Flood Frequency Analysis (Part-II)

PRINT VERSION MODULE

Top of page

MODULE OBJECTIVES

- To get familiarized with a few Statistical parameters
- To grasp difference between empirical vs. theoretical frequency distribution
- To understand & perform various tests to ensure fitness of data for flood frequency analysis
- To learn how to plot confidence band and its significance
- To grasp the meaning and significance of confidence band; confidence limit; outliers; expected probability etc.

Top of page

INTRODUCTION

The previous module on this topic provides elementary knowledge of flood frequency analysis. This module moves a step further, and enables the reader to handle complex problems related to this topic.

Estimates of extreme events of given recurrence interval are used for a host of purposes, such as design of dams, coffer dams, bridges, flood-plain delineation, flood control projects, barrages, and also to determine impact of encroachment of flood plain etc. Frequency analysis, if done manually, is burdensome, tedious, and leaves little manoeuvring space if something wrong is noticed at the end of calculation. It often requires calculations all over again. Accordingly, this module attempts at presenting some statistical parameters, its application in flood frequency analysis, and thereafter introduces HEC-SSP software that offers multiple functions to perform frequency analysis speedily and accurately.

Top of page

STATISTICAL TERMS/PARAMETERS OFTEN USED IN FREQUENCY ANALYSIS

Statistics

Statistics is concerned with the collection, ordering and analysis of data. Data consists of sets of recorded observations or values. It also provides criteria for judging the reliability of the correlation between variables; means for deriving the best relationship for predicting the one variable from known values of other variables. Any quantity that can have a number of values is a variable. A value that a variable takes is called 'Variate'. A variable can be either;

1. Discrete - a variable, whose possible values can be counted, e.g. number of rainfall days in a month or year. Number would take only integer values within zero and infinity, or
2. Continuous - a variable; which can take on any value within specified interval. Annual maximum discharge, for example, is a continuous variable as it could be any value between zero and infinity.

Top of page

Sample and Population

Any time set of recorded or observed data does not constitute the entire population. It is simply a fraction of entire population and is called a 'sample'. By deducing the characteristics exhibited by sample, inferences are drawn about the nature of entire population. In other words, collected samples help us predict the likely magnitude and occurrence of future events. It is obvious here that quality and length of sample used in analysis hugely impact the quality of forecast of ensuing events.

Top of page

Measure of central tendency

The arithmetic mean of a set of 'n' observations is their average:

$$\text{mean} = \frac{\text{sum of observations}}{\text{number of observations}} \quad \text{that is} \quad \bar{x} = \frac{\sum x}{n}$$

When calculating from a frequency distribution, this becomes:

$$\bar{x} = \frac{\sum xf}{n} = \frac{\sum xf}{\sum f}$$

In MS excel, for a given set of data, the mean can be determined by entering function 'average (a1:a20)' in formula bar. Here, a1:a20 indicates the range of cells from a1 to a20 containing sample data, if sample length is 20.

Mean is not a firm or fixed value; and fluctuates within a range with variation in length of samples. The range of this fluctuation is better expressed by another statistical parameter, i.e. Standard Error of Mean. Other measures of central tendency are median and mode.

Top of page

Dispersion Characteristics

Top of page

Range

The mean, mode and median give important information about the central tendency of data but they do not tell anything about the spread or dispersion of samples about the centre.

For example, let us consider the two sets of data:

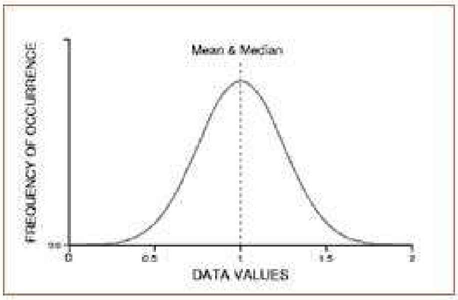
26, 27, 28, 29 30, and 5, 19, 20, 36, 60

The simplest measure of dispersion is the range - the difference between the highest and the lowest values. For these two set of data, both samples have a mean of 28, but range for first set is 4, for second it is 55. Obviously, one is clearly more tightly arranged about the mean than the other.

Top of page

Standard Deviation

The standard deviation, SD is most widely used measure of dispersion around Mean. It indicates the slope of distributed curve on either side of the mean. According to the nature of dispersal of data, slope could be either gentle or steep. A high SD indicates gentle slope, widely scattered around mean and higher range; while, converse is true, when SD is less. Based on this description, it can be presumed that first set of data will have smaller SD than that of the second set. A normally distributed curve slopes alike on either side of the mean as shown here. This aside, for normally distributed data, mean, median and mode, all coincide.



The variance of a set of data is the average of the square of the difference in value of a datum from the mean:

$$\text{variance} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n}$$

This has the disadvantage of being measured in the square of the units of the data. The standard deviation is the square root of the variance:

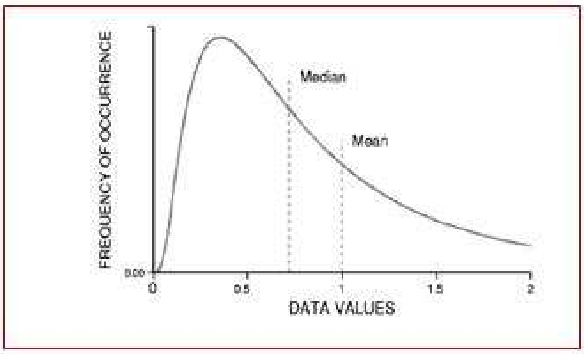
$$\text{standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

This formula with denominator 'n' indicates SD of entire population. However, for all practical purposes, we deal with 'samples' only, and in such case, denominator 'n' is replaced by (n-1) to account for limited length of data. Excel formula to estimate this parameter is =stdev(Range of data). Here, for two sets of data, SD computed is 1.58 & 21 respectively, which is consistent with our presumption made earlier.

Top of page

Skewness

In several cases, frequency of occurrence of variables is not normally distributed and plots either skewed +ve (right) (as shown in the fig.) or skewed -ve (left). In other words, slopes of the curve on either side are dissimilar. Unlike normally distributed data, mean, median and mode for skewed data do not coincide. Peaked point of skewed plot is the location of mode. For normally distributed curve, skewness is zero.



This parameter is determined by function skew(range of data) in MS Excel. It is evident, from table, that for evenly distributed data set, skewness is zero. Second set of data is positively skewed.

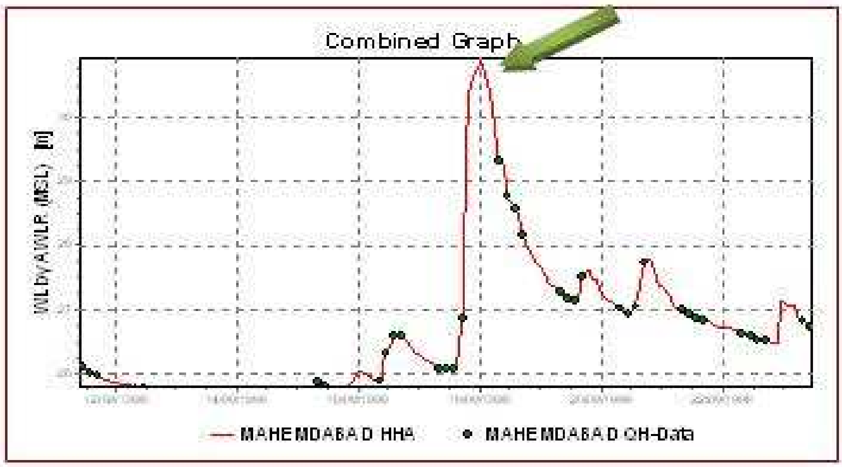
	26	5
	27	19
	28	20
	29	36
	30	60
Mean	28.0	28.0
SD	1.6	21.0
Skewness	0.0	0.9

HEC-SSP software itself computes these parameters and performs a number of tasks using them.

