Reference:

Raes, D., Willems, P. and GBaguidi, F. 2006. RAINBOW – a software package for analyzing data and testing the homogeneity of historical data sets. Proceedings of the 4th International Workshop on 'Sustainable management of marginal drylands'. Islamabad, Pakistan, 27-31 January 2006. (in press)

RAINBOW - a software package for hydrometeorological frequency analysis and testing the homogeneity of historical data sets

By Dirk RAES, Patrick WILLEMS and Félix GBAGUIDI

Abstract

A common problem in many areas of water resources engineering is that of analyzing hydrological and meteorological events for planning and design projects. For these purposes, information is required on rainfall events, flow depths, discharges, evapotranspiration levels, etc. that can be expected for a selected probability or return period. In the paper the software tool RAINBOW is presented which is designed to study meteorological or hydrologic records by means of a frequency analysis and to test the homogeneity of the record. After the selection or creation of a data set, an analysis on the data is performed. When opting for a frequency analysis, a menu is opened which contains various folders where a probability distribution can be selected, the data transformed, and results can be viewed or saved on disk. In RAINBOW the user can select a Normal, Log-Normal, Weibul, Gamma, Gumbel, Exponential or Pareto distribution. Apart from graphical methods (Probability plot and a Histogram of the data superimposed by the selected probability function) for evaluating the goodness of fit, RAINBOW offers also statistical tests for investigating whether data follow a certain distribution (Chi-square and the Kolmogorov-Smirnov test). When the goodness-of-fit is inadequate, one can either select another distribution or attempt to normalize the data by selecting a mathematical operator to transform the data. RAINBOW allows also to analyse time-series with zero or near zero events (the so called nil values) by separating temporarily the nil values from the non-nil values. By calculating the global probability, the nil and no-nil rainfall are combined again. When the probability distribution can be accepted, the user can view the calculated events that can be expected for selected probabilities or return periods. Frequency analysis of data requires that the data be homogeneous and independent. The restriction of homogeneity assures that the observations are from the same population. RAINBOW offers a test of homogeneity which is based on the cumulative deviations from the mean. By evaluating the maximum and the range of the cumulative deviations from the mean, the homogeneity of the data of a time series is tested.

The RAINBOW software itself is easy to install and use. It is menu driven and no specific computer knowledge is required. The software is freely available on the web. To DOWNLOAD go to http://www.iupware.be and select downloads and next software.

Introduction

The paper presents the software package RAINBOW with which magnitudes for events can be estimated that can be expected for a selected probability or return period. Such estimates can be obtained by means of a frequency analysis on historical data. Depending on the objective of the exercise, the type of data to be analysed can vary widely from one application to another. For hydrologic purposes, typically historical time series of meteorological and hydrological data are analysed to determine design rainfall depths, evapotranspiration levels, floods, etc that can occur with a selected probability. These estimates are required for the design of canals, pipelines, reservoirs, floodwater-spreading systems and hydraulic structures and for the proper management of floodwater and rainwater harvesting schemes, and irrigation and drainage projects. The selection of the probability or return period for design purposes is related to the damage the excess or the shortage of rainfall may cause, the risk one wants to accept and the lifetime of the project.

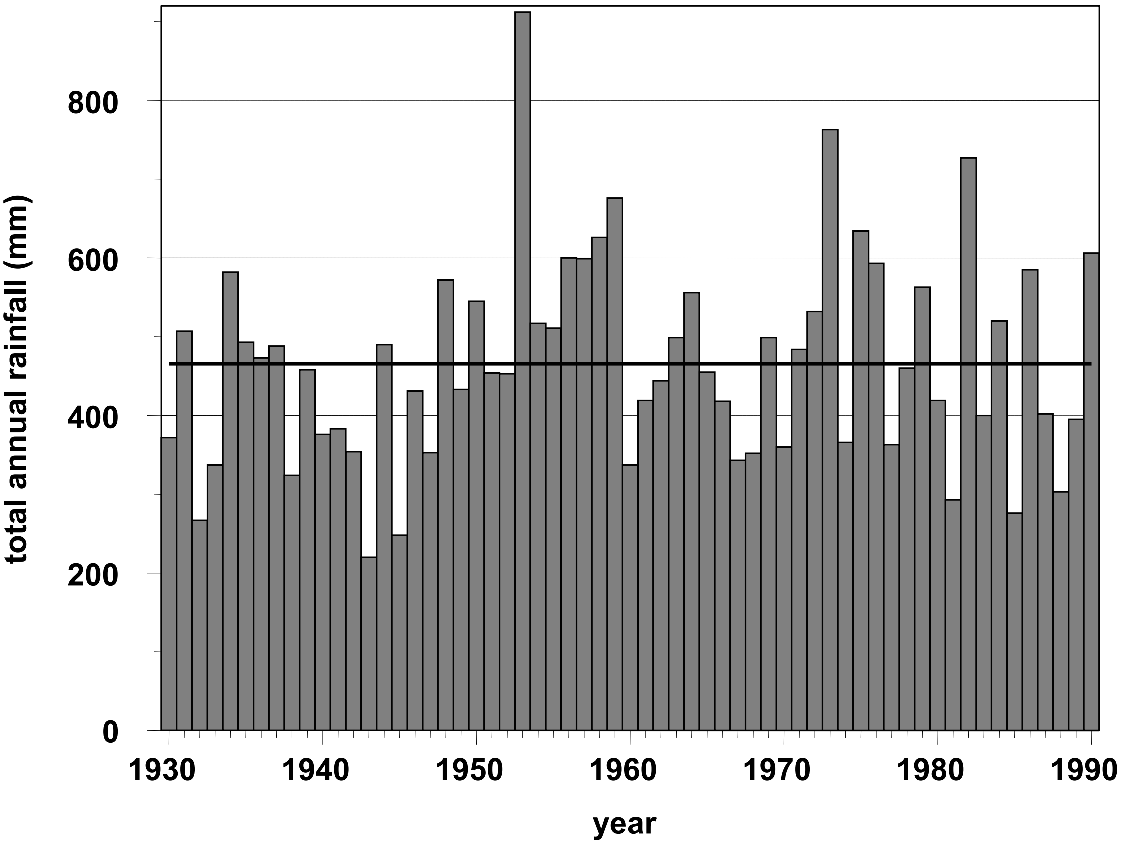


Figure 1. Total annual rainfall recorded in Tunis (Tunisia) for the period 1930-1990 with indication of the average rainfall (horizontal line).

