ENGINEER - Vol. LIII, No. 02, pp. [69-75], 2020
© The Institution of Engineers, Sri Lanka
DOI: http://doi.org/10.4038/engineer.v53i2.7414

# Screening of Annual Rainfall Time-Series Data in Kala Oya Basin: Case Study in Sri Lanka

**A.D.S. Iresh**

**Abstract:** Hydrometeorological data screening is essential analysis conducted prior to the use of such data for modelling and designing of water development schemes as inconsistencies can occur during the data collection or data entering. There are many methods adopted to perform screening of data by various publishers, but this paper presents a prominent method which has been used for the rainfall annual time-series of Kala Oya basin. Adequate numbers of rainfall gauging stations and outliers have been checked using statistical analysis and percentage significance level outlier constant for normal distribution. Trend analysis and absence of persistency have been estimated by the Mann-Kendall test and the first serial correlation coefficient method. Homogeneity of the region was checked by doing two statistical tests based on L-Moments of the time-series data. The first criterion was Discordancy measure (*Di*) and the second criterion was Heterogeneity measure (*Hi*) which have been based on L-moments of the time-series data. The results confirmed that the selected annual time-series data can be considered as homogeneous and consistent with minor deviations.

**Keywords:** Kala Oya, Annual Rainfall, Time-Series Data, Data Screening

## 1. Introduction

A longer time-series gives a greater chance of getting time-series as non-stationary, non-consistent or non-homogeneous. Hydrological time-series data can consist of errors or unconformities which cause inconsistencies and non-homogeneities due to causes of natural or manmade phenomena. Data-screening prior to modelling and designing of water development schemes is an essential requirement. Unscreened data may create false estimations, hence designs. Decisions made based on these false estimations or designs create irreversible damage to the environment, wildlife and humans. The importance of data-screening of time-series has been addressed by several researchers. Hosking and Wallis [3] reported that homogeneity measures provide an initial screening of data and indicate sites where the data may merit close examination. The aim of this paper is to elaborate on the complete data-screening procedure and apply it to the Kala Oya basin, Sri Lanka.

Seventeen rainfall gauging stations were selected for this study and, for each station, annual rainfall time-series was taken into the data-screening process. Selected time-series data for this study is from 1985 to 2018. Reliable data compensate reliable hydrological studies and enhance the quality of the results. The effort of this study is to obtain reliable rainfall time-series for the Kala Oya basin hydrological studies.

## 2. Study Area

Kala Oya stream originates from central mountains of Dambulla at an elevation about 870 metres above mean sea level. Kala Oya is the third longest river in Sri Lanka. The river flow generates from the central province and flows through the north-central province and falls into the sea from north-western province at a place called Gangewadiya which belongs to Wilpattu national wildlife park. Kala Oya watershed is located in four administrative districts, namely Anuradhapura, Matale, Kurunagala and Puttlam. Kala Oya basin location details are presented in Figure 1.

## 3. Methodology

The methodology adopted to screen the data starts from the selection of an adequate number of stations for the study. Time-series data exploratory analysis has been done by graphically plotting the time-series. Annual rainfall time-series high and low outliers have been calculated by using percentage significance level outlier constant for normal distribution.

*Eng. A. D. S. Iresh, C. Eng., MIE(SL), IESL Part I, II, III, M.Tech. (India), Dip(Irrigation Engineering.), Irrigation Engineer (Hydrology Division), Irrigation Department, Sri Lanka.*
*Email:shahikairesh@gmail.com*
*http://orcid.org/0000-0003-2345-3769*

