ENGINEER - Vol. XXXXV, No. 02, pp. [1-12], 2012

Key Issues of Data and Data Checking for Hydrological Analyses - Case Study of Rainfall Data in the Attanagalu Oya Basin of Sri Lanka

N. T. S. Wijesekera and L. R. H. Perera

Abstract: Inconsistencies and non-homogeneities in the hydrological and meteorological time series could be identified by incorporating statistical tests that detect trends and change points. Inconsistency which reflects systematic errors during recording and the non homogeneity that arises from either natural or man made changes to the gauging environment are both important for adequate time series analysis. It has also been identified that statistical tests together with physical or historical evidence and justifications from metadata need to be incorporated for a very detailed study. A case study was carried out for the rainfall data of Attanagalu Oya basin in the western province of Sri Lanka with a data set consisting of six stations having daily rainfall data for 30 years. According to Pettitt test, a significant change around 1977 & 1985 at Karasnagala and Pasyala could be found. However Pasyala is the most significant station for the change of rainfall pattern, which was confirmed by t-test. Knowledge of Meta data was found very important in order to make necessary corrections to shifts identified through Double Mass Analysis. This paper shows that statistical tests and rational judgements would enable suitable corrections even though it is common to find that most of the hydrological and meteorological data are either flagged for quality or poorly documented.

Keywords: Issues, Data Checking, Hydrological Analyses, Rainfall, Sri Lanka

1. Introduction

Water resources development and management is heavily dependent on hydrological and meteorological data. In order to make sure that the results obtained from these data are reliable for practical applications, such data should be, homogeneous and consistent either to carryout frequency analyses or to simulate a hydrological system [1]. In hydrologic analysis it is customary to search for long datasets since such data ensures that the sample taken represents the system performance. However, longer the time series the greater are the chances that the data series is neither stationary, consistent nor homogeneous. It is also necessary to identify the spatial representation of the data used in an analysis. In case of precipitation, spatial distribution of rain gauges is often non-representative since they are mostly located in the valleys where easy access is the main criteria. It has also been identified that in many mountainous catchments, the higher elevations receive more precipitation than the regions in the valley [2]. As such, prior to a responsible hydrological analysis, a suitable spatial and temporal analysis of data needs to be carried out through an efficient screening procedure.

As there are many organizations having different objectives perform data collection, there is also a necessity to check such observation data series for consistency and homogeneity. It is common to use statistical tests, either parametric or non-parametric, in order to detect the non-homogeneity in a time series. The choice between the two families of tests is based on the expected distribution of data involved. If data set is normally distributed, parametric tests are usually selected. If data set is expected to be non-normally distributed, non-parametric tests are preferred. Also it has been identified that some homogeneity tests depend on meta-data while the others are purely statistical. The presence of a single significant test result is considered as a weak evidence of change. In case of more results that are significant and not very similar, then they need to be taken as stronger evidence of change [3].

However it should be emphasized that application of more than one test to data may

***Eng. (Prof.) N. T. S. Wijesekera**, C. Eng., FIE(Sri Lanka), B.Sc. Eng. (Sri Lanka), M. Eng. (Tokyo), D. Eng. (Tokyo), Senior Professor of Civil Engineering, Department of Civil Engineering, University of Moratuwa, Sri Lanka*

***Eng. L. R. H. Perera,** B.Sc (Eng) (Moratuwa), M.Sc (Delft), C. Eng, MIE (Sri Lanka), MIAHS (UK), Chief Irrigation Engineer, Irrigation Department, Colombo, Sri Lanka*

make interpretation of the results rather complex. Due to differences in assumptions pertaining to each test, along with possible influence with the change of catchment condition, it has been identified that it is usually difficult to compare the results of different tests. Since it is particularly difficult to combine the results of different tests, distribution free testing methods are recommended for hydrological data which are often non-normally distributed [3].

2. Study Area & Data Availability

Attanagalu Oya basin which drains to the western coast of Sri Lanka (between 79° 50' & 80° 7'E and 6° 59' & 7° 17' N) is having a catchment area of 727km^2. The spatial coverage of the basin shows that it spreads over two provinces namely the Western and Sabaragamuwa and flows through the Gampaha and Kegalle administrative districts. The basin has an elevation of about 300m MSL as its highest. There are several large streams that combine to drain Attanagalu-Oya and they are namely, Kimbulapitiya Oya, Mapalan Oya, Dee-eli Oya and Uruwal Oya (Figure 1).

There are 18 rainfall gauging stations located either within the basin boundary or just outside the boundary. The rain gauging network maintained by Department of Meteorology consists of 17 stations, of which 16 do not possess automatic recording facilities but maintain daily records. The other one at Katunayaka in the vicinity of the catchment is a recording type. There is a recording type rain gauge at Karasnagala, maintained by Irrigation Department.

Based on the data availability and spatial coverage, daily data of six stations were selected for the study. This study considered data from 1970 to 2001. Station names and details of missing rainfall data during the said period are shown in Table 1.

3. Methodology

The following tests were carried out in this study with the use of the SPELL-Stat software [4].

