# HYDROINFORMATICS-HYDROCLIMATIC DATA REPRESENTATION

# INTRODUCTION

**Hydroclimatic variables:**

- Rainfall, Streamflow, Soil Moisture, Air temperature, Wind Speed, Relative Humidity

**Spatiotemporal Variability:**

- Hydroclimatic variables vary with space and time due to different hydrological/climatological phenomenon/processes. Ex: Rainfall variability over space and time
- Such processes are evolving continuously over time, studying the interdependence in hydroclimatic data with proper consideration of temporal information may lead to better insight into the governing processes.

# HYDROINFORMATION: CHANGES OVER TIME

Properties change over time ?

Hydroinformation changes over time: Increasing/Decreasing

- Rainfall
- Temperature
- Streamflows
- Water quality parameters
- Water Demands: Drinking, irrigation, low flow augmentation etc.

# MORE ABOUT DATA ANALYSIS

Statistics to be estimated based on World Climate Application Program (WCAP) World Meteorological Organization (WMO) (1998):

1. Mean
2. Standard error of mean
3. Standard deviation
4. Coefficient of variation
5. Coefficient of skew
6. Coefficient of kurtosis
7. Ranks for each month
8. Coefficient of autocorrelation
9. Standard error of coefficient of autocorrelation
10. Cumulative periodogram
11. Variance Spectrum
12. Confidence intervals for variance spectrum
13. Rescaled range
14. Hurst's coefficient
15. Number of runs
16. Trend in the mean
17. Trend in variance
18. Equality of subperiod means
19. Equality of subperiod variances
20. Jump in the mean
21. Gaussian filter

These statistics can be carried out for subsamples (or non-overlapping subperiods) of 5, 10, 20, and 30 years of length derived from the original time-series

# CENTRAL TENDENCY - ARITHMETIC MEAN

Sum of the observations divided by sample size.

Sample data with n observations $x_1, x_2, \ldots. x_n$ of a hydroclimatic variable X.

The central tendency of the data in terms of mean can be estimated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In case of grouped data, let us consider *k* as the number of groups, *n* as the total number of observations, $n_i$ as the number of observations in the *ith* group, and $x_i$ as the class mark of the *ith* group. Class mark is defined as midpoint of the group, i.e., mean of upper and lower bounds of group. For grouped data, the $\bar{x}$ is given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i$$

# CENTRAL TENDENCY

**Geometric Mean:**

Indicates the central tendency of any hydroclimatic data set by using the product of their values. It is defined as the nth root of the product of *n* observations.

The sample geometric mean = $\bar{x}_G (\prod_{i=1}^{n} x_i)^{1/n}$ $G.M = \sqrt[n]{x_1 \times x_2 \times \ldots x_n}$

**Weighted Mean:**

- Similar to arithmetic mean except some data points contribute more than others.
- Similar to arithmetic mean of grouped data with weighted factors as $\frac{n_i}{n}$.
- The weighted mean can be expressed as follows: $\bar{x}_w = \frac{\sum_{i=1}^{k} w_i \, x_i}{\sum_{i=1}^{k} w_i}$
- Where $w_i$ is the weights associated with the ith observation or group.

# CENTRAL TENDENCY

For population, considering $p_x(x_i)$ as the underlying distribution (Probability Mass function, PMF) of a discrete random variable, X, the population mean $\mu$ is expressed as:

$$\mu = \sum_{i=1}^{n} x_i \, p_x(x_i)$$

For a continuous random variable X, with underlying probability density function (pdf) as $f_x(x)$, the population mean can be estimated as:

$$\mu = \int_{-\infty}^{\infty} x \, f_x(x) \, dx$$

# MEASURE OF DISPERSION

Dispersion of a hydroclimatic variable represents how closely the values of a random variable are clustered or how widely it is spread around the central value. Dispersion – evaluation of variation (spread) in the data

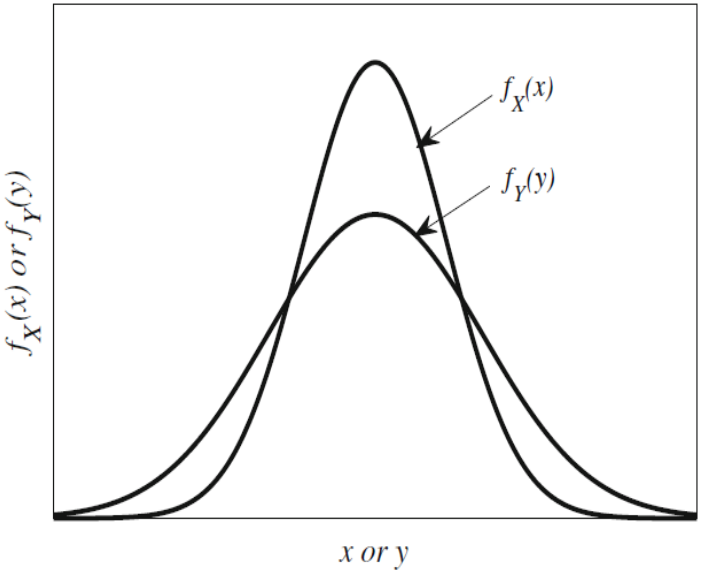
Ex: Two random variables, X and Y, with same mean but dispersion of Y is more than X.

**Range:**

- Minimum and maximum values convey the information about the variability present in the data.
- Difference between the maximum and minimum values in the sample
- Range has the disadvantage of not reflecting the frequency or magnitude of values that deviate either positively or negatively from the mean since only largest and smallest values are used in its determination.
- Relative range is used as the range divided by the mean

Range (R = Max $(x_i)$ – Min $(x_i)$ )

Mean absolute deviation (MAD = $\frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$) )



# MEASURE OF DISPERSION - VARIANCE

- Variance is a measure of the dispersion of a random variable taking mean as the central value.
- It is the average squared deviation from the sample mean.
- Considering X as a random variable and a sample $x_1, x_2, \ldots, x_n$ with sample mean $\bar{x},$ the differences $x_1 - \bar{x}, x_2 - \bar{x},$ …. $x_n - \bar{x}$ are called the deviations from the mean.
- The variance can be defined as the average of the squared deviations from the mean.
- The sample estimate of population variance $\sigma^2$ is denoted by $S^2$ and is given as:
  - $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Standard deviation, another measure of dispersion, is the positive square root of variance, and the unit of standard deviation is the same as the unit of the X.
- A dimensionless measure of dispersion is the coefficient of variation defined as the standard deviation divided by the mean. $C_v = \frac{S}{\bar{x}}$. A higher value indicates more dispersed data, i.e. high variability about the mean and vice versa.

