HYDROLOGY PROJECT
Technical Assistance

World Bank & Government of The Netherlands funded

Training module # SWDP - 41

How to analyse discharge data

New Delhi, November 1999

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: dhvdelft@del2.vsnl.net.in

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

Page

1. Module context

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	How to analyse discharge data
Target group	:	Assistant Hydrologists, Hydrologists, Data Processing Data Managers
Duration	:	One session of 120 minutes
Objectives	:	After the training the participants will be able to: • Analyse discharge data
Key concepts	:	• homogeneity of discharge data • basic statistics • min-mean-max series • frequency & duration curves • frequency distributions
Training methods	:	Lecture, Software
Training tools required	:	Board, OHS, Computer
Handouts	:	As provided in this module
Further reading and references	:	

3. Session plan

No	Activities	Time	Tools
1	***General*** • Important points	5 min	 OHS 1
2	***Computation of basic statistics*** • Basic statistics	5 min	 OHS 2
3	***Empirical frequency distributions (flow duration curves)*** • General • Tabulation for flow duration curve • Flow duration curves • Standardised flow duration curve	10 min	 OHS 3 OHS 4 OHS 5 OHS 6
4	***Fitting of frequency distributions*** 4.1 General description 4.2 Frequency distribution of extremes • Important points • Frequency distributions • Flood frequency plot	10 min	 OHS 7 OHS 8 OHS 9 OHS 10
5	***Time series analysis*** 5.1 Moving averages • General • Computations • Moving averages and original series plot 5.2 Mass curves and residual mass curves • Residual mass curve • Residual mass curve used in reservoir design 5.3 Run length and run sum characteristics • Definition diagram of run length and run sum 5.4 Balances	10 min	 OHS 11 OHS 12 OHS 13 OHS 14 OHS 15 OHS 16 OHS 17
6.	***Regression/relation curves***	2 min	OHS 18
7.	***Double mass analysis***	2 min	OHS 19
8.	***Series homogeneity testing***	2 min	OHS 20
9.	***Rainfall runoff simulation***	4 min	OHS 21
10.	***Exercise*** • Work out basic statistics of a long term monthly series • Use normal distribution for annual actual rainfall series • Use moving average and residual mass curve to see long term trends in the monthly rainfall data	70 min	

4. Overhead/flipchart master

5. Handout

Add copy of Main text in chapter 8, for all participants.

6. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

7. Main text

Contents

How to analyse discharge data

1. General

- **The purpose of hydrological data processing software is not primarily hydrological analysis. However, various kinds of analysis are required for data validation and further analysis may be required for data presentation and reporting. Only such analysis is considered in this module**

- **Analysis will be carried out at Divisional level or at State Data Processing Centres.**

- **There is a shared need for methods of statistical and hydrological analysis with rainfall and other climatic variables. Many tests have therefore already been described and will be briefly summarised here with reference previous Modules.**

- **The types of analysis considered in this module are:**

  - computation of basic statistics
  - empirical frequency distributions and cumulative frequency distributions (flow duration curves)
  - fitting of theoretical frequency distributions
  - Time series analysis
    - moving averages
    - mass curves
    - residual mass curves
    - balances
  - regression/relation curves
  - double mass analysis
  - series homogeneity tests
  - rainfall runoff simulation

2. Computation of basic statistics

Basic statistics are widely required for validation and reporting. The following are commonly used:

- arithmetic mean

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (1)$$

- median - the median value of a ranked series X_i
- mode - the value of X which occurs with greatest frequency or the middle value of the class with greatest frequency
- standard deviation - the root mean squared deviation S_x :

$$S_x = \sqrt{\frac{\sum (X_i - \overline{X})^2}{N-1}}$$

- skewness or the extent to which the data deviate from a symmetrical distribution

$$C_X = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^3}{S_X{}^3}$$

- kurtosis or peakedness of a distribution

$$K_X \quad = \quad \frac{(N^2 - 2N + 3)}{(N-1)(N-2)(N-3)} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^4}{S_X{}^4}$$

3. Empirical frequency distributions (flow duration curves)

A popular method of studying the variability of streamflow is through flow duration curves which can be regarded as a standard reporting output from hydrological data processing. Some of their uses are:

- in evaluating dependable flows in the planning of water resources engineering projects
- in evaluating the characteristics of the hydropower potential of a river
- in assessing the effects of river regulation and abstractions on river ecology
- in the design of drainage systems
- in flood control studies
- in computing the sediment load and dissolved solids load of a river
- in comparing with adjacent catchments.

A flow-duration curve is a plot of discharge against the percentage of time the flow was equalled or exceeded. This may also be referred to as a cumulative discharge frequency curve and it is usually applied to daily mean discharges. **The analysis procedure is as follows:**

Taking the N years of flow records from a river gauging station there are 365n daily mean discharges.