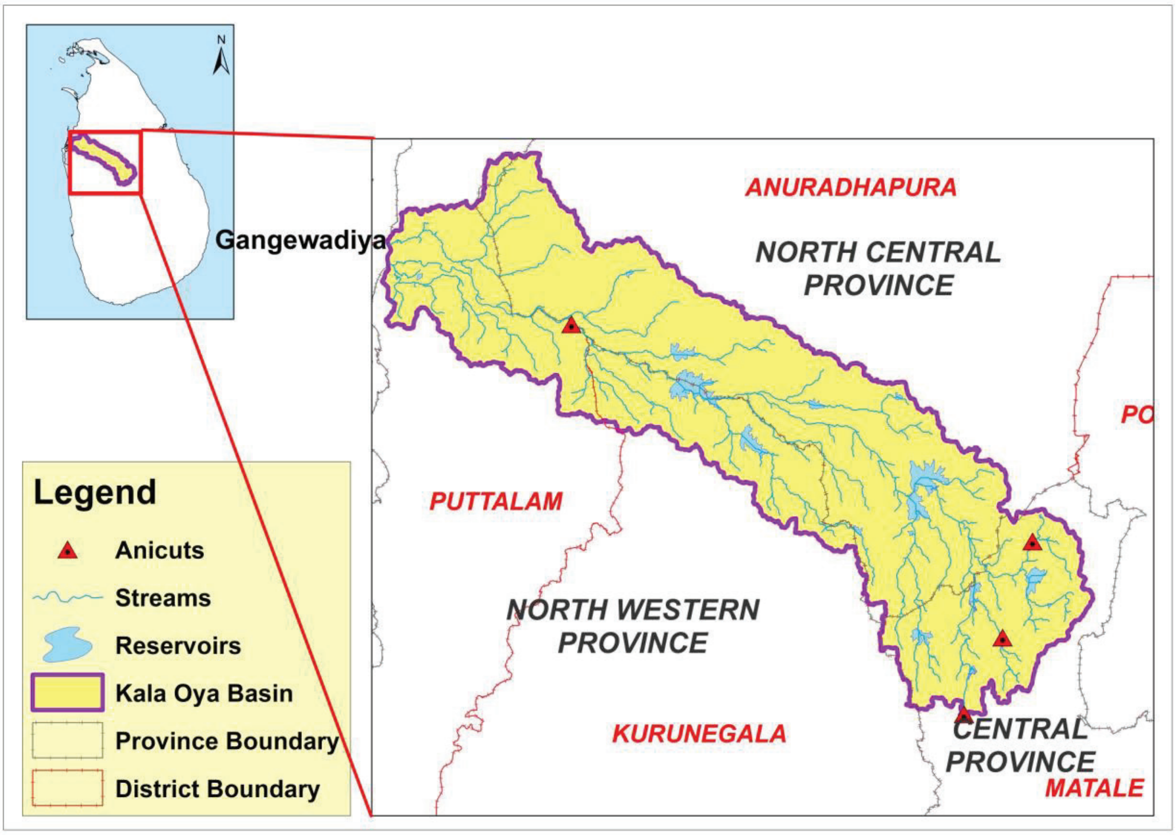

**Figure 1 – Kala Oya Basin Detail**

Modified Mann-Kendall test has been used to detect the short term trend of annual time-series. Absence of persistence has been examined by the first serial correlation coefficient method. If all the above criteria are satisfied, then homogeneity of the region has been analysed by doing two statistical tests based on L-Moments of the time-series. L-Moments have been estimated by probability weighted moments. If the selected stations time-series data satisfy all criteria, then the data can be used for basin hydrological studies. If the data series are not satisfying the criteria, then the time-series need to be changed or need to be corrected using a reliable correction procedure. Once changed, the time-series should undergo the screening procedure again. Step by step approach forms the complete data-screening procedure, which is illustrated by the flow chart in Figure 2.

### 3.1 Check for Adequacy

Data screening basic procedure begins with the check for adequacy of rain gauging stations inside the selected basin. The total number of stations inside the Kala Oya basin is nine. From the first adequacy check, it was identified that nine stations are not sufficient for the study. Hence numbers of stations have been increased by adding a 10 km buffer zone. The selected stations and buffer zone for the study basin are presented in Figure 3. Subramanya [1] provides a statistical method to obtain an optimal number of stations that should exist for a study area. The statistical test variables can be written as:

$$C_v = \frac{100 \times \sigma_{m-1}}{\bar{P}} \qquad \dots.(1)$$

$$\epsilon_{ex} = \frac{C_v}{\sqrt{m}} \qquad \dots.(2)$$

$$N = \frac{C_v}{\epsilon_{ex}} \qquad \dots.(3)$$

Where $C_v$ is the coefficient of variation. $\varepsilon_{ex}$ the expected error (in percentage), $\sigma_{m-1}$ the standard deviation of annual time-series data, $\bar{P}$ the mean precipitation, $m$ the number of stations selected and $N$ is the optimal number of stations required for the study.

