WORKSHOP

ON

MODELLING OF HYDROLOGIC SYSTEMS

4-8 September, 2000

Regional Flood Frequency Analysis

by

B. Venkatesh

Organised by

Hard Rock Regional Centre
National Institute of Hydrology
Belgaum-590 001 (Karnataka)

REGIONAL FLOOD FREQUENCY ANALYSIS

Venkatesh.B
Scientist 'C'
Hard Rock Regional Centre
National Institute of Hydrology
Belgaum-590 001 (Karnataka)

1.0 Introduction

Flood estimates are required for the design and economic appraisal of a variety of engineering works, including dam spillways, bridges and flood protection works. Flood estimates are also required for the safe operation of flood control structures, for taking emergency measures such as maintenance of flood levees, evacuating the people to safe localities etc. Two main approaches are available for flood estimation viz., deterministic approach and statistical approach. Deterministic approach assumes that input, say, the precipitation is related to the output in a predefined manner and there is no uncertainty involved in arriving at the output, say the discharge. Whereas the statistical approach treats the inter-relationship between processes as governed by theory of statistics. The inter-relationship between processes is established through the measure of correlation. The processes considered may be multivariate or univariate. For example, the rainfall-runoff process may be considered as multivariate while, the consideration of maximum annual peak series fall under univariate process. Flood frequency analysis deals with uni-variate process comprising of maximum peak flow values.

Hydrological processes such as rainfall, floods, droughts etc. are usually investigated by analysing their records of observations. Many characteristics of these processes may not represent definite relationship. For example, if you plot instantaneous peak discharges from each year for a river, a rather erratic graph is obtained. The variation of peak discharge from one year to another can not be explained by fitting a definite relationship, which we call as deterministic relationship. For the purposes of hydrological analysis, the annual peak discharge is then considered to be a random variable. Methods of probability and statistics are employed for analysis of random variables.

Frequency analysis is performed to determine the frequency of the likely occurrence of hydrologic events. This information is needed in the solution of a variety of water resources problems. Some pertinent examples include, design of reservoirs, flood ways, bridges, culverts, highways, levees, urban drainage systems, air field drainage, irrigation system, stream control works, water supply systems and hydro-electric power plants; zoning of flood protection projects; setting of flood insurance premium; drought- mitigation problems etc. Although, the frequency analysis of virtually every component of the hydrologic is required, our emphasis here will be on flood frequency analysis only.

Estimation of flood magnitudes and their frequencies for planning and design of water resources projects have been engaging attention of the engineers the world over since time immemorial. Estimation of magnitude of likely occurrence of floods is go a great importance for solution of a variety of water resources problems such as design of various hydraulic

structures, urban drainage systems, flood plain zoning and economic evaluation of flood protection works etc.

1.1 Objective of this lecture

After going through this lecture note, one will be able to

- define so of the important terms related to statistical and probabilistic methods in hydrology
- compute the statistics representing the measure of location, measure of dispersion and measure of symmetry
- compute the standard errors of some of the important sample statistics such as mean, standard deviation and co-efficient of skewness
- fit some of the theoretical frequency distributions such as Normal, Log normal, Extreme value type-I, Pearson Type-III and Log Pearson type-III distribution to the site data
- estimate the parameters of some of the distributions particularly by the method of moments
- compute standard error as well as confidence limits over the flood estimates
- identify the robust frequency distribution based on some of the goodness of fit tests.

The following assumptions are implicit in frequency analysis in order to have meaningful estimates from flood frequency analysis:

1. The data to be analysed describe random events.
2. It is homogeneous
3. The population parameters can be estimated from the sample data
4. It is of good quality

If the data available for analysis, do not satisfy any of the above listed assumptions then, much reliability can not be attached to the estimates.

For flood frequency analysis, either annual flood series or partial duration series may be used. The requirements with regard to data are that,

1. it should be relevant
2. it should be adequate and,
3. it should be accurate.

In general, an array of annual peak flood series may be considered as a sample of random and independent events. The non-randomness of the peak series will, however, increases the degree of uncertainty in the derived frequency relationship. Various tests are available to check the randomness of the peak flow data. The annual maximum flood series can generally be regarded as consisting of random events as the mean interval of each observed flood peak is 1 year. However, in the case of data used for partial duration series analysis, the independence among the data is doubtful. The peaks are selected in such a way that they constitute a random sample.

The term relevant means that data must deal with problem. For example, if the problem is of duration of flooding then data series should represent the duration of flows in excess of

some critical value. If the problem is of interior drainage of an area then data series must consist of the volume of water above a particular the should.

The term adequate primarily refers to length of data. The length of data primarily depends upon variability of data and hence there is no guideline for the length of data to be used for frequency analysis.

The term accurate refers primarily to the homogeneity of data and accuracy of the discharge figures. The data used for analysis should not have any effect of man made changes. Changes in the stage discharge relationship may render stage records non-homogeneous and unsuitable for frequency analysis. It is therefore preferable to work with discharge and if stage frequencies are required then most recent rating curve is used. Watershed history and flood record should be carefully examined to ensure that no major watershed changes have occurred during the period of record. Only records, which represent relatively constant watershed conditions should be used for frequency analysis.