To demonstrate the software, examples are worked out on time series of rainfall data extracted from the FAOCLIM databank (FAO, 2000). The total rainfall received in a given period at a particular location is highly variable from one year to another (Fig. 1).

The variability depends on the type of climate and the length of the considered period. Although time series of historic rainfall data are characterized by their average and standard variation, these values cannot be blindly used to estimate based on a normal distribution design rainfall depths that can be expected with a specific probability or return period. Applying this technique to a data set can produce misleading results since the actual characteristics of the distribution are ignored and it is assumed that they follow a particular distribution. To avoid this type of error, it is essential that the goodness of the assumed distribution be checked before design rainfall depths are estimated.

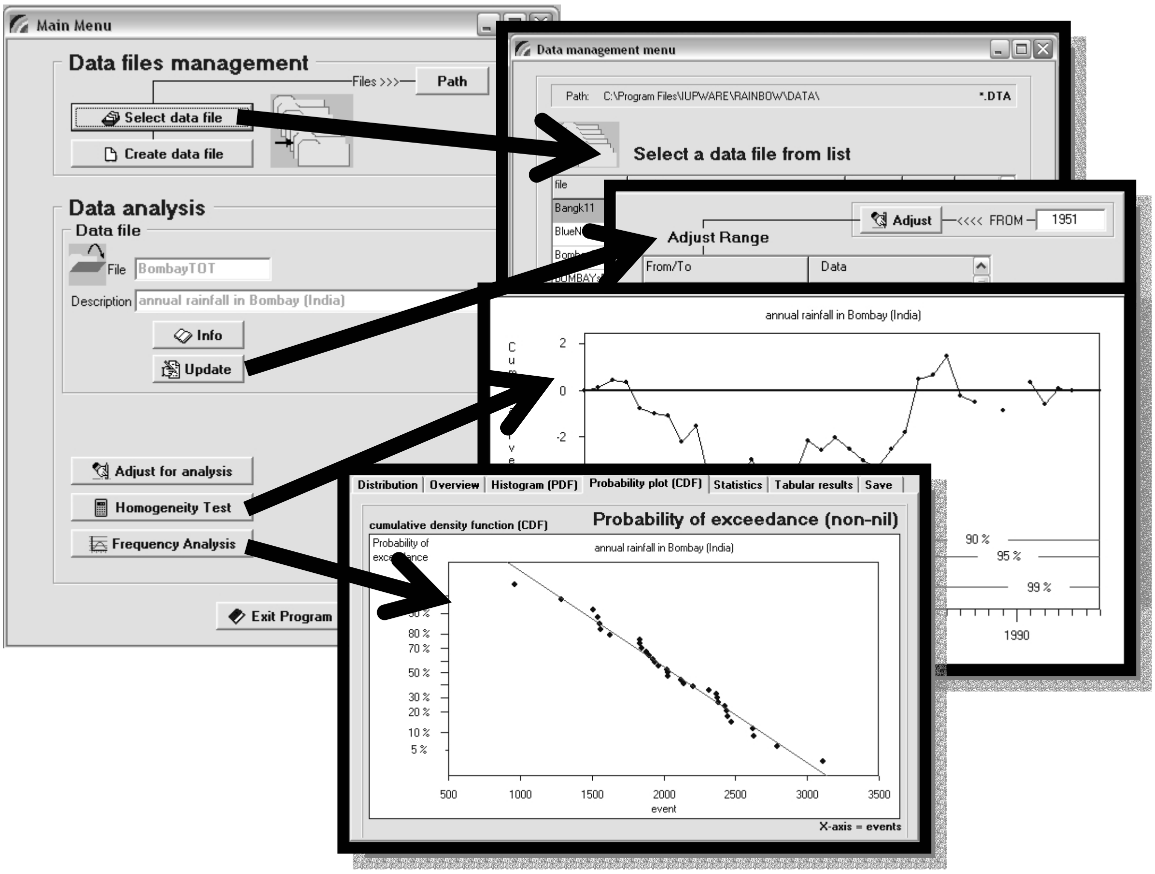
In a frequency analysis (Snedecor and Cochran 1980; WMO, 1981, 1983 and 1990; Haan, 2002) estimates of the probability of occurrence of future rainfall events are based on the analysis of historical rainfall records. By assuming that the past and future data sets are stationary and have no apparent trend one may expect that future time series will reveal frequency distributions similar to the observed one. It is obvious that the longer the data series the more similar the frequency distribution will be to the probability distribution. As the number of observations increases, the error in determining expected rainfall gradually diminishes. Although the required length of the time series depends on the magnitude of variability of the precipitation climate, a period of 30 years and over normally is thought to be very satisfactory. However, if interest lies in extreme rainfall events, larger number of years may be required.

Frequency analysis requires considerable computations and careful plotting. Efficiency can be gained by using software such as RAINBOW. This software has been specially designed to carry out frequency analyses and to test the homogeneity of data sets.

Structure of the RAINBOW program

The hierarchical structure of the RAINBOW program is presented in Figure 2. From the Main Menu, the user has access to the data and can perform an analysis. An analysis starts with the selection or creation of a data file. A rainfall data file contains typically historical observations of 10-daily, monthly, seasonal or yearly rainfall over a sufficient number of years. In stead of creating files when running RAINBOW, the user can also copy the available data from for example a spreadsheet and paste them in a data file as long as the user respects the structure and extension of the files. Data files are stored by default in the DATA subdirectory of the program, but with the help of the ‘Path’ button, files stored in other directories or drives can be accessed.

Once the data file is selected, an analysis on the data can be performed by selecting the ‘Homogeneity test’ or ‘Frequency analysis’. After the analysis, one returns to the Main menu to select other data files or perform other tests on the same data file.



