CHAPTER 4

METHODOLOGY ADOPTED

4.1 General

The preferred techniques for analyzing precipitation trends are nonparametric methods, for they are distribution-free, robust against outliers, and possess a relatively higher potential to tackle non-normally distributed data (Adhikary et al. 2015; Barua et al. 2013; Mu et al. 2012) (Roxy et al. 2017) observed a monotonic tendency with rising precipitation in the state's mountain and coastal areas while analyzing spatio-temporal trend and concentration of monsoon precipitation time series in West Bengal, India, from 1901 to 2002. Furthermore, the monsoon arrives sooner in North Bengal than in South Bengal, where precipitation is more evenly distributed than in the latter. They found that the monsoon precipitation trend in Gangetic West Bengal corresponds to the overall pattern in India, allowing for complete forecasting. Various academics across the globe have used both parametric and nonparametric tests to detect a trend, size of trend, and changing point of trend in hydrologic and meteorological variables on occasion. When the hydrological data series is shifted from the normal distribution, the nonparametric test produces better outcomes than the parametric test (RM Hirsch 1982). Trend analysis of the precipitation data seeks reliance on non-parametric statistical tools. Non-parametric statistical techniques are often unaffected by outliers and other types of non-normality (Lanzante.1996). They are a metric for monotonic linear dependency.

The effectiveness of various nonparametric methods applied was analysed for trend analysis. However, the rank-based Mann-Kendall (MK) method is commonly applied for hydro-meteorological data series.

The Mann-Kendall test is the most frequently used non-parametric test for detecting trends in hydrologic variables. The literature review also reveals the same fact. A statistically significant trend has to be identified in time series data using the non-parametric Mann-Kendall model, and Sen's slope estimation is typically used. Integration of these two methodologies has been relied upon by most of the researchers. Therefore, the same framework has been adopted in this study to carry out the analysis.

In order to carry out the analysis, data characteristics, such as mean, standard deviation, and percentage contribution to annual rainfall, have been determined for the annual and seasonal time series, i.e., pre-monsoon (March-May), monsoon (June-September), post-monsoon (October-November) and winter (December-February). The data collected for the five scenarios listed above were subjected to a 10-year trend line to determine long-term trends. Sen's technique was used to add a linear trend. Sen's slope estimate technique was used to conduct a non-parametric Mann-Kendall test. The details of the non-parametric test and Sen's methodology are explained hereunder.

The approach is then applied to hypothetical datasets, which involve trend and abrupt change detection methods. Theoretical datasets were created to reflect all conceivable changes in the chosen climatic variables to evaluate the change detection methods' applicability. This section also includes theoretical data analysis. The technique for analyzing the change in actual climatic datasets is then given based on this theoretical data analysis.

Statistical techniques were chosen as suitable instruments to detect changes in the specified variables. High-quality historical datasets, which were accessible from Meteorology databases, are required for appropriate statistical analysis results. Statistical tests were performed to determine whether or not there was a change in the chosen climatic variables and the flow chart of the methodology is illustrated in Figure 4.1.