### 3.2 Outlier and Unconformity Check

An outlying observation may be due to an extreme climatic event. If that is true, the value should be retained and used for the studies as other observations. Then again, an outlying observation may be the result of an error in collecting or recording the numeric value. In such cases, it may be desirable to check the outliers to ascertain the aberrant values. The aberrant value may even eventually be rejected to maintain the reliability of the data series. Outlying observation or "outlier" is one that appears to deviate markedly from other members of the sample in which it occurs [2]. Outliers of the annual rainfall time-series data have been checked by estimating the high and low outliers using percentage significance level outlier constant (*K*) for normal distribution. The constant *K* has been selected according to the data length of each time-series. The test criterion, High and Low outlier criteria can be written as:

$$High\ outlier\ (H.O) = Xmean + K * Stdx \quad ....(4)$$

$$Low\ outlier\ (L.O) = Xmean - K * Stdx \quad ....(5)$$

$X_{mean}$ is the mean value of the data series. *Stdx* is the standard deviation of the data series. *K* is the constant selected according to data length suggested by Grubbs and Beck [2]. When data length is 33, the *K* value has been selected as 2.79 and varied according to the data length.

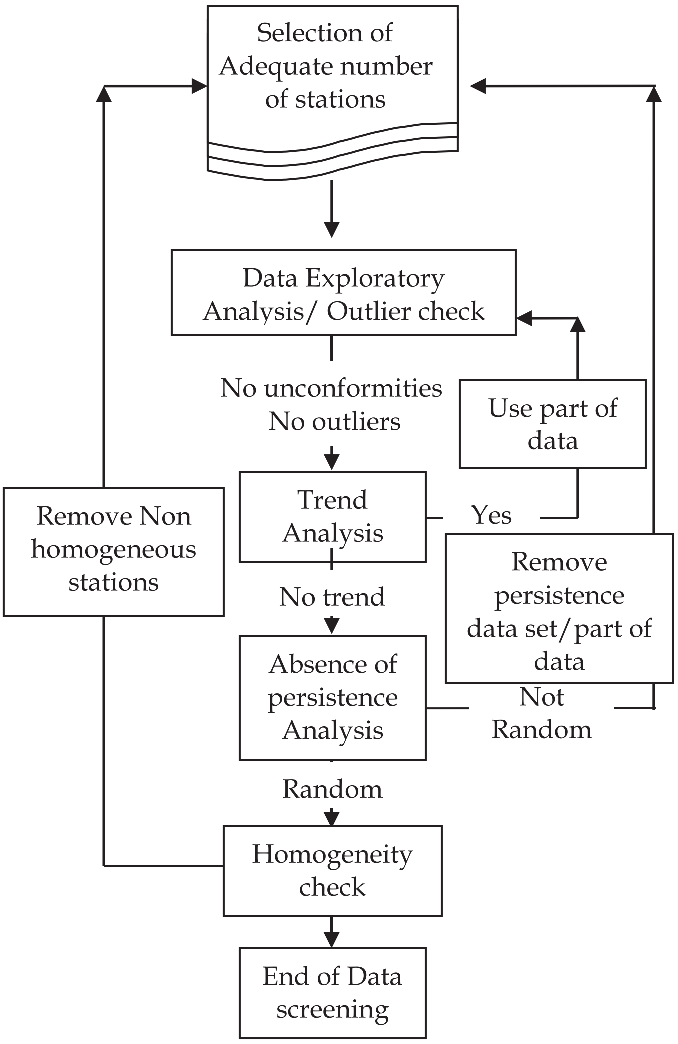
Selection of Adequate number of stations
Data Exploratory Analysis/ Outlier check
No unconformities No outliers
Use part of data
Trend Analysis
Yes
Remove Non homogeneous stations
No trend
Remove persistence data set/part of data
Absence of persistence Analysis
Not Random
Random
Homogeneity check
End of Data screening

**Figure 2 – Data Screening Flow Chart**

### 3.3 Trend Analysis

Mann–Kendall test (MK) with tie correction has been used to detect the monotonic trend of annual rainfall time-series [4]. For the observed annual rainfall time-series data of $X = x_1, x_2, x_3, \ldots, x_n$ MK statistic was estimated as follows.

