Government of India & Government of The Netherlands

Technical Assistance
HYDROLOGY PROJECT

DHV CONSULTANTS & DELFT HYDRAULICS with HALCROW, TAHAL, CES, ORG & JPS

VOLUME 8
DATA PROCESSING AND ANALYSIS

OPERATIONAL MANUAL - PART III

FINAL PROCESSING AND ANALYSIS

Table of Contents

1 INTRODUCTION

1.1 GENERAL

The prime objective of the Hydrology Project is to develop a sustainable Hydrological Information System for 9 states in Peninsular India, set up by the state Surface Water and Groundwater Departments and by the central agencies (CWC and CGWB) with the following characteristics:

- Demand driven, i.e. output is tuned to the user needs
- Use of standardised equipment and adequate procedures for data collection and processing
- Computerised, comprehensive and easily accessible database
- Proper infrastructure to ensure sustainability.

This Hydrological Information System provides information on the spatial and temporal characteristics of water quantity and quality variables/parameters describing the water resources/water use system in Peninsular India. The information needs to be tuned and regularly be re-tuned to the requirements of the decision/policy makers, designers and researchers to be able to take decisions for long term planning, to design or to study the water resources system at large or its components.

This manual describes the procedures to be used to arrive at a sound operation of the Hydrological Information System as far as hydro-meteorological and surface water quantity and quality data are concerned. A similar manual is available for geo-hydrological data. This manual is divided into three parts:

a) **Design Manual,** which provides information for the design activities to be carried out for the further development of the HIS

b) **Reference Manual,** including references and additional information on certain topics dealt with in the Design Manual

c) **Field/Operation Manual,** which is an instruction book describing in detail the activities to be carried out at various levels in the HIS, in the field and at the data processing and data storage centres.

The manual consists of ten volumes, covering:

1. Hydrological Information System, its structure and data user needs assessment
2. Sampling Principles
3. Hydro-meteorology
4. Hydrometry
5. Sediment transport measurements
6. Water Quality sampling
7. Water Quality analysis
8. Data processing
9. Data transfer, storage and dissemination, and
10. SW-Protocols.

This Volume 8 deals with data processing and consists of an Operation Manual and a Reference Manual. The Operation Manual comprises 4 parts, viz:

Part I: Data entry and primary validation

Part II: Secondary validation

Part III: Final processing and analysis

Part IV: Data management

This Part III deals with the final step in data processing and with data analysis and reporting. The procedures described in the manual are to be carried out in the Data Processing Centres to ensure uniformity in data processing throughout the Project Area and to arrive at high quality data.

2 DATA VALIDATION USING REGRESSION

2.1 INTRODUCTION

In regression analysis a relation is made between a dependent variable Y (i.e. the one one wants to estimate) and one or a number of independent variables X_i. The objective(s) of establishing a regression model may be manifold, like:

1. Making forecasts/predictions/estimates on Y based on data of the independent variable(s)
2. Investigation of a functional relationship between two or more variables
3. Filling in missing data in the Y-series
4. Validation of Y-series

In data processing at a number of occasions regression analysis is applied:

- for validation and in-filling of missing water level data a relation curve is established based on a polynomial relation between the observations at two water level gauging stations either or not with a time-shift
- for transformation of water levels into discharge series a discharge rating curve is created. The commonly used discharge rating curves are of a power type regression equation, where for each range of the independent variable (gauge reading) a set of parameters is established.
- for estimation of rainfall (or some other variable) on the grid points of a grid over the catchment as a weighted average of observations made at surrounding stations with the aid of kriging also falls into the category of regression.

For validation of rainfall data use is made of a linear relation between observations at a base station and surrounding stations. The weights given to the surrounding stations is inverse distance based. Because the weights are not determined by some estimation error minimization criterion as is the case in regression analysis but rather on the geographical location of the observation stations those relations are not regression equations.

In the above examples of applications of regression analysis linear as well as non-linear relations have been mentioned:

- a **linear regression** equation is an equation which is linear in its coefficients:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

How the variables X_i behave does not matter and they may for example form an i^{th} order polynomial; hence the relation between Y and X may be non-linear.

$$Y = \alpha X_1^{\beta 1} X_2^{\beta 2} ..... X_i^{\beta i}$$

- in a **non-linear regression** equation the coefficients also appear as a power, like e.g.:

By considering a logarithmic transformation on the equation an non-linear equation as shown above can be brought back to a linear one. Then, the error minimisation is carried out on the logarithm rather than on the original values. Note that far more complex non-linear regression models can be considered but this is outside the scope of hydrological data processing.

In this module at first attention will be given to linear regression equations. Dependent on the number of independent variables in the regression equation a further distinction is made between:

- **simple linear regression**, where the dependent variable is regressed on one independent variable, and
- **multiple and stepwise linear regression**: the dependent variable is regressed on more than one independent variable. The difference between multiple and stepwise regression is that in multiple linear regression all independent variables brought in the analysis will be included in the regression model, whereas in stepwise regression the regression equation is built up step by step taking those independent variables into consideration first, which reduce the error variance most; the entry of new independent variables is continued until the reduction in the error variance falls below a certain limit. In some stepwise regression tools a distinction is made between **free** and **forced** independent variables: a forced variable will always be entered into the equation no matter what error variance reduction it produces, whereas a free variable enters only if the error variance reduction criterion is met.

The type of regression equation that is most suitable to describe the relation depends naturally on the variables considered and with respect to hydrology on the physics of the processes driving the variables. Furthermore, it also depends on the range of the data one is interested in. A non-linear relation may well be described by a simple linear regression equation, within a particular range of the variables in regression, as applies for example to annual runoff regressed on annual rainfall. In Figure 2.1 the general nature of such a relationship is shown.

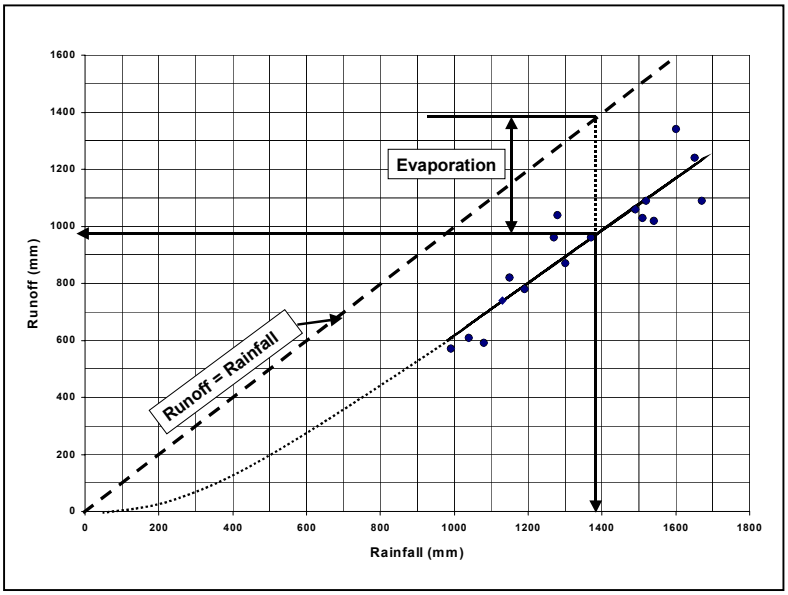


Figure 2.1:
General form of relation between annual rainfall and runoff

For low rainfall amounts the relation is highly non-linear in view of the strong varying rainfall abstractions due to evaporation. For very high rainfalls the abstraction is constant as it has reached its potential level; then the rainfall-runoff relation runs parallel to a line under 45^o with an offset equal to the potential evaporation and becomes a true linear relation. In between reaches may approximately be described by a linear equation. As long as the application of the relation remains within the observed range then there is no harm in using a linear relation, provided that the residuals distribute randomly about the regression equation over the range considered.

Another application of regression, which has not been discussed previously, is for validation of discharge data. A regression model is developed where runoff is regressed on rainfall (when monthly data are considered on rainfall in the same and in the previous month). By investigating the time-wise behaviour of the deviations from the regression line (i.e. the residuals) an impression is obtained about the stationarity of the rainfall-runoff relation (note: not of the stationarity of either rainfall or runoff!). Provided that the rainfall data are free of observation errors any non-stationary behaviour of the residuals may then be explained by:

- change in the drainage characteristics of the basin, or
- incorrect runoff data, which in turn can be caused by:
  - errors in the water level data, and/or
  - errors the discharge rating curve

Experience has shown that by applying double mass analysis on the observed and computed runoff (derived from rainfall) a simple but effective tool is obtained to validate the discharge data. (Alternatively, instead of using a regression model, also a conceptual rainfall-runoff model can be used but at the expense of a far larger effort.) Hence, a very important aspect of judging your regression model is to look carefully at the behaviour of the residuals, not only about the regression line as a function of X but also as a function of time. An example has been worked out on this application.

2.2 SIMPLE LINEAR REGRESSION

The most common model used in hydrology is based on the assumption of a linear relationship between two variables. Such models are called simple linear regression models, which have the following general form:

$$Y = \alpha + \beta X \tag{2.1}$$

Where: Y = dependent variable, also called response variable (produced by the regression model)

X = independent variable or explanatory variable, also called input, regressor, or predictor variable

α, β = regression coefficients

The actual observations on Y do not perfectly match with the regression equation and a residual ε is observed, see also Figure 2.2:

$$Y_i = \alpha + \beta X_i + \varepsilon \tag{2.2}$$

Hence:

$$Y_i - Y_i = \varepsilon_i \tag{2.3}$$

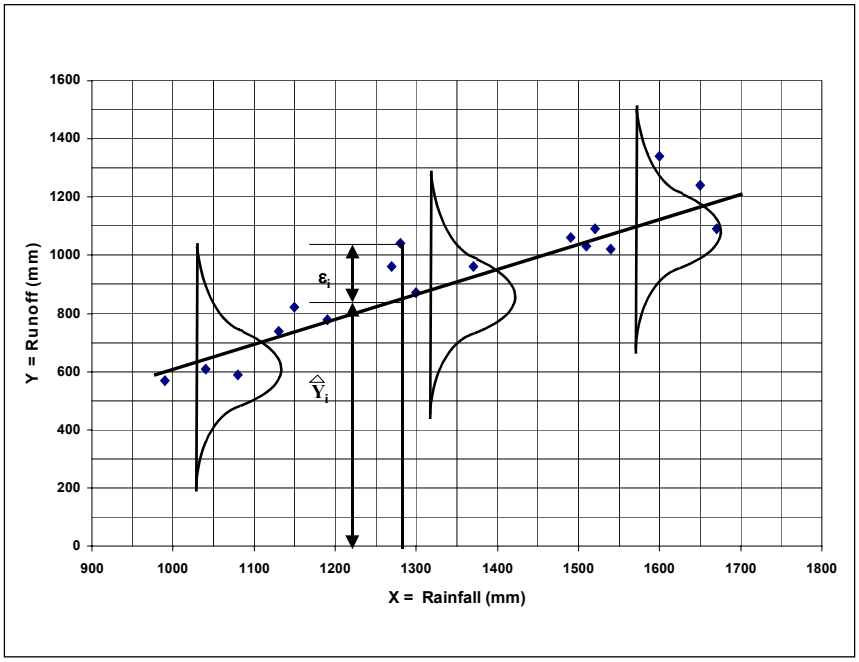


Figure 2.2: Explained Y_i and unexplained parts ε_i of Y under the assumption of a constant error distribution

The regression line will be established such that $E[(Y_i - Y)] = E[\varepsilon] = 0$, i.e. that it produces unbiased results and further that the variance of the residual σ_ε^2 is minimum. With respect to the residual it is assumed that its distribution about the regression line is normal and independent of X, hence for all values of X the distribution $F(\varepsilon)$ about the regression-line is the same, see Figure 2.2.

Now consider the following partitioning:

$$(Y - \overline{Y}) = (Y - \hat{Y}) + (\hat{Y} - \overline{Y}) = \varepsilon + (\hat{Y} - \overline{Y}) \quad \text{so:} \quad (Y - \overline{Y})^2 = (\varepsilon + (\hat{Y} - \overline{Y}))^2 \quad \text{and since:} \quad E[\varepsilon(\hat{Y} - \overline{Y})] = 0$$

$$E[(Y - \overline{Y})^2] = E[(\hat{Y} - \overline{Y})^2] + E[\varepsilon^2] \quad \text{or:}$$

$$\sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_\varepsilon^2 \tag{2.4}$$

Above equation expresses: **Total variance = explained variance + unexplained variance**

Hence, the smaller the unexplained variance (variance about regression) is, the larger the explained variance (or variance due to regression) will be. It also shows that the explained variance is always smaller than the total variance of the process being modelled. Hence the series generated by equation (2.1) will only provide a smoothed representation of the true process, having a variance which is smaller than the original, unless a random error with the characteristics of the distribution of the residual is added. Nevertheless, for individual generated values the estimate according to (2.1) is on average the best because $E[\varepsilon] = 0$. The root of the error variance is generally denoted as standard error.

In the following we will discuss:

- estimation of the regression coefficients
- measure for the goodness of fit
- confidence limits for the regression coefficients
- confidence limits for the regression equation
- confidence limits for the predicted values
- application of regression to rainfall-runoff analysis

Estimation of the regression coefficients

The estimators for the regression coefficients α and β , denoted by a and b respectively are determined by minimising $\Sigma\varepsilon^2$ Denoting the observations on X and Y by x_i and y_i this implies, that for:

$$M = \sum \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2 \tag{2.5}$$

to be minimum, the first derivatives of M with respect to a and b be set equal to zero:

$$\frac{\partial M}{\partial a} = -2\sum (y_i - a - bx_i) = 0 \tag{2.6a}$$

$$\frac{\partial M}{\partial b} = -2\sum x_i (y_i - a - bx_i) = 0 \tag{2.6b}$$

Above equations form the so called **normal equations.** From this it follows for a and b:

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})} = \frac{S_{XY}}{S_{XX}} \quad \text{and:} \quad a = \overline{y} - b\overline{x} \tag{2.7}$$

Since the procedure is based on minimising $\Sigma\varepsilon^2$, the estimators a and b for α and β are commonly called least squares estimators. This solution also satisfies $\Sigma\varepsilon = 0$ as is observed from (2.6a)

With 2.7 the simple regression equation can also be written in the form:

$$\hat{Y} - \overline{Y} = b(X - \overline{X}) \tag{2.8}$$

or with the definition of the correlation coefficient $r = S_{XY}/\sigma_X.\sigma_Y$:

$$\hat{Y} - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X}) \tag{2.9}$$

Measure for goodness of fit

By squaring (2.9) and taking the expected value of the squares it is easily observed by combining the result with (2.4) that the error variance can be written as:

$$\sigma_\varepsilon^2 = \sigma_Y^2(1 - r^2) \tag{2.10}$$

Hence, the closer r^2 is to 1 the smaller the error variance will be and the better the regression equation is in making predictions of Y given X. Therefore r^2 is an appropriate measure for the quality of the regression fit to the observations and is generally called the **coefficient of determination.**

It is stressed, though, that a high coefficient of determination **is not sufficient**. It is of great importance to investigate also the behaviour of the residual about the regression line and its development with time. If there is doubt about the randomness of the residual about regression then a possible explanation could be the existance of a non-linear relation. Possible reasons about absence of randomness with time have to do with changes in the relation with time as was indicated in the previous sub-chapter.

Confidence limits of the regression coefficients and model estimates

It can be shown, that, based on the sampling distributions of the regression parameters, the following estimates and confidence limits hold (see e.g. Kottegoda and Rosso, 1998).

Error variance

An unbiased estimate of the error variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-(a+bx_i))^2 = \frac{1}{n-2}\left(S_{YY}-\frac{S_{XY}^2}{S_{XX}}\right) \tag{2.11}$$

Note that n-2 appears in the denominator to reflect the fact that two degrees of freedom have been lost in estimating (α, β)

Regression coefficients

A (100-α) percent confidence interval for b is found from the following confidence limits:

$$CL_{\pm} = b \pm t_{n-2,1-\alpha/2}\frac{\hat{\sigma}_\varepsilon}{\sqrt{S_{XX}}} \tag{2.12}$$

A (100-α) percent confidence interval for a results from the following confidence limits:

$$CL_{\pm} = a \pm t_{n-2,1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}} \tag{2.13}$$

Regression line

A (100-α) percent confidence interval for the mean response to some input value x_0 of X is given by:

$$CL_{\pm} = a+bx_0 \pm t_{n-2,1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}} \tag{2.14}$$

Note that the farther away x_0 is from its mean the wider the confidence interval will be because the last term under the root sign expands in that way.

Prediction

A (100-α) percent confidence interval for a predicted value Y when X is x_0 follows from:

$$CL_{\pm} = a+bx_0 \pm t_{n-2,1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}} \tag{2.15}$$

It is observed by comparing (2.15) with (2.14) that in (2.15) full account of the error variance is added to last term. Hence, these confidence limits will be substantially wider that those for the mean regression line. Note however, since the multiplier of the standard error is under the root sign, the confidence limits in (2.15) are not simply obtained by adding t-times the standard error to the confidence limits of the regression line.

Example 2.1

In Table 2.1 some 17 years of annual rainfall and runoff data of a basin are presented. Regression analysis will be applied to validate the runoff series as there is some doubt about the rating curves applied before 1970. No changes took place in the drainage characteristics of the basin.

Year	Rainfall	Runoff	Year	Rainfall	Runoff	Year	Rainfall	Runoff
	(mm)	(mm)		(mm)	(mm)		(mm)	(mm)
1961	1130	592	1967	1670	872	1973	1650	1240
1962	1280	832	1968	1540	816	1974	1510	1030
1963	1270	768	1969	990	456	1975	1600	1340
1964	1040	488	1970	1190	780	1976	1300	870
1965	1080	472	1971	1520	1090	1977	1490	1060
1966	1150	656	1972	1370	960			

Table 2.1: Rainfall and runoff data (in mm) for the period 1961 to 1977

A time series plot of the rainfall and runoff series is presented in Figure 2.3a. A simple linear regression equation is established for R = f(P), see Figure 2.3b. The regression equation reads:

R = -530 + 1.025xP, with σ_ε = 130 mm and the coefficient of determination r^2 = 0.75.

From Figure 2.3c it is observed that the trend line for the residuals runs exactly parallel to the axis of the independent variable (=rainfall) and is zero throughout meaning that the regression was properly performed mathematically. It appears, though, that the assumption of a constant error distribution is not fulfilled: the variation about regression clearly increases with increase in the independent variable. The time series plot of the residuals when subjected to a trend analysis shows a clear upward trend. This looks like a gradual change in the rainfall-runoff relation in the period of observation. However, as stated above, no changes took place in the drainage characteristics of the basin. The plot of accumulated residuals shown in Figure 2.3e features a distinct change in the residuals as from 1970 onward. A double mass analysis on the observed runoff against the runoff computed by regression on the rainfall also shows a distinct break around 1970, see Figure 2.3f. From this analysis it is revealed that the runoff data prior to 1970 have been underestimated by 20%. Accordingly, a correction was applied to the runoff.

The corrected time series is shown in Figure 2.4a. The results of the regression analysis on the corrected data are presented in the Figures 2.4b to 2.4e. The regression equation now reads:

R = -303 + 0.920xP, with σ_ε = 88.3 mm and the coefficient of determination r^2 = 0.84. It is observed that the coefficient of determination has increased substantially and consequently the standard error has decreased; its value is now over 30% less. The behaviour of the residual as a function of the dependent variable and as a function of time are shown in Figures 2.4c and d. Figure 2.4c shows that the variance of the residual is now fairly constant with X. From Figure 2.4d it is observed that no time effect is present anymore. In Figure 2.4e the 95% confidence limits about the regression line and of the predictions are shown. The computations are outlined in Table 2.2.

Year	**X=Rainfall**	**Y=Runoff**	**$(X-Xm)^2$**	**Yest**	**CL1**	**CL2**	**UC1**	**LC1**	**UC2**	**LC2**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1961	1130	740	44100	737	64	199	801	673	936	538
1962	1280	1040	3600	875	47	194	922	827	1069	681
1963	1270	960	4900	866	48	194	914	818	1060	671
1964	1040	610	90000	654	78	204	732	576	858	450
1965	1080	590	67600	691	72	201	762	619	892	489
1966	1150	820	36100	755	61	198	816	694	953	557
1967	1670	1090	108900	1234	84	206	1317	1150	1440	1028
1968	1540	1020	40000	1114	62	198	1176	1052	1312	916
1969	990	570	122500	608	87	207	695	521	815	400
1970	1190	780	22500	792	56	196	848	736	988	596
1971	1520	1090	32400	1096	60	198	1155	1036	1293	898
1972	1370	960	900	958	46	194	1004	911	1152	764
1973	1650	1240	96100	1215	80	205	1295	1135	1420	1011
1974	1510	1030	28900	1086	58	197	1145	1028	1284	889
1975	1600	1340	67600	1169	72	201	1241	1098	1371	968
1976	1300	870	1600	893	46	194	940	847	1087	699
1977	1490	1060	22500	1068	56	196	1124	1012	1264	872
Xm	1340	S_{XX}	790200							

Table 2.2: Example computation of confidence limits for regression analysis

In the computations use is made of equations (2.14) and (2.15). In Column 2 the mean of X is computed and the sum of Column 4 is S_{XX}. In the Columns 6 and 7 the last term of equations (2.14) and (2.15) are presented. Note that $t_{n-2,1-\alpha/2} = 2.131$ and $\sigma_\varepsilon = 88.3$ mm. Column 6 and 7 follow from:

$$CL1 = t_{n-1,1-\alpha/2}\sigma_\varepsilon\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{XX}}} = 2.131x88.3\sqrt{\frac{1}{17}+\frac{(1130-1340)^2}{790200}} = 2.131x88x0.34 = 64 \text{ mm}$$

$$CL2 = t_{n-1,1-\alpha/2}\sigma_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{XX}}} = 2.131x88.3\sqrt{1+\frac{1}{17}+\frac{(1130-1340)^2}{790200}} = 2.131x88.3x1.06 = 199mm$$

The upper and lower confidence limits of the mean regression line then simply follow from Column 5 + 6 and Column 5 – 6, whereas the confidence limits for the predicted value (Columns 10 and 11) are derived from Column 5 + 7 and Column 5 – 7. It may be observed that the width of the confidence interval is minimum at the mean value of the independent variable. The variation of the width with the independent variable is relatively strongest for the confidence limits of the mean relation. The confidence limits for the prediction are seen to vary little with the variation in the independent variable, since the varying part under the root (i.e. the last term) is seen to be small compared to 1.

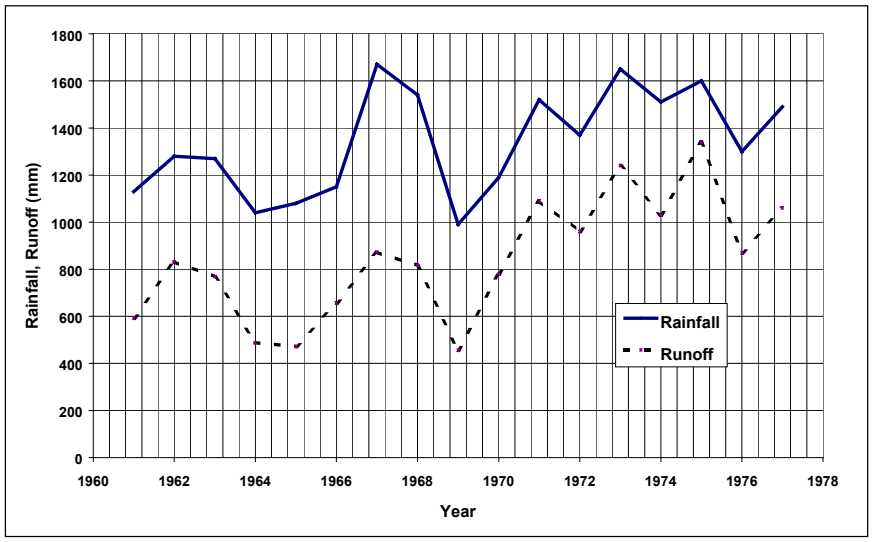


Figure 2.3a: Rainfall-runoff record 1961-1977

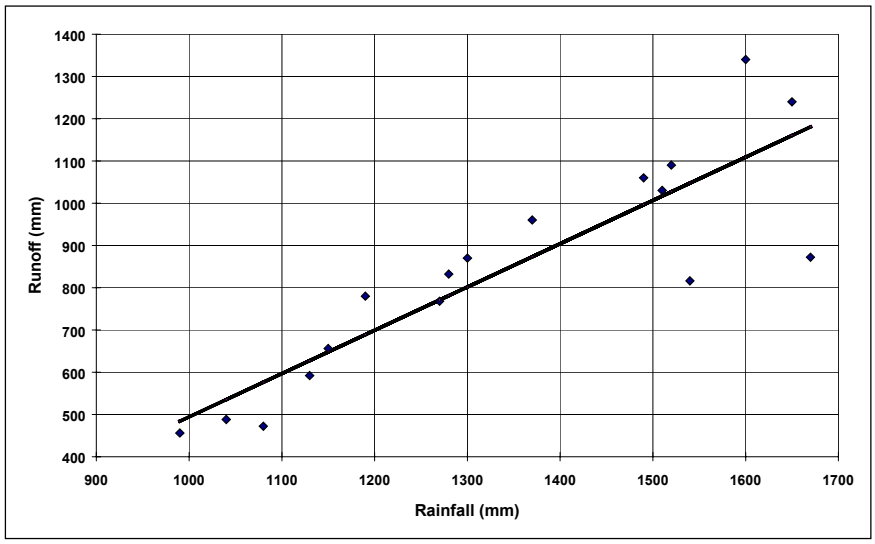


Figure 2.3b: Regression fit Rainfall-runoff

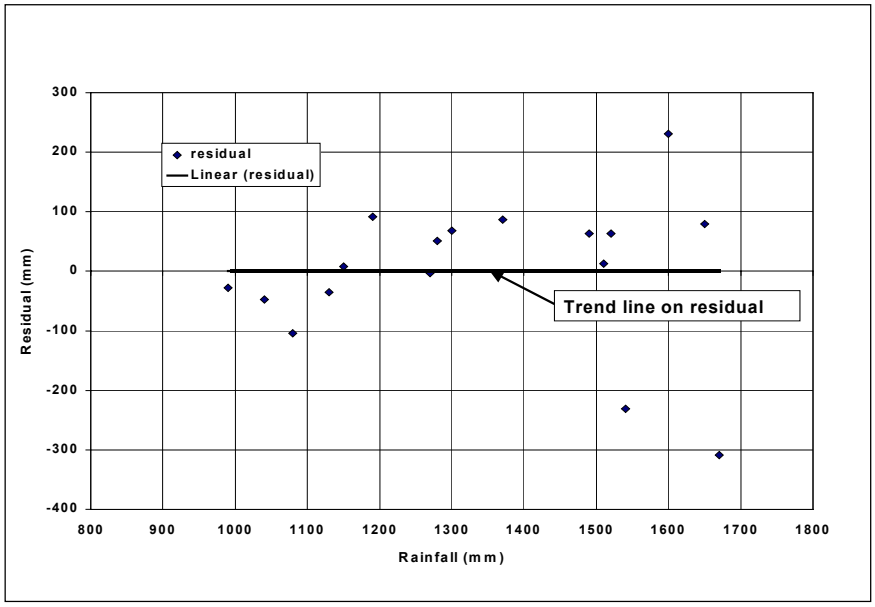


Figure2.3c: Plot of residual versus versus rainfall

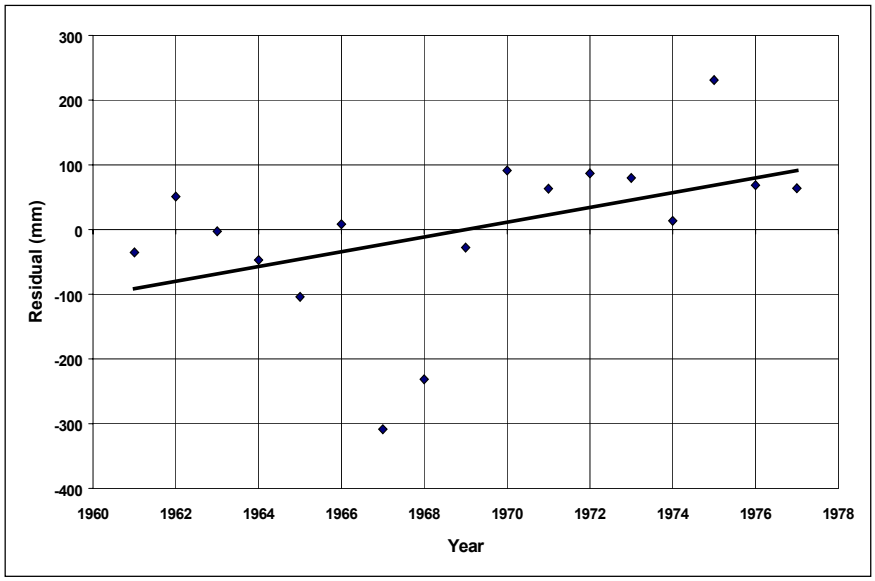


Figure 2.3d: Plot of residual versus time

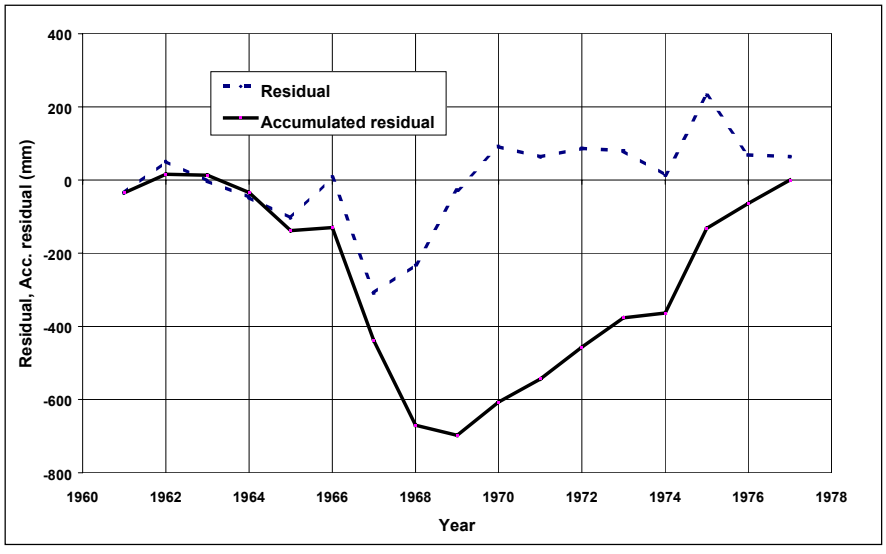


Figure 2.3e: Plot of accumulated residual

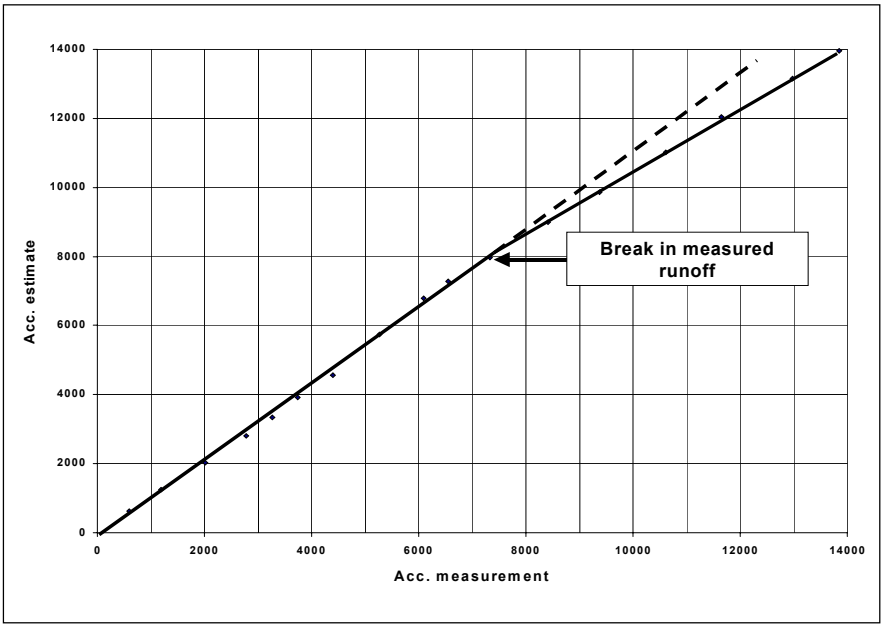


Figure 2.3f: Double mass analysis Observed versus computed runoff

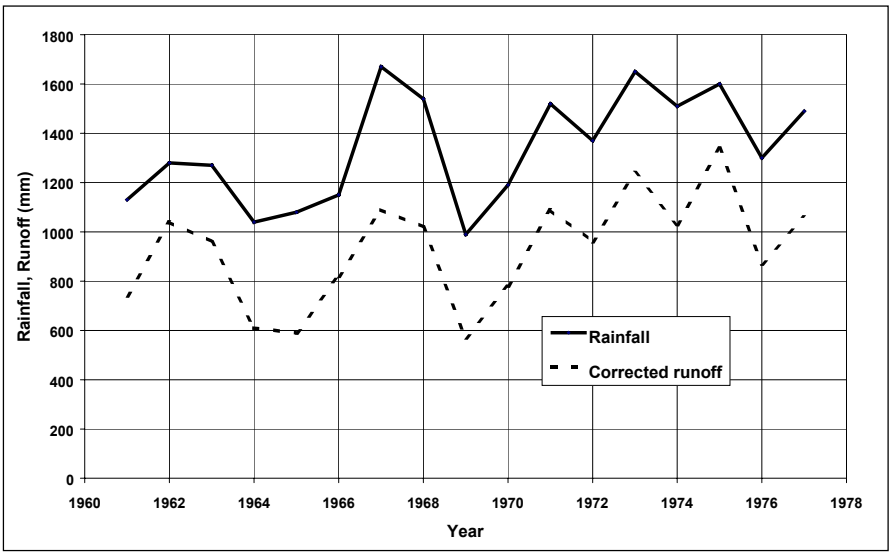


Figure 2.4a: Plot of rainfall and corrected runoff

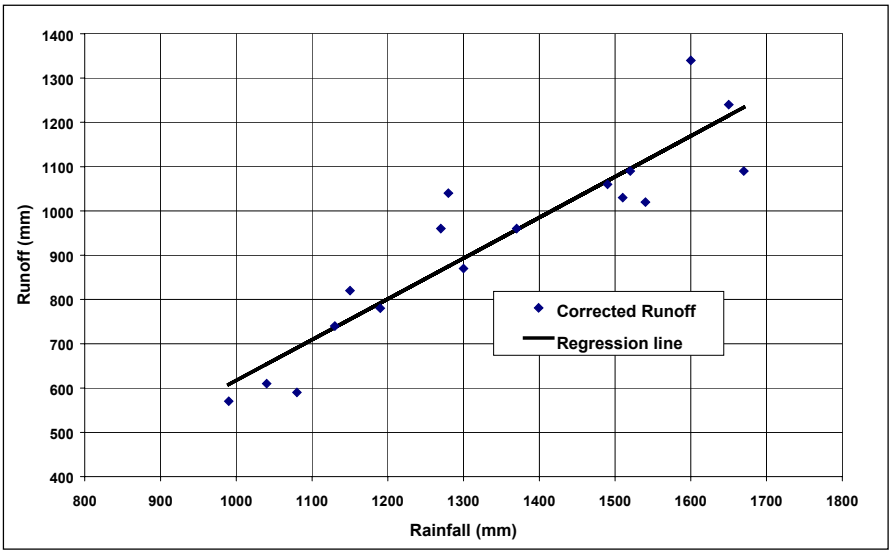


Figure 2.4b: Plot of rianfall-runoff regression, corrected runoff data

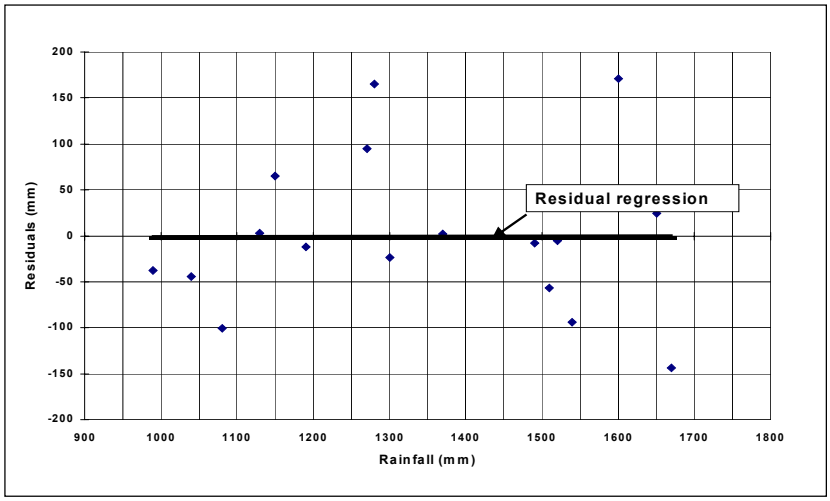


Figure 2.4c: Plot of residual versus rainfall

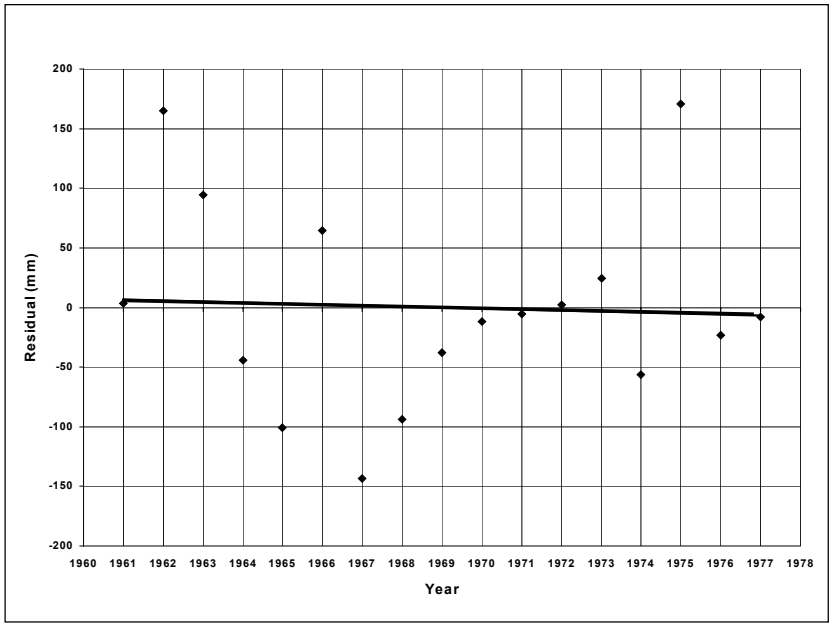


Figure 2.4d: Plot of residual versus time

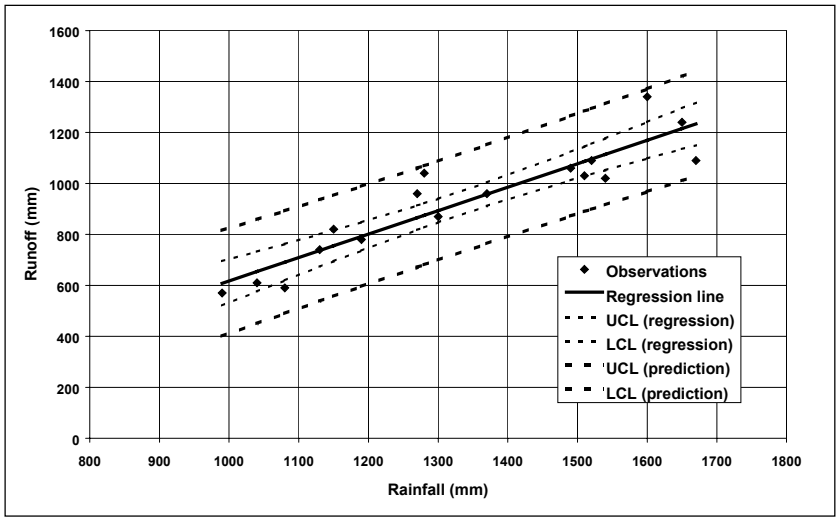


Figure 2.4e: Regression line with confidence limits for the mean regression and predicted values

Extrapolation:

The extrapolation of a regression equation beyond the range of X used in estimating α and β is discouraged for two reasons. First as can be seen from Figure 2.4e the confidence intervals on the regression line become wide as the distance from $\bar{X}$ is increased. Second the relation between Y and X may be non-linear over the entire range of X and only approximately linear for the range of X investigated. A typical example of this is shown in Figure 2.1.

2.3 MULTIPLE LINEAR REGRESSION

Often we wish to model the dependent variable as a function of several other quantities in the same equation. In extension to the example presented in the previous sub-chapter monthly runoff is likely to be dependent on the rainfall in the same month and in the previous month(s). Then the regression equation would read:

$$R(t) = \alpha + b_1 P(t) + \beta_2 P(t-1) + \ldots. \quad (2.16)$$

In this section the linear model is extended to include several independent variables.

A general linear model of the form:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots. \beta_p X_p + \varepsilon \quad (2.17)$$

Is discussed, where Y is a dependent variable, X_1, X_2, …. X_p are independent variables and β_1, β_2, ….. β_p are unknown parameters. This model is linear in the parameters β_j .Note that the form (2.16) can always be brought to the form (2.17) with the constant α by considering the variables Y and X_i centered around their mean values, similar to (2.8).

In practice n observations would be available on Y with the corresponding n observations on each of the p independent variables. Thus n equations can be written, one for each observation. Essentially we will be solving n equations for the p unknown parameters. Thus n must be equal to or greater than p. In practice n should be at least 3 or 4 times as large as p. The n equations then read

$$\mathbf{Y} = \mathbf{X}\ \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.18)$$

where: **Y** = (n x 1)-data column vector of the centered dependent variable $\left(Y_i - \bar{Y}\right)$

X = (n x p)-data matrix of the centered independent variables $\left(X_{i1} - \bar{X}_1\right),..\left(X_{ip} - \bar{X}_p\right)$

$\boldsymbol{\beta}$ = (p x 1)- column vector, containing the regression coefficients

$\boldsymbol{\varepsilon}$ = (n x 1)-column vector of residuals

The residuals are conditioned by:

$$E[\mathbf{e}] = 0 \quad (2.19)$$

$$Cov(\mathbf{e}) = \sigma_\varepsilon^2\, \mathbf{I} \quad (2.20)$$

where: **I** = (n x n) diagonal matrix with diagonal elements = 1 and off-diagonal elements = 0

σ_ε^2 = variance of (Y|X)

According to the least squares principle the estimates **b** of β are those which minimise the residual sum of squares $\varepsilon^T\varepsilon$. Hence:

$$\varepsilon^T\varepsilon = (Y - X\beta)^T (Y-X\beta) \quad (2.21)$$

is differentiated with respect to **b**, and the resulting expression is set equal to zero. This gives:

$$X^TXb = X^TY \quad (2.22)$$

called the **normal equations**, where β is replaced by its estimator **b**. Multiplying both sides with $(X^TX)^{-1}$ leads to an explicit expression for **b**:

$$b = (X^TX)^{-1} X^TY \quad (2.23)$$

The properties of the estimator **b** of β are:

$$E[b] = \beta \quad (2.24)$$

$$Cov(b) = \sigma_\varepsilon^2 (X^TX)^{-1} \quad (2.25)$$

By (2.21) and (2.22) the total adjusted sum of squares Y^TY can be partitioned into an explained part due to regression and an unexplained part about regression, as follows:

$$Y^TY = b^TX^TY + e^Te. \quad (2.26)$$

where: $(Xb)^TY$ = sum of squares due to regression

$e^Te.$ = sum of squares about regression, with ε replaced by **e** due to the replacement of β by **b**.

In words this reads:

Total sum of squares about the mean = regression sum of squares + residual sum of squares

The mean squares values of the right hand side terms in (2.26) are obtained by dividing the sum of squares by their corresponding degrees of freedom. If **b** is a (p x 1)-column vector, i.e. there are p-independent variables in regression, then the regression sum of squares has p-degrees of freedom. Since the total sum of squares has (n-1)-degrees of freedom (note: 1 degree of freedom is lost due to the estimation of $\bar{y}$), it follows by subtraction that the residual sum of squares has (n-1-p)-degrees of freedom. It can be shown that the residual mean square s_e^2:

$$s_e^2 = \frac{e^2e}{n - 1 - p}$$

Is an unbiased estimate of σ_ε^2 .The estimate s_e of σ_ε is the **standard error of estimate**.

The analysis of variance table (ANOVA) summarises the sum of squares quantities

Source	Sum of squares	Degrees of freedom	Mean squares
Regression (b_1, …., b_p)	$S_R = \mathbf{b}^T\mathbf{X}^T\mathbf{Y}$	p	$MS_R = \mathbf{b}^T\mathbf{X}^T\mathbf{Y}/p$
Residual (e_1, …., e_n)	$S_e = \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T\mathbf{Y} - \mathbf{b}^T\mathbf{X}^T\mathbf{Y}$	n-1-p	4 $MS_e = s_e^2 = \mathbf{e}^T\mathbf{e}/(n-1-p)$
Total (adjusted for $\bar{y}$)	$S_Y = \mathbf{Y}^T\mathbf{Y}$	n-1	$MSY = s_Y^2 = \mathbf{Y}^T\mathbf{Y}/(n-1)$

Table 2.3: Analysis of variance table (ANOVA)

As for the simple linear regression a measure for the quality of the regression equation is the **coefficient of determination**, defined as the ratio of the explained or regression sum of squares and the total adjusted sum of squares.

$$R_m^2 = \frac{\mathbf{b}^T\mathbf{X}^T Y}{\mathbf{Y}^T\mathbf{Y}} = 1 - \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{Y}^T\mathbf{Y}} \qquad (2.28)$$

The coefficient should be adjusted for the number of independent variables in regression. Then, instead of the sum of squares ratio in the most rand-hand side term the mean square ratio is used. So with the adjustment:

$$R_{ma}^2 = 1 - \frac{MS_e}{MS_Y} = 1 - \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{Y}^T\mathbf{Y}} \quad \frac{(n-1)}{(n-p-1)} = \cdots 1 - \left(1 - R_m^2\right)\left(\frac{n-1}{n-p1}\right) \qquad (2.29)$$

From this it is observed that $R_{ma}^2 < R_m^2$ except for $R_m = 1$ (i.e. a perfect model) where R_m is the multiple correlation coefficient and R_{ma} the adjusted multiple correlation coefficient.

Reference is made to the annex to the HYMOS manual for statistical inference on the regression coefficients.

Confidence Intervals on the Regression Line

To place confidence limits on Y_0 where $Y_0 = \mathbf{X}_0\mathbf{b}$ it is necessary to have an estimate for the variance of. $\hat{Y}_0$. Considering Cov**(b)** as given in (2.25) the variance Var($\hat{Y}_0$) is given by (Draper and Smith 1966):

$$\text{Var}(\hat{Y}_0) = s_e^2\mathbf{X}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0^T \qquad (2.30)$$

The confidence limits for the mean regression equation are given by

$$CL_{\pm} = X_0 b \pm t_{1-a/2,n-p}\sqrt{\text{Var}(\hat{Y}_0)} \qquad (2.31)$$

Comments

A common situation in which multiple regression is used is when one dependent variable and several independent variables are available and it is desired to find a linear model that is developed does not necessarily have to contain all of the independent variables. Thus the points of concern are: (1) can a linear model be used and (2) what independent variable should be included?

A factor complicating the selection of the model is that in most cases the independent variables are not statistically independent at all but are correlated. One of the first steps that should be done in a regression analysis is to compute the correlation matrix.

Retaining variables in a regression equation that are highly correlated makes the interpretation of the regression coefficients difficult. Many times the sign of the regression coefficient may be the opposite of what is expected if the corresponding variable is highly correlated with another independent variable in the equation.

A common practice in selecting a multiple regression model is to perform several regressions on a given set of data using different combinations of the independent variables. The regression that "best" fits the data is then selected. A commonly used criterion for the "best" fit is to select the equation yielding the largest value of R_{ma}^2.

All of the variables retained in a regression should make a significant contribution to the regression unless there is an overriding reason (theoretical or intuitive) for retaining a non-significant variable. The variables retained should have physical significance. If two variables are equally significant when used alone, but are not both needed, the one that is easiest to obtain should be used.

The number of coefficients estimated should not exceed 25 to 35 percent of the number of observations. This is a rule of thumb used to avoid "over-fitting" whereby oscillations in the equation may occur between observations on the independent variables.

2.4 STEPWISE REGRESSION

One of the most commonly used procedures for selecting the "best" regression equations is **stepwise regression.** This procedure consists of building the regression equation one variable at a time by adding at each step the variable that explains the largest amount of the remaining unexplained variation. After each step all the variables in the equation are examined for significance and discarded if they are no longer explaining a significant variation. Thus the first variable added is the one with the highest simple correlation with the dependent variable. The second variable added is the one explaining the largest variation in the dependent variable that remains unexplained by the first variable added. At this point the first variable is tested for significance and retained or discarded depending on the results of this test. The third variable added is the one that explains the largest portion of the variation that is not explained by the two variables already in the equation. The variables in the equation are then tested for significance. This procedure is continued until all of the variables not in the equation are found to be insignificant and all of the variables in the equation are significant. This is a very good procedure to use but care must be exercised to see that the resulting equation is rational.

The real test of how good is the resulting regression model, depends on the ability of the model to predict the dependent variable for observations on the independent variables that were not used in estimating the regression coefficients. To make a comparison of this nature, it is necessary to randomly divide the data into two parts. One part of the data is then used to develop the model and the other part to test the model. Unfortunately, many times in hydrologic applications, there are not enough observations to carry out this procedures.