Top of page

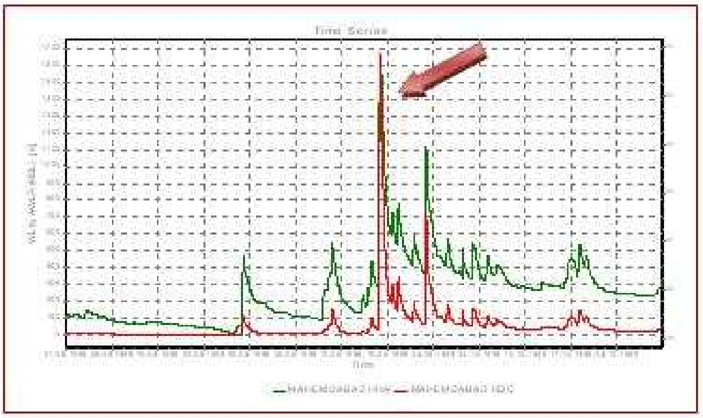
WHICH VALUE/DATA QUALIFIES AS AN ANNUAL PEAK OF A YEAR ?

Collection of a set of particular type of data is purpose driven. For frequency analysis of flood peaks corresponding to a return period of 50-yr or so, we look for collection of a set of instantaneous peak discharge of different years. Here, instantaneous peak discharge of a year means that discharge is highest of all discharge values flowed past a measuring section during the period. The question is how to gather this set of information. Following Para discusses this aspect.



Hourly discharge observation is not only expensive but also impracticable. Instead, a widely prevalent practice in India is to record discharge observation once a day (usually at 0800hr or so), and water level every hour. It is important to note that recorded discharge observation may or may not be the peak discharge of the day; and therefore, it can't be a true representative of an instantaneous peak discharge of a day. Let us understand it differently. In a plot shown here, water level hydrograph and the level when discharge was carried out have been shown together. It is easily noticeable here that peak water level (hence discharge) occurred between two observations. This means that if we pick up instantaneous peak discharge out of observed discharge recorded in a year, missing out true instantaneous peak can't be ruled out. Therefore, it had better look for all such peaks in a year, and pick up a corresponding discharge value that is highest of all. Followings are few approaches suggested to consider before finalizing an array of annual peaks.

1. Fit a rating curve (s) between observed discharge and corresponding water level. Rating curves so developed and hourly water level hydrograph together can be used to obtain a no-break/continuous discharge series of a particular year. A plot of water level and continuous discharge series, developed using HYMOS software, is displayed here. Peak of this series represents instantaneous annual peak of that year.



2. In absence of rating curve, a correlation between past observed discharge or mean daily discharge (maximum of a year) and instantaneous peak discharge can be developed. This relationship can be used to generate peak discharge corresponding to maximum observed discharge for subsequent years.
(for detailed discussion, pl refer to Hydrologic Frequency Analysis, Vol-3 published by US Army Corps of Engineers- 1975, http://www.hec.usace.army.mil/publications/IHDVolumes/IHD-3.pdf)

3. In some quarters, peak daily or peak mean daily discharge data are raised by certain

percentage, say 20 or 30%. This method is little ambiguous and subjective as all peak daily values may or may not touch instantaneous peak by application of a certain percentage.

Top of page

HOW TO ENSURE FITNESS OF DATA FOR FREQUENCY ANALYSIS?

Annual peaks gathered for frequency analysis must be a product of random factors only. Presence of one or more data influenced by manual and/or systematic errors gravely distorts the distribution of plot and its reliability, if go unnoticed in the analysis. So, it is essential that a suspected data should be detected and treated for its modification or retention or deletion before analysis. This apart, data should possess attributes, such as homogeneity, randomness, and stationarity. These attributes are explained in succeeding paragraphs.

a. **Homogeneity**

   Homogeneity implies that the sample is representative of same population. The homogeneous requirement means that each flood occurs under more or less similar conditions. Two flood events are homogeneous, if both are caused by same factor, such as rainfall. Flood peaks triggered by dam break, breach in embankment are isolated events, and should not be part of peaks created by rainfall. It is assumed that though peak flows of finite years' have been recorded; the same type of 'Statistical Character' (mean, standard deviation, and skewness) was always there and would behave alike in future too. For this reason, a set of data belonging to same population must closely exhibit the similar statistical behaviour with another set of data of same population. To test homogeneity of data, Student 't' test is normally performed.

b. **Independence/Randomness**

   This is explained in previous module on this topic. Independence or randomness is usually investigated by Turning Point test.

c. **Stationarity**

   In this the properties or characteristics of the sample do not fluctuate with time. Linear trend test determines this property of sample.

   If any of these is not an attribute of a sample, the use of probability/theoretical frequency distribution may lead to erroneous results. Accordingly, it is desirable that before any analysis, one must see that sample should conform to these attributes.

   HEC-SSP offers no tools to perform these tests. Nevertheless, interested users, can use HYMOS software to test if compiled set of data qualifies for flood frequency analysis. *For more details, we recommend reference to Hydrology Project-I Training Module no.43. This material is available as part of this week's module.*

Top of page

EMPIRICAL Vs. THEPRETICAL DISTRIBUTION CURVE

Absolute frequency - Supposing there is a variable which can take values from 0 to 100. A sample of this variable holds 50 different values. Let us group these data in five equal intervals, e.g., 0-20, 20-40,--- -- --, 80-100. There distribution across five groups is 'absolute frequency'. Absolute frequency, say n divided by N, is relative frequency or probability. Please notice that sum total of relative frequency is '1'. This concept is used a little later.

Class interval	Absolute Frequency, n	Relative Frequency, n/N
0-20	5	0.10
20-40	11	0.22
40-60	20	0.40
60-80	9	0.18
80-100	5	0.10
	N = 50	**∑ P = 1**

A relative frequency curve plotted on the basis of distribution of data in a sample presents a distribution curve known as empirical distribution curve. This distribution and its statistical parameters help an engineer fit a theoretical frequency distribution curve, as closely to the empirical distribution as possible to ensure mathematical tractability further.

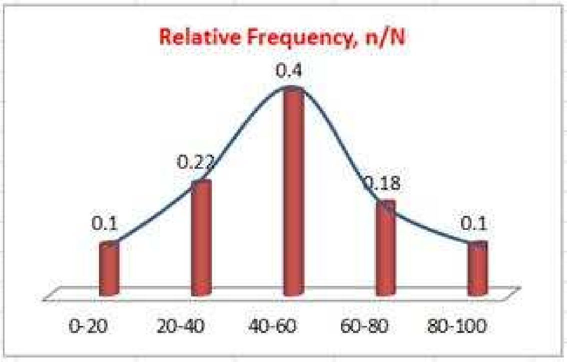


Fig 1

As understood a while ago, the probability or relative frequency is defined as the number of occurrences of a variate divided by the total number of occurrences, and is usually designated by P(x). The total probability for all variates should be equal to unity, that is, $\Sigma P(x) = 1$. Distribution of probabilities of all variates is called Probability Distribution, and is usually denoted as *f(x)* as shown in Fig.1.