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_j - x_i) \quad ....(6)$$

where, $\upsilon = x_j - x_i$

$sgn(\upsilon) = 1$ if $\upsilon > 0$; $sgn(\upsilon) = 0$ if $\upsilon = 0$;

$sgn(\upsilon) = -1$ if $\upsilon < 0$,

For $n \geq 8$, *S* is normally distributed as:

$$E(S) = 0 \quad ....(7)$$

$$Var(S) = \frac{n(n-1)(2n+5) - t_c}{18} \quad ....(8)$$

Tie correction define as:

$$t_c = \sum_{i=1}^{n} t_i \,.\, i.\,(i-1)(2i+5) \quad ....(9)$$

$$Z_M = \begin{cases} \dfrac{s-1}{(var(s))^{\frac{1}{2}}} & if \quad s > 0 \\ 0 & if \quad s = 0 \\ \dfrac{s+1}{(var(s))^{\frac{1}{2}}} & if \quad s < 0 \end{cases} \quad ....(10)$$

$Z_M$ is standard normally distributed with zero mean and unit variance. If the calculated $Z_M$ statistics lies within the limit of -1.96 and 1.96, then it is considered the null hypothesis of having no trend at 5% significance level.

### 3.4 Absence of Persistence Analysis

Time-series of yearly and seasonal totals are usually independent. But extreme rainfall events may create aberrant values. Such observations should be discarded. If a reliable correction procedure is available it may sometimes be corrected and retained. Hence it is essential to test the time-series for independence. The serial-correlation coefficient can assist with confirming the independence of a time-series. For this study, it is sufficient to compute lag 1 serial-correlation coefficient, i.e. the correlation between adjacent observations in a time-series. The first serial-correlation coefficients for the 17 stations annual time-series have been estimated. The estimated correlation coefficient $r_1$ was checked for the 5% significance level.