Path: C:\Program Files\IUPWARE\RAINBOW\DATA\ *.DTA

File: BombayTOT

Description: annual rainfall in Bombay (India)

FROM 1951

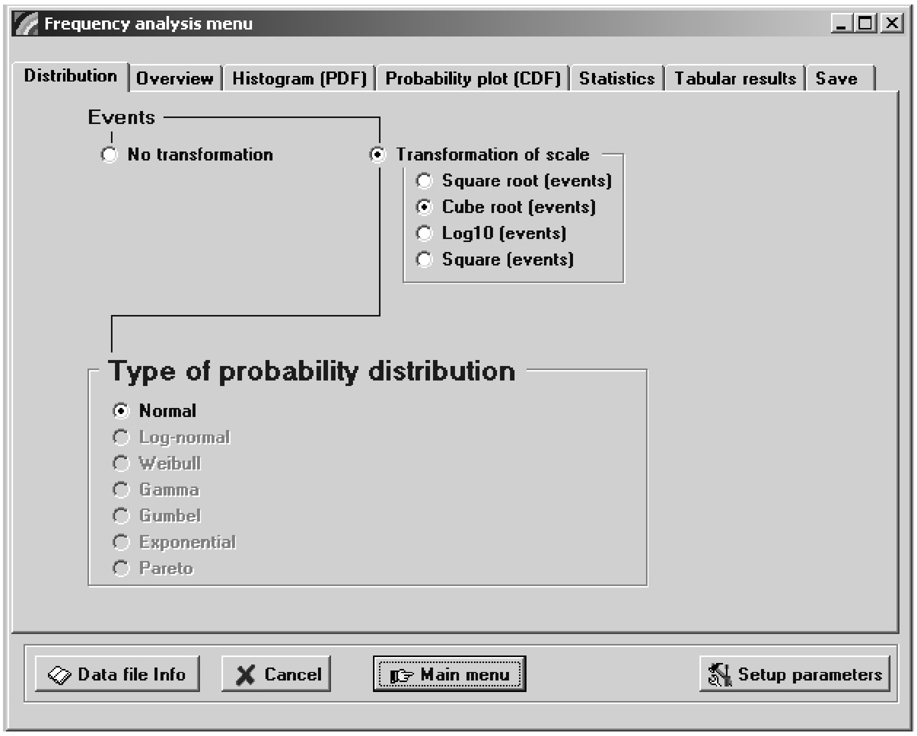
Figure 2. Structure of the RAINBOW program

Frequency analysis

When opting for a frequency analysis in the Main menu, the user is guided to the 'Frequency analysis' menu (Fig. 3) which contains various folders where the probability distribution can be selected, the data transformed, and results can be viewed or saved on disk. In the menu, the user can also setup options for assigning plotting positions and for estimating statistical parameters which are required when analysing the data.

From a frequency analysis, the estimates of rainfall depths for selected probabilities or return periods are obtained. The analysis consists in:

- ranking the historical data and assigning plotting positions by estimating the probability of exceedance with one or another method (Table 1);
- selecting a distributional assumption and plotting the data in a probability plot;
- verifying the goodness of the selected distribution. If unsatisfactory another distribution should be selected or the data should be transformed so that the transformed data follow the selected distribution;
- determining rainfall depths that can be expected for selected probabilities or return period from the probability plot.



Events

○ No transformation

● Transformation of scale

○ Square root (events)

● Cube root (events)

○ Log10 (events)

○ Square (events)

Type of probability distribution

● Normal

○ Log-normal

○ Weibull

○ Gamma

○ Gumbel

○ Exponential

○ Pareto

Figure 3. The frequency analysis menu with the various folders where the user can select options and view results.

Ranking data and assigning plotting positions:

The program ranks the historical data in descending order and assigns a serial rank number (r) ranging from 1 to n (number of observations) to the ranked data. Subsequently the probability of exceedance for each observation is estimated by one or another method. These probabilities will be the plotting positions for the ranked observations in the probability plot (Fig 4). Although the results will not differ profoundly from one to each other, RAINBOW allows the user to select a particular method (Tab. 1). The Weibull plotting position is the default setting.

Selecting a distributional assumption and plotting the data in a probability plot:

When selecting a distributional assumption, a frequency histogram superimposed by (corresponding scaled version of) the probability density function (Fig. 6) and probability plot (Fig. 4) are drawn in the corresponding folders. In RAINBOW the user can select a Normal (Haan, 2002), Log-Normal (Aitchison and Brown, 1957; Crow and Shimizu, 1988; Evans et al. 1993), Weibul (Haan, 2002), Gamma (Thom, 1951; Markovic, 1965; Mooley, 1973; Aksoy, 2000), Gumbel (Gumbel, 1958), Exponential (Haan, 2002) or Pareto (Norman et al., 1994) distribution

Table 1. Methods for estimating probabilities of exceedance (plotting positions) of ranked data, where r is the rank number and n the number of observations (Raes et al., 1996; Gbaguidi, 2005).

Method and (Source)	**Estimate of probability of exceedance (%)**
California (California State Department, 1923)	$\frac{r}{n}\ 100$
Hazen (Hazen, 1930)	$\frac{(r-0.5)}{n}\ 100$
Weibull (Weibul, 1939)	$\frac{r}{(n+1)}\ 100$
Cunnane Cunnane (1978)	$\frac{(r-0.4)}{(n+0.2)}100$
Gringorten (WMO, 1983)	$\frac{(r-0.44)}{(n+0.12)}\ 100$
Sevruk and Geiger (Sevruk and Geiger, 1981)	$\frac{(r-3/8)}{(n+1/4)}100$
Adamowski (Adamowski, 1981)	$\frac{(r-0.26)}{(n+0.5)}100$

A probability plot (Fig. 4) is a plot of the rainfall depths versus their probabilities of exceedance as determined by one or another method (Tab. 1). When the data are plotted in a graph where both axes have a linear scale, the data are not likely to be on a straight line but to follow a S-shaped curve. By selecting a probability distribution, the vertical axis of the probability plot is rescaled so that the data will fall on a straight line if it is distributed as selected (Fig. 5). On probability paper the cumulative distribution of the total population will fall on that straight line. This makes the verification of the goodness of selected distribution easier. Figure 5 refers to a normal distribution, but the same is true for other distributions. Only the rescaling of the vertical axis will be different.