1) Visual examination of Data
2) Outlier Testing
3) Homogeneity Testing with,
    - Test for serial Correlation
    - Test for Pre-Whitening
    - Test for Normality
    - Spearman's rank correlation test
    - Standard Normal Homogeneity test (SNHT)
    - Change point test (Pettitt test)
    - Test for stability of variance (F-test)
    - Test for stability of mean (t-test)
    - Double Mass Analysis
    - Method of Cumulative Residuals (Ellipse test)

Pattern of observed time series data was analysed [5] in order to: (1) Identify the nature of the phenomenon represented by the sequence of observations, and (2) Predict future values of the time series variables.

At the inception, Data were plotted for visual examination in order to identify any abrupt changes in the time series. Testing of high and low outliers was done using the equation $y_H = \overline{y} + K_n s_y$ and $y_L = \overline{y} - K_n s_y$ where y_H, y_L are high and low outlier thresholds in log and $\overline{y}$ is the mean, n is the sample size, s_y is the standard deviation and K_n is the parameter given in Chow et al.(1988)[6], for sample sizes varying from 10 to 140.

The serial correlation coefficient verifies the independence of a time series which in turn helps to ensure that each of the data have an equal probability of occurrence. If a time series is completely random, the population auto-correlation function will be zero for all lags other than zero. If all the data sets are perfectly correlated to each other then its value is unity. Sample serial correlation coefficients will deviate slightly from zero only because of sampling effects. In case of hydrological analysis, it is usually sufficient to compute the first lag serial correlation coefficient, i.e. the correlation between adjacent observations in a time series [1]. A confidence level of 95% was used for calculations.

Presence of serial correlation may also complicate the detection and evaluation of trends in hydrological time series. When a data set shows a drift towards higher (or lower values) over the period of record, the drift may be an indication of an underlying change or long term persistence. It could probably be that the data are dependent on some processes which are serially correlated. Several approaches have been suggested for removing the serial correlation from a data set prior to

applying the non-parametric tests. One of the most common approaches is the Pre-Whitening of the time series. The Pre-Whitening approach involves in the calculation of serial correlation and the removal of correlation if the calculated serial correlation is found significant at a level of 5% [3].

It is important to make sure that there is no correlation with the order in which the data have been collected and with an increase or a decrease in the magnitude of those data. It is also important that the selected testing periods are of sufficient length for test to be reliable [1]. A study of rainfall trends in Sri Lanka [7], which chose both Mann-Kendall rank statistic and the Spearman rank statistic, concluded that both tests have similar power in detecting a trend. In the present work, Spearman's rank correlation method is used to verify the absence of trend at a significance level of 5%.

Standard Normal Homogeneity Test (SNHT) & Pettitt test were chosen to identify any sudden shifts in the mean of the data sets thereby enabling the identification of change points. A critical probability level of 80% was chosen for acceptance of significant change points in the Pettitt test whereas critical confidence level of 90% was used in the SNHT [3].

Instability of the variance was tested to identify the existence of a non-stationarity of the time series. Ratio of the variances of two split, non-overlapping, sub sets of time series was selected as the test statistic. The region for test statistic of F_t was taken as, $F\{v_1, v_2, 2.5\%\} < F_t < F\{v_1, v_2, 97.5\%\}$; where, $v_1 = n_1\text{-}1$ (the number of degrees of freedom for the numerator), $v_2 = n_2\text{-}1$ (the number of degrees of freedom for the denominator), and n_1, n_2 equals the number of data in each sub set [1].

The t-test for stability of the mean was conducted after carrying out the F-test using same two non overlapping time series subsets. The test statistic t_t [1] was taken to be bounded as, $t\{v, 2.5\%\} < t_t < F\{v, 97.5\%\}$ where, $v = (n_1\text{-}1) + (n_2\text{-}1)$ (the degrees of freedom) including n_1 and n_2 data in each sub set.

In order to identify the employability of the parametric test procedure, the time series was tested for normality by computing probability of exceedence based on the Blom equation [8]. Estimation of the data X_{est} with standard variates was used to determine the variability of the quantile with 95% confidence limit.

Homogeneity of the time series was inspected with the method of cumulative residuals. The estimated cumulative residuals and the ellipse that relate with the probability level were plotted against years to find whether the cumulative residuals fall within the ellipse [9]

Double mass analysis was performed using plots of cumulative values of a station under investigation against the cumulative values of the particular station or cumulative values of the average of other stations over the same period of time. To identify the Relative Consistency of time series, detection of non-homogeneities was performed by identifying inflection points in the double mass plot. In case of significant changes, the annual values of an earlier portion of the record were adjusted to be consistent with the latter portion [10].

4. Results and Discussion

The present work conducted for rainfall data sets of six stations indicated the variation of results from different statistical tests which are commonly used for hydrological data testing. Annual rainfall data were plotted in order to find the presence of any abrupt changes. During the considered period of 30 years, abrupt changes or any dubious data were not apparent for all six stations. Results of statistical tests pertaining to each station are shown in Table 2. Missing data were filled with the use of single & multiple regression analysis. Computed best fit coefficients of determination (Table 3) were considered for data filling. Generation of missing data was carried out relative to a common data period (Table 3) in which the data were assumed as homogeneous for the computations.