2.0 Fundamental Terms used in the frequency analysis

In this section, some of the important statistical terms frequently used in the frequency analysis are defined

(a) Population

A population is a collection of persons or objects, e.g., (I) the pupil in a school, the workers in a factory, the people in a country, (ii) motor cars produced in a factory. Each unit of the population has many different possible attributes associated with it. These attributes might be (I) height, volume or weight which are measurable on a scale, or (ii) colour, condition which may not be numerically measurable.

(b) Sample data

Sample data are available data from the observation of an event.

(c) Random events

Events whose occurrence is not influenced by the occurrence of the same event earlier.

(d) Probability density function

Probability density function (P.D.F) is the probability of occurrence of an event

(e) Cumulative density function

Cumulative density function (C.D.F) is the probability of occurrence of all the events that are equal to or less than an event.

(f) Probability paper

A probability paper is a special graph paper on which the ordinate usually represents the magnitude of the variate and the abscissa represents the probability P, or the return period T. The ordinate and abscissa scales are so designed that the distribution plots more nearly a straight line permitting better definition of the upper and lower parts of the frequency curve. The probability paper is used to linearise the distribution so that data to be fitted appear close to the straight line. For Example, the extreme value and the log normal probability papers are used for linearization of the extreme value and log-normal distribution.

(g) Plotting Position

Determining the probability to assign a data point is commonly referred to as determining its plotting position.

(h) Return period

Return period (T) or recurrence interval is the time that elapses on an average between two events that equal or exceed a particular level. For example, T year flood will be equalled or exceeded on an average once in T years.

(I) Probability of Exceedence

If a coin is tossed once, on the average, its head will appear once in two time, or probability of exceedance (P) is P=1/2=0.5. The reciprocal of probability of occurrence is termed as return period, T.

$P = 1/T$

Thus a storm that has been exceeded on the average once inn 10 years has a probability of exceedance in any one year P=1/10 or 0.1

(j) Probability of Non-Exceedence

The probaiblity that the storm will not occur is termed as the probability of non-exceedance (P`).

$P` = 1-P = 1-1/T$

(k). Peak Annual Discharge

The peak annual discharge is defined as the maximum instantaneous volumetric rate of discharge during a year

(l) Annual flood series

The annual flood series is the sequence of the peak annual discharge for each year of the record.

(m) Partial Duration Flood Series

The partial series consists of all recorded floods above a particular threshold regardless of the number of such floods occurring each year

(n) Design Flood

Design flood is the maximum flood which any structure can safely pass. It is the adopted flood to control the design of a structure.

3.0 Sample statistics

In any analysis of statistical data in general and of hydrolytic data in particular, certain calculations are usually made in order to determine some of the basic properties inherent in the data. For instance, the sample mean and variance are two statistics defining the most important characteristics of a given set of statistical data. In general sample statistics provide the basic information about the variability of a given data set. The most useful sample statistics measure the following characteristics,

- the central tendency or value around which all other values are clustered.
- the spread of the sample values around mean
- the asymmetry or skewness of the frequency distribution and
- the flatness of the frequency distribution

These statistical properties are determined by sample statistics are described below:

3.1 Measure of Central Tendency or Measure of Location

In statistics various measures of location are described. One of the important measure of central tendency is mean

(l) Mean

The sample mean is the measure of central tendency of a given data set. If X_1, X_2, X_3, ... X_n represent a sequence of observations, the mean of the sequence is given by:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

3.2 Measure of Dispersion or Variation

(I) Standard Deviation

The unbiased estimate of population standard deviation (S) from the sample is given as:

$$S = \left[\frac{1}{(N-1)} \sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2 \right]^{1/2}$$

(ii) Variance

Variance represents dispersion about the mean. Mathematically for sample it is expressed as:

$$\text{Variance} = S^2 = \frac{\sum_{i-1}^{N} \left(X_i - \bar{X} \right)^2}{N}$$

(iii) The coefficient of Variation

The coefficient of variation C is a dimensionless dispersion parameter and is equal to the ratio of the standard deviation and the mean

$$C_v = S / \bar{X}$$

This coefficients extensively used in hydrology particularly as a regionalisation parameter.

The range and mean deviation have the same units (dimension) as the original data. The variance has the square of the units of the original data and hence can not be directly compared with the data. Therefore, the standard deviations used because its dimensions are that of the data.

In many samples of hydrological data, especially in flood hydrology the largest value is very much larger than the second largest. Therefore the range R might not be a good indicator of the scatter inherent in the data as a whole.

The mean deviation is a good measure of spread but can not be handled easily in mathematical statistics because of the absolute value sing while the same applies to the interquartile range. The variance is more easily handled mathematically and holds a prominent place. The interquartile range is easy to evaluate but is very difficult from a mathematical point of view and hence is not much used even though it is quite good at describing spread.