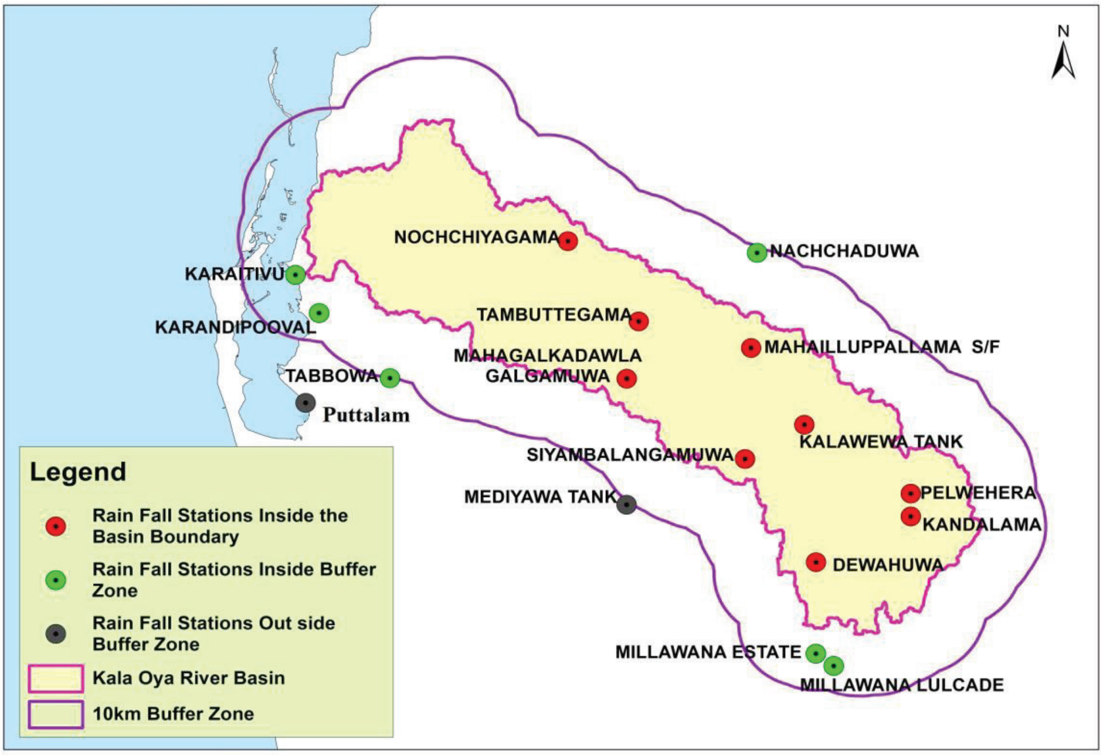

**Figure 3 – Selected Rainfall Station Locations**

For: $X = x_2, x_3, x_4$ ….. , …… , $x_n$ and $Y = x_1, x_2, x_3$ ,…. , …… , $x_{n-1}$, the test parameters were estimated as follows.

$$r_n = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \quad ....(11)$$

where $r_1$ upper and lower limits defined as,

$$r_{1\,upper} = \frac{[-1 + 1.96(n-2)^{0.5}]}{(n-1)} \quad ....(12)$$

$$r_{1\,lower} = \frac{[-1 - 1.96(n-2)^{0.5}]}{(n-1)} \quad ....(13)$$

$r_1$ is standard normally distributed with zero mean and unit variance. If the values of the $r_1$ statistics calculated lie within the limit of $r_{1upper}$ and $r_{1lower}$, and then it is considered the null hypothesis of having no persistence at 5% significance level.

### 3.5 Homogeneity Check

For a basin hydrological study, all the sites located in the basin must have a homogeneous time-series. For these reasons, the homogeneity of the time-series needs to be tested. Hasking and Wallis [3], [5] introduced two statistical tests based on L-moments of time-series data to check the homogeneity of the data series.

The first statistic is discordancy measure ($D_i$) [3], [5], [6], [7]. The discordancy statistical test, discordancy index $D_i$ is defined by:

$$D_i = \frac{1}{3}\langle (u_i - \bar{u})^T (u_i - \bar{u}) S^{-1} \rangle \quad ....(14)$$

where $i$ is the site number, $u_i$ is the L-moment vector of $i^{th}$ site and $u_i$ is the vector which includes sample L-moments of $t$, $t_3$ and $t_4$. Then $u_i$ is defined as:

$$u_i = [t^i, \quad t_3^i, \quad t_4^i] \quad ....(15)$$

$\bar{u}$ is the average L-moment vector of $n_s$ (number of sites) and can be written as:

$$\bar{u} = \frac{1}{n_s}\sum_{i=1}^{n_s} u_i \quad ....(16)$$

$S_D$ is the sample covariance matrix and can be defined by:

$$S_D = (n_s - 1)\sum_{i=1}^{n_s} (u_i - \bar{u})(u_i - \bar{u})^T \quad ....(17)$$

For any site, if $D_i \geq 3$, it is considered as disharmonious [3], [5], [6], [7].

The second statistic is heterogeneity measure ($H_i$) [3], [5], [6], [7].

Heterogeneity measure is defined by three statistics, $H_1$, $H_2$, and $H_3$. The statistic $H_i$ is estimated as follows:

$$V_1 = \frac{\sum_{i=1}^{n_s} n_i \, (t^i - \bar{t})^2}{\sum_{i=1}^{n_s} n_i} \qquad \ldots.(18)$$

Where, $n_s$ is the number of sites and $n_i$ is the recorded length of each site. $\bar{t}$ is the average of $t^i$ values defined as:

$$\bar{t} = \frac{\sum_{i=1}^{n_s} n_i \, t^i}{\sum_{i=1}^{n_s} n_i} \qquad \ldots.(19)$$

$$V_2 = \frac{\sum_{i=1}^{n_s} n_i \left[ (t^i - \bar{t})^2 + \left[ t_3^i - \bar{t}_3 \right]^2 \right]^{1/2}}{\sum_{i=1}^{n_s} n_i} \qquad \ldots.(20)$$

$$V_3 = \frac{\sum_{i=1}^{n_s} n_i \left[ \left( t_3^i - \bar{t}_3 \right)^2 + \left[ t_4^i - \bar{t}_4 \right]^2 \right]^{1/2}}{\sum_{i=1}^{n_s} n_i} \qquad \ldots.(21)$$

$$H_i = \frac{\lfloor V_i - \mu_u \rfloor}{\sigma_u} \qquad \ldots.(22)$$

Where $\mu_u$ and $\sigma_u$ are the mean and standard deviation of artificially developed data using four-parameter Kappa distribution. $H_1$, $H_2$, and $H_3$ statistics estimate the degree of heterogeneity in a group of sites as reasonably homogeneous if $H_i < 1$, the region is fairly homogeneous if $1 \leq H_i \leq 2$ and if $H_i > 2$ the region is absolutely heterogeneous [3], [5], [6], [7].