The co-efficient of determination with regression analyses is relatively good for the stations at Pasyala (0.91) and Vincit (0.88), whereas other stations showed to have relatively low values (Table 3). It was assumed that the period considered for regression (i.e. 01.05.1981 – 31.03.1982) is homogeneous. Selecting a homogeneous period is entirely dependant on the available metadata. In this study, the considered homogeneous period for regression is less than one year. Therefore, it was felt reasonable to assume that a minimum or no changes could occur to the station during the selected period. Based on these facts, the above assumption could be treated as realistic.

Minimum values of annual rainfall which were lower than low outlier were corrected with the low outlier. High outliers showed higher values in case of the maximum values of annual rainfall (Table 4). Tabular comparison of annual rainfall showed that the minimum values should be filled with the low outlier except for the minimum value of Pasyala. Annual rainfall comparisons with values shown in Figure 4, were used to identify any abrupt changes. From this data set, annual rainfall at Pasyala station which has more issues than other stations, and with t-test confirmed change points, is selected to discuss the issues related to rainfall data. Statistical analysis & homogeneity test results of Pasyala annual rainfall are shown in Figures 2 (a-f) and 3 (a-f). In these Figures, graphs before Double Mass analysis correction are shown by letters a, c and e, whereas the letters b, d and f, show the results after the Double Mass Analysis.

The presence of significant changes around 1977 & 1985 at both Karasnagala and Pasyala could be identified (Table 2) from the Pettitt test. If there are no Meta data then it is difficult to conclude whether the changes around 1977 &1985 are due to a situation as a result of climatic change, or due to some other natural or man made changes to the environment during the period of record or systematic errors associated with the recording of the data for Attanagalu Oya Basin. Even though the t-test results confirm the presence of a change in Pasyala around 1977 and 1985, the non-availability of Meta data prevented from incorporating Double Mass corrections with sufficient confidence.

In order to identify the possibility of data use, an alternative option was considered. Since there is no strong evidence that the mean state of rainfall in Sri Lanka has changed significantly over the past decades, it was assumed that the effect of climate change had not significantly affected the rainfall of Attanagalu Oya. Accordingly the application of Double Mass curve for correction of change points was considered realistic and the same was utilized for Pasyala in order to correct the change which was present at 1985. Reduction of the trend could be observed after Double Mass correction. The change at Pasyala in 1977 was insignificant after carrying out Double Mass correction for 1985 (Table 2, Figure 2).

Homogeneity test shows that the annual rainfall at Henerathgoda, Karasnagala and Vincit is homogeneous at 85% non-exceedence probability level whereas for Halgahapitiya, Katunayake and Pasyala it is at 90% non-exceedence probability level. Homogeneity test results, before & after Double Mass corrections for Pasyala are shown in Figure 3. It could be observed that Pasyala rainfall data set is homogeneous at 90% non-exceedence probability level even after the Double Mass correction. Homogeneity test showed that the acceptable probability level of Pasyala data set is 90% since all residuals were found to be within the 90% probability ellipse after Double Mass correction (Figure 3). As the t-test results did not confirm the results of the Pettitt test, rest of the stations were not subjected to correction.

Correlogram shows that 1st lag serial correlation for all datasets had fallen within the 95% confidence limit. Therefore, all these annual rainfall time series are with a satisfactory level of randomness and independence. As a result, pre-whitening of annual rainfall time series was not necessary prior to performing statistical tests. In order to select the need of parametric or non-parametric testing, normality testing was carried out and it was identified that all stations follow the normal distribution pattern except Katunayake which exceeds the 95% confidence limits.

It can be observed that a decreasing rainfall pattern is prevailing in Attanagalu Oya basin. Also the change of rainfall pattern around 1977 & 1985 is common for some stations. Mass curve shows that the change in the slope is not significant. Therefore, it suggests that the data from each station are satisfying consistency. As such, likely reasons for the changes around above years are mainly due to man made changes to the environment and most probably due to change of instruments.

Some of the homogeneity tests depend on meta-data while the other tests are purely statistical. The presence of a single significant test result could be identified as weak evidence of change. If more tests not similar to one another lead to significant test results, then it provides stronger evidence of change. Carrying out similar tests which would provide multiple-significance is not an extra proof of change. However, application of more than one test to the data may make interpretation of results complex. Since the differences in assumptions of the tests and the possible influence of change in the catchment condition, it is usually difficult

to compare and in particular to combine the results of different tests.

Meta data plays a major role when making firm conclusions with regards to data checking. In Sri Lanka most gauging stations are maintained without proper documentation of meta data and it is common knowledge that the most of the hydrological and meteorological data are poorly documented and quality flagged. This is a big challenge faced by hydrologists when attempts are taken to analyze rainfall data. It is known that for situations where no meta data are available, hydrologists need to consider regional and global changes to rainfall during that period. However it is difficult to address micro climatic changes without meta data. In many stations of Sri Lanka it is not a difficult task to obtain a 30 year long rainfall dataset. These data are bound to be with missing data periods, non-homogeneity and other inconsistencies. Hydrologists need to identify the purpose of data and perform checks to ensure reliability of results produced with such data. The present study presents an attempt taken to identify measures that can be taken when data checking is carried out in case of a Sri Lankan situation.