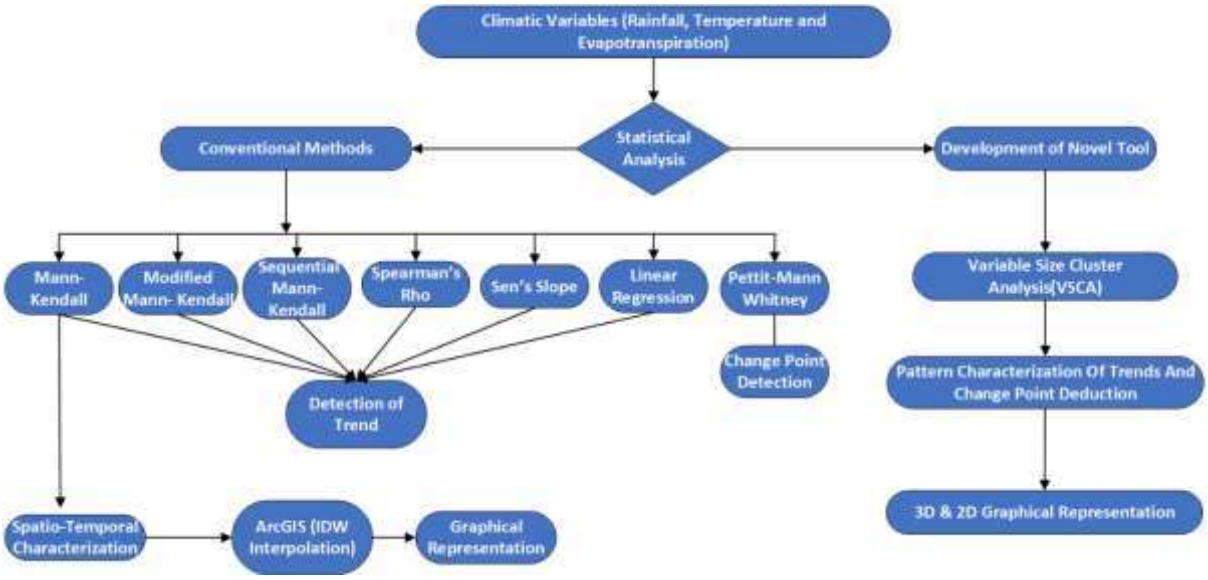

Figure 4.1 Flow Chart of Methodology

4.2 Statistical Techniques

4.2.1 Autocorrelation (Serial dependency check and removal)

Autocorrelation checks serial dependency between the long-term time series of climatic observations. Serial correlation having a positive value characterizes persistence, which means the data are dependent on each other in the hydro-climatological time-series studies, and MK statistic is underestimated. Serial correlation is the persistent problem in determining and understanding trends in a particular climatological time series. Nevertheless, this method has been commonly implemented in various past works (Kundu et al. 2015; Türkeş et al. 2002).

Data is normalized using the following expression:

$$Y_t = (y_t - \bar{y}) / \sigma$$

y_t - observed time series.

$\bar{y}$ - mean of annual/seasonal long-term time series.

σ- standard deviation of annual/seasonal long-term time series.

However, as suggested by (Basistha et al. 2008), serial correlation coefficients are calculated up to lag3. The coefficient between Y_t and Y_{t+L} can be equated as;

$$\rho_L = \frac{\sum_{t=1}^{n-L}(Y_t - \bar{Y}_t).(Y_{t+L} - \bar{Y}_{t+L})}{\left[\sum_{t=1}^{n-L}(Y_t - \bar{Y}_t)^2.\sum_{t=1}^{n-L}(Y_{t+L} - \bar{Y}_{t+L})^2\right]^{1/2}}$$

Eq. (12) below gives the results of the significance test,

$$\rho_k = \frac{-1 \pm t_g (n-k-1)^{1/2}}{n-k}$$

Different confidence intervals such as 90, 95, and 99 % having the values of t_g is 1.645, 1.965, 2.326, respectively. For $\rho_{(L=1)} \geq \rho_{(k)}$, the null hypothesis H_0: the randomness of the climatic time series data is rejected against the serial correlation; the series is then considered free from persistence. Therefore, the blended series (Z_t) can be obtained with additive white noise.

4.2.2 Pre-whitening

Before implementing the MK test, removing the effect of serial correlation using the pre-whitening procedure (Dash et al. 2009) was done. Trend test with the pre-whitening method (Yue et al. 2001);(Arora et al. 2005; Kishore et al. 2016; Sharma et al. 2019; Suryavanshi et al. 2014) has been implemented to identify a significant trend and persistence in this study.