The cumulative probability curve, F*(x)* is of the type as shown in Fig.2.

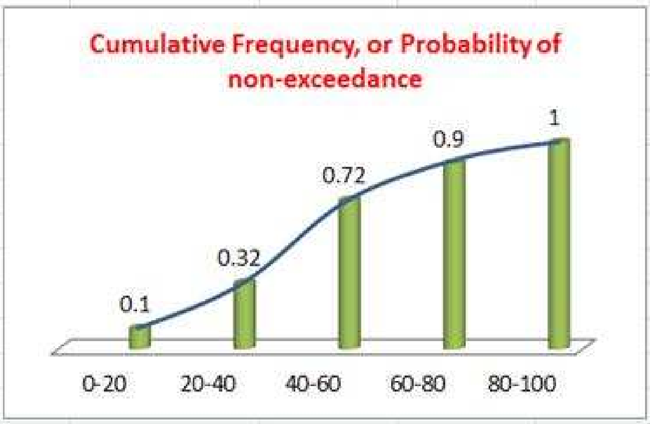


Fig 2

The cumulative probability or 'probability of non-exceedance', designated as $P(x < x)$, represents the probability that the random variable has a value less than certain assigned value *x*. Additive inverse of $P(x < x)$, or $P(x \geq x)$, is termed as Exceedance Probability. Reciprocal of exceedance probability is return 100 times the Exceedance Probability is called as Exceedance Frequency. Now, glance at **Table1**; and read what the probability of 60 not getting exceeded is.

Table 1

Class interval	Absolute Frequency, n	Relative Frequency, n/N	Cumulative Frequency, or Probability of non-exceedance	Exceedance Probability	Exceedance Frequency	Return Period in year = 1/Exceedance Probability
0-20	5	0.10	0.10	0.90	90	1.1
20-40	11	0.22	0.32	0.68	68	1.5
40-60	20	0.40	0.72	0.28	28	3.6
60-80	9	0.18	0.90	0.18	18	5.6
80-100	5	0.10	1.00	0.00	0	
	N = 50	∑P = 1				

In the context of flood frequency analysis, we apply above concepts by assuming the instantaneous yearly flood peaks as the variate 'x'. Then, if the functions f(x) or F(x) becomes known by fitting a theoretical distribution, it is possible to find out the probability (or return period) of a flood peak, or conversely, a flood magnitude of desired return period (also return interval or recurrence interval).

There are a number of probability distribution functions f(x), which have been suggested by statisticians. HEC-SSP supports following distribution functions.
(Reader can download and install HEC-SSP software from site,
http://www.hec.usace.army.mil/software/hec-ssp/downloads.html)

Without log transformation

I. ***Normal &***
II. ***Pearson type III***

With log transformation

I. ***Log normal &***
II. ***Log Pearson type III***

Another often used distribution is **Gumbel method.** Even if, HEC-SSP software does not include this method, user can readily use mean and standard deviation to estimate flood peak corresponding to a return period, T = (1/P) by use of formula placed below:

$$X_T = M + B * (-l_n (-l_n (1-P)))$$

Where,

$M = X_{mean}$ - 0.45005 * Standard Deviation
B = 0.7797 * Standard Deviation

However, this method is recommended when length of data is fairly large, say more than 100 (ref: Patra K C, Hydrology and Water Resources Engineering). Alternatively, when data is scarce, i.e., data length is below 100, user may use Gumbel table, which features in almost every hydrology book, to read K, frequency factor for given sample size and return period. In this case, X_T is estimated by

$$X_T = X_{mean} + K * St\ Deviation$$

Top of page

PLOTTING POSITION

To assign a probability to a sample data (also called variate) and to determine its 'plotting position' on probability sheet, sample data consisting of N values is arranged in descending order. Each data (say the event X) of the ordered list is then assigned a rank 'm' starting with 1 for the highest up to N for the lowest of the order. The exceedance probability of a certain value x is estimated by formula presented below:

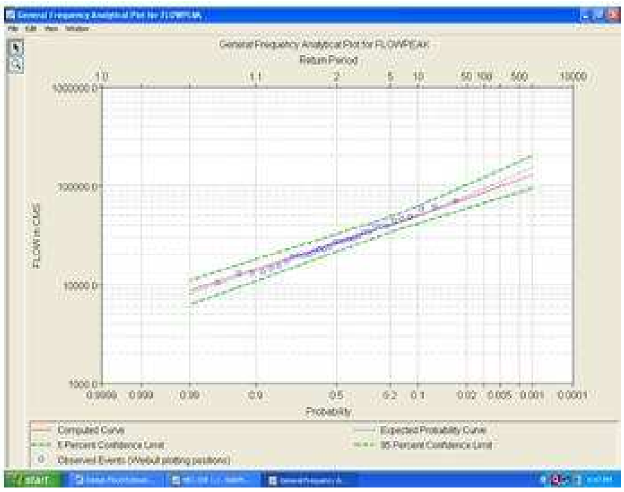
$$p = (m-a)/(N-a-b+1)$$

Where, m is rank of the sample data in the array; N represents the size of sample; and 'a' and 'b' are constants. For different methods, a and b assume different values. For Weibull method, a & b equal zero; and hence, P reduces to m/(n+1). HEC-SSP, by default, uses Weibull method to show dispersion of data. Nevertheless, option is available for alternate methods by defining appropriate value of a & b. Of these, the Weibull formula is most commonly used, because it is simple and intuitively easily understood to determine the probability. (For detailed discussion on the choice of a particular method, reader may refer to Applied Hydrology by Ven T Chow, p -).

Top of page

WHICH DISTRIBUTION FITS WELL ?

HEC-SSP offers graphical plot displaying scatter of sample data in addition to computed curve. Here, user has choice to choose method of plotting position and a theoretical curve of his choice. Graphical plot is a visual aid of determining worthiness of choice broadly; and therefore, conclusion based on merely eye judgment is hugely subjective. To overcome this limitation, user can analyze the result distilled by software and employ any one of the following tests to measure the strength of fitness. However, such analysis needs to be done outside; as HEC-SSP contains no built-in function of this kind. This module presents steps to perform D-test only. Details with regard to others, users may refer to Hydrology Project Training Module no.43.



- Chi-square test
- Kolmogorov-Smirnov test
- Binomial goodness of fit test, and
- D-index test

Once a particular distribution is found the best, it is adopted for calculation of peak floods in future.

D-index is calculated by

$$D\text{-}index = \Sigma 1 to 6\ (abs(Xi_{observed} - Xi_{computed})/(mean\ of\ sample)$$

where,

$Xi_{observed}$ = observed value of a given p, exceedance probability

$Xi_{computed}$ = for identical p, value determined by distribution curve

D-index test is shown later in this module.