# NUMERICAL SUMMARY STATISTICS

1. Central tendency – Mean, mode, median (measure of location)

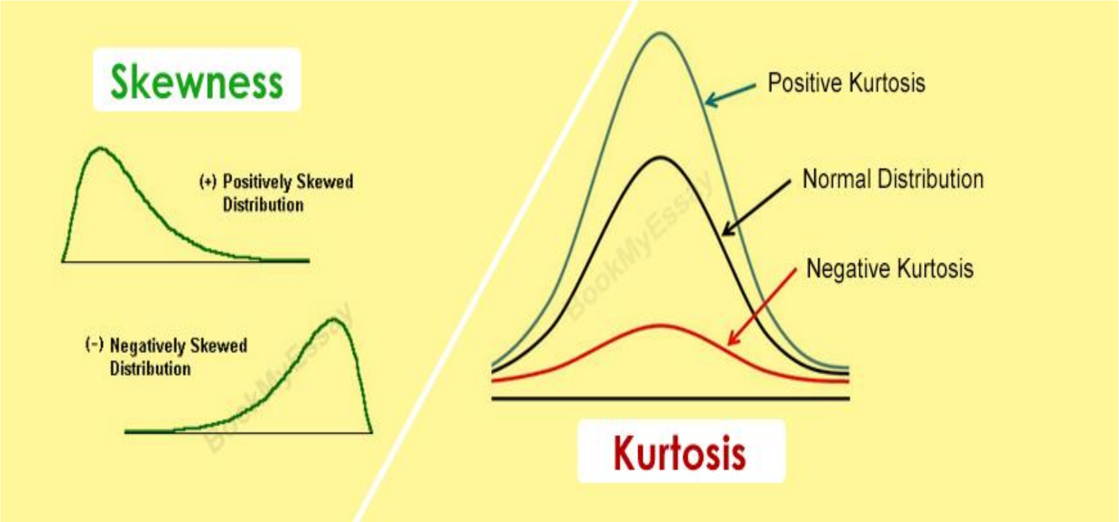
2. 3. Shape – Skewness and kurtosis parameter of the dataset.

$$g = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)S^3}$$

$$\text{k} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)S^4}$$

S = Standard Deviation

- There are three kurtosis categories—mesokurtic (normal), platykurtic (less than normal), and leptokurtic (more than normal).



Distributions with a large kurtosis have more tail data than normally distributed data, which appears to bring the tails in toward the mean. Distributions with low kurtosis have fewer tail data, which appears to push the tails of the bell curve away from the mean.

# STATISTICAL INFORMATION - HYDROLOGICAL DATA

Statistical Properties

- Mean
- Standard deviation/variance
- Correlation coefficients - Joint dependency between variables
- Probability distributions: Probability density function, cumulative distribution function
- Trends

# TIME SERIES

- A time series is a set of chronological observations recorded at successive times or over successive periods of times.
- A set of data depending on the time

Ex: Hourly temperature announced by IMD

Ex: Daily rainfall recorded

Ex: Monthly dissolved oxygen level or pH, nitrates, etc. measured at a location

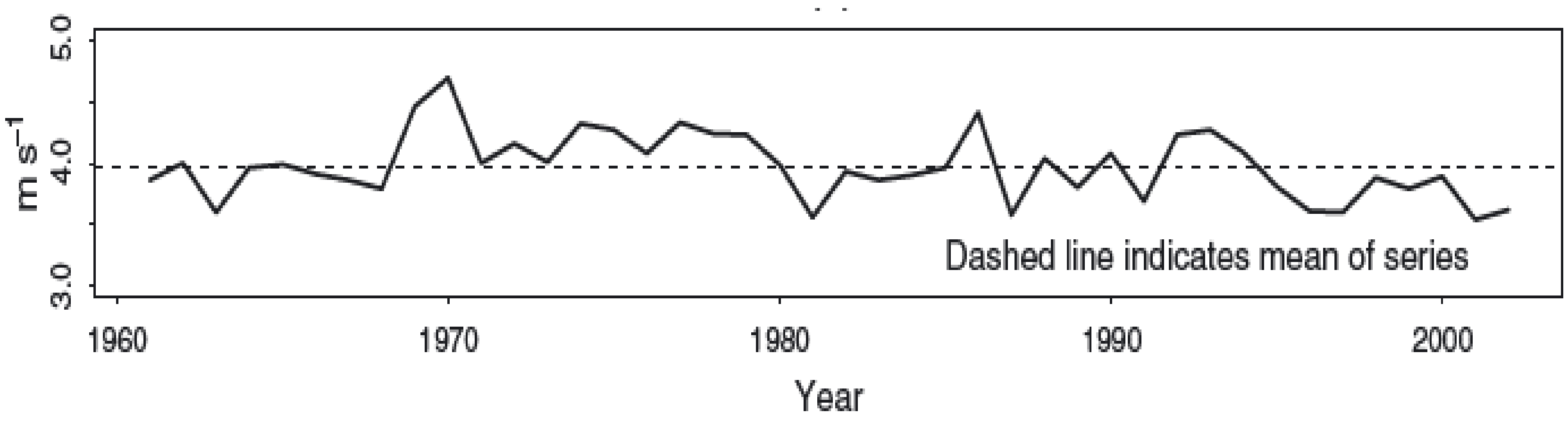
Ex: Daily river flow measured at a location

Ex: Annual drinking water consumption of Hyderabad city

# TIME SERIES PLOTS

Observation Vs Time

Ex: Time series plot of wind speed data: Annual time series data from 1961 to 2002



# HYDROINFORMATICS - TREND ANALYSIS

- To describe the past behavior of a process: nature and extent of changes in a region’s water and climate related information
- To understand the mechanism behind observed changes: dependability of variables, such as human activity rather than natural processes.
- To assess possible future scenarios: extrapolating past changes into future
- To monitor the effectiveness of environmental control and adaptive policies