2.5 TRANSFORMING NON LINEAR MODELS

Many models are not naturally linear models but can be transformed to linear models. For example

$$Y = \alpha X^{\beta} \tag{2.32}$$

is not a linear model. It can be linearized by using a logarithmatic transformation:

$$\ln Y = \ln \alpha + \beta \ln X \tag{2.33}$$

or

$$Y_T = \alpha_T + \beta_T X_T \quad (2.34)$$

Where: $Y_T = \ln Y$
$\alpha_T = \ln \alpha$
$\beta_T = \beta$
$X_T = \ln X$

Standard regression techniques can now be used to estimate α_T and β_T for the transformed equation and α and β estimated from the logarithmatic transformation. Two important points should be noted.:

- First the estimates of α and β obtained in this way will be such that $\sum(Y_{Ti} - \hat{Y}_{Ti})^2$ is a minimum and not such that $\sum(Y_i - \hat{Y}_i)^2$ is a minimum.
- Second the error term on the transformed equation is additive ($Y_T = \alpha_T + \beta_T X_T + \varepsilon_T$) implying that it is multiplicative on the original equation i.e. $Y = \alpha X^{\beta}\varepsilon$. These errors are related $\varepsilon_T = \ln \varepsilon$. The assumptions used in hypothesis testing and confidence intervals must now be valid for ε_T and the tests and confidence intervals made relative to the transformed model.

In some situations the logarithmic transformation makes the data conform more closely to the regression assumptions. The normal equations for a logarithmic transformation are based on a constant percentage error along the regression line while the standard regression is based on a constant absolute error along the regression line

2.6 FILLING IN MISSING DATA

An important application of regression analysis is the use of a regression equation to fill in missing data. In Part II of Volume 8 attention has been given to fill in missing rainfall and water level data. In this section attention will be given to filling in missing runoff data using rainfall as input. Typically, such techniques are applied to time series with time intervals of a decade, a month or larger.

Generally a regression of the type presented in equation (2.16) is applicable. Assume that the objective is to fill in monthly data. The regression coefficients are likely to be different for each month, hence the discharge in month k of year m is computed from:

$$Q_{k,m} = a_k + b_{1k}P_{k,m} + b_{2k}P_{k-1,m} + s_{e,k}e \quad (2.35)$$

It is observed that the regression coefficients are to be determined for each month in the year. The last term is added to ensure that the variance of the discharge series is being preserved. It represents the unexplained part of the functional relationship. Omitting the random component will result in a series with a smaller variance, which creates an inhomogeneity in the series and certainly does affect the overall monthly statistics. Dependent on the application it has to be decided whether or not to include the random components. If, however, a single value is to be estimated the random component should be omitted as the best guess is to rely on the explained part of equation (2.35); E[e] = 0. Note that the calibration of such a model will require at least some 15 to 20 years of data, which might be cumbersome occasionally.

Experience has shown that for a number of climatic zones the regression coefficients do not vary much from month to month, but rather vary with the wetness of the month. Two sets of parameters are then applied, one set for wet conditions and one for dry conditions with a rainfall threshold to discriminate between the two parameter sets The advantage of such a model is that less years with concurrent data have to be available to calibrate it, with results only slightly less than with (2.35) can be achieved. The use of a threshold is also justifiable from a physical point of view as the abstractions from rainfall basically create a non-linearity in the rainfall-runoff relationship.

Concurrent rainfall and runoff data should be plotted to investigate the type of relationship applies. One should never blindly apply a particular model.

3 DATA VALIDATION USING A HYDROLOGICAL MODEL

3.1 GENERAL

3.1.1 WHAT IS A HYDROLOGICAL MODEL

A physical or mathematical model is a simplified version of reality that is amenable to testing.

A hydrological rainfall runoff model is a means of representing the transformation of an input of rainfall over a catchment area to runoff at a specified outflow point. To simplify the complex processes operating over the catchment and beneath its surface, the hydrology of the catchment is conceived as a series of interlinked processes and storages.

Storages are considered as reservoirs for which water budgets are kept and the processes which control the transfer of water from one storage to the next are described mathematically by logical rules and equations to define, initiation, rate and cessation. Storages are allocated a total capacity and an actual content at any particular moment in time.

Complex catchment processes can be simplified and represented in a wide variety of ways and a large number of models have been developed. The selection of a model type depends on the uses to which it will be put and the availability of measured information on inputs, outflows and storages. The data processing software HYMOS has selected and extended/adapted the Sacramento Model which has had previous wide use and testing. It is physically realistic, it can operate with the amount of information typically available and requires limited computer power. Many more sophisticated models exist but all are limited by availability and quality of data and for most applications there is little to be gained by the use of more sophisticated models.

3.1.2 OPTIMISATION, CALIBRATION, VERIFICATION AND APPLICATION

For a particular catchment the operation of the model depends on the selection of the value of storage capacities and the parameters of the linking equations. This may be done by estimation based on the physical properties of the catchment, e.g. soil type and impermeable area, or they may be computed by the process of optimisation.

Optimisation is the means by which, using a measured input of rainfall (and evapotranspiration) and successive computer runs, the parameters of the model are progressively adjusted to improve the correspondence between the gauged outflow (Q_{gaug}) and the outflow simulated by the computer run (Q_{sim}). Optimisation may be done by manual adjustment of parameters or by automatic optimisation. Optimisation makes use of quantitative measures of goodness of fit (the objective function) such as:

$$F=\sum_{i=1}^{n}(Q_{t,gaug}-Q_{t,sim})^2$$

for the n values of the time series being optimised. In automatic optimisation the objective function (F) is minimised by a search through the parameter space in a defined and efficient way. The model is run with a given set of parameters, the objective function is calculated, the parameters are adjusted and the process repeated until the value of F shows no further improvement.

The entire process of parameter estimation and optimisation using measured time series of input rainfall and outflow is referred to as calibration. Calibration is subject to uncertainty in simulation and results in disagreement between recorded and simulated output. Following may be the sources of uncertainities:

a) random or systematic errors in the input data, e.g. precipitation or evapotranspiration used to represent the input conditions in time and space for the catchment
b) Random or systematic errors in recorded output data, i.e. measured discharges for comparison with simulated discharges
c) Errors due to non-optimal parameter values
d) Errors due to incomplete or biased model structure.

During calibration only error source (c) is minimised, whereas the disagreement between simulated and recorded output is due to all four error sources. Measurement errors (a) and (b) serve as "background noise" and give a minimum level of disagreement below which further parameter or model adjustments will not improve the results. The objective of calibration is therefore to reduce error source (c) until it insignificant compared with the error sources (a) and (b).

It is usual to withhold a part of the measured data from calibration. This can then be used to verify the performance of the model by using the calibrated parameters with the new data (without optimisation) to determine the objective function and goodness of fit. Verification is a means of ensuring that the optimised parameters are a true representation of the physical behaviour of the catchment and not simply a consequence of the model structure.

The calibrated and verified model is then ready for application where the rainfall input is known but the outflow is unknown.

3.1.3 USES OF HYDROLOGICAL RAINFALL RUNOFF MODELS

Rainfall runoff models have a wide variety of uses which include:

- filling in and extension of discharge series
- validation of runoff 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.)

The use of the model for the HIS is normally limited to the filling in of missing values in discharge series and the correction of suspect values. It is not usually applied to short sequences of missing data but to gaps of several months in length. The time and effort involved in the calibration of the model does not normally justify application to short gaps; though the model may be thus used if it has previously been calibrated for the same catchment.

3.2 THE SACRAMENTO MODEL

3.2.1 OUTLINE OF MODEL COMPONENTS

The application of the Sacramento model as integrated in HYMOS is based on a semi-distributed approach. It implies that a catchment is divided into a number of segments, which are interconnected by channel reaches as shown in Figure 3.1.

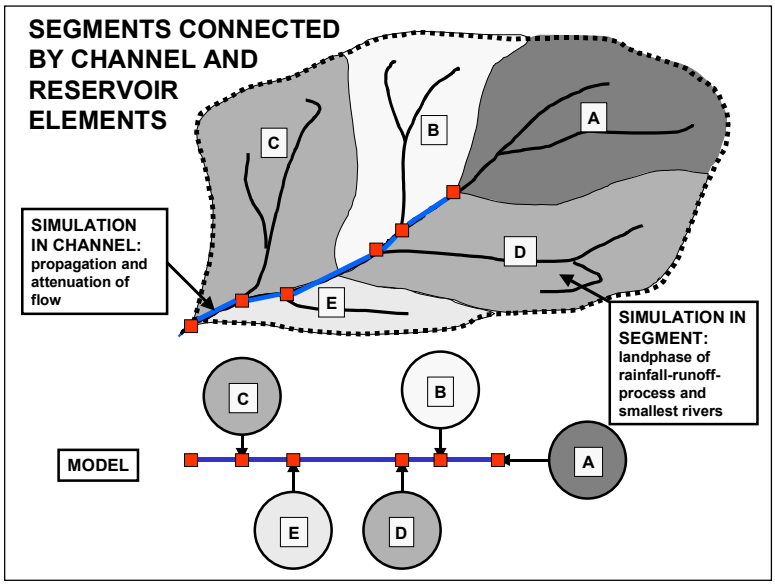


Figure 3.1:
Semi-distributed approach towards rainfall-runoff simulation

In a segment rainfall is transformed into runoff to the main river system. An explicit moisture accounting lumped parameter model is used to carry out the transformation. Important elements in the segment phase is the computation of the rainfall abstractions and the response time of the catchment to rainfall input, for which the time of concentration is an indicator, see Figure 3.2. Within a segment areal homogeneity of rainfall input and basin characteristics is assumed. The contributions of the segments to the main river are routed through the river network where the main features are travel time and flood wave damping. Generally a Muskingum layer approach or unit hydrograph technique is used for the routing.

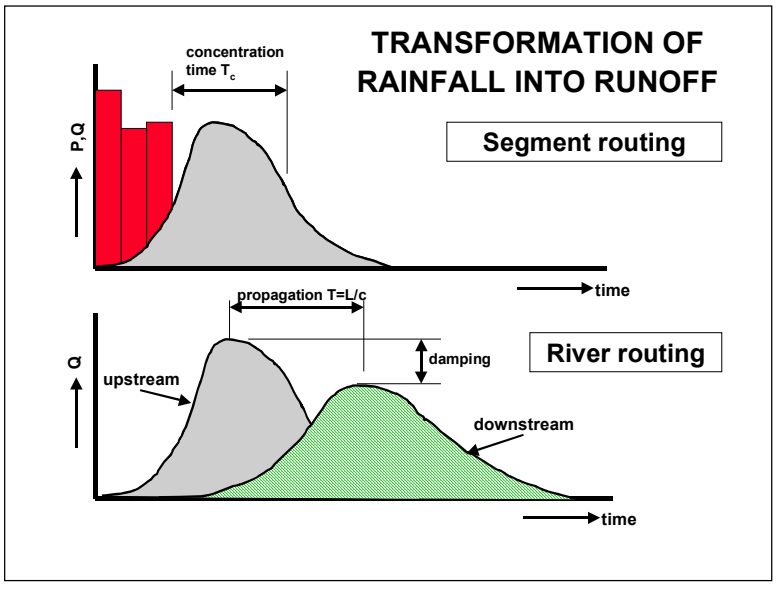


Figure 3.2:
Features of segment and river routing

Basic data input requirements are time series of rainfall, evapo-transpiration and the observed discharge as well as the catchment or segment area. The data time interval depends on the objective of the simulation and is generally taken as 1 hour or 1 day. The model simulates the rainfall runoff process with a time step, which is less than the data time interval.

All parameters and storage capacities have also to be initially estimated on the basis of physical properties of the segment and the river system. Some then remain fixed whilst other are recommended for optimisation.

3.2.2 THE SEGMENT MODULE

The segment module simulates the rainfall-runoff process in part of the catchment, where the attention is on the land-phase of the rainfall-runoff process. It is assumed that the open water system in the segments contributes little to the shaping of the hydrograph. The conceptualisation of the processes as described in the segment module is presented in Figure 3.3.

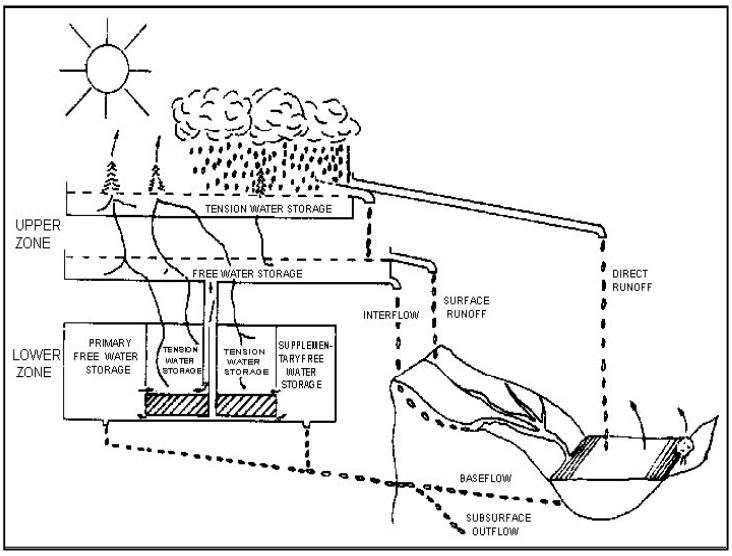


Figure 3.3:
Conceptualisation of the rainfall runoff process in a segment

The segment module is divided into the following components, (see also Figure 3.4):

Impervious area	with transfer to	direct runoff
Previous area		
Upper zone		
Tension storage	with transfer to	evaporation, free water storage
Free water storage	with transfer to	evaporation, percolation, surface runoff and interflow
Lower zone		
Tension storage	with transfer to	evaporation, free water storage
Free water storage	with transfer to	base flow

From the *impervious* areas, precipitation immediately discharges to the channel. However, impervious areas, which drain to a pervious part before reaching the channel, are not considered impervious. Both zones have a tension and a free water storage element. Tension water is considered as the water closely bound to soil particles. Generally first the tension water requirements are fulfilled before water enters the free water storage.

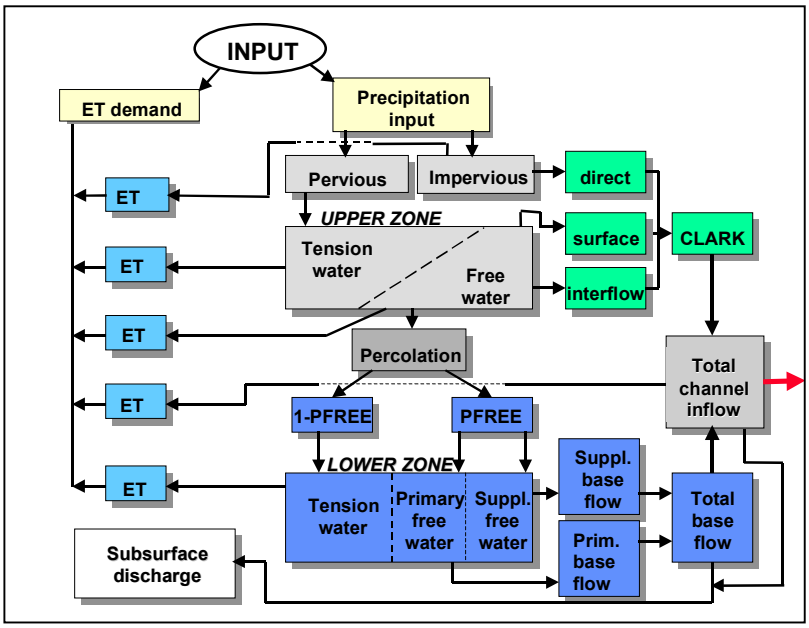


Figure 3.4: Schematisation of rainfall runoff process in a segment

In the following sub-sections the various components will be described in detail

Upper zone storage

The upper zone tension storage represents that precipitation volume required under dry conditions:

1. to meet all interception requirements, and
2. to provide sufficient moisture to the upper soil so that percolation can begin.

If the maximum storage capacity of the upper-zone tension storage is exceeded, water becomes available for the upper zone free water storage, a temporary storage from which water percolates to the lower zone system and from which water discharges to the channel via the interflow component. The preferred flow direction from the upper zone is the vertical direction, i.e. percolation to the lower zone system.

Interflow occurs only when the precipitation rate exceeds the percolation rate. The upper zone is treated as a linear storage element which is emptied exponentially: discharge = storage * storage depletion coefficient. The upper zone free water storage depletion coefficient is denoted by *UZK* and the upper zone free water content by *UZFWC* then the interflow takes place at a rate:

$$Q_{interflow} = UZFWC * UZK \qquad (3.1)$$

When the precipitation intensity exceeds the percolation intensity and the maximum interflow drainage capacity, then the upper zone free water capacity *(UZFWM)* is completely filled and the excess precipitation causes surface runoff.

Lower zone storage

The lower zone consists of the tension water storage, i.e. the depth of water held by the lower zone soil after wetting and drainage (storage up to field capacity) and two free water storages: the *primary and supplemental* storage elements representing the storages leading to a slow and a fast groundwater flow component, respectively. The introduction of two free lower zone storages is made for greater flexibility in reproducing observed recession curves caused by groundwater flow.

Percolation from upper to lower zones

The percolation rate from the upper zone to the lower zone depends on the one hand on the lower zone demand, i.e. requirements determined by the lower zone water content relative to its capacity and on the other hand on the upper zone free water content relative to its capacity.

The lower zone percolation demand is denoted by $PERC_{act.dem}$. The upper zone free water content relative to its capacity is *UZFWC/UZFWM*. Hence, the actual percolation intensity then reads:

$$PERC = PERC_{act.dem} * UZFWC/UZFWM \quad (3.2)$$

The lower zone percolation demand has a lower and an upper limit:

- the minimum lower zone percolation demand, and
- the maximum lower zone percolation demand.

The minimum lower zone percolation demand occurs when all three lower zone storages are completely filled. Then by continuity the percolation rate equals the groundwater flow rate from full primary and supplemental reservoirs. Denoting the minimum demand by *PBASE* then it follows:

$$PERC_{min.dem} = PBASE = LZFPM * LZPK + LZFSM * LZSK \quad (3.3)$$

where: *LZFPM* = lower zone primary free water storage capacity

LZFSM = lower zone supplemental free water storage capacity

LZPK = drainage factor of primary storage

LZSK = drainage factor of supplemental storage

The maximum lower zone percolation demand takes place if the lower zone is completely dried out i.e. if its content = 0. Then the maximum percolation rate is expressed as a function of *PBASE*:

$$PERC_{max.dem} = PBASE (1 + ZPERC) \quad (3.4)$$

with: *ZPERC* >> 1 usually.

The actual lower zone percolation demand depends on the lower zone content relative to its capacity. Computationally it means that *ZPERC* has to be multiplied by a function *G* of the relative lower zone water content such that this function:

- equals 1 in case of a completely dry lower zone
- equals 0 in case of a completely saturated lower zone
- represents an approximate exponential decay of the percolation rate in case of a continuous recharge.

In the Sacramento model this function has the following form:

$$G = \left(\frac{\sum(\text{lower zone capacities} - \text{lower zone content})}{\sum(\text{lower zone capacities})} \right)^{REXP} \quad (3.5)$$

and the *actual percolation demand* is given by (see Figure 3.5):

$$PERC_{act.dem} = PBASE\ (1 + ZPERC * G) \quad (3.6)$$

Distribution of percolated water from upper zone

The percolated water drains to three reservoirs, one tension and two free water reservoirs. Based on the preceding comments one would expect that the lower zone tension storage is filled first before percolation to the lower zone free water storages takes place. However, variations in soil conditions and in precipitation amounts over the catchment cause deviations from the average conditions. This implies that percolation to the free water reservoirs and hence groundwater flow takes place before the tension water reservoir is completely filled. The model allows for this to let a fraction of the infiltrated water percolate to the two free water storages. When the tension water reservoir is full, all percolated water drains to the primary and supplemental free water storage in a ratio corresponding to their relative deficiencies.

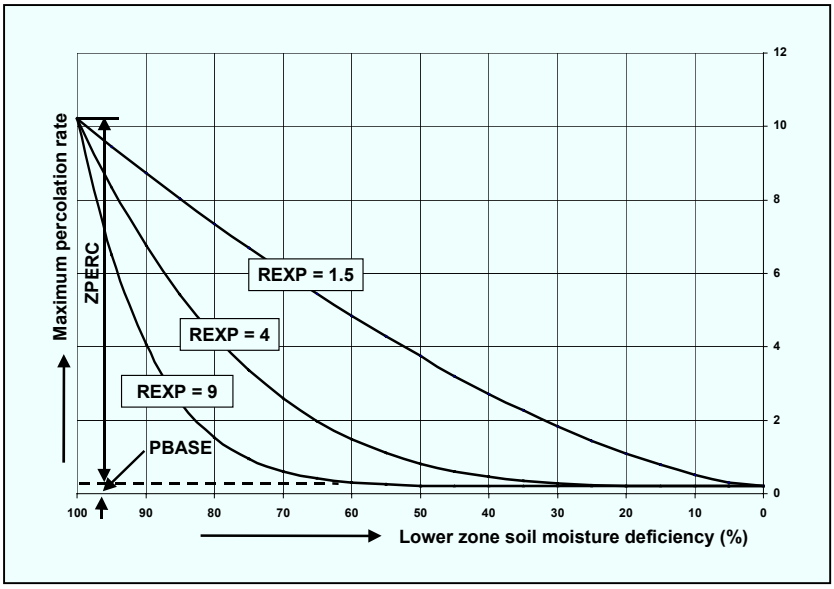


Figure 3.5:
Actual percolation demand representation

Groundwater flow

Baseflow to the river from groundwater depends on the contents of the two lower zone free water storages and two drainage constants expressed in fractions of the content per day. If the actual contents of the primary and supplemental free water zones are denoted by *LZFPC* and *LZFSC* respectively then the total base flow *QBASE* becomes, in accordance with the linear reservoir theory:

$$QBASE = LZFPC * LZPK + LZFSC * LZSK \quad (3.7)$$

The drainage factors *LZPK* and *LZSK* can be determined from the recession part of the hydrograph by plotting that part of the hydrograph on semi-logarithmic paper (Fig. 3.6). In the lowest part of the recession curve only the slow base flow component is acting while in the higher stages both base flow components contribute.

The drainage factor LZPK follows from:

$$K = \left(\frac{QP_{t_0+\Delta t}}{QP_{t_0}} \right)^{1/\Delta t} \quad (3.8)$$

and

$$LZPK = 1 - K \quad (3.9)$$

where: K = recession coefficient of primary base flow for the time unit used

Δt = number of time units, generally days

$QP_{t0+\Delta t}$ = a discharge when recession is occurring at the primary base flow rate

QP_{t0} = the discharge t time units later

If QP_{max} represents the maximum value of the primary base flow, then the maximum water content of the lower zone becomes:

$$LZFPM = QP_{max} / LZPK \quad (3.10)$$

and similarly the supplemental lower zone free water capacity is determined; at least this procedure provides first estimates of the lower zone free water capacities (Figure 3.6).

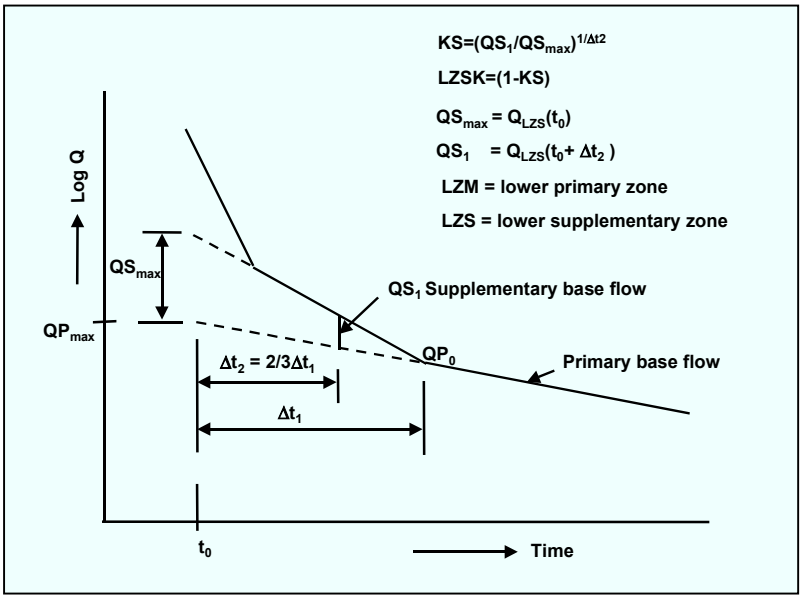


Figure 3.6: Principle of computation of lower zone recession coefficient

The total base flow contributes completely or in part to the channel flow. A complete contribution occurs if subsurface discharge (i.e. discharge from the segment, which is not measured at the outlet) is absent. Otherwise a fraction of the total base flow represents the subsurface flow.

Actual evapotranspiration

Evaporation at a potential rate occurs from that fraction of the basin covered by streams, lakes and riparian vegetation. Evapotranspiration from the remaining part of the catchment is determined by the relative water contents of the tension water zones. If *ED* is the potential evapotranspiration, then the actual evapotranspiration from the upper zone reads:

$$E_1 = ED * (UZTWC / UZTWM) \quad (3.11)$$

i.e. the actual rate is a linear function of the relative upper zone water content. Where $E_1 < ED$ water is subtracted from the lower zone as a function of the lower zone tension water content relative to the tension water capacity:

$$E_2 = (ED - E_1) * LZTWC / (UZTWM + LZTWM) \quad (3.12)$$

If the evapotranspiration should occur at such a rate that the ratio of content to capacity of the free water reservoirs exceeds the relative tension reservoir content then water is transferred from free water to tension water such that the relative loadings balance. This correction is made for the upper and lower zone separately. However, a fraction *RSERV* of the lower zone free water storage is unavailable for transpiration purposes.

Impervious and temporary impervious areas

Besides runoff from the pervious area, the channel may be filled by rainwater from the impervious area. With respect to the size of the impervious area it is noted that in the Sacramento model a distinction is made between permanent and temporary impervious areas where temporary impervious areas are created when all tension water requirements are met, i.e. an increasing fraction of the catchment assumes impervious characteristics.

Routing of surface runoff

Before the runoff from the impervious areas, the overland- and interflow reach the channel, they may be transformed according to a unit hydrograph leading to an adapted time distribution of these flow rates.

Use can be made here of the Clark method, which is a combined time-area and storage routing method. The model requires the construction of a time-area diagram. For this isochrones are constructed representing points of equal travel time to the segment outlet, see Figure 3.7. The areas between successive isochrones is determined and subsequently properly scaled by the time of concentration T_c. The latter is defined as the time required to have the effect of rainfall fallen in the most remote part felt at the segment outlet. The time-area diagram can be thought of as the outflow from the segment if only translation and no deformation takes place of an instantaneous unit supply of rain over the entire segment. Subsequently, the time area diagram flow is routed through a linear reservoir, which characterises the effect of storage in the open drainage system of the segment. This reservoir is represented by the second parameter: the recession coefficient k. It is noted that the output from the reservoir represents the instantaneous unit hydrograph *(IUH).* This has to be transformed into say a 1-hour unit hydrograph, dependent on the chosen routing interval.

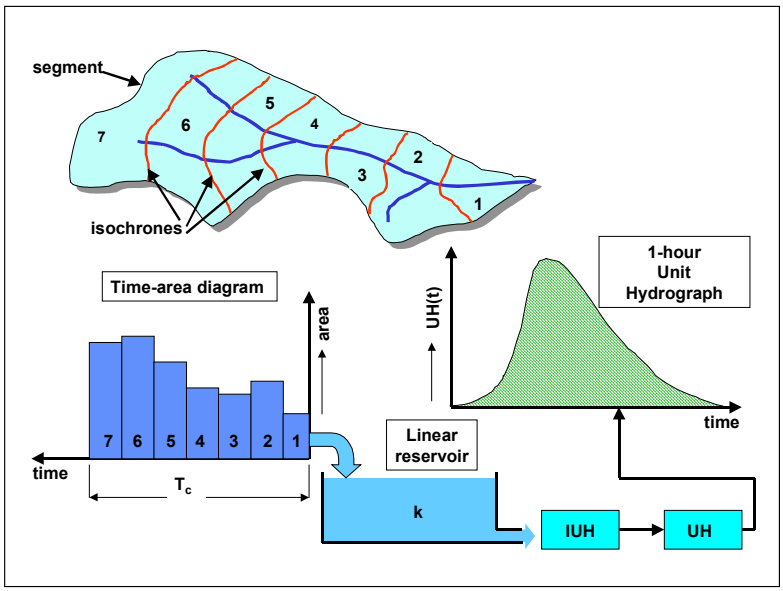


Figure 3.7:
Principles of the Clark method for simulating surface runoff and interflow

The two parameters T_c and k can be obtained from observed rainfall and discharge hydrographs. The time of concentration is equal to the time interval between cessation of rainfall and the time the hydrograph has receded to its inflection point (see Figure 3.2). Alternatively it is determined from physical features of the segment as length and slope. A large number of empirical formulas are available which relate the time of concentration to topographical features of the basin. It is noted, though, that these formulas have generally only local validity. The best is to estimate the celerity from the flow velocities in the drainage system taking account of the following characteristics of celerity:

- If the rivulet remains inbank the celerity is about 1.5 to 1.7 times the cross-sectional flow velocity
- If the flow becomes overbank the above celerity has to be multiplied with the ratio of the drain width and the total width of the flow at the water surface (i.e. inclusive of the floodplain)

To the time required to travel through the drainage system one has to add the overland flow time.

The recession coefficient k is determined from the slope of recession part of the surface runoff hydrograph, similar to the procedure for groundwater.

3.2.3 THE CHANNEL MODULE

Contributions to the channel flow component are made by:

- runoff from impervious areas,
- overland flow from the pervious areas,
- interflow, and
- base flow (completely or in part).

The propagation and attenuation of the segment outflows in channel branches can be described by:

- Unit hydrograph technique
- Muskingum routing
- Structure and reservoir routing

Unit hydrograph technique

To propagate and attenuate riverflow through a channel reach for each channel branch a unit hydrograph can be defined. It describes how the inflow to the branch will be redistributed in time while travelling through the branch. Let the inflow to the branch be denoted by I_i and let the ordinates of the unit hydrograph be U_1, U_2, etc., with $U_i = 1$, then the outflow from the branch O_i becomes:

$$O_i = I_i \times U_1 \qquad (3.13)$$

$$O_{i+1} = I_i \times U_2 + I_{i+1} \times U_1$$

$$O_{i+2} = I_i \times U_3 + I_{i+1} \times U_2 + I_{i+2} \times U_1, \text{ etc.}$$

If e.g. the travel time through the reach is exactly 1 time interval and there is no attenuation then:

$U_1 = 0$, $U_2 = 1$.

This option provides a simple means to combine segment outflows entering the river at different locations, when the computational interval is too large for proper channel routing using the

Muskingum approach. E.g. the travel time through a branch is 15 hours, but the computational interval is 1 day as rainfall data were only available as daily totals. Then within that day (24-15)/24x100% arrives and the rest the next day, so $U_1 = 0.375$, $U_2 = 0.625$. Routing with daily intervals is very acceptable when one is interested in 10-daily or monthly flow data and not in the finest details of the hydrograph.

Muskingum routing

The Muskingum procedure is based on the following routing equation:

$$O(t+\Delta t) = c_1 I(t) + c_2 I(t+\Delta t) + c_3 O(t) \qquad (3.14)$$

where: I = Inflow to channel reach

O = Outflow from a channel reach

Since equation (3.14) is derived from S = K(xI+(1-x)O), where S = storage in the reach, it is observed that for x = 0 a simple linear reservoir concept follows: S = KO. With x = 0.5 there is no attenuation and the inflow is passed on through the end of the channel reach without any attenuation in time K.

The routing interval should be less than or equal to *K* as otherwise peaks will be missed at the downstream end of the reach. Often a value of Δt = ½ to ¼ of K is advised. However, taking Δt too small then c_2 becomes negative, which will lead to negative outflows when the inflow hydrograph suddenly rises. To avoid negative outflows the routing interval is conditioned by: $2Kx \leq \Delta t \leq K$.

Unfortunately, this leaves little freedom in the selection of Δt when x is close to 0.5. Note that when x=0.5 and Δt = K, it follows from (3.14) that c_1=1, c_2=0 and c_3=0; hence: $O(t+\Delta t) = O(t+K) = c_1 I(t)$, i.e the inflow is passed on to the outlet time K later, unaltered.

Flood wave celerity and attenuation changes drastically when the river reaches the flood plain. To cope with these changes a layered Muskingum approach can be used. The principle of the layered Muskingum procedure is displayed in Figure 3.8, in which the meaning of the various parameters is explained. By applying different sets of parameters for the inbank flow and overbank flow the reduction of the flood wave celerity in case of wide flood plains can be taken into account.

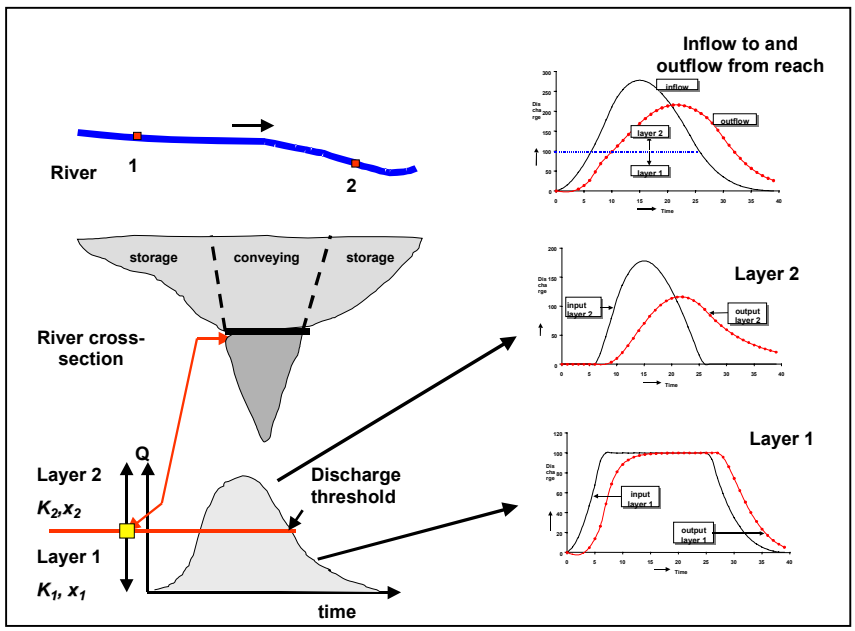


Figure 3.8: Principle of layered Muskingum approach

Structures and reservoirs

Some features may be present in the river which affect the shape of the hydrograph, like culverts and reservoirs:

- culvert

  A culvert limits the capacity of the river. Basically, it chops the peak of the hydrograph beyond the capacity of the culvert. In the model the shape of the hydrograph is altered such that upon passage of the floodwave the maximum downstream hydrograph value is kept at the culvert capacity until the entire volume in the upstream hydrograph above the capacity of the culvert has passed. It is noted that some old bridges may act also more or less like a culvert.

- reservoir

  The model includes a number of reservoir routing options where the flow is controlled by overflow (ogee and glory type) structures and underflow structures. For routing a third order Runge-Kutta scheme is used.

3.2.4 ESTIMATION OF SEGMENT PARAMETERS

Overview of parameters

The following groups of parameters can be distinguished for a particular segment:

Segment

Segment area (km^2)

Direct runoff

PCTIM	Permanently impervious fraction of segment contiguous with stream channels
ADIMP	Additional impervious fraction when all tension water requirements are met
SARVA	Fraction of segment covered by streams, lakes and riparian vegetation

Upper soil moisture zone

UZTWM	Capacity of upper tension water zone (mm)
UZFWM	Capacity of upper free water zone (mm)
UZK	Upper zone lateral drainage rate (fraction of contents per day)

Percolation

ZPERC	Proportional increase in percolation from saturated to dry conditions in lower zone
REXP	Exponent in percolation equation, determining the rate at which percolation demand changes from dry to wet conditions

Lower zone

LZTWM	Capacity of lower zone tension water storage (mm)
LZFPM	Capacity of lower zone primary free water storage (mm)
LZFSM	Capacity of lower zone supplemental free water storage (mm)
LZPK	Drainage rate of lower zone primary free water storage (fraction of contents per day)
LZSK	Drainage rate of lower zone supplemental free water storage (fraction of contents per day)
PFREE	Fraction of percolated water, which drains directly to lower zone free water storages

RSERV	Fraction of lower zone free water storages which is unavailable for transpiration purposes
SIDE	Ratio of unobserved to observed baseflow
SSOUT	Fixed rate of discharge lost from the total channel flow (mm/t)

Surface runoff

Unit hydrograph ordinates

Internal routing interval

PM	Time interval increment parameter
PT1	Lower rainfall threshold

Basically two procedures are available to get first estimates for the majority of the segment parameters:

- from observed rainfall and runoff records: this method is usually applied and works well provided that the model concepts are applicable and that reliable records are available for some time covering the majority of the range of flows
- from soil characteristics: this method is particularly suitable if no runoff records are available, i.e. for ungauged catchments.

With respect to gauged catchments the following grouping of parameters according to the method of estimation can be made:

1. Parameters computed and estimated from basin map solely:
   segment area and *SARVA*
2. Parameters estimated from observed rainfall and runoff records:
   readily: *LZFPM, LZPK, LZFSM, LZSK, PCTIM*
3. approximately: *UZTWM, UZFWM, UZK, LZTWM, SSOUT* and *PFREE*
   Parameters estimated from topographic maps and rainfall and runoff records:
   unit hydrograph ordinates obtained from Clark method
   Parameters to be obtained through trial runs:
4. *ZPERC, REXP, SIDE, ADIMP, RSERV*
5. Internal routing parameters, as per requirement:
   PM, PT1, PT2

In the next sub-sections guidelines are given for the determination and estimation of the segment parameters for gauged catchments.

Segment parameter estimation for gauged catchments

The estimation of the segment parameters is presented according to their order of appearance in the previous sub-section. The sequence in which the estimation is done in practice is different from this order, for which reference is made to the end of the sub-section.

Segment:

Segment area

To allow a good comparison between the observed and simulated runoff from the basin, the segment area (km^2) should refer to the total segment area draining upstream of the gauging station. Any difference between total segment area up to the main stream and the area upstream of the gauging station can be accommodated for in the channel routing part.

Direct runoff:

PCTIM

Permanently impervious fraction of the basin contiguous with stream channels. It can be determined from small storms after a significant period of dry weather. Then the volume of direct runoff (=observed runoff - baseflow) divided by the volume of rain gives the percentage impervious fraction of the basin. *PCTIM* should not be close to 1!

An example is given below.

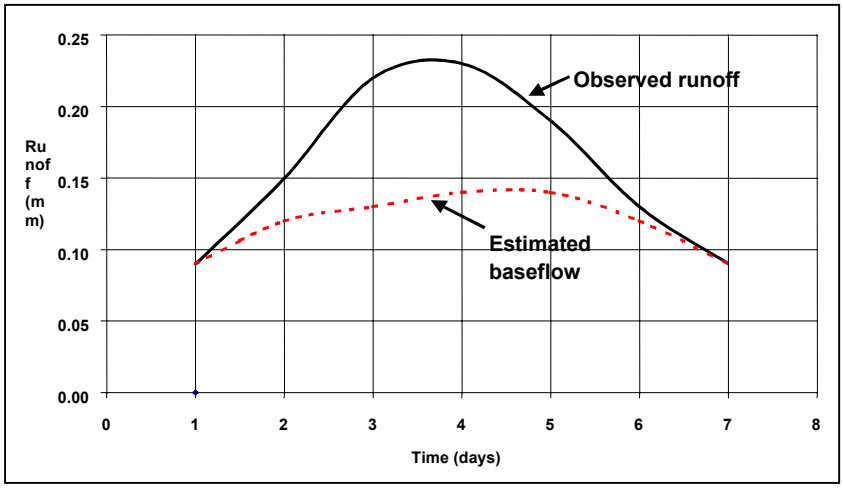


Figure 3.9: Calculation of PCTIM

ADIMP

Fraction of the basin, which becomes impervious as all tension water requirements are met. It can be estimated from small storms after a very wet period. As before, the volume of direct runoff divided by the volume of rain gives the total percentage of impervious area. The estimate for *ADIMP* follows from:

ADIMP = Total Percentage Impervious – PCTIM (3.15)

SARVA

Fraction of the basin covered by streams, lakes, and riparian vegetation, under normal circumstances. The *SARVA* area is considered to be the same as or less than *PCTIM* (see below). Detailed maps may be referred to in order to estimate the extent of paved areas, which drain directly to the streams so that differences between *PCTIM* and *SARVA* can be approximated. Generally, *SARVA* appears to range between 40% and 100% of the *PCTIM* value.

Upper soil moisture zone:

UZTWM - ***the upper tension storage capacity***

The depth of water, which must be filled over non-impervious areas before any water becomes available for free water storage. Since upper zone tension water must be filled before any streamflow in excess of the impervious response can occur, its capacity can be approximated from hydrograph analysis. Following a dry period when evapotranspiration has depleted the upper soil moisture, the capacity of upper zone tension water can be estimated. That volume of rainfall, which is retained before runoff from the pervious fraction is visible, is identified as *UZTWM.* To that rainfall volume the losses to evaporation during the considered period should be added. All periods of rain following a dry

period should be checked for estimation of this parameter. Generally the capacity of the upper zone tension will vary between 25 and 175 mm, depending on the soil type.

Following the logic of the Curve Number method, where the initial abstraction before rainfall becomes effective is estimated as 20% of the potential maximum retention, the *UZTWM* becomes:

UZTWM = 50.8(100/CN - 1) (mm)

CN-values range from 30 to about 90 for rural areas and are a function of:

- soil type (soil texture and infiltration rate)
- hydrological soil groups A-D are distinguished
- land use,
- type of land cover,
- treatment, and
- hydrologic or drainage condition

It is also a function of antecedent moisture condition, for which the condition “dry” should be taken in view of the meaning of *UZTWM*. Based on this assumption *UZTWM* would vary between 120 and 6 mm, values which are in the range of those given above, particularly if one realises that the 20% of the potential maximum as initial abstraction is an average value. Reference is made to standard textbooks on hydrology for CN-values

UZFWM - *the upper free water storage capacity*

Upper zone free water represents that depth of water, which must be filled over the non-impervious portion of the basin in excess of *UZTWM* in order to maintain a wetting front at maximum potential. This volume provides the head function in the percolation equation and also establishes that volume of water, which is subject to interflow drainage. Generally its magnitude ranges from 10-100 mm. It is not generally feasible to derive the magnitude of the upper zone free water from direct observations, and successive computer runs are required in order to establish a valid depth.

However, if a rough estimate of UZK is available (see below), then a rough value of UZFWM can be obtained from the hydrograph at the time of the highest interflow, by reducing the flow value with primary and supplemental baseflow.

UZK - *the upper zone lateral drainage rate*

The upper zone lateral drainage rate is expressed as the ratio of the daily withdrawal to the available contents. Its range is roughly 0.18 to 1.0, with 0.40 generally serving as an effective initial estimate. Though basically, this factor is not capable of direct observation and must be determined by successive computer runs, Peck (1976) suggests the following approximate procedure. *UZK* is roughly related to the amount of time that interflow occurs following a period with major direct and surface runoff. A long period of interflow results in a small value for *UZK*. Assuming that interflow is observed during N consecutive days and that interflow becomes insignificant when it is reduced to less than 10% of its maximum value it follows:

$$(1 - UZK)^N = 0.10 \quad \text{or} \quad UZK = 1 - 0.1^{1/N} \qquad (3.16)$$

Values for UZK as a function of N can be read from Figure 3.10.

Percolation:

ZPERC

The proportional increase in percolation from saturated to dry condition is expressed by the term ZPERC. The value of ZPERC is best determined through computer trials. The initial estimate can be derived by sequentially running one or two months containing significant hydrograph response following a dry period. The value of ZPERC should be initially established so that a reasonable determination of the initial run-off conditions is possible.

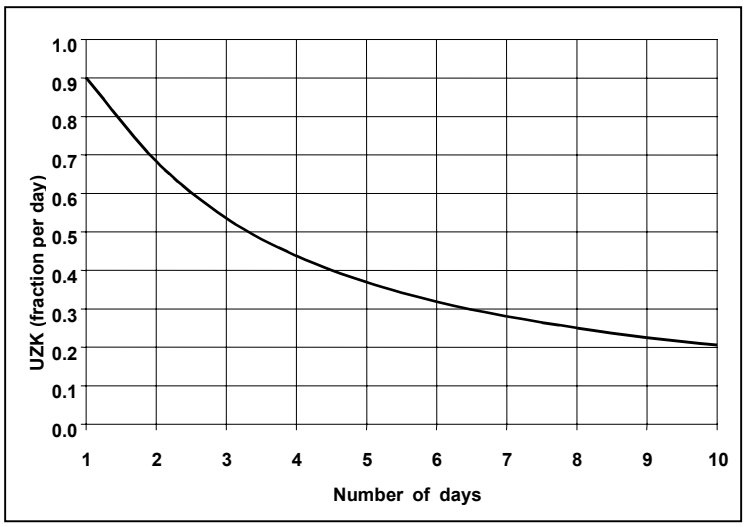


Figure 3.10:
UZK as function of number of days with significant interflow

Amstrong (1978) provides a procedure to derive *ZPERC* from the lower zone tension and free water reservoir capacities and drainage rates, using equations (3.3) and (3.4). The maximum percolation takes place when the upper zones are full and the lower zones are empty. Assuming that the maximum daily percolation will be the maximum contents of the lower zones, from equation (3.4) it follows for *ZPERC*:

$$\mathrm{ZPERC} = \frac{\mathrm{LZTWM} + \mathrm{LZFPM} + \mathrm{LZFSM} - \mathrm{PBASE}}{\mathrm{PBASE}} \qquad (3.17)$$

If data would be available on maximum percolation rates *ZPERC* can be estimated using equation (3.4). Values for *ZPERC* ranging from 5 to 80 have been used.

REXP

The exponent in the percolation equation which determines the rate at which percolation demand changes from the dry condition, *(ZPERC + 1)*PBASE*, to the wet condition, *PBASE*. Figure 3.5 illustrates how different values of the exponent affect the infiltration rate. It is recommended that an initial estimate of this exponent is made from the same record which is used in determining an initial estimate of *ZPERC*. The interaction between *PBASE, ZPERC* and *REXP* may require a shift of all three terms whenever it becomes clear that a single term should be changed. Visualising the percolation curve generated by these three terms helps to ascertain the necessary changes. The observed range of *REXP* is usually between 1.0 and 3.0. Generally a value of about 1.8 is an effective starting condition. Values for *REXP* for different soils are given by Amstrong (1978) and are presented in Table 3.1.

Soil classification	REXP
Sand	1.0
Sandy loam	1.5
Loam	2.0
Silty loam	3.0
Clay, silt	4.0

Table 3.1: Perlocation exponent REXP for different soil types

Lower zone:

LZTWM - *lower zone tension water capacity (mm)*

This volume is one of the most difficult values to determine effectively. Inasmuch as carryover moisture in this storage may exist for a period of many years, its total capacity may not be readily discernible from available records. If a drought condition during the period of record in the basin or in the area being studied has been sufficient to seriously affect the transpiration process of deep rooted plants, then the period of record is usually sufficient to determine the maximum storage value of lower zone tension water. Often, however, field data is not adequate for this purpose. As a result, unless great care is taken, the depth of lower zone tension water storage may inadvertently be set near the maximum deficit experienced during the period of record rather than the actual capacity of the zone. It has been noted that the plant growth of an area is a relatively effective indicator of the capacity of the lower zone tension water zone. In heavily forested regions of deep-rooted conifers, this zone may be approximately 600 mm in magnitude. In areas of deep-rooted perennial grasses this depth is more likely to be close to 150 mm. Where vegetation is composed primarily of relatively shallow-rooted trees and grasses, this depth may be as little as 75 mm. It should be realised that this zone represents that volume of water, which will be tapped by existing plants during dry periods.

An approximate procedure to estimate *LZTWM* from a water balance analysis is presented by Peck (1976). For this a period is selected with direct and/or surface runoff following an extended dry spell. The selected period is bounded by the times t_1 and t_2. At both times t_1 and t_2 only baseflow occurs. The start t1 is selected immediately prior to the occurrence of direct/surface runoff and t_2 immediately following a period of interflow. The times t_1 and t_2 can best be selected from a semi-log plot of the runoff, see Figure 3.11.

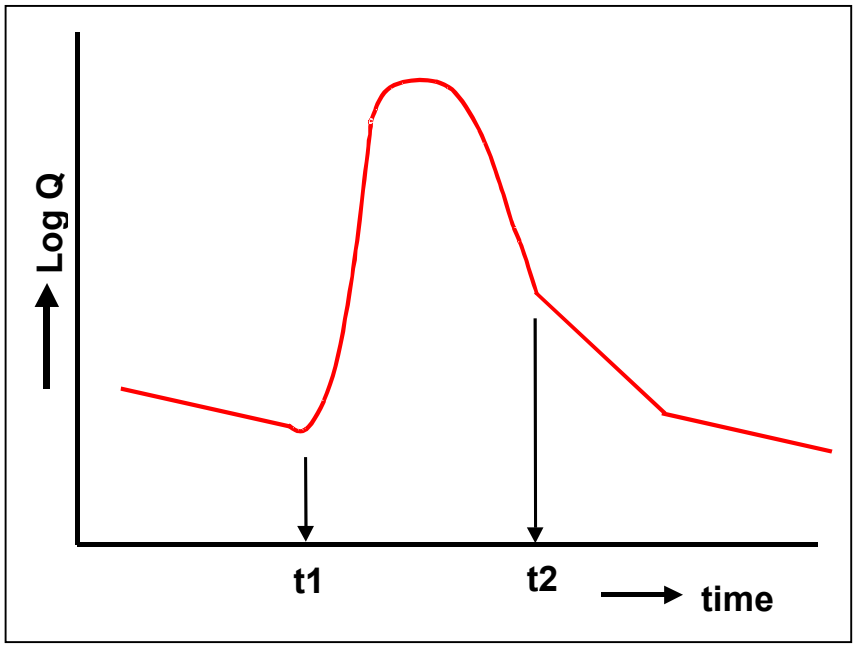


Figure 3.11: Selection of period for LZTWM estimation

Assuming that UZTW is full and UZFWC is empty at times t_1 and t_2 the water balance for the period 1-t_2 then reads:

$$P - R - E - \Delta LZFPC - \Delta LZFSC = \Delta LZTWC \quad (3.18)$$

where: *P* = precipitation from t_1 to t_2 (mm)

R = total runoff from t_1 to t_2 (mm)

E = segment evaporation (mm); this amount would small for most wet period and may be neglected

Δ*LZFPC* = change in storage in LZ primary free water reservoir from t_1 to t_2 (mm)

Δ*LZFSC* = change in storage in LZ supplemental free water reservoir from t_1 to t_2 (mm)

Δ*LZTWC* = change in the lower zone tension water (mm)

Δ*LZTWC* is a lower limit of *LZTWM* since:

- The lower zone tension water reservoir may not have been fully empty at t_1
- The lower zone tension water reservoir may not have been completely filled at t_2

Hence some 10 to 20% (or more) may be added to the value obtained through (3.18). If such ideal cases as assumed above cannot be found, water balances for periods of 3 to 4 months may be considered.

