Statistical And Trend Analysis Of Rainfall And River Discharge: Yala River Basin, Kenya

Githui F. W.*[+], A. Opere* and W. Bauwens[+]

*Department of Meteorology, University of Nairobi, P O Box 30197 Nairobi, Kenya.
fgithui@vub.ac.be, aopere@uonbi.ac.ke

[+]Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussels, 1050 Brussels. wbauwens@vub.ac.be

Abstract

Yala River is one of the several rivers that drain into Lake Victoria in East Africa. Lake Victoria is the World's second largest freshwater lake and is the source of the Nile River. The Yala River catchment is centered about 35^{o}E, 0.1^{o}N. Rainfall and flow data for the period 1963-1998 were used for this study. The time series of monthly and annual values of rainfall and discharge were analyzed using statistical methods. Trend analysis was performed and the best-fitted models were determined using the statistical criteria of Mean Error (ME), Mean Square Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percent Error (MAPE). This was done in an attempt to determine whether or not there have been any significant changes in rainfall and discharge over this catchment. Probability distributions were fitted and the ?2, Anderson-Darling and the Kolmogorov-Smirnov tests were used to select the theoretical distribution, which best fitted the data. The analogue year's plots were also carried out in order to distinguish the years with near normal, above normal and below normal conditions, using the long term mean of the variables. While rainfall data analysis for the stations analyzed shows on average a decreasing trend, river discharge shows a decreasing slope in the upstream station and an increasing trend in the downstream station.

Introduction

Hydrologic time series almost always exhibit seasonality due to the periodicity of the weather. In Kenya, this arises greatly from seasonal variations in precipitation volume, as well as the rate of evapotranspiration. For observed data that exhibit high seasonality, methods to analyze trends should be those that incorporate the seasonal component. Spatial differences in trends can occur as a result of spatial differences in the changes in rainfall and temperature and spatial differences in the catchment characteristics that translate meteorological inputs into hydrological response (Burn and Elnur, 2002). Here we look at river discharge and rainfall in the Yala basin. Two river gauging stations, 1FG01 (upstream) and 1FG02 (downstream), and 10 rainfall stations (Kakamega Forest Station, Kaimosi Tea Estates Ltd., Esirwa Kaimosi Farms, Kaimosi Settlement, Nabkoi Forest Station, Eldoret

Kibabet Estates Ltd., Nandi Hills Kibwari Tea Estate, Maseno Veterinary Station, Maseno Siriba GTC and Bondo Water Supply) were selected.

This study seeks to determine if their values generally increase or decrease. In statistical terms this is a determination of whether or not the probability distribution from which they arise has changed with time. We also quantify the amount or rate of change, in terms of changes in the median as a central value. The Seasonal Kendall method, a non-parametric test, is used in this study for trend analysis because there are very few underlying assumptions about the structure of the data making them robust against departures from normality (Helsel & Hirsch, 1991). In addition, the use of ranks rather than actual values makes them insensitive to outliers and missing values. Hirsch, et. al, (1982) suggest that the Seasonal Kendall, a non-parametric test, is preferred to the simple or seasonal regression tests when data are skewed, cyclic, and serially correlated.

The Yala River is one of the rivers that drain into Lake Victoria, the second largest freshwater lake in the world. It has a basin area of 3280 km^2. This paper examines (a) the time series of monthly values of discharge and rainfall at the selected stations, (b) the existence of trends and the evaluation of the best fitted trend models, and (c) their probability distributions.

Methodology

The variables of study chosen are river discharge and rainfall. River discharge is known to reflect an integrated response of the entire river basin while rainfall serves as one of the major input into the runoff processes. All the stations that were chosen had a long record of data (>30yrs) for validity of the time series and trend analysis results. Missing data in the data sets were filled using linear interpolation, the Maintenance of Variance Extension Type1 (MOVE.1) and the Ordinary Least Squares (OLS) methods. The methods of analysis include time series analysis, fitting of trends, fitting probability distributions and analogue plots and they are described in the next section.