1. The frequency or number of occurrence in selected classes is counted (Table 1). The class ranges of discharge do not need to be the same.
2. The class frequencies are converted to cumulative frequencies starting with the highest discharge class.
3. The cumulative frequencies are then converted to percentage cumulative frequencies. The percentage frequency represents the percentage time that the discharge equals or exceeds the lower value of the discharge class interval.
4. Discharge is then plotted against percentage time. Fig. 1 shows an example based on natural scales for the data in Table 1. A histogram plot may also be made of the actual frequency (Col. 2) in each class, though this is not as useful as cumulative frequency.
5. The representation of the flow duration curve is improved by plotting the cumulative discharge frequencies on a log-probability scale (Fig. 2). If the daily mean flows are log normally distributed they will plot as a straight line on such a graph. It is common for them to do so in the centre of their range.

Table 1 Derivation of flow frequencies for construction of a flow duration graph

Daily discharge class 1	Frequency 2	Cumulative frequency 3	Percentage cumulative frequency 4
Over 475	3	3	0.21
420-475	5	8	0.44
365-420	5	13	0.89
315-365	8	21	1.44
260-315	25	46	3.15
210-260	36	82	5.61
155-210	71	153	10.47
120-155	82	235	16.08
105-120	52	287	19.64
95-105	42	329	22.52
85-95	50	379	25.94
75-85	58	427	29.91
65-75	83	520	35.59
50-65	105	625	42.78
47-50	72	697	47.71
42-47	75	772	52.84
37-42	73	845	57.84
32-37	84	929	63.59
26-32	103	1032	70.64
21-25	152	1184	81.04
16-21	128	1312	89.80
11-16	141	1453	99.45
Below 11	8	1461	100.00
	Total days = 1461		

From Fig 2 percentage exceedence statistics can easily be derived. For example the 50% flow (the median) is 45 m^3 /sec and flows less than 12 m^3 / sec occurred for 2% of the time

The slope of the flow duration curve indicates the response characteristics of a river. A steeply sloped curve results from very variable discharge usually for small catchments with little storage; those with a flat slope indicate little variation in flow regime.

Comparisons between catchments are simplified by plotting the log of discharge as percentages of the daily mean discharge (i.e. the flow is standardised by mean discharge)(Fig. 3). A common reporting procedure is to show the flow duration curve for the current year compared with the curve over the historic period. Curves may also be generated by month or by season, or one part of a record may be compared with another to illustrate or identify the effects of river regulation on the river regime.

Flow duration curves provide no representation of the chronological sequence. This important attribute, for example the duration of flows below a specified magnitude, must be dealt with in other ways.

4. Fitting of frequency distributions

4.1 General description

The fitting of frequency distributions to time sequences of streamflow data is widespread whether for annual or monthly means or for extreme values of annual maxima or minima. The principle of such fitting is that the parameters of the distribution are estimated from the available sample of data, which is assumed to be representative of the population of such data. These parameters can then be used to generate a theoretical frequency curve from which discharges with given probability of occurrence (exceedence or non-exceedence) can be computed. Generically, the parameters are known as location, scale and shape parameters which are equivalent for the normal distribution to:

- location parameter mean (first moment)
- scale parameter standard deviation (second moment)
- shape parameter skewness (third moment)

Different parameters from mean, standard deviation and skewness are used in other distributions. Frequency distributions for data averaged over long periods such as annual are often normally distributed and can be fitted with a symmetrical normal distribution, using just the mean and standard deviation to define the distribution. Data become increasingly skewed with shorter durations and need a third parameter to define the relationship. Even so, the relationship tends to fit least well at the extremes of the data which are often of greatest interest. This may imply that the chosen frequency distribution does not perfectly represent the population of data and that the resulting estimates may be biased.

Normal or log-normal distributions are recommended for distributions of mean annual flow.

4.2 Frequency distributions of extremes

Theoretical frequency distributions are most commonly applied to extremes of time series, either of floods or droughts. The following series are required:

- maximum of a series: The maximum instantaneous discharge value of an annual series or of a month or season may be selected. All values (peaks) over a specified threshold may also be selected. In addition to instantaneous values maximum daily means may also be used for analysis.

- minimum of a series: With respect to minimum the daily mean or period mean is usually selected rather than an instantaneous value which may be unduly influenced by data error or a short lived regulation effect.

The object of flood frequency analysis is to assess the magnitude of a flood of given probability or return period of occurrence. Return period is the reciprocal of probability and may also be defined as the average interval between floods of a specified magnitude.

A large number of different or related flood frequency distributions have been devised for extreme value analysis. These include:

- Normal and log-normal distributions and 3-parameter log-normal

- Pearson Type III or Gamma distribution
- Log-Pearson Type III
- Extreme Value type I (Gumbel), II, or III and General extreme value (GEV)
- Logistic and General logistic
- Goodrich/Weibull distribution
- Exponential distribution
- Pareto distribution

Different distributions fit best to different individual data sets but if it is assumed that the parent population is of single distribution of all stations, then a regional best distribution may be recommended. A typical graphical output of flood frequency distribution is shown in Fig. 4.