## 4. Results and Discussion

### 4.1 Check for Adequacy

During the check for adequacy of rainfall stations, it was identified that when the number of stations is selected as nine (which was actual stations inside the basin), the standard error estimated using equations 1 and 2 becomes 8.83% ($\sigma_{m-1}$ obtained as 336.97 and $C_v$ obtained as 26.49). When the $\varepsilon_{ex}$ is 8.83 the optimum number of stations calculated is 17. Accordingly, a 10 km buffer zone from the boundary of the basin has been considered to account for more stations for the study. Also, another 2 stations closer to the 10 km buffer zone boundary have been selected to have a uniform distribution of the stations inside the study area and to increase the total number of stations to 17. Then again, the equations 1, 2, and 3 are used to estimate the optimal number of stations ($N$). Hence $\sigma_{m-1}$ was obtained as 52.88 and $\varepsilon_{ex}$ obtained as 4.16 with the standard error of 1.0%.

### 4.2 Outlier and Unconformity Check

Estimated high and low outlier values for 17 stations were plotted with the annual time-series data to detect the outliers. High and low outlier margins estimated for Mahailuppallama time-series data were graphically plotted with annual time-series data and shown in Figure 4. The graphs plotted with high and low outliers with the time-series data confirmed that there are no unconformities or considerable outliers detected for all the 17 stations annual time-series. Table 2 shows the high and low outliers estimated for 17 annual time-series data with maximum and minimum values.

### 4.3 Trend Analysis

Mann–Kendall test with tie correction has been applied to 17 stations annual rainfall time-series data to detect the monotonic trend. The test results are given in Table 3. The test results confirmed that 15 stations time-series data have no trend whereas two stations (Puttalam and Mahagalkadawala) time-series data have a trend. $Z_M$ statistics obtained for Puttalam and Mahagalkadawala are 2.88 and 2.37, respectively. Since these values are positive, two stations time-series data has an upward trend. But these values are closer to the 5% significance level of 1.96. If we increase the significance level to 0.5% then these stations also can consider as no trend.

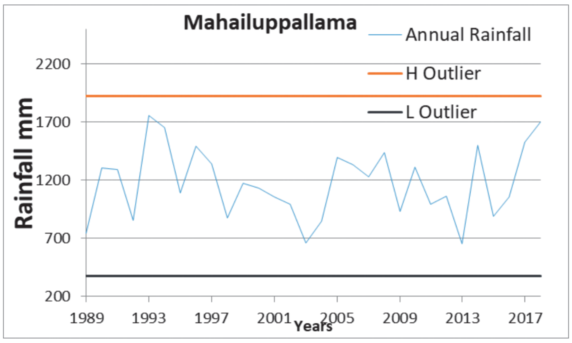

**Figure 4 – Mahailuppallama Station Time-series Data with High and Low Outliers**

### 4.4 Absence of Persistence Analysis

The first serial correlation coefficient for lag one was estimated to check the absence of the persistence of time-series data and the test results for 17 stations are presented in Table 1. The results confirmed that 15 stations estimated correlation coefficient for lag 1 lies within upper and lower confidence limits. But two stations (Millawana Lulkade and Mediyawa Tank) calculated $r_1$ statistics (0.319 and 0.269) do not lie within upper and lower limits of 0.245 and -0.275. But the increased values are not that significant.