In equation (3.18) ΔLZFPC and ΔLZFSC are computed as follows:

$$\Delta LZFPC = LZFPC(t_2) - LZFPC(t_1), \text{ where } LZFPC(t) = QP(t)/LZPK \quad (3.19)$$

$$\Delta LZFSC = LZFSC(t_2) - LZFSC(t_1), \text{ where } LZFSC(t) = QS(t)/LZSK \quad (3.20)$$

The primary baseflows *QP* at times t_1 and t_2 are estimated by extrapolation from other periods. Let the discharges at t_1 and t_2 be denoted by Q_1 and Q_2, then the supplemental baseflows follow from:

$$QS(t_1) = Q_1 - QP(t_1) \quad \text{and} \quad QS(t_2) = Q_2 - QP(t_2) \quad (3.21)$$

LZFPM - *lower zone primary free water storage*

The maximum capacity of the primary lower zone free water, which is subject to drainage at the rate expressed by *LZPK*. The value of the lower zone primary free water maximum can be approximated from hydrograph analysis. For this the primary base flow, obtained from a semi-log plot of the lower end of the recession curve, is extended backward to the occurrence of a peak flow. Assuming that the primary free water reservoir is completely filled then, so that it outflow is at maximum (*QPmax*), its value is determined from equation (3.10). The effectiveness of this computation in determining the maximum capacity is dependent upon the degree to which the observed hydrograph provides a representation of the maximum primary baseflow. If only a portion of the groundwater discharge is observable in the stream channel, the estimated capacity based upon surface flows must be increased to include the non-channel components by applying the term *SIDE* (See below).

LZFSM - *lower zone supplemental free water storage*

The maximum capacity of the lower zone supplemental free water reservoir, which is subject to drainage at the rate expressed by *LZSK*. A lower limit of the lower zone free water supplemental maximum can be approximated from hydrograph analysis. Fig. 3.6 illustrates the computation of the lower zone free water supplemental maximum. Note that first the primary base flow has to be identified and corrected for, see also equation (3.21). The effectiveness of this computation in

determining the maximum capacity is dependent upon the degree to which the observed hydrograph provides a representation of the maximum baseflow capability of the basin. If only a portion of the groundwater discharge is observable in the stream channel, the estimated capacity based upon surface flows must be increased to include the non-channel components by applying the term *SIDE* (See below).

LZPK - *lateral drainage of the lower zone primary free water reservoir*

Lateral drainage rate of the lower zone primary free water reservoir expressed as a fraction of the contents per day. The coefficient is determined from the primary base flow recession curve. Selecting flow values from this curve at some time interval Δt apart provides with the help of equations (3.8) and (3.9) the required estimate, see also Figure 3.6.

LZSK - *lateral drainage of the lower zone supplemental free water reservoir*

Lateral drainage rate of the lower zone supplemental free water reservoir, expressed as a fraction of the contents per day. Its computation is outlined in Figure 3.6. The procedure is similar to that of *LZPK*, with the exception that the flow values have to be corrected for the primary base flow.

PFREE

The fraction of the percolated water, which is transmitted directly to the lower zone free water aquifers. Its magnitude cannot generally be determined from hydrograph analysis. An initial value of 0.20 is suggested. Values will range between 0 and 0.50. The analysis of early season baseflow allows an effective determination of *PFREE*. The relative importance of *PFREE* can be determined from storms following long dry spells that produce runoff (*UZTW* completely filled). If the hydrograph returns to approximately the same base flow as before then little filling of the lower zone free water reservoirs did take place and hence the *PFRE*E-value can be rated small, 0 to 0.2. If, on the contrary, the base flow has increased significantly a *PFREE*-value as high as 0.5 may be applicable.

RSERV

Fraction of the lower zone free water, which is unavailable for transpiration purposes. Generally this value is between zero and 0.40 with 0.30 being the most common value. This factor has very low sensitivity.

SIDE

Represents that portion of base flow, which is not observed in the stream channel. When the soil is saturated, if percolation takes place at a rate, which is greater than the observable baseflow, the need for additional soil moisture drainage becomes manifest. *SIDE* is the ratio of the unobserved to the observed portion of base flow. When the saturated soils do not drain to the surface channel, *SIDE* allows the correct definition of *PBASE*, in order that the saturated percolation rate may be achieved. In an area where all drainage from baseflow aquifers reaches surface channels, *SIDE* will be zero. Zero or near zero values occur in a large proportion of basins. However, in areas subject to extreme subsurface drainage losses, *SIDE* may be as high as 5.0. It is conceivable that in some areas the value of *SIDE* may be even higher.

SSOUT

The sub-surface outflow along the stream channel, which must be provided by the stream before water is available for surface discharge. This volume expressed in mm/time interval is generally near zero. It is recommended that the value of zero be utilised, and *SSOUT* is applied only if the log Q vs.

time plot requires a constant addition in order to achieve a valid recession characteristic. If constant volumes of flow are added to observed stream flow, the slope of the discharge plot will be altered. That value, which is required to linearize the primary recession, is the appropriate value of *SSOUT*. It should be realised that where *SSOUT* is required, an effective determination of lower zone free water storages and discharge rates will require inclusion of the *SSOUT* value (mm/Δt)

Surface runoff

Unit hydrograph ordinates for the routing of flow from the impervious and pervious surfaces as well as interflow towards the segment outlet can be obtained through standard unit hydrograph procedures. It requires the selection of rainfall events (corrected for losses) with their resulting flood hydrographs (corrected for base flow). Note that for each event the net rainfall amount should match with the surface runoff and interflow amount. Various procedures are available to arrive at a unit hydrograph. If the rainfall intensity during the storm varies, multiple linear regression and discrete convolution techniques may be applied. The regression technique is readily available in spreadsheet software. The resulting unit hydrographs generally will show some irregularities and hence requires some smoothing afterwards. Unit hydrographs from various storms may appropriately be averaged to arrive at a representative unit hydrograph for the segment.

Another option is to use the Clark method. The principle of the Clark method was dealt with in Sub-section 3.2.2. First requirement is the derivation of a time-area diagram. If a Digital Elevation Model (DEM) is available from a catchment with appropriate software automatic calculation of the time-area diagram is possible. In the absence from a DEM the time-area diagram is derived from a basin map. By estimating travel times to the basin outlet (from river and terrain slopes, assumed roughness and flow depth) isochrones can be determined. The areas between successive isochrones is determined leading to a first estimate of the time-area diagram. The total time base of the time-area diagram should be the concentration time T_c, but due to inaccurate assessment of celerities in the basin it may differ from that. Therefore, the time base of the time-area diagram is scaled by a more appropriate estimate of T_c. An estimate for T_c may be obtained as the time lapse between the cessation of rainfall and the occurrence of recession on the falling limb of the hydrograph of surface runoff. The time base of the time-area diagram is scaled by this time lapse. Alternatively, the concentration time is estimated from an empirical formula applicable to the region. E.g. for a number of small catchments in the Indus basin the following equation applies:

$$T_c = \frac{1}{119}\frac{L}{\sqrt{S}} \qquad (3.22)$$

where: T_c = concentration time (hrs)
L = length of river (km)
S = slope of main river

The units of the time-area diagram (km^2) are converted into m^3/s by multiplication with 0.278/Δt, with Δ t in hours. Subsequently, the time-area diagram is routed through a linear reservoir, with reservoir coefficient k, estimated from the slope of the recession curve of the surface water hydrograph. The routing is carried out by the following equation:

$$O_{i+1} = c_1 I_{av} + c_2 O_i$$
$$\text{with}: I_{av} = \frac{1}{2}(I_i + I_{i+1});\ c_1 = \frac{\Delta t}{k + \Delta t/2};\ c_2 = \frac{k - \Delta t/2}{k + \Delta t/2};\ c_1 + c_2 = 1 \qquad (3.23)$$

where: I_{av} = average inflow during Δt (input is in form of histogram)
O = outflow from the linear reservoir

The outflow from the reservoir is the Instantaneous Unit Hydrograph (IUH) for the basin, which has to be transformed by averaging or S-curve technique into the Unit Hydrograph resulting from a rainfall of duration equal to the routing interval.

Internal routing interval

PM Time interval increment parameter

PT_1 Lower rainfall threshold

PT_2 Upper rainfall threshold

In case the time step used in the model is larger than 1 hour, the model simulates the redistribution of water between the various reservoirs with a time step, which is smaller than the time interval of the basic data. Particularly for the infiltration process this effect could be important. Also the rainfall will be lumped to that smaller interval. The number of increments in the time interval is derived from:

$$N_{\Delta t} = 1 + PM * (UZFWC * F + P_{eff}) \quad (3.24)$$

where:

$$F = 1 \quad \text{for } P_{eff} < PT_1 \quad (3.25)$$

$$F = 1/2\ P_{eff} / PT_2 \quad \text{for } PT_1 \leq P_{eff} \leq PT_2 \quad (3.26)$$

$$F = 1 - 1/2PT_2 / P_{eff} \quad \text{for } P_{eff} > PT_2 \quad (3.27)$$

The most important parameter is seen to be *PM*. Taking a very small value for *PM* (say *PM = 0.01*), then $N_{\Delta t}$ remains approximately1. If e.g. *PM = 0.1* then $N_{\Delta t}$ becomes substantially larger than 1. To limit the increase of $N_{\Delta t}$ a low value for PT_1 is to be chosen in combination with a large value of PT_2, which will reduce the value of *F*.

Sequence of parameter estimation

From the presentation above it will be clear that certain parameters should be estimated before other can be assessed. The following sequence is recommended of which the first three steps are mandatory:

1. Segment area
2. Lower zone primary free water parameters *LZPK* and *LZFPM*
3. Lower zone supplemental free water parameters *LZSK* and *LZFSM*
4. Impervious fraction *PCTIM*
5. Upper zone parameters *UZTWM*, *UZK* and *UZFWM*
6. Lower zone tension capacity *LZTWM*
7. Percolation parameters ZPERC and REXP
8. Remaining parameters

Linear reservoirs

An essential feature of the Sacramento model is that the free water reservoirs are considered as linear reservoirs, i.e. there is a linear relation between the reservoir storage S and the outflow Q:

$$S = kQ \tag{3.28}$$

If the recharge is indicated by I, the continuity equation for the linear reservoir reads:

$$dS/dt = I - Q \tag{3.29}$$

Eliminating S from above equations results in a linear first order differential equation in *Q*:

$$\frac{dQ}{dt} + \frac{1}{k}Q - \frac{1}{k}I = 0 \tag{3.30}$$

With I constant and at $t = t_0$ $Qt = Qt_0$ the solution to (3.30) reads:

$$Q_t = I\left(1 - \exp\left(-\frac{(t-t_0)}{k}\right)\right) + Q_{t_0}\exp\left(-\frac{(t-t_0)}{k}\right) \text{for} t \geq t_0 \tag{3.31}$$

When there is no recharge to the reservoir (*I = 0*) equation (3.31) reduces to:

$$Q_t = Q_{t_0}\exp\left(-\frac{(t-t_0)}{k}\right) \tag{3.32}$$

This equation can be compared with (8), using the same notation:

$$Q_t = Q_{t_0}K^{(t-t_0)} \tag{3.33}$$

Hence:

$$K = \exp\left(-\frac{1}{k}\right) \text{or :} k = -\frac{1}{\ln K} \tag{3.34}$$

Expressing time in days, then the amount of water released from the reservoir in 1day amounts according to equation (3.28):

$$S_0 - S_1 = kQ_0 - kQ_1 = kQ_0\left(1 - \exp\left(-\frac{1}{k}\right)\right) = S_0(1-K) \tag{3.35}$$

This is seen to match with e.g. the equations for the lower zone primary free water reservoir, where:

S_0 = LZFPC and 1-K = LZPK (3.36)

Equation (3.34) provides a means to express the lower zone free water parameters in terms of dimensions and physical properties of aquifers. Consider the phreatic aquifer shown in Figure 3.12, which has the following dimensions and properties:

1. The width of the aquifer perpendicular to the channel is L
2. The water table at the divides is h_0 above the drainage base
3. The specific aquifer yield is μ
4. The aquifer transmissivity is T.

The amount of water stored above the drainage base per unit length of channel available for drainage is:

$S = \mu c_1 L h_0$ with ½<c1<1

The discharge to the channel per unit length of channel according to Darcy with the Dupuit assumption

$Q = -2Tdh/dx = 2Tc_2h_0/(L/2)$ with $c_2>1$

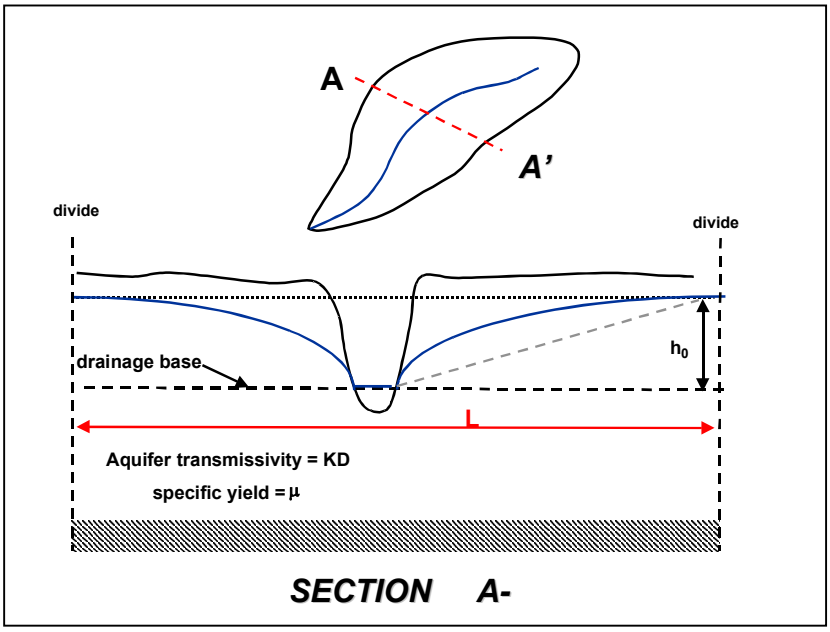


Figure 3.12: Schematic cross section through basin aquifer

Combining the above two equations by eliminating h_0 and bringing it in the form of the linear storage discharge relation (3.28):

$$S = \frac{\mu L^2}{4cT} Q \text{ with : } c = \frac{c_2}{c_1} > 1$$

Hence for the reservoir coefficient k in (3.28) it follows:

$$k = \frac{\mu L^2}{4cT} \qquad (3.37)$$

The reservoir coefficient k is seen to be proportional to the square of the aquifer width and inversely proportional to T, which is logical as k is a measure for the reside-time of the percolated water in the groundwater zone. The value of c varies between 2 and 2.5 dependent on the shape of the water table. For the parameters K and LZPK for the lower zone primary free water storage it then follows:

$$K = \exp(-\frac{4cT}{\mu L^2}) \text{and :} LZPK = 1 - \exp(-\frac{4cT}{\mu L^2}) \qquad (3.38)$$

A similar story applies for the lower zone free supplemental reservoir, which can be viewed as representing the drainage from the shallower based denser network of the smaller channels, see Figure 3.13. Since its main difference is with the aquifer width L, which is much smaller than for the deeper based primary channel network, its reservoir coefficient will be smaller than of the primary free water storage and consequently LZSK >> LZPK.

Note that similar differences in a basin between fast and slow draining aquifers if different soils are present leading to different transmissivities.

Note also that from equation (3.33) it follows for $t - t_0 = 1$ that $K = Q_1/Q_0$. Hence, by deriving this ratio for the recession part of the hydrograph, the parameter K can be obtained from the lowest part of the recession curve where the ratio becomes constant.

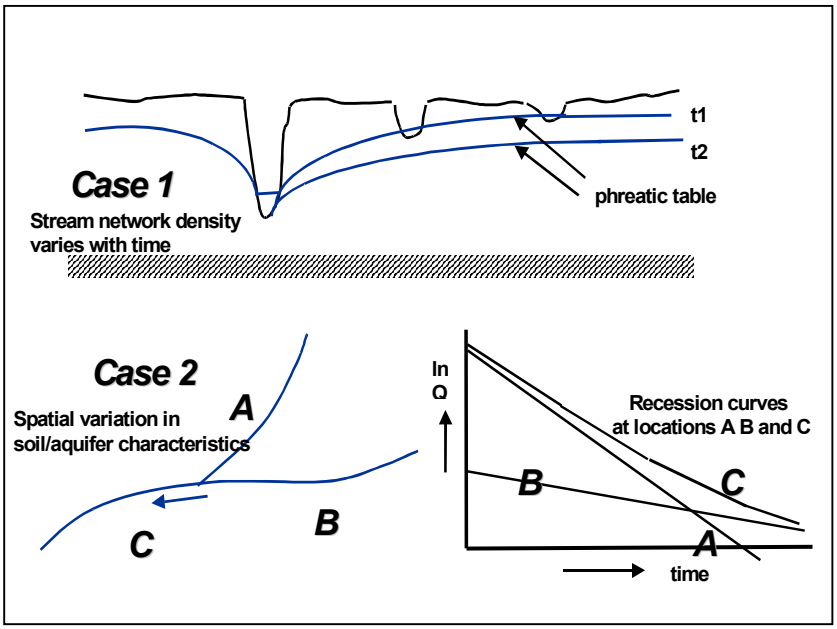


Figure 3.13: Cases of multiple exponential decay of recession curve

3.2.5 REQUIRED INPUT

The input required to run the model for simulation of rainfall-runoff process in a segment is presented in Figure 3.14, which shows the HYMOS screen for running the Sacramento model.

To run the channel routing module the routing parameters, as presented in Section 3.2.3, have to be entered for each distinguished branch. For each branch it has to be specified which hydrograph has to be routed to the next node, which may be:

1. Segment outflow from one or more segments, draining at the upstream channel node
2. Outflow from one or more upstream channel branches
3. Hydrograph presented by the user, e.g. the outflow from a reservoir

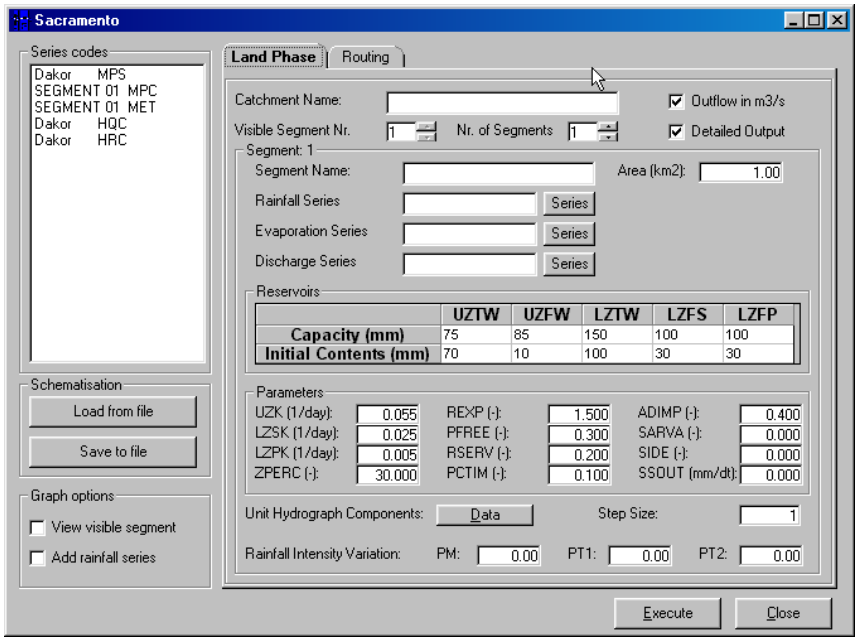


Figure 3.14: Input screen for running Sacramento rainfall-runoff model in HYMOS

Example 3.1 Case study 1:Rainfall-runoff simulation for Dakor basin

A worked out example has been prepared for a part of the KHEDA catchment. The selected basin is located upstream of the stream gauging station Dakor: 'Dakor basin'.

Basin layout and available data

Case study 1 is carried out for Dakor catchment, see Figure 3.15, located in the south-eastern part of the basin indicated in the database as KHEDA catchment, see Figure 3.16.

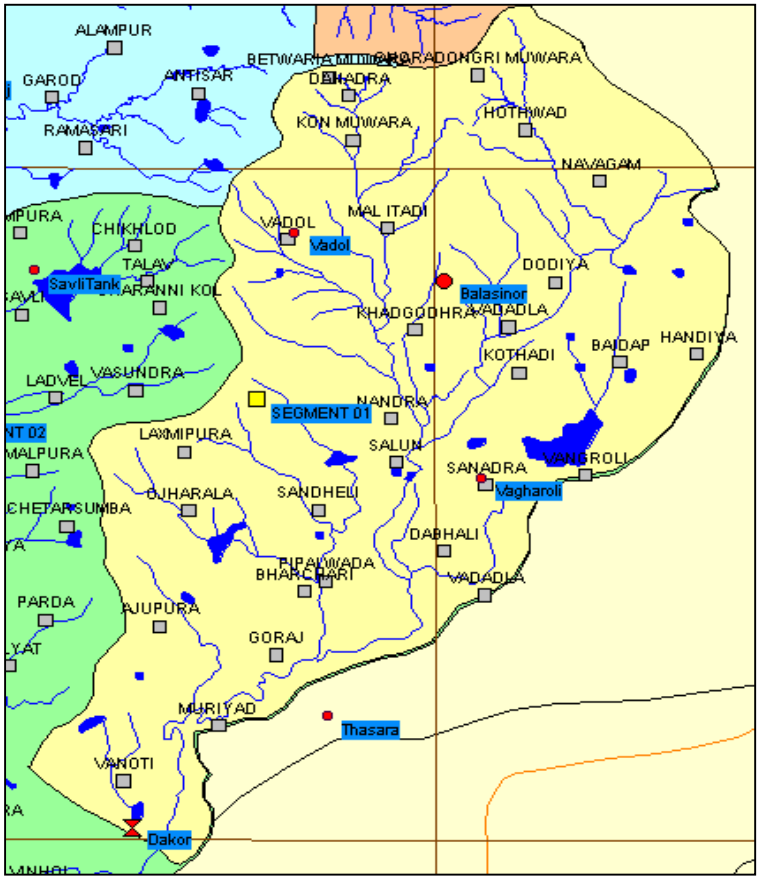


Figure 3.15: Layout of Dakor catchment

The basin measures 430.59 km^2 upstream of gauging station Dakor. The length of the river is 53 km and the slope is approximately 10^{-3}. At the measuring site at Dakor the river is about 60 m wide. The river bed is at 45 m +MSL. From the basin map it is observed that some storage tanks are present in the area. The area contains sandy soils, which dry out quickly.

In and around Dakor basin the following stations are of importance:

Rainfall: Vadol
Balasinor
Savli Tank
Mahisa
Vagharoli
Thasara
Dakor

Evaporation: Anand
Valsad

Streamflow: Dakor

Daily rainfall data is available for quite a number of years for the above-mentioned stations. Hourly rainfall data, however, is lacking. Also a long record of pan evaporation data is available from two stations a little south of the basin, but which are considered representative for the basin. Hourly water level data is available for a large number of monsoon seasons. Prior to and after the monsoon no water level records are available as the river runs generally dry. During the monsoon season four times per day flow measurements are being carried out.

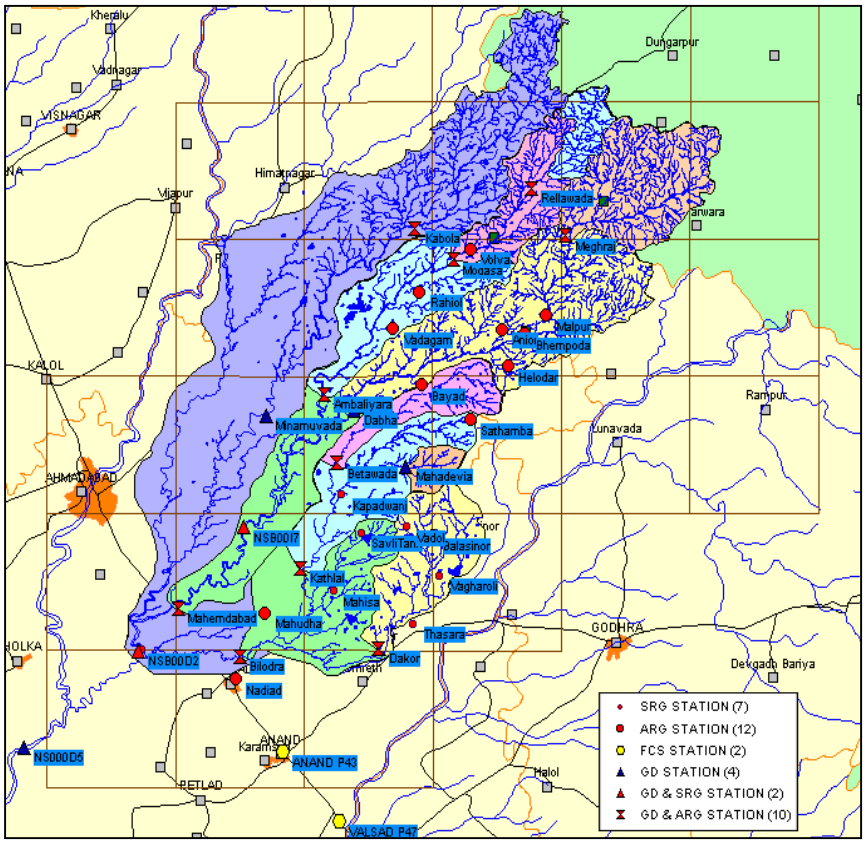


Figure 3.16: Map showing the location of the Dakor basin in the KHEDA catchment

Objective

The objective of this example is to demonstrate the development of a basin rainfall-runoff model based on Sacramento model available in HYMOS as a tool for creation of long series of runoff based on climatic data. Emphasis will be on the steps involved in model calibration and verification. Though the final acceptable result may involve a number of trials, this number can be limited if the initial estimates for the parameter values are carefully made. One should also get an indication of the possible range of the parameters for the basin under study.

The model will be developed using daily data on rainfall, evaporation and runoff. In this case we solely concentrate on segment rainfall-runoff simulation. River routing will not be considered as the interval of one day is too large for meaningful routing in such a small basin. For that hourly data should have been present for rainfall. For water resources analysis routing with an interval of one day will be sufficient.

Basin reconnaissance and input data preparation

After having collected and studied the topographic, geologic, soils and land use maps of the area as well as the characteristics of the hydraulic infrastructure it is imperative that a field visit precedes the model development. Based on differences is drainage characteristics, it may be decided to subdivide the basin in segments. The question on sub-division comes again when dealing with the spatial variability of the rainfall. How far sub-division should take place depends basically on the objective of the study in relation spatial variability. Segment areas in practice vary from a few hundred square kilometers to a few thousand. For water resources assessment studies where an exact reproduction of the shape of the hydrograph is not of importance segments will generally not be small; matching with the locations where flow data are required then also plays a role. Furthermore, practicalities as the availability of a gauging station with calibration data matters. The basin itself acts diffusive and smoothes the differences. Here it assumed that the Dakor catchment is sufficiently homogeneous to be covered by one segment.

Next the input and calibration data are being prepared including catchment rainfall data, potential evapotranspiration and runoff. It is noted here that in view of the objective of the course, being familiarisation with the Sacramento model, data validation is not given the attention it deserves but should be given due attention in actual model development. Completely erroneous models may result from poorly validated data.

For calibration and verification purposes representative periods have to be selected, which include the full gamma of flows. For the case study the year 1994 will be considered for calibration purposes. Flows have been very large that year and also a distinct recession curve is available for parameter estimation.

Catchment rainfall series 1994

All rainfall stations mentioned above were considered for calculation of the areal rainfall. An example of the variability of the point rainfall data on a daily basis is illustrated in Figure 3.17. The Figure also shows that occasionally day shifts in the rainfall data seem to be present. Such errors may deteriorate the quality of the catchment rainfall. Though in this area it is hard to say whether such errors are present as the correlation distance of rainfall events here is are rather small, careful analysis of the daily record casts doubt on the time of occurrence of the rainfall events as reported. Another impression of the spatial variability of the data is obtained from the annual totals as listed for the years 1993 and 1994 in Table 3.2.

Station	Annual Rainfall (mm) 1993	Annual Rainfall (mm) 1994	Thiessen weights
Vadol	590	1317	0.17
Balasinor	574	1485	0.33
Savli Tank	837	1193	0.02
Mahisa	700	775	0.02
Vagharoli	924	1577	0.24
Thasara	991	1775	0.14
Dakor	672	1252	0.08

Table 3.2: Annual rainfall of years 1993 and 1994 and Thiessen weights for areal rainfall computation

From the table it is observed that the spatial variability in the rainfall amounts even at short distances is rather large. The low annual value for Mahisa in 1994 compared to its neighbours is mainly due to the fact that an extremely large rainfall, which occurred in the region, was not available in the Mahisa record (erroneously or not). From the table one can also see that rainfall totals from one year to another may vary considerably.

Thiessen method has been applied to compute the daily areal rainfall in the Dakor basin, see also Figure 3.18 and Table 3.2, where the station weights are presented. It is observed that the contributions of Savli Tank and Mahisa in the areal total for the Dakor basin are small, hence the doubts on the Mahisa record for 1994 will not greatly affect the computed areal average.

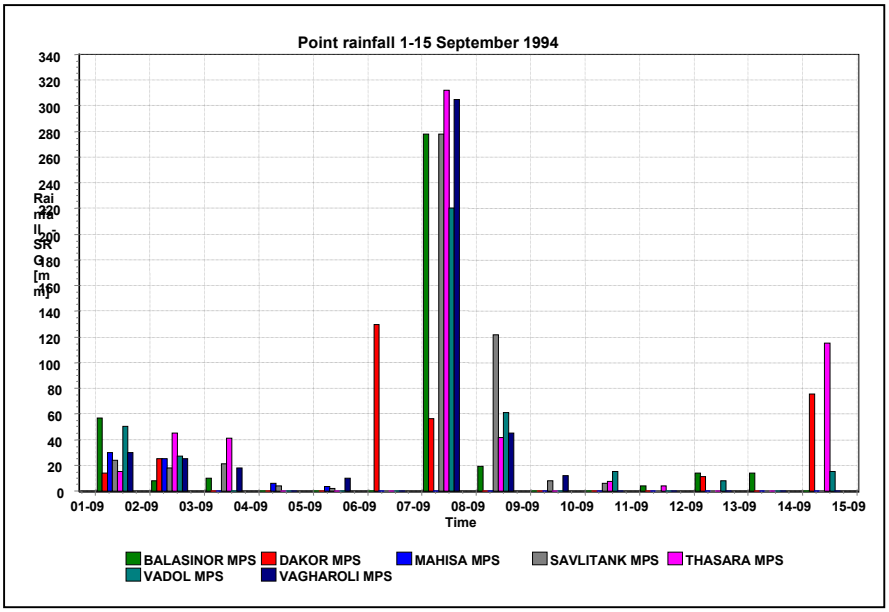


Figure 3.17: Example of daily point rainfall data at the selected stations

The resulting daily average rainfall for the year 1994 is presented in Figure 5. It may be observed that the rainfall occurs from mid June till mid September only. The annual total amounts 1483 mm.

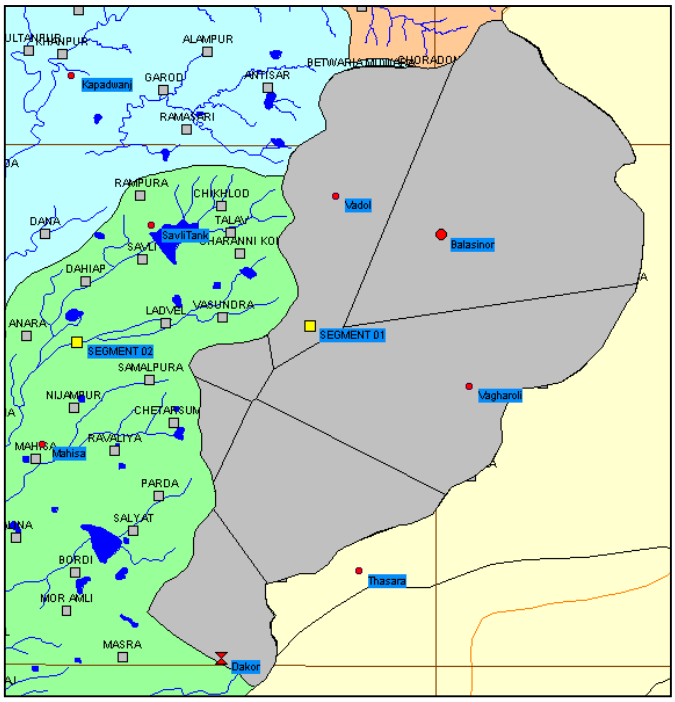


Figure 3.18: Thiessen polygon for Dakor basin rainfall

Potential evapotranspiration series 1994

The potential evapotranspiration in Dakor basin is derived from the pan-evaporation records available from the stations Anand and Valsad. To transform pan evaporation into potential evapotranspiration generally pan coefficients ranging from 0.6 to 0.8 are being applied. Here, an average value of 0.7 is used. The variation of the potential evapotranspiration through the year is presented in Figure 6. Make sure that the evapotranspiration series does not include missing data. The annual total potential evapotranspiration for 1994 amounts 1483 mm, which is coincidentally exactly equal to the computed basin average rainfall. It is observed that the potential evapotranspiration during the monsoon season drops to about 2 mm/day in July and August, with an average of 2.9 mm/day from June till September.

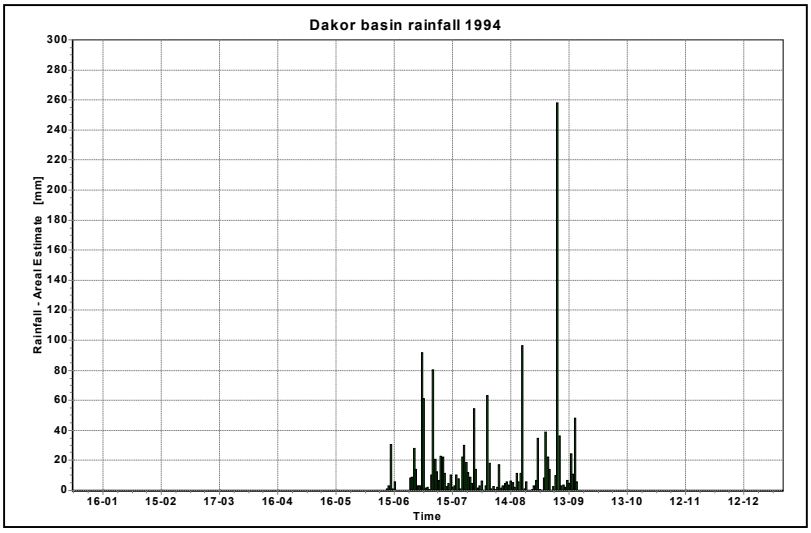


Figure 3.19 Daily rainfall in Dakor basin for the year 1994

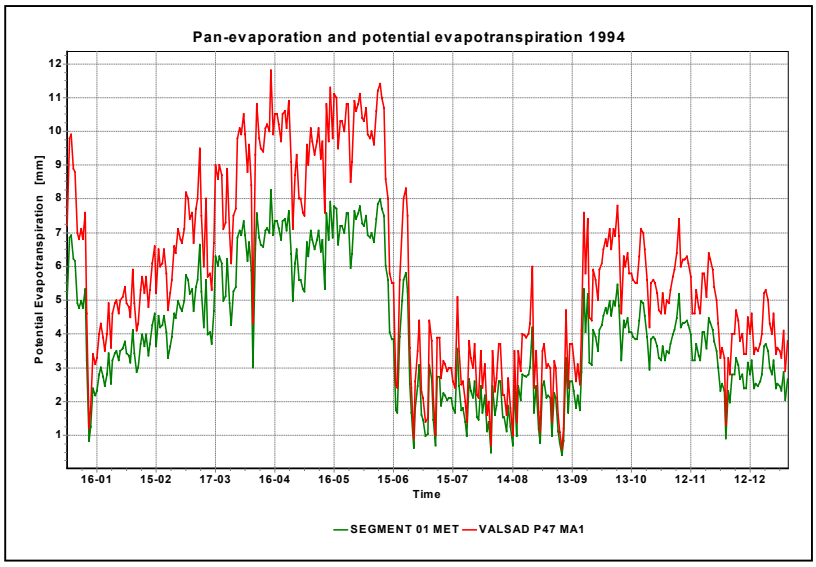


Figure 3.20: Pan evaporation and potential evapotranspiration near Dakor 1994

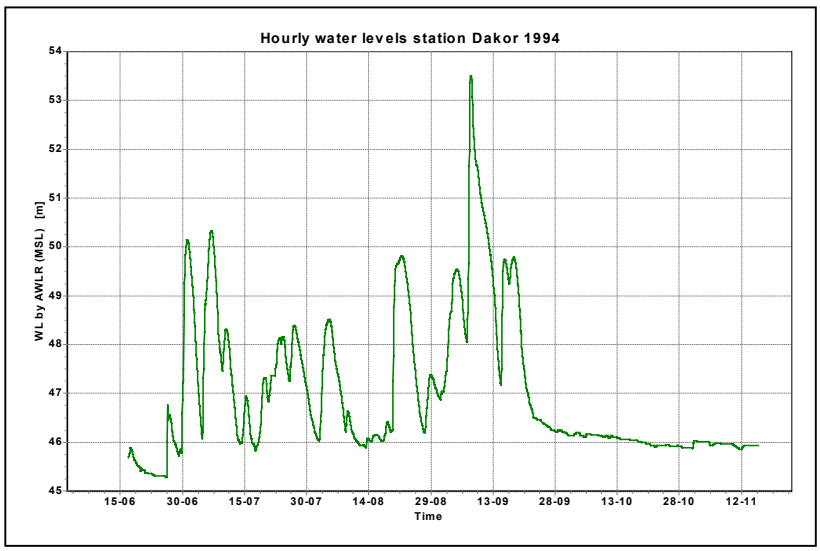


Figure 3.21: Hourly water levels Dakor 1994

Water levels and discharges for 1994

The runoff series for Dakor are derived from the hourly water level record available for that station. The entire water level record for 1994 is presented in Figure 3.21.

The water levels are seen to vary between 45 to over 53 m+MSL at Dakor, i.e. 8 m difference. One also observes that the peaks are rounded, quite different from the water level record at Bilodra, available in your database.

The water level data are transformed into discharges by means of stage-discharge relations. The relations were fitted to the data presented in Figure 3.22. Two rating curves were developed, one valid till 14 July and one valid thereafter. About the cause of the change no further investigations were made, but the changes are most likely caused by shifts in the control section due to morphology. Note that the change takes place after the occurrence of the first peaks in 1994. The rating curves fitted to the data from 15 July onward are shown in Figure 3.23.

Note that a few measurements in the higher flow region have been omitted. These data referred to a falling stages but these data plotted to the right of the curve matching with the highest flows during steady stages. For a stable channel the data should have plotted to the left of the curve and from that point of view were considered inconsistent. One reason for plotting right might have been that the downstream control section has drastically eroded during the passage of the flood wave.

The resulting hourly discharge hydrograph at Dakor for 1994 is shown in Figure 3.24.

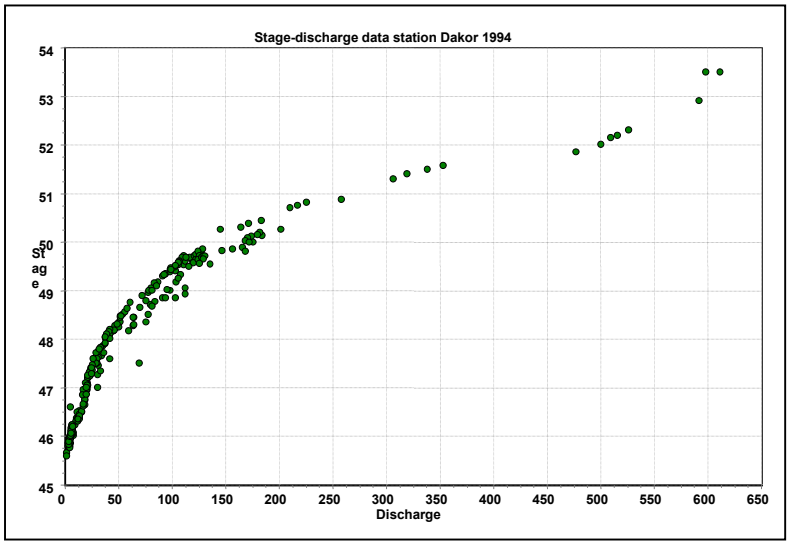


Figure 3.22: Stage-discharge measurements of 1994 for station Dakor

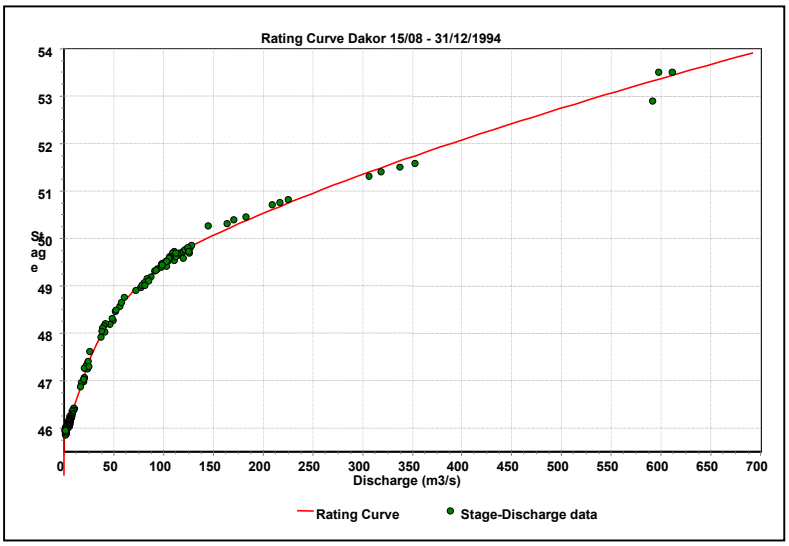


Figure 3.23: Stage-discharge relation for station Dakor (15/8 – 31/12/1994)

The discharges in m^3/s have subsequently been transformed to hourly runoff values in mm/hr by multiplying the discharges with 3600 (s)/area(km^2) x 10^{-3}. Subsequently, the hourly runoff values have been aggregated to daily values, in a manner equal to the way daily rainfall data are treated, i.e. from 8.00 hrs at day 1 to 8.00 hrs at day 2, reported as a daily value for day 2. This requires special attention while executing the aggregation.

The total runoff for the year 1994 amounted 1062 mm, i.e. a runoff coefficient of 72%.

The daily rainfall, potential evapotranspiration and runoff data for 1994 are tabulated in Tables 3.3, 3.4 and 3.5.

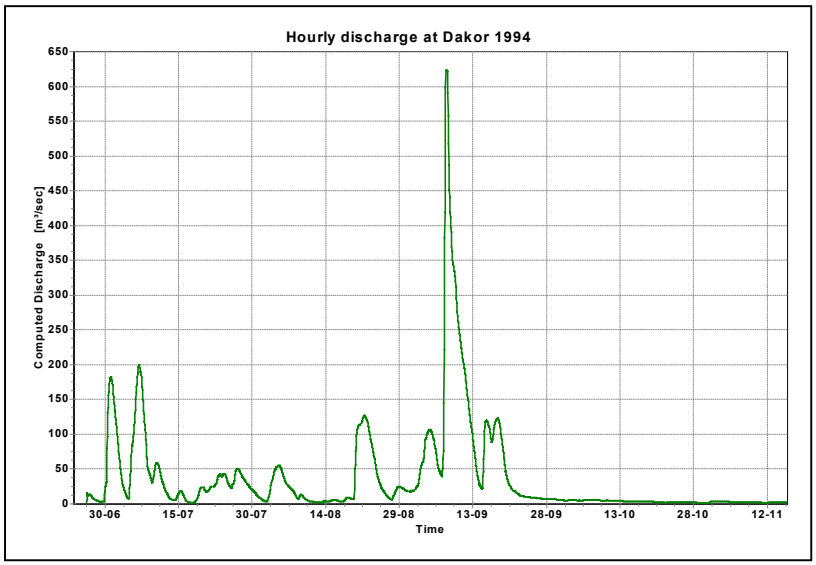


Figure 3.24: Hourly runoff at Dakor for 1994

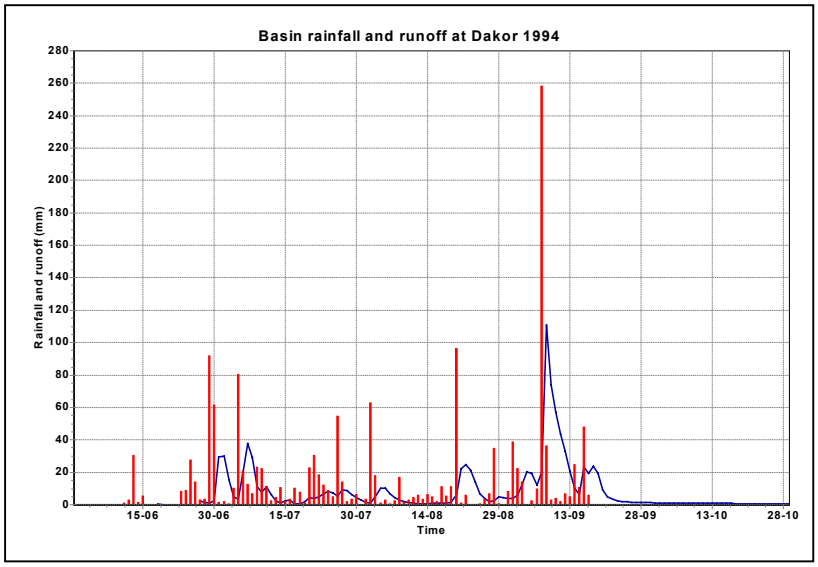


Figure 3.25: Daily rainfall and runoff for 1994 for Dakor basin

Estimation of parameters

The model parameter estimation based on the data for the year 1994 is carried along the lines as presented in the text in Chapter 3.2.4.

Segment area

Segment area is derived from the basin boundary data in HYMOS:

Segment area = 430.59 km^2.

Lower zone primary free water storage parameters *LZPK* and *LZFPM*

Reference is made to the semi-logarithmic plot of the runoff series, shown in Figure 3.26. Lowest runoff values with an exponential decay showing as a straight line in the plot are present at the end of October and in November, see Figure 3.27. A straight line is fitted the observations and the drainage factor *LZPK* is determined from the runoff values at 31/10 and 12/11, which are solely attributed to runoff from the lower zone primary free water storage. It then follows from equation

$$KP = \left(\frac{R_{(12/11)}}{R_{(31/10)}}\right)^{1/12} = \left(\frac{0.23}{0.27}\right)^{1/12} = 0.986 \qquad (3.8)$$

Hence the drainage factor becomes with equation (3.9):

LZPK = 1 – KP = 1 – 0.986 = 0.014

To arrive at a value for the capacity of the lower zone primary free water storage an estimated maximum runoff value from that reservoir is required RP max. Assuming that after a very wet period this maximum is achieved one can estimate this maximum by extrapolating the primary baseflow recession curve backward in time up to the runoff peak on 8 September. Under the presumption that this storm has completely filled the lower zone primary free water storage a maximum value of approximately 0.6 mm/day can be read from the semi-log plot. (This value can also be computed from the value at 31/11 using KP and the time interval from 8/9 till 31/10: $QP(8/9) = QP(31/11)KP^{-53} = 0.57$ mm/day).

$$LZFPM = \frac{QP\,max}{LZPK} \approx \frac{0.6}{0.014} \approx 45mm$$

Hence the LZFPM becomes with equation (3.10):

Note that the value is rounded to the nearest 5 mm, in view of the uncertainties involved.

It is noted here that extrapolation of the primary baseflow recession curve backward from 31/10 to 8/9 is not entirely correct as in between some recharge may have occurred by the storm between 16 and 18/9. On the other hand we are not entirely sure that on 8/9 the lower free water zone was completely filled.

Abstractions from river flow may affect the result. Abstractions may be observed from recession on semi-log plot failing to fall to a straight line (it curves downward in the course of time). By adding a constant amount a straight line can often be obtained. The effect is illustrated in Figure 3.28.

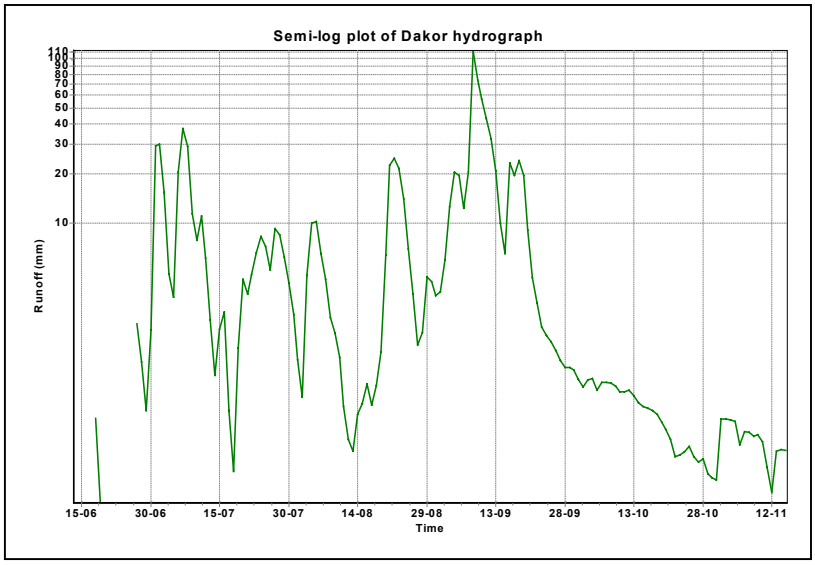


Figure 3.26: Semi-logarithmic plot of runoff series for Dakor 1994

It is observed that the result is rather sensitive to abstractions and due care should be given to this phenomenon. If abstractions are present and no corrections are made than the estimated *LZPK*-value will be too high.