Top of page

CASE STUDY

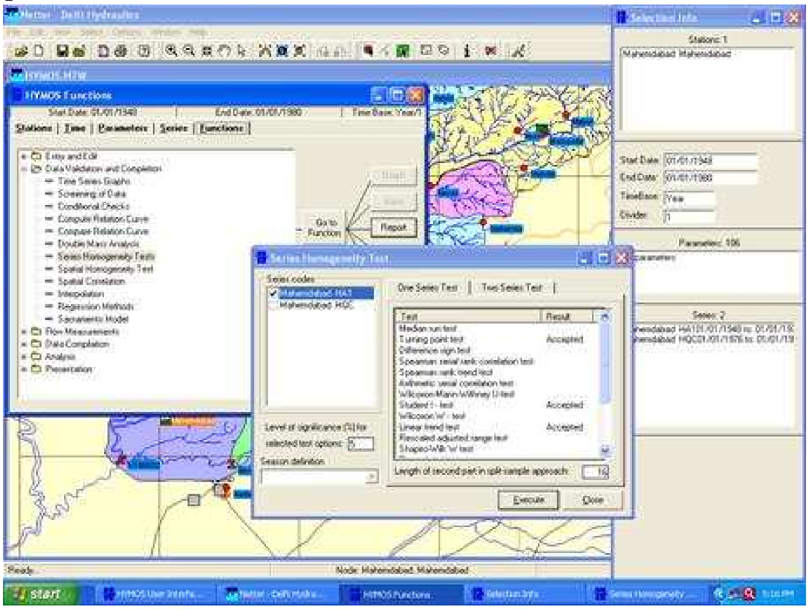
This point forward, a real sample (Table 2) has been collected for its frequency analysis with ***HEC-SSP software***. The application of the method of plotting and fitting a theoretical distribution curve, analysis of output will help reader grasp the functions of this software speedily. The software outputs a series of additional information, which have been discussed at appropriate locations.

Table 2
Annual Flood Series in cumec from 1948 to 1979 for the river Narmada at Garudeswar

Year	Annual Peaks in Cumecs	Year	Annual Peaks in Cumecs
1948	23890	1964	19560
1949	26810	1965	15250
1950	45630	1966	13000
1951	10380	1967	22670
1952	13290	1968	58100
1953	17100	1969	31170
1954	28650	1970	69400
1955	29150	1971	18980
1956	12910	1972	17980
1957	26700	1973	61350
1958	19700	1974	27300
1959	38800	1975	33750
1960	21250	1976	19500
1961	43360	1977	22700
1962	38880	1978	34260
1963	14250	1979	38200

Step 1

As quoted earlier, this set of data is required to be investigated to confirm its adherence to desired attributes of sample data, i.e. homogeneity, randomness and stationarity. Following is screenshot of HYMOS software which is used to conduct series homogeneity test of a given series. A pop-up window in the middle of this screenshot indicates results of this series as 'accepted'. In all three tests, hypothesis, that series is random, is not rejected. This implies that the current sample is a collection of random data.




```
Date of first element in series= 1948  0   0   0   1
Number of data                 =   32

 Turning Point Test
 ------------------
Number of turning points (=Nt)    =       20
Mean of Nt                        =   20.000
Standard deviation of Nt          =    2.317
Test statistic [u] (abs.value)    =     .000
Prob(u.le.[u])                    =     .500
Hypothesis: H0: Series is random
            H1: Series is not random
            A two-tailed test is performed
            Level of significance is 5.00 percent
            Critical value for test statistic 1.960
Result:     H0 not rejected

Date of first element in series= 1948  0   0   0   1
Number of data                 =   32
```

```
Test for Significance of Linear Trend
--------------------------------------
Intercept parameter      (=b1)    = 21530.610
Slope parameter          (=b2)    = .4813E+03
St.dev. of b2            (=sb2)   = .2762E+03
St.dev. of residual      (=se)    = .1445E+05
Test statistic [t]   (abs.value)  =     1.743
Degrees of freedom                =        30
Prob(t.le.[t])                    =      .954
Hypothesis: H0: Series is random
            H1: Series is not random
            A two-tailed test is performed
            Level of significance is 5.00 percent
            Critical value for test statistic 2.042
Result:     H0 not rejected

Date of first element in series= 1948  0  0  0  1
Number of data                 =   32

Student t-Test with Welch modification
---------------------------------------
Number of data in first set       =        16
Number of data in second set      =        16
Test statistic [t]  (abs.value)   =     1.480
Degrees of freedom                =        30
Prob(t.le.[t])                    =      .925
Mean of first set         (mA)    = 25666.880
St.dev. of first set      (sA)    = 11229.590
Mean of second set        (mB)    = 33323.130
St.dev. of second set     (sB)    = 17385.350
Var. test stat.   (Qi=sA^2/sB^2)  =      .417
Prob(Q.le.Qi)                     =      .051
Hypothesis: H0: Series is random
            H1: Series is not random
            A two-tailed test is performed
            Level of significance is 5.00 percent
            Critical value for test statistic 2.042
Result:     H0 not rejected
```

Step 2

Subsequent steps begin with creation and saving of an EXCEL sheet with two columns - first for year and second for discharge. This file is imported (Fig.4) in HEC-SSP software to carry out frequency analysis. Interested reader is suggested to go through 'User's Manual' of this software (p 4-7 to p 4-9 to learn how to import data from MS excel), which is available under 'Help' menu of software.

This manual is also available at ***http://www.hec.usace.army.mil/software/hec-ssp/documentation/HEC-SSP_20_Users_Manual.pdf*** .

Optionally, user can directly input data by selecting 'Manual' button on 'Data Importer' window (Fig.4). To open 'Data importer' window, click on 'Data' menu followed by choosing 'New'.

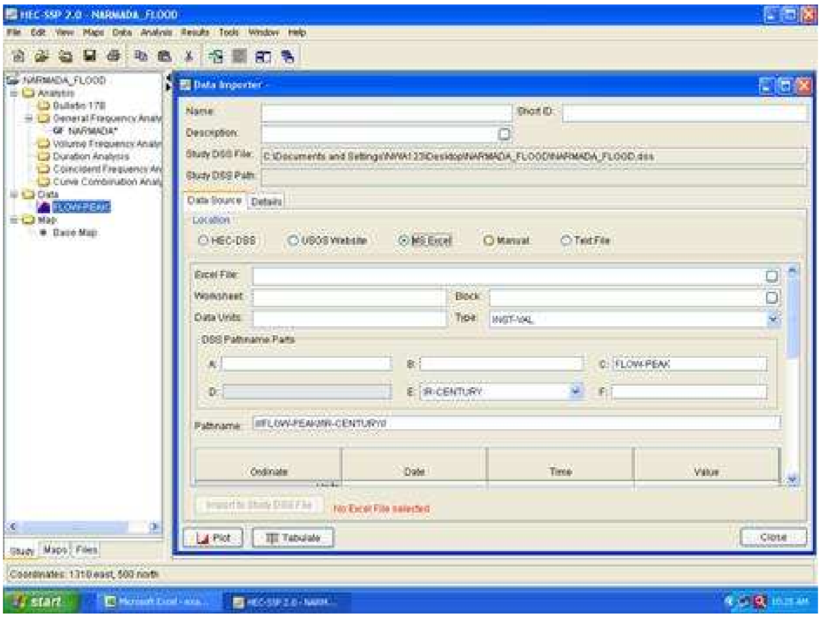


Fig 4

Step 3

Once, data is available, Chapter 6 of 'User's Manual' help user finish frequency analysis. 'General Frequency Analysis Editor' window as shown in Fig.5 can be activated by selecting Analysis New - General Frequency Analysis option on the menu. An analysis report (Table 3) along with distribution curve (Fig.6) generated by the software for this set of data using Log Pearson type III distribution is placed next. Before, we delve into results; let us familiarize ourselves with a couple of lines appearing on the plot. Later, we will discuss their significance, and how they are estimated.