It is clear that there is no single distribution that represents equally the population of annual floods at all stations, and one has to use judgement as to which to use in a particular location depending on experience of flood frequency distributions in the surrounding region and the physical characteristics of the catchment. No recommendation is therefore made here.

A standard statistic which characterises the flood potential of a catchment and has been used as an 'index flood' in regional analysis is the mean annual flood, which is simply the mean of the maximum instantaneous floods in each year. This can be derived from the data or from distribution fitting. An alternative index flood is the median annual maximum, similarly derived. These may be used in reporting of general catchment data.

Flood frequency analysis may be considered a specialist application required for project design and is not a standard part of data processing or validation. Detailed descriptions of the mathematical functions and application procedures are not described here. They can be found in standard mathematical and hydrological texts or in the HYMOS manual.

5. Time series analysis

Time series analysis may be used to test the variability, homogeneity or trend of a streamflow series or simply to give an insight into the characteristics of the series as graphically displayed. The following are described here:

- moving averages
- residual series
- residual mass curves
- balances

5.1 Moving averages

To investigate the long term variability or trends in series, moving average curves are useful. A moving average series Y_i of series X_i is derived as follows:

$$Y_i = \frac{1}{(2M+1)} \sum_{j=i-M}^{j=i+M} X_j$$

where averaging takes place over 2M+1 elements. The original series can be plotted together with the moving average series. An example is shown in Table 2 and Fig. 5

Table 2. Computation of moving averages (M = 1)

I 1	Year 2	Annual runoff (mm) 3	Totals for moving average $=X_{i-1} + X_i + X_{i+1}$ 4	Moving average Y_i = Col 4 / 3 5
1	1970	520		
2	1971	615	520+615+420 = 1555	518.3
3	1972	420	615+420+270 = 1305	435.0
4	1973	270	420+270+305 = 995	331.7
5	1974	305	270+305+380 = 955	318.3
6	1975	380	305+380+705 = 1390	463.3
7	1976	705	380+705+600 = 1685	561.7
8	1977	600	705+600+350 = 1655	551.7
9	1978	350	600+350+550 = 1500	500.0
10	1979	550	350+550+560 = 1460	486.7
11	1980	560	550+560+400 = 1510	503.3
12	1981	400	560+400+520 = 1480	493.3
13	1982	520	400+520+435 = 1355	451.7
14	1983	435	520+435+395 = 1350	450.0
15	1984	395	435+395+290 = 1120	373.3
16	1985	290	395+290+430 = 1115	371.7
17	1986	430	290+430+1020 =1740	580.0
18	1987	1020	430+1020+900 =2350	783.3
19	1988	900		

5.2 Mass curves and residual mass curves

These methods are usually applied to monthly data for the analysis of droughts.

For mass curves, the sequence of cumulative monthly totals are plotted against time. This tends to give a rather unwieldy diagram for long time series and should not be used. Residual mass curves or simply residual series are an alternative procedure and has the advantage of smaller numbers to plot. An example is shown in Fig. 6. With respect to reservoir design (Fig. 7), each flow value in the record is reduced by the mean flow and the accumulated residuals plotted against time. A line such as AB drawn tangential to the peaks of the residual mass curve would represent a residual cumulative constant yield that would require a reservoir of capacity CD to fulfil the yield, starting with the reservoir full at A and ending full at B.

5.3 Run length and run sum characteristics

Related properties of time series which are used in drought analysis are run-length and run-sum. Consider the time series X_1 ....X_n and a constant demand level y as shown in Fig. 8. A negative run occurs when X_t is less than y consecutively during one or more time intervals. Similarly a positive run occurs when X_t is consecutively greater than y. A run can be defined by its length, its sum or its intensity. The means, standard deviation and the maximum of run length and run sum are important characteristics of the time series.

5.4 Storage analysis

Use of sequent peak algorithm can be made for computing water shortage or equivalently the storage requirements without running dry for various draft levels from the reservoir. The procedure used in the software is a computerised variant of the well known graphical Ripple technique. The algorithm considers the following sequence of storages:

$$S_i = S_{i-1} + (X_i - D_x)\ C_f \qquad \text{for } i = 1, 2N \ ; \qquad S_0 = 0$$

where:

X_i = inflow
D_i = $D_L\ m_x$
m_x = average of x_i , i = 1, N
D_L = draft level as a fraction of m_x
C_f = multiplier to convert intensities into volumes (times units per time interval)

The local maximum of S_i larger than the preceding maximum is sought. Let the locations be k2 and k1 respectively with k2 > k1. Then the largest non-negative differences between S_{k1} and S_i , i = k1 …, K2 …, is determined, which is the local range. This procedure is executed for two times the actual series $X_i = X_{N+i}$. In this way initial effects are eliminated.