**Table 1 – First Serial Correlation Coefficient Test Results**

| Station Name | $r_1$ | $r_{upper}$ | $r_{lower}$ | persistency Condition |
|---|---|---|---|---|
| Mahailuppallama | 0.14 | 0.24 | -0.27 | No Persistence |
| Kalawewa Tank | 0.03 | 0.24 | -0.27 | No Persistence |
| Mahagalkadawala | 0.01 | 0.24 | -0.27 | No Persistence |
| Siyabalangamuwa | 0.02 | 0.24 | -0.28 | No Persistence |
| Pelwehera | 0.05 | 0.24 | -0.27 | No Persistence |
| Dewahuwa | 0.21 | 0.24 | -0.27 | No Persistence |
| Kandalama | 0.17 | 0.24 | -0.27 | No Persistence |
| Nochchiyagama | 0.1 | 0.24 | -0.28 | No Persistence |
| Karaitivu | 0.07 | 0.24 | -0.28 | No Persistence |
| Karandipooval | -0.09 | 0.24 | -0.28 | No Persistence |
| Nachchaduwa | 0.07 | 0.24 | -0.28 | No Persistence |
| Tabbowa | 0.09 | 0.24 | -0.28 | No Persistence |
| Millawana-Lulcade | 0.32 | 0.24 | -0.28 | Persistence Exist |
| Millawana-Estate | 0.17 | 0.24 | -0.28 | No Persistence |
| Mediyawa Tank | 0.27 | 0.24 | -0.28 | Persistence Exist |
| Puttalam | 0.08 | 0.24 | -0.28 | No Persistence |
| Thabuththegama | 0.04 | 0.24 | -0.28 | No Persistence |

**Table 2 – Outlier Test Results**

| Station Name | Outlier (mm) | | Observed (mm) | |
|---|---|---|---|---|
| | High | Low | Max | Min |
| Mahailuppallama | 1926 | 373 | 1759 | 654 |
| Kalawewa Tank | 1916 | 640 | 1865 | 845 |
| Mahagalkadawala | 2094 | 448 | 2020 | 778 |
| Siyabalangamuwa | 1878 | 412 | 1675 | 621 |
| Pelwehera | 2310 | 776 | 2328 | 1074 |
| Dewahuwa | 2257 | 808 | 1981 | 875 |
| Kandalama | 1992 | 832 | 1834 | 1074 |
| Nochchiyagama | 1635 | 531 | 1567 | 659 |
| Karaitivu | 1628 | 456 | 1475 | 493 |
| Karandipooval | 1755 | 169 | 1529 | 108 |
| Nachchaduwa | 1758 | 285 | 1667 | 477 |
| Tabbowa | 1829 | 626 | 1733 | 736 |
| Millawana-Lulcade | 1936 | 446 | 1593 | 386 |
| Millawana-Estate | 2378 | 1242 | 2385 | 1349 |
| Mediyawa Tank | 2161 | 743 | 2021 | 927 |
| Puttalam | 1936 | 441 | 1822 | 718 |
| Thabuththegama | 1864 | 691 | 1717 | 891 |

### 3.5 Homogeneity Check

Homogeneity of the 17 time-series data have been checked by estimating the discordancy measure and heterogeneity measure.

**Table 3 – Mann-Kendall Test Result for Trend**

| Station Name | Mann-Kendall Statistics | | Observed Trend |
|---|---|---|---|
| | S | $Z_m$ | |
| Mahailuppallama | 69 | 1.04 | No Trend |
| Kalawewa Tank | 43 | 0.87 | No Trend |
| Mahagalkadawala | 119 | 2.37 | Trend Exist |
| Siyabalangamuwa | -46 | -0.89 | No Trend |
| Pelwehera | 48 | 0.97 | No Trend |
| Dewahuwa | 21 | 1.2 | No Trend |
| Kandalama | -9 | -0.72 | No Trend |
| Nochchiyagama | 58 | 1.46 | No Trend |
| Karaitivu | -5 | -0.11 | No Trend |
| Karandipooval | -12 | -1.36 | No Trend |
| Nachchaduwa | 33 | 0.96 | No Trend |
| Tabbowa | 8 | 0.41 | No Trend |
| Millawana-Lulcade | 31 | 1.75 | No Trend |
| Millawana-Estate | -16 | -0.37 | No Trend |
| Mediyawa Tank | -1 | 0 | No Trend |
| Puttalam | 101 | 2.88 | Trend Exist |
| Thabuththegama | 8 | 1.11 | No Trend |

The results obtained by doing the discordancy measure are presented in Table 4. The results confirmed that 12 stations time-series data are not discordant, but 5 stations (Pelwehera, Karaitiu, Karandipooval, Tabbowa, and Millawana Lulkade) time-series data as discordant.