3.3 Measures of Symmetry

If the data re exactly symmetrically displaced about the mean than the measure of symmetry should be zero. If the data to the right of the mean (large) are more spread out from the mean than those on the left then by convention, the asymmetry is positive and vice versa for negative asymmetry.

(I) Skewness Coefficient

The skewness coefficient or coefficient of skewness represents a non-dimensional measure the asymmetry of the frequency distribution of the data. An unbiased estimate of the coefficient is given by:

$$C_s = \frac{N\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^3}{(N-1)(N-2)S^3}$$

The skewness coefficient has an important meaning since it gives indication of the symmetry of the distribution of the data. Symmetrical frequency distributions have very small or negligible sample skewness coefficients C_S, while asymmetrical frequency distributions have either positive or negative coefficients. often a small value of C_S, indicate that the frequency distribution of the sample may be approximated by the normal distribution function since C_S = 0 for this function.

Note that because of the third central moment has dimension equal to the cube of the data, it is not of direct use. It also depends on the units of the original data. The co-efficient of skewness does not have this disadvantage and is therefore preferred. The interquartile measure of symmetry (I_{AS}) is not dimensionless

3.4 Measures of Peakdeness or Flatness

The kurtosis coefficient measures the peakedness or the flatness of the frequency distribution near its centre. An unbiased estimate of this coefficient is given by

$$C_k = \frac{N^2\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^4}{(N-1)(N-2)(N-3)S^4}$$

A related coefficient called the excess coefficient denoted by E is defined by

$E = C_K - 3$

Positive values of E indicate that a frequency distribution is more peaked around its centre than the normal distribution. Frequency distribution is known as LEPTOKURTIC, the

negative values of E indicate that a given frequency distribution is more flat around its centre than the normal. Frequency distribution is known as PLATYPURTIC.

Normal distribution is said to be MESOKURTIC. Both kurtosis and excess coefficient are seldom used in statistical hydrology.

4.0 Probability Distribution

A distribution is an attribute of a statistical population. If each element of a population has a value of X then the distribution describes the constitution of the population as seen through its X values. It tells whether they are in general very large or very small, that is their location on the axis. It tells whether they are bunched together or spread out and whether they are symmetrically disposed on the X axis or not. These three things are described by the mean, standard deviation and skewness.

Distribution also tells the relative frequency or proportion of various X values in the population in the same way that a histogram gives that information about the sample. These relative frequencies are also probabilities and hence the distribution tells us the probability, $P_r(X < x)$, that the X value on an element drawn randomly from the population would be less than a particular value x. Knowing $P_r(X < x)$ for all x values, the laws of probability may then be used to deduce the probability of any proposition about the behaviour of a random sample of X values drawn from the population.

When the population is sufficiently large the histogram of its X values can be made with very small class intervals and the histogram can be replaced by a smooth curve, the area enclosed by any two vertical ordinates being the relative frequency or probability of x values between those ordinates.

Because of this probability interpretation, a relative frequency distribution is also called a probability distribution and the curve describing it is called a probability density function (p.d.f) whose cumulative function is called as the cumulative distribution function (c.d.f).

In flood frequency analysis, the sample data is used to fit probability distribution which in turn is used to extrapolate from recorded events to design events either graphically or analytically by estimating the parameters of the distribution. A large number of peak flow distribution are available in literature.

A large number of frequency distributions are available in literature. Some of the commonly used probability distribution in the area of hydrology and water resources are: log-normal (two parameters and three parameters), Extreme value type-I (Gumbel or EVI), Pearson type-III, Log-Pearson type-III, General extreme value, Gamma and Exponential distributions. The probability density functions (P.D.F) and Cumulative density function (C.D.F) and other properties of these distribution are given in Table.1. Here the normal, log-normal, extreme value type-I, Pearson type-III and log-Pearson type-III distributions have been discussed. The probability density functions (P.D.F), cumulative density functions (C.D.F) and other properties of these distribution are given in table.1.

4.1 Normal Distribution

The normal distribution is one of the important distribution in statistical hydrology. This is a bell shaped symmetrical distribution having coefficient of skewness equal to zero. The normal distribution enjoys unique position in the field of statistics due to central limit theorem. This theorem states that under certain very-broad conditions, the distribution of sum of random variables tends to a normal distribution irrespective of the distribution of random variables, as the number of terms in the sum increases. The probability density function (PDF) and Cumulative density function (CDF) of the distribution are given in table.1.

4.2 Log-normal Distribution

The causative factors for may hydrologic variables act multiplicatively rather than additively and so the logarithms of these variables which are the product of these causative factors follow the normal distribution.

If $Y= \log_e(X)$ follows normal distribution, the X is said to follow log normal distribution. If the variable X has lower bound X_0, different from zero and the variable $Y= \log_e(X-X_0)$ follows normal distribution then X is log normally distributed with three parameters. The PDF and DCF of the distribution area given in Table.1.