Tiny circular points in blue are annual peaks occupying their position on the plot (also probability sheet) according to probability assigned to them by 'Weibull method'. As discussed earlier in the module, this scattering is 'Empirical Frequency Distribution'. A line in red denotes Log Pearson Type-III 'Theoretical Distribution Curve'. Could you read on the plot what return period for circular point farthest to the right is? It is roughly 30yrs. If we desire to ascertain peak discharge of still higher return period sticking to empirical distribution, no information is available. For a majority of hydrological and hydraulic related studies, flood magnitude of return period of 50 yrs or more is needed. Such estimations are extracted with the help of theoretical distribution plot, which is mathematically extended further.

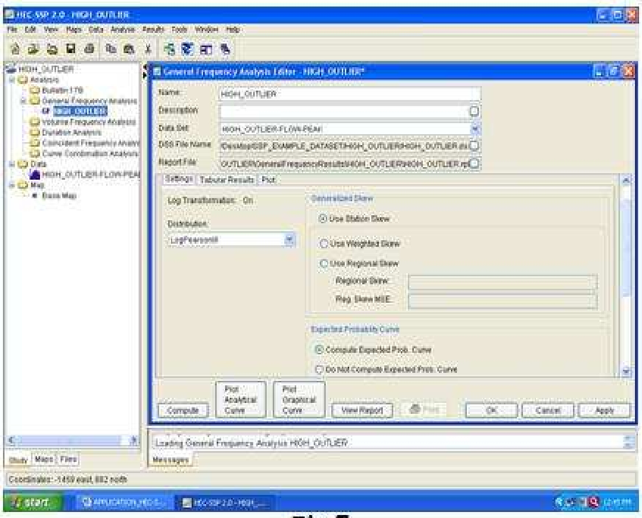


Fig.7

Fig 5

- A dotted line in blue is expected probability curve. This aspect is discussed later.
- A pair of two lines in green on either side of plot is 90% confidence band. This aspect is also covered later.

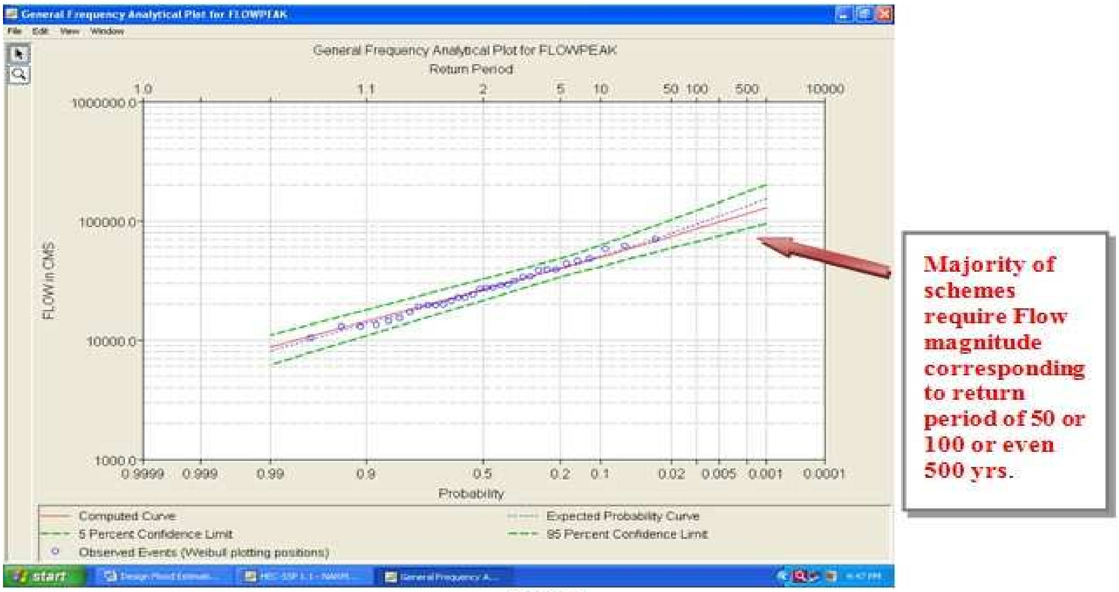


Fig.6

Table 3

-- Final Results ---

<< Plotting Positions >>

FLOW-PEAK

Events Analyzed Day Mon Year	FLOW CMS	Ordered Events Rank	Water Year	FLOW CMS	Weibull Plot Pos %
31 Dec 1947	23,890.0	1	1970	69,400.0	**3.03** = **Rank/(N+1)**
31 Dec 1948	26,810.0	2	1973	61,350.0	6.06
31 Dec 1949	45,630.0	3	1968	58,100.0	9.09
31 Dec 1950	10,380.0	4	1972	47,980.0	12.12
31 Dec 1951	13,290.0	5	1950	45,630.0	15.15
31 Dec 1952	17,100.0	6	1961	43,360.0	18.18
31 Dec 1953	28,650.0	7	1962	38,800.0	21.21
31 Dec 1954	29,150.0	8	1959	38,800.0	24.24
31 Dec 1955	12,910.0	9	1979	38,200.0	27.27
31 Dec 1956	26,700.0	10	1978	34,260.0	30.30
31 Dec 1957	19,700.0	11	1975	33,750.0	33.33
31 Dec 1958	38,800.0	12	1969	31,170.0	36.36
31 Dec 1959	21,250.0	13	1955	29,150.0	39.39
31 Dec 1960	43,360.0	14	1954	28,650.0	42.42
31 Dec 1961	38,800.0	15	1974	27,300.0	45.45
31 Dec 1962	14,250.0	16	1949	26,810.0	48.48
31 Dec 1963	19,560.0	17	1957	26,700.0	51.52
31 Dec 1964	15,250.0	18	1948	23,890.0	54.55
31 Dec 1965	13,000.0	19	1977	22,700.0	57.58
31 Dec 1966	22,670.0	20	1967	22,670.0	60.61
31 Dec 1967	58,100.0	21	1960	21,250.0	63.64
31 Dec 1968	31,170.0	22	1958	19,700.0	66.67
31 Dec 1969	69,400.0	23	1964	19,560.0	69.70
31 Dec 1970	18,980.0	24	1976	19,500.0	72.73
31 Dec 1971	47,980.0	25	1971	18,980.0	75.76
31 Dec 1972	61,350.0	26	1953	17,100.0	78.79
31 Dec 1973	27,300.0	27	1965	15,250.0	81.82
31 Dec 1974	33,750.0	28	1963	14,250.0	84.85
31 Dec 1975	19,500.0	29	1952	13,290.0	87.88
31 Dec 1976	22,700.0	30	1966	13,000.0	90.91
31 Dec 1977	34,260.0	31	1956	12,910.0	93.94
31 Dec 1978	38,200.0	32	1951	10,380.0	96.97

Observed data in red are used for D-test later in the module.

<< Skew Weighting >>

Based on 32 events, mean-square error of station skew = 0.165

Mean-square error of regional skew = -?