The Seasonal Kendall Test

The Seasonal Kendall test (Hirsch et al., 1982) accounts for seasonality by computing the Mann-Kendall test on each of m seasons (m represents months) separately, and then combining the results. This means that January data are compared only with January, February only with February, etc. No comparisons are made across season boundaries. Kendall's S statistic S_i for each season, are summed to form the overall statistic S_k.

$$S_k = \sum_{i=1}^{m} S_i \tag{1}$$

Kendall's S is calculated by subtracting the number of (Y,T) pairs (M), where Y (the variable, in this case rainfall and river discharge) decreases as T (time) increases, from the number of (Y,T) pairs (P), where Y increases with increasing T.

$$S = P - M \tag{2}$$

where P = the number of times the Y's increase as the T's increase, or the number of $Yi < Yj$ for all $i < j$, M = the number of times the Y's decrease as the T's increase, or the number of Yi > Yj for i < j .

for all i = 1,....(n - 1) and j = (i+1),.....n. (n is the number of data pairs)
When the product of number of seasons and number of years is more than about 25, (Helsel & Hirsch, 1991), the distribution of S_k can be approximated quite well by a normal distribution with expectation equal to the sum of the expectations (zero) of the individual S_i under the null hypothesis, and variance equal to the sum of their variances. S_k is standardized using (eq. 3) and the result is evaluated against a table of the standard normal distribution.

$$Z_{Sk} = \begin{cases} \frac{S_k - 1}{s_{Sk}} & if\ S_k > 0 \\ 0 & if\ S_k > 0 \\ \frac{S_k + 1}{s_{Sk}} & if\ S_k < 0 \end{cases} \qquad (3)$$

where $m_{Sk} = 0$

$$s_{Sk} = \sqrt{\sum_{i=1}^{m} (n_i/18)(n_i - 1)(2n_i + 5)} \qquad (4)$$

n_i = number of data in the i[th] season.

The null hypothesis is rejected at significance level a if $|Z_{Sk}|$ > Zcrit where Zcrit is the value of the standard normal distribution with a probability of exceedance of a /2.

An estimate of the trend slope b for Y over time T is computed as the median of all slopes between data pairs within the same season, where b is given by

$$b = Median\left[\frac{Y_j - Y_i}{T_j - T_i}\right] \qquad \text{for all } i<j \qquad (5)$$

Therefore no cross-season slopes contribute to the overall estimate of the Seasonal Kendall trend slope. All possible slopes within each season are calculated with the median slope being the Seasonal Kendall slope.

The approach described above assumes a single pattern of trend across all seasons. This can fail to reveal differences in behavior between different seasons. The Y variable may exhibit a strong trend in one season while showing no trend in the other seasons. It is possible that trends in different seasons may cancel each other out, resulting in an overall seasonal Kendall test statistic stating no trend. It is thus useful to perform analysis on each season.

For each season i (i=1,2,...m), Z_i is computed as $Z_i = S_i / Var(S_i)$. These are summed to compute the "total" chi-square statistic. The "trend" and "homogeneous" chi-squares are then computed as.

$$\chi^2_{total} = \sum_{i=1}^{m} Z_i^2 \quad (6)$$

$$\chi^2_{trend} = m * \overline{Z}^2 \quad (7)$$

where $$\overline{Z} = \frac{\sum_{i=1}^{m} Z_i}{m} \quad (8)$$

$$\chi^2_{homogeneous} = \chi^2_{total} - \chi^2_{trend} \quad (9)$$

The ?2 (homogeneous) values are compared to tables of the chi-square distribution with m-1 degrees of freedom to determine whether or not the seasons are homogeneous with respect to trend. If it exceeds the critical value for the pre-selected α, the null hypothesis that the seasons are homogeneous with respect to trend is rejected.

Fitting trend lines

The trend analysis applied here was an attempt to fit the linear, quadratic and exponential models to the time series of river dis charge and rainfall.

$$Y = a + b * T \quad (10)$$

$$Y = EXP\left(a + b/T\right) \quad (11)$$

$$Y = a + b * Ln(T) \quad (12)$$

$$Y = a + b * T + c * T^2 \quad (13)$$

The best-fitted model of time trend for each variable was chosen based on the Mean Error (ME), Mean Square Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percent Error (MAPE), (Makridakis et al., 1986). The MSE is the average squared difference between the regression estimates and observation pairs. It directly reflects the accuracy of the resulting forecasts since it indicates the variability of the observed values around the regression line (Wilks, 1995). This is similar to MAE except that the squaring function is used rather than the absolute value function. Since the MSE is computed by squaring forecast errors, it is more sensitive to larger errors than MAE.