4.3 Pearson type-III Distribution

Pearson type-III distribution is a three-parameter distribution. This is also known s Gamma distribution with three parameters. The PDF and CDF of the distribution are given in Table.1.

4.4 Exponential Distribution

Exponential distribution is a special case of Pearson type-III distribution when shape parameter $\gamma=1$, The PDF and CDF are given in Table.1.

4.5 Gumbel Extreme Value (Type-I) distribution (EVI)

One of the most commonly used distributions in flood frequency analysis is the double exponential distribution (known as Gumbel distribution or extreme value type-I or Gumbel EVI distribution). The PDF and CDF of the distribution are given in Table.1.

4.6 Log-Pearson type-III Distribution (LP3)

If $Y= \log_e(X)$ follows Pearson Type-III distribution then X is said to follow log-Pearson Type-III distribution. In 1967, the US Water Resources Council recommended that the log-Pearson type-III distribution should be adopted as the standard flood frequency distribution by all US federal government agencies. The PDF and CDF of the distribution are given in Table.1.

4.7 Extreme Value Distribution

Just as there is a family of Pearson type-III distributions, each member being characterised by a value of γ , there is also a family of EV distributions, each member of which is characterised by the value of a parameter denoted by k. The family can be divided into three lasses, corresponding to different ranges of k values. The three classes are referred as Fisher Tippett type-I, type-II and Type-III. They are also known as EV-I, EV-II, and EV-III distributions. In practice, k value lie in the range -0.6 to + 0.6. For EV-I distribution, k is zero and coefficient of skewness is equal to 1.139. For EV-II distribution, the value of k is -ve and skewness is greater than 1.139. for EV-III distribution the value of k is +ve and coefficient of skewness is less than 1.139. EV-I and EV-II distributions are known s Gumbel and Frechet distributions respectively.

4.8 General Extreme Value Distribution

The General Extreme Value Distribution is a generalised 3 parameter extreme value distribution proposed by Jenkinson (1955). The theory and the applicability of GEV are reviewed in the British flood studies report (NERC, 1975). The Probability Density Function (PDF) and cumulative density function (CDF) for GEV distribution is given below.

$$F(X) = e^{-\left(1-k\left((x-\mu)/\alpha\right)\right)^{\frac{1}{k}}}$$

where μ,α and k are location, scale and shape factors of GEV distribution respectively.

4.9 Wakeby distribution

Houghton in 1978 introduced a five-parameter distribution, which have named as wakeby, as a substitute for traditional F distributions. The mathematical structure of Wakeby distribution is given as below

$$X = m + a\left[1-\left(1-F(X)\right)^{b}\right] - c\left[1-\left(1-F(X)\right)^{-d}\right]$$

where, F = F(X) = P (Q$\leq$ Q_T), is uniform (0,1) variate. The equation is so written that the parameters a,b,c, and d are always positive and m is sometimes positive

Wakeby distribution is potentially useful to flood frequency analysis in particular and to flow frequency analysis in general for several reasons. First, it offers a simple explanation of the condition of separation. Second, it is characterised by five parameters suggesting better capability of fitting data than that of distributions characterised by fewer parameters. Moreover, because the left tail of the distribution is more strongly influenced by parameter **b** and the right tail by parameter **d,** the distribution can accommodate various types of flows ranging from low flows to floods.

5.0 Method of Parameter Estimation

Using the statistical methods does the estimation of the parameters of distributions when the stream flow records are available at a site under consideration. These methods will be discussed with reference to the annual maximum series throughout because this series is most often used in practice. Since no deductive method exist for deciding on the form of a distribution, the data would follow, the suitability of each possible candidate distribution must be examined. Each distribution is considered in turn to be the correct distribution and some numerical index is calculated expressing the agreement or lack of fit between the assumption and the information about the distribution appropriate in flow records. In this case the information in the flow record is that contained in the appropriate series. First the parameter of the distribution must be estimated from the data. A numerical measure of the difference between the two distributions, fitted theoretical and observed, should be used to make a decision between different forms of distributions. This in choosing between the different form of distribution at a given site of river having historical data, there are two steps:

1. estimation of parameters, and
2, calculation of the numerical measure of agreement

In this section, only method of parameter estimation is discussed.

There are four well know parameters estimation techniques, viz;

1. Graphical
2. Least square
3. Method of Moments, and
4. Method of maximum likelihood

The recent literature provides some more methods of parameter estimation which are being used for wide application in flood frequency analysis, they are,

1. Probability Weighted Moments (PWM)
2. L- Moments

5.1 Graphical Method

In graphical estimation, the variate under consideration is regard as a function of standardised or reduced variate of a known distribution. The sample data is plotted as series of N discrete points on an ordinary graph paper with abscissa being the reduced variate of the probability distribution under consideration and the ordinate being the variate. The variates are plotted against the corresponding probability or reduced variate or return period determined using the appropriate plotting position formulae. The plotted points on the graph paper represent the sample distribution and a line drawn through these is considered as an estimate of the population relation. Then this straight line is projected to arrive at the flood magnitude of desired return period.