5. Conclusions

1. The present work using daily rainfall data identified the variation of results from different statistical tests indicating the necessity to compare and rationalize the outputs prior to practical use.
2. In the Attanagalu Oya basin, the co-efficient of determination from regression analyses showed relatively good values for the Pasyala and Vincit stations with values of 0.91 and 0.88 respectively.
3. A decreasing rainfall pattern prevails in Attanagalu Oya Basin and a significant change in rainfall pattern could be identified around 1977 & 1985 for some stations.
4. Tests conducted confirm that rainfall data of Pasyala station has an inhomogeneity, and rectification could be carried out to achieve a 90% confidence level.
5. A significant change in Pasyala was identified around 1977 & 1985 by the Pettitt test and confirmed by t tests whereas the changes at other stations were identified only by one test or none. Pasyala was identified as the most significant station out of those used for the study.
6. Testing clearly identified the need of supporting tests for confirmation of indications made by a particular test, while raising the issue of testing carried out with the use of several tests.
7. Data checking enabled identification of confidence limits for data use thereby providing the most important information to assess the validity of using the resulting hydrologic outputs for reliable conclusions.

References

1. Dahmen, E.R. and Hall, M.J., 1990. Screening of Hydrological Data, Tests for Stationarity and Relative Consistency, ILRL, The Netherlands.
2. Uhlenbrook, S., 2006.Catchment Hydrology with Satellites, Models and Rubber boots. UNESCO-IHE, Delft, The Netherlands.
3. Tu, M., 2006. Assessment of the Effect of Climate Variability and Land use Change on the Hydrology of the Meuse River Basin. A.A Balkema Publishers, The Netherlands
4. Guzman, J.A. and Chu, M.L., 2003. SPELL-Stat Statistical Analysis Program. Universidad Industrial de Santander, Colombia.
5. Ma, L.C.A, 2003. Assessment of the Long Term Rainfall Runoff Relation of the Geul Catchment. M.Sc. Thesis (HH 444), UNESCO-IHE, Delft, The Netherlands (Unpublished).
6. Chow, V.T., Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. McGraw Hill book Company, Singapore.
7. Jayawardene H.K.W.I., Sonnadara D.U.J., and Jayewardene D.R., 2005 Trends of Rainfall in Sri Lanka over the Last Century, Sri Lankan Journal of Physics, Vol.6 (2005) 7-17.
8. Cunnane, C., 1978. Unbiased plotting position – a review journal of Hydrology 37:205-222
9. Perera, L.R.H., 2007. Detecting the Impacts of Climate Variability on Meteorological Parameters and Evaporation**.** M.Sc. Thesis (WSE - HWR - 07.06), UNESCO-IHE, Delft, The Netherlands (Unpublished).
10. Dingman, S.L., 2002. Physical Hydrology, 2nd edition, Prentice Hall, New Jersey, U.S.A.

Acknowledgement

This research was supported by University of Moratuwa Senate Research Grant Number 202. Encouragement given by the University of Moratuwa and the Senate Research Committee is gratefully acknowledged.