The correlation coefficient reflects linear association between two variables but it does not account for biases that may be present in the forecasts. The ME is simply the difference between the average forecast and average observation and therefore expresses the bias of the forecasts. Forecasts that are on average too high will exhibit ME>0, and forecasts that are on average too low will exhibit ME<0. The bias however does not give information about the typical magnitude of individual forecast errors and is therefore not an accuracy measure. Forecasts that are more highly correlated with the observations will exhibit lower MSE, other factors being equal. The closer the average ME is to zero and the smaller the values of MSE, MAE and MAPE, the better the model fits the time-series.

The MAPE is commonly used because it produces a measure of relative overall fit. The absolute values of all the percentage errors are summed up and the average is computed. In comparison to ME which is determined simply as the average error value and affected by outliers (large positive and negative errors can cancel each other out resulting in a zero error), or MAE, which de-emphasizes outliers by their average, the MAPE is a more meaningful measurement. The MAPE also de-emphasizes outliers, and calculates the average absolute error in percentage terms.

Analogue plots

The analogue year's plots were drawn by getting the cumulative long term annual means (LTM) and then comparing these with the cumulative annual means. This was done in order to distinguish the years with near normal, above normal and below normal conditions. Values lying near the LTM were considered to have near normal flows (NN), those lying above as above normal flows (AN) and those lying below as below normal flows (BN).

Probability Distributions

Nine probability distribution models were applied to the available data. These are: the Normal distribution, the two and three-parameter Log-Normal distribution, the two and three-parameter Gamma distribution, Beta distribution, Weibull, Pearson V and the Log Pearson Type III distribution. The ?2, the Kolmogorov-Smirnov (K-S) (Shahin and Lange, 1993) and the Anderson-Darling (A-D) tests were used to test the goodness of fit of the distributions to the rainfall and stream flow data. The A-D test for goodness-of-fit is designed to detect discrepancies in the tails of distributions. It is more powerful than the K-S test against many alternative distributions. Unlike the K-S statistic, which focuses in the middle of the distribution, the A-D statistic highlights differences between the tails of the fitted distribution and input data. The K-S test tends to be more powerful than chi-square tests against many alternative distributions. A weakness of the ?2 statistic is that there are no clear guidelines for selecting the number and location of the bins (the number of intervals into which the range of data is divided). In some situations, one can reach different conclusions from the same data depending on how the bins were specified.

Results and Discussion

River Discharge at Stations 1FG01 and 1FG02

Trend analysis

The flow data was analyzed for the period 1963-1998 and was obtained from the Ministry of Water, Kenya. Summary statistics which are the mean, standard deviation (StDev), minimum, maximum, median, skewness, kurtosis and coefficient of variation (CV) are given in Table 1 below. Results of the trend analysis performed on the annual total river flows are given in Table 2 The probability values (p-value) obtained against the null hypothesis $H_{0,}$ of no trend in the trend analysis, have shown that there is no significant trend neither at 0.1, 0.01 nor 0.05 significant levels. The slope estimator shows a decrease of 51.20 (m^3/s)/year for the upstream station 1FG01 and an increase of 34.60 (m^3/s)/year for the downstream station 1FG02. Linear regression indicates slopes of -26.59 and 32.27 (m^3/s)/year for 1FG01 and 1FG02 respectively. This was done to compare the magnitudes of the slope obtained using the two methods and it can be seen that there is a marked difference in the slope magnitudes obtained for 1FG01 using the two methods.

Table (1): Annual mean flow statistics (Units of flow: (m^3/s))

	Mean	StDev	Minimum	Median	Maximum	Skewness	Kurtosis	CV
1FG01	872.9	234.3	460.5	888.7	1334.8	0.212	-0.417	27%

1FG02	992.2	292.4	569.0	897.1	1826.6	0.587	0.173	29%

Table (2): Annual trend statistics for flow at stations 1FG01 and 1FG02

	1FG01	1FG02
Kendall Slope	-51.20	34.60
Z- test statistic, Z_{Sk}	-0.62	0.65
p-value (for Kendall slope)	0.53	0.51
Linear regression slope	-26.59	32.27
p-value (linear regression slope)	0.58	0.59
Is there trend? (Significance level, 5%)	No	No

The residuals shown in Fig, 1 are those calculated using the Kendall method.

