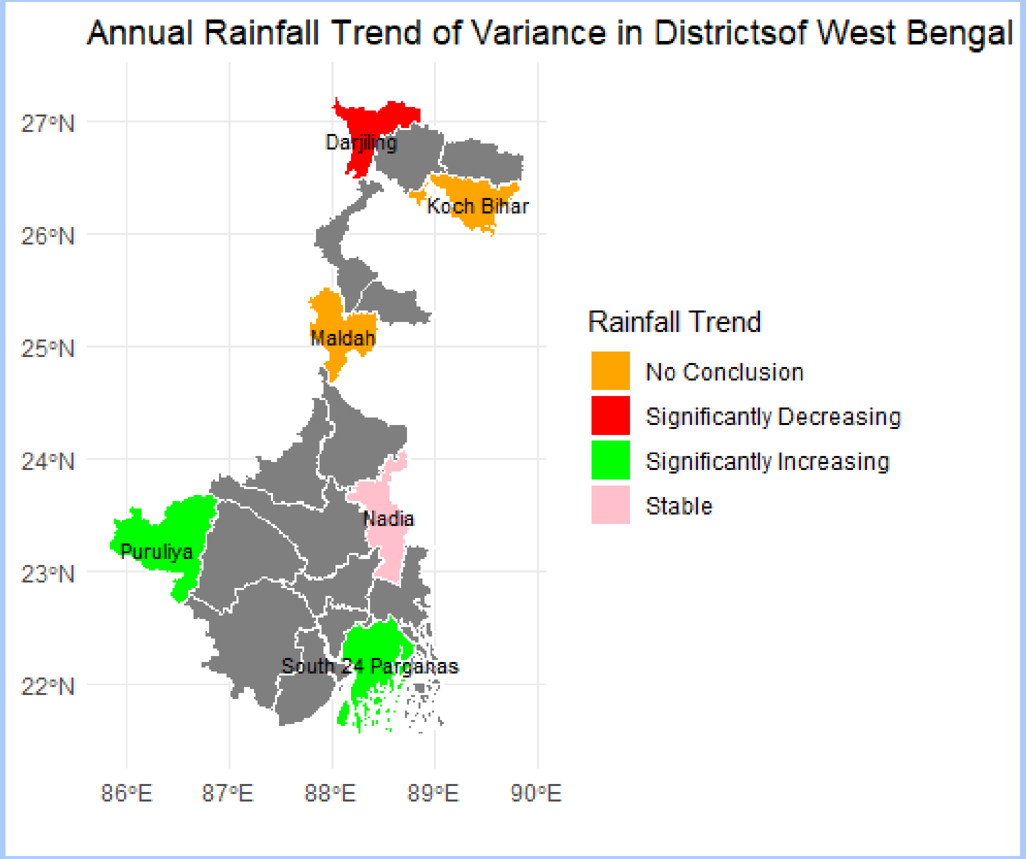
The forecasts are plotted here. Also 80% and 95% confidence intervals are also shown.



DISCUSSION & CONCLUSION

- Coochbehar and Darjeeling receive highest and second highest mean Monsoon rainfall of 2503.6081 mm and 2503.92 mm, respectively, which are substantially higher than the overall state mean of 1385.4474 mm.
- It can also be seen that Darjeeling has lower SD (360.5978 mm), which means Darjeeling receives high rainfall consistently than Coochbehar, which has an SD of 594.4453 mm.
- Nadia receives the lowest mean Monsoon rainfall of 956.942 mm, followed by Purulia at 1077.317 mm.

District	Beta Coefficient (Regression)	P-Value (Beta Coefficient)	MK Test Statistics	P-Value (MK Test)	Sen's Slope Estimator
Darjeeling	−1.413	0.0014	-1810	0.00002	−1.409
Coochbehar	1.369	0.149	56	0.4504	0.1109
Malda	0.7936	0.175	2	0.4991	0.0018
Manbhum Purulia	0.8289	0.0034	1012	0.0109	0.5006
Nadia	0.0872	0.409	-88	0.4218	-0.0411
South 24 Parganas	2.0805	0.0021	1594	0.0002	1.2252



DISCUSSION & CONCLUSION

- MK Test gives a negative value of TSSE (−1.409) with a p-value of 0.00002 for Darjeeling. Regression test also gives a negative value of slope coefficient (−1.413) with a p-value of 0.0014, which also supports the decline obtained from MK Test. This agreement highlights a consistent and statistically significant decreasing trend in Monsoon variance over 120 years in Darjeeling.
- For Coochbehar, MK Test gives a positive value of TSSE (0.1109) with a p-value of 0.4504. But Regression test gives a positive value of slope coefficient (1.369) with a small p-value of 0.149, which is more or less contradictory with MK Test. So though the Monsoon variance has a increasing trend in Coochbehar, it's significance cannot be concluded properly.
- For Malda, MK Test gives a positive value of TSSE (0.0018) with a high p-value of 0.4991. But Regression test gives a positive value of slope coefficient (0.7936) with a small p-value of 0.175, which is more or less contradictory with MK Test. So though the Monsoon variance has a increasing trend in Malda, it's significance cannot be concluded properly.
- MK Test gives a positive value of TSSE (0.5006) with a small p-value of 0.0109 for Purulia. Regression test also gives a positive value of slope coefficient (0.8289) with a small p-value of 0.0034, which also supports the increment obtained from MK Test. This agreement highlights a consistent and statistically significant increasing trend in Monsoon variance over 120 years in Purulia.

DISCUSSION & CONCLUSION

- For Nadia, MK Test gives a negative value of TSSE (−0.0411) with a high p-value of 0.4218. So it can be said from MK Test that Nadia shows a non-significant decreasing trend in Monsoon variance over 120 years. But Regression test gives a positive value of slope coefficient (0.0872) with a high p-value of 0.409, which is opposite of with MK Test. So for Nadia, Monsoon variance is stable over 120 years.
- MK Test gives a positive value of TSSE (1.2252) with a small p-value of 0.0002 for South 24 Parganas. So it can be said from MK Test that South 24 Parganas shows a significantly increasing trend in Monsoon variance over 120 years. Regression test also gives a positive value of slope coefficient (2.0805) with a small p-value of 0.0021 which also supports the increment obtained from MK Test. This agreement highlights a consistent and statistically significant increasing trend in Monsoon variance over 120 years in South 24 Parganas.

FUTURE WORKS

The research on this project is ongoing, with several key areas identified for further exploration.

- To identify potential change points in the rainfall patterns of West Bengal. Detecting these change points will provide insights into shifts in Monsoon and their impacts on the region's rainfall.
- To propose an alternative measure of dispersion for rainfall data, beyond the traditional use of variance. Developing new dispersion metrics could offer a more nuanced understanding of rainfall variability and better capture the complexities inherent in rainfall patterns.
- Rainfall in any region is influenced by a multitude of climatic factors, including temperature, humidity, wind direction, and others. Future studies will focus on incorporating these variables into the analysis to provide a more comprehensive understanding of rainfall characteristics. By integrating these climatic factors, we can develop more robust models that account for the interplay between different atmospheric conditions and rainfall.

These future directions will enhance the overall understanding of rainfall patterns in West Bengal and contribute to more accurate predictions and better-informed climate adaptation strategies for the region.

THANK YOU !!!