In the plot of the frequency histogram (Fig. 6), RAINBOW constructs a frequency histogram of the observed data and superimposes it with the selected probability density function (after rescaling it to represent frequencies). The class interval is selected by the program as such that at least five observations belong to one class.

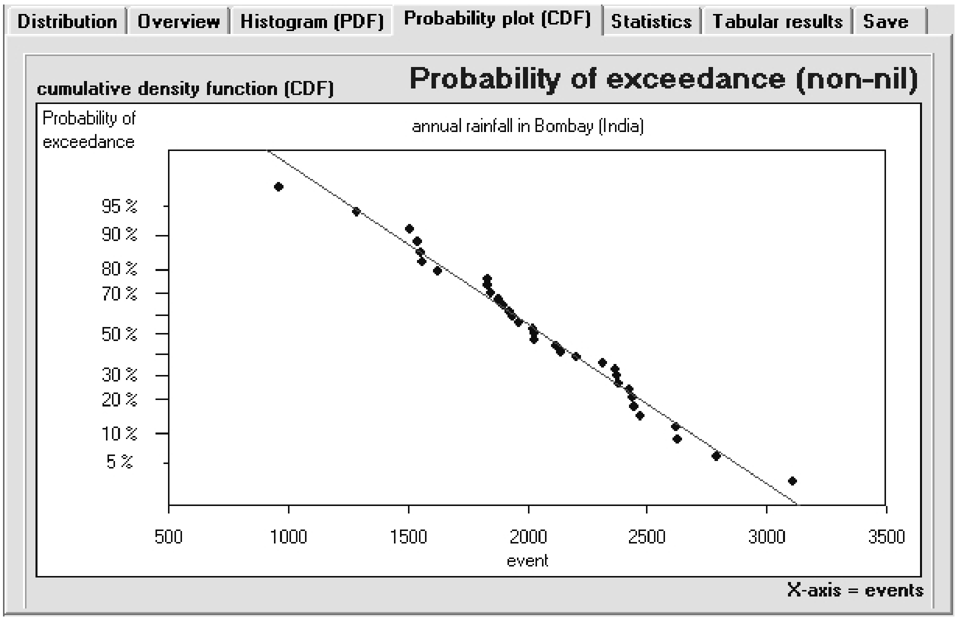


Figure 4. Probability plot for the annual rainfall (1960 – 1996) in Bombay (Normal distribution).

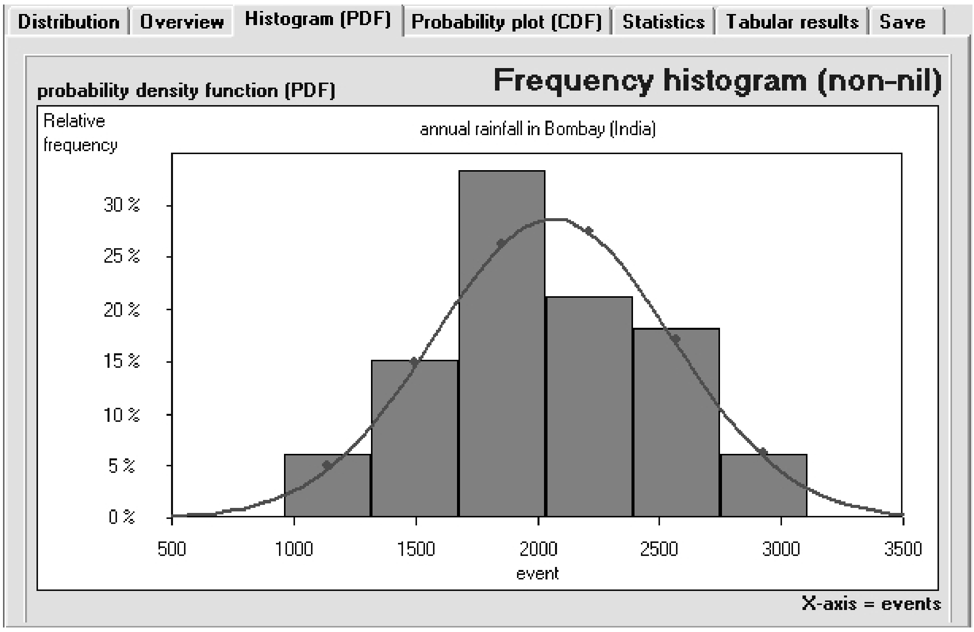


Figure 6. Histogram superimposed by the probability density function for the annual rainfall (1960 – 1996) in Bombay (Normal distribution).