# WHAT IS A TREND ?

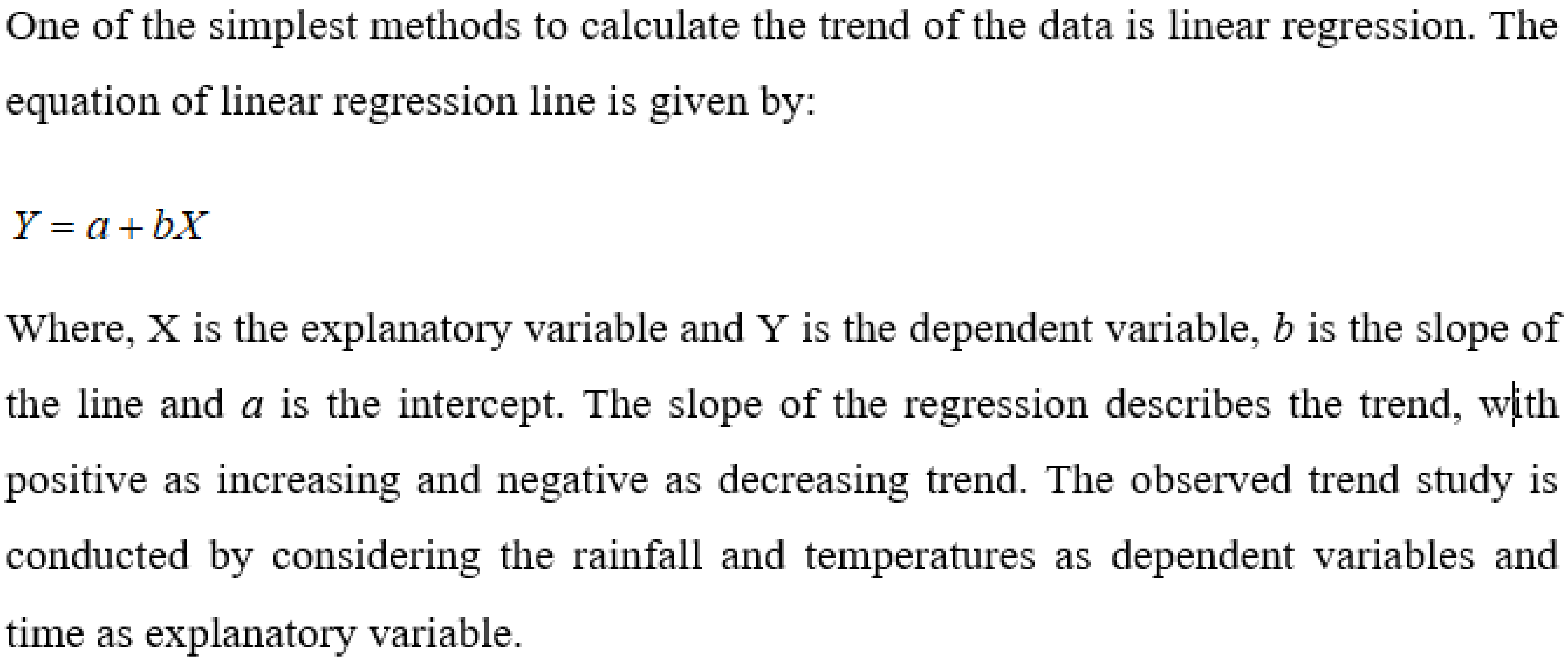
- Investigation of changes in a system over a period of time

- Change in the mean level of a series or long-term change in the mean level or long-term movement -- Change in the mean level

- “Trend is long-term temporal variation in the statistical properties of a process, where long-term depends on the application”

- Changes in frequency of extreme events such as floods

# LINEAR TREND ANALYSIS

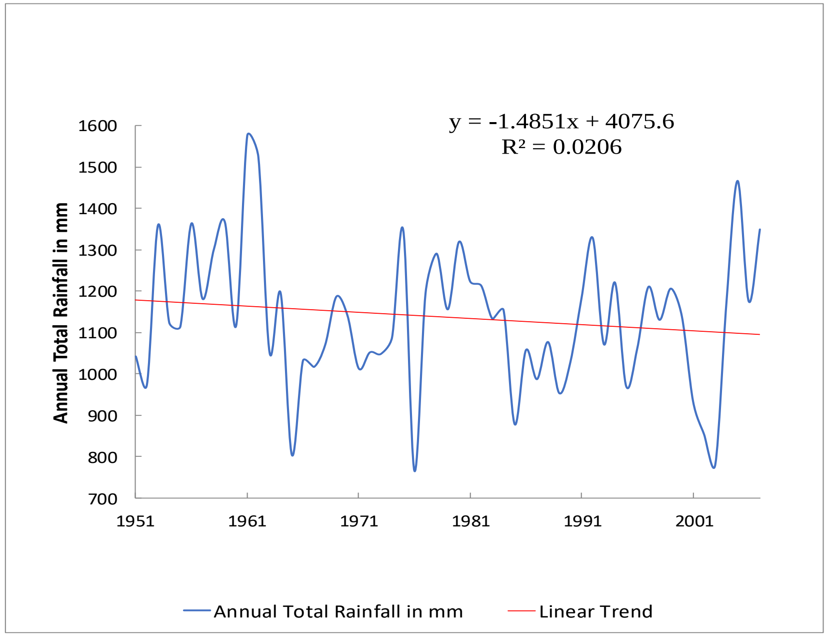


One of the simplest methods to calculate the trend of the data is linear regression. The equation of linear regression line is given by:

$$Y = a + bX$$

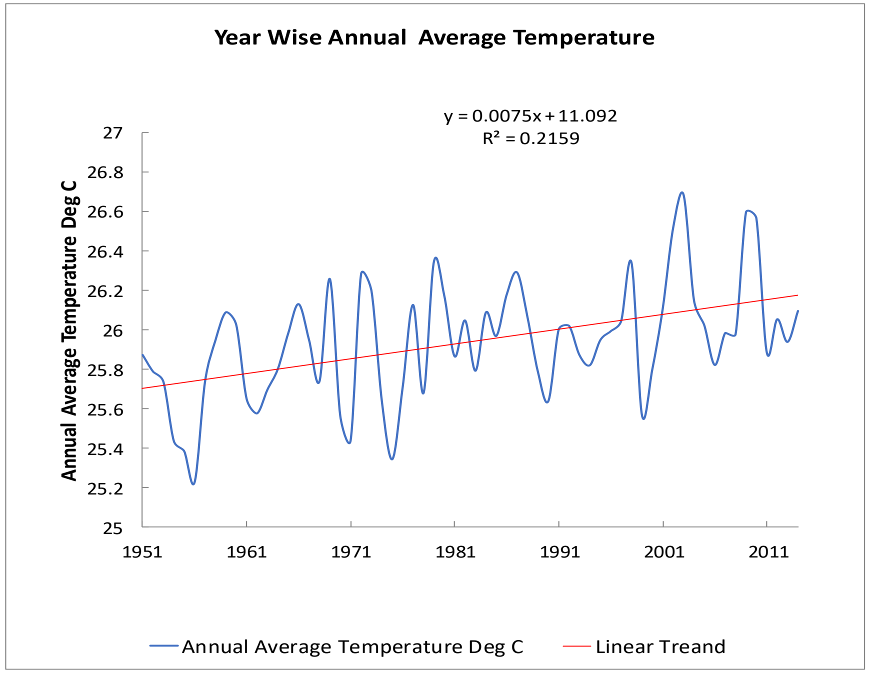
Where, X is the explanatory variable and Y is the dependent variable, $b$ is the slope of the line and $a$ is the intercept. The slope of the regression describes the trend, with positive as increasing and negative as decreasing trend. The observed trend study is conducted by considering the rainfall and temperatures as dependent variables and time as explanatory variable.