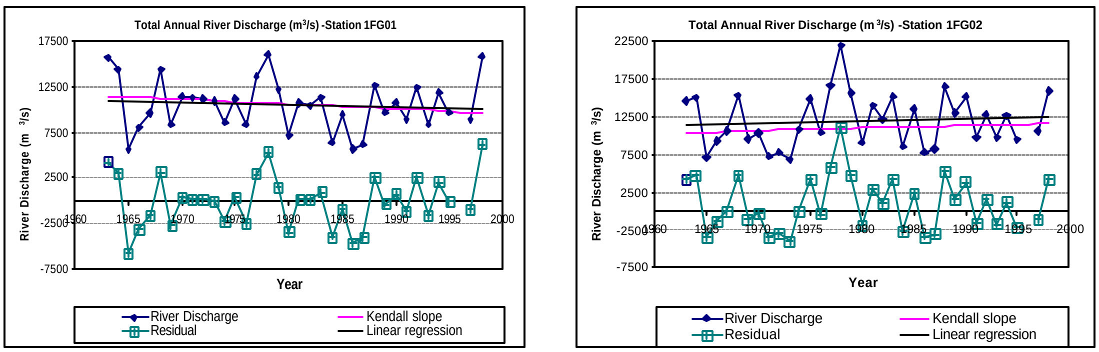


Figure 1 (a) and (b): Total Annual River Discharge showing Kendall slope and Linear regression slopes for (a) Station 1FG01 and (b) Station 1FG02

The means of the residuals are not significantly different from zero at 5% significance level, showing that they are from a random distribution. This implies that a linear model may be applied to the annual total flow, although there is large discrepancy between the slope values obtained using the Kendall method and the linear regression for station 1FG01.

The results from the Seasonal Kendall trend analysis performed on the monthly totals are shown in Table 3 below. Station 1FG02 shows a significant trend at 10% significance level with a slope of 3.32(m^3/s)/month which translates to a change in annual total flow of 39.85 (m^3/s)/year.

Table (3): Monthly Seasonal Kendall trend statistics for stations 1FG01 and 1FG02

	1FG01	1FG02
Kendall Slope	-1.37	3.32
Z- test statistic, Z_{Sk}	-0.88	1.66
p-value (for Kendall slope)	0.38	0.10
Is there trend? (Significance level, 10%)	No	Yes

The station 1FG01 shows a decreasing trend of 1.37 (m^3/s)/month equivalent to a change in annual total flow of -16.40 (m^3/s)/year. Again we see the large difference between the slopes obtained using the annual totals and monthly totals for the series 1FG01.

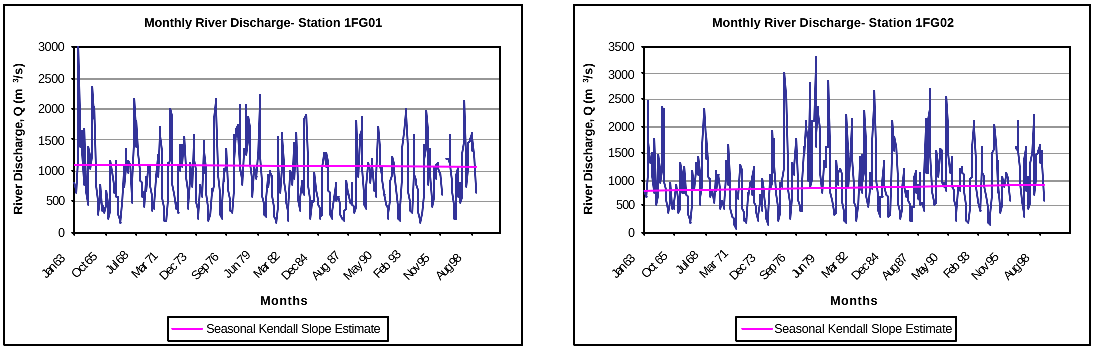


Figure 2(a) and (b): (a) Monthly river discharge for station 1FG01 (b) Monthly river discharge for station 1FG02