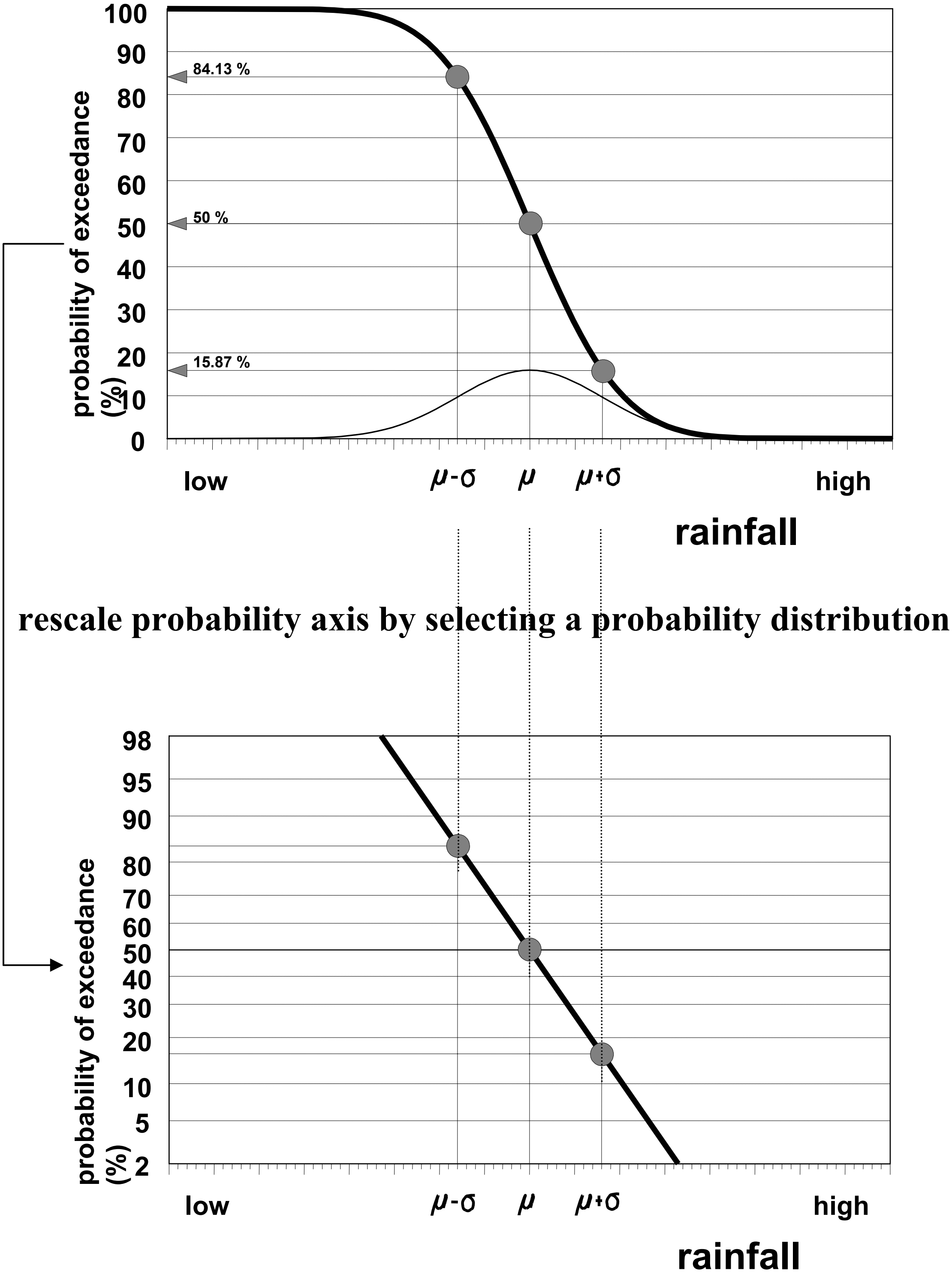
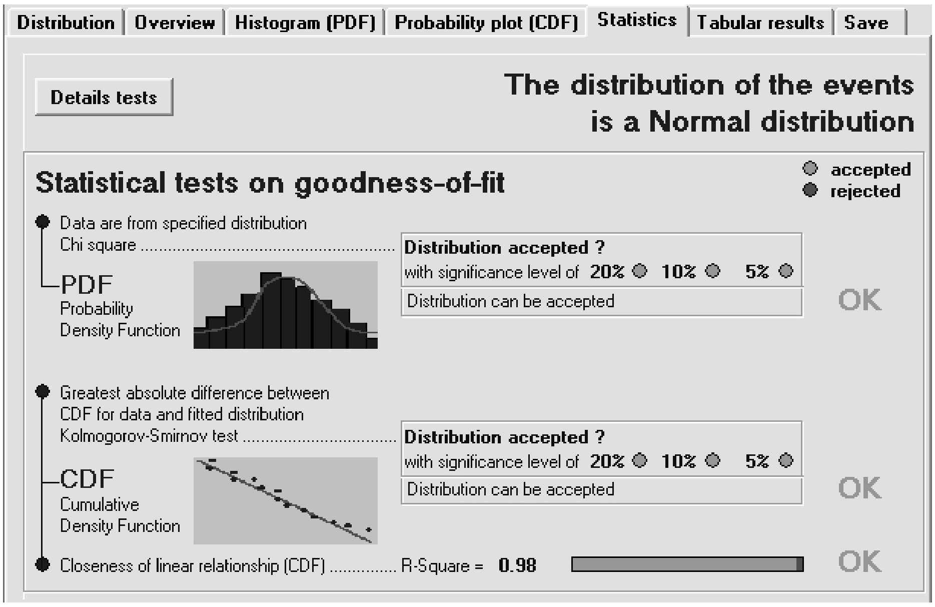


Figure 5. Effect of the rescaling of the vertical axis of a probability plot

Statistical parameters, describing the characteristics of the data set, are required in the frequency analysis. The parameters are the mean and standard deviation (Normal and Lognormal distribution), the shape and scale parameters (Gamma and Weibull distribution) or the one parameter for the Exponential (λ) and Pareto distribution (α). RAINBOW offers different options for parameter estimation: the Method of Moment (Atwood et al., 2003), the Maximum Likelihood Method (Law and Kelton, 1991) and the Regression Method. For most distributions the estimated parameters will vary somewhat with the selected method. The most commonly used method is the Maximum Likelihood Method that Law and Kelton (1991) qualified as the preferred method of parameter estimation for distribution fitting. When selecting the Regression Method, the mean and standard deviation of the data set are obtained from the best fitted line through the data in the probability plot (Normal and Lognormal distribution).



The distribution of the events is a Normal distribution

Statistical tests on goodness-of-fit

○ accepted
● rejected

Data are from specified distribution
Chi square ....

Distribution accepted ?
with significance level of 20% ○ 10% ○ 5% ○
Distribution can be accepted

Greatest absolute difference between CDF for data and fitted distribution
Kolmogorov-Smirnov test ....

Distribution accepted ?
with significance level of 20% ○ 10% ○ 5% ○
Distribution can be accepted

Closeness of linear relationship (CDF) .... R-Square = 0.98

Figure 7. Result from statistical tests evaluating the goodness of fit for the Normal distribution of the annual rainfall (1960 – 1996) of Bombay.