# EXAMPLE



The aerially averaged annual total rainfall trend over Tunga-Bhadra River basin for the observed period of 1951 to 2007

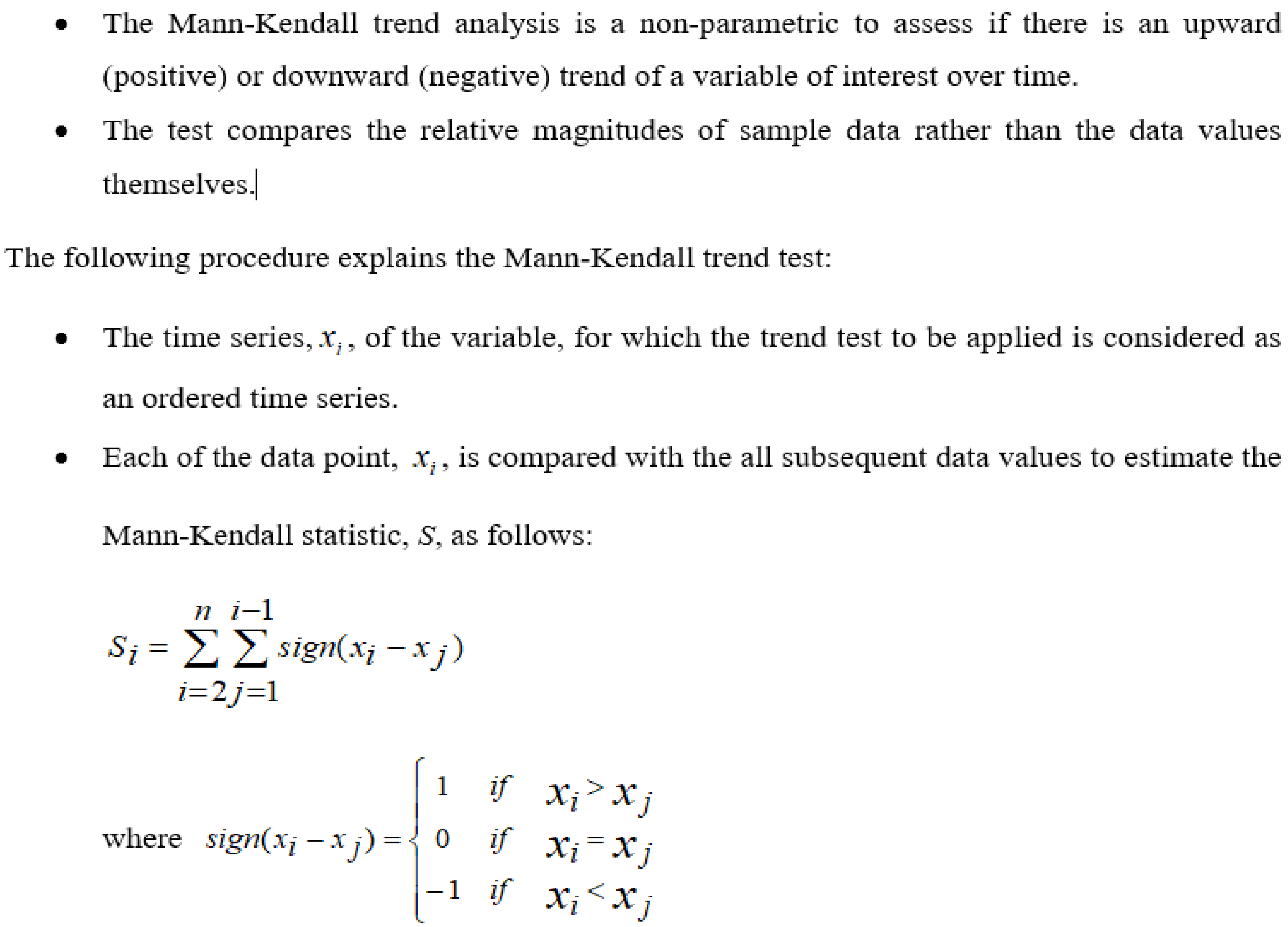
14.85 mm/decade of decrease in annual total rainfall over the basin



Spatial annual average temperature trend over Tunga-Bhadra River basin for the observed period of 1951 to 2014

0.1 °C/decade of increase in annual average temperature over the basin

# MANN-KENDALL TREND TEST



- The Mann-Kendall trend analysis is a non-parametric to assess if there is an upward (positive) or downward (negative) trend of a variable of interest over time.
- The test compares the relative magnitudes of sample data rather than the data values themselves.

The following procedure explains the Mann-Kendall trend test:

- The time series, $x_i$, of the variable, for which the trend test to be applied is considered as an ordered time series.
- Each of the data point, $x_i$, is compared with the all subsequent data values to estimate the Mann-Kendall statistic, $S$, as follows:

$$S_i = \sum_{i=2}^{n} \sum_{j=1}^{i-1} sign(x_i - x_j)$$

where $sign(x_i - x_j) = \begin{cases} 1 & if \quad x_i > x_j \\ 0 & if \quad x_i = x_j \\ -1 & if \quad x_i < x_j \end{cases}$

- A very high positive value of $S$ is an indicator of an increasing trend, and a very low negative value of S indicates a decreasing trend.
- From the Mann-Kendall statistic, $S$, the normalized test statistics, Z, is computed as follows:

$$Z = \frac{S-1}{[VAR(S)]^{1/2}} \qquad \text{if} \quad S > 0$$

$$Z = 0 \qquad \text{if} \quad S = 0$$

$$Z = \frac{S+1}{[VAR(S)]^{1/2}} \qquad \text{if} \quad S < 0$$

where $VAR(S)$ is the variance of $S$. According to (Kendall, 1975) $VAR(S)$ can be written as follows:

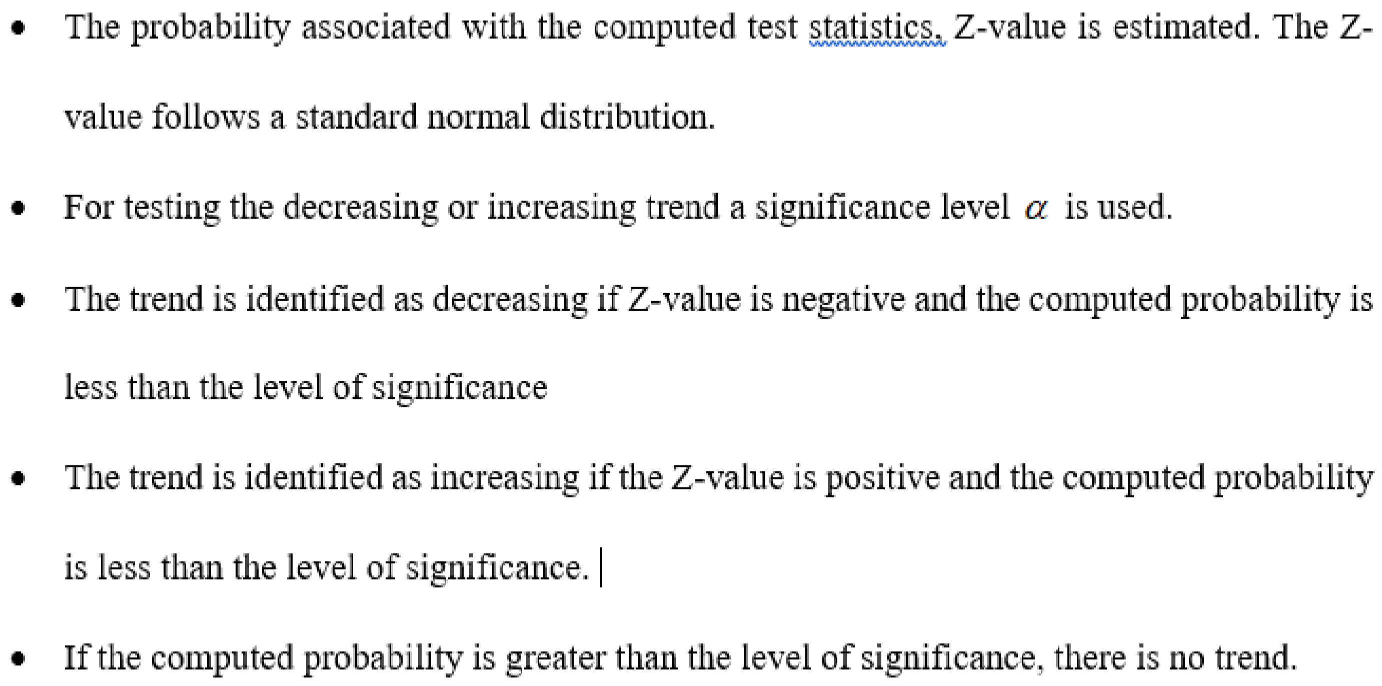
$$VAR\ (S) = \frac{1}{18}\left[ n(n-1)(2n+5) - \sum_{p=1}^{g} t_p(t_p - 1)(2t_p + 5) \right]$$

where $n$ is the number of data points, $g$ is the number of tied groups (a tied group is a set of sample data having the same value), and $t_p$ is the number of data points in the $P^{\text{th}}$ group.

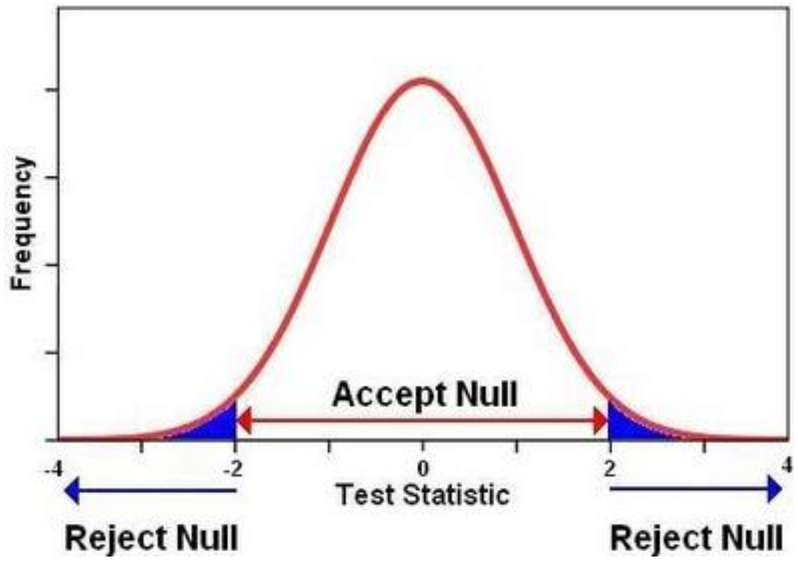
- A very high positive value of $Z$ is an indicator of an increasing trend, and a very low negative value of Z indicates a decreasing trend.

Null Hypothesis, H0: No monotonic trend

Alternative Hypothesis: Upward or Downward monotonic trend



- The probability associated with the computed test statistics, Z-value is estimated. The Z-value follows a standard normal distribution.
- For testing the decreasing or increasing trend a significance level $\alpha$ is used.
- The trend is identified as decreasing if Z-value is negative and the computed probability is less than the level of significance
- The trend is identified as increasing if the Z-value is positive and the computed probability is less than the level of significance.
- If the computed probability is greater than the level of significance, there is no trend.