<< Frequency Curve >>
FLOW-PEAK

Computed Curve FLOW, CMS	Expected Probability	Percent Chance Exceedance	Confidence Limits 0.05 FLOW, CMS	Confidence Limits 0.95 FLOW, CMS
128,850	155,033	0.1	202,011	95,032
114,835	133,685	0.2	175,013	86,216
97,583	109,310	0.5	142,963	75,091
85,421	93,258	1.0	121,241	67,029
73,951	78,936	2.0	101,499	59,223
67,382	71,046	3.0	90,555	54,646
56,853	58,897	6.1	73,629	47,108
50,892	52,273	9.1	64,428	42,706
46,744	47,736	12.1	58,208	39,569
43,550	44,303	15.2	53,532	37,106
40,937	41,522	18.2	49,787	35,054
8,691	8,071	99.0	10,957	6,232

For first data on the list here, % exceedance is 0.1. This means exceedance probability is 0.1/100 = 0.001, or return period is 1/0.001 = 1000 yrs. against this, Q_{1000} is 128850cumec. In the same row, Q = 202,011 indicates a figure which has exceedance probability of 0.05 for return period of 1000yrs. 1-0.05 = 0.95 is its probability of non-exceedance.

Computed data in red are used for D-test later in the module

<< Systematic Statistics >>
FLOW-PEAK

Log Transform: FLOW, CMS		Number of Events	
Mean	4.419	Historic Events	0
Standard Dev	0.213	High Outliers	0
Station Skew	0.104	Low Outliers	0
Regional Skew	---	Zero Events	0
Weighted Skew	---	Missing Events	0
Adopted Skew	0.104	Systematic Events	32

Of several useful contents generated by software, two of them need special attentions. These are:

I. Confidence Limits, and

II. Expected Probability

Top of page

CONFIDENCE BANDS AND CONFIDENCE LIMITS

The record of annual peak flow at a site is a random sample collected over a period of time. A varied nature of causative factors and complex interactions among them bring about randomness in the sample. Therefore, in all likelihood, a different set of samples of same population results in different estimate of the frequency curve. Thus, an estimated flood frequency curve can be only an approximation to the true frequency curve of the population of annual flood peaks. To gauge the accuracy of this approximation, one may construct an interval or a range of hypothetical frequency curves that, with a high degree of confidence, contains the population frequency curve. Such intervals are called confidence intervals and their end points are called confidence limits. This is analogous to standard error of mean or standard error of mean relationship concept.

The two limits of 0.05 and 0.95, or 5% and 95% chance exceedance curve,(pl see the result in table 3), imply that there is 90% chance/probability that discharge value will lie/occur between these bounds; and only 10% of observation may fall outside this band. If we put it differently, upper limit suggests a flow with 5% of exceedance probability, or (100-95), i.e. 5% non-exceedance probability. If certainty of this degree is warranted for any project, flow of this magnitude can be chosen for design, but at the cost of escalation in project cost. In fact, this choice is a trade-off between cost of the project and safety of the structure. Similar conclusion can be drawn about lower limit

The confidence band width is determined by a formula given below:

$$Q_{U,L} = Q_{mean} \pm K_{U,L} * St\ Deviation$$

Where,

$K_{U,L}$ is a function of exceedance probability, sample size, skewness coefficient and confidence interval opted by the user. The value of $K_{U,L}$ declines with rise in sample size. This brings two lines representing QU & QL closer to each other, and therefore, a narrower band will appear. HEC-SSP assumes exceedance probability of 0.05 and 0.95 by default and returns the output. User, at his discretion, can select any other value instead. For more details about $K_{U,L}$, reader may refer to 'Reference 2'.

Top of page

EXPECTED PROBABILITY

The expected probability adjustment is necessitated to account for a bias introduced in the distribution curve on account of shortness of data. Factually, all distributions assume spread of data from $-\infty$ to $+\infty$; while in reality, this is far from real. This calls for measures to address short length of data. **Table 4 is an excerpt from** Applied Hydrology by Ven Te Chow listing correction factors for different return periods.

Table 4

T years	*Exceedance Probability*	*Expected Probability*
10000	0.0001	0.0001*(1.0+1600/N^1.72)
1000	0.001	0.001*(1.0+280/N^1.55)
100	0.01	0.01*(1.0+26/N^1.16)
20	0.05	0.05*(1.0+6/N^1.04)
10	0.10	0.10*(1.0+3/N^1.04)
3.33	0.30	0.30*(1.0+0.46/N^0.925)

Where, N is number of sample data used in the analysis. Please notice that as N approaches infinity, expected probability equals exceedance probability. Here too, HEC-SSP offers both alternatives to compute or not to compute expected probability and corresponding flood values for various exceedance probabilities (Fig.7).

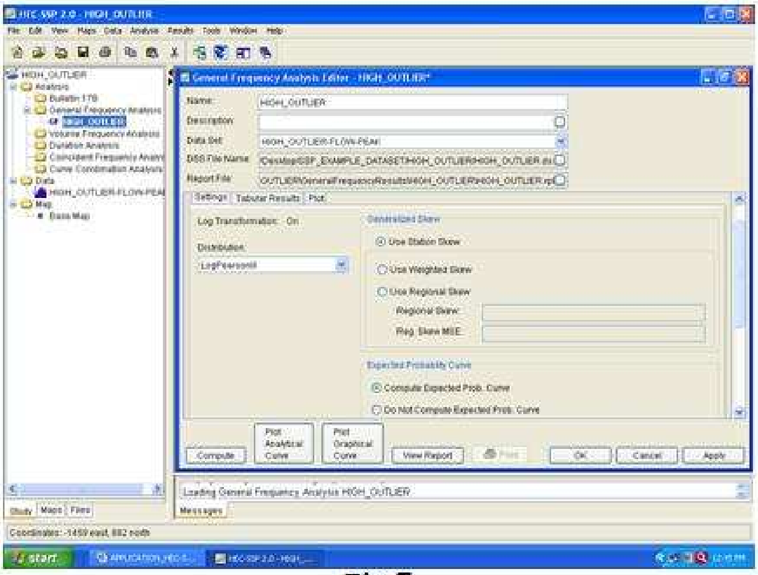


Fig.7

Top of page

HOW TO PERFORM D-INDEX TEST

HEC-SSP software, by default, outputs flood peaks of a few exceedance frequencies like 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0, 80.0, 90.0, 95.0, and 99.0. However, appropriate part of window, shown at Fig.8, can be suitably adjusted by the user to gather flood peaks of desired exceedance frequency, usually matching with what tabulated by the software using Weibull method. (pl refer to tabular result under Table 3).

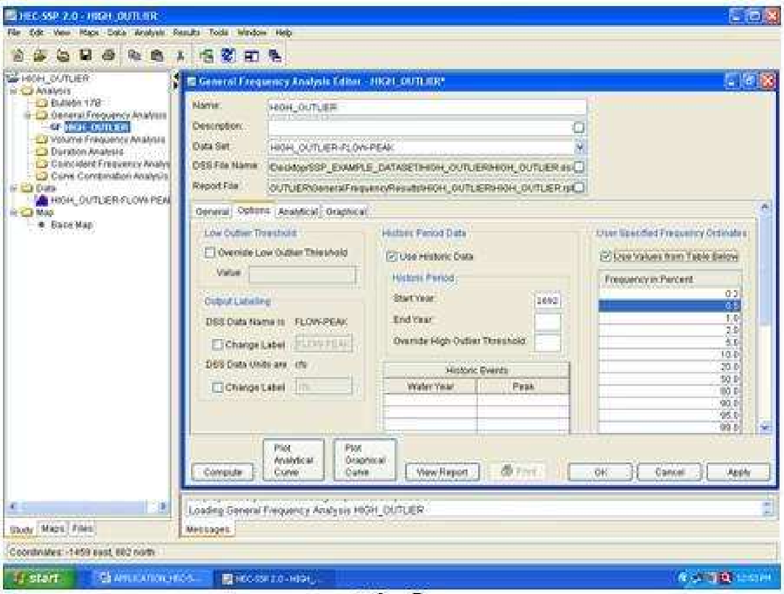


Fig.8

An attempt to compute D-index values for this set of data, outside the HEC-SSP environment, is placed at Table 5. Please mark that data, as highlighted in red in Table 3, populate this table for calculation of D-test. It could be seen, lower the value of D-test, the better the fit is.