The Locally Weighted Scatterplot Smoothing technique, LOWESS (Helsel & Hirsch, 1991) was used to compute residuals. LOWESS is a smoothing technique that describes the relationship between two variables without assuming linearity or normality of residuals. It describes the data pattern whose form changes as the smoothing coefficient is altered. The LOWESS pattern is chosen in a way that produces less local minima and maxima, but not so smooth as to eliminate true changes in slope.

The Seasonal Kendall test was applied to the residuals that resulted from the fitted LOWESS curve. The S_k statistic was tested to see if it was significantly different from zero. The test for S_k is the test for trend. Results show that station 1FG01 does not show any significant trend (p-value = 0.32) whereas 1FG02 shows a significant increasing trend (p-value = 0.10) at the 0.1 significance level. This could be due to factors such as changes in land cover and land use among others. With LOWESS smoothing, it was shown that 1FG01 does not show monotonic trend. Thus the data was divided into two proportions. The first proportion showed a significant decreasing trend while the second proportion showed a positive slope that was not significant.

Results of month-by-month analysis using the Mann Kendall method shown in Table 4 reveal a significant trend at 10% significance level for station 1FG02, for the month of June (Table 4). The month of August for station 1FG01 is marginally significant. For 1FG01, seven months exhibit negative slopes while for 1FG02, nine months exhibit positive slopes.

Table 4(a) and (b): Kendall Slope (S), Z-statistic and p-value for (a) Station 1FG01 and (b) 1FG02

(a)

	Kendall slope	Z_{Sk}	p-value
Jan	-2.40	-0.45	0.65
Feb	-1.79	-0.80	0.43
Mar	-2.25	-0.48	0.63
Apr	-3.13	-0.34	0.73
May	0.65	0.06	0.95
Jun	4.44	0.67	0.50
Jul	1.72	0.20	0.84
Aug	-11.23	-1.27	0.21
Sep	-7.92	-0.94	0.35
Oct	-1.90	-0.29	0.77
Nov	3.60	0.53	0.60
Dec	0.30	0.04	0.97

(b)

	Kendall slope	Z_{Sk}	p-value
Jan	1.28	0.17	0.86
Feb	-2.06	-0.48	0.63
Mar	-0.21	-0.06	0.95
Apr	-3.33	-0.31	0.75
May	8.71	0.99	0.32
Jun	12.06	1.76	0.08
Jul	8.53	1.05	0.29
Aug	0.74	0.04	0.97
Sep	1.17	0.12	0.90
Oct	3.05	0.42	0.67
Nov	8.21	1.19	0.24
Dec	3.97	0.72	0.47

In order to determine whether or not the seasons behave differently, the ?2 (homogeneous) values were computed and these are shown in Table 5 The results, which do not exceed the ?2 critical value at 0.1 level of significance (?2 critical=17.28) show that the seasons are homogeneous with respect to trend.

Table (5): ?2 homogeneity test for trend

	1FG01	1FG02
?2 (total)	4.70	7.86
?2 (trend)	0.80	2.70
?2 (homogeneous)	3.90	5.16

Trend line fitting

The results of the model that best fitted the data are given in the table below. It has been found that the quadratic model fits better than the others mentioned in the methodology section. This however does not mean this model can be used to forecast ahead of the data range used for model fitting. Rather it should be used to predict values only within that given data range.

Table 6: Trend models that describe the river discharge trends, with the values of the statistical tests for the goodness-of-fit (Q-stream flow and T is time).