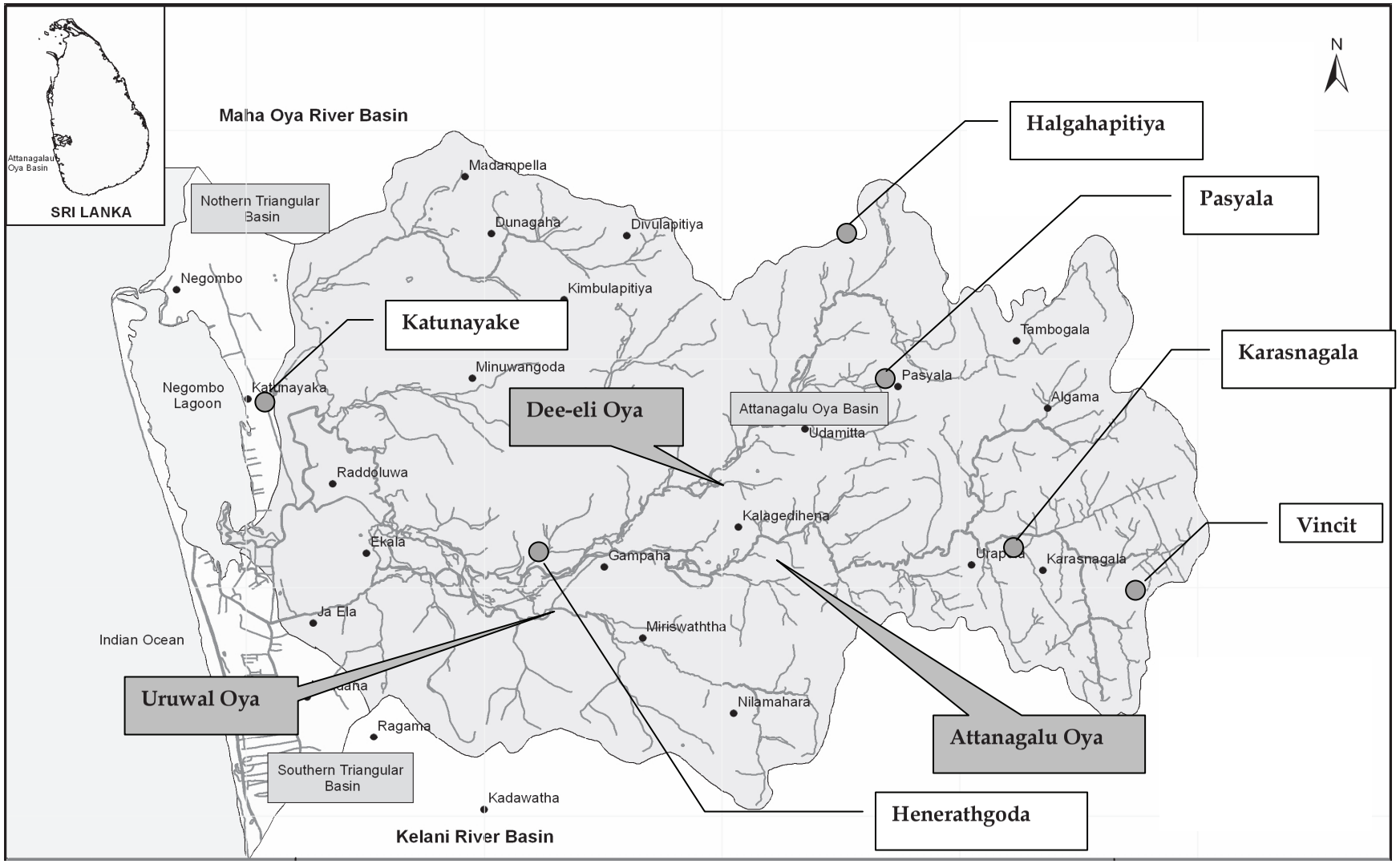

Figure 1 - Study Area, Stream Network and Rain Gauging Stations

Statistical Analyses Results at Pasyala (Correction in 1985)

Before Double Mass Corrections (1970 – 2001)	After Double Mass Corrections (1970 – 2001)
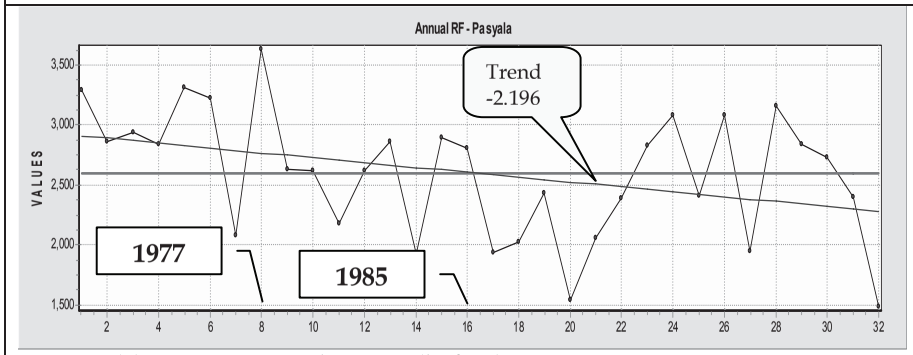	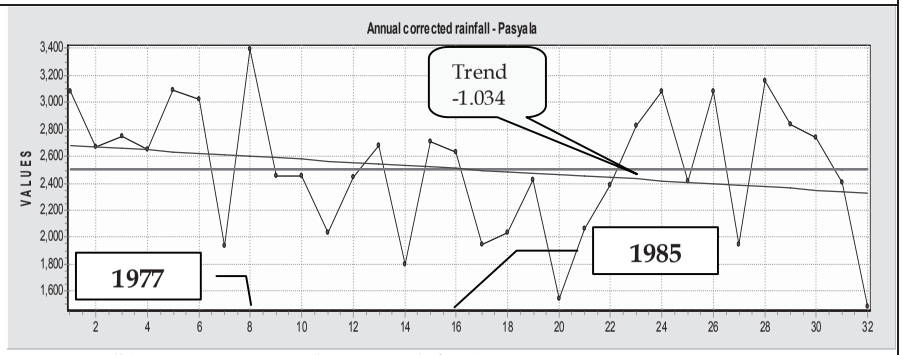
Figure 2(a) Time Series Behaviour (before)	Figure 2 (b) Time Series Behaviour (after)
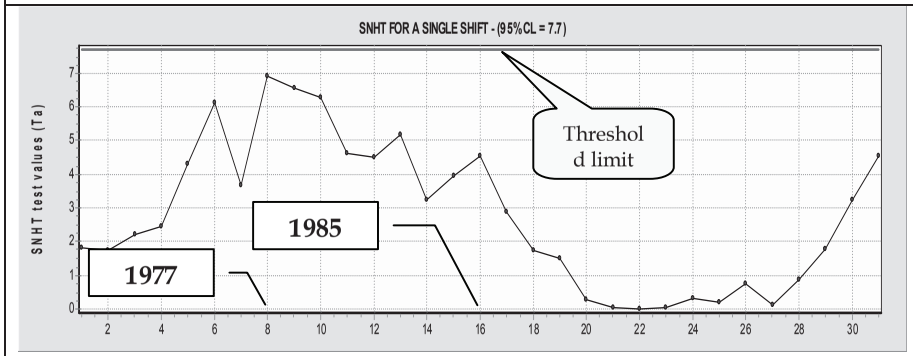	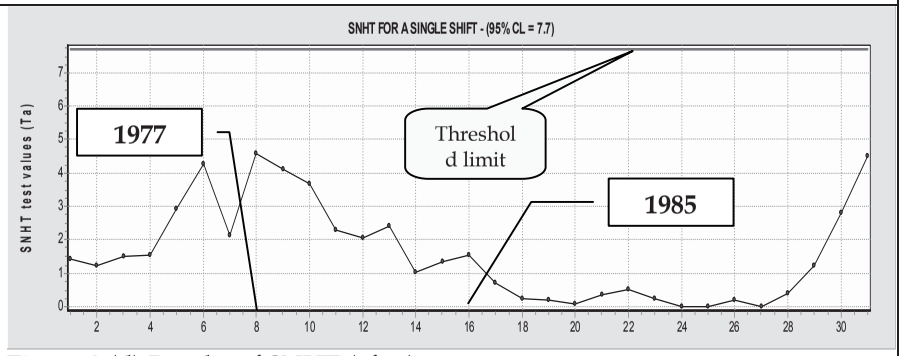
Figure 2 (c) Results of SNHT (before)	Figure 2 (d) Results of SNHT (after)