5.5 Balances

This method is used to check the consistency of one or more series with respect to mass conservation. Water balances are made of discharge series at successive stations along a river or of stations around a junction. The method has already been described in detail with an example in Module 36

6. Regression /relation curves

Regression analysis and relation curves are widely used in validation and for the extension of records by the comparison of the relationship between neighbouring stations. Procedures have been described with respect to climate in Module 17 and more fully with respect to discharge in Modules 37 and 39 and are not discussed further here.

7. Double mass analysis

The technique of double mass analysis is again widely used in validation of all climatic variables and is described in Module 9 for rainfall, Module 17 for climate and Module 36 for discharge. The method is not discussed further here.

8. Series homogeneity tests

Series homogeneity tests with respect to climate are described in Module 17 for the following:

- Student's t test for the stability of the mean
- Wilcoxon-W test on the difference in the means
- Wilcoxon-Mann-Whitney U-test

Series homogeneity tests may also be applied to streamflow but it should also be recognised that inhomogeneity of streamflow records can arise from a variety of sources including:

- data error
- climatic change
- changes in land use in the catchment
- changes in abstractions and river regulation

9. Rainfall runoff simulation

Rainfall runoff simulation for data validation is described in Module 38 with particular reference to the Sacramento model which is used by HYMOS. The uses of such models are much wider than data validation and include the following:

- filling in and extension of discharge series
- generation of discharges from synthetic rainfall
- real time forecasting of flood waves
- determination of the influence of changing landuse on the catchment (urbanisation, afforestation) or the influence of water use (abstractions, dam construction, etc.)