Station	**Model**	**ME**	**MSE**	**MAE**	**MAPE**
1FG01	$Q=4587.2-0.25T+4.2E\text{-}06T^2$	-4.8	230726.3	389.9	44.3
1FG02	$Q=109.3+0.05T-6.4E\text{-}07T^2$	-4.3	331575.9	452.03	45.1

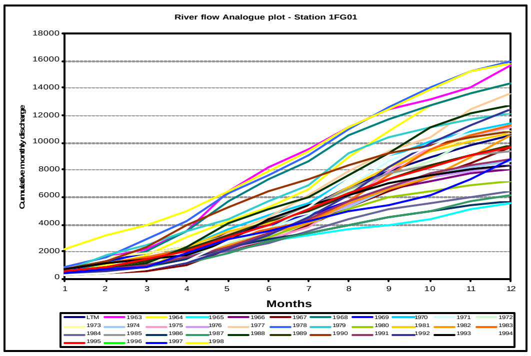
Probability distributions

Nine probability distributions were applied to the monthly total data. The best fitted distributions are shown in the Table 7 below. The station 1FG02, which showed an increasing trend (see Table 3) was further investigated by dividing the time series into two portions where the breaking point was determined by LOWESS smoothing. The probability distributions were then fitted to each portion. This was done to establish whether or not the probability distribution of the flow had changed with time. The first portion, 1963-1982, showed that the Gamma distribution fitted best while for the second portion, 1983-1998, the Log Pearson (III) fitted best. This is in contrast with 1FG01, which fitted Log Pearson (III) for both portions. This goes to show that the probability distribution for 1FG02 may have changed with time according to the results obtained herein. Lack of a significant trend for 1FG01 could be attributed to lack of a monotonic trend which was established through LOWESS smoothing. The distribution of flows within the year depends on a number of factors e.g. area of the basin, relief climate, soil conditions and land cover among others.

Table (7): The best-fitted distribution, parameters and the values of the goodness-of-fit tests

Station	Distribution	Distribution Parameters	A-D A^2	K-S D	?2 $?^2$
1FG01	1.Log Pearson (III)	a=6.998, ß=-0.239, ?=8.278	0.371	0.025	14.265
	2. Gamma	a=3.116, ß=281.240	0.754	0.039	15.576
1FG02	1.Log Pearson (III)	a=11.099, ß=-0.191, ?=8.845	0.161	0.017	8.363
	2. Gamma	a=2.922, ß=342.248	0.188	0.020	13.515

The results from the analogue years shown below indicate that 1963-64, 1968, 1977-79, 1988, 1990 and 1998 had above normal flows while the years with below normal flow were 1965, 1966, 1971, 1973, 1980, 1984, 1986 and 1987. The most severe floods occurred in 1968, 1977-78 and 1998. There was a marked reduction of flow in the period 1984-1987 recorded at 1FG01 and this was also illustrated by LOWESS smoothing. This period was a drought period and factors such as increased domestic and agricultural usage of river water could have contributed to this. After 1987, the flow increased again and this could be attributed to increased rainfall in 1988 as shown in the analogue plots coupled with less river water usage. People would use less of the river water since there would be enough rain for agricultural activities and they would also harvest rainwater for other uses.



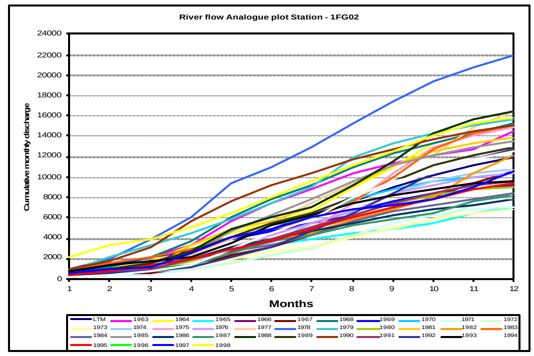


Figure 3 (a) and (b): Analogue plots for (a) Station 1FG01 and (b) Station 1FG02

Rainfall results for 10 Stations

Daily rainfall data was obtained from the Kenya Meteorological Department for the period 1963-1998. A few of the stations however had few data because of missing values. The data was aggregated into monthly totals and these were analyzed for trend using the same procedure as the river discharge and the results are presented in Table 8 below. Eight out of the ten selected stations showed negative slopes. Of these, four were found to have significant decreasing trend. These four stations are found within or near forested areas. In contrast, two of the stations that showed positive slopes, though not significant are located within a forested area. There was no clear pattern identified for the stations which showed positive or negative slopes. Thus there is need to have a more dense network of the rainfall stations which is hampered by lack of enough data.