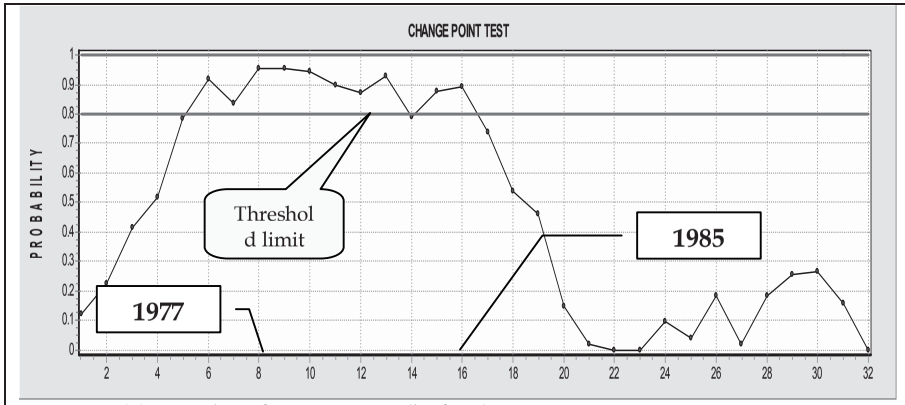

Figure 2 (e) Results of Pettitt Test (before)

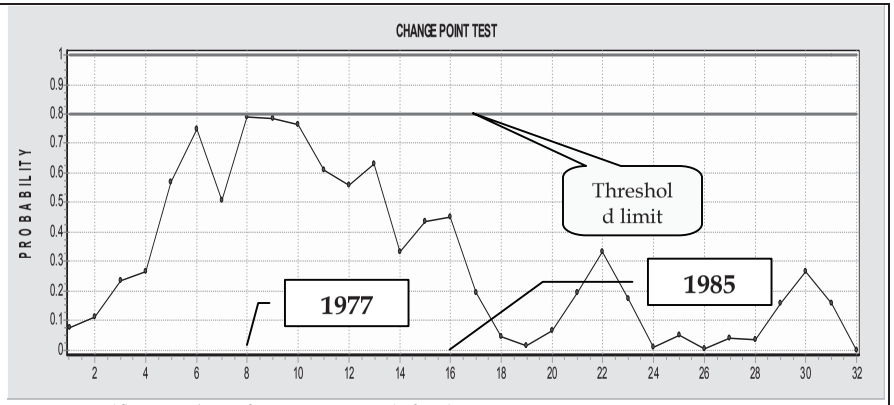

Figure 2 (f) Results of Pettitt Test (after)

Figure 2 - Statistical Analyses Results at Pasyala (Correction in 1985)