In graphical estimation the line is subjectively place and could vary with analyst. This subjectivity is regarded as a major drawback by hydrologists.

5.2 Least Squares Estimation

In the least square estimation technique a simple linear regression equation is fitted between the variate under consideration and the corresponding frequency factor K. The form of Chow's general equation is used as the liner regression equation and it is written as;

$$X_i = a + b K_i + \varepsilon_i$$

X_i = is i^{th} variate
K_i = the frequency factor corresponding to the i^{th} variate.
ε_I = error term with mean = 0 and standard deviation $\sigma\varepsilon$

It is not correct to interpret a and b as mean and standard deviation of the X_i series as these parameters are estimated in the least squares sense and as such they can never become equal to mean and standard deviation of the sample data.

This method has not been accepted as a standard method in practice as it involves the use of plotting position formulae to determine the frequency factor K_i and due to the assumption that the error variance $\sigma^2 \varepsilon$ remain same for all observations. The defect due to the former assumption could be eliminated by using the appropriate plotting position method. However, the later assumption makes the method more defective as the higher events recorded have more error variance than the recorded lower events. All these assumptions affect the correct parameter estimation of a and b.

5.3 Method of Moments

The method of moments makes use of the fact that if all the moments of a distribution are known then everything about the distribution is know. For all the distribution in common usage, four moments or fewer are sufficient to specify all the moments. For instance, two moments, the first together with any moment of even order are sufficient to specify all the moments of the normal distribution and therefore the entire distribution. Similarly, in the Gumbel EV type-I distribution, the first two moments are sufficient to specify all the moments and hence the distribution. In these cases the number of moments needed to specify all the moments and hence the distribution equals the number of parameters.

The method of moments estimation is dependent on the assumption that the distribution of variate values in the sample is representative of the population distribution. Therefore, a representation of the former provides an estimate for the later. Given that the form of the distribution is known or assumed, the distribution which the sample follows is specified by its first two or three moments calculated from the data.

Having estimated the mean and the standard deviation for two parameter distributions, the magnitude of the required return period flood is computed using Chow's general frequency equation as;

$$X_T = \mu + K_T \sigma$$

in which,

X_T = the magnitude of flood at required return period T
K_T = the frequency factor corresponding to T
μ and σ = mean and standard deviation of the population, which would be replaced by the sample statistics.

5.4 Method of Maximum Likelihood

The principle of Maximum likelihood states that for a distribution with probability density function $p(x_i\,\alpha_i\beta_i)$, where α_i,β_i are the distribution parameters to estimated, the probability of obtaining a given value of x_i is proportional to $p(x_i\,\alpha_i\beta_i)$ and the joint probability, L, of obtaining a sample of N values $x_1, x_2, \ldots\ldots x_n$ is proportional to the product.

$$L = \prod_{i=1}^{n} p(x_i\,\alpha_i\beta_i)$$

This is called the likelihood. The method of maximum likelihood is to estimate $\alpha_i,\beta_i\ldots\ldots$ such that L is maximised. This is obtained by partially differentiation L with respect to each of the parameters and equating to zero. Frequently ln(L) is used instead of L to simplify computations.

5.5 Method of Probability Weighted Moments

Probability weighted moments, a generalisation of the usual moments of a probability distributions, were introduced by Greenwood et al (1979). There are several distributions whose parameters can be conveniently estimated from their probability weighted moments. Landwehr et al (1979) investigated the small sample properties of probability weighted moment estimators of parameters and quantiles for the Gumbel distribution and found them superior in many respects to the conventional moment and maximum -likelihood estimators. The PWM's are defined by the general expression as given below

$$M_{i,j,k} = E\,X^i F(X)^j (1-F(X))^k, \quad \text{i,j,k are real}$$

Note that, $M_{1,0,0}$ are ordinary moments and equals the population first moment or mean. $M_{1,0,0}$ can be estimated by the sample mean while $M_{1,0,1}$ is estimated by a weighted linear function of the ranked sample values

$$x_1 \le x_2 \le \ldots\ldots x_i \le \ldots\ldots x_n$$

Thus

$$\hat{M}_0 = \bar{Q}$$

$$\hat{M}_k = \sum_{i=1} w_i^k x(i)/N$$

where W_i is weighting factor corresponding to (1-F(x)) in the definition of $M_{1,0,1}$. The unbiased estimation for $M_{(1)}$ is obtained by using

$$w_i^k = (1 - F_i)^k \quad \text{with } F_i = \frac{i - 0.35}{N}$$

use of this introduces some bias but this is negligible for quantile estimation

5.6 L-Moments

The L-moments are estimated by linear combinations of order statistics (hence the prefix L). Theoretically, L-moments are able to characterize a wider range of distributions than conventional moments. Practically, they are less subjected to bias in estimation, and they approximate their asymptotic normal distribution more closely. The main advantage of L-moments over conventional moments is that L-moments suffer less from the effects of sampling variability; they are more robust to outliers in the data. A unified approach to the use of order statistics for the statistical analysis of univariate probability distributions, based on L-momens, has been developed by Hosking (1990).