Table (8): Seasonal Kendall slope and trend

Station	Slope Magnitude and direction (mm/yr)	Is there Trend? (Significance level, 5%)
Kakamega Forest Station	-8	Yes
Kaimosi Tea Estates Ltd.	2	No
Esirwa Kaimosi Farms	-7	Yes
Kaimosi Settlement	3	No
Nabkoi Forest Station	-2	No
Eldoret Kibabet Estates Ltd.	-3	No
Nandi Hills Kibwari Tea Estate	-7	Yes
Maseno Veterinary Station	-2	No
Maseno Siriba GTC	-12	Yes
Bondo Water Supply	-2	No

Table (9): Slope direction and trend (+ positive slope, - negative slope, grey shade is significant slope)

Station	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Kakamega Forest Station	+	-	-	-	-	-	-	-	+	-	-	-
Kaimosi Tea Estates Ltd.	+	-	-	+	+	+	+	-	-	+	+	-
Esirwa Kaimosi Farms	-	-	-	-	+	+	+	-	-	-	-	-
Kaimosi Settlement	+	-	+	-	+	+	+	+	+	+	+	+
Nabkoi Forest Station	-	-	-	-	-	+	+	-	-	-	+	+
Eldoret Kibabet Estates Ltd.	-	-	+	-	+	+	-	-	-	-	+	-
Nandi Hills Kibwari Tea Estate	+	-	+	-	+	+	-	-	-	-	+	-
Maseno Veterinary Station	+	-	+	-	-	-	-	-	+	-	-	-
Maseno Siriba GTC	+	-	+	-	-	-	+	+	-	+		+
Bondo Water Supply	+	+	+	-	-	+	+	+	+	+	+	+

Of the four different trend models explained in the methodology section, the best-fitted trend model was the linear model which gave the least values for MSE, MAE and MAPE. On average, we can say that the rainfall series shows a decreasing trend. A month-by-month Kendall analysis was done and the results are shown in Table 9. It can be seen that the month of December has experienced changes in the monthly rainfall amounts. Probability distribution fitted showed that the monthly rainfall data is well fitted by the Three-parameter Log Normal, Weibull and Pearson V distributions.

Conclusion

The null hypothesis H_0 that "there is no trend" was assumed. However, any given test for trend comes with it a definition of what is meant by "no trend". This includes assumptions usually related to the type of distribution and serial correlation. Failing to reject H_0 does not mean that it has been adequately proved that there is no trend. Rather, it implies that the evidence available is not sufficient to conclude that there is a trend. A number of factors that could contribute to these are the quality of data, amount of data and other exogenous variables. In this study, efforts were made to include as much data as possible, which was quality, controlled before usage. The rainfall data however had few missing values, although this was less than 10% of the total number of observations. The results obtained herein are an indication of climate variability in this region, which in the long term would constitute climate change. It has been shown that rainfall data shows a decreasing trend while river discharge shows a negative slope in the upstream station and an increasing significant trend in the downstream station. It is important that any trend or lack thereof should be investigated further through models that can account for external factors such as human and industrial activities. The trend significance determinations studied here are considered reasonable and appropriate indications of trend.

Acknowledgements

We acknowledge the support of Vlaamse Interuniversitaire Raad (VLIR), Belgium, in form of a Study Scholarship for the leading author of this paper.

References

Burn, D.H. and M.A. Hag Elnur. 2002. "Detection. Of Hydrologic Trends and Variability. Journal of Hydrology 255 (2002), 107-122.

Helsel, D.R. and Hirsch, R.M., 1991. Statistical Methods in Water Resources. Book 4, Hydrologic Analysis and Interpretation.

Hirsch, R.M., Slack, J.R. and Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data, Water Resour. Res. 18, 107–121.

Makridakis, S., Wheelwright, S.C. and Megee, V.E., 1986. Forecasting: Methods and Application, 2nd edition. Wiley, New York.

Shahin M., van Oorschot H.J.L., and de Lange S.J., 1993. Statistical Analysis in Water Resources Engineering. Balkema Publishers, Rotterdam, Netherlands.

Wilks D.S., 1995. Statistical Methods in Atmospheric Sciences.