Table 5

Rank	Qi Observed	Exceedance Probability as per Weibull	Qi Computed From frequency curve	Abs(col 2 – col 3)
1	69,400.0	0.0303	67382	2018
2	61,350.0	0.0606	56853	4497
3	58,100.0	0.0909	50892	7208
4	47,980.0	0.1212	46744	1236
5	45,630.0	0.1515	43550	2080
6	43,360.0	0.1818	40937	2423
			Total	19462

Average/Mean of the samples = 29495

D-Index = 19462/29495 = 0.659

Top of page

OUTLIERS

Outliers are values in a data set which plot significantly away from remainder of sample data (main body of the plot), and their deletion, retention and modification warrants prudent considerations of all of the factors giving birth to them. In Paragraph to follow, this aspect has been discussed at length.

The following equation is used to detect outliers:

$$Q_{High}, Q_{Low} = Q_{mean} \pm K_N * St\ Deviation$$

Where,

K_N is a frequency factor and varies according to sample size.

HEC-SSP automatically performs detection process; reports and analyzes the set of data accordingly.

Top of page

HANDLING DIVERSE SCENARIOS

The study covered in this module plots all annual peaks more or less closely aligned to theoretical distribution line (see Fig.6). It also means the absence of even a single peaks straying from rest of peaks. So, the number of outlier for this case is zero. Nevertheless, samples not as coherent as cited here are always a possibility; and it is likely that they may contain outliers - both high and low or either of the two; i.e. zero flows; or even historical floods outside the systematic (also continuous) records of annual peaks.

In dealing with such records, one, however, must be convinced about the authenticity of data, and should guard against entries of all inflated or dubious values in the analysis.

In HEC-SSP, presence of zero flows and low outliers are automatically detected and counted out by the software, and a conditional probability adjustment, to account for truncated values, is employed to estimate revised plotting position. Software also modifies values of statistical parameters to define theoretical distribution curve.

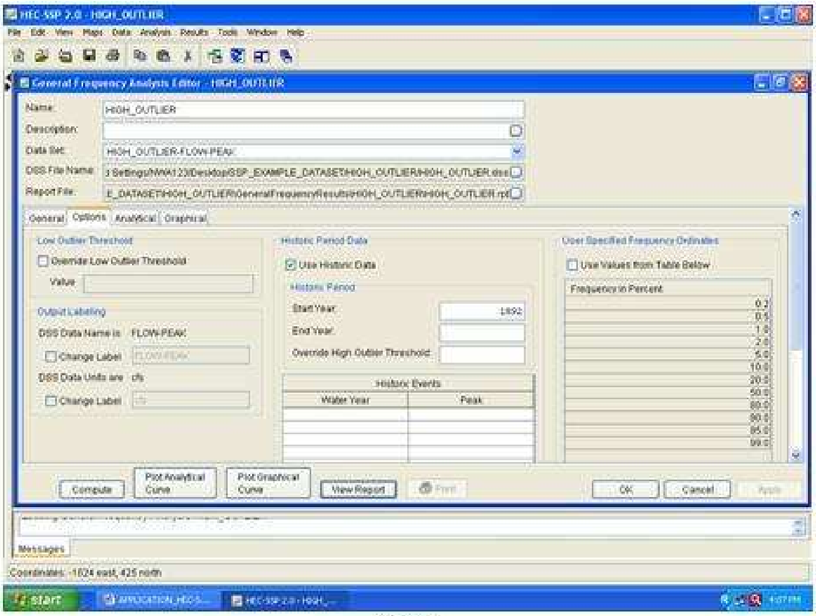


Fig.9

In a deviation from above, high outliers, so long as they are not suspected values, are not eliminated from the record as they are invaluable piece of the flow record and might be representative of longer period of record. For example, a flood value in a set of data, detected by software as outlier, could be the largest flood that has ever occurred in an extended period of time backward. Like other cases, HEC-SSP detects high outlier as well, and presents the analysis accounting for revised length of time period entered by user and number of high outliers detected by software itself. A computed curve returned by the software utilizes modified statistical parameters, i.e. mean, standard deviation, and skewness coefficient. Fig.9 is one of the windows of the software that lets user make appropriate entry to define Historical Period, if a high outlier falls beyond the systematic record. To gather more information about mathematical steps involved in dealing with varying cases such as cited here, interested users should refer to material referenced against Sl. No. 2, at the end of this chapter.

Here, we place sample data set (Table 6 & 7) for Flood Frequency Analysis under different conditions. User may key in this set of data in HEC-SSP to perform frequency analysis for different cases.

Table 6

High Outliers		***Low Outliers***	
27 Jun 1935	1,460	16 Apr 1929	8,750
09 Mar 1936	4,050	22 Oct 1929	15,500
26 May 1937	3,570	07 May 1931	4,060
14 Sep 1938	2,060	03 Feb 1939	6,300
11 Mar 1939	1,300	19 Apr 1940	3,130
04 Jun 1940	1,390	05 Apr 1941	4,160
10 Mar 1941	1,720	21 May 1942	6,700
03 Jun 1942	6,280	14 Oct 1942	22,400
16 Jun 1943	1,360	23 Mar 1944	3,880
12 May 1944	7,440	17 Sep 1945	8,050
11 Mar 1945	5,320	02 Jun 1946	4,020
28 Feb 1946	1,400	14 Mar 1947	1,600
24 Jun 1947	3,240	13 Apr 1948	4,460
16 Mar 1948	2,710	30 Dec 1948	4,230
04 Mar 1949	4,520	01 Feb 1950	3,010
18 Jun 1950	4,840	04 Dec 1950	9,150
27 Mar 1951	8,320	27 Apr 1952	5,100
30 Mar 1952	13,900	21 Nov 1952	9,820
07 Jun 1953	71,500	01 Mar 1954	6,200
21 Jun 1954	6,250	18 Aug 1955	10,700
09 Jul 1955	2,260	14 Mar 1956	3,880
12 Jul 1956	318	09 Feb 1957	3,420
04 Jul 1957	1,330	26 Mar 1958	3,240
30 Jun 1958	970	02 Jun 1959	6,800
31 May 1959	1,920	08 May 1960	3,740
28 Mar 1960	15,100	18 Feb 1961	4,700
01 Mar 1961	2,870	21 Mar 1962	4,380
28 Mar 1962	20,600	19 Mar 1963	5,190
01 Jun 1963	3,810	09 Jan 1964	3,960
08 Sep 1964	726	05 Mar 1965	5,600
01 Apr 1965	7,500	31 Dec 1965	4,670
09 Feb 1966	7,170	31 Dec 1966	7,080
18 Jun 1967	2,000	31 Dec 1967	4,640
20 Jul 1968	829	31 Dec 1968	536
04 Apr 1969	17,300	31 Dec 1969	6,680
03 Mar 1970	4,740	31 Dec 1970	8,360
31 Dec 1970	13,400	31 Dec 1971	18,700
31 Dec 1971	2,940	31 Dec 1972	5,210
31 Dec 1972	5,660		