Verifying the goodness of the selected distribution:

If the data in a probability plot (Fig. 4) fall in a reasonable alignment, it may be assumed that the data can be approximated by the assumed distribution. Since the data is only a sample of the total population it would be rare for a set of data to plot exactly on a line and a decision must be made as to whether or not the deviations from the line are random variations or represent true deviations indicating that the data does not follow the given probability distribution. By fitting a line through the points an indication of the goodness of fit is given by the coefficient of determination (R^2) of the fitted line. Owing to sampling variations, the points will depart somewhat from the line even with data that follow perfectly the assumed distribution. When the points in the probability plot do not fall in a reasonable alignment, the data is most likely not distributed as the selected distribution especially if the points deviate from the straight line in some systematic matter.

Apart from graphical methods (Probability plot and Histogram) for evaluating the goodness of fit, RAINBOW offers also statistical tests (Haktanir and Holacher, 1993; Kottegoda, 1980) for investigating whether data follows a certain distribution (Fig. 7). The null hypothesis (H_0) is that the data comes from a distribution of the assumed form. The Chi-square test (Snedecor and Cochan, 1980) is based on the probability density function (Fig. 6). The smaller the value of the χ^2 statistics, the better the expected model fit to the sample at hand (Topaloglu, 2000). The χ^2 finds evidence against the null hypothesis in terms of a probability (P-value of the test). The smaller the P-value the stronger the evidence against H_0. RAINBOW tests with significance levels of $\alpha = 0.20$, 0.10 and 0.05. The Kolmogorov-Smirnov test (Topaloglu, 2000) is based on the cumulative density function (Fig. 4). The statistic used is the greatest absolute difference between the cumulative density function for the data and the fitted distribution. The difference is compared with critical values selected according to the significance levels of $\alpha = 0.20$, 0.10 and 0.05. If the difference is smaller than the critical value the assumed probability distribution is accepted with that level of significance.

Transformation of the data

When the goodness-of-fit is inadequate, one can either select another distribution or attempt to normalize the data by selecting a mathematical operator to transform the data (Raes et al., 1996). Since dealing with a normal distribution has several practical advantages, it is common practice to transform data that are not normally distributed so that the resulting normalized data can be presented by the normal curve. The transformation of the data will change the scale of the records (i.e. the abscissa of the probability plot).

For positively skewed data a transformation can be used to reduce higher values by proportionally greater amounts than smaller values. This transformation will rescale the magnitude of the records and the transformed data might be closer to the normal distribution than the original data (Fig. 8). Operators available in RAINBOW to rescale the data are the square root (resulting in a fairly moderate transformation), the cube root and the logarithm (resulting in a substantive transformation).

Data sets with zero rainfall

For months at the onset or cessation of the rainy season, or for small periods such as weeks or 10-day periods, rainfall data might be zero or near zero in some of the years. As such the rainfall data is bounded on the left by zero or near zero values. If the occurrence of low rainfall is high, the frequency distribution becomes severely skewed. A method to analyse time-series with zero or near zero rainfall (the so called nil values) is to separate temporarily the nil values from the non-nil values. RAINBOW allows the specification of a nil value different from zero. By excluding the nil's from the frequency analysis, the frequency distribution becomes less skewed to the left, and the data can be analysed. By calculating the global probability, the nil and no-nil rainfall are combined. This type of mixed distribution with a finite probability that X = nil and a continuous distribution of probability for X > nil is discussed by Haan (2002).

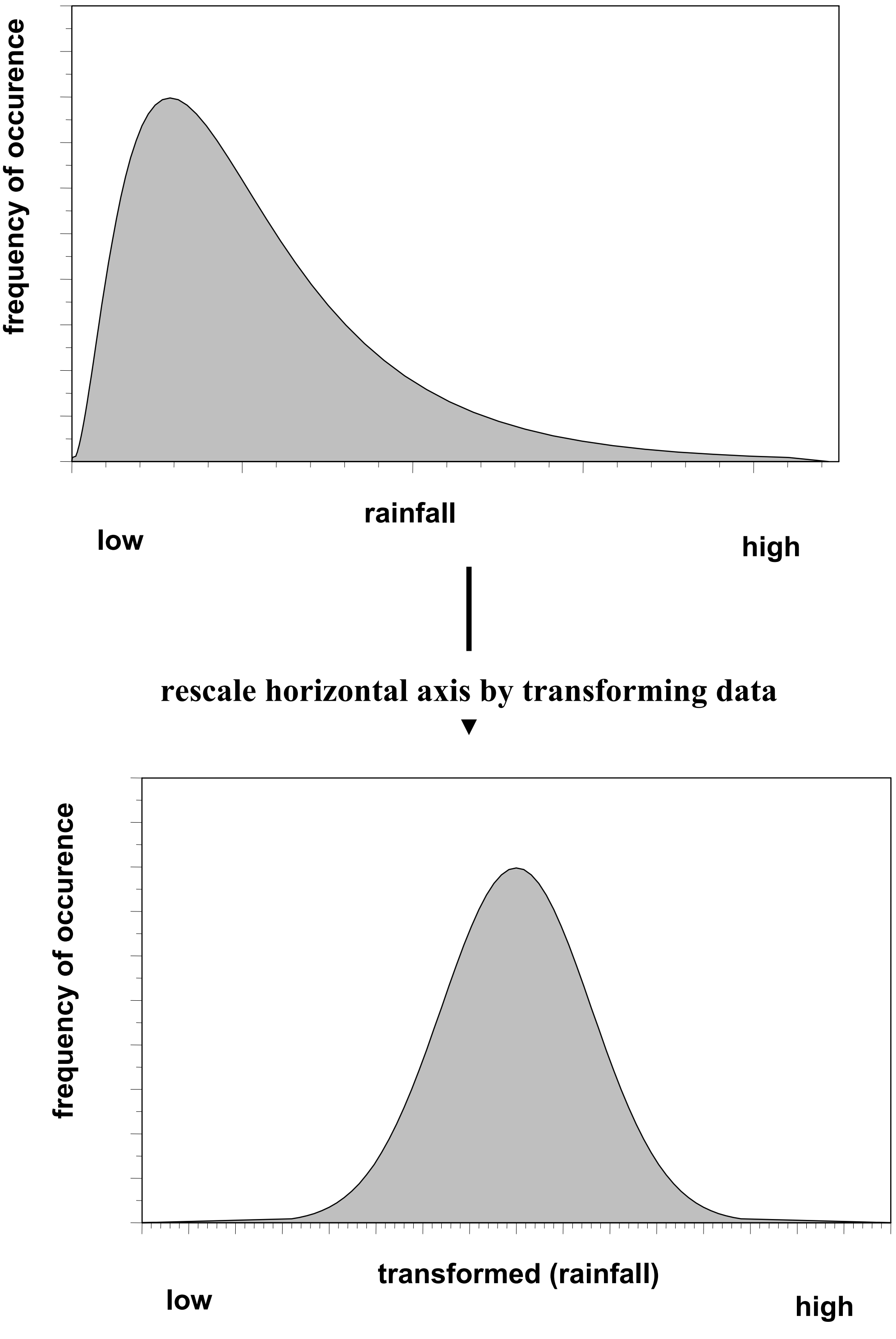
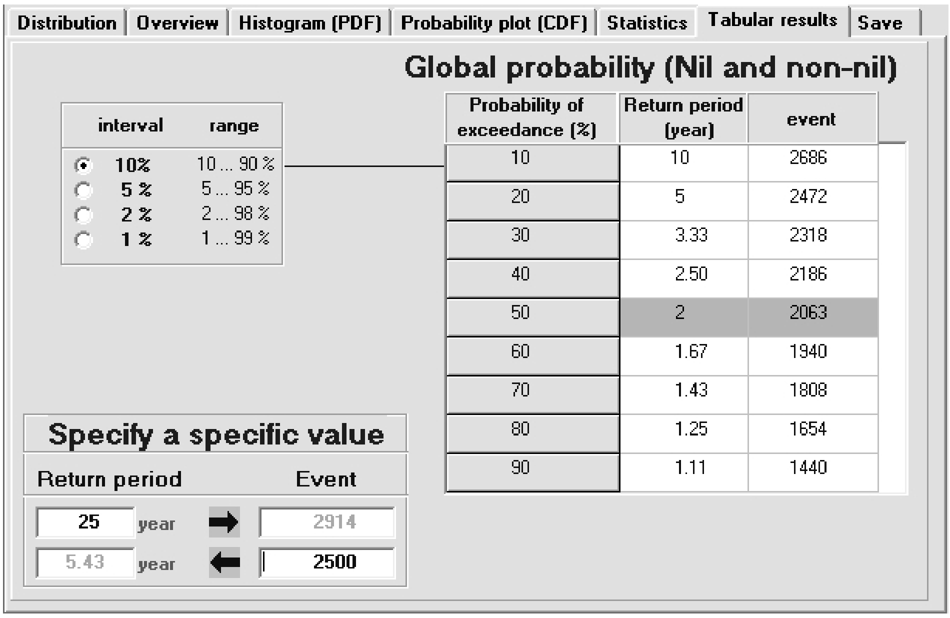