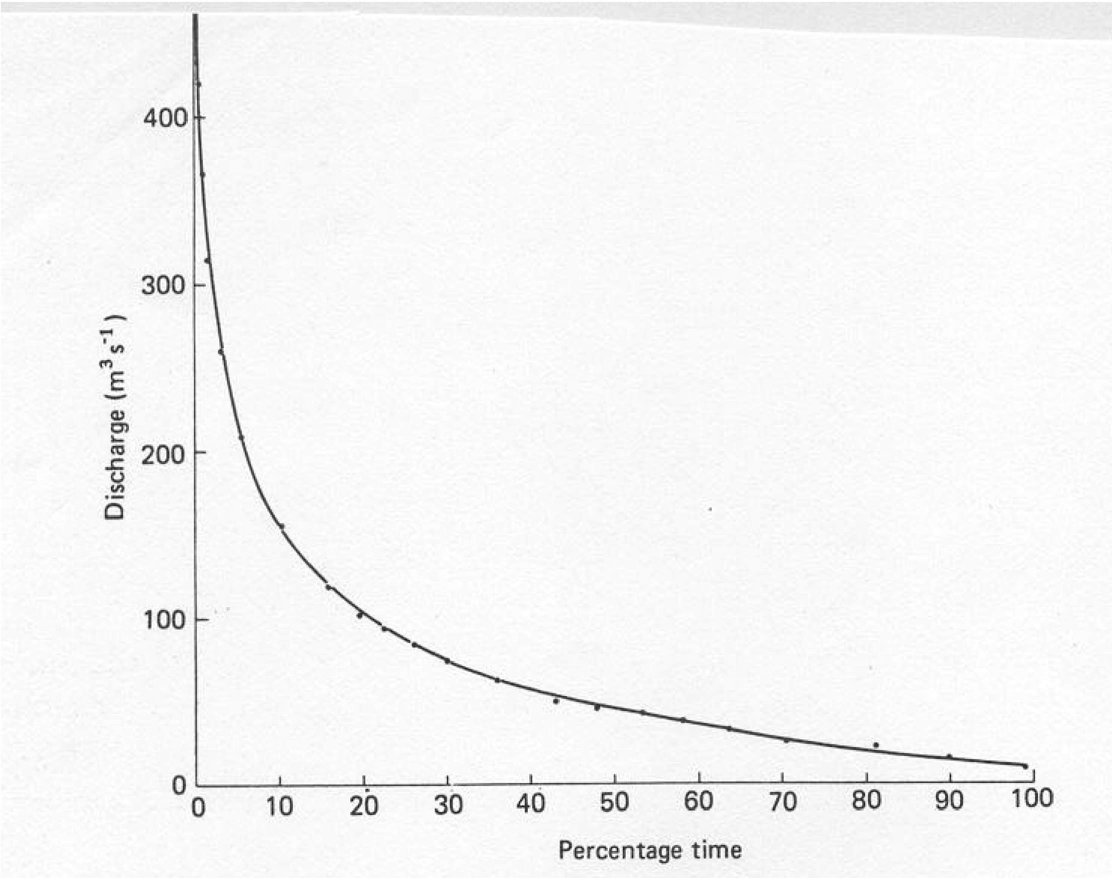

Fig. 1 Flow duration curve plotted on a natural scale

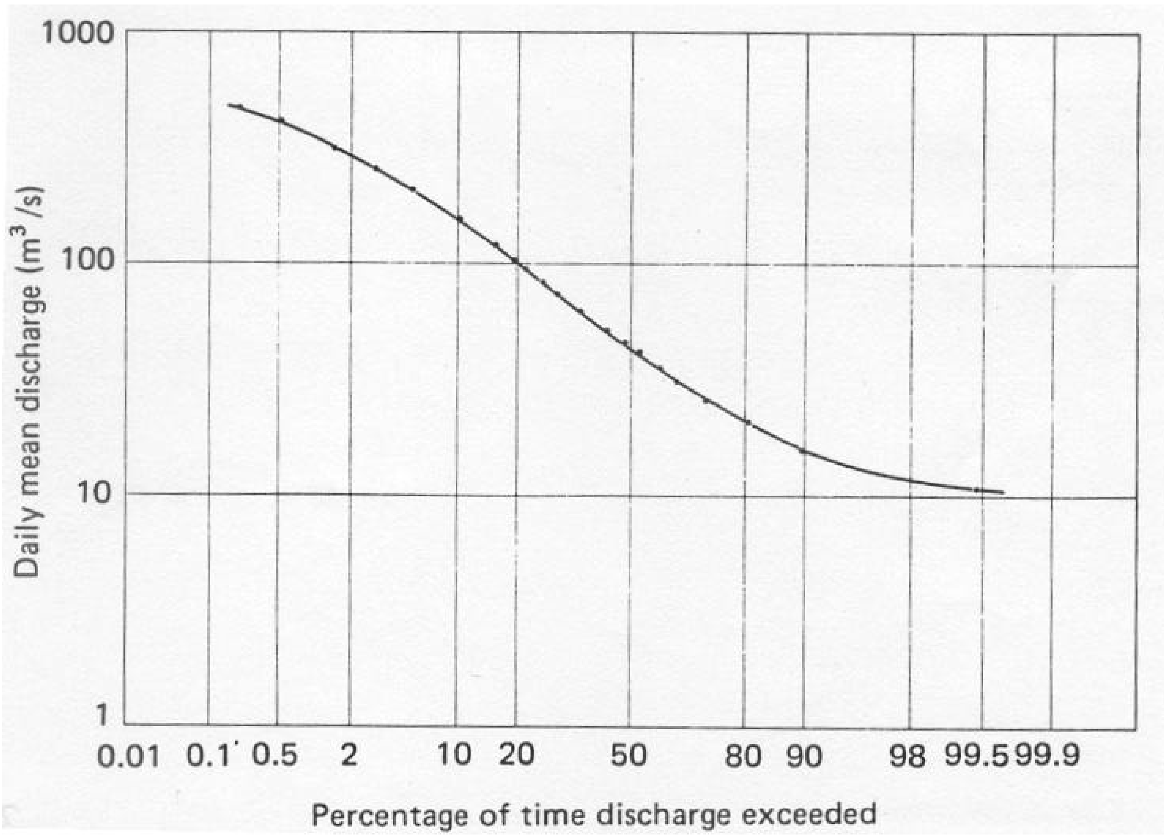

Fig. 2 Flow duration curve plotted on a log-probability scale (after Shaw 1995)