L-moments for Data Samples

Probability weighted moments, defined by J.A. Greenwood et al. are precursors of L-moments. Sample probability weighted moments, computed from data values $x_1, x_2, \ldots\ldots\ldots\ldots x_n$, arranged in increasing order, are given by

$$b_0 = n^{-1} \sum_{j=1}^{n} x_j$$

$$b_r = n^{-1} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots\ldots(j-r)}{(n-1)(n-2)\ldots\ldots(n-r)} X_j$$

L-moments are certain liner combinations of probability weighted moments that have simple interpretations as measures of the location, dispersion and shape of the data sample. The first few L-moments are defined by

$$l_1 = b_0$$
$$l_2 = 2b_1 - b_0$$
$$l_3 = 6b_2 - 6b_1 + b_0$$
$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0$$

(the coefficients are those of the "shifted Legendre polynomials")

The first L-moment is the sample mean, a measure of location. The second L-moments is (a multiple of) Gini's mean difference statistic, a measure of the dispersion of the data values about their mean.

By dividing the higher-order L-moments by the dispersion measure, we obtain the L-moments ratios,

$t_r = l_r/l_2$

These are dimensionless quantities, independent of the units of measurements of the data. t_3 is a measure of skewness and t_4 is a measure of kurtosis-- these are respectively the L-skewness and L-kurtosis. They take values between -1 and +1 (exception: some even-order L-moment ratios computed for very small samples can be less than -1).

The L-moment analogue of the coefficient of variation (standard deviation divided by the mean), is the L-CV; defined by

$t = l_2/l_1$, it takes the values between 0 and 1.

6.0 Tests of Goodness of Fit

The validity of probability distribution function proposed to fit the empirical frequency distribution of a given sample may be tested graphically or by analytical method. Graphical approaches are usually based on comparing visually the probability density function with the corresponding empirical density function of the sample under consideration. In other words, model CDF is compared with empirical CDF. often these CDF graphs are made on specially designed paper such that the model CDF plots as a straight line. Often, graphical approaches for judging how good a model is, are quite subjective. A number of analytical tests have been proposed for testing the goodness of fit of proposed models. Three of these test are presented subsequently.

6.1 Chi-Square Test

The Chi-square goodness of fit test in one of the most commonly used tests for testing the goodness of fit of probability distribution functions to empirical frequency distributions.

Assume that it is desired to test the goodness of fit of a probability model, with density function $f_y(y, \theta')$ and CDF $f_y(y, \theta')$, to the empirical distribution of a sample $y_1 \ldots\ldots\ldots y_n$, where N is the sample size, $\theta' = (\theta_1' \ldots\ldots\ldots \theta_n')$ is the set of parameters estimated form the sample, and p= number of parameters. The probability space (100% probability) is divided into m intervals (class intervals) with probability $p_1 \ldots\ldots\ldots p_m$ in each class intervals such that $p_1 + \ldots\ldots\ldots + p_m = 1$. If such probabilities are the same then p_j=1/m and the m cumulative probabilities are $\frac{1}{m}, \frac{2}{m}, \frac{3}{m} \ldots\ldots 1$. For the first (m-1) cumulative probabilities the corresponding values of y are determined from the model. Let $y_1', \ldots\ldots\ldots y_{m-1}'$ be the set of y's corresponding to probabilities $\frac{1}{m}, \ldots\ldots\ldots, \frac{1}{m-1}$ which are also the upper class limits for the first m-1 class intervals.

Now let the sample $y_1 \ldots\ldots\ldots y_n$ be arranged in increasing order of magnitude and let N, be the number of sample values that fall in the j-th class interval for j=1,.....m. Since the probability of p_j for the j-th class interval, the expected number of sample value that would fall in the j-th interval is equal to p_j. N. Then, it may be shown that the (test) statistic

$$C = \sum_{j=1}^{m} \frac{\left(N_j - N P_j\right)^2}{N P_j}$$

is approximately Chi-square distributed with m-1-p degrees of freedom. Since p_j=1/m, then

$$C = \frac{m}{N} \sum_{j=1}^{m} N_j^2 - N$$

The number of classes are selected in such a way that theoretical frequency of each class is not less than 5. The number of classes should not be less than 6 and more than 20 though these rules donot have theoretical basis. The length of class intervals should be selected in such a way that the main characteristic features of the observed distribution are enphasized and chance variations are obscured.

The above equations may be sued to test the hypothesis of good fit of a given model to the empirical frequency distribution of sample by comparing the computed test statistic C with the tabulated Chi-square statistic $^{2}_{1-\alpha,(m-1-p)}$ in which α is the significance level and (m-1-p) is the number of degrees of freedom. Then, the hypothesis of good for is accepted at the γ= 1-α, probability level if $C \leq \aleph^2_{1-\alpha,(m-1-p)}$

The critical values of ^{2} for different probability levels and degrees of freedom are read form the statistical tables.