Time-series data is de-trended, Y_t' by using the normalized dataset Y_t, and the median values of the slope, β. Before analyzing the trend, the time series data is divided by the sample mean, $\overline{Y_t}$ such that the properties remain unchanged with mean equals to unity. If the slope tends to zero, then there is no requirement for analysis of the trend. The slope is assumed to be linear with a value differing from zero, de-trending is carried out using;

$$Y_t^{'} = Y_t - T_t = Y_t - \beta * t$$

The original time series data are supposed to be successively independent, and then if ρ_1 is approximately equalled to zero, the MK test can directly be implemented to the sample data. For serially correlated data, ρ_1 is removed from the $Y_t^{'}$ by

$$Z_t^{'} = Y_t^{'} - \rho_1 . Y_{t-1}^{'}$$

The independent residual series, Z_t' obtained after the trend-free pre-whitening procedure, is the one that does not contain any trend. In the blended series, Z_t has a linear dependency with the identified trend, T_t, and the residual. The blended series can be used for trend estimation.

$$Z_t = Z_t^{'} + T_t$$

Twelve statistical test methods are proposed hereafter, chosen based on the World Meteorological Organization (WMO)/UNESCO World Climate Programme (WCP) Expert Workshop on "Detecting Trend and Other Changes in Hydrological Data" and the Cooperative Research Center in Catchment Hydrology (CRCCH) "Hydrological Recipes," Australia. Change in a time series can occur steadily (a trend), abruptly (a step-change), or in a more complex form. It may affect the mean, median, variance, or other aspects of the time series.

The test methods are categorized as parametric and non-parametric ones. The distinction between these two should be clear from their name. Parametric test procedure proved to be more powerful than its counterpart. The test types used are listed, and each test method is described in detail in the subsequent sections.

The starting point in each test's procedures is the definition of a null hypothesis *(H_0)* and an alternative one *(H_1).* The definitions of these two hypotheses are based on what was intended to be tested. For example, suppose H_0 is defined as "no change in the mean of the flow data,"

and H_1 as "the mean is either increasing or decreasing with time." Data with an insignificant length of the records are dropped out of the tests. Significance levels (α) of 10, 5, and 1 % are applied for inference purposes and comparisons between a test statistic and a critical value, obtained from the calculation and appropriate statistical tables.

It is necessary to carry out this analysis because most water resources systems have been designed with the basic assumption of stationary hydrology. It is important to study the stationarity properties and possible changes/trends in our time series of hydro climatologic and hydrological data; otherwise, we may be attempting to underestimate the flood or low flow magnitudes over-or.

4.3 Trend Detection Methods

For trend identification, many statistical tests are available. The preferred techniques for analyzing precipitation trends are nonparametric methods, for they are distribution-free, robust against outliers, and possess a relatively higher potential to tackle non-normally distributed data. (Yue et al. 2001) (Sonali and Nagesh Kumar 2013) have been analyzing the efficiency of several non-parametrical techniques for trend analysis; however, the Mann-Kendall (MK)-based method is often used in hydrometeorology, while the rho test (SR) of Spearman are seldom utilized (Hirsch et al. 1982). While linear regression (LR) is parametric, it implies that information is usually distributed (Jamle and Meshram 2019). Sen's slope estimator is employed to estimate the size of the trend. In short-term rainfall, extreme, continuous, or monotonic trends may be detected using non-parametric Mann-Kendall trend statistics at 5 percent and 10 percent ties.

Non-parametric trend tests have been used to evaluate trends in climatological variables for the study station. Non-parametric tests are "distribution-free" and, as such, can be used for non-normal variables, whereas parametric tests are those that make assumptions about the

parameters of the data distribution from which the sample is drawn. This is often the assumption that the meteorological data are normally distributed. Before applying the MK test, we implemented lag1 autocorrelation to check the serial correlation. The positive serial correlation (persistence) has been diminished from climatic data using pre-whitening and applied the trend test to it. The SQMK test has also been conducted to understand the fluctuations in trends over the study period. Sen's slope estimator has calculated the magnitude of the trend. MK test has been used wherever the autocorrelation in data is found to be insignificant. The following sections below discussed the statistical methods.