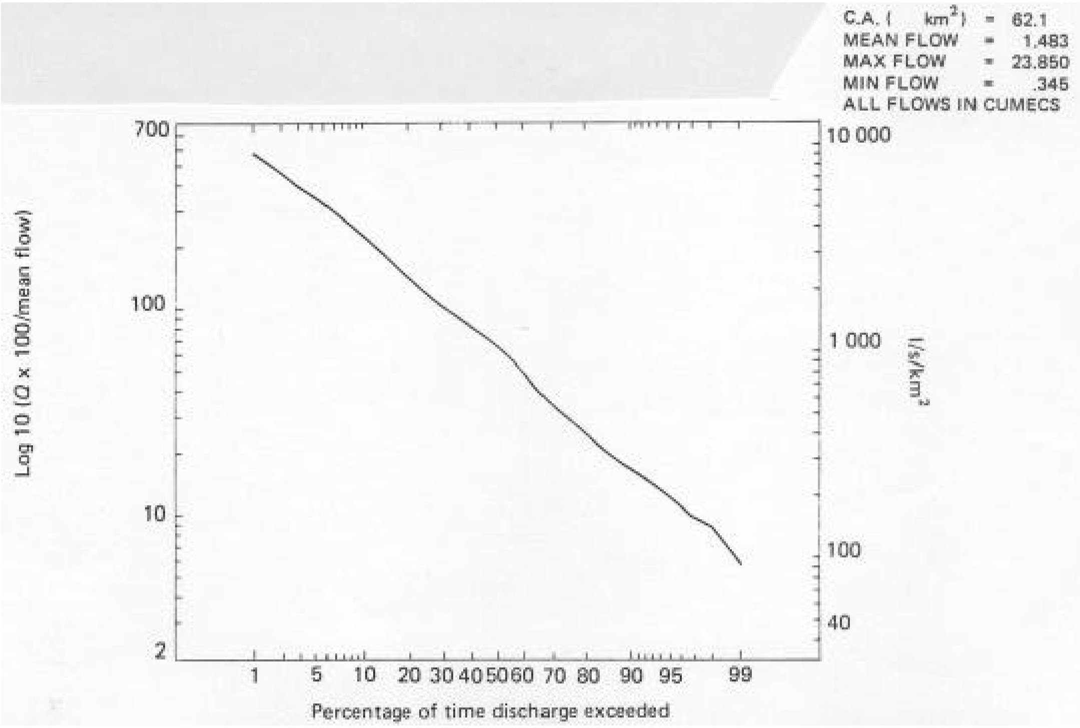

Fig. 3 Flow duration curve standardised by mean flow (after Shaw 1995)

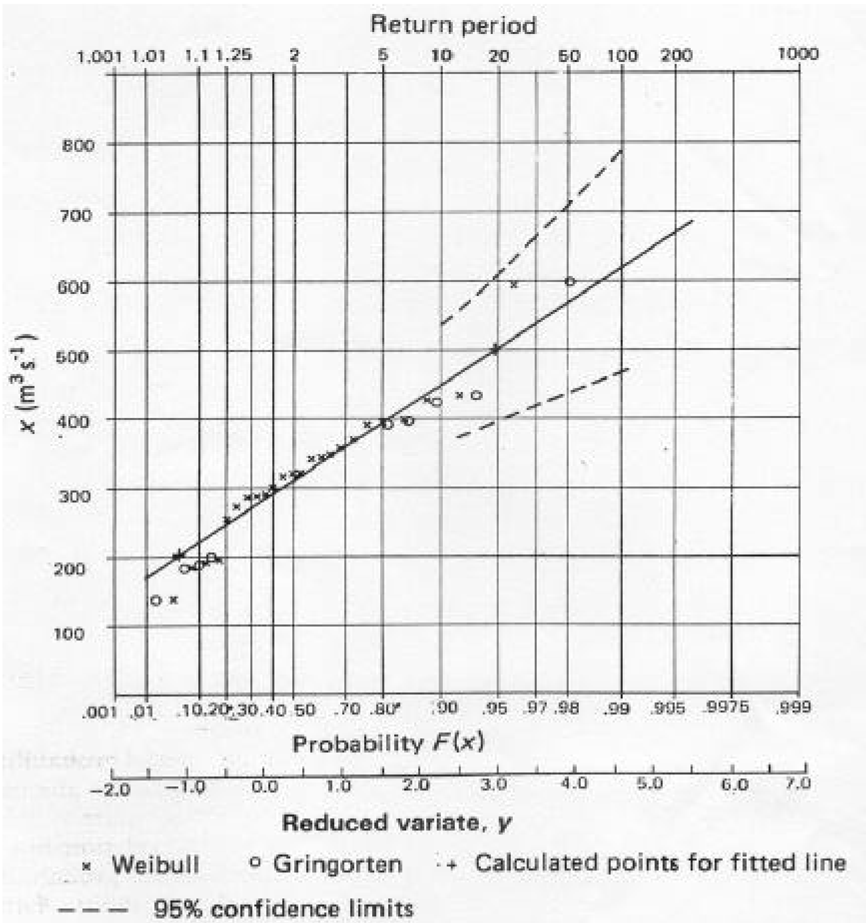

Fig. 4 Flow frequency curve showing discharge plotted against return period (top) and probability (Lower) (after Shaw, 1995)