Table 7

Zero Flows		Historic Events	
07 Feb 1932	4,260	**1843**	**15000**
28 Jan 1933	345		
31 Dec 1933	516	27 Mar 1932	**891**
07 Apr 1935	1,320	22 Aug 1933	**2,680**
12 Feb 1936	1,200	04 Mar 1934	**1,080**
12 Feb 1937	2,180	08 Jul 1935	**3,000**
10 Feb 1938	3,230	02 Jan 1936	**1,590**
08 Mar 1939	115	21 Feb 1937	**770**
26 Feb 1940	3,440	22 Jul 1938	**3,320**
03 Apr 1941	3,070	02 Feb 1939	**978**
23 Jan 1942	1,880	14 Mar 1940	**1,770**
20 Jan 1943	6,450	06 Feb 1941	**746**
28 Feb 1944	1,290	12 Aug 1942	**1,000**
01 Feb 1945	5,970	29 Dec 1942	**980**
24 Dec 1945	782	05 Jan 1944	**865**
29 Sep 1947	0	17 Sep 1945	**1,040**
29 Sep 1948	0	25 Dec 1945	**1,000**
11 Mar 1949	335	21 May 1947	**483**
04 Feb 1950	175	04 May 1948	**740**
02 Dec 1950	2,920	29 Dec 1948	**1,040**
11 Jan 1952	3,660	02 Aug 1950	**1,590**
06 Dec 1952	147	24 Nov 1950	**5,720**
29 Sep 1954	0	10 Mar 1952	**1,490**
18 Jan 1955	16	21 Nov 1952	**918**
22 Dec 1955	5,620	13 Dec 1953	**670**
23 Feb 1957	1,440	17 Aug 1955	**4,390**
01 Apr 1958	10,200		
15 Feb 1959	5,380	**End Year**	**1974**
09 Feb 1960	448		
29 Sep 1961	0		
14 Feb 1962	1,740		
31 Jan 1963	8,300		
21 Jan 1964	156		
29 Sep 1965	560		
29 Dec 1965	128		
20 Jan 1967	4,200		
29 Sep 1968	0		
24 Jan 1969	5,080		
28 Feb 1970	1,010		
20 Dec 1970	584		
29 Sep 1972	0		
10 Feb 1973	**1,510**		

As outlined in one of the preceding paragraphs, HEC-SSP has the ability to detect low outliers and/or zero flows and projecting the probability curve by introducing conditional probability adjustment. Contrary to this, analysis of high outliers and historical data do need a few entries by user. **Fig.10** deals with high outliers, where a peak discharge of 71,500 cumec is labeled as a high outlier by software, and an entry of 1892 by user in a cell by start year implies this peak is highest known value since year 1892. **Fig.11** deals with historical data; where user has entered historical flood value along with corresponding year. An entry of 1974 against end year signifies no significant flood since regular discharge recording ceased in year 1955.

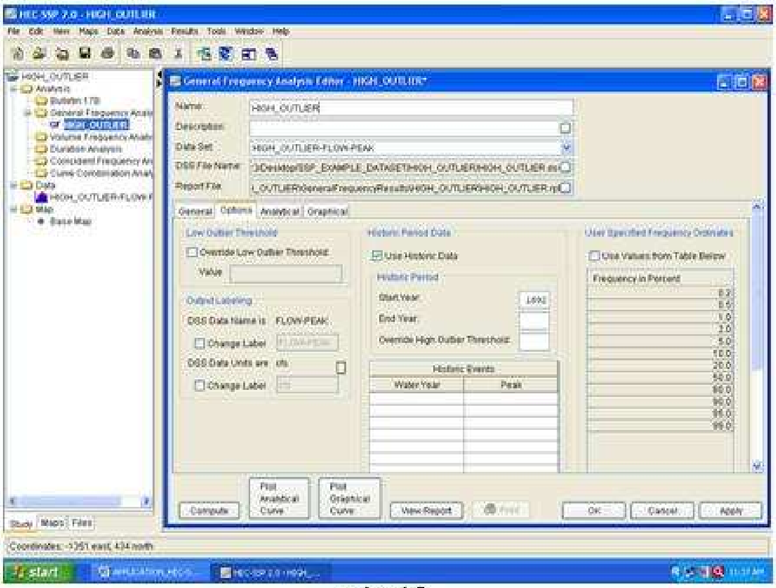


Fig.10

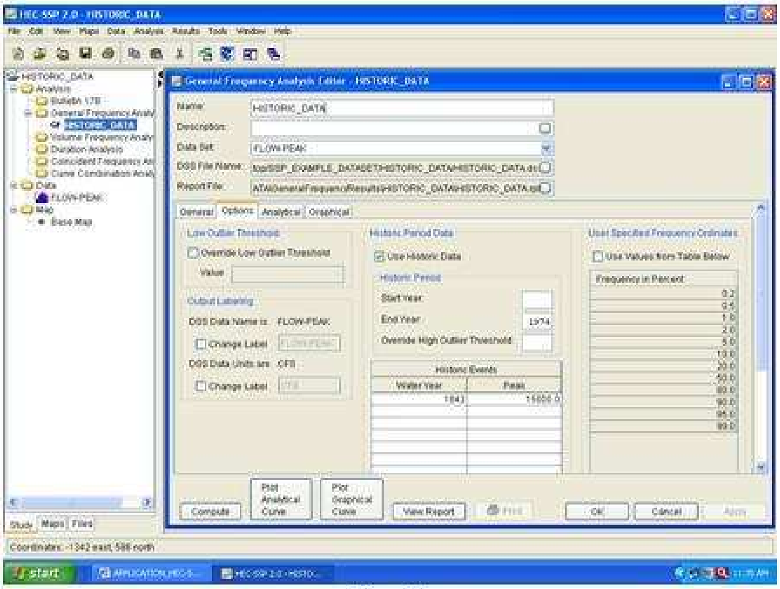


Fig.11

Top of page

REFERENCES

1. HEC-SSP User's Manual, available at http://www.hec.usace.army.mil/software/hec-ssp/documentation/HEC-SSP_20_Users_Manual.pdf
2. Guidelines for Determining Flood Flow Frequency- Bulletin 17B of the Hydrology Sub-Committee - A publication by US Department of the Interior Geological Survey Office of Water Data Coordination, http://water.usgs.gov/osw/bulletin17b/bulletin_17B.html
3. Ven Te Chow, David R Maidment, Larry W Mays, (International Edition 1988), Applied Hydrology, McGraw-Hill Book Company
4. Patra, K C, (2001), Hydrology & Water Resources Engineering, Narosa Publishing House
5. Hydrologic Frequency Analysis, Vol-3 published by US Army Corps of Engineers- 1975, http://www.hec.usace.army.mil/publications/IHDVolumes/IHD-3.pdf
6. Mutreja, K N, Applied Hydrology, Tata McGraw Hill Publishing Company Limited, N Delhi

7. Hydrology Project- Phase I (India), Training Module no.43

Top of page

CONTRIBUTOR

Anup Kumar Srivastava
Director
National Water Academy, Pune, India

Top of page

ACKNOWLEDGEMENT

Author of this module hereby acknowledges the invaluable support received from Shri D S Chaskar, and Dr R N Sankhua, both Directors, National Water Academy, CWC, Pune in preparation and presentation of this module in current shape.