4.3.1 A brief description of the MK test

All of the hydrologic variables' time series were analyzed using the Mann-Kendall non-parametric trend test. Mann was the first to use this test, while Kendall was the first to calculate the test statistic distribution. The observed patterns' importance may be determined (generally adopted as five percent). The technique is very resistant to the effects of extremes and works well with skewed variables. It can handle missing data values efficiently. Furthermore, many researchers have discovered that this technique is most often used in rainfall trend detection investigations and provides trustworthy findings (Ghosh et al. 2009; Karpouzos et al. 2010) This test has also been utilized in other hydro-meteorology time-series investigations such as streamflow (Kumar et al. 2017). (Zhang et al. 2010) used the Mann-Kendall test to investigate the impact of data series length on streamflow trend detection. To evaluate the trend pattern of the data series, they recommended using the Mann-Kendall test repeatedly with different origin and destination times. The Mann-Kendall test statistic for a time series displaying data as R_1, R_2 …, R_n is given by:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn\left(R_j - R_i\right)$$

$$sgn\left(R_j - R_i\right) = \begin{cases} +1 & if\ R_i < R_j \\ -1 & if\ R_i > R_j \end{cases}$$

In a consistently increasing trend of a dataset, the test statistic displays a positive value. In contrast, a consistently decreasing trend in time series would yield a negative value of 'S.' However, a larger magnitude of 'S' indicates that the trend is relatively more consistent in its direction. The variance of the Mann-Kendall statistic is determined by:

$$E(S) = 0$$

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

Where t_p is the cluster size of p_{th} tied groups and q is the number of tied groups, and n is the length of time-series data.

The test statistic Z is determined as follows:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{\text{var}(S)}} & if\ S > 0 \\ \dfrac{S+1}{\sqrt{\text{var}(S)}} & if\ S < 0 \end{cases}$$

The hypothesis of no trend (H_0) is rejected if $|Z| > Z_{1-\alpha/2}$, where Z is picked from the standard normal distribution table and α is the level of significance (Barnard 1947). That signifies that there is a probability α for the trend being falsely identified. The induction of expected variance of S for the determination of Z provides the Mann-Kendall test with an ability to reject what might appear to be a trend over small periods as the fractional data series poses to exhibit a minor trend. The importance level is the criteria for rejecting the null hypothesis in hypothesis testing (H_0). Although the level of significance decision is still mainly subjective, analysts have used both 0.05 (often referred to as 5% of significance) and 0.01 (1% of

significance). The lower threshold of significance indicates that more data should differ substantially from the zero hypothesis. 0.05 of significance has been taken as more cautious than 0.01 thresholds in the current rainfall variability analysis.

4.3.2 Assumptions

The MK test is based on the following assumptions:

• When no trend exists, the measurements (observations or data) collected over time are independent and identically distributed. The condition of independence implies that the observations are not serially connected across time.

• The observations acquired throughout time are indicative of the actual circumstances at sampling periods.

• The sample collection, processing, and measurement procedures offer unbiased and representative observations of the underlying populations across time.

There is no need that the measures be normally distributed or that the trend, if present, be linear. The MK test can be calculated even if there are missing data and values below one or more limits of detection (LD), but the test's performance will suffer as a result. The assumption of independence demands that the interval between samples be sufficiently long such that there is no correlation between data taken at various times.

The Mann-Kendall (MK) test determines if a variable of interest has a monotonic upward or decreasing trend over time. A monotonic upward (downward) trend indicates that the variable continuously rises (decreases) with time, although the trend may or may not be linear. The MK test may be used instead of parametric linear regression analysis, determining if the slope of the predicted linear regression line is greater than zero. The regression analysis requires that the residuals from the fitted regression line be normally distributed; this assumption is not needed by the MK test, which is a non-parametric (distribution-free) test. According to

(Hirsch et al. 1982) the MK test is best regarded as an exploratory study and should be used to discover stations with substantial magnitude changes and to quantify these results.

4.3.3 Modified Mann–Kendall test

In the presence of autocorrelation, pre-whitening has been used to identify a trend in a time series. On the other hand, pre-whitening has been shown to decrease the detection rate of significant trends in the MK test (Yue et al. 2001) As a result, the MMK test was used to identify trends in an autocorrelated dataset. After removing a non-parametric trend estimate such as Theil and Sen's median slope from the data, the autocorrelation across ranks of the observations, k is assessed. Because the variance of S is overestimated when the data are positively autocorrelated, only significant values of k are utilized to compute the variance correction factor n/n S:

$$\frac{n}{n_S^*} = 1 + \frac{2}{n(n-1)(n-2)} \times \sum_{k=1}^{n-1} (n-k)(n-k-1)(n-k-2)\rho\kappa$$

N is the actual number of observations, n∗ s is considered an ‘effective number of observations to account for autocorrelation in the data, and ρk is the autocorrelation function of the ranks of the observations. Several for significant autocorrelation in data. The corrected variance is then computed as

$$V^*(S) = V(S) \times \frac{n}{n_S^*}$$

V (S) is from Equation (6); the rest is in the MK test.