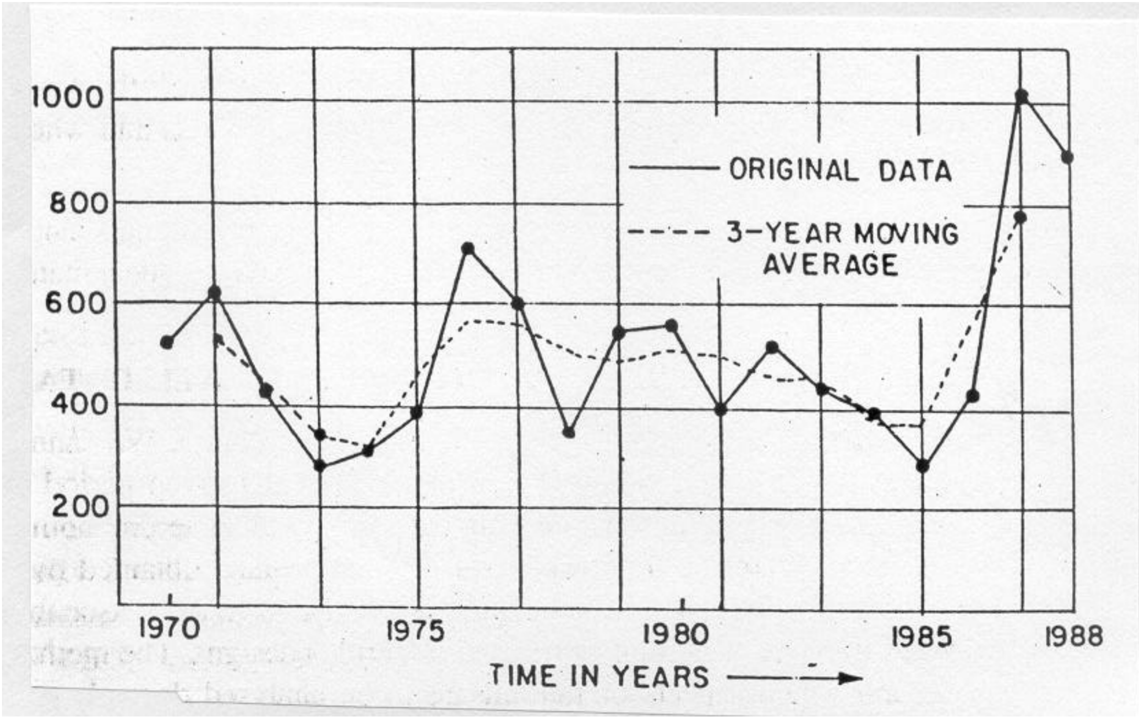

Fig. 5 Moving average of annual runoff

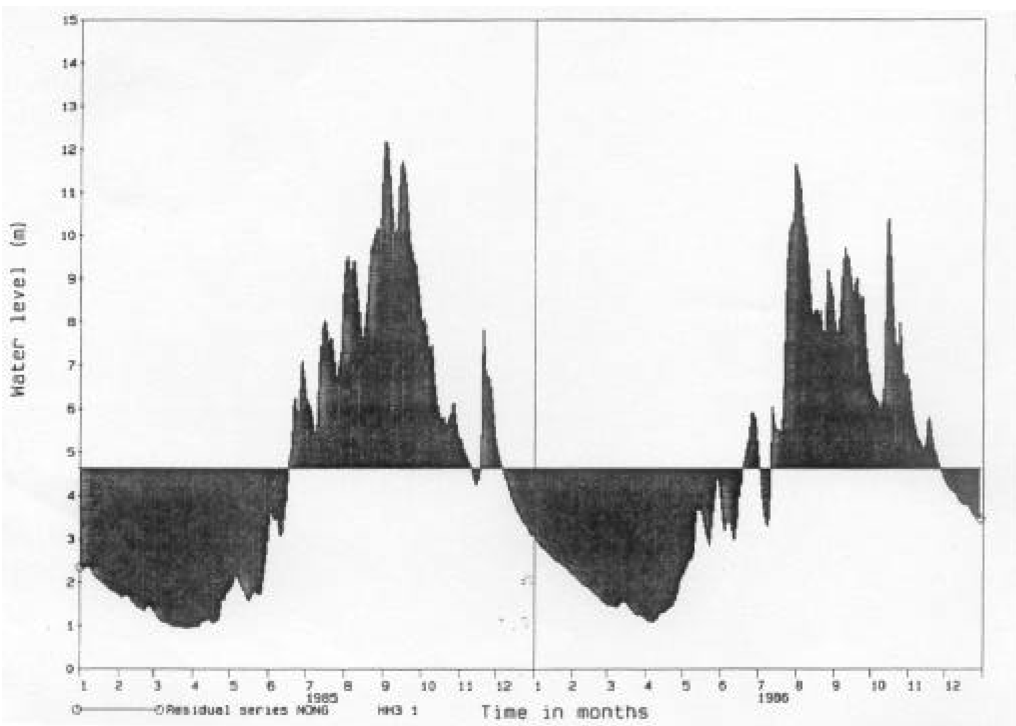

Fig. 6 Example of a residual mass curve