Question: what would be the estimates for *LZPK* and *LZFPM* if 0.1 mm/day is added to the recession curve??

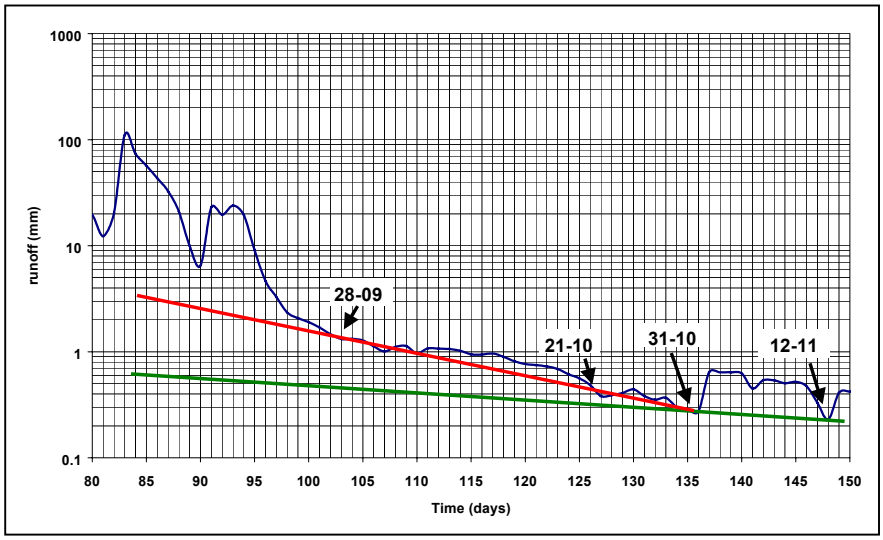


Figure 3.28: Detail of semi-log plot of Dakor hydrograph for estimation of lower zone free water storage parameters

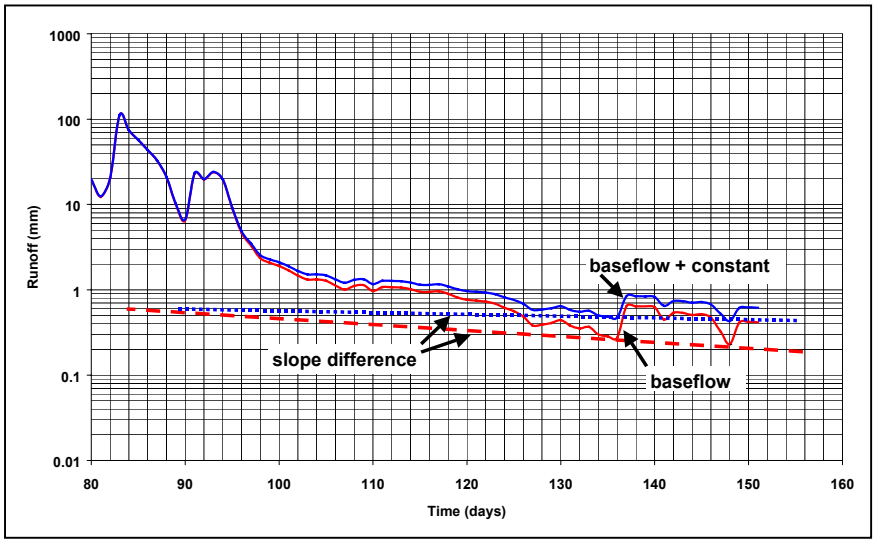


Figure 3.29: Effect of abstraction of river flow on primary baseflow parameters

Daily data and statistics of series SEGMENT 01 MPC Year = 1994

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	.00	.00	.00	.00	.00	.00	1.54	3.23	38.90	.00	.00	.00
2	.00	.00	.00	.00	.00	.00	1.88	63.02	22.35	.00	.00	.00
3	.00	.00	.00	.00	.00	.00	.57	18.01	13.99	.00	.00	.00
4	.00	.00	.00	.00	.00	.00	10.19	1.00	.20	.00	.00	.00
5	.00	.00	.00	.00	.00	.00	80.43	2.83	2.52	.00	.00	.00
6	.00	.00	.00	.00	.00	.00	20.56	.41	9.69	.00	.00	.00
7	.00	.00	.00	.00	.00	.00	12.47	2.21	258.01*	.00	.00	.00
8	.00	.00	.00	.00	.00	.00	6.53	17.11	36.08-	.00	.00	.00
9	.00	.00	.00	.00	.00	.00	23.00	1.44	3.08	.00	.00	.00
10	.00	.00	.00	.00	.00	.00	22.04	2.92	3.64	.00	.00	.00
11	.00	.00	.00	.00	.00	1.00	11.28	4.42	1.90	.00	.00	.00
12	.00	.00	.00	.00	.00	3.00	2.38	5.58	6.83	.00	.00	.00
13	.00	.00	.00	.00	.00	30.36	4.53	3.55	4.67	.00	.00	.00
14	.00	.00	.00	.00	.00	1.28	10.52	6.26	24.53	.00	.00	.00
15	.00	.00	.00	.00	.00	5.55	2.08	5.02	10.62	.00	.00	.00
16	.00	.00	.00	.00	.00	.11	3.33	2.10	48.08	.00	.00	.00
17	.00	.00	.00	.00	.00	.00	10.20	11.23	5.60	.00	.00	.00
18	.00	.00	.00	.00	.00	.00	7.94	5.51	.00	.00	.00	.00
19	.00	.00	.00	.00	.00	.00	.84	11.15	.00	.00	.00	.00
20	.00	.00	.00	.00	.00	.00	22.50	96.36	.00	.00	.00	.00
21	.00	.00	.00	.00	.00	.00	30.31	.96	.00	.00	.00	.00
22	.00	.00	.00	.00	.00	.00	18.47	5.93	.00	.00	.00	.00
23	.00	.00	.00	.00	.00	8.25	11.95	.00	.00	.00	.00	.00
24	.00	.00	.00	.00	.00	8.55	8.74	.00	.00	.00	.00	.00
25	.00	.00	.00	.00	.00	27.80	4.92	.70	.00	.00	.00	.00

Table 3.3: Daily rainfall in Dakor segment for 1994

26	.00	.00	.00	.00	.00	13.94	54.55	3.34	.00	.00	.00	.00
27	.00	.00	.00	.00	.00	3.12	13.91	6.81	.00	.00	.00	.00
28	.00	.00	.00	.00	.00	3.22	1.79	34.75	.00	.00	.00	.00
29	.00	*********	.00	.00	.00	91.97	3.22	.33	.00	.00	.00	.00
30	.00	*********	.00	.00	.00	61.31	6.37	.00	.00	.00	.00	.00
31	.00	*********	.00	*********	.00	*********	.00	8.26	*********	.00	*********	.00
Data	31	28	31	30	31	30	31	31	30	31	30	31
Eff.	31	28	31	30	31	30	31	31	30	31	30	31
Miss	0	0	0	0	0	0	0	0	0	0	0	0
Sum	.00	.00	.00	.00	.00	259.46	409.04	324.44	490.70	.00	.00	.00
Mean	.00	.00	.00	.00	.00	8.65	13.19	10.47	16.36	.00	.00	.00
Min.	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Max.	.00	.00	.00	.00	.00	91.97	80.43	96.36	258.01	.00	.00	.00

Annual values:

Data	365 * Sum	1483.63 * Minimum	.00 * Too low	0	
Effective	365 * Mean	4.06 * Maximum	258.01 * Too high	1	
Missing	0				

Daily data and statistics of series SEGMENT 01 MET Year = 1994

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	5.04	3.36	4.97	6.93	5.25	7.49	.98	1.12	2.17	4.62	3.43	2.31
2	6.86	3.15	5.74	6.16	6.72	6.93	1.05	1.40	2.10	4.97	3.71	1.96
3	6.93	4.13	5.60	6.72	6.30	6.86	3.08	.49	.98	4.55	3.99	2.80
4	6.23	3.43	5.18	5.88	7.07	7.00	2.66	2.45	2.24	4.97	4.13	2.80
5	6.16	2.87	5.32	3.01	6.79	6.72	1.47	1.61	2.10	4.83	4.48	3.29
6	4.90	3.01	4.69	6.51	6.51	7.42	.70	1.89	1.12	5.46	5.18	3.08
7	4.76	3.71	5.39	7.56	6.79	7.84	2.73	2.59	.77	4.83	4.20	2.66
8	4.97	3.99	5.60	6.86	7.07	7.98	2.73	2.59	.42	3.22	4.34	2.80
9	4.76	3.64	6.65	6.65	6.44	7.70	1.89	1.54	.84	4.41	4.34	2.38
10	5.32	3.99	5.25	6.58	6.79	7.49	2.24	1.54	3.29	4.20	4.41	2.38
11	3.22	3.36	4.20	7.07	5.32	6.02	2.17	1.12	1.68	4.48	4.27	3.15
12	.84	3.71	5.60	7.14	7.56	5.60	2.03	1.89	2.59	4.06	3.99	2.80
13	1.26	4.27	3.99	7.00	6.79	4.06	2.10	1.26	2.59	4.06	3.22	3.22
14	2.38	4.62	4.06	8.26	7.91	3.85	2.10	.70	2.31	3.92	3.22	2.38
15	2.17	3.64	3.71	6.93	6.86	3.85	1.82	2.45	1.82	3.85	3.71	2.52
16	2.31	4.55	4.69	7.35	7.77	1.75	1.68	.98	2.17	3.85	3.36	2.45
17	2.80	4.20	6.30	7.35	7.70	1.68	3.57	2.45	1.75	4.41	3.22	2.59
18	3.01	4.27	6.02	7.14	6.65	3.92	2.45	2.03	2.87	4.97	4.06	2.80
19	2.73	4.55	6.30	6.79	7.21	4.97	1.75	2.80	5.32	4.90	4.06	3.64
20	2.45	4.06	6.09	7.35	7.21	5.60	1.82	2.77	4.06	4.55	3.57	3.71
21	2.80	3.29	4.97	7.42	7.00	5.81	1.40	2.73	5.18	3.92	4.48	3.50
22	3.43	3.64	5.11	7.07	7.56	5.25	.98	2.80	3.15	2.94	4.34	3.01
23	2.52	3.92	6.23	7.63	7.56	2.52	2.66	3.01	3.08	3.85	4.13	2.80
24	3.22	4.62	4.90	6.37	5.95	1.68	2.31	4.20	4.13	3.92	3.78	3.22
25	3.43	4.48	4.27	4.97	6.37	.63	2.10	1.68	3.92	3.85	3.50	2.38

26	3.50	4.97	5.25	6.09	7.63	1.82	2.59	2.45	3.50	3.64	3.01	2.52
27	3.22	4.76	5.39	6.51	7.42	2.45	1.54	1.19	4.13	3.29	2.31	2.45
28	3.50	4.69	6.86	5.60	7.56	3.08	1.47	.77	4.27	3.22	2.52	2.31
29	3.57	*********	7.07	5.60	7.77	1.61	2.45	2.45	4.55	3.64	2.31	2.87
30	3.78	*********	6.93	5.32	7.28	1.47	1.68	2.59	4.76	3.22	.91	2.03
31	3.43	*********	7.35	*********	7.21	*********	2.17	2.10	*********	3.50	*********	2.66
Data	31	28	31	30	31	30	31	31	30	31	30	31
Eff.	31	28	31	30	31	30	31	31	30	31	30	31
Miss	0	0	0	0	0	0	0	0	0	0	0	0
Sum	115.50	110.88	169.68	197.82	216.02	141.05	62.37	61.63	83.86	128.10	110.18	85.47
Mean	3.73	3.96	5.47	6.59	6.97	4.70	2.01	1.99	2.80	4.13	3.67	2.76
Min.	.84	2.87	3.71	3.01	5.25	.63	.70	.49	.42	2.94	.91	1.96
Max.	6.93	4.97	7.35	8.26	7.91	7.98	3.57	4.20	5.32	5.46	5.18	3.71

Annual values:

Data	365 * Sum	1482.56 * Minimum	.42 * Too low	0		
Effective	365 * Mean	4.06 * Maximum	8.26 * Too high	0		

Daily data and statistics of series Dakor HRC Year = 1994

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	29.42	1.48	3.83	1.13	.64	-999.99*
2	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	30.15	.87	5.99	1.01	.64	-999.99*
3	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	15.36	4.82	12.58	1.11	.64	-999.99*
4	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	4.93	10.00	20.27	1.13	.62	-999.99*
5	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	3.55	10.23	19.47	.97	.45	-999.99*
6	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	20.30	6.55	12.29	1.07	.54	-999.99*
7	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	37.48	4.55	20.33	1.07	.54	-999.99*
8	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	29.02	2.68	110.92*	1.06	.51	-999.99*
9	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	11.34	2.14	73.86	1.02	.52	-999.99*
10	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	7.84	1.53	56.92	.95	.47	-999.99*
11	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	11.02	.77	43.44	.94	.33	-999.99*
12	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	6.11	.48	32.67	.96	.23	-999.99*
13	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	2.58	.41	20.85	.90	.41	-999.99*
14	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	1.18	.69	9.99	.81	.42	-999.99*
15	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	2.27	.80	6.51	.77	.42	-999.99*
16	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	2.89	1.05	23.10	.75	-999.99*	-999.99*
17	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	.72	.78	19.58	.73	-999.99*	-999.99*
18	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	.65	.31	1.03	23.90	.69	-999.99*	-999.99*
19	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	.20	1.73	1.65	19.49	.61	-999.99*	-999.99*
20	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	4.56	6.39	9.09	.56	-999.99*	-999.99*
21	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	3.68	22.37	4.67	.49	-999.99*	-999.99*
22	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	4.86	24.74	3.29	.38	-999.99*	-999.99*
23	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	6.62	21.46	2.35	.39	-999.99*	-999.99*
24	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	8.27	14.03	2.07	.41	-999.99*	-999.99*
25	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	7.23	7.00	1.91	.44	-999.99*	-999.99*

26	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	5.18	3.72	1.69	.38	-999.99*	-999.99*
27	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	2.44	9.25	1.82	1.47	.35	-999.99*	-999.99*
28	-999.99*	-999.99*	-999.99*	-999.99*	-999.99*	1.43	8.47	2.18	1.32	.37	-999.99*	-999.99*
29	-999.99**********		-999.99*	-999.99*	-999.99*	.73	6.16	4.69	1.32	.30	-999.99*	-999.99*
30	-999.99**********		-999.99*	-999.99*	-999.99*	2.23	4.33	4.38	1.28	.28	-999.99*	-999.99*
31	-999.99**********		-999.99**********		-999.99**********		2.79	3.63	*********	.27	*********	-999.99*

Data	31	28	31	30	31	30	31	31	30	31	30	31
Eff.	0	0	0	0	0	6	31	31	30	31	15	0
Miss	31	28	31	30	31	24	0	0	0	0	15	31
Sum	-999.99	-999.99	-999.99	-999.99	-999.99	7.69	289.61	168.95	566.45	22.30	7.37	-999.99
Mean	-999.99	-999.99	-999.99	-999.99	-999.99	1.28	9.34	5.45	18.88	.72	.49	-999.99
Min.	-999.99	-999.99	-999.99	-999.99	-999.99	.20	.31	.41	1.28	.27	.23	-999.99
Max.	-999.99	-999.99	-999.99	-999.99	-999.99	2.44	37.48	24.74	110.92	1.13	.64	-999.99

Annual values:

Data	365 * Sum	1062.37 * Minimum	.20 * Too low	0		
Effective	144 * Mean	7.38 * Maximum	110.92 * Too high	1		

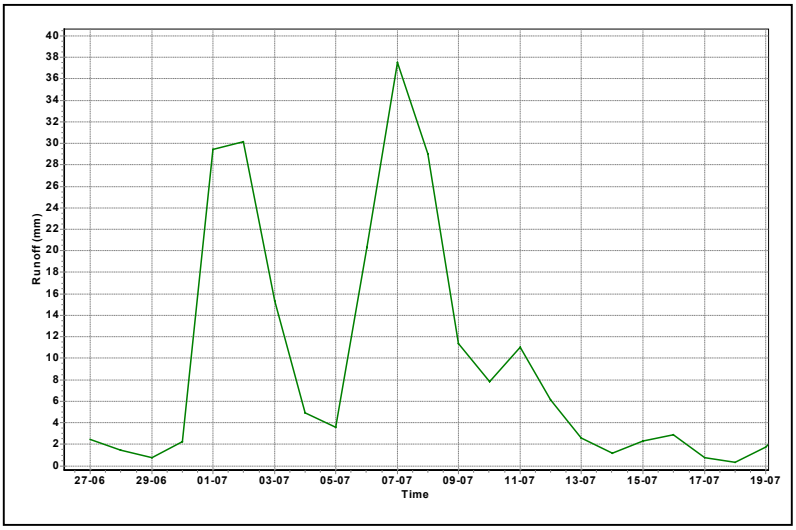


Figure 3.29: Part of hydrograph used for LZTWM

Assuming that changes in lower zone free water storages as well as the upper zone storages are small, from a simple waterbalance computation for the period 29/6 to 14/7 the following change in the LZTW storage is observed

ΔLZTWC = ΣP – ΣR - ΣE = 361 – 217 – 31 = 113 mm

Note that the evaporation is taken as potential as the UZTW storage was filled.

The value obtained in this manner is certainly a lower limit as by mid July there is still capacity in the lower zone. The first estimate is therefore set as:

LZTWM = 150 mm

Extending the period to 24/9 leads to a value of 1377 – (1016 + 180) = 181 mm, but then corrections for the free lower zone storages have to be taken into account as well. Also the errors may occur, stemming from the fact that the evaporation is not at its potential rate etc.

Percolation parameters ZPERC and REXP

From equation 17 a first estimate for ZPERC can be obtained. It requires the value of PBASE:

PBASE = LZFPM x LZPK + LZFSM x LZSK = 45 x (0.014 + 0.067) = 3.64mm/day

Hence with equation 17:

$$\text{ZPERC} = \frac{\text{LZTWM} + \text{LZFSM} + \text{LZFPM} - \text{PBASE}}{\text{PBASE}} = \frac{150 + 45 + 45 - 3.64}{3.64} = 57$$

So as a first approximation a value of 60 is assumed.

REXP is estimated at 1.5 as the soils appear to be sandy, see Table 3.1.

Unit hydrograph parameters

The concentration time for the Dakor basin can be estimated by assuming a celerity between 2 to 3 m/s or 7 to 10 km/hr during floods. With a total river length of 53 km it implies that the concentration time will be in the order of 5 to 8 hrs, which is much smaller than the time interval to be used in the simulation. The hydrograph though shows that the surface runoff is considerably delayed. An approximation for the unit hydrograph components based on inspection of the runoff compared to the rainfall gives the following hydrograph values:

0.15, 0.40, 0.30, 0.15

The Clark procedure could also be used here. This is discussed below.

From Figure 3.30 a concentration time is computed from a comparison of the rainfall and runoff record. The time between the cessation of rainfall to the inflection point on the falling limb of the hydrograph is a good indicator for the time of concentration T_c. From Figure 3.30 a value of 2 days ($\pm$ 0.5 days) is read.

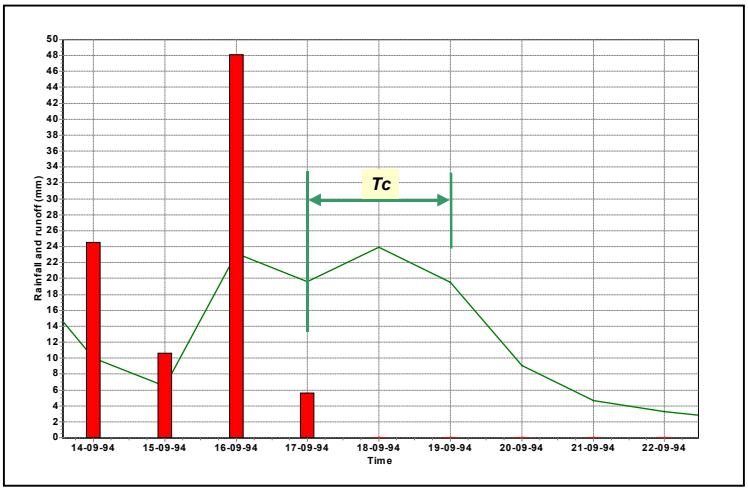


Figure 3.30: Estimation of the time of concentration from rainfall and runoff record.

In the time area diagram 2 intervals are considered. The time area diagram presents the hydrograph resulting from an instantaneous supply of 1 mm over the catchment. Since we consider two intervals only one isochrone is considered. The isochrone separate the segment into two parts where, given the shape of the segment, the lower part constitutes about 40% of the segment and the upper part 60%. Again great detail is not required as we are dealing with daily data in a small segment, with a T_c value somewhere between 1.5 and 2.5 days. The approximate time area diagram is presented in Figure 3.32.

The next step is to estimate the reservoir coefficient k. This coefficient is obtained from the slope of the recession of the surface water hydrograph. For this Figure 3.31 is observed. If the baseflow part is subtracted from the actual flow values then the surface runoff is seen to reduce from 18 mm/day to 2 mm/day in 2 days. Hence, k is obtained from:

$$Q_{t2} = Q_{t1} \exp(-\frac{(t2-t1)}{k})$$

$$k = -\frac{(t2-t1)}{\ln(Q_{t2}/Q_{t1})} = -\frac{2}{\ln(2/18)} = \frac{2}{-2.20} = 0.91 \text{days}$$

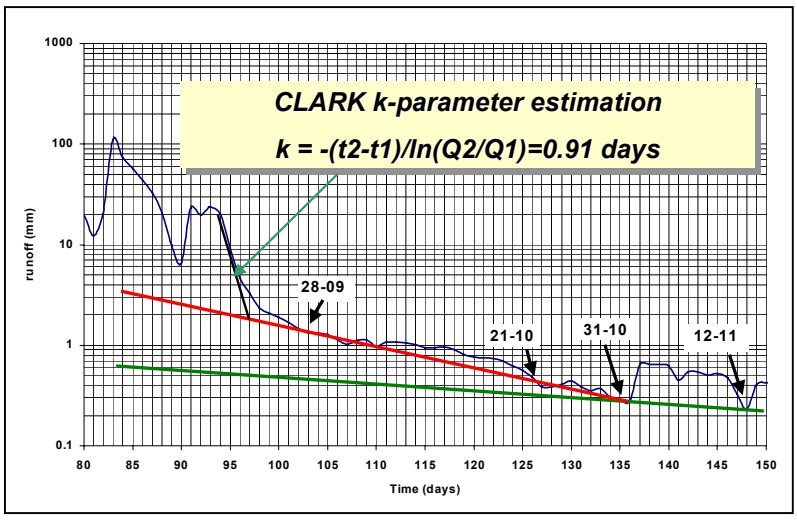


Figure 3.31: Estimation of reservoir coefficient parameter in Clark method

Note that for the estimation of k one should particularly concentrate on the surface runoff part and not on the interflow part as interflow is already delayed by the UZFW reservoir.

The routing coefficients then become according to equation 23 with Δt = 1 day and k = 0.91 days:

$$c1 = \frac{\Delta t}{k + \Delta t/2} = \frac{1}{0.91 + 1/2} = 0.71; \; c2 = 1 - c1 = 1 - 0.71 = 0.29$$

Hence, the routing equation becomes:

Qi+1 = c1 x Iav + c2 x Qi = 0.71 x Iav + 0.29 x Qi, etc.

The result is the instantaneous unit hydrograph. The 1-day unit hydrograph is obtained by averaging over successive intervals: Qday, i = ½(Qinst,i + Qinst,i-1).

The routing is carried out in the Table below, and the result is presented in Figure 3.32.

Time	Input I_{av}	$Q_{out\text{-}inst}$	$Q_{out\text{-}day}$
0	**0**	**0.00**	**0.00**
1	**0.4**	**0.28**	**0.14**
2	**0.6**	**0.51**	**0.40**
3	**0**	**0.15**	**0.33**
4	**0**	**0.04**	**0.10**
5	**0**	**0.01**	**0.03**
6	**0**	**0.00**	**0.01**
7	**0**	**0.00**	**0.00**

Table 3.6 Conversion of time area diagram into 1-day unit hydrograph

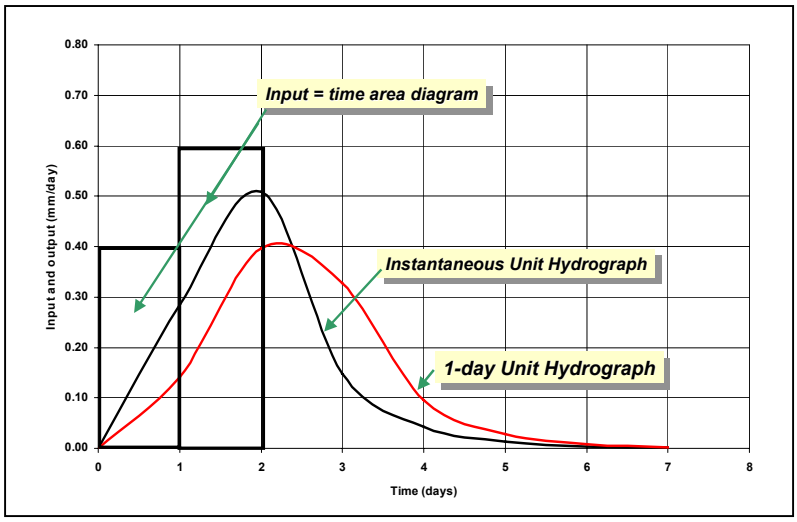


Figure 3.32: Time-area diagram and instantaneous and 1 day unit hydrograph

The values of the last column in the Table are input to the model as the 1-day unit hydrograph.

Other parameters

The *SIDE* parameter needs special attention in the Dakor case as the actual groundwater tables are far below the drainage base. It implies that water from the lower zone free water reservoirs will percolate further down to the deep groundwater table. To estimate the value of *SIDE* the unobserved portion of groundwater should be determined. Say, 100 mm is withdrawn from aquifer by mining and the groundwater tables are declining annually with 2 m, and the specific yield is 0.3. Then 100 mm + 0.3 x 2 m = 160 mm is withdrawn from the aquifer. This value has to be compared with the observed baseflow. Then *SIDE* follows from:

$$\text{SIDE} = \frac{\text{unobserved base flow}}{\text{observed base flow}} = \frac{160}{\text{observed base flow}}$$

The mined amount of groundwater (100 mm) is to be added to the model as rainfall.

The other parameters are set to their nominal values as the hydrograph do not permit estimation of e.g. *PCTIM* or *ADIMP*. For *PCTIM* the very first period in the monsoon would have been appropriate, but water level observations started too late for that and some days have missing values. The total list of first estimate of the parameters is shown below.

The initial contents when starting the run on 1/1/94 by assuming that in September the previous year all free base flow reservoirs were full. The potential evapotranspiration from September to January amounts about 400 mm so the tension water reservoirs are expected to have dried up.

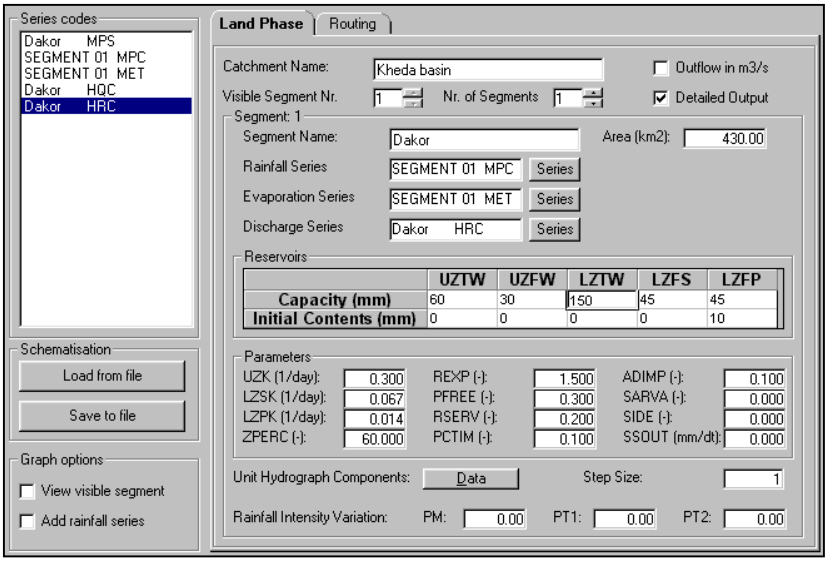


Figure 3.33: List of input parameters for Sacramento model

First runs

The results of the first run are shown in Figure 3.34. The accumulated difference between observed and simulated run off closes at – 41 mm. Matching the water balance is the most important first step while continuing with calibration. By adding 50 mm to the *LZTWM* the difference is seen to be nearly eliminated, Figure 3.35. Then the fine-tuning can start. Particular attention is required to parameter *SIDE*. From the first run an observed base flow of 288 mm. With the assumed removal of 160 mm an estimate for the *SIDE* parameter would be 160/288 = 0.56.

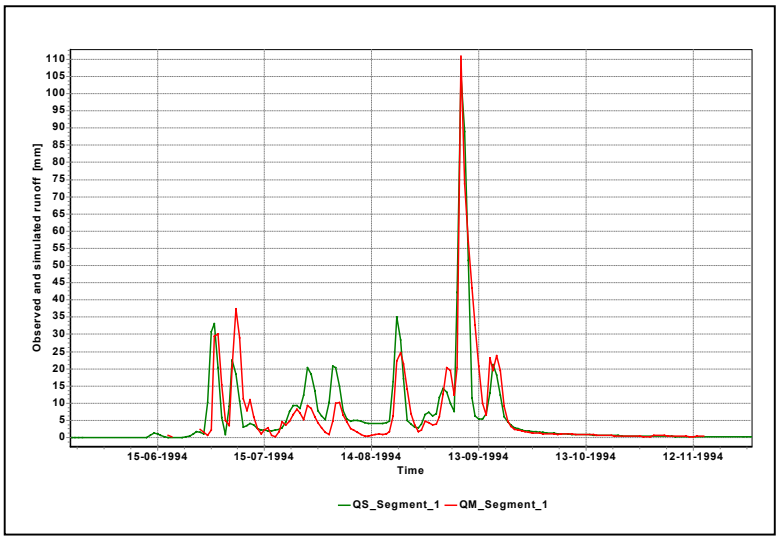


Figure 3.34: Observed and simulated runoff, first run

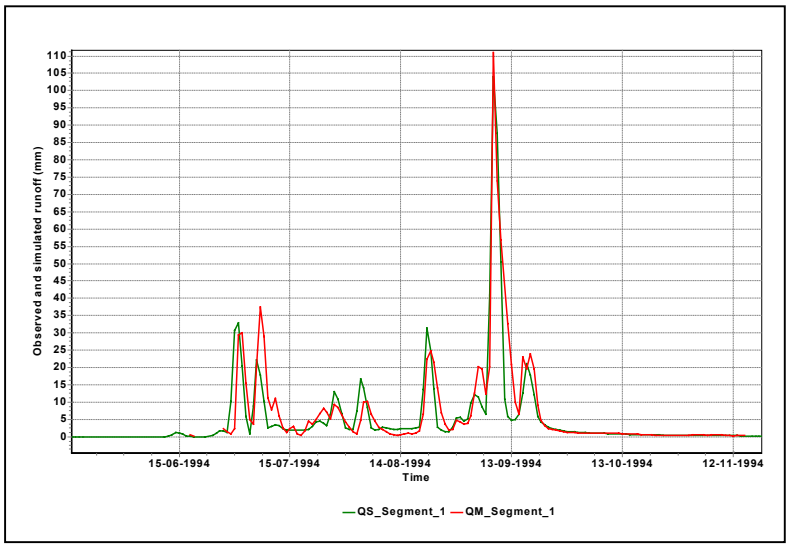


Figure 3.35: Observed and simulated flow, LZTWM increased to 200 mm

```
Simulation of rainfall-runoff process: Land Phase

     Catchment name    = Kheda basin

     Number of segments:  1

 Rainfall-runoff simulation of segment  1

 Catchment: Kheda basin             Segment: Dakor

Rainfall series   : SEGMENT 01  MPC
Evaporation series: SEGMENT 01  MET
Discharge series  : Dakor       HRC

                   UZTW  UZFW  LZTW LZFSW LZFPW
 Capacaty  (mm)    60.0  30.0 200.0  45.0  45.0
 Initial content     .0    .0    .0    .0  10.0

 UZK   =      .3000 (1/DAY)     RSERV =      .2000 (-)
 LZSK  =      .0670 (1/DAY)     PCTIM =      .1000 (-)
 LZPK  =      .0140 (1/DAY)     ADIMP =      .1000 (-)
 ZPERC =    60.0000 (-)         SARVA =      .0000 (-)
 REXP  =     1.5000 (-)         SIDE  =      .0000 (-)
 PFREE =      .3000 (-)         SSOUT =      .0000 (mm/dt)

 Given unit hydrograph components:
     .150    .400    .300    .150

 Applied unit hydrograph components:
     .150    .400    .300    .150    .000

 Given Rainfall Intensity Components (PT1,PT2):
     .000    .000
```

Time step results

Year	Mo	Da	Ho	PRECIP	UZTWC	UZFWC	LZTWC	LZFSC	LZFPC	MDISCH	CDISCH	ACCDIFF
1994	6	16	0	.11	35.58	.00	1.27	.00	.31	-999.99	.72	.00
1994	6	17	0	.00	34.59	.00	1.26	.00	.30	-999.99	.19	.00
1994	6	18	0	.00	32.33	.00	1.26	.00	.30	.65	.09	.56
1994	6	19	0	.00	29.65	.00	1.24	.00	.29	.20	.00	.76
1994	6	20	0	.00	26.88	.00	1.23	.00	.29	-999.99	.00	.76
1994	6	21	0	.00	24.28	.00	1.22	.00	.29	-999.99	.00	.76
1994	6	22	0	.00	22.15	.00	1.20	.00	.28	-999.99	.00	.76
1994	6	23	0	8.25	29.47	.00	1.19	.00	.28	-999.99	.13	.76
1994	6	24	0	8.55	37.20	.00	1.19	.00	.27	-999.99	.46	.76
1994	6	25	0	27.80	60.00	4.60	1.19	.00	.27	-999.99	1.01	.76
1994	6	26	0	13.94	60.00	12.12	4.41	.69	.95	-999.99	1.70	.76
1994	6	27	0	3.12	60.00	.67	12.89	2.47	2.75	2.44	1.61	1.59
1994	6	28	0	3.22	60.00	.14	13.36	2.40	2.81	1.43	1.17	1.85
1994	6	29	0	91.97	60.00	30.00	13.46	2.27	2.80	.73	10.12	-7.55
1994	6	30	0	61.31	60.00	30.00	34.12	6.57	7.15	2.23	30.75	-36.07
1994	7	1	0	1.54	60.00	.89	54.89	10.64	11.45	29.42	33.01	-39.66
1994	7	2	0	1.88	60.00	.83	55.51	10.06	11.42	30.15	20.41	-29.91
1994	7	3	0	.57	57.49	.00	56.09	9.51	11.38	15.36	5.85	-20.40
1994	7	4	0	10.19	60.00	5.13	56.07	8.88	11.22	4.93	.91	-16.38
1994	7	5	0	80.43	60.00	30.00	59.66	9.08	11.80	3.55	8.98	-21.81
1994	7	6	0	20.56	60.00	20.10	80.49	13.14	15.90	20.30	22.39	-23.90
1994	7	7	0	12.47	60.00	9.74	94.56	15.44	18.53	37.48	18.12	-4.53
1994	7	8	0	6.53	60.00	3.80	101.38	15.97	19.63	29.02	10.48	14.01
1994	7	9	0	23.00	60.00	21.11	104.04	15.51	19.88	11.34	2.75	22.60
1994	7	10	0	22.04	60.00	19.80	118.82	17.93	22.48	7.84	3.29	27.15
1994	7	11	0	11.28	60.00	9.11	132.68	20.01	24.82	11.02	3.81	34.36
1994	7	12	0	2.38	60.00	.35	139.06	20.21	25.67	6.11	3.49	36.98
1994	7	13	0	4.53	60.00	2.43	139.30	18.91	25.35	2.58	2.54	37.03
1994	7	14	0	10.52	60.00	8.42	141.00	18.07	25.30	1.18	2.10	36.11
1994	7	15	0	2.08	60.00	.26	146.89	18.33	26.00	2.27	2.13	36.25
1994	7	16	0	3.33	60.00	1.65	147.07	17.15	25.67	2.89	1.99	37.15
1994	7	17	0	10.20	60.00	6.63	148.23	16.29	25.51	.72	1.90	35.98

1994	7	18	0	7.94	60.00	5.49	152.87	16.40	25.95	.31	2.08	34.21
1994	7	19	0	.84	59.09	.00	156.71	16.29	26.24	1.73	2.15	33.80
1994	7	20	0	22.50	60.00	19.80	156.70	15.20	25.87	4.56	2.30	36.05
1994	7	21	0	30.31	60.00	28.91	170.56	17.82	27.81	3.68	3.51	36.22
1994	7	22	0	18.47	60.00	17.49	190.79	21.98	30.74	4.86	4.90	36.18
1994	7	23	0	11.95	60.00	13.96	198.36	22.54	31.52	6.62	5.44	37.36
1994	7	24	0	8.74	60.00	11.43	200.00	24.30	32.98	8.27	5.20	40.43
1994	7	25	0	4.92	60.00	7.42	200.00	25.79	34.26	7.23	4.83	42.82
1994	7	26	0	54.55	60.00	30.00	200.00	25.95	34.79	5.18	8.88	39.12
1994	7	27	0	13.91	60.00	25.64	200.00	31.65	38.14	9.25	16.56	31.82
1994	7	28	0	1.79	60.00	13.19	200.00	34.44	39.95	8.47	15.36	24.92
1994	7	29	0	3.22	60.00	7.80	200.00	34.32	40.35	6.16	11.35	19.73
1994	7	30	0	6.37	60.00	8.86	200.00	33.34	40.31	4.33	6.21	17.84
1994	7	31	0	.00	57.83	4.68	200.00	32.68	40.34	2.79	5.01	15.63
1994	8	1	0	3.23	59.98	2.45	200.00	31.33	40.08	1.48	4.19	12.92
1994	8	2	0	63.02	60.00	30.00	200.00	29.72	39.69	.87	9.29	4.50
1994	8	3	0	18.01	60.00	30.00	200.00	34.09	41.29	4.82	19.79	-10.47
1994	8	4	0	1.00	59.37	15.74	200.00	36.86	42.36	10.00	19.23	-19.69
1994	8	5	0	2.83	60.00	9.45	200.00	36.77	42.49	10.23	14.19	-23.65
1994	8	6	0	.41	58.52	5.31	200.00	35.75	42.31	6.55	7.32	-24.42
1994	8	7	0	2.21	58.20	2.95	200.00	34.17	41.95	4.55	5.15	-25.02
1994	8	8	0	17.11	60.00	14.40	200.00	32.35	41.49	2.68	4.38	-26.72
1994	8	9	0	1.44	59.90	7.63	200.00	32.92	41.67	2.14	4.64	-29.23
1994	8	10	0	2.92	60.00	5.36	200.00	32.13	41.47	1.53	4.83	-32.53
1994	8	11	0	4.42	60.00	6.14	200.00	31.01	41.17	.77	4.52	-36.27
1994	8	12	0	5.58	60.00	6.88	200.00	30.17	40.94	.48	4.07	-39.86
1994	8	13	0	3.55	60.00	5.81	200.00	29.60	40.76	.41	3.93	-43.38
1994	8	14	0	6.26	60.00	8.51	200.00	28.87	40.54	.69	3.94	-46.63
1994	8	15	0	5.02	60.00	6.83	200.00	28.84	40.50	.80	4.03	-49.85
1994	8	16	0	2.10	60.00	4.54	200.00	28.43	40.36	1.05	4.02	-52.83
1994	8	17	0	11.23	60.00	11.03	200.00	27.56	40.09	.78	3.97	-56.01
1994	8	18	0	5.51	60.00	8.86	200.00	28.32	40.27	1.03	4.20	-59.18
1994	8	19	0	11.15	60.00	12.73	200.00	28.45	40.28	1.65	4.61	-62.14
1994	8	20	0	96.36	60.00	30.00	200.00	29.43	40.54	6.39	16.12	-71.87
1994	8	21	0	.96	59.14	14.67	200.00	33.79	41.79	22.37	35.03	-84.52

YEAR	MO	ND		PRECIP	E-POT	E-ACT	Runoff	Baseflow	Storage	UZTWC	UZFWC	LZTWC	LZFSC	LZFPC	ADIMC
1994	8	26	0	3.34	53.43	.81	200.00	28.63	40.50	3.72	3.03	-70.96			
1994	8	27	0	6.81	59.18	.40	200.00	26.82	39.97	1.82	2.69	-71.84			
1994	8	28	0	34.75	60.00	30.00	200.00	25.11	39.43	2.18	3.90	-73.56			
1994	8	29	0	.33	58.70	13.55	200.00	31.08	41.05	4.69	6.59	-75.46			
1994	8	30	0	.00	56.17	7.03	200.00	31.70	41.24	4.38	7.31	-78.39			
1994	8	31	0	8.26	60.00	6.14	200.00	30.88	41.03	3.63	6.29	-81.04			
1994	9	1	0	38.90	60.00	30.00	200.00	30.06	40.81	3.83	6.94	-84.15			
1994	9	2	0	22.35	60.00	30.00	200.00	34.24	41.98	5.99	11.83	-90.00			
1994	9	3	0	13.99	60.00	29.49	200.00	37.11	42.82	12.58	14.48	-91.90			
1994	9	4	0	.20	58.37	16.45	200.00	39.02	43.40	20.27	13.32	-84.95			
1994	9	5	0	2.52	58.84	9.51	200.00	38.65	43.37	19.47	9.96	-75.43			
1994	9	6	0	9.69	60.00	12.91	200.00	37.39	43.09	12.29	7.43	-70.57			
1994	9	7	0	258.01	60.00	30.00	200.00	36.83	42.97	20.33	42.06	-92.30			
1994	9	8	0	36.08	60.00	30.00	200.00	38.98	43.51	110.92	105.37	-86.75			
1994	9	9	0	3.08	60.00	19.71	200.00	40.51	43.91	73.86	89.00	-101.89			
1994	9	10	0	3.64	60.00	11.93	200.00	40.36	43.90	56.92	51.47	-96.43			
1994	9	11	0	1.90	60.00	7.23	200.00	39.22	43.65	43.44	11.39	-64.38			
1994	9	12	0	6.83	60.00	8.44	200.00	37.59	43.27	32.67	6.29	-38.01			
1994	9	13	0	4.67	60.00	6.90	200.00	36.34	42.96	20.85	5.30	-22.46			
1994	9	14	0	24.53	60.00	26.10	200.00	35.00	42.62	9.99	5.40	-17.87			
1994	9	15	0	10.62	60.00	23.24	200.00	37.07	43.09	6.51	6.84	-18.20			
1994	9	16	0	48.08	60.00	30.00	200.00	38.15	43.35	23.10	12.93	-8.03			
1994	9	17	0	5.60	60.00	21.31	200.00	39.87	43.77	19.58	21.36	-9.81			
1994	9	18	0	.00	57.13	12.45	200.00	40.05	43.83	23.90	18.16	-4.07			
1994	9	19	0	.00	52.06	7.27	200.00	38.88	43.57	19.49	12.33	3.09			
1994	9	20	0	.00	48.54	4.19	200.00	36.98	43.12	9.09	6.08	6.10			
1994	9	21	0	.00	44.35	2.36	199.82	34.70	42.57	4.67	4.55	6.22			
1994	9	22	0	.00	42.02	1.29	199.55	32.50	42.00	3.29	3.53	5.98			

etc.

Summary of values (in mm)

YEAR	MO	ND	PRECIP	E-POT	E-ACT	Runoff	Baseflow	Storage	UZTWC	UZFWC	LZTWC	LZFSC	LZFPC	ADIMC
1994														
	1	396	.00	115.50	14.07	.92	.92	26.11	.06	.00	27.14	.00	2.09	26.81
	2	424	.00	110.88	8.56	.54	.54	17.00	.01	.00	17.66	.00	1.41	17.42
	3	455	.00	169.68	7.67	.40	.40	8.93	.00	.00	9.13	.00	.91	9.00
	4	485	.00	197.82	4.41	.25	.25	4.27	.00	.00	4.22	.00	.59	4.16
	5	516	.00	216.02	2.16	.17	.17	1.94	.00	.00	1.82	.00	.38	1.79
	6	546	259.46	141.05	29.13	51.18	.62	181.10	60.00	30.00	34.12	6.57	7.15	136.21
	7	577	409.04	62.37	56.05	237.92	39.78	296.17	57.83	4.68	200.00	32.68	40.34	248.38
	8	608	324.44	61.64	55.20	265.62	66.07	299.79	60.00	6.14	200.00	30.88	41.03	255.94
	9	638	490.70	83.86	71.68	482.12	69.15	236.69	24.04	.00	189.27	18.78	37.56	209.61
	10	669	.00	128.10	76.79	23.91	23.91	135.98	2.63	.00	125.22	2.19	24.26	125.40
	11	699	.00	110.18	40.44	8.23	8.23	87.31	.39	.00	82.42	.27	15.89	81.20
	12	730	.00	85.47	21.04	4.70	4.70	61.57	.09	.00	59.29	.03	10.27	58.22

4 ANALYSIS OF RAINFALL DATA

4.1 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.
- The types of processing considered in this module are:
- checking data homogeneity
- computation of basic statistics
- annual exceedance rainfall series
- fitting of frequency distributions
- frequency and duration curves
- Most of the hydrological analysis for purpose of validation will be carried out at the Divisional and State Data Processing Centres and for the final presentation and reporting at the State Data Processing Centres.

Reference is made to Annex 1 for a review of statistics relevant for rainfall data.

4.2 CHECKING DATA HOMOGENEITY

For statistical analysis rainfall data from a single series should ideally possess property of homogeneity - i.e. properties or characteristics of different portion of the data series do not vary significantly.

Rainfall data for multiple series at neighbouring stations should ideally possess spatial homogeneity.

Tests of homogeneity is required for validation purposes and there is a shared need for such tests with other climatic variables. The following test have been described in Volume 8, Operational Manual Part II: Secondary Validation:

- Secondary validation of rainfall data
  a) Spatial homogeneity testing
  b) Consistency tests using double mass curves

- Correcting and completing rainfall data
  a) Adjusting rainfall data for long-term systematic shifts - double mass curves

- Secondary validation of climatic data
  a) Single series tests of homogeneity, including trend analysis, mass curves, residual mass curves, Student's t and Wilcoxon W-test on the difference of means and Wilcoxon-Mann-Whitney U-test to investigate if the sample are from same population.
  b) Multiple station validation including comparison plots, residual series, regression analysis and double mass curves.

4.3 COMPUTATION OF BASIC STATISTICS

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

- arithmetic mean
- 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 :

$$S_X = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1}} \qquad (4.1)$$

- skewness and kurtosis

In addition empirical frequency distributions can be presented as a graphical representation of the number of data per class and as a cumulative frequency distribution. From these selected values of exceedence probability or non-exceedence probability can be extracted, e.g. the daily rainfall which has been excceded 1%, 5% or 10% of the time.

Example 4.1

Basic statistics for monthly rainfall data of MEGHARAJ station (KHEDA catchment) is derived for the period 1961 to 1997. Analysis is carried out taking the actual values and all the months in the year. The results of analysis is given in Table 4.1 below. The frequency distribution and the cumulative frequency is worked out for 20 classes between 0 and 700 mm rainfall and is given in tabular results and as graph in Fig. 4.1. Various decile values are also listed in the result of the analysis.

Since actual monthly rainfall values are considered it is obvious to expect a large magnitude of skewness which is 2.34 and also the sample is far from being normal and that is reflected in kurtosis (=7.92). The value of mean is larger then the median value and the frequency distribution shows a positive skew. From the table of decile values it can may be seen that 70 % of the months receive less than 21 mm of rainfall. From the cumulative frequency table it may be seen that 65 percent of the months receive zero rainfall (which is obvious to expect in this catchment) and that there are very few instances when the monthly rainfall total is above 500 mm.

```
First year  =  1962                Last year   =  1997
Actual values are used

Basic Statistics:
      Mean              = .581438E+02
      Median            = .000000E+00
      Mode              = .175000E+02
      Standard deviation= .118932E+03
      Skewness          = .234385E+01
      Kurtosis          = .792612E+01
      Range             = .000000E+00 to .613500E+03
      Number of elements=         420
      Decile                Value
        1               .000000E+00
        2               .000000E+00
        3               .000000E+00
        4               .000000E+00
        5               .000000E+00
        6               .000000E+00
        7               .205100E+02
```

8 .932474E+02
9 .239789E+03

Cumulative frequency distribution and histogram

Upper class limit	Probability	Number of elements
.000000E+00	.658183	277.
.350000E+02	.729543	30.
.700000E+02	.769981	17.
.105000E+03	.815176	19.
.140000E+03	.836584	9.
.175000E+03	.867507	13.
.210000E+03	.881779	6.
.245000E+03	.903187	9.
.280000E+03	.917460	6.
.315000E+03	.934110	7.
.350000E+03	.950761	7.
.385000E+03	.955519	2.
.420000E+03	.967412	5.
.455000E+03	.981684	6.
.490000E+03	.986441	2.
.525000E+03	.988820	1.
.560000E+03	.995956	3.
.595000E+03	.995956	0.
.630000E+03	.998335	1.
.665000E+03	.998335	0.