Figure 8. Transformation of positively skewed rainfall data.



Global probability (Nil and non-nil)

	interval	range
●	10%	10 ... 90 %
○	5 %	5 ... 95 %
○	2 %	2 ... 98 %
○	1 %	1 ... 99 %

Probability of exceedance (%)	Return period (year)	event
10	10	2686
20	5	2472
30	3.33	2318
40	2.50	2186
50	2	2063
60	1.67	1940
70	1.43	1808
80	1.25	1654
90	1.11	1440

Specify a specific value

Return period		Event
25 year	➡	2914
5.43 year	⬅	2500

Figure 9. Tabular results of the frequency analysis on the annual rainfall (1960 – 1996) of Bombay.

Determining rainfall depths that can be expected for selected probabilities or return periods

When the probability distribution can be accepted, the user can find the rainfall depths (X_P) that can be expected for selected probabilities in a Table (Fig. 9). The probability refers to the probability of exceedance and it specifies the likelihood that the actual rainfall will be equal to or higher than the estimated rainfall depth X_P. The return period (also called the recurrence interval) is the average time between successive years where the value of X_P is exceeded. It is the reciprocal value of the probability when expressed as a fraction.

Homogeneity test of time series

Frequency analysis of data requires that the data be homogeneous and independent. The restriction of homogeneity assures that the observations are from the same population. One of the tests of homogeneity (Buishand, 1982) is based on the cumulative deviations from the mean:

$$S_k = \sum_{i=1}^{k}\left(X_i - \overline{X}\right) \qquad\qquad k = 1, \dots, n$$

where X_i are the records from the series $X_1, X_2, \dots, X_n$ and $\overline{X}$ the mean. The initial value of $S_{k=0}$ and last value $S_{k=n}$ are equal to zero (Figure 10). When plotting the S_k's (also called a residual mass curve) changes in the mean are easily detected. For a record X_i above normal the $S_{k=i}$ increases, while for a record below normal $S_{k=i}$ decreases. For a homogenous record one may expect that the S_k's fluctuate around zero since there is no systematic pattern in the deviations of the X_i's from their average value $\overline{X}$.