4.3.4 Sequential Mann–Kendall analysis

Detection of Trend by MK test at the closing of any time period does not give an overall view of trend (trend structure) for a whole time series. The central idea of the SQMK tests is to check trend fluctuations over the time period, and this analysis considered the proportionate

values of each term in the time series (y_1, y_2 ... y_n). Sequential progressive series u(t) and backward series u'(t) obtained from the Sequential Mann-Kendall Rank Statistic (SMKRS) were used to recognize the abrupt changes (Chen et al. 2016). Herein, u(t) standardized variable fluctuates in its sequential behavior around zero, and u(t) will be similar to the Z_{MK} value. There is a statistically significant trend if the two series cross each other, and at that time, it diverges and surpassed particular threshold values (Bandyopadhyay et al. 2009; OI Abdul Aziz 2006; Shadmani et al. 2011). Sometimes, the positive and negative trend is always not noteworthy, so it can be detected more significant trend by using SQMK graphs (Rahman et al. 2016).

The mean annual time series, y_j (j = 1, 2 ...n) are compared with y_i (i= 1, 2 ...j-1) are the sequential values of SMKRS in the series. The number of circumstances $y_j > y_i$ is counted at each comparison and represented by n_j. The statistics t_j is thus calculated as

$$t_j = \sum_{i=1}^{j} n_j$$

The Mean $E(t_j)$ and variance $Var(t_j)$ of statistics t_j is the calculated using following equations

$$E(t_j) = n(n-1)/4 \text{ and}$$

$$Var(t_j) = \left[j(j-1)(2j+5) \right] / 72$$

The sequential progressive value u(t) of the statistic is given by

$$u(t) = \frac{t_j - E(t_j)}{\sqrt{\text{var}(t_j)}}$$

Likewise, the value of backward sequential (u'(t)) statistic of the SMKRS is estimated beginning from the last of the series. In the present study, this method is applied to recognize the turning points of a trend.

4.3.5 Spearman's rho method

Spearman's Rho test is another non-parametric method to measure the strength of association, represented by 'r,' between the two variables. The values of 'r' lie in the range of ±1, where '*r*' = 1 means a perfect positive correlation and -1 represents a perfectly negative. The application of Spearman's Rho methodology seeks a few prerequisites on a dataset, and they are given as below:

1. Scale of measurement must be ordinal (or interval, ratio).

2. Data must be in the form of matching pairs.

3. The association between the datasets must be monotonic, i.e., variables either increase in values together, or one increases while the other decreases.

Applying this method in the trend analysis of precipitation is relatively lesser; hence, it has not been attempted in the present study, and further description of the method is not being dealt with further.

Spearman's rho test is another rank-based nonparametric method used for trend analysis (Yue et al. 2001). The null hypothesis (H_0) indicates no trend over time; the test statistics R_{sp} and standardized statistics Z_{sp} are defined as follows:

$$R_{sp} = 1 - \frac{6\sum_{i=1}^{n}(D_i - i)^2}{n(n^2 - 1)}$$

$$Z_{sp} = R_{sp}\sqrt{\frac{n-2}{1-R_{sp}^2}}$$

Where D_i is the rank of i_{th} observation, i is the chronological order number, n is the length of the time series data, and Z_{sp} is Student's t-distribution with $(n-2)$ degree of freedom. The positive values of Z_{sp} represent an increasing trend across the hydrologic time series whereas,

negative values represent a decreasing trend. The critical value at a 0.05 significance level of Student's t-distribution table is defined as ($n-2$, $1-\alpha/2$). If $|Z_{sp}| > (n-2, 1-\alpha/2)$, ($H_0$) is rejected for $|Z_{sp}|>2.08$, and a significant trend exists in the hydrologic time series.