Table 4.1: Computational results of the basic statistics for monthly rainfall at MEGHARAJ

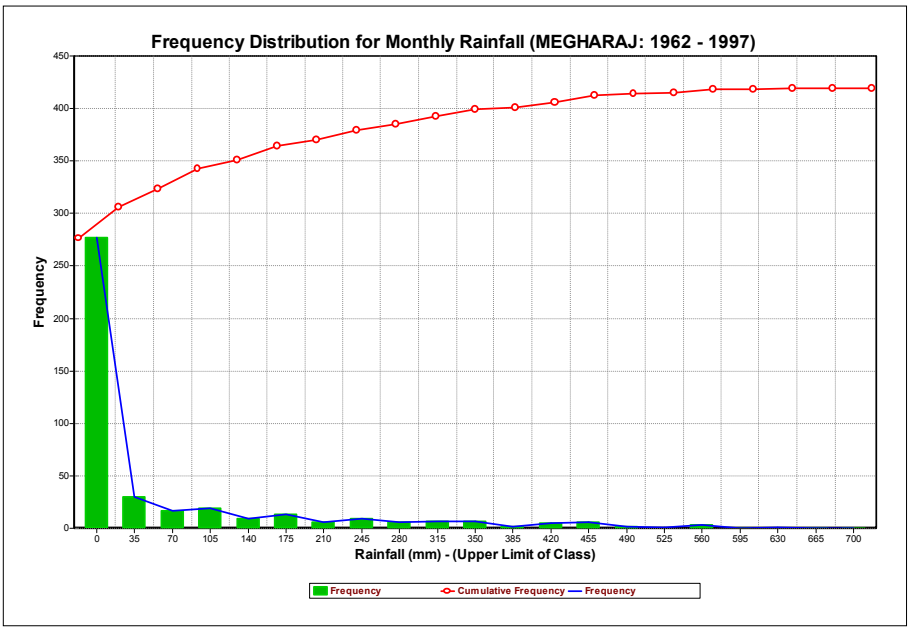


Figure. 4.1: Frequency and cumulative frequency plot of monthly rainfall at MEGHARAJ station

4.4 ANNUAL EXCEEDANCE RAINFALL SERIES

The following are widely used for reporting or for subsequent use in frequency analysis of extremes:

- maximum of a series. The maximum rainfall value of an annual series or of a month or season may be selected using HYMOS. All values (peaks) over a specified threshold may also be selected. Most commonly for rainfall daily maxima per year are used but hourly maxima or N-hourly maxima may also be selected.
- minimum of a series. As the minimum daily value with respect to rainfall is frequently zero this is useful for aggregated data only.

4.5 FITTING OF FREQUENCY DISTRIBUTIONS

A common use of rainfall data is in the assessment of probabilities or return periods of given rainfall at a given location. Such data can then be used in assessing flood discharges of given return period through modelling or some empirical system and can thus be applied in schemes of flood alleviation or forecasting and for the design of bridges and culverts.

Frequency analysis usually involves the fitting of a theoretical frequency distribution using a selected fitting method, although empirical graphical methods can also be applied. The fitting of a particular distribution implies that the rainfall sample of annual maxima were drawn from a population of that distribution. For the purposes of application in design it is assumed that future probabilities of exceedence will be the same as past probabilities. However there is nothing inherent in the series to indicate whether one distribution is more likely to be appropriate than another and a wide variety of distributions and fitting procedures has been recommended for application in different countries and by different agencies. Different distributions can give widely different estimates, especially when extrapolated or when an outlier (an exceptional value, well in excess of the second largest value) occurs in the data set. Although the methods are themselves objective, a degree of subjectivity is introduced in the selection of which distribution to apply.

These words of caution are intended to discourage the routine application and reporting of results of the following methods without giving due consideration to the regional climate. Graphical as well as numerical output should always be inspected. Higher the degree of aggregation of data, more normal the data will become.

The following frequency distributions are available in HYMOS:

- Normal and log-normal distributions
- Pearson Type III or Gamma distribution
- Log-Pearson Type III
- Extreme Value type I (Gumbel), II, or III
- Goodrich/Weibull distribution
- Exponential distribution
- Pareto distribution

The following fitting methods are available for fitting the distribution:

modified maximum likelihood

- method of moments

For each distribution one can obtain the following:

estimation of parameters of the distribution

a table of rainfalls of specified exceedance probabilities or return periods with confidence limits
results of goodness of fit tests

- a graphical plot of the data fitted to the distribution

Example 4.2

Normal frequency distribution for rainfall data of MEGHARAJ station (KHEDA catchment) is fitted for three cases: (a) annual maximum values of daily data series, (b) annual maximum values of monthly data series and (c) actual annual rainfall values.

Fig. 4.2 shows the graphical fitting of normal distribution for annual maximum values of daily series. The scatter points are the reduced variate of observed values and a best fit line showing the relationship between annual maximum daily with the frequency of occurrence on the basis of normal distribution. The upper and lower confidence limit (95 %) are also shown, the band width of which for different return periods indicate the level of confidence in estimation.

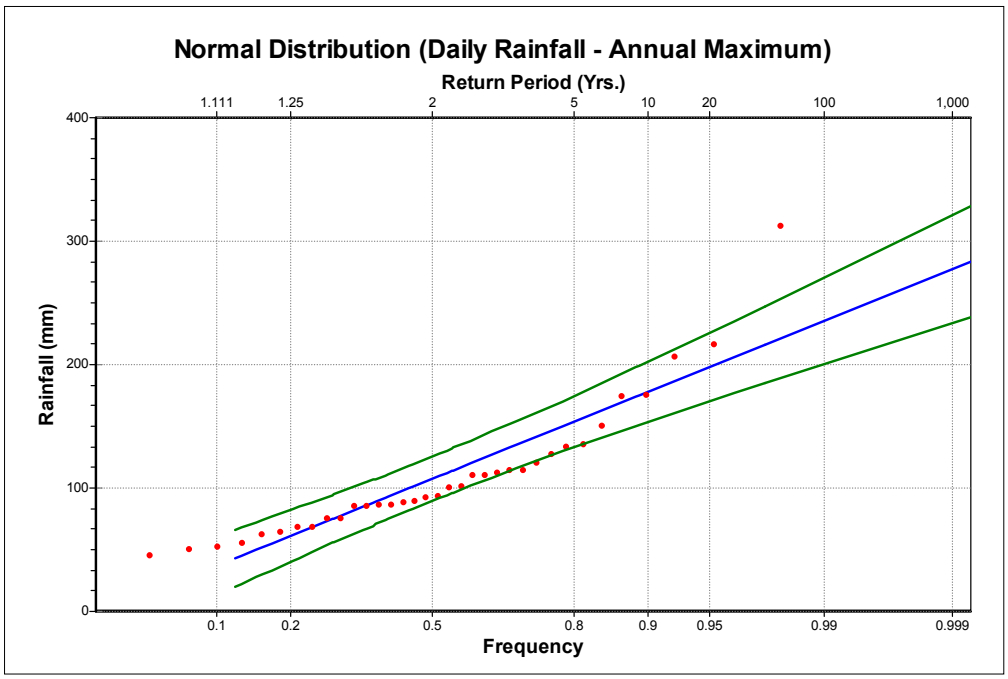


Figure 4.2: Normal distribution for annual maximums of daily rainfall at MEGHARAJ station

Figure 4.3 and Figure 4.4 shows the normal distribution fit for annual maximum of monthly data series and the actual annual values respectively. The level of normality can be seen to have increased in the case of monthly and yearly data series as compared to the case of daily data.

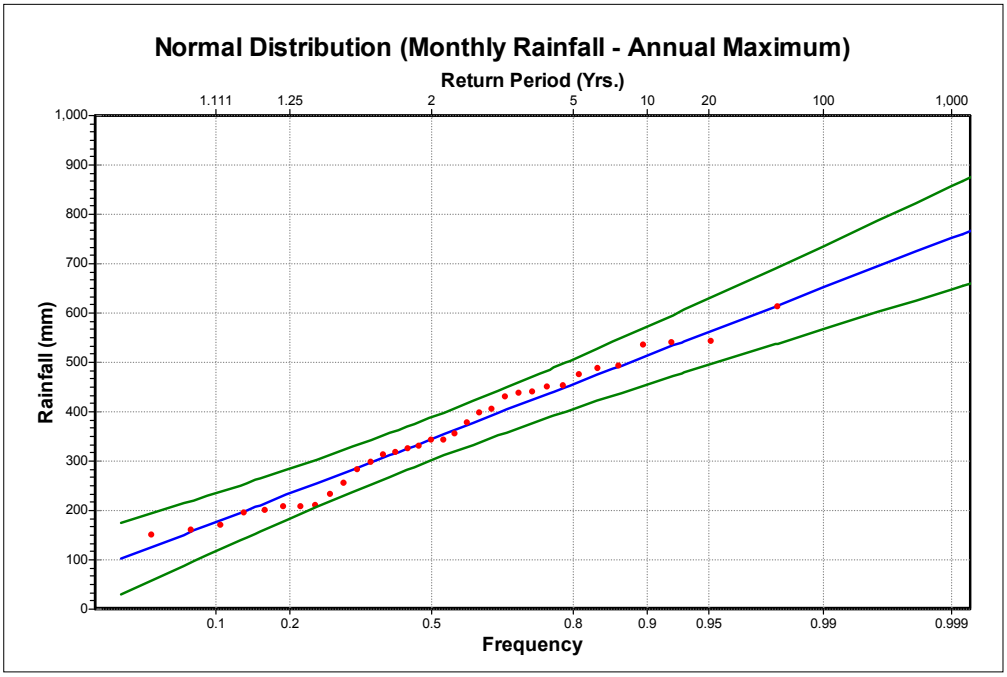


Figure 4.3: Normal distribution for annual maximums of monthly rainfall at MEGHARAJ station

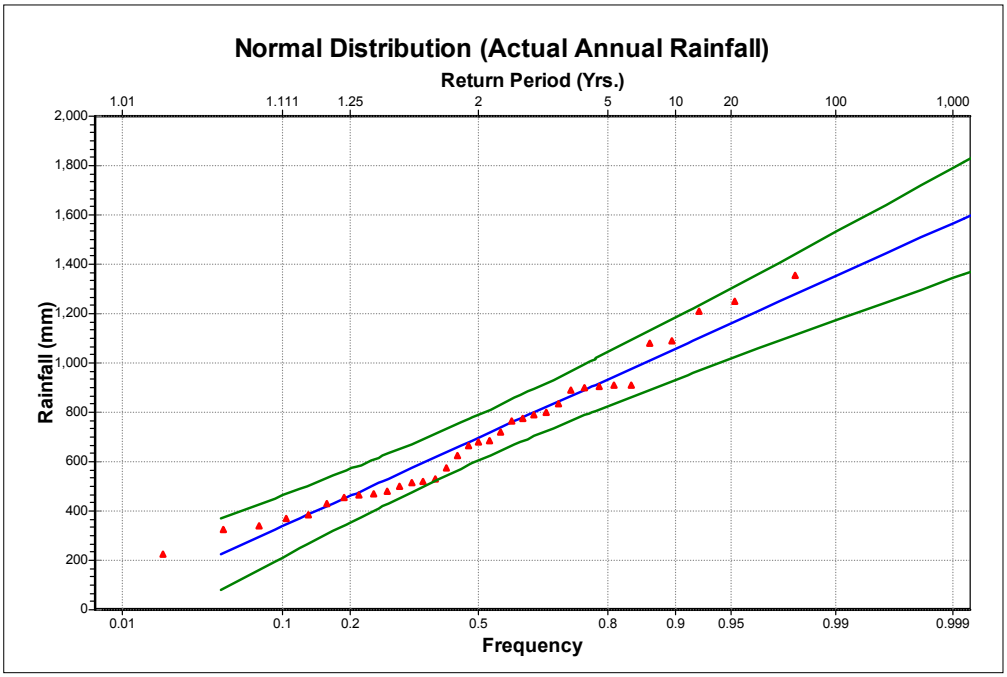


Figure 4.4: Plot of normal distribution for actual annual rainfall at MEGHARAJ station

The analysis results for the case of normal distribution fitting for actual annual rainfall is given in Table 5.1. There are 35 effective data values for the period 1962 to 1997 considered for the analysis. The mean annual rainfall is about 700 mm with an standard deviation of 280 mm and skewness and kurtosis of 0.54 and 2.724 respectively. The observed and theoretical frequency for each of the observed annual rainfall value is listed in the increasing order of the magnitude. The results of a few tests on good of fit is also given in table. In the last, the rainfall values for various return periods from 2 to 500 years is given alongwith the upper and lower confidence limits.

```
 First year  =  1962
 Last year   =  1997

 Basic statistics:
 Number of data     =    35
 Mean               =    697.725
 Standard deviation =    281.356
 Skewness           =       .540
 Kurtosis           =      2.724
```

year	observation	obs.freq.	theor.freq.p	return-per.	st.dev.xp	st.dev.p
13	225.700	.0198	.0467	1.05	73.7978	.0256
11	324.300	.0480	.0922	1.10	65.2262	.0383
18	338.000	.0763	.1005	1.11	64.1160	.0401
25	369.500	.1045	.1217	1.14	61.6525	.0443
5	383.300	.1328	.1319	1.15	60.6155	.0460
3	430.500	.1610	.1711	1.21	57.2863	.0517
17	456.000	.1893	.1951	1.24	55.6434	.0546
24	464.500	.2175	.2036	1.26	55.1224	.0554
23	472.500	.2458	.2117	1.27	54.6448	.0563
20	481.300	.2740	.2209	1.28	54.1341	.0571
8	500.000	.3023	.2411	1.32	53.1019	.0588
4	512.900	.3305	.2556	1.34	52.4337	.0599
2	521.380	.3588	.2654	1.36	52.0147	.0606
29	531.500	.3870	.2773	1.38	51.5363	.0614
7	573.800	.4153	.3298	1.49	49.8067	.0641
10	623.500	.4435	.3960	1.66	48.3757	.0663
30	665.500	.4718	.4544	1.83	47.7129	.0672
31	681.000	.5000	.4763	1.91	47.5996	.0674
27	686.000	.5282	.4834	1.94	47.5784	.0674
1	719.100	.5565	.5303	2.13	47.6261	.0673
34	763.500	.5847	.5924	2.45	48.2011	.0665
33	773.000	.6130	.6055	2.53	48.3988	.0662
22	788.000	.6412	.6258	2.67	48.7632	.0657
19	799.000	.6695	.6406	2.78	49.0705	.0652
15	833.800	.6977	.6857	3.18	50.2575	.0634
21	892.000	.7260	.7551	4.08	52.9196	.0591
6	900.200	.7542	.7641	4.24	53.3571	.0584
26	904.000	.7825	.7683	4.32	53.5647	.0581
28	911.500	.8107	.7763	4.47	53.9833	.0574
16	912.000	.8390	.7768	4.48	54.0117	.0573
12	1081.300	.8672	.9136	11.58	66.0629	.0370
32	1089.500	.8955	.9181	12.21	66.7472	.0359
14	1210.300	.9237	.9658	29.20	77.5678	.0209
9	1248.000	.9520	.9748	39.61	81.1752	.0170
35	1354.000	.9802	.9902	101.67	91.7435	.0086

```
Results of Binomial goodness of fit test
variate dn = max(|Fobs-Fest|)/sd =     1.4494 at Fest= .2773
prob. of exceedance P(DN>dn)     =      .1472
number of observations           =    35

Results of Kolmogorov-Smirnov test
variate dn = max(|Fobs-Fest|)    =      .1227
prob. of exceedance P(DN>dn)     =      .6681

Results of Chi-Square test
variate = chi-square             =     5.2000
prob. of exceedance of variate   =      .2674
number of classes                =     7
number of observations           =    35
degrees of freedom               =     4
```

Values for distinct return periods

Return per.	prob(xi<x) p	value x	st. dev. x	confidence lower	intervals upper
2	.50000	697.725	47.558	604.493	790.957
5	.80000	934.474	55.340	825.987	1042.961
10	.90000	1058.347	64.184	932.521	1184.173
25	.96000	1190.401	75.692	1042.014	1338.788
50	.98000	1275.684	83.867	1111.271	1440.096
100	.99000	1352.380	91.565	1172.876	1531.885
250	.99600	1444.011	101.085	1245.846	1642.177
500	.99800	1507.611	107.852	1296.180	1719.043

Table 4.2: Analysis results for the normal distribution fitting for annual rainfall at MEGHARAJ

4.6 FREQUENCY AND DURATION CURVES

A convenient way to show the variation of hydrological quantities through the year, by means of frequency curves, where each frequency curve indicates the magnitude of the quantity for a specific probability of non-exceedance. The duration curves are ranked representation of these frequency curves. The average duration curve gives the average number of occasions a given value was not exceeded in the years considered. The computation of frequency and duration curves is as given below:

4.6.1 FREQUENCY CURVES

Considering "n" elements of rainfall values in each year (or month or day) and that the analysis is carried out for "m" years (or months or days) a matrix of data X_{ij} {for i=1,m and j=1,n} is obtained. For each $j = j_0$ the data $X_{I,jo}$, {for i=1,m} is arranged in ascending order of magnitude. The probability that the i^{th} element of this ranked sequence of elements is not exceeded is:

$$F_i = \frac{i}{m+1} \tag{4.2}$$

The frequency curve connects all values of the quantity for j=1,n with the common property of equal probability of non-exceedance. Generally, a group of curves is considered which represents specific points of the cumulative frequency distribution for each j. Considering that curves are derived for various frequencies F_k {k=1,n_f}, then values for rainfall $R_{k,j}$ is obtained by linear interpolation between the probability values immediately greater (F_I) and lesser (F_{i-1}) to n_k for each j as:

$$R_{k,jo} = R_{1-1} + (R_i - R_{i-1}) \frac{F_k - F_{i-1}}{F_i - F_{i-1}} \tag{4.3}$$

4.6.2 DURATION CURVES

When the data $R_{k,j}$, k=1,n_f and j=1,n is ranked for each k, the ranked matrix represents the duration curves for given probabilities of non-exceedance.

When all the data is considered without discriminating for different elements j {j=1,n} and are ranked in the ascending order of magnitude, then the resulting sequence shows the average duration curve. This indicates how often a given level of quantity considered will not be exceeded in a year (or month or day).

Example 4.3

A long-term monthly rainfall data series of MEGHARAJ station (KHEDA catchment) is considered for deriving frequency curves and duration curves. Analysis is done on the yearly basis and the various frequency levels set are 10, 25, 50, 75 and 90 %.

Figure 4.5 shows the frequency curves for various values (10, 25, 50, 75 and 90%) for each month in the year. Monthly rainfall distribution in the year 1982 is also shown superimposed on this plot for comparison. Minimum and maximum values for each month of the year in the plot gives the range of variation of rainfall in each month. Results of this frequency curve analysis is tabulated in Table 4.3.

Figure 4.6 shows the plot of duration curves for the same frequencies. The plot gives values of monthly rainfall which will not be exceeded for certain number of months in a year with the specific level of probability. The results of analysis for these duration curve is given in Table 4.4.

The average duration curve, showing value of rainfall which will not be exceeded "on an average" in a year for a certain number of months is given as Figure 4.7.

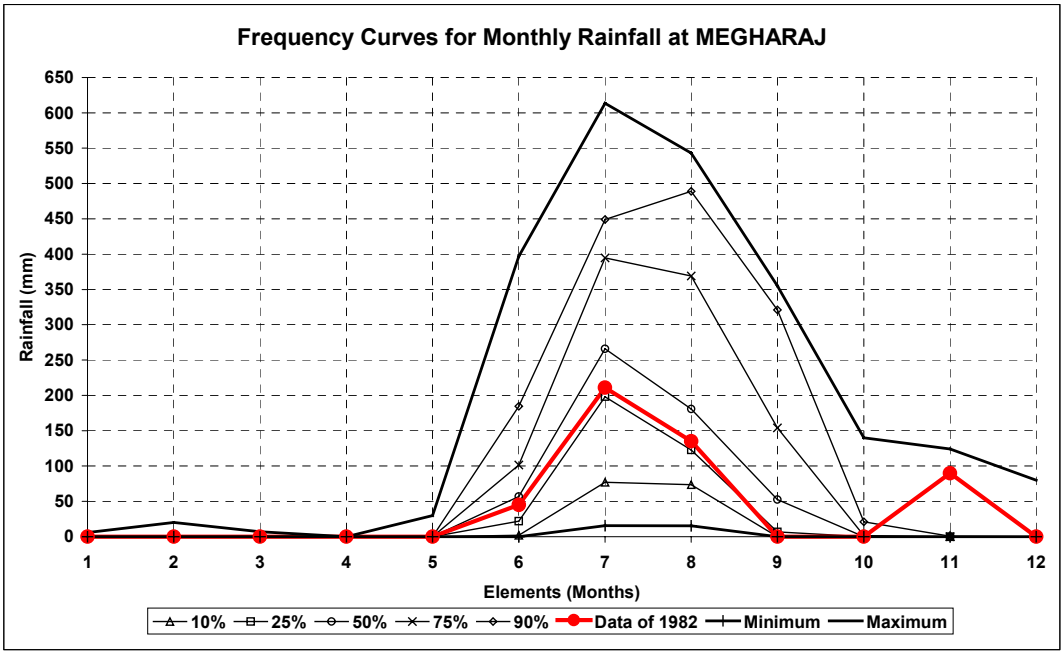


Figure 4.5: Monthly frequency curves for rainfall at MEGHRAJ station

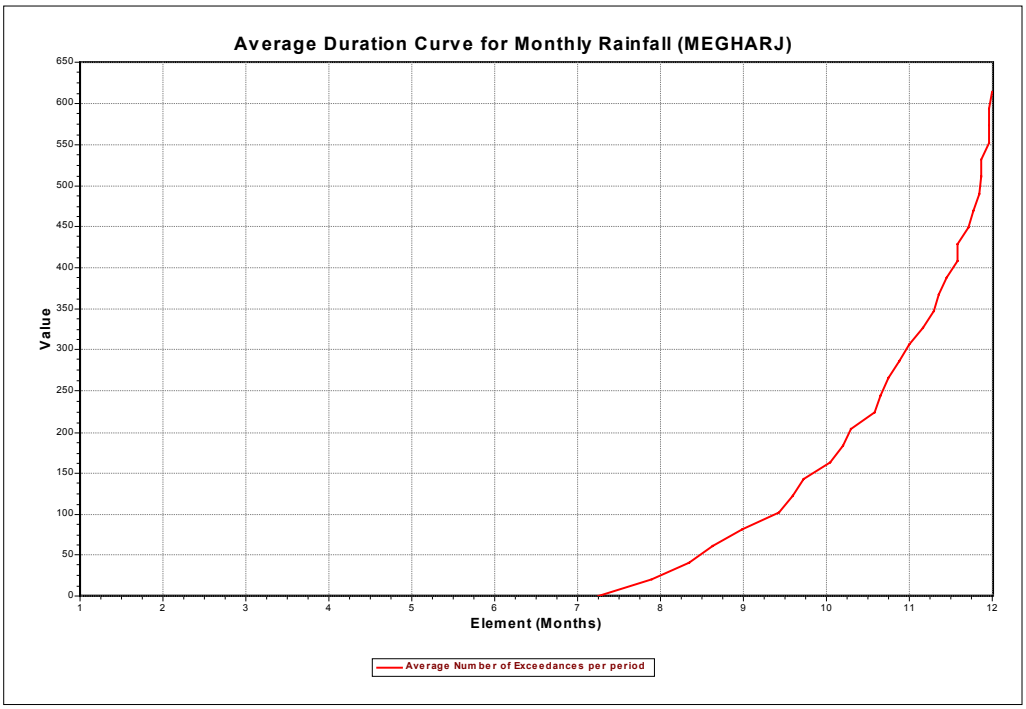


Figure 4.6: Monthly duration curves for rainfall at MEGHARAJ station

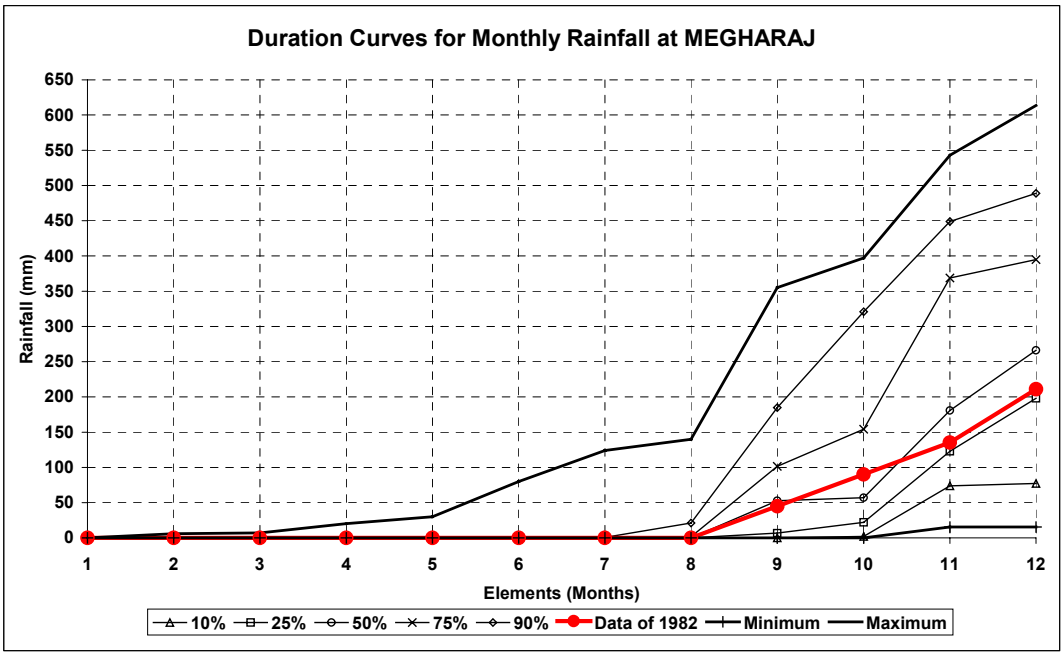


Figure 4.7: Average monthly duration curves for rainfall at MEGHARAJ station

Element	No. of Data	Frequency					Year 1982	Min.	Max.
		0.01	0.25	0.05	0.75	0.09			
1	29	0	0	0	0	0	0	0	6
2	29	0	0	0	0	0	0	0	20
3	29	0	0	0	0	0	0	0	7
4	29	0	0	0	0	0	0	0	0
5	29	0	0	0	0	0	0	0	30
6	33	2	22	57	101.54	184.8	45	0	397
7	36	77.01	198	266	394.75	448.93	211	15.05	613.05
8	36	73.64	122.77	190.75	368.87	488.92	135.3	15.03	543
9	36	0	6.63	52.5	154	320.09	0	0	355.09
10	30	0	0	0	1.37	21.05	0	0	140
11	29	0	0	0	0	0.08	90	0	124
12	29	0	0	0	0	0	0	0	80

Table 4.3: Results of analysis for frequency curves for monthly data for MEGHARAJ station (rainfall values in mm)

No. of Elements	Frequency					Year 1982	Min.	Max.
	0.1	0.25	0.5	0.75	0.9			
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	6
3	0	0	0	0	0	0	0	7
4	0	0	0	0	0	0	0	20
5	0	0	0	0	0	0	0	30
6	0	0	0	0	0	0	0	80
7	0	0	0	0	0.8	0	0	124
8	0	0	0	1.37	21.05	0	0	140
9	0	6.63	52.05	101.54	184.08	45	0	355.09
10	2	22	57	154	320.09	90	0	397
11	73.64	122.77	190.75	368.87	448.93	135.3	15.03	543
12	77.01	198	266	394.75	488.92	211	15.05	613.5

Table 4.4: Results of analysis for duration curves for monthly data for MEGHARAJ station (rainfall values in mm)

1	Rainfall Value	0	20.45	40.9	61.35	81.08	102.25	
	No. of Exceedances	7.25	7.89	8.34	8.63	8.98	9.43	
2	Rainfall Value	122.07	143.15	163.6	184.05	204.05	224.95	
	No. of Exceedances	9.59	9.72	10.04	10.02	10.03	10.59	
3	Rainfall Value	245.04	265.85	286.03	306.75	327.02	347.65	
	No. of Exceedances	10.65	10.75	10.88	11.01	11.17	11.29	
4	Rainfall Value	368.01	388.55	409	429.45	449.09	470.35	
	No. of Exceedances	11.36	11.45	11.58	11.58	11.71	11.78	
5	Rainfall Value	490.08	511.25	531.07	552.15	572.06	593.05	613.05
	No. of Exceedances	11.84	11.87	11.87	11.97	11.97	11.97	12

Table 4.5: Results of analysis for average duration curves for monthly data for MEGHARAJ station (rainfall values in mm)

4.7 INTENSITY-FREQUENCY-DURATION ANALYSIS

If rainfall data from a recording raingauge is available for long periods such as 25 years or more, the frequency of occurrence of a given intensity can also be determined. Then we obtain the intensity-frequency-duration relationships. Such relationships may be established for different parts of the year, e.g. a month, a season or the full year. The procedure to obtain such relationships for the year is described in this section. The method for parts of the year is similar.

The entire rainfall record in a year is analysed to find the maximum intensities for various durations. Thus each storm gives one value of maximum intensity for a given duration. The largest of all such

values is taken to be the maximum intensity in that year for that duration. Likewise the annual maximum intensity is obtained for different duration. Similar analysis yields the annual maximum intensities for various durations in different years. It will then be observed that the annual maximum intensity for any given duration is not the same every year but it varies from year to year. In other words it behaves as a random variable. So, if 25 years of record is available then there will be 25 values of the maximum intensity of any given duration, which constitute a sample of the random variable. These 25 values of any one duration can be subjected to a frequency analysis. Often the observed frequency distribution is well fitted by a Gumbel distribution. A fit to a theoretical distribution function like the Gumbel distribution is required if maximum intensities at return periods larger than can be obtained from the observed distribution are at stake. Similar frequency analysis is carried out for other durations. Then from the results of this analysis graphs of maximum rainfall intensity against the return period for various durations such as those shown in Figure 4.8 can be developed.

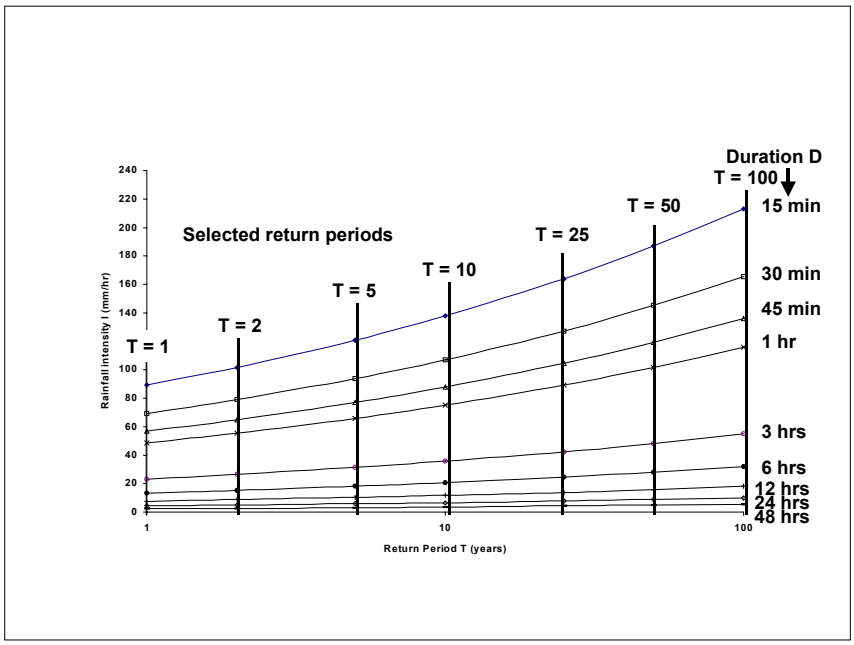


Figure 4.8: Intensity-frequency-duration curves

By reading for each duration at distinct return periods the intensities intensity-duration curves can be made. For this the rainfall intensities for various durations at concurrent return periods are connected as shown in Figure 4.9.

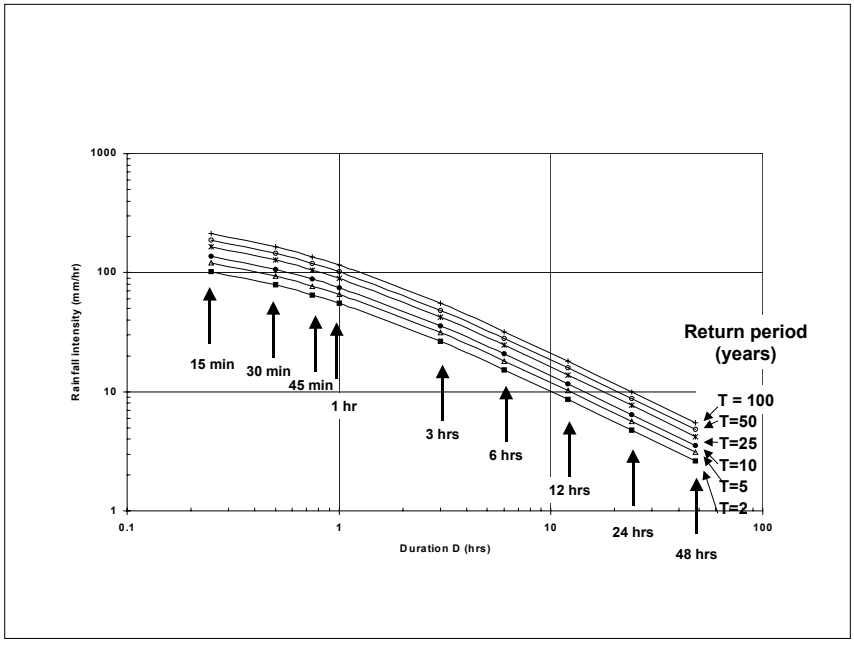


Figure 4.9: Intensity-frequency-duration curves for various return periods

From the curves of Figure 4.9 the maximum intensity of rainfall for any duration and for any return period can be read out.

Alternatively, for any given return period an equation of the form.

$$I = \frac{c}{(D+a)^b} \tag{4.4}$$

can be fitted between the maximum intensity and duration

where I = intensity of rainfall (mm/hr)

D = duration (hrs)

c, a, b are coefficients to be determined through regression analysis.

One can write for return periods T_1, T_2, etc.:

$$I = \frac{c_1}{(D+a_1)^{b_1}}; I = \frac{c_2}{(D+a_2)^{b_2}}; \text{etc} \tag{4.5}$$

where c_1, a_1 and b_1 refer to return period $T_1$and c_2, a_2 and b_2 are applicable for return period T_2, etc. Generally, it will be observed that the coefficients a and b are approximately the same for all the return periods and only c is different for different return periods. In such a case one general equation may be developed for all the return periods as given by:

$$I = \frac{KT^d}{(D+a)^b} \tag{4.6}$$

where T is the return period in years and K and d are the regression coefficients for a given location. If a and b are not same for all the return periods, then an individual equation for each return period may be used. In Figure 4.8 and 4.9 the results are given for Bhopal, as adapted from Subramanya, 1994. For Bhopal with I in mm/hr and D in hours the following parameter values in equation 4.6 hold:

K = 69.3; a = 0.50; b = 0.878 and d = 0.189

When the intensity-frequency-duration analysis is carried out for a number of locations in a region, the relationships may be given in the form of equation 4.6 with a different set of regression coefficients for each location. Alternatively, they may be presented in the form of maps (with each map depicting maximum rainfall depths for different combinations of one return period and one duration) which can be more conveniently used especially when one is dealing with large areas. Such maps are called isopluvial maps. A map showing maximum rainfall depths for the duration of one hour which can be expected with a frequency of once in 50 years over South India is given in Figure 4.10.

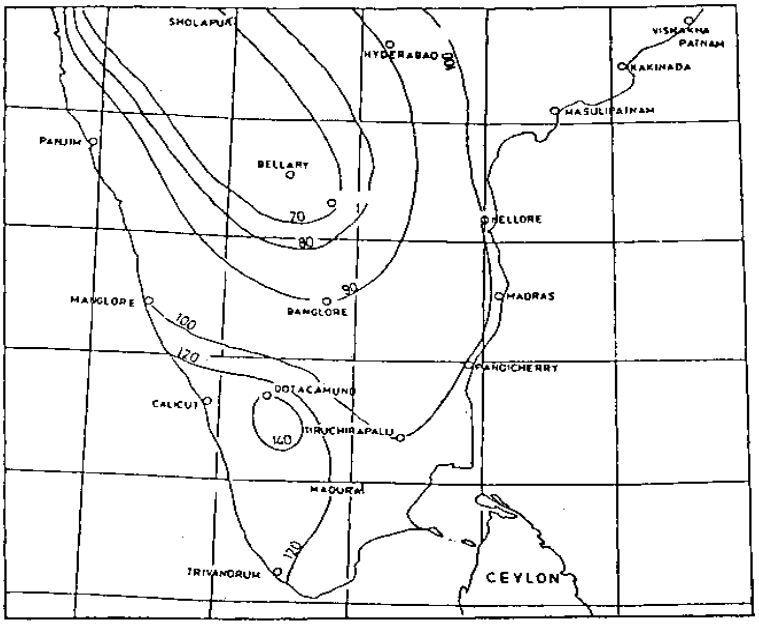


Figure 4.10:
Isopluvial map of 50 years 1 hour rainfall over South India

Annual maximum and annual exceedance series

In the procedure presented above annual maximum series of rainfall intensities were considered. For frequency analysis a distinction is to be made between annual maximum and annual exceedance series. The latter is derived from a partial duration series, which is defined as a series of data above a threshold. The maximum values between each upcrossing and the next downcrossing (see Figure 4.11) are considered in a partial duration series. The threshold should be taken high enough to make successive maximums serially independent or a time horizon is to be considered around the local maximum to eliminate lower maximums exceeding the threshold but which are within the time horizon. If the threshold is taken such that the number of values in the partial duration series becomes equal to the number of years selected then the partial duration series is called annual exceedance series.

Since annual maximum series consider only the maximum value each year, it may happen that the annual maximum in a year is less than the second or even third largest independent maximum in another year. Hence, the values at the lower end of the annual exceedance series will be higher than those of the annual maximum series. Consequently, the return period derived for a particular I(D) based on annual maximum series will be larger than one would have obtained from annual exceedances. The following relation exists between the return period based on annual maximum and annual exceedance series (Annexure II. Equation (4.158)):

$$T_E = \frac{1}{\ln\left(\frac{T}{T-1}\right)} \qquad (4.7)$$

where: T_E = return period for annual exceedance series

T = return period for annual maximum series

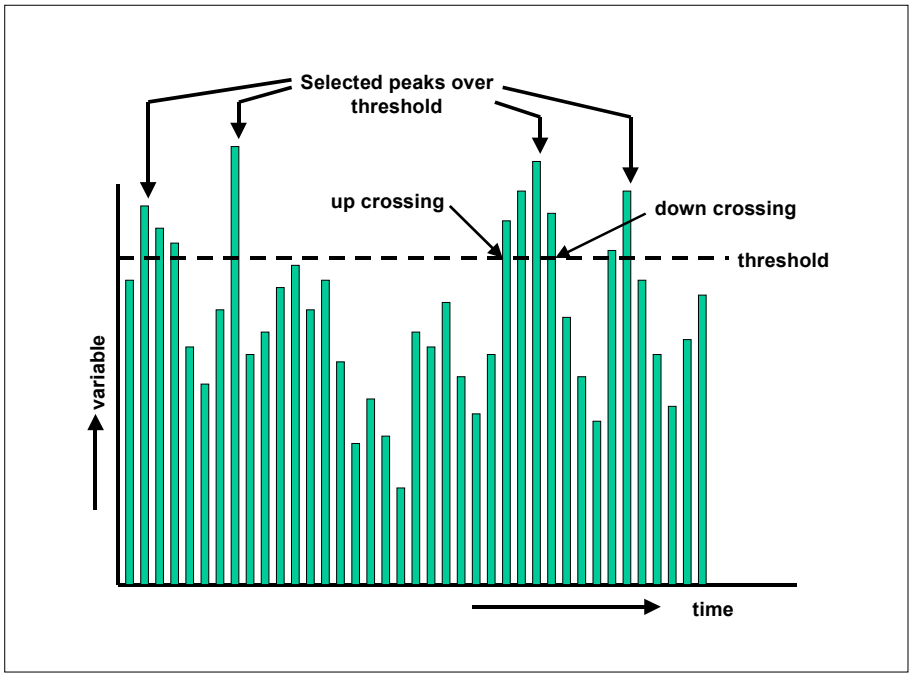


Figure 4.11: Definition of partial duration series

The ratio T_E/T is shown in Figure 4.12. It is observed that the ratio approaches 1 for large T. Generally, when T < 20 years T has to be adjusted to T_E for design purposes. Particularly for urban drainage design, where low return periods are used, this correction is of importance.

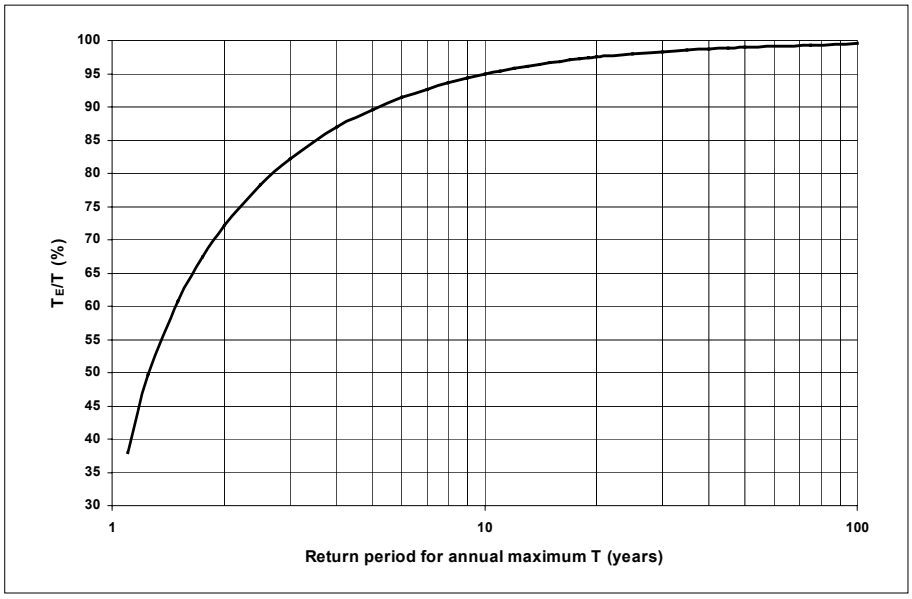


Figure 4.12: Relation between return periods annual maximum (T) and annual exceedance series (T_E).

In HYMOS annual maximum series are used in the development of intensity-duration-frequency curves, which are fitted by a Gumbel distribution. Equation (4.7) is used to transform T into T_E for T < 20 years. Results can either be presented for distinct values of T or of T_E.

Example 4.4

Analysis of hourly rainfall data of station Chaskman, period 1977-2000, monsoon season 1/6-30/9. First, from the hourly series the maximum seasonal rainfall intensities for each year are computed for rainfall durations of 1, 2, …, 48 hrs. In this way annual maximum rainfall intensity series are obtained for different rainfall durations. Next, each such series is subjected to frequency analysis using the Gumbel or EV1 distribution, as shown for single series in Figure 4.13. The IDF option in HYMOS automatically carries out this frequency analysis for all rainfall durations. The results are presented in Table 4.6. The fit to the distribution for different rainfall durations is shown in Figure 4.14. It is observed that in general the Gumbel distribution provides an acceptable fit to the observed frequency distribution.

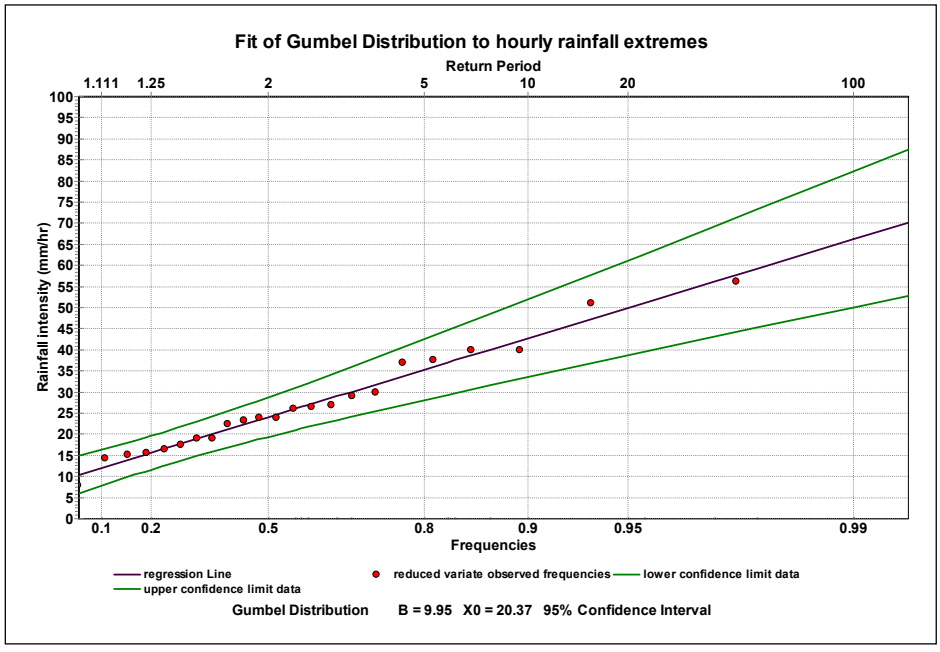


Figure 4.13: Fitting of Gumbel distribution to observed frequency distribution of hourly annual maximum series for monsoon season

```
Intensity - Duration - Frequency Relations
Input Timestep: 1 hour
Duration computation in: hour
Year: 1998 contains missing values, analysis may not be correct
Year: 1999 contains missing values, analysis may not be correct
Year: 2000 contains missing values, analysis may not be correct

Period from '01-06' to '01-10'
Start year: 1977
End year  : 2000

Maximum Intensities per year for selected durations

                              Durations (hour)
Year      1       2       3       4       6       9       12      18      24      48

1977    14.40   9.70    7.60    6.55    4.37    2.91    2.18    1.68    1.56    1.03
1978    19.00   14.45   10.03   7.65    5.12    3.41    2.56    1.71    1.66    1.27
1979    23.30   19.40   14.50   10.95   7.30    5.00    3.78    2.81    2.37    1.86
.
1997    37.00   29.25   27.50   25.50   23.08   18.06   15.71   11.97   9.25    5.41
1998    22.40   15.00   10.00   8.88    6.17    4.46    3.50    2.56    2.38    1.47
1999    30.00   22.60   16.17   12.38   8.30    5.53    4.15    2.77    2.07    1.08
2000    24.00   12.30   8.53    6.70    4.62    4.38    3.28    2.19    2.29    1.48

Parameters of Gumbel distribution

Duration   X0        BETA    Sd1     Sd2

1                 20.372     9.955  2.140  1.584
2                 13.623     7.250  1.558  1.154
3                 10.322     5.745  1.235  0.914
4                 8.316      4.703  1.011  0.748
6                 6.308      3.567  0.767  0.568
9                 4.634      2.673  0.574  0.425
12                3.605      2.166  0.465  0.345
18                2.541      1.520  0.327  0.242
24                2.192      1.211  0.260  0.193
48                1.329      0.701  0.151  0.112

              IDF-data: Annual Maximum

Duration                              Return Periods
                    1        2        4        10        25      50      100

1                 11.666   24.020   32.774   42.773    52.212  59.214  66.164
2                 7.282    16.280   22.656   29.938    36.813  41.912  46.975
3                 5.297    12.428   17.480   23.251    28.698  32.740  36.751
4                 4.203    10.040   14.175   18.899    23.358  26.666  29.949
6                 3.188     7.615   10.752   14.334    17.716  20.225  22.716
9                 2.296     5.614    7.964   10.648    13.183  15.063  16.929
12                1.711     4.399    6.303    8.479    10.532  12.056  13.568
18                1.212     3.098    4.435    5.961     7.403   8.472   9.533
24                1.133     2.636    3.700    4.916     6.064   6.916   7.761
48                0.716     1.586    2.202    2.906     3.571   4.064   4.553
```

Table 4.6: Example of output file of IDF option

Note that in the output table first a warning is given about series being incomplete for some years. This may affect the annual maximum series. Comparison with nearby stations will then be required to see whether extremes may have been missed. If so, the years with significant missing data are eliminated from the analysis.

Next, the table presents an overview of the annual maximum series, followed by a summary of the Gumbel distribution parameters x_0 and β, with their standard deviations (sd1, sd2) and for various rainfall durations the rainfall intensities for selected return periods. The latter values should be compared with the maximum values in the annual maximum series.

Note that Figure 4.14 gives a row-wise presentation of the last table, whereas Figure 4.15 gives a column-wise presentation of the same table. This figure is often presented on log-log scale, see Figure 4.16.