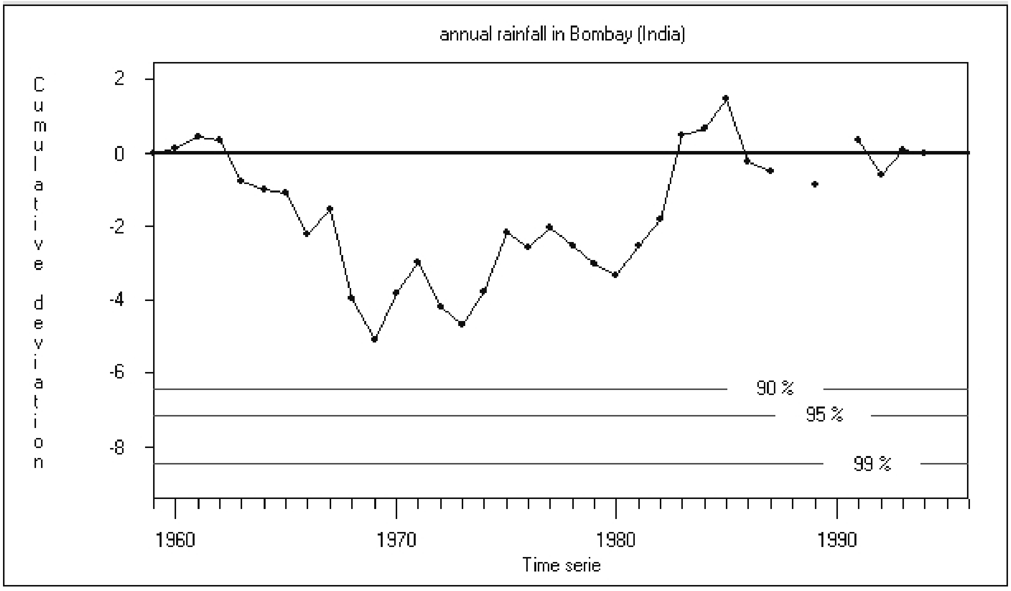


Figure 10. Rescaled cumulative deviations from the mean for the total annual rainfall (1960 – 1996) for Bombay. When the deviation crosses one of the horizontal lines the homogeneity of the data set is rejected with respectively 90, 95 and 99% probability.

To test the homogeneity of the data set, RAINBOW rescales the cumulative deviations by dividing the S_k's by the sample standard deviation value. By evaluating the maximum (Q) and the range (R) of the rescaled cumulative deviations from the mean, the homogeneity of the data of a time series can be tested. High values of Q or R are an indication that the data of the time series is not from the same population and that the fluctuations are not purely random. Critical values for the test-statistic which test the significance of the departures from homogeneity are plotted as well (Fig. 10).

References

Adamowski, K., 1981. Plotting formula for flood frequency. Water Resources Bulletin 17(2): 197-202.

Aitchison, J., and Brown, J.A.C. 1957. The lognormal distribution, Cambridge University Press, Cambridge, UK.

Aksoy, H. 200. Use of gamma distribution in hydrological analysis. Turk; J. Engin. Environ. Sci. 24: 419-428.

Atwood, C.L., Lachance J.L, Martz, H.F., Anderson, D.J., Englehendt, M., Whitehead, D., and Wheeler, T. 2003. Handbook of parameter estimation for probability risk assessment. US. Nuclear regulatory commission. Office of nuclear research. Washington, DC.

Buishand, T.A. 1982. Some methods for testing the homogeneity of rainfall records. Journal of Hydrology, (58): 11 – 27.

California State Department of Public works. 1923. Flow in California streams. Bulletin 5. Chapter 5 (cited by Haan, 2002).

Crow, E.L., and Shimizu, K. 1988. Lognormal distribution, theory and applications, Marcel Dekker, New York, USA.

Cunnane, C. 1978. Unbiased plotting positions – a review. Journal of hydrology 37(3/4): 205-222 (cited by Haan, 2002).

Evans, M., Hastings, N. and Peacock, B. 1993. Statistical Distributions, Second Edition, John Wiley and Sons, New York, USA.

FAO. 2000. FAOCLIM – World-wide agroclimatic database. Environment and Natural Resources Working Paper No. 5. (CD-ROM). FAO - Agro meteorology group, Rome, Italy.

Gbaguidi, F. 2005. Hydroclimatological statistics for water engineering applications. MSc dissertation, Interuniversity Programme in Water Resources Engineering (IUPWARE), VUB and K.U.Leuven, Belgium. 87 pp.

Gumbel, E.J. 1958. Statistics of extreme values. Columbia University Press, New York, USA.

Haan, C.T. 2002. Statistical methods in Hydrology. Second Edition. Iowa State University Press, Ames, Iowa, USA: 128 -160.

Haktanir, T. and Holacher, H.B. 1993. Evaluation of various distributions for flood frequency analysis. Hydrol. Sco. J. 38(1): 15-32.

Hazen, A. 1930. Flood flow, a study of frequencies and magnitudes. John Wiley and Sons, INC. New York (as cited by Haan, 2002).

Kottegoda, N.T. 1980. Stochastic water resources technology. Dept. of Civil Engineering. University of Birmingham. The Macmillan Pres, Ltd. London.

Law, A.M., and Kelton, W.D. 1991. Simulation modeling and analysis. Mc.Graw-Hill InC., New York, USA.

Markovic, R.D. 1965. Probability functions of best fit distribution of annual precipitation and runoff. Hydro-paper 8. Colorado State University, Colorado, USA.

Mooley, D.A. 1973. Gamma distribution probability model for Asian summer monthly rainfall. Monthly weather review 101(2): 160-176.

Norman, L.J., Kortz, S., and Balakrishman, N. 1994. Continuous univariate distributions, Vol 1, John Wiley & sons.

Raes, D., Mallants, D. and Song Z. 1996. RAINBOW – a software package for analysing hydrologic data. In Blain W.R. (Ed.) Hydraulic Engineering Software VI. Computational Mechanics Publications, Southampton, Boston: 525-534.

Sevruk, B. and Geiger, H. 1981. Selection of distribution types for extremes of precipitation. World Meteorological Organisation, Operational Hydrology Report, No. 15, WMO-No. 560, Geneva.

Snedecor, G.W. and Cochran, W.G. 1980. Statistical methods. Iowa State university press, USA. 507 pp.

Thom, H.C.S. 1951. A note on the Gamma distribution. Monthly weather review 86(4): 117-122 (as cited by Haan, 2002).

Topaloglu, F. 2000. Determining suitable probability models for flow and precipitation series of the Seyhan river basin. Turk J. Agric. 26: 187-194.

Weibull, W. 1939. A statistical study of the strength of material. Ing. Vetenskaps Akad. Handl. (Stockholm) Vol. 151, pp. 15 (as cited by Haan, 2002).

WMO. 1981. Guide to agricultural meteorological practices. World Meteorological Organization, WMO – No. 134. Geneva, Switzerland.

WMO. 1983. Guide to climatological practices. World Meteorological Organization, WMO – No. 100. Geneva, Switzerland.

WMO. 1990. On the statistical analysis of series of observations. World Meteorological Organization, WMO –N° 415. Geneva, Switzerland. 192 pp.