Rejection region: is the set of values of the test statistics that cause us to reject the null hypothesis

Significance level: probability that the test statistics will fall in the critical region when the null hypothesis is accurate

# QUANTIFICATION OF MAGNITUDE OF TREND ?

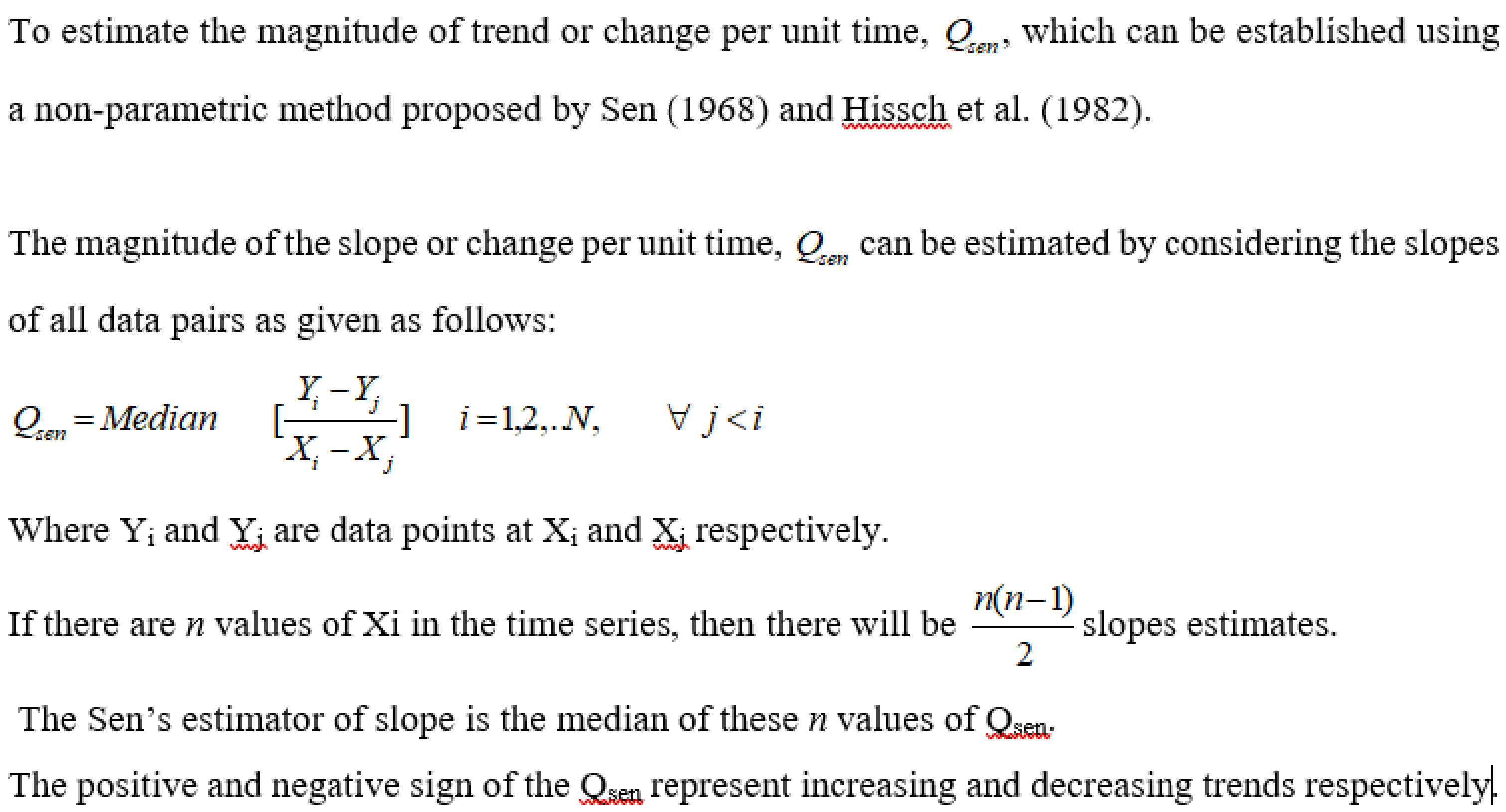
Mann-Kendall trend test can indicate if the trend is present or not

Quantification of trend is not possible

Sen slope estimator -

- median of the pairwise slopes
- Index of change over a unit time period

# MAGNITUDE OF TREND: SEN’S SLOPE TEST



To estimate the magnitude of trend or change per unit time, $Q_{sen}$, which can be established using a non-parametric method proposed by Sen (1968) and Hissch et al. (1982).

The magnitude of the slope or change per unit time, $Q_{sen}$ can be estimated by considering the slopes of all data pairs as given as follows:

$$Q_{sen} = Median \quad [\frac{Y_i - Y_j}{X_i - X_j}] \quad i = 1,2,.N, \qquad \forall \; j < i$$

Where $Y_i$ and $Y_j$ are data points at $X_i$ and $X_j$ respectively.

If there are $n$ values of Xi in the time series, then there will be $\frac{n(n-1)}{2}$ slopes estimates.

The Sen’s estimator of slope is the median of these $n$ values of $Q_{sen}$.

The positive and negative sign of the $Q_{sen}$ represent increasing and decreasing trends respectively.

$$\tau = \frac{\text{(number of concordant pairs)} - \text{(number of discordant pairs)}}{\frac{1}{2}n(n-1)}$$

# CHANGE POINT DETECTION TEST -PETTITT-TEST

- The change-point detection method identifies a change point at T, in a time series of data.
- With a null hypothesis as there is no change in the distribution of a sequence of random variables of $X_1$, $X_2$,…$X_T$.
- The alternative hypothesis is considered as the distribution function, $F_1(x)$, of the variables from $x_1$ to $x_t$ is different from the distribution function $F_2(x)$ of the random variables from $X_{t+1}$ to $X_T$. The test statistics $K_T = \max_{1 \le t < T} |S_i|$

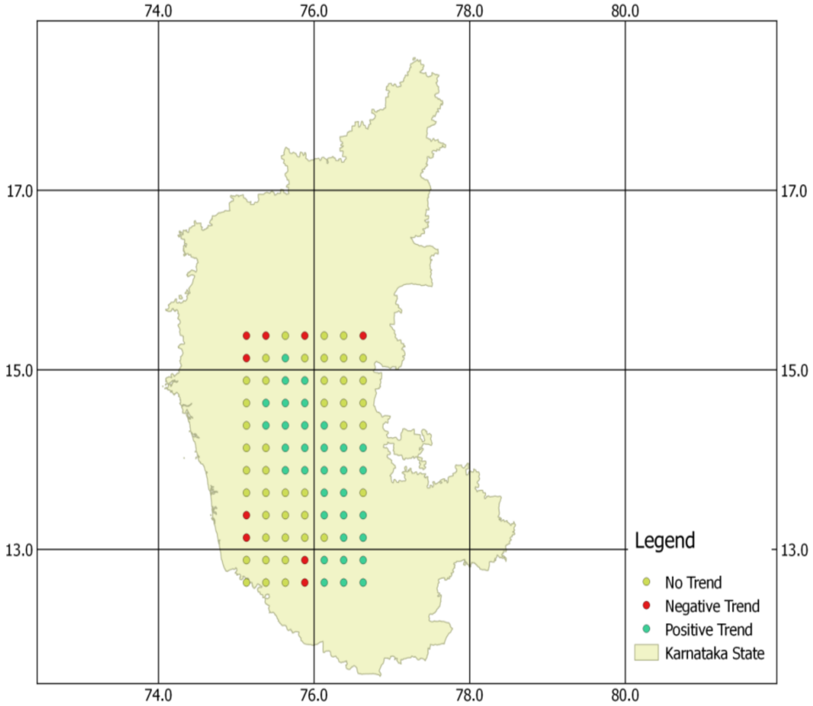
$$\sum_{i=1}^{t} \sum_{j=t+1}^{n} sign\,(x_t - x_j)$$