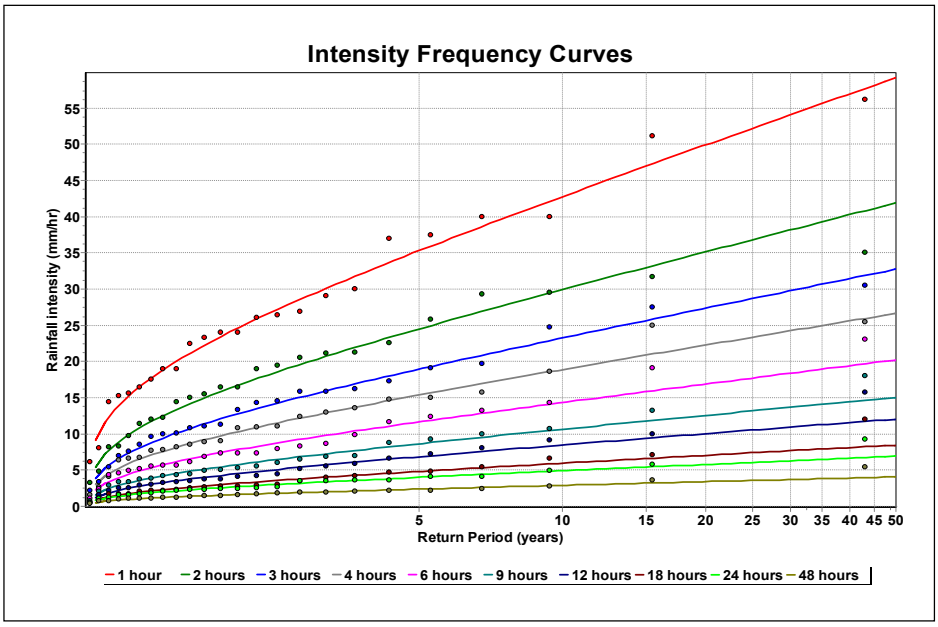


Figure 4.14: Intensity Frequency curves for different rainfall durations, with fit to Gumbel distribution

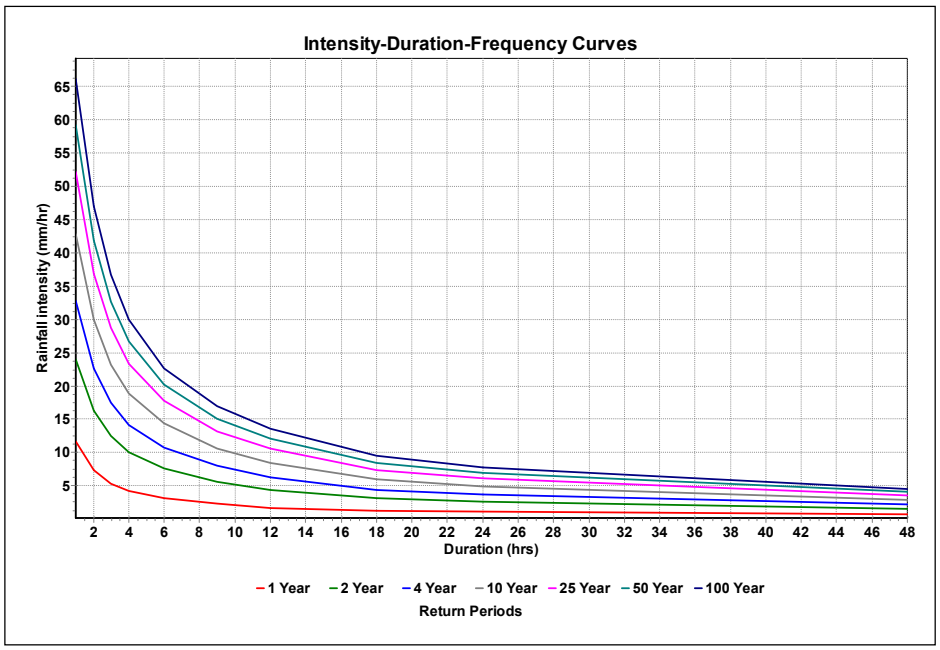


Figure 4.15: Intensity-Density-Frequency curves for Chaskman on linear scale (Annual maximum data)

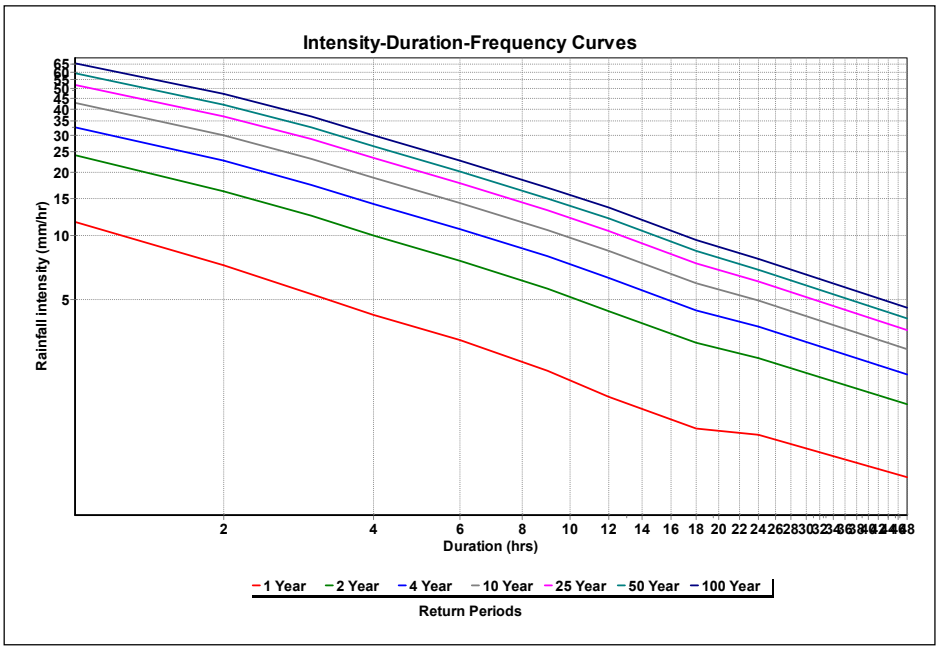


Figure 4.16: Intensity-Density-Frequency curves for Chaskman on double-log scale (Annual maximum data)

The IDF-option in HYMOS also includes a procedure to convert the annual maximum statistics into annual exceedance results by adapting the return period according to equation (4.7). Hence, rather than using the annual exceedance series in a frequency analysis the annual maximum series' result is adapted. This procedure is useful for design of structures where design conditions are based on events with a moderate return period (5 to 20 years). An example output is shown in Table 4.7 and Figures 4.16 and 4.17. Compare results with Table 4.6 and Figures 4.14 and 4.15.

```
IDF  - data : Annual Exceedences

         Te       Tm
         1        1.581977
         2        2.541494
         4        4.520812
         10       10.50833
         25       25.50333
         50       50.50167
         100      100.5008
```

Duration	Return Periods						
	1	2	4	10	25	50	100
1	20.372	27.272	34.172	43.293	52.414	59.314	66.214
2	13.623	18.648	23.674	30.317	36.960	41.986	47.011
3	10.322	14.304	18.286	23.551	28.815	32.797	36.780
4	8.316	11.576	14.836	19.145	23.454	26.713	29.973
6	6.308	8.780	11.252	14.521	17.789	20.261	22.733
9	4.634	6.487	8.339	10.788	13.237	15.089	16.942
12	3.605	5.106	6.607	8.592	10.576	12.078	13.579
18	2.541	3.595	4.648	6.041	7.433	8.487	9.540
24	2.192	3.031	3.870	4.980	6.089	6.928	7.767
48	1.329	1.815	2.301	2.943	3.585	4.071	4.557

Table 4.7: Example of output of IDF curves for annual exceedances.

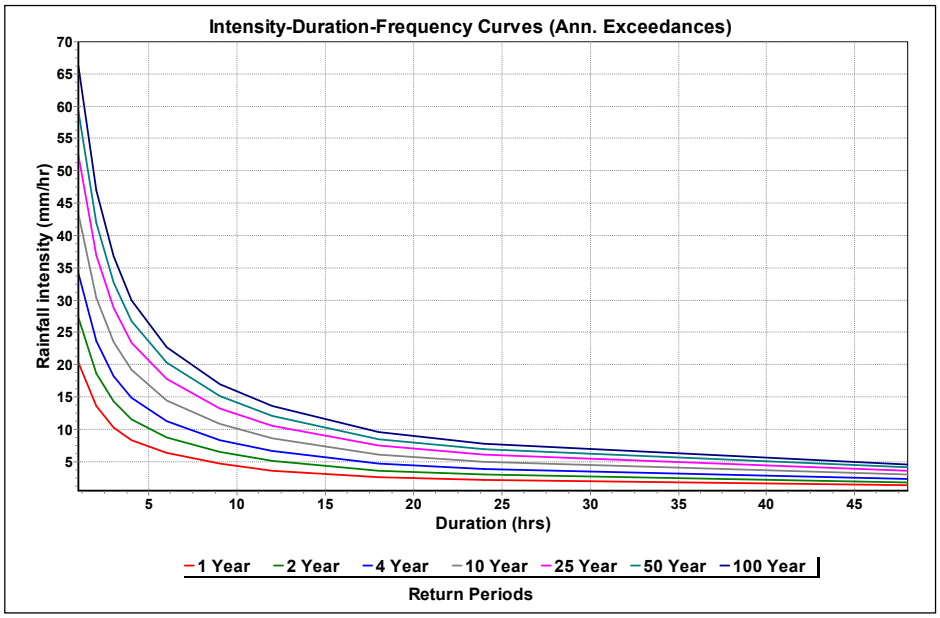


Figure 4.17: Intensity-Density-Frequency curves for Chaskman on linear scale (Annual exceedances)

Finally, the Rainfall Intensity-Duration curves for various return periods have been fitted by a function of the type (4.4). It appeared that the optimal values for "a" and "b" varied little for different return periods. Hence a function of the type (4.6) was tried. Given a value for "a" the coefficients K, d and b can be estimated by multiple regression on the logarithmic transformation of equation (4.6):

$$\log I = \log K + d \log T - b \log(D + a) \qquad (4.12)$$

By repeating the regression analysis for different values of "a" the coefficient of determination was maximised. The following equation gave a best fit (to the logaritms):

$$I = \frac{32.8T^{0.27}}{(D + 0.65)^{0.81}} \qquad R^2 = 0.993$$

Though the coefficient of determination is high, a check afterwards is always to be performed before using such a relationship!! A comparison is shown in Figure 4.19. A reasonable fit is observed.

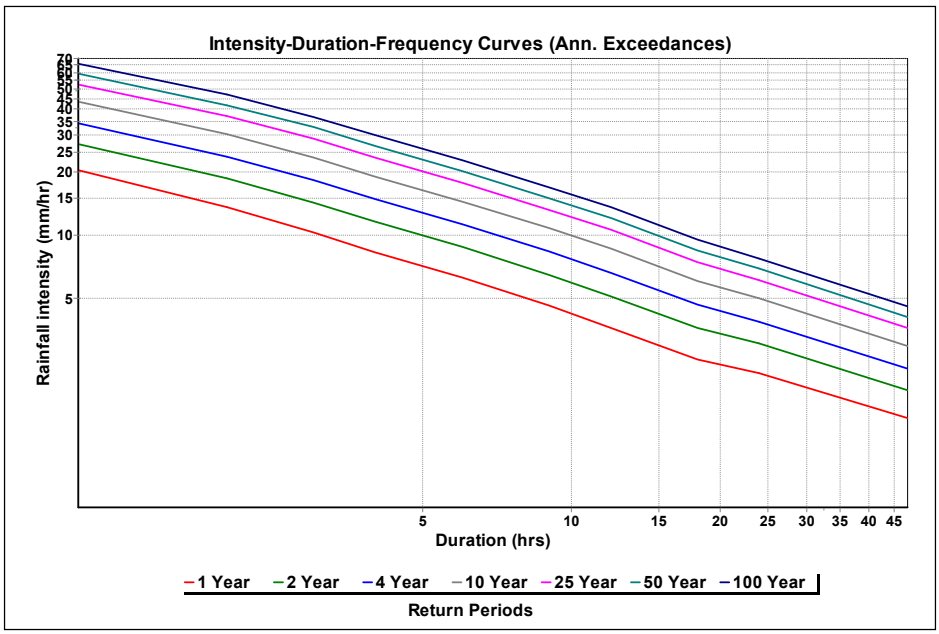


Figure 4.18: Intensity-Density-Frequency curves for Chaskman on double-log scale (Annual exceedances)

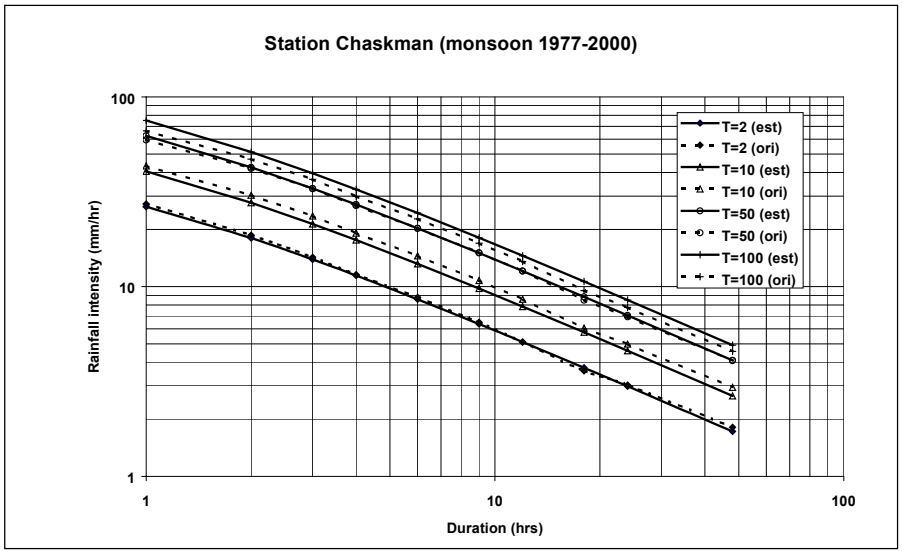


Figure 4.19: Test of goodness of fit of IDF-formula to IDF-curves from Figure 4.17 and 4.18

4.8 DEPTH-AREA-DURATION ANALYSIS

In most of the design applications the maximum depth of rainfall that is likely to occur over a given area for a given duration is required. Wherever possible, the frequency of that rainfall should also be known. For example, the knowledge of maximum depth of rainfall occurring on areas of various sizes for storms of different duration is of interest in many hydrological design problems such as the design of bridges and culverts, design of irrigation structures etc.

A storm of given duration over a certain area rarely produces uniform rainfall depth over the entire area. The storm usually has a centre, where the rainfall P_o is maximum which is always larger than the average depth of rainfall P for the area as a whole. Generally, the difference between these two values, that is $(P_o – P)$, increases with increase in area and decreases with increase in the duration. Also the difference is more for convective and orographic precipitation than for cyclonic. To develop quantitative relationship between P_o and P, a number of storms with data obtained from recording raingauges have to be analysed. The analysis of a typical storm is described below (taken from Reddy, 1996).

The rainfall data is plotted on the basin map and the isohyets are drawn. These isohyets divide the area into various zones. On the same map the Thiessen polygons are also constructed for all the raingauge stations. The polygon of a raingauge station may lie in different zones. Thus each zone will be influenced by a certain number of gauges, whose polygonal areas lie either fully or partially in that zone. The gauges, which influence each zone along with their influencing areas, are noted. Next for each zone the cumulative average depth of rainfall (areal average) is computed at various time using the data of rainfall mass curve at the gauges influencing the zone and the Thiessen weighted mean method. In other words in this step the cumulative depths of rainfall at different times recorded at different parts are converted into cumulative depths of rainfall for the zonal area at the corresponding times. Then the mass curves of average depth of rainfall for accumulated areas are computed starting from the zone nearest to the storm centre and by adding one more adjacent to it each time, using the results obtained in the previous step and using the Thiessen weight in proportion to the areas of the zones. These mass curves are now examined to find the maximum average depth of rainfall for different duration and for progressively increasing accumulated areas. The results are then plotted on semi-logarithmic paper. That is, for each duration the maximum average depth of rainfall on an ordinary scale is plotted against the area on logarithmic scale. If a storm contains more than one storm centre, the above analysis is carried out for each storm centre. An enveloping curve is drawn

for each duration. Alternatively, for each duration a depth area relation of the form as proposed by Horton may be established:

$$P = P_o \, e^{-kA^n} \qquad (8.1)$$

where: P_o = highest amount of rainfall at the centre of the storm (A = 25 km^2) for any given duration

P = maximum average depth of rainfall over an area A (> 25 km^2) for the same duration

A = area considered for P

k, n = regression coefficients, which vary with storm duration and region.

Example 4.5

The following numerical example illustrates the method described above. In and around a catchment with an area of 2790 km^2 some 7 raingauges are located, see Figure 4.20. The record of a severe storm measured in the catchment as observed at the 7 raingauge stations is presented in Table 4.8 below:

Time in hours	Cumulative rainfall in mm measured at raingauge stations						
	A	B	C	D	E	F	G
4	0	0	0	0	0	0	0
6	12	0	0	0	0	0	0
8	18	15	0	0	0	6	0
10	27	24	0	0	9	15	6
12	36	36	18	6	24	24	9
14	42	45	36	18	36	33	15
16	51	51	51	36	45	36	18
18	51	63	66	51	60	39	18
20	51	72	87	66	66	42	18
22	51	72	96	81	66	42	18
24	51	72	96	81	66	42	18

Table 4.8: Cumulative rainfall record measured for a severe storm at 7 raingauges (A to G)

The total rainfall of 51, 72, 96, 81, 66, 42 and 18 mm are indicated at the respective raingauge stations A, B, C, D, E, F and G on the map. The isohyets for the values 30, 45, 60 and 75 mm are constructed. Those isohyets divide the basin area into five zones with areas as given in Table 4.9. The Thiessen polygons are then constructed for the given raingauge network [A to G] on the same map. The areas enclosed by each polygon and the zonal boundaries for each raingauge is also shown in Table 4.9.

Zone	Area	Raingauge station area of influence in each zone (km^2)						
	km^2	A	B	C	D	E	F	G
I	415	0	105	57	253	0	0	0
II	640	37	283	0	20	300	0	0
III	1015	640	20	0	0	185	170	0
IV	525	202	0	0	0	0	275	48
V	195	0	0	0	0	0	37	158

Table.4.9: Zonal areas and influencing area by rain gauges

As can be seen from Figure 4.20 Zone I (affected by the rainfall stations with the highest point rainfall amounts) is the nearest to storm centre while Zone V is the farthest.

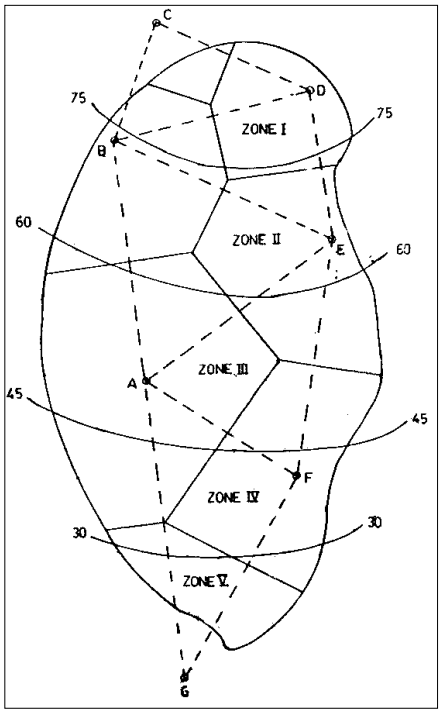


Figure 4.20:
Depth-area-duration analysis

The cumulative average depth of rainfall for each zone is then computed using the data at Raingauge stations A, B, C, D, E, F and G and the corresponding Thiessen weights. For example, the average depth of rainfall in Zone I at any time, PI is computed from the following equation.

$$P_I = \frac{105 \times P_B \times 57 \times P_C + 253 \times P_D}{(105 + 57 + 253)}$$

where P_B, P_C and P_D are the cumulative rainfalls at stations B, C and D at any given time. That is

$$P_I = 0.253\ P_B + 0.137\ P_C + 0.610\ P_D$$

Similarly for Zone II, we have:

$$P_{II} = \frac{37 \times P_A \times 283 \times P_B + 20 \times P_D + 300 \times P_E}{(37 + 283 + 20 + 300)}$$

or:

$$P_{II} = 0.058\ P_A + 0.442\ P_B + 0.031\ P_D + 0.469\ P_E \text{ and so on.}$$

These results are shown in Table 4.11 and Figure 4.22.

Time (hours)	Zone I	Zone II	Zone III	Zone IV	Zone V
4	0	0	0	0	0
6	0	0.70	7.60	4.62	0
8	3.80	7.67	12.66	10.07	1.14
10	6.07	24.07	21.66	18.80	7.71
12	15.23	29.44	31.81	27.26	11.85
14	27.30	39.77	39.47	34.83	18.42
16	41.85	46.31	46.86	40.14	21.42
18	56.09	60.53	50.87	41.71	21.99
20	70.40	64.78	52.65	43.28	22.56
22	80.78	68.25	52.65	43.28	22.56
24	80.78	68.25	52.65	43.28	22.56

Table 4.11: Cumulative average depths of rainfall in various zones in mm.

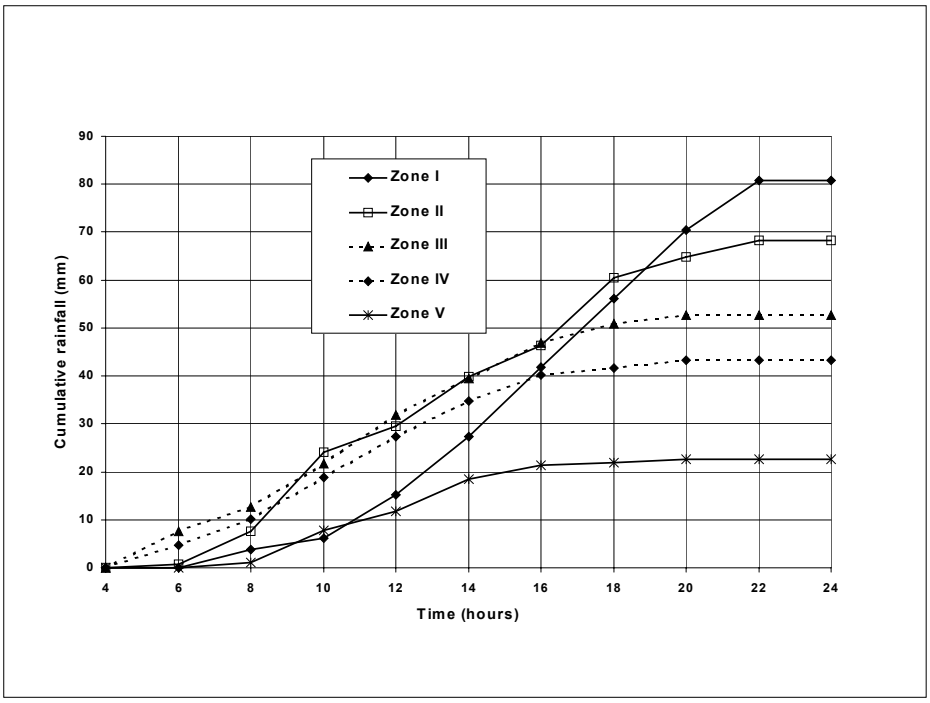


Figure 4.22: Cumulative average depths of rainfall in Zones I to V

In the next step the cumulative average rainfalls for the progressively accumulated areas are worked out. Here the weights are used in proportion to the areas of the zones. For example, the cumulative average rainfall over the first three zones is given as

$$P_{I+II+III} = \frac{415 \times P_I + 640 \times P_{II} + 1015 \times P_{III}}{415 + 640 + 1015}$$

$$= 0.2\ P_I + 0.31\ P_{II} + 0.49\ P_{III}$$

The result of this step are given in Table 4.11 and Figure 4.22.

Time hours	I 415 km^2	I + II 1055 km^2	I + II + III 2070 km^2	I + II + III + IV 2595 km^2	I + II + III+ IV + V 2790 km^2
4	0	0	0	0	0
6	0	0.43	3.94	4.08	3.79
8	3.80	6.15	9.34	9.49	8.91
10	6.07	17.00	19.28	19.18	18.38
12	15.23	23.86	27.76	27.66	26.55
14	27.30	34.87	37.12	36.66	35.38
16	41.85	44.56	45.69	44.57	42.95
18	56.09	58.79	54.91	52.24	50.12
20	70.40	66.99	59.96	56.59	54.21
22	80.78	73.17	63.12	59.11	56.55
24	80.78	73.17	63.12	59.11	56.55

Table 4.11: Cumulative average rainfalls for accumulated areas in mm

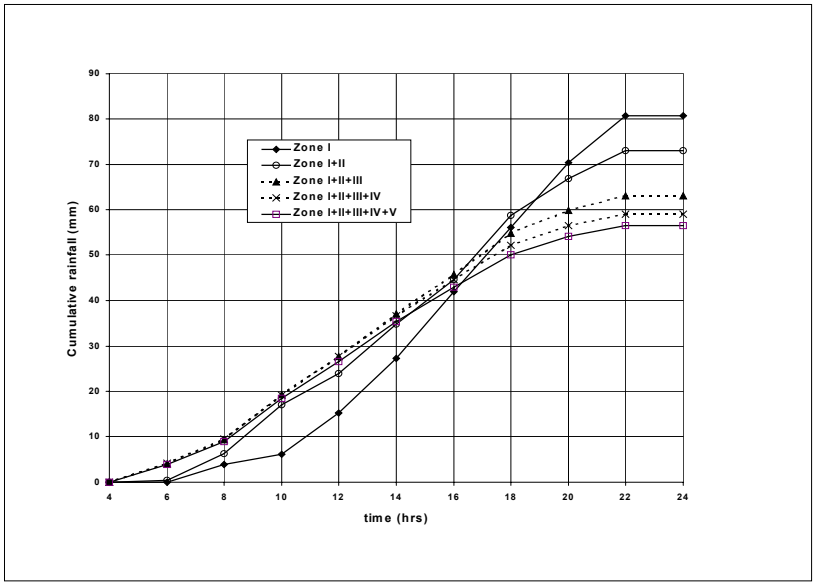


Figure 4.22: Cumulative average depths of rainfall in cumulated areas Zones I to I+II+III+IV+V

Now for any zone the maximum average depth of rainfall for various durations of 4, 8, 12, 16 and 20 h can be obtained from Table 4.10 by sliding a window of width equal to the required duration over the table columns with steps of 2 hours. The maximum value contained in the window of a particular width is presented in Table 4.12.

Duration in hours	Maximum average depths of rainfall in mm				
	415 km^2	1055 km^2	2070 km^2	2595 km^2	2790 km^2
4	28.79	23.92	18.42	18.17	17.64
8	55.17	43.13	36.35	35.08	34.04
12	74.71	60.84	50.97	48.16	46.33
16	80.78	72.74	59.18	56.59	54.21
20	80.78	73.17	63.12	59.11	56.55

Table 4.12: Maximum average depths of rainfall for accumulated areas

For each duration, the maximum depths of rainfall is plotted against the area on logarithmic scale as shown in Figure 4.23.

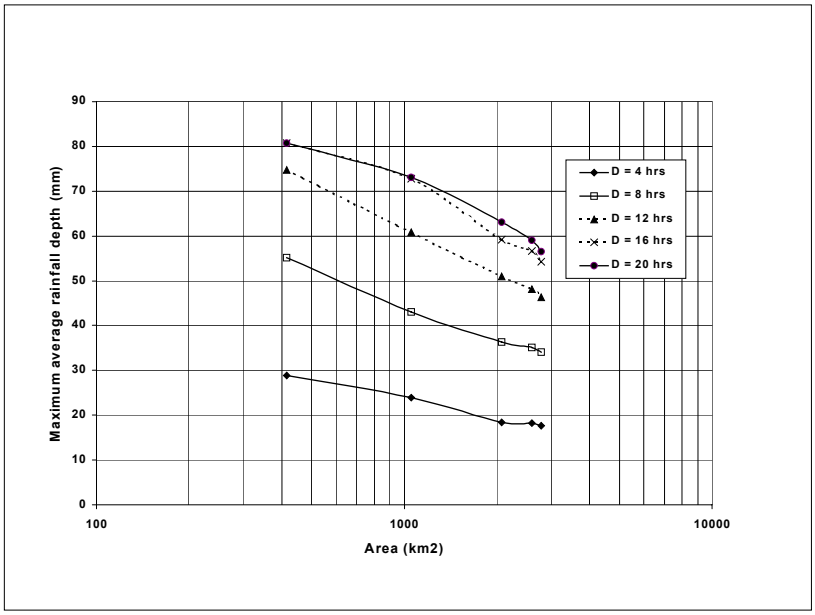


Figure 4.23: Depth-area-duration curves for a particular storm

By repeating this procedure for other severe storms and retrieving from graphs like Figure 4.23 for distinct areas the maximum rainfall depths per duration, a series of storm rainfall depths per duration and per area is obtained. The maximum value for each series is retained to constitute curves similar to Figure 4.23. Consequently, the maximum rainfall depth for a particular duration as a function of area may now be made of contributions of different storms to produce the overall maximum observed rainfall depth for a particular duration as a function of area to constitute the depth-area-duration (DAD) curve. For the catchment considered in the example these DAD curves will partly or entirely exceed the curves in Figure 4.23 unless the presented storm was depth-area wise the most extreme one ever recorded.

Areal reduction factor

If the maximum average rainfall depth as a function of area is divided by the maximum point rainfall depth the ratio is called the Areal Reduction Factor (ARF), which is used to convert point rainfall extremes into areal estimates. ARF-functions are developed for various storm durations. In practice, ARF functions are established based on average DAD's developed for some selected severe representative storms.

These ARF's which will vary from region to region, are also dependent on the season if storms of a particular predominate in a season. Though generally ignored, it would be of interest to investigate whether these ARF's are also dependent on the return period as well. To investigate this a frequency analysis would be required to be applied to annual maximum depth-durations for different values of area and subsequently comparing the curves valid for a particular duration with different return periods.

In a series of Flood Estimation Reports prepared by CWC and IMD areal reduction curves for rainfall durations of 1 to 24 hrs have been established for various zones in India (see e.g. CWC, Hydrology Division, 1994). An example is presented in Figure 4.24 (zone 1(g)).

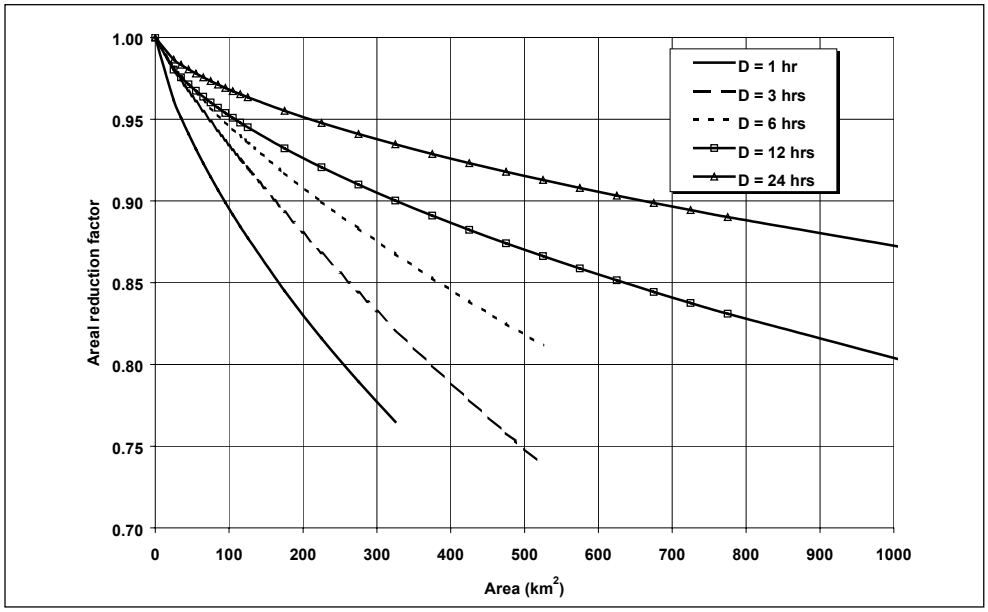


Figure 4.24: Example of areal reduction factors for different rainfall durations

Time distribution of storms

For design purposes once the point rainfall extreme has been converted to an areal extreme with a certain return period, the next step is to prepare the time distribution of the storm. The time distribution is required to provide input to hydrologic/hydraulic modelling. The required distribution can be derived from cumulative storm distributions of selected representative storms by properly normalising the

horizontal and vertical scales to percentage duration and percentage cumulative rainfall compared to the total storm duration and rainfall amount respectively. An example for two storm durations is given in Figure 4.25, valid for the Lower Godavari sub-zone – 3 (f). From Figure 4.25 it is observed that the highest intensities are occurring in the first part of the storm (about 50% within 15% of the total storm duration). Though this type of storm may be characteristic for the coastal zone further inland different patterns may be determining. A problem with high intensities in the beginning of the design storm is that it may not lead to most critical situations as the highest rainfall abstractions in a basin will be at the beginning of the storm. Therefore one should carefully select representative storms for a civil engineering design and keep in mind the objective of the design study. There may not be one design storm distribution but rather a variety, each suited for a particular use.

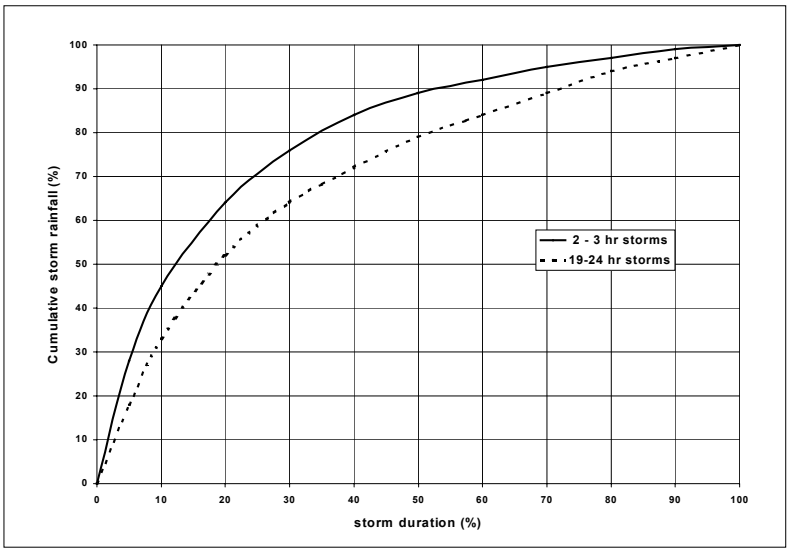


Figure 4.25:
Time distributions of storms in Lower Godavari area for 2-3 and 19-24 storm durations

5 ANALYSIS OF CLIMATIC DATA

5.1 GENERAL

Evaporation is the process by which water is lost to the atmosphere in the form of vapour from large open free water bodies like ponds, rivers, lakes and reservoirs.

Transpiration is the process by which water leaves the body of a living plant and reaches the atmosphere as water vapour. The water present as soil moisture in the root zone is extracted by the vegetation through its roots and is passed through the stem and branches and is eventually lost as transpiration from the leaves. For hydrological purposes, evaporation and transpiration processes are commonly considered together as evapotranspiration

Potential evapotranspiration (PE) is usually defined as the water loss which will occur from a surface fully covered by green vegetation if at no time there is a deficiency of water in the soil for the use of vegetation. It is primarily dependent on climatic conditions.

Actual evapotranspiration (AE) is the real evapotranspiration at a location dependent on the available moisture in the soil which is in turn dependent on soil characteristics. It may be calculated from PE for the specific conditions at the site.

Evaporation from a free water surface and potential evapotranspiration are the principal variables of interest in hydrology. Evaporation estimates may be based on measurement of losses from an evaporation pan or on theoretical and empirical methods based on climatological measurements. Practical estimation of potential evapotranspiration depends on estimation from climatological data. Several researchers have developed empirical formulae for estimation of evaporation and evapotranspiration from climatic data. These formulae range from simple regression type equations to more detailed methods such as those representing water budget, energy budget and mass transfer

approaches; the principal methods in use are based on the Penman equation and methodology as discussed in full.

Climatic data (with the exception of measured pan evapotranspiration) are thus not themselves of interest in hydrology but they are required for the estimation of evaporation from open water and evapotranspiration.

5.2 ANALYSIS OF PAN EVAPORATION

5.2.1 PANS FOR ESTIMATING OPEN WATER EVAPORATION

The standard Class A pan used in India, the method of measurement, typical errors and error detection have been described in Operational Manual Part I, Volume 8. Evaporation measured by pans does not represent the evaporation from large water bodies such as lakes and reservoirs. Pans have the following limitations:

- Pans differ from lakes and reservoirs in the heat storage characteristics and heat transfer. Pans exposed above ground are subject to heat exchange through the sides
- The height of rim in an evaporation pan affects the wind action over the surface.
- The heat transfer characteristics of the pan material is different from that of the reservoir

Since heat storage in pans is small, pan evaporation is nearly in phase with climate, but in the case of very large and deep lakes the time lag in lake evaporation may be up to several months. Estimates of annual lake evaporation can be obtained by application of the appropriate lake – pan coefficient to observed pan evaporation.

The lake – pan coefficient is given by E_l / E_p where E_l is the evaporation from the lake and E_p is the evaporation from the pan. Pan - lake coefficients show considerable variation from place to place and from month to month for the same location (WMO Technical Note 126). The variation from month to month precludes the use of a constant pan coefficient.

Monthly pan coefficient depends on climate, and lake size and depth, and range will generally vary from 0.6 to 0.8. For dry seasons and arid climates the pan water temperature is less than the air temperature and the coefficient may be 0.60 or less whilst for humid seasons and climates where the pan water temperature is higher than air temperature pan coefficients may be 0.80 or higher. The average value used is generally 0.7. Based on the studies carried out in India, the average pan - lake coefficient for the Indian Standard pan was found to be 0.8 ranging from 0.65 to 1.10. Ramasastri (1987) computed open water evaporation using pan – lake coefficients for whole of India based on the evaporation data of 104 US Class A pan evaporimeters.

5.2.2 EFFECTS OF MESH SCREENING

The top of the standard pan in use in India is covered fully with a hexagonal wire netting of galvanised iron to protect the water in the pan from birds. The screen has an effect to reduce pan evaporation by about 14 % as compared to that from an un-screened pan. Although a correction factor of 1.144 is commonly applied, it seems preferable, to retain the originally measured values in the archive, to indicate that this is the case in reports, and to leave mesh corrections to users. This is to allow for the possibility that future amendments may be made to the correction factor.

5.2.3 PANS FOR ESTIMATING REFERENCE CROP EVAPOTRANSPIRATION

Provision is made in HYMOS for the estimation of reference crop evapotranspiration from:

$$E_t = K_p \, E_{pan} \tag{5.1}$$

where

K_P = pan coefficient (FAO (1977) publication No 24)

E_{pan} = pan evaporation in mm / day

The pan coefficient is a function of relative humidity, daily windrun and the fetch. The fetch depends on the dryness or wetness of the upwind land surface as illustrated in Fig. 5.1. There are two cases:

- for case 1 the fetch is the length of the upwind green crop from the pan
- for case 2 the fetch is the length of the upwind dry surface between the crop and the pan

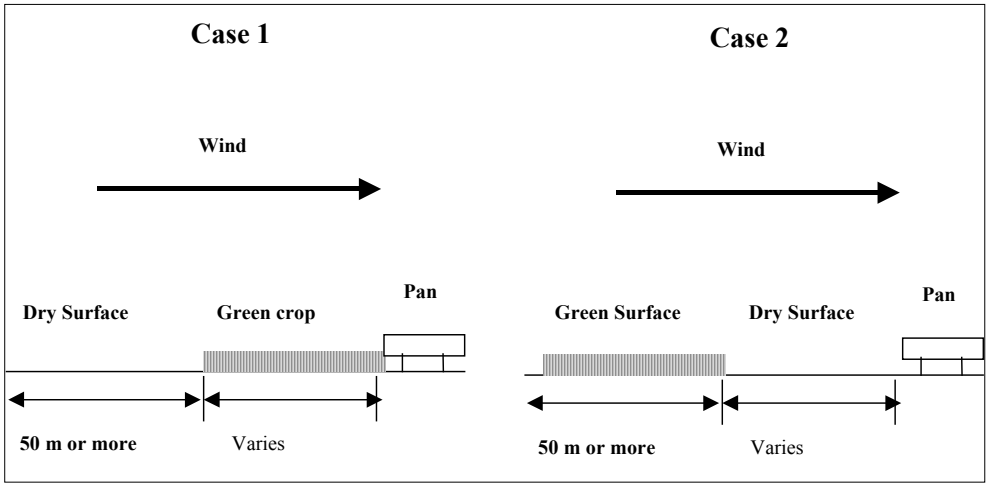


Figure 5.1: Definition sketch for computing pan coefficient

5.2.4 PAN EVAPORATION REFERENCES

Gupta, Shekhar; Vasudev and P. N. Modi (1991) 'A regression model for potential evapotranspiration estimation', Journal of Indian Water Resources Society, Vol 11, No 4 pp 30 – 32

Ramasastri, K. S. (1987) 'Estimation of evaporation from free water surfaces' Proceedings of National Symposium on Hydrology Roorkee (India), pp II – 16 to II – 27

Venkataraman, S. and V. Krishnamurthy (1965) 'Studies on the estimation of Pan evaporation from meteorological parameters', Indian Journal of Meteorology and Geophysics, Vol.16 , No.4 pp 585 - 602

World Meteorological Organisation (1966) 'Measurement and estimation of evaporation and evapo-transpiration' WMO Technical Note No. 83

World Meteorological Organisation (1973) 'Comparison between pan and lake evaporation WMO Technical Note No. 126

5.3 ESTIMATION OF POTENTIAL EVAPOTRANSPIRATION

5.3.1 GENERAL

The Penman method, in wide use for estimation of potential evapotranspiration arose from earlier studies of methods to estimate open water evaporation. In turn, both depend on the combination of two physical approaches which have been used in calculating evaporation from open water:

- the mass transfer method, sometimes called the vapour flux method, which calculates the upward flux of water vapour from the evaporating surface
- the energy budget method which considers the heat sources and sinks of the water body and air and isolates the energy required for the evaporating process

The disadvantage of these methods is the requirement for data not normally measured at standard climatological stations. To overcome this difficulty Penman (1948) developed a formula for calculating open water evaporation, combining the physical principles of the mass transfer and energy budget methods with some empirical concepts incorporated, to enable standard meteorological observations to be used. The method was subsequently adapted to estimate potential evapotranspiration and to substitute alternative more commonly measured climatic variables for those less commonly measured. Reference is made to Volume 3 for a derivation of the Penman equation (Volume 3, Design Manual, Hydro-meteorology, Chapter 2).

5.3.2 THE PENMAN METHOD

The Penman formula may be presented in a number of formats but may be conveniently expressed as follows:

$$E = \frac{\Delta}{\Delta+\gamma} R_n + \frac{\gamma}{\Delta+\gamma} f(u)(e_s - e_a) \qquad (5.2)$$

where:

E = reference crop evapotranspiration (mm/day)
Δ = slope of e_s - t curve at temperature t ($kP_a/^oC$)
γ = psychometric constant ($kP_a/^oC$)
R_n = net radiation (mm/day)
f(u) = wind related function
e_s = saturation vapour pressure at mean air temperature (kP_a)
e_a = actual vapour pressure (kP_a)

The vapour pressure-temperature gradient Δ is computed from:

$$\Delta = \frac{de_s}{dT} = \frac{4098\, e_s}{(237.3 + T)^2} \qquad (5.3)$$

where

T = t + 273.16 (K)
t = air temperature (^oC)

and

$$e_s(T) = 0.6108 \exp\left(\frac{17.27T}{T+237.3}\right) \qquad (5.4)$$

The psychmetric constant γ ($kP_a/°C$) is computed from:

$$\gamma = \frac{e_s(T_w) - e_a(T_a)}{T - T_w} = \frac{c_p\, p}{\varepsilon\, \lambda} \quad (5.5)$$

where

c_p = specific heat of air (=1.005 kJ hg^{-1} $°C^{-1}$)
p = atmospheric pressure (kPa)
ε = ratio of molecular masses of water vapour and dry air = 0.622
λ = latent heat of vaporisation ($kJkg^{-1}$)

Where the air pressure is not measured, it is estimated as:

$$p = 101.3\left(\frac{T - 0.0065H}{T}\right)^{5.256} \quad (5.6)$$

where

H = elevation relative to m.s.l (m)

Where net radiation R_n is not available (as is normally the case in India), it can be substituted in turn by net shortwave and net longwave radiation, and then by bright sunshine totals which are more commonly measured at standard climatological stations. Thus net radiation can be computed from:

$$R_n = R_{ns} - R_{nl} \quad (5.7)$$

where

R_{ns} = net shortwave radiation
R_{nl} = net longwave radiation

and in turn net shortwave radiation is:

$$R_{ns} = (1-\alpha)R_s \quad (5.8)$$

where

α = albedo
R_s = shortwave radiation

If the shortwave radiation is not available it is computed from:

$$R_s = R_a\,(a_1 + b_1\, n/N) \quad (5.9)$$

where

R_a = extra terrestrial radiation (available from tables dependent on latitude and time of year)
n/N = actual to maximum bright sunshine duration (from Campbell Stokes sunshine recorder)
a_1 , b_1 = coefficients

If the net longwave radiation is not available it is estimated from:

$$R_{nl} = \sigma T^4\,(a_2 - b_2\sqrt{e_a})(a_3 + b_3\, n/N) \quad (5.10)$$

where
σ = Boltzmann constant ($\sigma = 2.10^{-9}$)
a_2, b_2 = coefficients in vapour term
a_3, b_3 = coefficients in radiation term

The wind function f(u), as proposed by FAO is given by:

$$f(u) = 0.26\,(1+U_{24}/100) \tag{5.11}$$

where

U_{24} = 24 hour wind run (km/day) measured at 2m above ground level

The actual vapour pressure is computed by one of the following three formulae depending on which time series is available. For current data the formula using wet and dry bulb temperature is used even if relative humidity and dew point have already been calculated by the observer; this is to avoid incorporating observer's calculation errors. The other formulae may be required for historic data where wet and dry bulb temperatures are no longer available.

$$e_a = e_s\ rh/100 \tag{5.12}$$

$$e_a = e_s\,(t_{wb}) - \gamma\,(t_{db} - t_{wb}) \tag{5.13}$$

$$e_a = e_s\,(t_{dew}) \tag{5.14}$$

where: rh = relative humidity in %
t_{wb}, t_{db} = wet and dry bulb temperature (^oC)
t_{dew} = dew point temperature (^oC)

Daily potential evapotranspiration using the Penman formula may thus be computed using the following observations at standard Indian climatological stations:

t_{max} , t_{min}	to obtain t_{mean} as (t_{max} ,+ t_{min})/2 in (^oC)
t_{wb}, t_{db}	to obtain actual and saturated vapour pressures (e_a , e_s)
U_{24}	to obtain the wind function *f(u)*
n	actual daily bright sunshine duration using Campbell Stokes sunshine recorder to compute net shortwave and net longwave radiation (R_{ns} , R_{nl})

For current data these series must be available for calculation of evapotranspiration to be carried out. Other constants and coefficients required by the method are held in HYMOS.

5.4 OTHER POTENTIAL EVAPOTRANSPIRATION FORMULAE

A large number of empirical and theoretical formulae have been proposed for the calculation of potential evapotranspiration and several of these are available in HYMOS. These will not form a part of routine processing but may be used for special applications. The following methods are available:

- Christiansen method
- FAO radiation method
- Makkink radiation method
- Jensen-Haise method
- Blaney-Criddle method
- Mass transfer method

The minimum requirements of observed variables to obtain estimates using the above methods is shown in Table 5.1.

	Christiansen	Radiation	Makkink	Jensen-Haise	Blaney-Criddle	Mass Transfer
Air pressure						
Temp. max.	✓	✓	✓	✓	✓	
Temp. min.	✓	✓	✓	✓	✓	
Temp. db.	✓	✓			✓	✓
Temp. wb.	✓	✓			✓	✓
Wind run	✓	✓			✓	✓
Sunshine hrs	✓	✓	✓	✓	✓	
Altitude	✓			✓		

Table 5.1: Minimum series requirements to obtain PE estimates by various methods

6 ANALYSIS OF WATER LEVEL DATA

Water levels of rivers do basically not form a homogeneous set of data, suitable for statistical analysis. A water level in a river is the result of an upstream discharge and the hydraulic conditions of the downstream control, i.e. the hydraulic characteristics of the control channel and in case of backwater a downstream water level.

For a river with a flood plain one will observe that a fixed increase of discharge at stages where the river flows inbank will produce a much larger increase in the water level than when the river flows overbank; in case of a very wide flood plain the levels will hardly rise in response of an increase of discharge, as depicted in Figure 6.1.

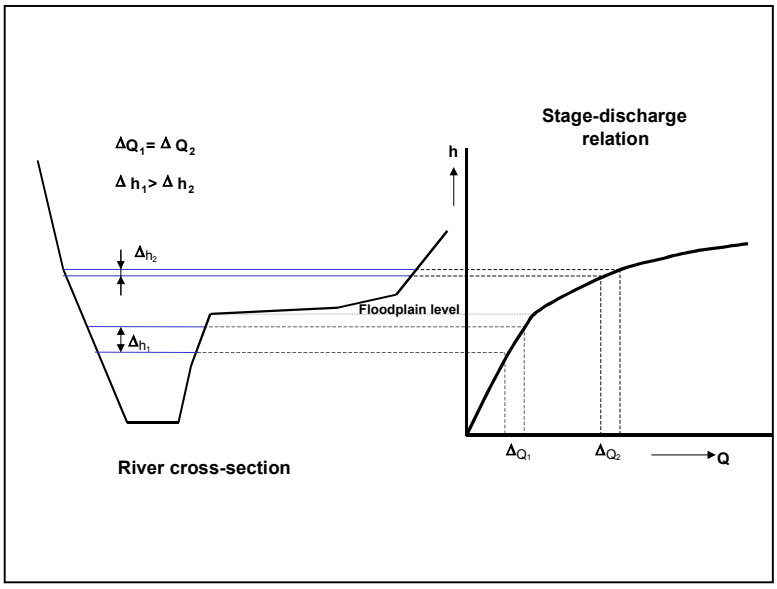


Figure 6.1: Response of river stage to fixed increase of discharge for inbank and overbank flow

It implies that **extrapolation** of a frequency distribution of river stages is **not justified** as the hydraulic characteristics of the downstream control rather than statistics determine the behaviour of the tail of the frequency distribution. The correct procedure would then be to apply frequency analysis to the discharge. Subsequently, the appropriate stage-discharge relation for the river location is used to transform the frequency distribution of the discharge in one for the water level, as shown in Figure 6.2.

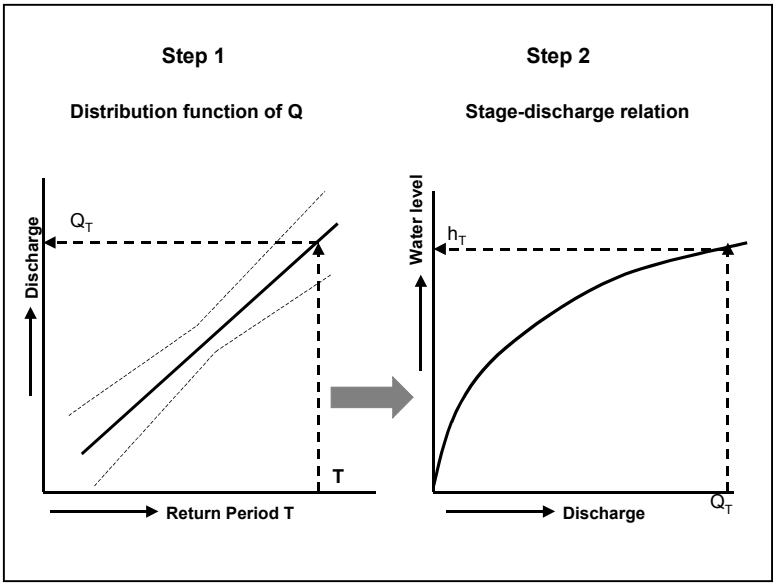


Figure 6.2:
Procedure for derivation of design stages from frequency analysis on discharges

Hence for estimation of the height of a water level at a very low exceedance probability or a large return period T, e.g. for designing the height of dikes to give protection against flooding, the design level is to be obtained via the discharge. Often the design discharge (either as a single value or as a hydrograph) is used as input to a mathematical hydraulic model to arrive at the design level along the river. For correct results, the model should properly represent the geometry of the river and flood plain and their hydraulic roughness.