6.2 Kolmogorov- Smirnov Test

This is a distribution free test widely used in statistical hydrology. It is based on the maximum difference between the cumulative empirical distribution $F_e(y)$ and the cumulative probability distribution being fitted $F_{y'}(y:\theta')$. Consider the statistic

$$D = \max_{i=1}^{N} \left(F_e(y) - F_{y'}(y_i, \theta') \right)$$

Where $F_e(y)$ and $F_{y'}(y:\theta')$ represent the empirical and model cumulative distribution, respectively, corresponding to the observation value Y_i, I=1,....., N which has been arranged in increasing (or decreasing) order of magnitude, N= Sample size and θ' is the parameter set of the model estimated from the given sample. In the Kolmogorov-Simirnov test, the empirical CDF $F_e(y)$ is based on the plotting position i/N, although in practice the plotting position i/(N+!) is often used.

The goodness of fit test of the selected probability model to the empirical distribution is accepted if

$$D \leq d_{\alpha}(N)$$

Where, d_{α} (N) is the Kolmogorov-Smirno statistic corresponding to the sample size N and confidence level γ= 1-α,. The statistic d_{α} (N) is read from the table.

6.3 D-Index Method

In order to compare the relative fit of different distributions to hydrological data, USWRC (United States Water Resources Council) has suggested the following procedure:

Weibull plotting position formula estimates the Probability of exceedance of observation.

$$P(X \geq x) = m/(N+1)$$

where, P is the probability of exceedance
m is the rank of the flood values arranged in the descending order of magnitude, and
N is the number of observations.

Flood peaks are estimated for a specific series of recurrence intervals viz, 2,5,10,15,20 and 30 years. For each recurrence interval, the historical value is obtained by interpolation in terms of recurrence interval between the two floods of record of adjacent recurrence intervals. The discharge corresponding to the same recurrence interval is also calculated on the basis of fitted distribution. The D index for comparison purposes of the fit of different distributions is given as

$$D - Index = \frac{1}{\bar{x}} - \sum_{i=1}^{6} ABS\left(x_{i,obs} - x_{i,com}\right.$$

in which $\bar{x}$ is the mean of the observed series.

Often instead of using flood values corresponding to recurrence intervals of 2,5,10,15,20 and 30 years, highest six observations/flood values are used, as the aim is to see the fit of the distribution in the upper tail region. The distribution which gives the minimum D-Index is considered as the best fit distribution.

7.0 Case Studies

The above described techniques are used to carry out flood frequency studies using the data of Krishna basin and Wainganga basin. Here in these case studies, the PWM based Gumbel distribution, General Extreme Value distribution and Wakeby distributions are used. The descriptive goodness of fit criteria is used to get the best fitted distribution for each of the basin. The regional frequency formulae for each of these basin are given in table.2.

Table.2. Regional frequency formulae for different classifications considered in Krishna Basin

Region	Method	Regional flood formula
Medium Catchment	Index-flood	$Q_T = (37.93+19.73(-\ln(-\ln(1-1/T))))A^{0.39}$
	EV-I Distribution	$Q_T = (36.94+21.45(-\ln(-\ln(1-1/T))))A^{0.39}$
	GEV Distribution	$Q_T=(117.43(-\ln(1-1/T))^{-0.155} -81.87)A^{0.39}$
	Wakeby Distribution	$Q_T=(2.219+21.65(1-(1.F)^{16.095} - 457.7(1-(1-F)^{-0.055}))\ A^{0.39}$
Large Catchment	Index-flood	$Q_T = (4.136+1.809(-\ln(-\ln(1-1/T))))A^{0.64}$
	EV-I Distribution	$Q_T = (4.032+1.974(-\ln(-\ln(1-1/T))))A^{0.64}$
	GEV Distribution	$Q_T=(51.56(-\ln(1-1/T))^{-0.037} -47.56)A^{0.64}$
	Wakeby Distribution	$Q_T=(1.318+1.695(1-(1.F)^{5.25} - 49.31 (1-(1-F)^{-0.55}))\ A^{0.64}$
Considering basin as a unit	Index-flood	$Q_T = (13.71+7.03(-\ln(-\ln(1-1/T))))A^{0.52}$
	EV-I Distribution	$Q_T = (13.58+7.25(-\ln(-\ln(1-1/T))))A^{0.52}$
	GEV Distribution	$Q_T=(65.60(-\ln(1-1/T))^{-0.1} -52.35)A^{0.52}$
	Wakeby Distribution	$Q_T=(3.16+6.16(1-(1.F)^{9.165} - 4211.54 (1-(1-F)^{-0.002}))\ A^{0.52}$

Table.3. Regional frequency formulae developed using different distribution for Wainganga basin.