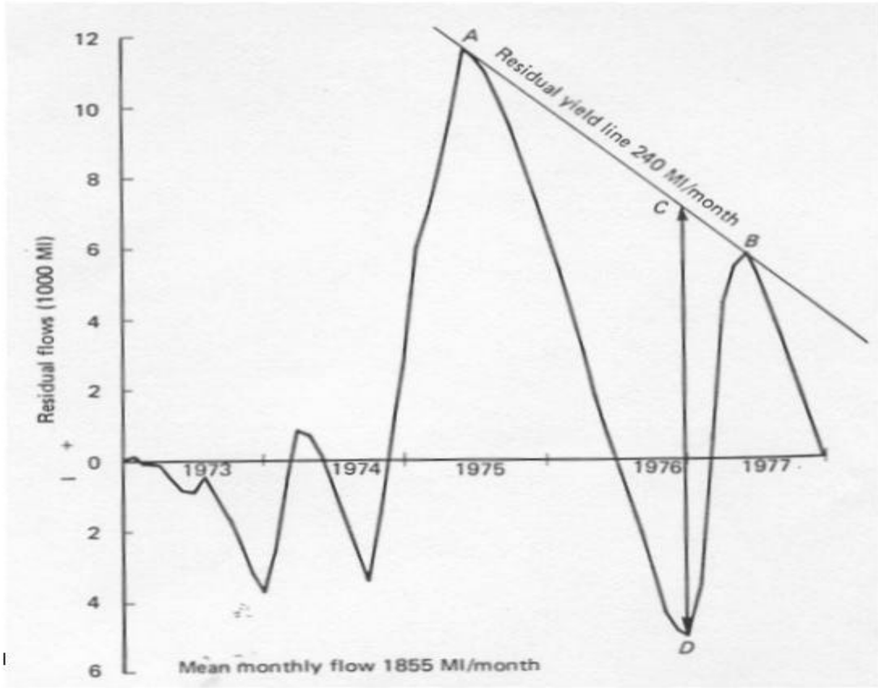

Fig. 7 **Residual mass curve used in drought analysis**

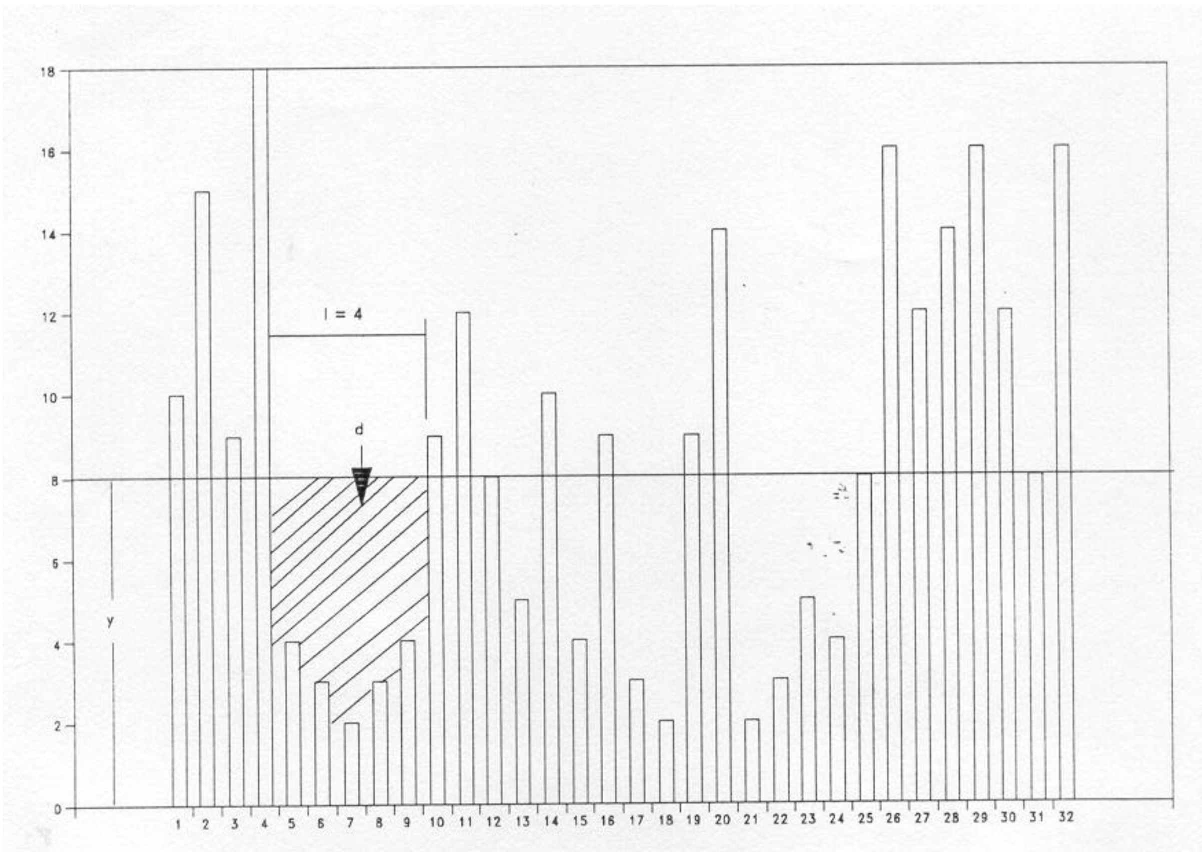

Fig. 8 **Definition diagram of run-length and run-sum**