In the coastal area the derivation of a frequency distribution of water levels becomes more complex as besides the discharge from upstream also the tide at sea affects the water level. Then the joint occurrence of river discharge and tide at sea, possibly to be extended with wind setup has to be considered for making statistical interference on river stages. The application of a mathematical hydraulic model, run for a variety of combinations of inflow hydrographs and dowstream tidal levels, is in such cases is indispensable to arrive at the design stage.

Frequency analysis of river stages is only justified when **interpolation** takes place and no extrapolations are being made **and the downstream control of the river location has not changed in the course of time**. Such analysis is useful e.g. for navigation purposes, where the number of days in a year that a certain water level is exceeded and hence a Least Available Depth can be guaranteed is of importance, or when intake levels have to be determined for water abstractions.

7 ANALYIS OF DISCHARGE DATA

7.1 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

Reference is made to Annex 1 for a review of statistics relevant for the analysis of run-off data.

7.2 COMPUTATION OF BASIC STATISTICS

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

- arithmetic mean

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \qquad (7.1)$$

- 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 Sx :

$$S_x = \sqrt{\frac{\sum (X_i - \overline{X})^2}{N-1}} \qquad (7.2)$$

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

$$C_X = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^3}{S_X{}^3} \qquad (7.3)$$

- kurtosis or peakedness of a distribution

$$K_X = \frac{(N^2 - 2N + 3)}{(N-1)(N-2)(N-3)} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^4}{S_X{}^4} \qquad (7.4)$$

7.3 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 7.1). 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. 7.1 shows an example based on natural scales for the data in Table 7.1. 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. 7.2). 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.

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		

Table 7.1: Derivation of flow frequencies for construction of a flow duration graph

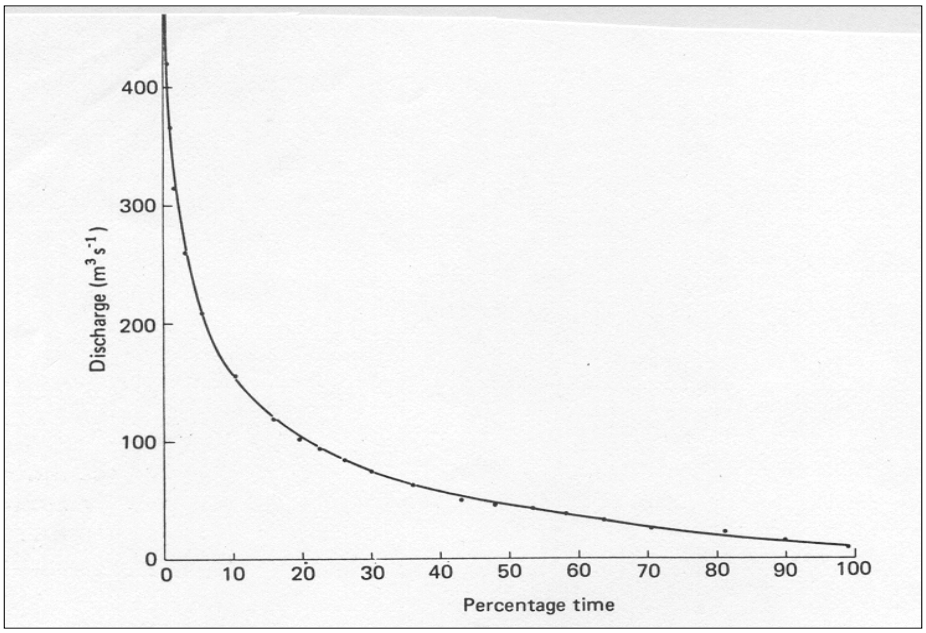


Figure 7.1: Flow duration curve plotted on a natural scale

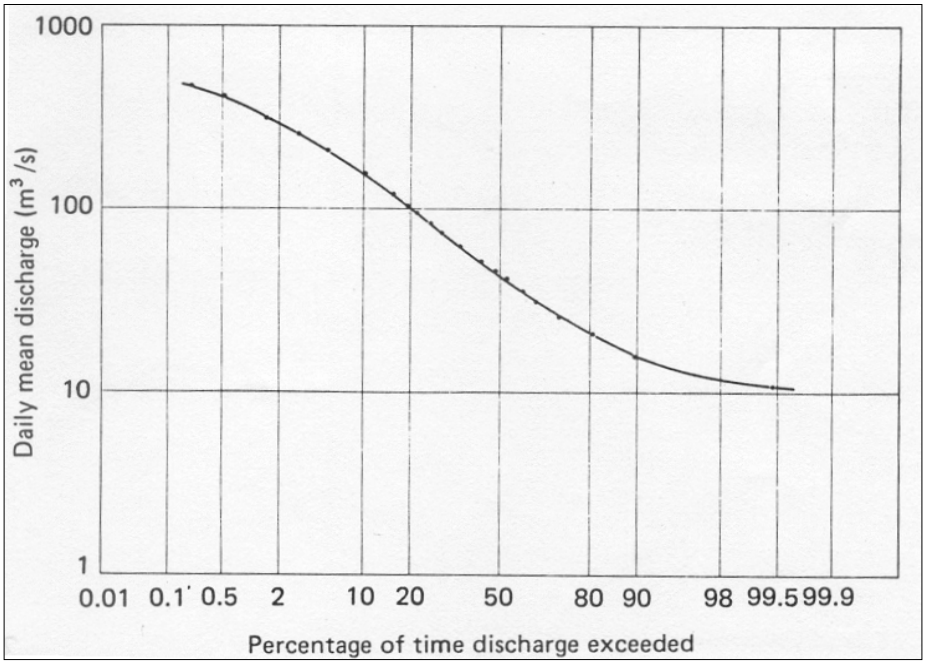


Figure 7.2: Flow duration curve plotted on a probability scale

From Figure 7.2 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)(Figure 7.3). 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.

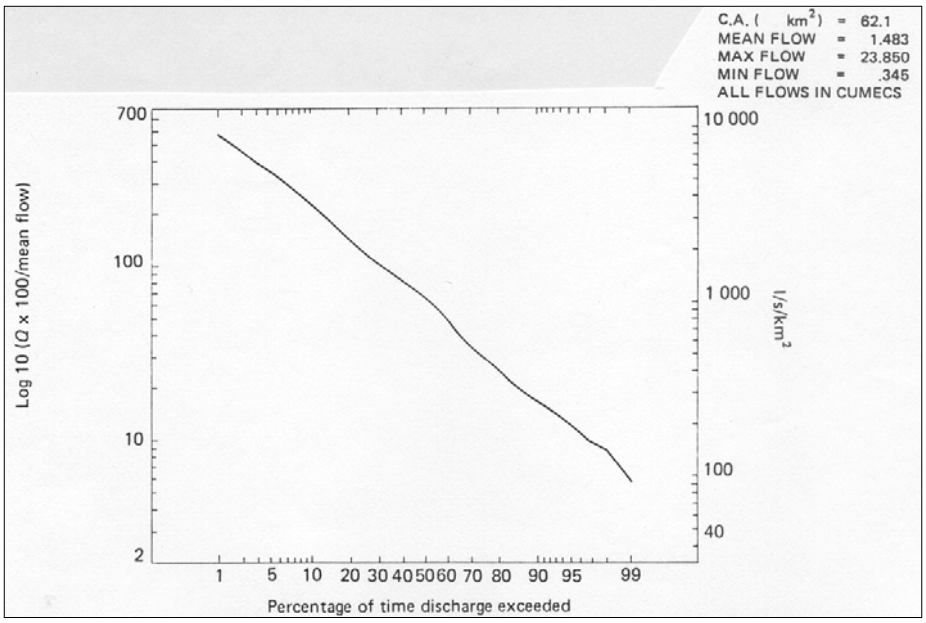


Figure 7.3: Flow duration curve standardised by mean flow (after Shaw 1995)

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.

7.4 FITTING OF FREQUENCY DISTRIBUTIONS

7.4.1 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.

7.4.2 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 Figure 7.4.

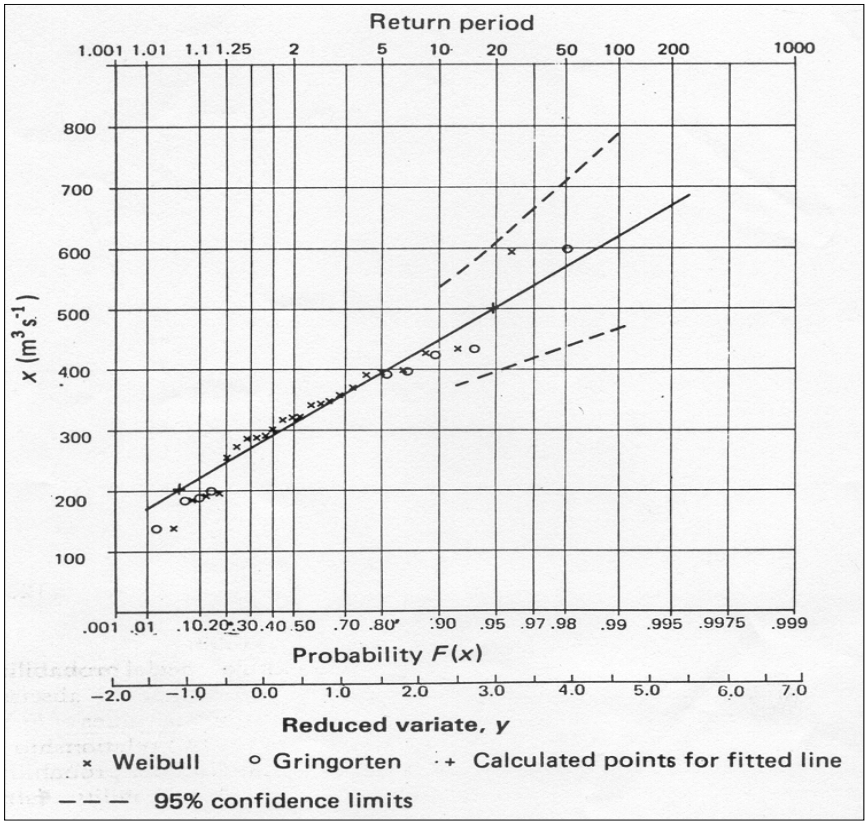


Figure 7.4:
Flow frequency curve showing discharge plotted against return period (top) and probability (lower) (after Shaw, 1995)

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.

7.5 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

7.5.1 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:

$$Y_i = \frac{1}{(2M+1)} \sum_{j=i-M}^{j=i+M} X_j \qquad (7.5)$$

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 7.2 and Figure 7.5

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		

Table 7.2: Computation of moving averages (M = 1)

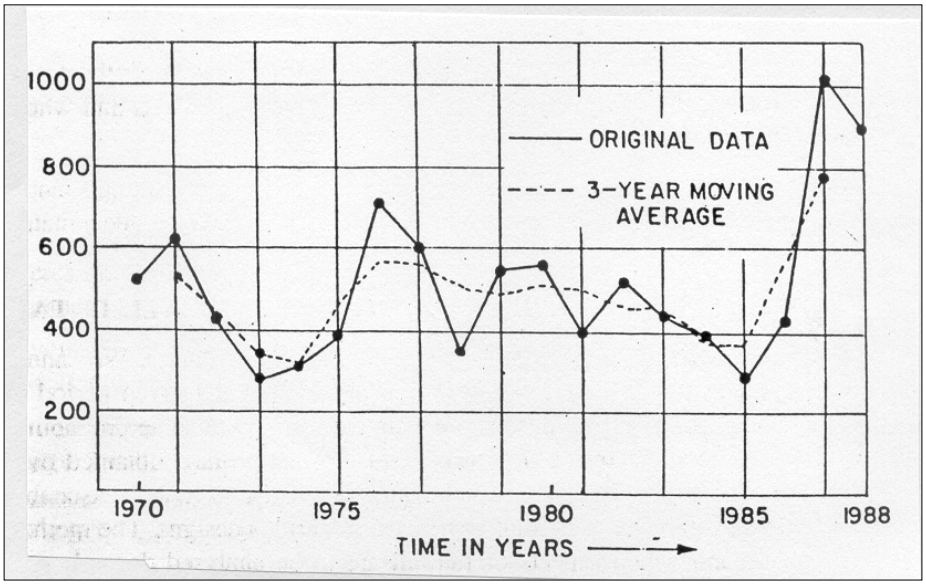


Figure 7.5: Moving average of annual runoff

7.5.2 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 Figure 7.6, 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.

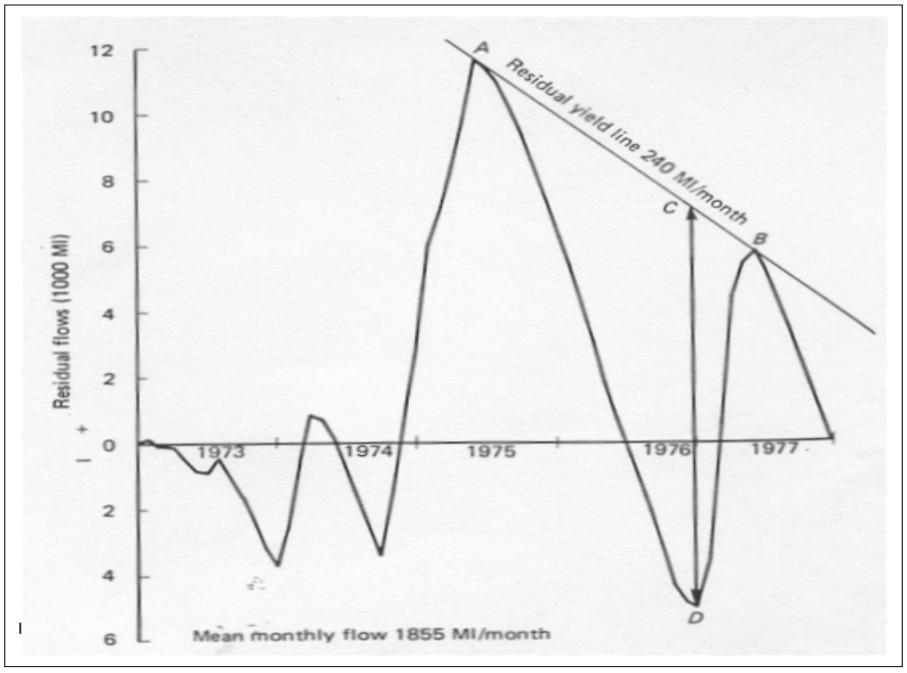


Figure 7.6: Residual mass curve used in drought analysis

7.5.3 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 Figure 7.7. 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.

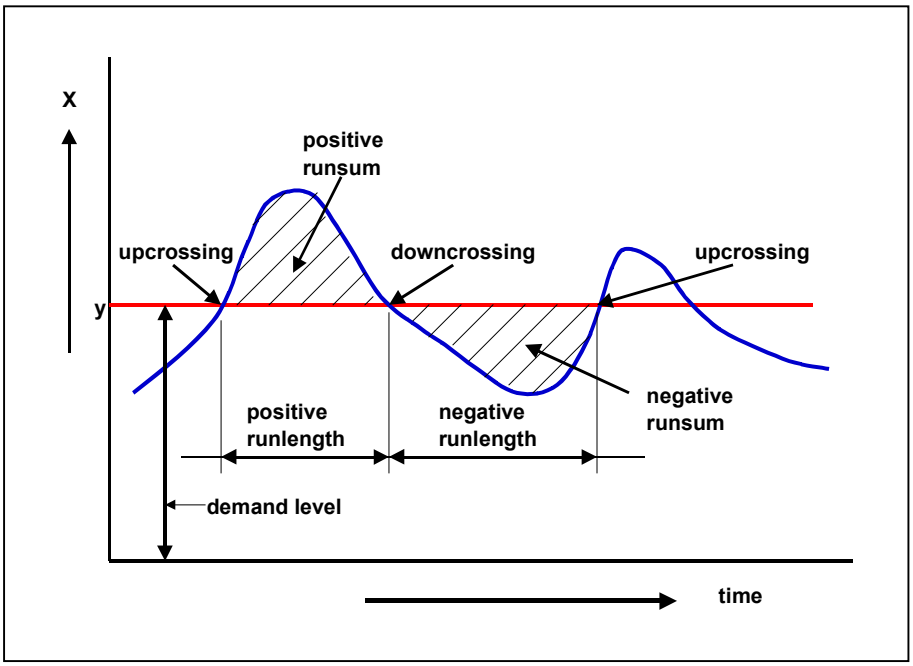


Figure 7.7: Definition diagram of run-length and run-sum

7.5.4 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:

$$S_i = S_{i-1} + (X_i - D_x)\, C_f \qquad \text{for } i = 1, 2N \;; S_0 = 0 \qquad (7.6)$$

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.

7.5.5 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 Part II, Chapter 12.

7.6 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, water level and discharge, see part II.

7.7 DOUBLE MASS ANALYSIS

The technique of double mass analysis is again widely used in validation of all climatic variables and is described in Part II, for rainfall, for climate and for discharge. For discharge series the technique is particularly useful for validation purposes by comparing an observed discharge series with a computed discharge series, as shown in Chapter 2. The method is not discussed further here.

7.8 SERIES HOMOGENEITY TESTS

Series homogeneity tests with respect to climate are described in Part II, Chapter 2 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

7.9 RAINFALL RUNOFF SIMULATION

Rainfall runoff simulation for data validation is described in Chapter 3 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.)

8 ANALYSIS OF WATER QUALITY DATA

8.1 INTRODUCTION

8.1.1 OBJECTIVES

This chapter presents aspects of data analysis that are relevant to the surface water quality data collected by central and state organisations. It is prepared for water quality specialists in the data centres, who are responsible for analysing water quality data compiled from different laboratories. While data analysis is done for multiple purposes, and many different types of analyses are possible, this chapter focuses on the data analysis needed for the production of a standardised yearbook for surface water quality as presented in chapter 14. As such, it is not a complete overview of all the possible tests and analyses that can be done for water quality data – rather is it a selection geared for regular production of yearbooks.

This chapter presents data validation and different statistical analyses, discusses the relevance and possible applications of each analysis and gives examples using some water quality data sets. In most cases, the examples have been processed using the HYMOS software. This chapter is ***NOT*** intended to be a detailed HYMOS tutorial for water quality data analysis, but presents the method, results of analyses and results as calculated in HYMOS.

8.1.2 RELATION HYMOS AND SWDES

The HP water quality laboratories all have customised software for data entry of analytical results (SWDES). SWDES also has the functionality to make some simple data analysis and graphical presentation. However, for more advanced data analysis, also with possibilities for comparison of data from different laboratories, HYMOS is the required software.

The procedure for data analysis is as follows:

- Laboratories enter analytical data in SWDES
- Laboratories can make initial data analysis and graphical presentation of their data analysis results
- On a regular basis, laboratories should export data from SWDES and send a data diskette to the State Data Center where HYMOS is installed.
- Complete analysis of water quality data and production of yearbooks will be done by the water quality specialist at the State Data Center, using HYMOS.

8.1.3 SAMPLE DATA SETS

Some sample water quality data sets are used in many of the examples presented in this chapter. These data sets are given at the end of this chapter and are referred to in the text when they are used in examples. An overview of data sets including the Station and Series name in HYMOS (Water Quality Test Data) is given in Table 8.1.

No.	Source	In Report	HYMOS		For:
			Station	Series	
1	(modified after) Gilbert, p. 190	Table 8.18, Fig. 8.1, Tab. 8.2	Outliers	Ros	Rosner - outliers
2	(modified after) McBean p.120	Table 8.6 Also used for Rosner since n=25	Outliers	DX2	Dixon – outlier (also Rosner)
3	McBean	Example 8.3	Outliers	DIX	Dixon – outlier
4	Gilbert, p144	Table 8.19	BasicStat	NO3	Basic Statistics
5	Stowa1 Cd at Neerbeek	Table 8.20	STOWA1	QCD	• Basic Stats • Linear Trend • Box-Whisker
6	Stowa2 PO4 Bergsche Achterplas	Table 8.21	Amstdampl as	Ros	Step trend on *paired* data

Table 8.1: Overview of Data Sets used

8.2 VALIDATION AND SCREENING

Validation and screening are important first steps in data processing. HYMOS offers a range of tabular, graphical, computational and statistical validation techniques for these purposes, such as:

- plotting of time series
- identifying and flagging of outliers

Various options are available for series completion. Among these are interpolation techniques which use series relations derived with e.g. regression techniques or spatial relations. Much emphasis has been given to facilities for administration of the data processing steps and data qualification features. The latter include assignment of labels to individual data indicating the source - e.g. original or corrected - and reliability of data.

Screening of data aims at detecting outliers. An outlier can be defined as an observation that does not matches with the pattern of earlier observations. Outliers may be the result of mistakes in the data generation process but could also be an indication of a true change in the system under study, such as an accidental spill on a river stretch. A shear infinity of causes of mistakes is possible in the starting with sample collection, transport, storage and analysis upto entering into files and computers. Since it is best to detect mistakes as soon as possible after they have been made, extensive data checking is done upon entering data in the laboratories (using SWDES software). Part of these checks are again described in Section 8.2.1. Section 8.2.2 describes how historical data from the same location can be used to check newly entered data. Section 8.2.3 illustrates Rosner's test for detecting one or more outliers in a series of observations.

After identification of one or more outliers, a decision has to be made what to do with the data point. If an obvious mistake is detected than, if possible, the corrected value wil be entered. Otherwise the data point may be excluded. Section 8.2.4. describes the options and how to keep track of original data and corrected data in HYMOS.

8.2.1 CONSISTENCY CHECKS

Primary data checking is done when data are entered into SWDES software in the laboratories. Part of the primary checking is repeated in HYMOS. There is the reason for this duplication: if in SWDES an outlier is detected and there is no traceable error in the sampling or analysis, the observation will remain in the database. In HYMOS this observation should be checked again. At the level of HYMOS, a more powerful analysis is possible because of more statistical tests and because more data (from different stations or agencies) may be available.

8.2.2 CONTROL CHARTS

Control charts should be established for all monitoring stations and all parameters based upon historical data. These plots serve as a guide for when investigation action is required.

The control chart is usually made based on individual data points. Values for the previous 3 years are used to calculate the mean, and upper/lower limits. New values can then be compared to the historical range of data.

If there are a lot (many years) of data, annual or seasonal means can be used instead of individual data points. In this case, it will be better to use the previous 5-10 years of data to make the control chart, in order to have enough data values. Note: control charts assume that the data are normally distributed (Gilbert, (1887), p. 194).

Construction of the control chart for historical data is similar to the preparation of the Shewart control chart used for within-laboratory AQC (see WQ training module no. 49). The control chart shows a set of data points, together with the central line (mean of the data), warning limits ($\pm$2 sd) and control limits ($\pm$3 sd).

If a set of water quality data is obtained for a specific station, some variation of the observed values will be evident. However, over a period of time (e.g. 3 years), the expected mean concentration, range of values and variations can be determined. If there are no external influences, any new observations should fit the data distribution established by the historical data (e.g. last 3 years). The function of a control chart is to show the historical data, together with the mean (central line) the warning limits and the control limits. New data can be plotted on the established control chart to identify any deviation from the historical pattern. Before a control chart can be constructed, it is necessary to define a meaningful historical data set. For surface water quality monitoring programme with at least 6 observations a year, a period of 3 years should provide a meaningful historical data set. For groundwater quality monitoring, only one or two observations a year are typically available. A control chart for this type of data may be less meaningful, or would need

Reference: page 194-202 Gilbert and HP Water Quality training modules on AQC.

8.2.3 OUTLIERS

Data outliers are extreme (high or low) values that do not conform with the main body of a data set. Outliers in water quality data sets can occur due to practical mistakes or instrumental failure in all aspects of water quality sampling and analysis: from sample collection, transport, storage, analysis or data entry. They may result from transcription or keypunch errors, or can be the result of instrument breakdowns, calibration problems or power failures.

The presence of one or more outliers within a data set may greatly influence any calculated statistics and yield biased results. Thus outliers should be identified, flagged, and *possibly* removed from a data set. The handling of outliers is discussed in Sub-section 8.2.4.

Several procedures have been developed as alternative methods for detecting outliers, including statistical tests to determine whether an observation appears extreme and does not fit the distribution of the rest of the data. Suggestions for identifying outliers are:

- Graphical analysis of data (visual),
- Rosner's test (statistical, $n \geq 25$)
- Dixon's test (statistical, $n \leq 25$);

Graphical Analysis of Data

Graphical analysis of data is a useful first step to visually check if there are any outliers in a data set. Graphical analysis can identify suspected outliers, which should then be checked with a formal statistical analysis. Graphical analysis can be made by ranking the data (in increasing order) and preparing a probability plot or linear plot of ranked data. Figure 8.1 shows a simple linear plot of the ranked data in Table 8.18, as concentration vs. data rank. The highest and lowest values stand out from the rest of the data and may be outliers. The data set can be analysed with Rosner's test to see if the data values are indeed outliers.

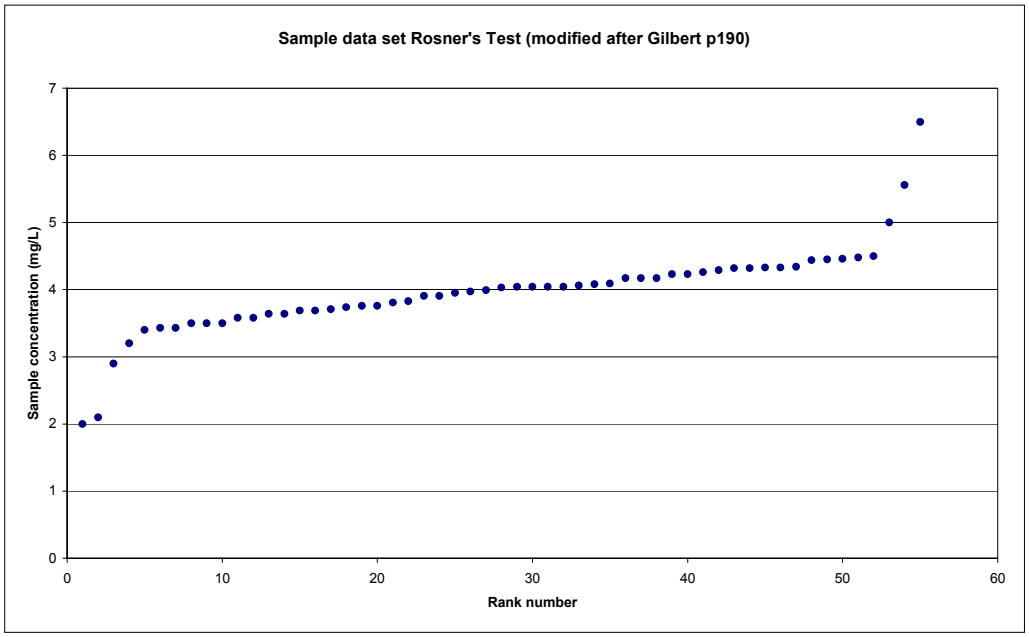


Figure 8.1: Plot of sample concentrations (from Table 8.18 concentration vs. rank), showing possible outliers.

Rosner's Test

Introduction and applicability

Rosner's test is a sequential procedure for identifying up to 10 outliers. The test assumes that the population has a normal distribution. If a lognormal distribution is more plausible, all computations should be performed on the logarithms of the data. Rosner's approach is designed to avoid masking of one outlier by another. Masking occurs when an outlier goes undetected because it is very close in value to another outlier.

The maximum number of outliers detected is 10. The procedure repeatedly deletes the value farthest from the mean and re-computes the test statistic after each deletion. A table is used to evaluate the test statistic when $n \geq 25$ and $n \leq 5000$. Rosner's test is 'two-tailed' since the procedure identifies either suspiciously large or suspiciously small data.

Rosner's test can be applied to a dataset when:

- The data is normally distributed
- The data points are independent of each other
- The number of values (n) in the dataset is ≥ 25
- The dataset is ideally without trend or periodicity

Calculation

The value of Rosner's test is calculated with the following formulas:

$$\overline{X}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i} x_j$$

$$s^{(i)} = \sqrt{\frac{1}{n-i-1}\sum_{j=1}^{n-i}\left(x_j - \overline{X}^{(i)}\right)^2}$$

$$R_{i+1} = \frac{\left|x^{(i)} - \overline{X}^{(i)}\right|}{s^{(i)}} \qquad (8.1)$$

Where:

$R_{(i+1)}$ =	Test statistic for deciding whether the i+1 most extreme values in the data set are outliers from a normal distribution
$\overline{X}^{(i)}$	The mean of the n-i observations in the data set that remain after the i most extreme observation(s) have been deleted
$s^{(i)}$ =	The standard deviation of the n-i observations in the data set that remain after the i most extreme observations have been deleted
$x^{(i)}$ =	The observation which is the furthest from the mean (largest difference with the mean) after i most extreme data (large or small) have been removed.
x_j =	Observation value (with rank j)
n =	The total number of observations in the data set
i =	Index number for the number of extreme observations removed from the dataset

Rosner's test is applied to the following hypothesis:

- H0: The entire data set is from a normal distribution, there are no outliers
- H1: The data set contains 1, 2… or 10 outliers.

The critical values for $\lambda_{(i+1)}$ are tabulated (and computerised in HYMOS) for comparison with $R_{(i+1)}$.

The following procedure is followed (starting with the full dataset, i=0)

1. Compute the mean and standard deviation
2. Compute the Rosner's test statistic $R_{(i+1)}$, where $x^{(i)}$ is the most extreme value in the dataset with n-i data

$$R_{i+1} = \frac{\left|x^{(i)} - \overline{X}^{(i)}\right|}{s^{(i)}} \qquad (8.2)$$

3. Retrieve the tabled critical value $\lambda(i+1)$.
4. Compare the two test values, reject H0 when $R(i+1) > \lambda(i+1)$.
5. Remove the most extreme value from the data (raise i with 1) and redo the procedure, starting from step 1 up to the maximum of 10 iterations.
6. Find the highest value of i where $R_{(i+1)} > \lambda_{(i+1)}$: if i=0 accept H0: dataset contains no outliers, if $i \neq 0$ then accept H1: dataset contains i+1 outliers.

Example 8.1 Rosner's test for outliers

The data set in Table 8.18, see Figure 8.1 is tested for outliers. The observations of nitrate concentration are ordered from smallest to largest (n=55), and are evaluated using Rosner's test in HYMOS. We have already seen that the highest data value may be an outlier.

The test proceeds by first checking the complete data set (n=55, i=0) to see if one outlier is present. After that the test proceeds to test if two outliers are present until the maximum of 10 outliers is reached. The test stops when the maximum number of outliers to test is reached.

i=0	no extreme value is removed	complete dataset is tested if 1 outlier exists
i=1	1 extreme value is removed	reduced dataset is tested if 2 outliers exist in the dataset,
i=2	2 extreme value is removed	reduced dataset is tested if 3 outliers exist in the dataset
etc. up to maximum of i=9		

```
Statistical Tests on Data Homogeneity and Randomness
====================================================

One Series Test
---------------

Series code: Outliers    Ros

Date of first element in series= 1983 5 19 0 0
Number of data                 =  55

Rosner's Test
--------------------
Hypothesis: H0: Series contains no Outliers
            H1: Series contains Outliers


Level of significance is 5 Percent
```

n	mean	st.dev	rem.outl.	extreme	date	R-calc	R-table	Result
55	3.965	0.6665	0	6.50	3-Jan-97	3.80	3.52	H1 Accepted
54	3.918	0.5737	1	2.00	19-May-83	3.34	3.51	H0 Accepted
53	3.955	0.5132	2	2.10	14-Sep-83	3.61	3.505	H1 Accepted
52	**3.990**	**0.4470**	**3**	**5.56**	**2-Jan-97**	**3.51**	**3.50**	**H1 Accepted**
51	3.959	0.3919	4	2.90	8-Dec-83	2.70	3.49	H0 Accepted
50	3.980	0.365	5	5.00	1-Jan-97	2.79	3.48	H0 Accepted
49	3.960	0.338	6	3.20	15-Mar-84	2.25	3.47	H0 Accepted
48	3.976	0.3223	7	3.40	10-May-84	1.79	3.46	H0 Accepted
47	3.988	0.3143	8	3.43	12-Sep-84	1.77	3.45	H0 Accepted
46	4.000	0.3065	9	3.43	10-Oct-84	1.86	3.43	H0 Accepted

```
Data set contains: 4 outliers
```

Table 8.2: Results of Rosner's tests from HYMOS (Analysis of data in Table 8.18)

For each value of i (up to i=9), the key values of Rosner's test (R calc) are computed and compared to the tabulated critical values (R table) as listed in Table 8.2 and 8.3.

- For i=0, the complete data set is evaluated. In this case, 6.50 is the most extreme value in the dataset ($x_j^{(i)}$) as compared to the mean ($\overline{X}^{(i=0)}$ = 3.965).
- For i=1, 1 extreme value (6.50 measured on 3-1-97) is thrown out. In this case, 2.00 is the most extreme value in the remaining dataset ($x_j^{(i)}$) as compared to the mean ($\overline{X}^{(i=1)}$ = 3.918).

The Rosner's test value (R_{calc}) and critical values (R_{table}) for α =0.01 in this example (default α = 0.05) are calculated in HYMOS. The null hypothesis, H0, is tested for each value of I is shown in Table 8.3:

I	Relation R_{calc}, R_{table}	Null Hypothesis	Conclusion on outliers
i=0	$R_{calc} > R_{table}$	H1 is accepted	presence of 1 outlier is shown
i=1	$R_{calc} < R_{table}$	H0 is accepted	presence of 2 outliers cannot be shown
i=2	$R_{calc} > R_{table}$	H1 is accepted	presence of 3 outliers is shown
i=3	**$R_{calc} > R_{table}$**	**H1 is accepted**	**presence of 4 outliers is shown**
i=4	$R_{calc} < R_{table}$	H0 is accepted	presence of 5 outliers cannot be shown
Rows for i=5 to i=9 (testing for 6 to 10 outliers) is not shown because H0 is accepted for all of them (see Table 8.2)			

Table 8.3 Overview of Rosner's test results

The final conclusion is that there is are 4 outliers in the data set. Note from Table 8.2 that outliers no 3 and 4 (2.10 from 14-sept-83 and 5.56 from 2-jan-97) are 'masked' by the first two extreme values (6.5 and 2.00). Upon removing extreme values, the test becomes more sensitive since the standard deviation strongly decreases and the test statistic R becomes higher.

Reference

Statistical Methods for Environmental Pollution Monitoring, R.O. Gilbert, 1987, John Wiley & Sons Inc.

Dixon's Test

Introduction and applicability

Dixon's test uses individual data points at the high and low end of a sorted data set to check for outliers. It should be used for small data sets, i.e. $n \leq 25$. Because Dixon's uses the extremes of a data set (both the highs and the lows), if portions of the data set are censored (see Sub-section 8.2.5) the procedure cannot be utilised.

Calculation

In Dixon's test, the set of observations is first ranked, in increasing order. The ratio of the difference of an extreme (high or low) values from on of its nearest neighbour values is then calculated, using a formula that varies with sample size (see Table 8.4), and varies according to whether the suspected outlier is the smallest or largest value. This ratio is then compared to a tabulated critical value (see Table 8.5) and, if found equal or greater, the extreme value is considered an outlier at the given confidence level.

N	Test criterion (r)	
	High outlier	Low Outlier
3 4 5 6 7	$\frac{X_n - X_{n-1}}{X_n - X_1}$	$\frac{X_2 - X_1}{X_n - X_1}$
8 9 10	$\frac{X_n - X_{n-1}}{X_n - X_2}$	$\frac{X_2 - X_1}{X_{n-1} - X_1}$
11 12 13	$\frac{X_n - X_{n-2}}{X_n - X_2}$	$\frac{X_3 - X_1}{X_{n-1} - X_1}$
14 15 .. .. 24 25	$\frac{X_n - X_{n-2}}{X_n - X_3}$	$\frac{X_3 - X_1}{X_{n-2} - X_1}$

Table 8.4 Formula for test criteria (r) for Dixon's test

n	Level of significance α						
	0.30	*0.20*	*0.10*	*0.05*	*0.02*	*0.01*	*0.005*
3	0.684	0.781	0.886	0.941	0.976	0.988	0.994
4	0.471	0.560	0.679	0.765	0.846	0.889	0.926
5	0.373	0.451	0.557	0.642	0.729	0.780	0.821
6	0.318	0.386	0.482	0.560	0.644	0.698	0.740
7	0.281	0.344	0.434	0.507	0.586	0.637	0.680
8	0.318	0.385	0.479	0.554	0.631	0.683	0.725
9	0.288	0.352	0.441	0.512	0.587	0.635	0.677
10	0.265	0.325	0.409	0.477	0.551	0.597	0.639
11	0.391	0.442	0.517	0.576	0.638	0.679	0.713
12	0.370	0.419	0.490	0.546	0.605	0.642	0.675
13	0.351	0.399	0.467	0.521	0.578	0.615	0.649
14	0.370	0.421	0.492	0.546	0.602	0.641	0.674
15	0.353	0.402	0.472	0.525	0.579	0.616	0.647
16	0.338	0.386	0.454	0.507	0.559	0.595	0.624
17	0.325	0.373	0.438	0.490	0.542	0.577	0.605
18	0.314	0.361	0.424	0.475	0.527	0.561	0.589
19	0.304	0.350	0.412	0.462	0.514	0.547	0.575
20	0.295	0.340	0.401	0.450	0.502	0.535	0.562
21	0.287	0.331	0.391	0.440	0.491	0.524	0.551
22	0.280	0.323	0.382	0.430	0.481	0.514	0.541
23	0.274	0.316	0.374	0.421	0.472	0.505	0.532
24	0.268	0.310	0.367	0.413	0.464	0.497	0.524
25	0.262	0.304	0.360	0.406	0.457	0.409	0.516

Table 8.5 Critical values for the Dixon test of outliers

Example 8.2: Dixon's Test for Outliers (1)

No.	*Conc.*	*No.*	*Conc.*
1	0.5	13	12.4
2	0.6	14	12.4
3	0.8	15	12.5
4	1.0	16	12.5
5	1.1	17	13.1
6	1.9	18	14.5
7	2.9	19	20.2
8	4.6	20	22.1
9	8.8	21	24.7
10	9.2	22	24.9
11	11.1	23	44.0
12	12.1	24	46.9
		25	57.0

Table 8.6:
Measured concentrations (mg/l) at a monitoring site (n=25)

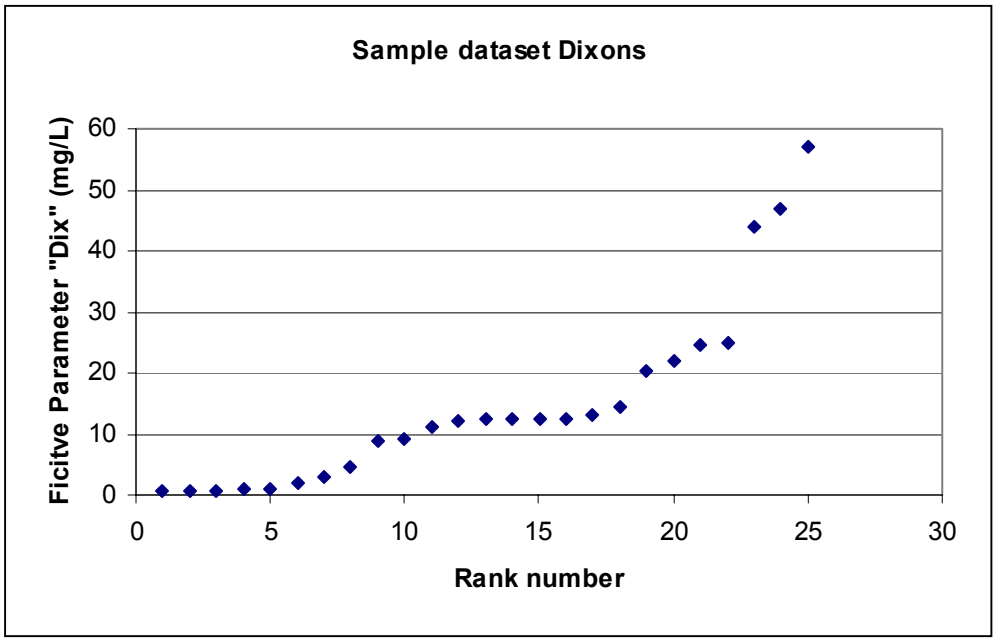


Figure 8.2: Plotted concentrations for Dixons test presented in Table 8.6

Using the data in Table 8.6, to test if the highest value (x_{25}=57) or lowest value (x_1=0.5) is an outlier, Dixon's test criteria are calculated:

$$r_{23} = \frac{x_{25} - x_{23}}{x_{25} - x_3} = \frac{57 - 44}{57 - 0.8} = 0.23 \quad r_1 = \frac{x_3 - x_1}{x_{23} - x_1} = \frac{0.8 - 0.5}{44 - 0.5} = 0.0069$$

The critical value for the highest value at n=25 at 95% is 0.406. Because r < 0.406, then x_{25} is *not* considered an outlier at the 95% level of significance.

The critical value of the lowest value for n=25, 95% is 0.0069. Because r < 0.406, then x_1 is *not* considered an outlier at the 95% level of significance.

Note

Since n=25 in this example, the data set may be analysed using Rosner's test also. In that case the three most extreme values are identified as outliers at the 95% confidence level! This difference occurs because Dixon's test does not compensate for 'masking', e.g. if the most extreme value were removed from the data (leaving 24 data points) Dixon's test would identify the value 46.9 as an outlier also (Dixons's coefficient

becomes 0.477 compared to a tabulated value of 0.413 at the 95% confidence level). Also the then next extreme value of 44 could be identified as outlier if Dixon's test were 'iterative' as is Rosner's test.

Example 8.3 Dixon's test for outliers (2)

Concentration measurements for dissolved oxygen (mg/l) in surface water are as follows (data are sorted in increasing order, n=11):

0.50, 3.77, 3.80, 3.90, 3.92, 4.45, 4.95, 5.44, 5.61, 6.21, 9.51

The data set is plotted in Figure 8.3 and Dixon's test is applied to determine whether the highest and lowest values are outliers:

Since both r11 (=0.6794) and r1 (=0.5779) are larger than the critical value of 0.576 at n=11 and 5% (Table 8.5), both the highest and the lowest value are considered as an outlier.

$$r_{11} = \frac{x_{11} - x_9}{x_{11} - x_2} = \frac{9.51 - 5.61}{9.51 - 3.77} = 0.6794 \quad r_1 = \frac{x_3 - x_1}{x_{10} - x_1} = \frac{3.80 - 0.50}{6.21 - 0.50} = 0.5779$$

At 1% confidence level the test criterion becomes 0.679 (Table 8.5) and therefore the highest value is still considered an outlier (0.6794 > 0.679) whereas the lowest value of 0.5 is not considered an outlier (See calculated HYMOS result at the 1% confidence level in Table 8.7). The test is thus somewhat more sensitive for outlier detection for extreme high values

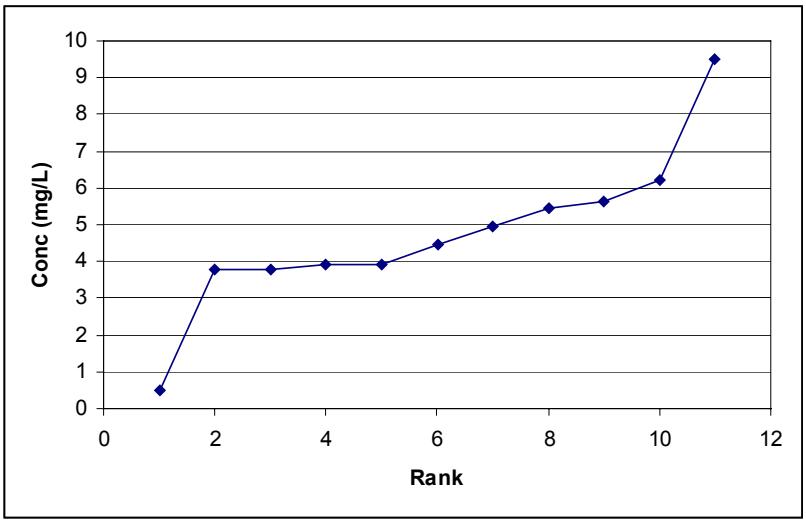


Figure 8.3 Sample data for Dixon's test, in example 8.3

```
Statistical Tests on Data Homogeneity and Randomness
====================================================

One Series Test: Dixon Test for outliers
----------------------------------------

Series code  : Outliers    DIX

Date of first element in series = 01-01-1983
Number of valid data            =  11

Level of significance is 1 Percent

Hypothesis: H0: The outlier belongs to the sample
            H1: The outlier does not belong to the sample

Ranked test data
----------------------
       1    0.5
       2    3.77
       3    3.8
       4    3.9
       5    3.92
       6    4.45
       7    4.95
       8    5.44
       9    5.61
      10    6.21
      11    9.51

Test statistics of lowest value
-------------------------------
Calculated critical value  = 0.5779334
Table critical value       = 0.679

Result: 0.5779334 < 0.679  H0 not rejected

Test statistics of highest value
--------------------------------
Calculated critical value  = 0.6794425
Table critical value       = 0.679

Result: 0.6794425 > 0.679  H0 rejected
```

Table 8.7: Sample data for Dixon's test, example 8.3

Reference

- McBean and Rovers, 1998. Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment, Prentice Hall PTR Environmental Management and Engineering Series.
- Gopal K. Kanji, 1999. 100 statistical tests. SAGE Publications.

8.2.4 HANDLING OF OUTLIERS

After an outlier has been identified, one must decide what to do with it. If an outlier is found, the outlier should be flagged[1] in the data set. For further analysis of the data set, a decision must be made as to what to do with the data value:

- Leave the outlier data value in the data set, and use statistical tests that are not so sensitive to outliers.
- Remove the outlier data value for all further statistical tests.

Caution should be used in removing outliers! Outliers should not be removed on the basis of statistical tests (as Rosner's) only. First of all there is always a chance (the level of confidence of the test, α) that the test incorrectly declares an observation as outlier. Multiple unusually high outliers may indicate that the data should be modeled by a skewed distribution such as the lognormal distribution. Outliers may always be a true representation of a rare event in the field, such as rain, flood, religious bathing, extra factory effluent etc. etc.

In HYMOS the original data point is always kept in the database and marked as such. A corrected value can therefore always be distinguished from the original.

8.2.5 CENSORED DATA

This section is relevant for data near the detection limit of the analytical method. This situation often occurs for trace contaminants analysed by advanced equipment like AAS or GC.

In some environmental sampling situations, the true concentration of the sample being measured may be very near zero, in which case the value may be lower than the measurement limit of detection (LOD). The limit of detection (LOD) is defined as the 'lowest concentration level that can be determined to be statistically different from a blank'.

In this situation, analytical laboratories may report them as not detected (ND) values, zero, or less-than (LT) values. The analytical laboratory may also report the value as <LOD, where LOD is given a numerical value. This is the procedure which is incorporated in SWDES. Data sets containing these types of data are said to be *censored* because the data values below the LOD are not available.

For statistical analysis of the data, the missing data make it difficult to summarize and compare data sets and can lead to biased estimates of means, variances, trends and other values. For analysis of data, a value of ½ LOD is to be utilised. When a value of <LOD is entered in SWDES, this value is automatically converted to ½ LOD for analysis within HYMOS.

8.3 BASIC STATISTICS

8.3.1 PROPERTIES OF THE DATA SET

There are two important properties of the data set which should be known before extensive data analysis begins:

[1] Flagging the data means adding a check mark or 'flag' in the data set next to a specific data point, to indicate that it has been indicated as an outlier. This is not the same as deleting the data, because the flagged data *may* continue to be used in further analysis.

- is the data equidistant or non-equidistant?
- is the data normally distributed?

These properties influence the statistical tests that can be applied in data analysis.

Equidistant data

Data which are equidistant are present in the data set at regular intervals, of e.g. 1 per day, 1 per month, 3 per month, 1 per season, 6 per year, etc. Certain statistical analyses require that data is equidistant. In the event that one or more data values is missing (not measured or not reported) then the missing value must be filled by one or more 'data filling' methods.

Ideally, water quality data should be equidistant. The water quality monitoring being conducted is expected to take place at regular intervals for the identified baseline, trend and surveillance stations. For example it recommended that baseline stations should be monitored a minimum of once every 2 months (see HIS Volume 6). CWC is generally monitoring its water quality stations 3 times per month.

In practice, there may be logistical problems in regularly obtaining a water quality sample for a particular station, or some rivers may only be flowing seasonally. Therefore, many water quality data sets are 'non-equidistant'.

Normally distributed data

Most frequency distributions for *random* observations, when the total number of observations is very large tending to be the same as the population, N, conform to the normal distribution. The distribution is give by the theoretical equation:

$$y = \frac{e^{-(x_i-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \qquad (8.3)$$

where y = frequency of observations

μ = the mean

σ = standard deviation

The standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum x_i^2 - \left(\sum x_i\right)^2/N}{N}} \qquad (8.4)$$

For a normal distribution, 68.3 % of the total observations lie in the range $\mu \pm \sigma$, 95.5% in the range $\mu \pm 2\sigma$ and 99.7% in the range $\mu \pm 3\sigma$. This is illustrated in Figure 8.4.

Many statistical analyses require that the data being studied are normally distributed. Such statistical test are called *parametric* analyses. Unfortunately, most water quality data are NOT normally distributed. Analysis of the data is therefore best done using special *non-parametric* statistical tests.