Region	Method	Regional flood formula
Wainganga catchment	Index-flood	$Q_T = (4.359+2.973(-\ln(-\ln(1-1/T))))A^{0.7}$
	EV-I Distribution	$Q_T = (4.246+2.869(-\ln(-\ln(1-1/T))))A^{0.7}$
	GEV Distribution	$Q_T=(4.40+2.84\ (-\ln(1-1/T))^{-0.1}\ )A^{0.7}$
	Wakeby Distribution	$Q_T=(0.7113+4.03(1-(1.F)^{2.54} - 21.04(1-(1-F)^{-0.116}\ )\ A^{0.7}$

8.0 Remarks

The purpose of the frequency analysis is to estimate the design flood for desired recurrence interval assuming the sample data follow a theoretical frequency distribution. It is assumed that the sample data is a true representative of the population. It is generally seen that minimum 30 to40 years of requires are needed in order to carry out the flood frequency analysis to the at site data for estimating the floods in extrapolation range, somewhat, within the desired accuracy. In case the length of records are too short, it represents inadequate data situation and at site flood frequency analysis fails to provided the reliable and consistent flood estimates. The regional flood frequency curves together with at sire mean is generally able to provide more reliable and consistent estimates of floods under the inadequate data situation. For ungauged catchment, the regional flood frequency analysis approach is the only way to estimate the flood, for desired recurrence interval for which a regional relationship between mean annual peak flood and catchment characteristics is developed along with the regional frequency curves.

Most of the goodness of fit test used for judging the best fit distribution represents the descriptive ability of the theoretical distributions. Generally the frequency distributions show larger deviations in extrapolation range. Thus a distribution which may pass the goodness of fit criteria not necessarily be able to estimate the floods for higher recurrence intervals to the desired accuracy. In order to judge the performance of the theoretical distributions in predicting the higher recurrence interval floods the predictive ability tests must be taken up using the Monte Carlo experiments with the data, generated by the selected distributions based on the descriptive ability criteria. Such generation studies provide a better opportunity for understanding the characteristics of the theoretical distributions.

Reference

1. Chow, V.T., 1964. Hand book of Applied Hydrology, Mc-Graw Hill New York

2. Haan,C.T., 1977. Statistical Methods in Hydrology, The Iowa State Univeristy Press, USA.

3. Kite, G.W., 1977. Frequency and Risk Analysis in hydrology, Water Resources Publications, Colorado.

4. Linseley, R.K., Kohler, H.A., and Paulhus, J.L.H., 1975. Hydrology for Engineers, Mc-Graw Hill, International book Company.

5. McGuess, R.H. and Snyder,W.M., 1985. Hydrologic Modelling, Statistical Methods and Applications, Prentice-Hall,Englewood Cliffes, New Jersey.

6. Venkatesh,B. and Singh,R.D., 1999. Development of Regional Flood Formula for Krishna Basin, Vol.5, No.2, Journal of Indian Society of Hydraulics, pp 44.54.
7. Venkatesh.B, 1999. Flood frequency analysis of Wainganga basin, Presented in XVIII Annual Convention and National Seminar on 'Water Resources and Sustainable Development in Next Century' held at Solapur, during 27-28 December 1999.

8. Yevjevich,V., 1972. Probability and Statistics Hydrology, Water Resources Publications, Fort Collins, Colorado

Table 1. Distribution used in flood frequency analysis

Distribution Name	Probability density function f(x) or Distribution function F(x)	Variate Range
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	
Log-normal	$f(x) = \frac{1}{\sigma_y \sqrt{2\pi} x} \exp\left[-\frac{1}{2}\left(\frac{\log_e x - \mu_y^2}{\sigma_y}\right)^2\right]$	
Pearson Type- III	$f(x) = \frac{(x-m/a)^b}{\lvert a\rvert \Gamma(b+1)} \exp\left\{-\left(\frac{x-m}{a}\right)\right\}$	$m \le x$ if $a > 0$ $x \ge m$ if $a < 0$
Exponential	$F(x) = 1 - \exp\left(-\frac{x-m}{a}\right)$	$m \le x$
Gumbel or EV-I	$F(x) = \exp\left(-e^{-(x-\mu/\alpha)}\right)$	$-\infty \le x \le \infty$ $\alpha > 0$
Extreme Value	$F(x) = \exp\left\{-\left(\frac{\mu-\varepsilon}{x-\varepsilon}\right)^k\right\}$	$k>0$, $\varepsilon \le x$ $0 \le \varepsilon < \mu$
General Extreme Value	$F(x) = \exp\left\{-\left(1-k\left(\frac{x-\mu}{\alpha}\right)\right)^{1/k}\right\}$	$\alpha > 0$ $\mu + \frac{\alpha}{k} \le x \le \infty$ if $k<0$ $-\infty < x \le \mu + \frac{\alpha}{k}$ if $k>0$
Wakeby	$X = m + a\left[1-(1-F(X))^b\right] - c\left[1-(1-F(X))^{-d}\right]$	