Before Double Mass corrections (1970 – 2001)	After Double Mass corrections (1970 – 2001)
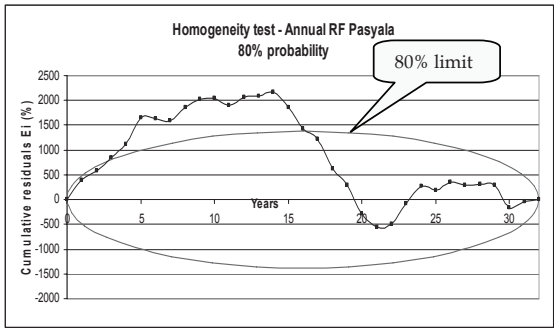	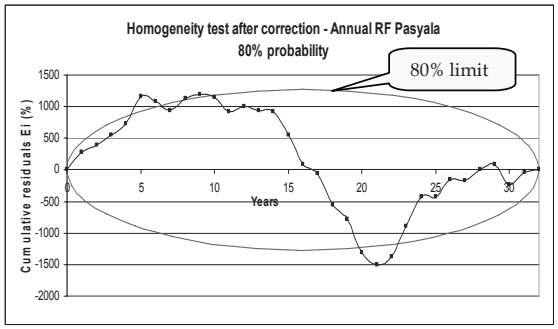
Figure 3(a) Variation of Residuals with 80% limit	Figure 3(b) Variation of Residuals with 80% limit
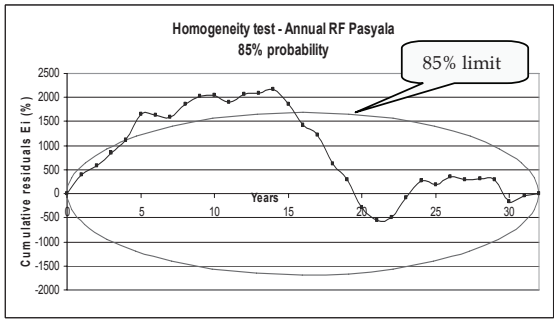	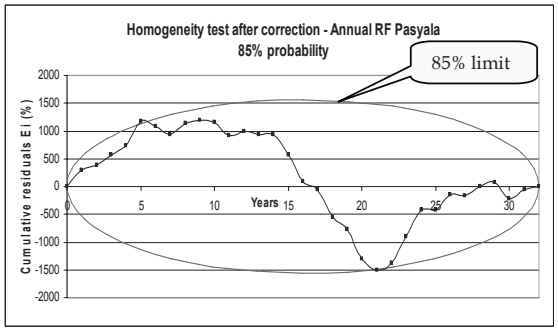
Figure 3(c) Variation of Residuals with 85% limit	Figure 3 (d) Variation of Residuals with 85% limit
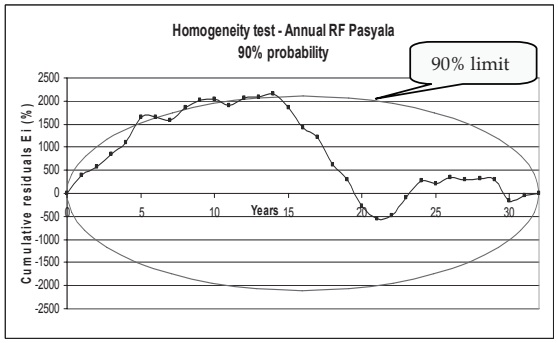	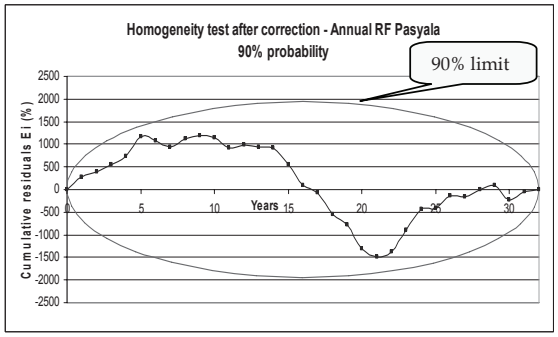
Figure 3 (e) Variation of Residuals with 90% limit	Figure 3 (f) Variation of Residuals with 90% limit

Figure 3 - Homogeneity Test for Pasyala (Correction in 1985)