4.3.6 Sen's slope estimator

Sen's slope is also a non-parametric test used to calculate the magnitude of the linear trend represented by the median value of the slope values of the data sets. Having determined the nature of the trend in the time series data of precipitation, Sen's estimator predicted the magnitude of the trend. The slope (T_k) for the data sets/ pairs is determined from the equation proposed by Sen (1968) as:

$$T_k = \frac{y_j - y_i}{j - i} \qquad \text{for k = 1, 2, …N}$$

y_j is a *j-th* data point, and y_i is an *i-th* data point in time series, and j should be greater than i, i.e., $j > i$. Sen's slope estimator's main advantage lies in the global median value used, making it more resilient to the consequence of extreme values in the time series. The median of these values of slopes (T_k) is calculated by

$$\beta = \begin{cases} T_{N+1/2} & if\ N\ is\ odd \\ \frac{1}{2}(T_{N/2} + T_{N+2/2}) & if\ N\ is\ even \end{cases}$$

The positive β value represents an increasing trend, and the negative β value indicates a decreasing trend in the time series. Having determined T_k as above, Sen's slope estimator β is confirmed by a two-sided test at 100*(1 – α) % confidence interval.

4.3.7 Linear regression method

A linear regression method was used to check whether there is a significant relationship between the variables under consideration. The regression line is used to estimate the slope.

The slope indicates the mean temporal change of the studied variable. Positive slope values show increasing trends, while negative slope values indicate decreasing trends. A linear regression line is an equation of the form:

$$y = a + bx$$

x is the explanatory variable, y is the dependent variable, is the slope of the line, and a is the intercept (Gocic and Trajkovic, 2013).

Pettitt Mann Whitney (PMW) *Ut*, t-test is usually applied to identify the point of significant change. Details for this method are described in brevity as:

4.3.8 Pettitt Mann Whitney Test

Pettitt Mann Whitney (PMW) Ut, t-test is usually applied to identify the point of significant change. This method is briefly described here.

Pettitt Mann Whitney (1979) test is a nonparametric protocol that discerns changes in the mean (median) of a time series with the unclear and non-evident presence of change points and is commonly applied to detect a single change-point in hydrological series or climate series with continuous data. It tests the null hypothesis (H_0), which represents the variables *(T)* following one or more distributions that have the same location parameter (no change), against the alternative hypothesis (H_a), which represents that a change point exists. The non-parametric statistic K_T is defined as:

$$K_T = \max |U_{t,T}|$$

Where

$$U_{t,T} = \sum_{i=1}^{t} \sum_{j=t+1}^{T} \operatorname{sgn}(X_j - X_i)$$

$U_{t,T}$ is the Pettitt test index, and the change point of the series is located at K_T when the statistic is significant.

Further,

$$\operatorname{sgn}(X_j - X_i) = \begin{cases} +1 & ...if\ X_j > X_i \\ -1 & ...if\ X_j < X_i \end{cases}$$

The significance probability of K_T is approximated for $p \leq 0.05$ as:

$$p \approx 2\exp\left(\frac{-6K_T^2}{T^3 + T^2}\right)$$

4.3.9 Proportionate Change

The equation is as follows for the calculation of proportionate change (PC) of different climatic variables.

$$PC = \frac{n * \beta}{|y|} * 100$$

Where |y| = absolute average value, while β, n is the median of the slope of the time series data and length of the time period.

Nevertheless, trend detection methods have their pros and cons. Hence, it is important to consider the limitations of the methods. A more technically sound and robust method should be developed to overcome the limitations and improve the trend detection analysis.

4.4 Summary

The results of trend analysis would aid in the formulation of effective water management strategies and develop appropriate mitigation measures to protect water resources. The extreme value trend analysis results would aid in forecasting the pattern and intensity of future extreme occurrences in further depth. That would make it easier to create efficient

safety preventive measures, such as increasing drainage capacity and redesigning stormwater drainage systems for floods, and (ii) developing appropriate drought mitigation methods for drought, which has a high human cost in the contemporary day.