- A change point occurs at time t when the test statistic $K_T$ is different from zero for a given significance level, which is given by:
  - $p = 2.\exp\left(\dfrac{-6K_T^2}{T^2 + T^3}\right)$
- If the p-value estimated is less than the selected significance level, $\alpha$ , then reject the null hypothesis and divide the data into two interval to locate the change point in the time series with two different distribution functions.
- The two-time intervals before and after the change –point refers to the heterogeneous characteristics from each other.

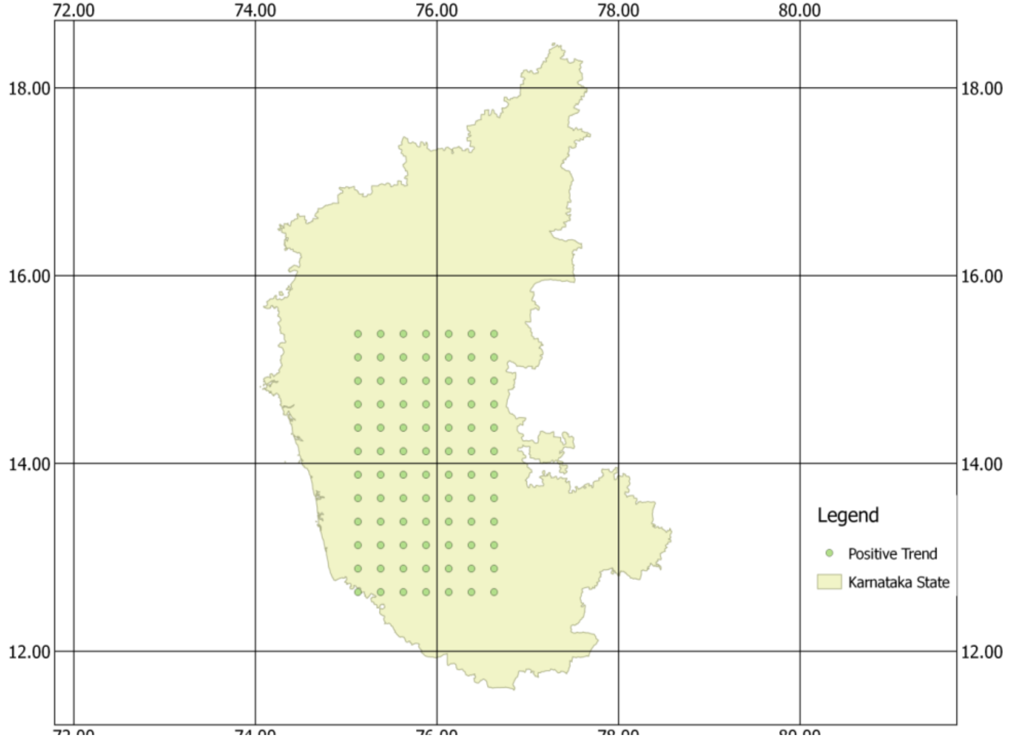
# STUDY OF EXTREME CLIMATE OF INDIA- CONTD..

1. Linear trend Analysis
2. Monotonic Trend - Increasing/Decreasing - Mann-Kendall Trend test
3. Magnitude of Trend- Sen’s slope test
4. Change point - Pettit Change point Detection Test

# EXAMPLE



Rainfall trend analysis



Temperature trend analysis

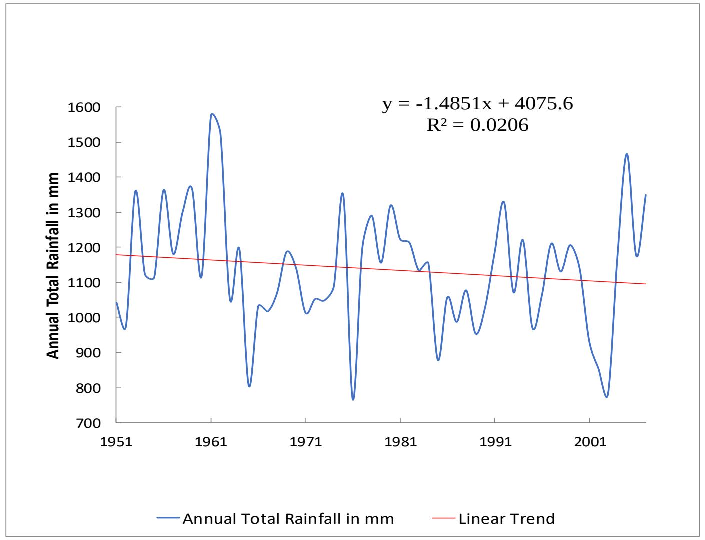
trend analysis with daily data of rainfall for period of 1951 to 2007 with Mann-Kendall trend test. Positive Trend - Increasing rainfall trend; Negative Trend – Decreasing rainfall trend; No Trend – no trend at 5% significance level