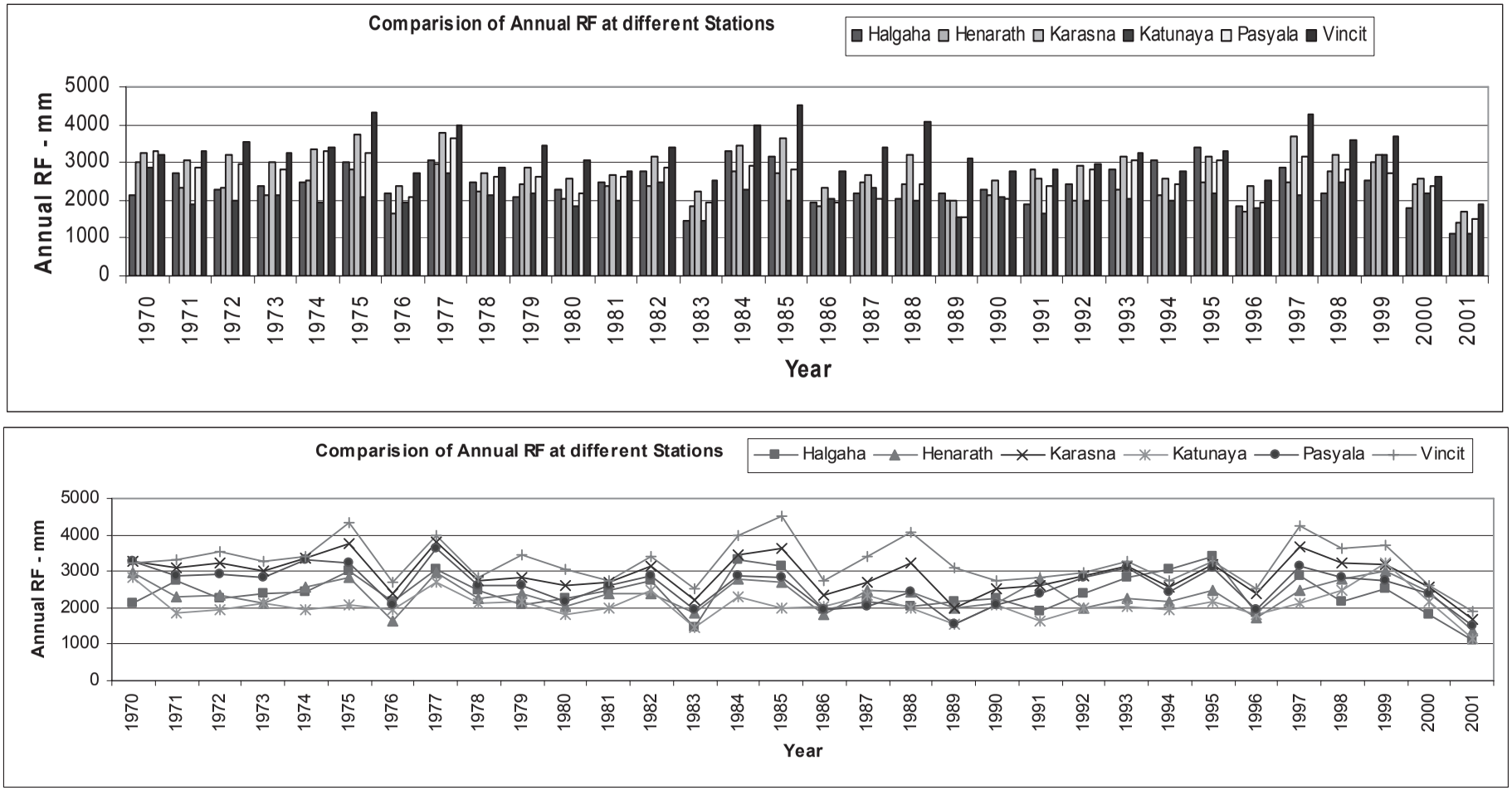

Figure 4 - Annual Rainfall Comparison at Selected Stations

Table 1 - Rain Gauging Stations & Missing Data

Station Name	Missing Duration	No. of Missing Days
Halgahapitiya	1 Sep 83 – 30 Sep 83	30
	1 Jul 88 – 31 March 89	274
Karasnagala	None	None
Henarathgoda	1 Oct 77 - 31 Dec 77	92
	1 Jun 80 – 30 Jun 80	30
	1 Jun 82 – 30 Jun 82	30
	1 Dec 82 – 31 Dec 82	31
	1 Sep 83 – 30 Sep 83	30
Katunayaka	7 Apr 87 – 20 Apr 87	14
Pasyala	1 Oct 73 – 31 Dec 73	92
	1 Sep 79 – 30 Sep 79	30
	1 Apr 89 – 30 Apr 89	30
	1 Jun 89 – 31 Jul 89	61
Vincit	1 Apr 81 – 30 Apr 81	30
	1 Apr 82 – 30 Apr 82	30

Table 2 - Summary of the Statistical Analysis for Annual Rainfall

Station & Duration	Year of change (Pettitt test)	Pettitt test probability (80%)	Spearman linear trend t-value (95%)	F-test F-value (95%)	t-test t-value (95%)	SNHT T_0 (95%)
Halgahapitiya 1970-2001	1985	0.6605	-0.891	1.301	1.232	1.492
Henerathgoda 1970-2001	1975	0.3132	-0.437	1.504	1.054	1.388
Karasnagala 1970-2001	**1977** **1985**	**0.8053** **0.7917**	-1.616	1.105 1.172	1.734 1.847	4.285
Katunayake 1970-2001	1996	0.3471	-0.03	-*	-*	-
Pasyala 1970-2001	**1977** **1985**	**0.8053** **0.8922**	**-2.196**	1.484 1.32	**2.533** **2.267**	6.923
Pasyala corrected for 1985 1970-2001	1977	0.7917	-1.034	1.676	**2.523**	4.575
Vincit 1970-2001	1988	0.7057	-1.19	1.151	1.243	2.713

Values in parenthesis are thresholds (Confidence Level) for each test.

- Not enough data for the split record test*

Table 3 - Summary of the Regression Analyses -Best Fit Coefficient of Determination for Missing Data Estimation

Data Missing Station	RF Stations Considered for Regression	Period Considered for Regression	Coefficient of Determination
Halgahapitiya	Karasnagala, Katunayake, Pasyala	01.05.1981 – 31.03.1982	0.62
Henerathgoda	Karasnagala, Pasyala, Katunayake	01.05.1981 – 31.03.1982	0.49
Katunayake	Halgahapitiya, Karasnagala, Pasyala, Henerathgoda	01.05.1981 – 31.03.1982	0.62
Pasyala	Karasnagala	01.05.1981 – 31.03.1982	0.91
Vincit	Karasnagala	01.05.1981 – 31.03.1982	0.88

Table 4 - Tabular Comparison of Annual RF Prior to Correction for Outliers

RF Station	Halgahapitiya	Henerathgoda	Karasnagala	Katunayake	Pasyala	Vincit
Mean	2400.1	2342.4	2908.5	2078.1	2596.1	3251.6
Max	3412.2	3020.3	3795.0	3223.7	3632.5	4494.1
Min	**1122.6**	**1427.7**	**1686.8**	**1130.8**	1486.5	**1902.1**
High Outlier	4329.6	3719.3	4655.3	3361.3	4482.8	5199.9
Low Outlier	1263.9	1427.7	1758.7	1240.8	1438.5	1966.3
Skewness	-0.91	-0.75	-0.76	-0.53	-0.76	-0.32

Minimum values shown in bold font were corrected with low outlier values