The second statistic of homogeneity check is heterogeneity measure. The heterogeneity statistics of $H_1$, $H_2$ and $H_3$ have been estimated for the 17 rainfall gauging stations annual time-series using L-moments and the results are presented in Table 5. Since the heterogeneity statistic $H_1$, $H_2$ and $H_3$ values are less than 1, the region can be considered as reasonably homogeneous.

**Table 4 – Discordancy Measure Statistical Test Results**

| Station Name | Record length | $D_i$ | $D_i >= 3$ Criterion |
|---|---|---|---|
| Mahailuppallama | 34 | 2.419 | satisfy |
| Kalawewa Tank | 28 | 1.619 | satisfy |
| Mahagalkadawala | 28 | 1.733 | satisfy |
| Siyabalangamuwa | 28 | 1.089 | satisfy |
| Pelwehera | 28 | 3.834 | not satisfy |
| Dewahuwa | 14 | 0.042 | satisfy |
| Kandalama | 10 | 1.487 | satisfy |
| Nochchiyagama | 24 | 1.512 | satisfy |
| Karaitivu | 22 | 7.375 | not satisfy |
| Karandipooval | 8 | 9.584 | not satisfy |
| Nachchaduwa | 22 | 0.540 | satisfy |
| Tabbowa | 16 | 3.916 | not satisfy |
| Millawana-Lulcade | 14 | 6.491 | not satisfy |
| Millawana-Estate | 24 | 2.866 | satisfy |
| Mediyawa Tank | 18 | 0.396 | satisfy |
| Puttalam | 22 | 1.972 | satisfy |
| Thabuththegama | 8 | 0.804 | satisfy |

**Table 5 – Heterogeneity Measure Statistical Test Results**

| No of Stations | Heterogeneity statistics | | | Homogeneous condition |
|---|---|---|---|---|
| | $H_1$ | $H_3$ | $H_2$ | |
| 17 | -3.576 | -3.576 | -3.576 | Reasonably homogeneous |

## 5. Conclusions

All statistical check results confirmed that 17 stations annual time-series data are statistically homogeneous with minor deviations and can be used for the Kala Oya basin hydrological studies with satisfaction. The data screening method introduced by this paper can be used for the other river basins of Sri Lanka to check the hydrometeorological data consistency and homogeneity before using it for modelling and designing of water development schemes.

## Acknowledgement

Author wishes to thank Irrigation Department, Sri Lanka, for the support given, by providing necessary data for the research.

## References

1. Subramanya, K., *Engineering Hydrology,* 4th ed., McGraw Hill Education Private Limited, India, 30 p.

2. Grubbs, F. E., and Beck, G., "Extension of Sample Sizes and Percentage Point for Significance Tests of Outlying Observations", *J. Technometrics.*, Vol. 14, No. 04, 1972, pp. 847-854.

3. Hosking, J. R. M., and Wallis, J. R., "Some Statistics Useful in Regional Flood Frequency Analysis", *J. Water Resources Research.*, Vol. 29, No. 02, February, 1993, pp. 271-281.

4. Hamed, K. H., and Rao, A. R., "A Modified Mann-Kendall Trend Test for Autocorrelated Data", *J. Hydrology., Elsevier,* Vol. 204, September, 1997, pp. 182-196.

5. Hosking, J. R. M., and Wallis, J. R., "The Effect of Intersite Dependence on Regional Flood Frequency Analysis", *J. Water Resources Research.*, Vol. 24, No. 04, April, 1988, pp. 588-600.

6. Eslamian, S. S., and Feizi, H., "Maximum Monthly Rainfall Analysis Using L-Moments for an Arid Region in Isfahan Province, Iran", *J. Applied Meteorology and Climatology., American Meteorological Society,* Vol. 46, April, 2007, pp. 494-503.

7. Malekinezhad, H., and Zare-garizi, A., "Regional Frequency Analysis of Daily Rainfall Extremes Using L-moments Approach", *J. Faculty of Natural Resources., University Boulevard, Safayieh,Yazd, Islamic republic of Iran,* Vol. 27, No. 04, September, 2014, pp. 411-427.