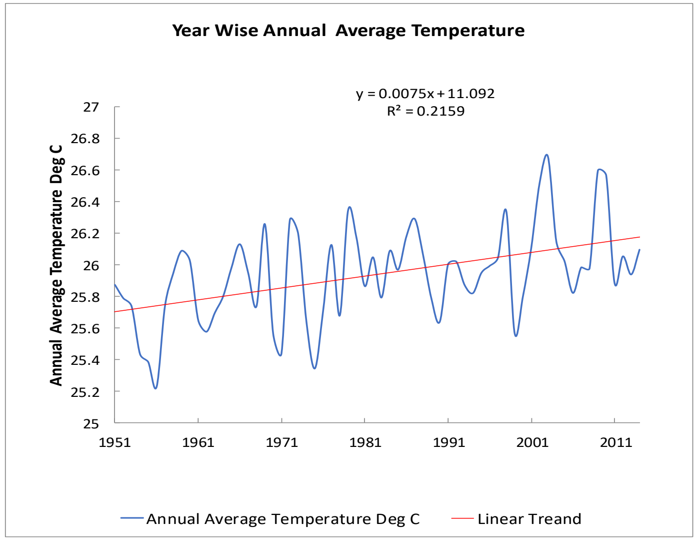
# LINEAR TREND

Spatial average of all valid grid points

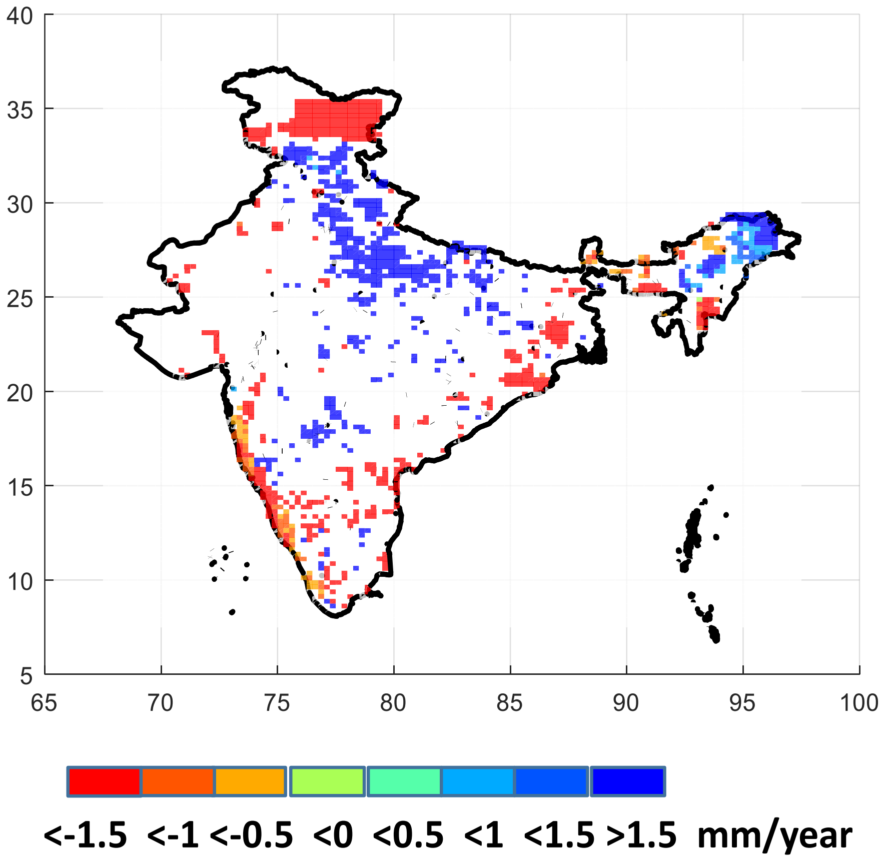
Time period - common data period of 1951 to 2014 has to be picked

Outcomes: Linear trend slope to ger the basic Idea of trend over India



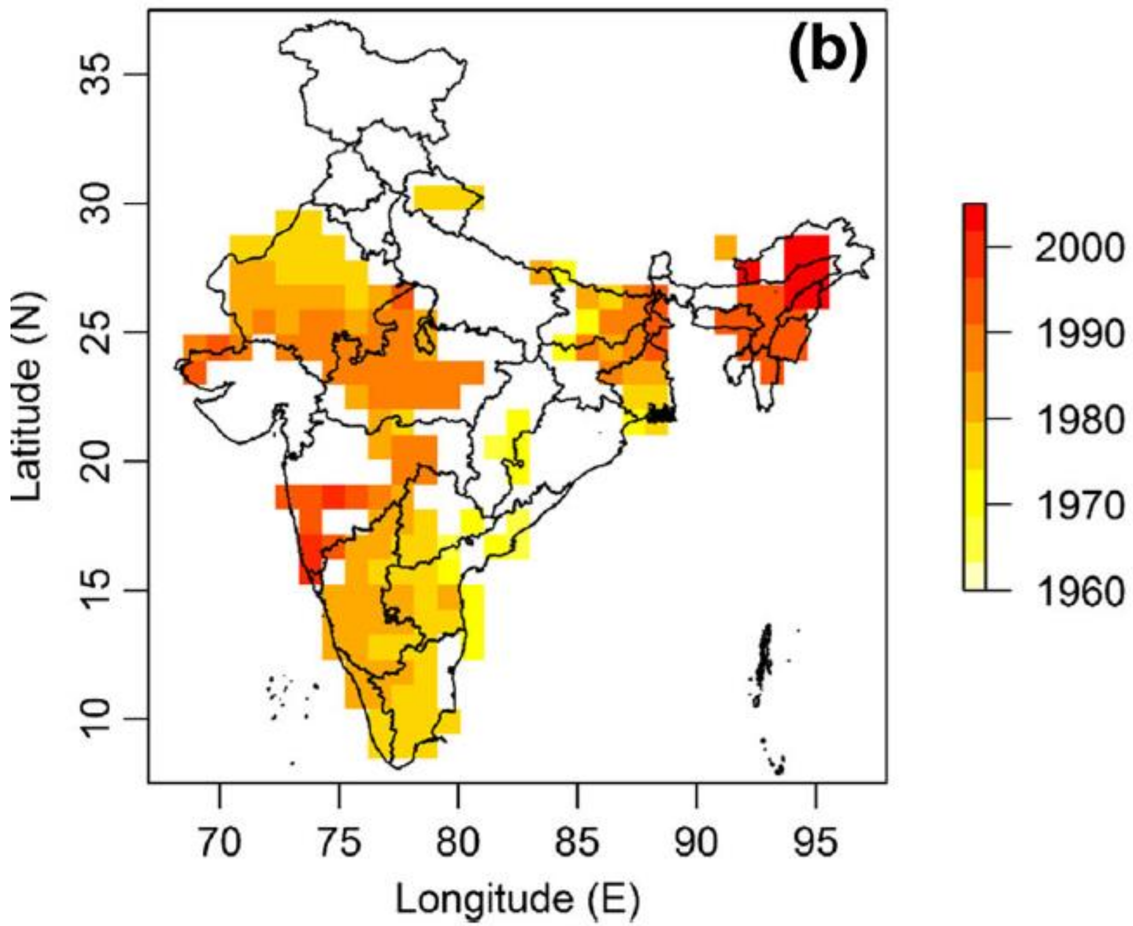


# MONOTONIC TREND



Sen’s slope for statistically significant (5% significance level) positive and negative trends with Mann-Kendall trend test at each grid point from 1951 to 2014

# CHANGE POINT DETECTION



# THANK YOU