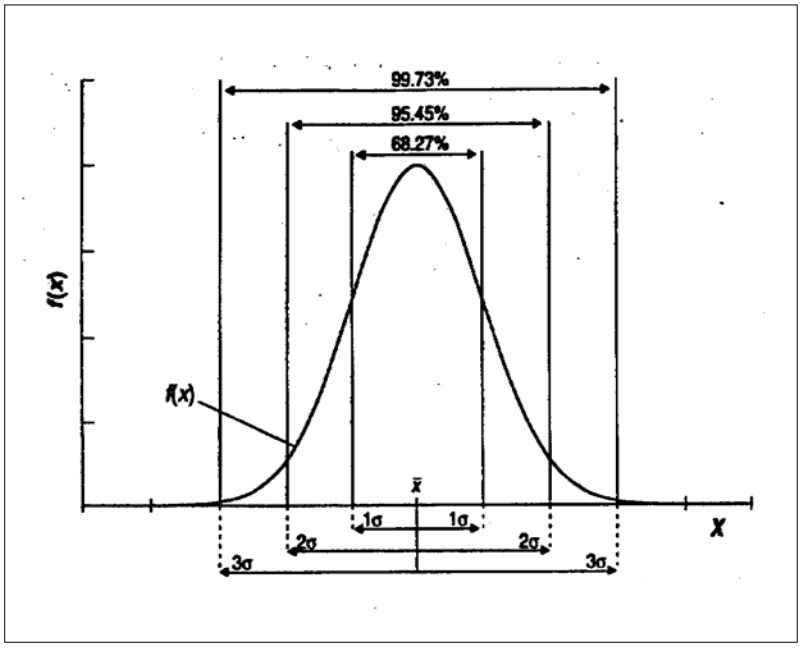


Figure 8.4:
Normal distribution of a set of random observations

The common properties of water quality data, i.e. non-equidistant and non-normally distributed are accommodated in HYMOS, which can analyse both equidistant and non-equidistant data, and which has both parametric and non-parametric statistical tests. In the following chapters, several statistical tests are described for analysis of water quality data. Most of the tests have been selected because they are applicable to the special properties of water quality data.

8.4 SUMMARY STATISTICS

For data set of x_i(i=1, n), HYMOS calculates a number of basic summary statistics which are described below. Calculation of these statistics is usually the first step of any data analysis after data validation procedure. The data sets in Table 8.19 and 8.20 are used to illustrate most of the principles of basic statistics presented in this chapter.

Arithmetic mean

The arithmetic mean, $\bar{x}$, of a set of data is calculated by adding all the observed values and dividing the sum by the total number of observations, n:

$$\bar{x} = (x_1 + x_2 + ...x_n)/n \tag{8.5}$$

or, in different notation:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8.6}$$

where x_1, x_2,.....x_n are the observed values and n is the total number of observations.

The arithmetic mean is the most common measure of the 'central tendency' of a data set, i.e. its center point.

Geometric mean

When there are a few very high values or very low, such as in the cases of bacteriological analysis, the arithmetic mean is not necessarily representative of the central tendency. In such cases the geometric mean, g, is used:

$$g = (x_1 \times x_2 \times x_3 \ldots\ldots\ldots\ldots x_n)^{1/n} \quad (8.7)$$

Median

The median is the middle value of a ranked set of data. If the sample size, n, is an odd number, one-half of the values exceed the median and one-half are less. When n is even, the median is the average of the two middle terms.

Standard deviation

The tendency of observations to cluster (or not to cluster) around the mean value, is measured by *standard deviation*, s.

Standard deviation is calculated as:

$$s = \sqrt{\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots(x_n - \bar{x})^2\}/(n-1)} \quad (8.8)$$

or

$$s = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2/n}{(n-1)}}$$

A small value of *s* signifies that most of the observations are close to the mean value. A large value indicates that the observed values are spread over a larger range. The standard deviation has the same units as the quantity measured, i.e. for a set of concentration observations in mg/l, the standard deviation is also in units of mg/l.

Range

The range gives the minimum and the maximum values in the data set, where:

- Minimum is the minimum value in a set of data observations: $X_{min} = \min(x_1, x_2, x_3, \ldots, x_n)$
- Maximum is the maximum value in a set of data observations: $X_{max} = \max(x_1, x_2, x_3, \ldots, x_n)$

where $x_1, x_2, \ldots\ldots x_n$ are the observed values and n is the total number of observations.

Range = <minimum> to <maximum>

Example 8.4: Basic Statistics in HYMOS

The example data set in Tables 8.19 and 8.20 are used to illustrate calculations of the 'Basic Statistics'. Results of a HYMOS analyses are presented in Table 8.8 and 8.9 (Section 1, summary statistics).

Caution

It should be noted that the formulas used in HYMOS often differ slightly from those in other statistical analysis packages (e.g. Excel). Thus calculated results may vary slightly.

8.4.1 QUANTILES AND PROPORTIONS

Quantiles and proportions are related in that they both are concerned with the percentage of a sample population relative to a specific concentration. Consider the generic cumulative distribution function (cdf) of a set of data, Figure 8.5.

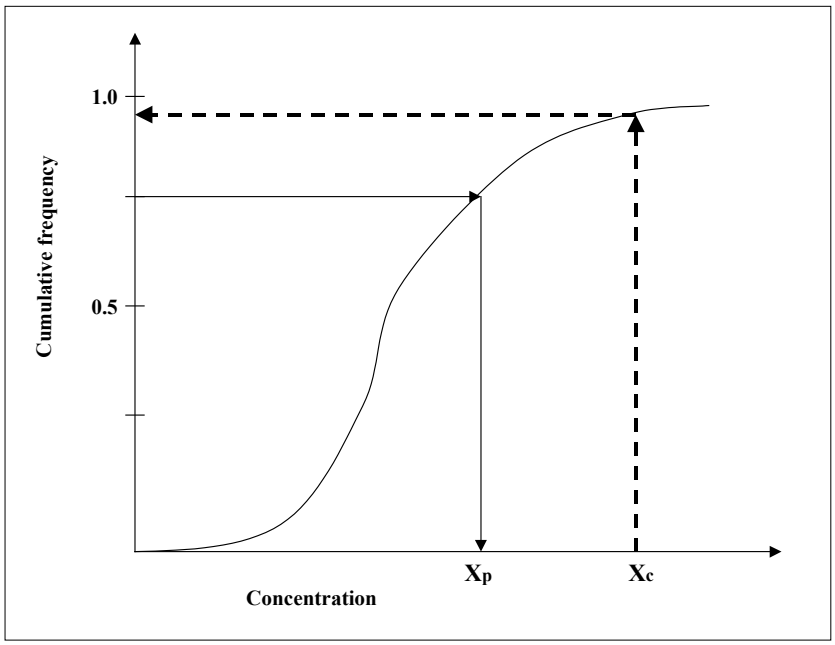


Figure 8.5: Cumulative distribution function (cdf), illustrating quantiles and percentiles

For quantiles (also called percentiles), we are interested in finding: "what is the concentration (x_p) that corresponds to a specific (e.g the 75th) percentile of the data set?". From the vertical (frequency) axis of the cdf plot, you can start at the 75th percentile, read horizontally to the function, then read down to the corresponding concentration (x_p). This is illustrated with the solid line.

For proportions, we are interested in finding: "what proportion of the data is below a specific concentration (x_c)?". From the horizontal (concentration) axis of the cdf plot, you can start at the concentration of interest (x_c), read up to the function, then read horizontally to the corresponding frequency or proportion. This is illustrated with the dashed line.

It is possible to calculate quantiles and proportions without drawing the cdf for a data set, as described below.

Percentiles or quantiles

Introduction and applicability

Percentiles are also called quantiles. Formally, the pth percentile is such that the there is a probability (p), that an observation in the data set will have a concentration less than x_p. For example:

- the median is the 50th percentile; There is a probability of 0.5 that an observation will have a concentration less than the median (or, 50% of the data are less than the median).
- 25th percentile: There is a probability of 0.25 that an observation will have a concentration less than the 25th percentile, or 25 % of data are less than this value
- 75th percentile: There is a probability of 0.75 that an observation will have a concentration less than the 75th percentile, or 75% of data are less than this value

The 25th and 75th percentile are especially important as these values are used in the box and whisker plots (see Section 8.5.3) which are commonly used to present water quality data. HYMOS calculates the 25% and 75% percentiles (quantiles) as part of the 'Basic Statistics' analysis (e.g. Table 8.8, Section 1).

HYMOS also calculates the 10%, 20%, …, 90% percentiles, which are called 'Decile Values' in the 'Basic Statistics' analysis (e.g. Table 8.8 and 8.9, Section 3).

Calculation

The calculation of a specific percentile is fairly straightforward.

For a ranked set of data ($x_1 < x_2 < x_3 < x_n$) the non-exceedance probabilities assigned to the k^{th} largest observation is:

$$p = \frac{k}{n+1}$$

Hence:

$k = p(n+1)$

where: k = the k_{th} largest observation in the ranked data set

p = the non-exceedance probability of the percentile value, in the range $0 \leq p \leq 1$

n = number of data points

and

x_p is the concentration relating to rank *k*

If k is not an integer, x_p is obtained by linear interpolation between the two closest order statistics.

An example calculation is made using the data in Table 8.19 – Nitrate concentrations measured at station 'BasicStat'. A cumulative distribution frequency (cdf) plot of the data is given in Figure 8.6.

To calculate the 90th percentile (x_{90}) for a data set where n= 46

p = 0.9

n = 46

k = 0.9 * (46+1) = 42.3

The 90th percentile ($x_{0.90}$) is the concentration between n=42 (x_{42}) and n=43 (x_{43}) in the ranked data set. Checking the data set in Table 8.20, we see the 90th percentile concentration must be between 14.97 (n=42) and 15.13 (n=43), or 15.0 mg/l. To be even more exact, we could make a proportional estimate (0.3) between the two values, but this is not really necessary.

Various alternatives to calculate proportions are available (see plotting distribution in HYMOS Manual, Chapter 10.2 'fitting distributions'). The method by Chegodayev is presented in the basic statistics and assigns a non-exceedance probability to the k^{th} largest value of:

$$p = \frac{k - 0.3}{n + 0.4}$$

Hence:

k= p(n+0.4)+0.3

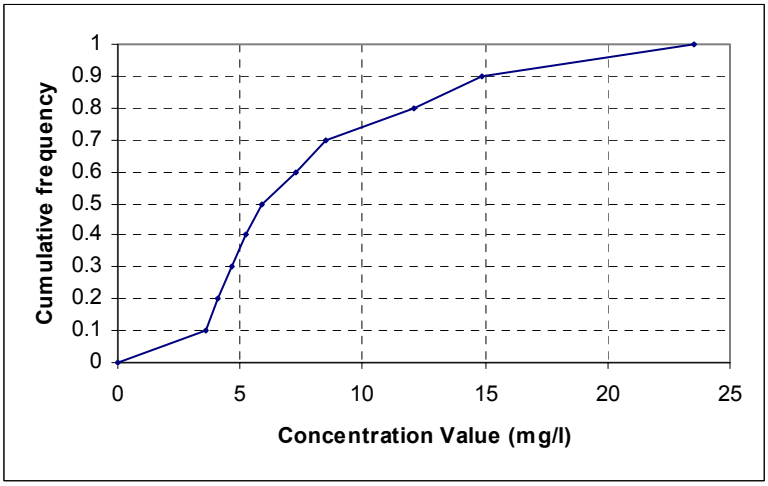


Figure 8.6: Cumulative frequency of 'Nitrate' concentrations from location Basic Statistic (data in Table 8.19)

The 90th percentile value can also be read from the graph, Figure 8.6. Reading horizontally from cdf = 0.9, and then down to the x-axis, we also obtain a value of 15 mg/l as the 90th percentile.

Example 8.5 – Basic statistics in HYMOS – Percentiles and Distribution frequency (1)

An example of percentiles calculation in HYMOS is given, using the same data (Table 8.19). The HYMOS analysis results of basic statistics are given in table form (Table 8.6, Section 3, Deciles) and also as a graph of cumulative frequency and histogram, (Figure 8.7). This graph is helpful in visualising the percentiles. It also clearly shows that the data are not normally distributed.

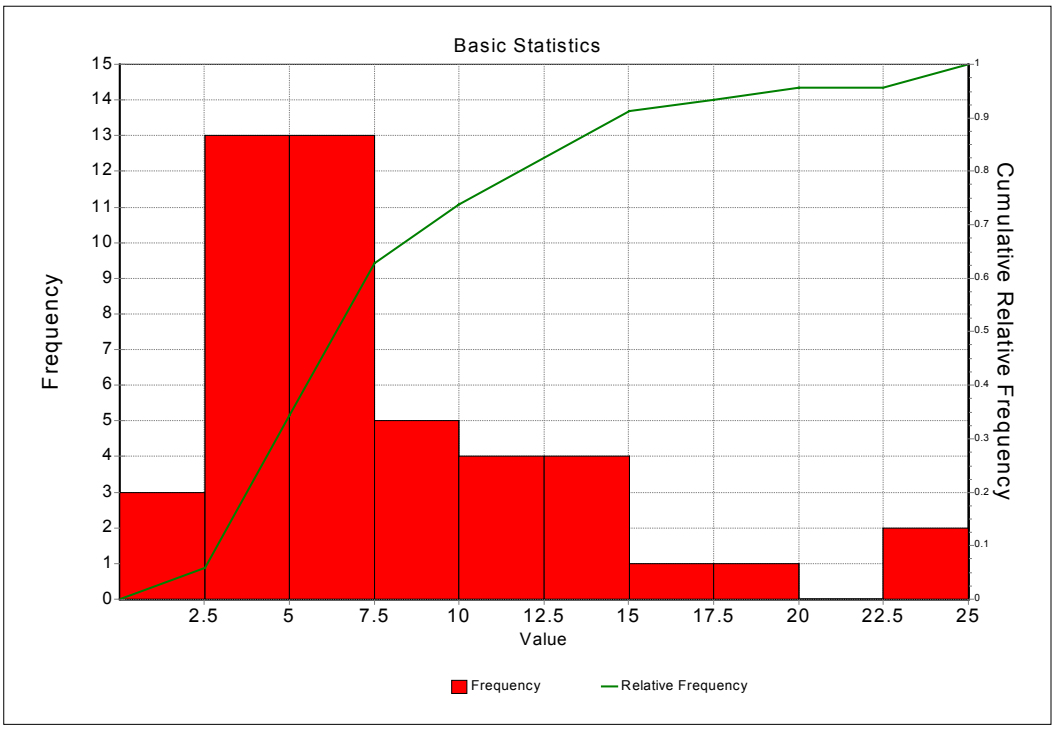


Figure 8.7: Histogram and Cumulative Frequency Distribution for concentrations in Table 8.19, results from HYMOS 'Basic Statistics' analysis (X-axis 'Value' is conc. in mg/l) with 10 classes giving concentration intervals of 2.5 mg/L.

Example 8.6 - Basic statistics in HYMOS – Percentiles and Distribution frequency (2)

An example of percentiles calculation is given, using the Cadmium data of station Neerbeek in Table 8.19. The HYMOS analysis results of basic statistics are given in table form (Table 8.9, Section 3, Deciles) and also as a graph of cumulative frequency and histogram, (Figure 8.6). This graph is helpful in visualising the percentiles. It also clearly shows that the data are not normally distributed.

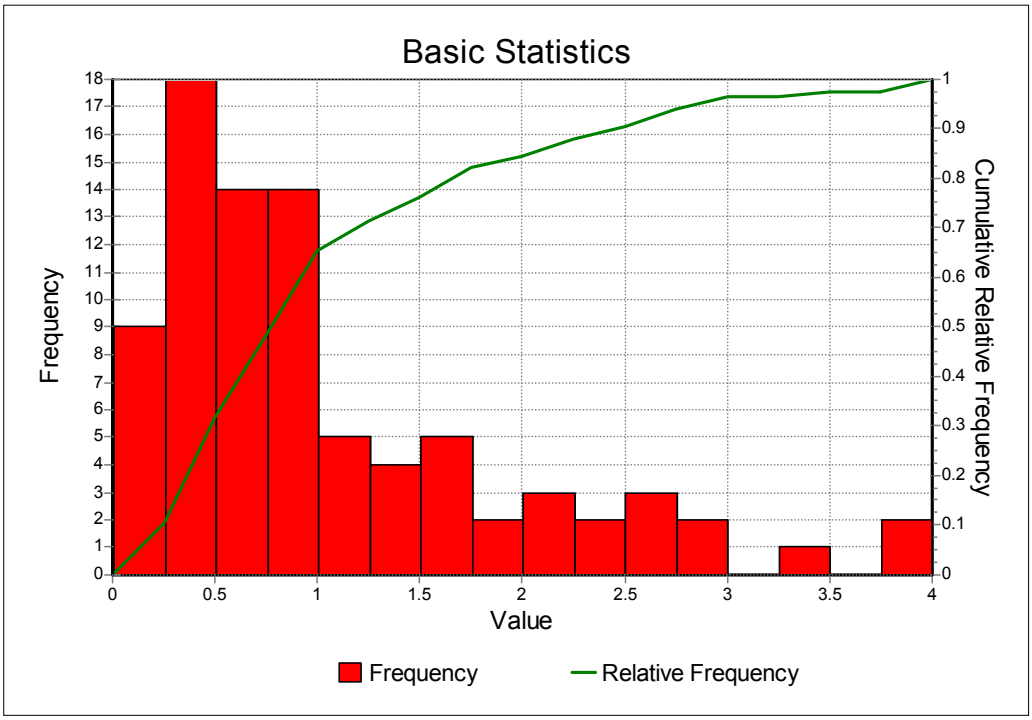


Figure 8.8: Histogram and Cumulative Frequency Distribution for Cd concentration, Station Neerbeek (Table 8.20), results from HYMOS 'Basic Statistics' analysis (X-axis 'Value' is conc. in μg/l) with 16 classes giving concentration intervals of 0.25 μg/L.

From the graph, percentile values can be determined using the cumulative frequency distribution (cdf) line, for example:

Percentile	Cdf (right axis graph)	Conc. reading from graph (bottom axis)	Hymos calculation (from Table 8.9)
90th percentile	0.9	2.5 μg/l	2.54 μg/l
75th percentile	0.75	1.4 μg/l	1.5 μg/l
50th percentile	0.5	0.7 μg/l	0.78 μg/l

Reference

R.O. Gilbert, 1987, Statistical Methods for Environmental Pollution Monitoring, John Wiley & Sons Inc.

Proportions

Introduction

In proportions, the goal is to estimate the proportion of a data set that is less than (or more than) a specific concentration (x_c). Such an analysis is important if water quality regulations specify that the proportion of the population exceeding a specified concentration x_c, (upper limit) must be less than some specified value.

Calculation

A non-parametric calculation of proportions is quite simple, based on ranked data values. To estimate P_{xc}, the proportion of the population exceeding x_c, we calculate:

$P_{xc} = u/n$

where:
n = no. of observations
u = no. of observation exceeding x_c

An simple example calculation is made using the data in Table 8.19 – Nitrate concentrations to illustrate:

- What is the proportion of the data which exceed 15 mg/l ?
- What is the proportion of the data which a less than 10 mg/l?

From the ranked data set, we know that n= 46 and 4 observations exceed 15 mg/l, while 12 observations exceed 10 mg/l. Therefore:

$P_{15} = u/n = 4/46 = 0.087$

$P_{10} = u/n = 10/46 = 0.217$

The proportion of data *exceeding* 15 mg/l is 0.087 (8.7%).

The proportion of data *less than* 15 mg/l is (1-0.087) or 0.913 (91.3%)

The proportion of data *exceeding* 10 mg/l is 0.217 (21.7%).

The proportion of data *less than* 10 mg/l is (1-0.217) or 0.783 (78.3%)

These same results can also be obtained from using the plotted data in Figure 8.6. For example, by starting on the X-axis at 10 mg/l, reading up to the cdf and then to the left axis, we find that the proportion of data less than 10 mg/l is ~0.75. The graphical analysis is somewhat less accurate than the calculation.

Example 8.7 – Basic Statistics in HYMOS – Proportions (1)

The data in Table 8.19, concentrations at location BasicStat, have been analysed with HYMOS. As part of the 'Basic Statistics' results, HYMOS also calculates proportions at a certain number of concentration class limits (usually 10). This is shown in Table 8.6, Section 4, 'Cumulative distribution frequency and histogram'.

For the concentrations given in 'Upper class limit', the proportion (probability) of data less than this concentration is given. For example, the proportion of data less than 10 mg/l is 0.739. For estimating proportions other than those listed as class limits, the cumulative frequency plot of the data (Figure 8.7) can be used.

Example 8.8 – Basic Statistics in HYMOS – Proportions (2)

The data in Table 8.20, cadmium concentrations at location Stowa1, have been analysed with HYMOS. As part of the 'Basic Statistics' results, HYMOS also calculates proportions at a certain number of concentration class limits (usually 10 but in this case 16). This is shown in Table 8.7, Section 4, 'Cumulative distribution frequency and histogram'.

For the concentrations given in 'Upper class limit', the proportion (probability) of data less than this concentration is given. For example, the proportion of data less than 2.5 μg/l is 0.905. For estimating proportions other than those listed as class limits, the cumulative frequency plot of the data (Figure 8.8) can be used.

Reference

R.O. Gilbert, 1987, Statistical Methods for Environmental Pollution Monitoring, John Wiley & Sons Inc.

```
Series code = BasicStat   NO3
 First year  =  1983
 Last year   =  2000

 Actual values are used

1.     Basic Statistics of series BasicStat   NO3
      ========================================


      Mean              =      7.9602
      Median            =      5.9050
      Mode of classes   =      3.7585
      Standard deviation=      5.2606
      Variance          =     27.6740
      Skewness          =      1.4062
      Kurtosis          =      4.4432
      Range             =      1.8500 to     23.5300
      25% Quantile      =      4.2475
      75% Quantile      =     11.2000
      Number of elements=          46

2.              95% Confidence Intervals
           6.3980 <     Mean     <         9.5224
           9.7291 <   Variance   <        38.9165
           3.8074 < 25% Quantile <       5.0252
           7.4897 < 75% Quantile <      14.8310

3.     Percentiles
     Decile                Value        Value
                                        Chegadayev

        1                  2.816569    2.881928
        2                  3.700831    3.773879
        3                  4.585093    4.665830
        4                  5.469354    5.557780
        5                  6.353616    6.449731
        6                  7.237877    7.341682
        7                  9.106361    9.396246
        8                 11.755300   12.142650
        9                 13.504610   15.153950

4.     Cumulative frequency distribution and histogram

     Upper class limit   Probability      Probability   Number of elements
                                          Chegadayev

           .010000         .000000         .000000            0.
          2.509000         .065217         .058190            3.
          5.008000         .347826         .338362           13.
          7.507000         .630435         .618535           13.
         10.006000         .739130         .726293            5.
         12.505000         .826087         .812500            4.
         15.004000         .913043         .898707            4.
         17.503000         .934783         .920259            1.
         20.002000         .956522         .941810            1.
         22.501000         .956522         .941810            0.
         25.000000        1.000000         .984914            2.
                                                              0
```

Table 8.8: Results in HYMOS 'Basic Statistics' analysis of Nitrate data (Table 8.19)

```
Series code = STOWA 1     QCD
 First year  =  1983
 Last year   =  1996

 Actual values are used

1.     Basic Statistics of series STOWA 1   QCD
      ========================================

      Mean              =      1.0825
      Median            =       .7800
      Mode of classes   =       .6425
      Standard deviation=       .8893
      Variance          =       .7908
      Skewness          =      1.4671
      Kurtosis          =      4.7885
      Range             =       .0500 to      4.0000
      25% Quantile      =       .4925
      75% Quantile      =      1.5000
      Number of elements=          84

2.              95% Confidence Intervals
            .8895 <     Mean     <       1.2755
            .5962 <   Variance   <       1.0996
            .6500 <    Median    <       1.0000
            .3547 < 25% Quantile <        .6000
           1.0970 < 75% Quantile <       2.0530

3.     Percentiles
     Deciles for data range
     Decile               Value

        1                  .245000
        2                  .400000
        3                  .500000
        4                  .650000
        5                  .780000
        6                 1.000000
        7                 1.100000
        8                 1.700000
        9                 2.500000

     Deciles for classes
     Decile               Value        Value
                                       Chegodayev
        1                  .205921     .210911
        2                  .382632     .386374
        3                  .528558     .530381
        4                  .657692     .658604
        5                  .786827     .786827
        6                  .981071     .979379
        7                 1.220893    1.217507
        8                 1.788000    1.773781
        9                 2.538501    2.519541
```

Table 8.9: Results of HYMOS 'Basic Statistics' analysis of STOWA1 Cd data (Table 8.20)
Continued on next page)

4. Cumulative frequency distribution and histogram

Upper class limit	Probability	Probability Chegodayev	Number of elements
.050000	.011765	.008294	1.
.445000	.235294	.233412	19.
.840000	.541176	.541469	26.
1.235000	.705882	.707346	14.
1.630000	.776471	.778436	6.
2.025000	.835294	.837678	5.
2.420000	.882353	.885071	4.
2.815000	.941176	.944313	5.
3.210000	.952941	.956161	1.
3.605000	.964706	.968009	1.
4.000000	.988235	.991706	2.
			0

8.4.2 CONFIDENCE INTERVALS

Confidence intervals are used to indicate the uncertainty associated with the estimated value of a particular parameter. The following confidence intervals are commonly calculated, and are part of the 'Basic Statistics' function in HYMOS (Table 8.8 and Table 8.9):

- confidence interval about the mean
- confidence interval about the variance
- confidence interval about the 25^{th} and 75^{th} quantile

In all cases, HYMOS calculates the 95% confidence interval, e.g. there is 95% confidence that the mean/variance/quantile is within the given interval.

Confidence interval about the mean

Introduction

The confidence intervals about the mean gives the uncertainty associated with the calculated value of the mean for a particular data set. Statistical confidence limits define an interval such that there is a specified probability (1-α), that a parameter will be contained in the interval (e.g. 95% for α=0.05). The width of the confidence interval increases when:

- more confidence is requested, e.g. 99% (α=0.01) instead of 95% (α=0.05);
- the standard deviation is higher
- the sample size is smaller
- serial correlation is present

Calculation

For normally distributed data, the confidence interval about the mean is defined as (see HIS Manual, Volume 2, Chapter 3):

$$\bar{x} - t_{\alpha/2,\nu} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,\nu} \times \frac{s}{\sqrt{n}} \quad \text{with :} \quad \nu = n - 1 \tag{8.9}$$

or:

$$\mu_{lcl} = \bar{x} - t_{\alpha/2,\nu} \times \frac{s}{\sqrt{n}}$$

$$\mu_{ucl} = \bar{x} + t_{\alpha/2,\nu} \times \frac{s}{\sqrt{n}}$$

Where: μ_{lcl} = lower confidence limit of mean

μ_{ucl} = upper confidence limit of mean

$\bar{x}$ = mean

α = significance level

$t_{\alpha,\nu}$ = critical value of student - t distribution for a confidence level of 1-α, and ν = n-1 degrees of freedom

s = standard deviation

n = number of observations

Note that the sampling distribution of the mean is very nearly normal for n > 30, even when the population is non-normal.

The confidence intervals about the variance are derived from:

$$\frac{\nu\, s^2}{X^2_{\nu,1-\alpha/2}} \leq \sigma^2 \leq \frac{\nu\, s^2}{x^2_{\nu,\,\alpha/2}} \quad \text{with :} \quad \nu = n - 1 \tag{8.10}$$

where: $\chi^2_{\nu,\,1-\alpha/2}$, $\chi^2_{\nu,\,\alpha/2}$ upper and lower critical values of χ^2 – distribution for a confidence level of 1 - α and ν = n – 1 degrees of freedom

8.5 PRESENTATION

For preparation of Water Quality yearbooks a number of data presentations are recommended to illustrate the important aspects of the water quality status. The presentations (graphs) can all be made using HYMOS. Some of the simple presentations can also be made in the Data Entry Software (SWDES, water quality). A description and example of each presentation method is given below.

8.5.1 TIME SERIES

The time series graph is the first and most important presentation of the data that should be made.

The time series is a plot of the measured concentration (Y-axis) as a function of time (X-axis). A time series plot is usually made for *one* parameter, at *one* location, as shown in Figure 8.9, which shows the measured values for cadmium at the location Stowa1 for the period 1983-97 (data in Table 8.20).

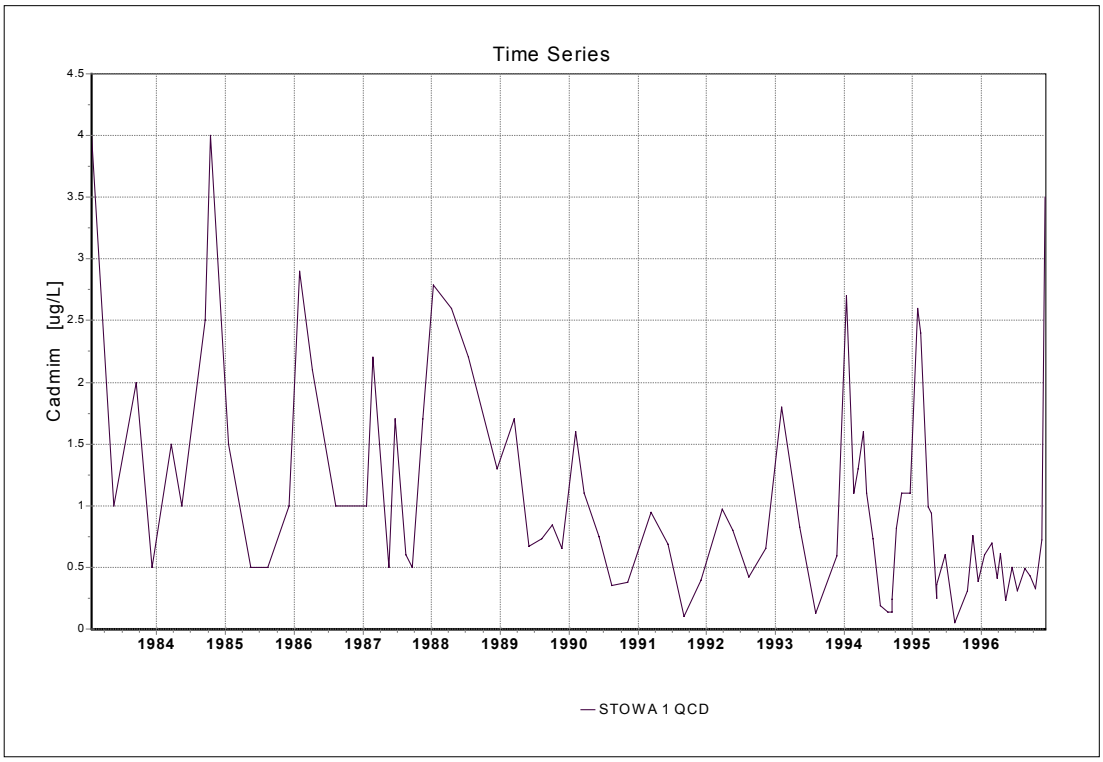


Figure 8.9: Time series of cadmium at location STOWA1

From a visual analysis of a time series plot, you can get an idea about the variability in the data, extreme high and low values, and possible trends. This time series shows that the cadmium concentration is highly variable, and that there seems to be a general decreasing trend. The statistical significance of this trend should be further investigated, as discussed in Sub-section 8.6.2.

For some special types of data analyses, it may be desirable to plot results for the *same parameter* from *2 different locations* on one graph. This would show whether the parameter concentration is showing the same concentration pattern at two different locations, maybe indicating a correlation between the locations. If a correlation seems to exist based on visual analysis of the time series, then further statistical analysis could be conducted.

Another special analysis is to plot results for *two different parameters* from *one location*. This would show whether there is a possible correlation between two different parameter concentrations. If a correlation seems to exist, based on visual analysis of the time series, then further statistical analysis could be conducted.

8.5.2 LONGITUDINAL PLOTS

A longitudinal plot is made to show the concentration of a water quality parameter at different locations along a river. The plot depicts the measured concentration (Y-axis) as a function of distance along the river (X-axis). This distance is often called ‘boating distance’, measured in km from the mouth of the river.

8.5.3 BOX AND WHISKERS PLOT

Introduction

A box and whiskers plot is a useful way of summarising the range and spread of data for a given parameter. The box and whiskers plot may be made for one or more parameters or stations, using results for a period of time. The selected data will be summarised over a selected period, which can be a season, year or several years. The box and whiskers plots present a useful and quick graphical summary of data from different locations or from one time period to another.

The box and whiskers plot depicts the following statistics for the data:

- minimum and maximum(the end of the 'whiskers' – this shows the full range of data in the period of interest)
- the 25th and 75th percentile (lower and upper end of the 'box' – this gives an indication of the spread)
- mean of the data in the period of interest (+)
- median of the data in the period of interest (-)

Yearly box and whisker plot

The purpose of the yearly box and whiskers plot is to show the variation in a water quality parameter by comparing the results for different years. The plot is made for data of one *water quality parameter* at *one location*. For each year, the statistics are calculated and plotted. An example is shown below in Figure 8.10 for Cd data in Table 8.20 (Station STOWA1).

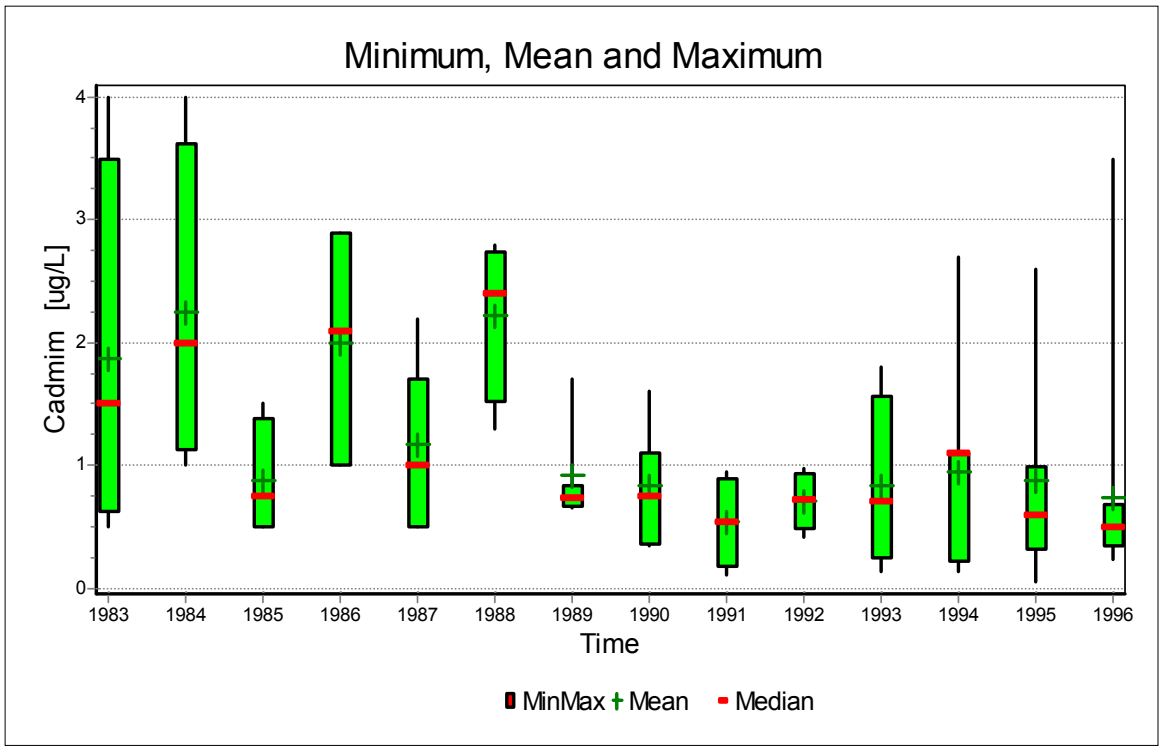


Figure 8.10: Yearly box and whisker plot for Cd at Station STOWA1 (1983-1996)

Seasonal box and whiskers Plot

The purpose of the seasonal box and whiskers plot is to show the variation in a water quality between seasons. The plot is made for data of one *water quality parameter* at *one location*. The data is analysed per season, for example: winter, summer, monsoon. The beginning and end date of each season has to be clearly defined (this can be done in HYMOS). For each season, the available data are used to calculate the statistics and make the box and whiskers plot. An example is shown in Figure 8.11, showing the Cd concentrations at location STOWA1 in three different seasons: summer, monsoon and winter. The HYMOS summary report is given in Table 8.10 and 8.11.

The summary report on box and whiskers analysis in HYMOS for the processing period 1-11-1982 to 28-02-1988 is given (actual data set is longer, see Table 8.20). Note, the effect of the combination of processing period and seasons definition (summer: 28-2/31-5, monsoon: 1-6/31-10, winter: 1-11/28-2) in this example. The value of 4.00 (19-01-1983, see Table 8.20) is considered part of the winter season of 1982! Values for summer and monsoon are missing for 1982. If the processing period starts somewhere in 1983, the value 4.00 measured on 19-01-1983 is omitted from analysis. Other software packages may not handle seasons correctly.

```
Computation of minimum, maximum, mean, standard deviation, median and percentiles

Series ID: STOWA 1     QCD

MINIMUM
        28-02/31-05    31-05/01-11    01-11/28-02    Year
1982      -999.990       -999.990          4.000          4.000
1983         1.000          2.000          0.500          0.500
1984         1.000          2.500          1.500          1.000
1985         0.500          0.500          1.000          0.500
1986         2.100          1.000          1.000          1.000
1987         0.500          0.500          1.700          0.500
1988      -999.990       -999.990       -999.990       -999.990

SUMMARY      0.500     0.500     0.500     0.000

MAXIMUM
        28-02/31-05    31-05/01-11    01-11/28-02    Year
1982      -999.990       -999.990          4.000          4.000
1983         1.000          2.000          0.500          2.000
1984         1.500          4.000          1.500          4.000
1985         0.500          0.500          2.900          2.900
1986         2.100          1.000          2.200          2.200
1987         0.500          1.700          2.790          2.790
1988      -999.990       -999.990       -999.990       -999.990

SUMMARY      2.100     4.000     4.000     4.000

MEAN
        28-02/31-05    31-05/01-11    01-11/28-02    Year
1982      -999.990       -999.990          4.000          4.000
1983         1.000          2.000          0.500          1.167
1984         1.250          3.250          1.500          2.100
1985         0.500          0.500          1.950          1.225
1986         2.100          1.000          1.600          1.575
1987         0.500          0.933          2.245          1.298
1988      -999.990       -999.990       -999.990       -999.990

SUMMARY      1.100     1.600     1.954     0.765
……continued for Standard deviation, median, 25% and 75% fraction.
```

Table 8.10: Example of HYMOS report for min-mean-max series, with 3 seasons

	QC+ Maximum	QC- Minimum	QC* Average	QC/ Median	QC1 25% Quartile	QC2 75% Quartile
28-02	2.6	0.23	1.014783	0.97	0.5	1.3
1-05	4	0.05	0.7799999	0.5	0.31	0.8025
01-11	4	0.38	1.47	1.1	0.65	2.2

Table 8.11: HYMOS example of View for min-mean-max series. These values are used to construct the seasonal summary box and whiskers (Figure 8.11)

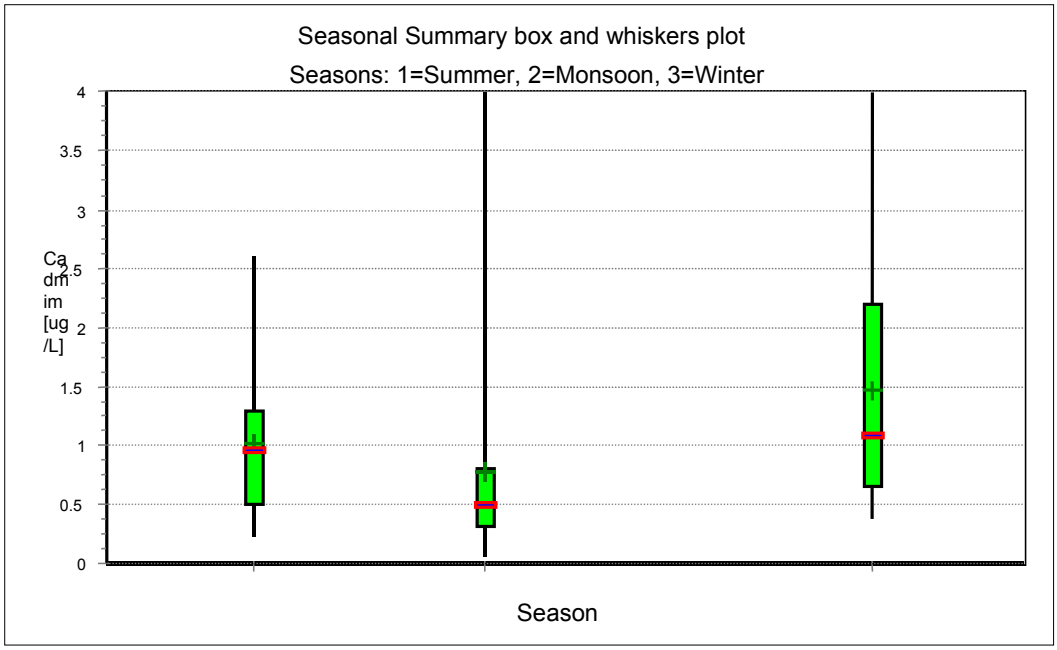


Figure 8.11: Seasonal box and whiskers summary plot for data from Table 8.20 summary statistics for all summer, monsoon and winter seasons in the period of 1983-1996.

Location-wise box and whiskers plot

The purpose of the location-wise box and whiskers plot is to show the variation in a water quality parameter at several different locations, i.e. How does the concentration vary from one location to the next. The plot is made for data of *one water quality parameter* for one or possibly more years (the number of years should be the same for different locations, otherwise the data are not comparable). The data is analysed per location. For each location, the statistics are calculated and plotted.

8.5.4 STANDARDS COMPARISON

For many of the water quality parameters, it is useful to compare the measured concentrations with Indian or international water quality standards. Such standards are often defined for drinking water or irrigation water purposes.

Comparison of the measured concentrations with standards can be done graphically, by plotting the time series of the data together with the standard, which is shown as a horizontal line at the acceptable limit. In this manner, it is very easy to see which measurements are above the acceptable standard and get a feel for how often this occurs.

In Figure 8.12, the cadmium concentrations at location STOWA1 are plotted together with the water quality standard for cadmium, 0.5 μg/l. From this graphical presentation of data, it is easy to see that from 1983 up until 1990, almost *all* measurements were higher than the water quality standard. In 1996, most measurements are below the standard, but some higher values were recorded.

For an more exact analysis of the water quality results compared to the standard, use of the percentiles or proportion analyses can be made (Sub-section 8.4.1).

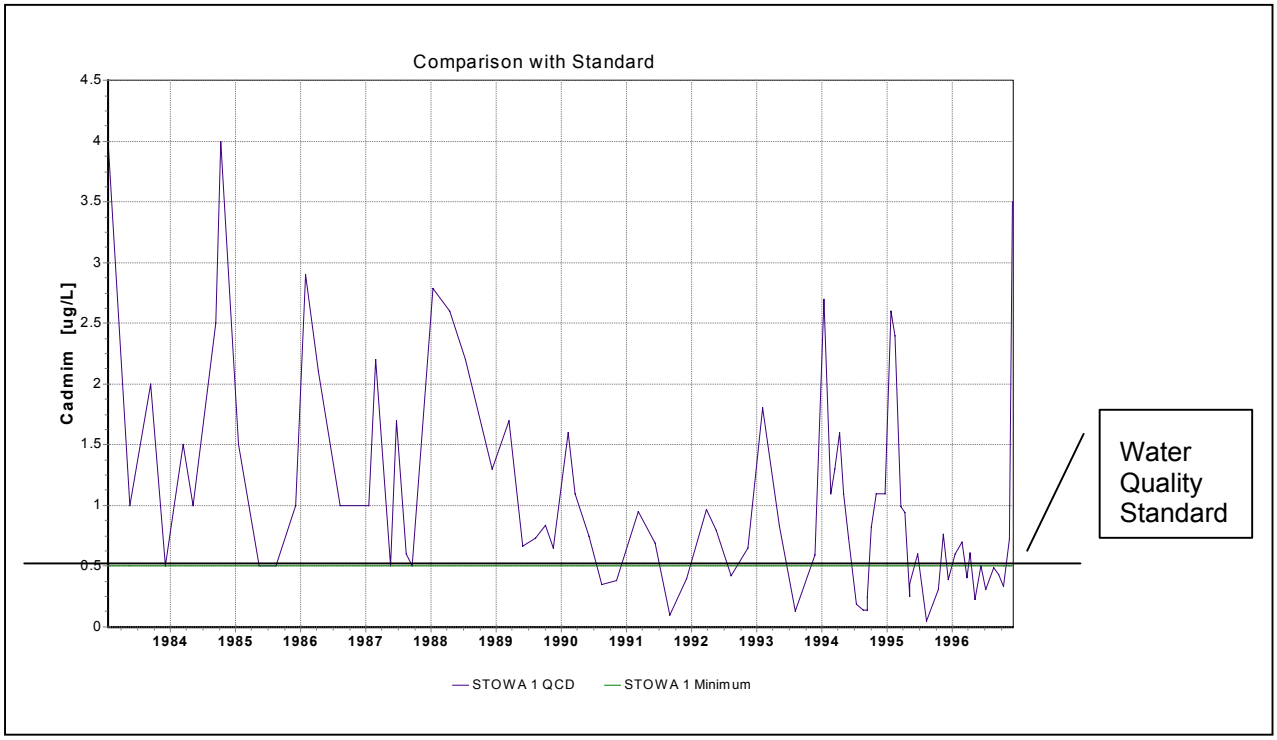


Figure 8.12: Comparison with Standard, Cadmium at location STOWA1

8.6 TRENDS

An important objective of many water quality monitoring programmes is to detect changes or trends in the water quality over time. By trend we mean a noticeable change in the measured concentration of a water quality parameter, usually over a period of a few years. The purpose may be to see if pollution in water is increasing, perhaps due to growth of industry, or to check if the water quality is improving, perhaps following new waste water control programmes.

8.6.1 TYPES OF TRENDS

Different types of trends exist and are discussed in this and the next chapter:

- linear trend: This is a common type of trend which we want to analyse, showing an increase or decrease in a water quality parameter concentration over time (Sub-section 8.6.2).
- step trend: This type of trend shows a sudden and long-lasting change in the concentration of a water quality parameter (Section 8.7). Such a change can come from several reasons:
  - a new industrial effluent in a river (could cause a sudden increase in a pollutant concentration);
  - a new wastewater treatment facility (could cause a sudden decrease in a pollutant concentration); or

  - a new procedure in analysis (different laboratory conducting the analysis, new analytical method or new equipment being used in a laboratory).
  - The method to test for a step trend is described in Section 8.7, Comparing populations.
- cycles: This type of 'trend' shows a cyclical pattern in the water quality concentration. Cycles may be caused by seasonal fluctuations in a natural process or human activity which affects the concentration of a particular parameter. Cycles are not real trends because they do not indicate a long-term change. Special care must be taken in analysing trends if cycles are present in the data (Sub-section 8.6.2, Seasonality)

Some data will show no real trend, but will simply have random variations in measured concentration over time.

8.6.2 ANALYSIS OF A LINEAR TREND

Most often in the analysis of trend, we are interested if there is a *linear trend*, i.e. a slow but steady change in the water quality over the period of several years. Assessing if such a trend exists can often be difficult, due to a number of complicating factors which are characteristic of water quality data, as described briefly below.

Seasonality

The variation added by seasonal or other cycles makes it difficult to detect long-term trends. This problem can be solved by:

- removing the cyclic effects before applying other tests to study the trend.
- using a trend test which is unaffected by cycles

For avoiding the complications caused by seasonality in water quality data, we recommend using the non-parametric Seasonal Kendall test for analysis of a linear trend. This test is available in HYMOS, and is unaffected by seasonality or other cycles.

Flow correction

The detection of trends in river water quality is difficult when the concentrations are related to river discharge, which is often the case. Some compensation for the flow should therefore be made. One method is to convert the concentration values to instantaneous *loads* (flow $\times$ concentration). The trend analysis can then be made using the loads and not concentrations. Alternatively, a correction for the effect of flow on concentration can be made, comprising several steps:

1. The correlation between concentration and flow must be established, with a regression equation for the concentration-flow relationship
2. The *residuals* of the concentration data are then determined.
3. The residuals can then be checked for trend.

Both methods require that flow data are available at the time of each concentration measurement.

Correlated data

Data which are measured within very short time intervals are likely to be correlated, i.e. they are not independent of one another. Most statistical tests require uncorrelated (independent) data observations. The standard water quality data is most likely uncorrelated, because the sampling frequency is not high.

Trend estimation

Reference is made to Annex II, Chapter 6, for a description of the linear trend test, where the significance of the slope of the trend line is investigated.

Example 8.9 – Linear Trend in HYMOS

The cadmium concentrations from station STOWA1 (Table A.3) are used to illustrate calculation of a trend with linear regression.

```
Statistical Tests on Data Homogeneity and Randomness
 ====================================================

 One Series Test
 ---------------

 Series code: STOWA 1     QCD

 Date of first element in series= 1983  1 19  0  0
 Number of data                 =   84

 Test for Significance of Linear Trend
 -------------------------------------
 Intercept parameter     (=b1)    =     1.807
 Slope parameter         (=b2)    =-.2345E-03
 St.dev. of b2           (=sb2)   = .3681E-02
 St.dev. of residual     (=se)    = .8181E+00
 Test statistic [t]   (abs.value) =      .064
 Degrees of freedom               =        82
 Prob(t.le.[t])                   =      .525
 Hypothesis: H0: Series is random
             H1: Series is not random
             A two-tailed test is performed
             Level of significance is  1.00 percent
             Critical value for test statistic 2.637
 Result:     H0 not rejected
```

Table 8.12: Results of HYMOS calculation on Linear Trend analysis

The calculated slope is: m= - 0.2345E-03 per day

This slope is tested with the t-statistic at 1% significance. The null hypothesis is not rejected, thus the slope is not significant.

Seasonal Kendall Slope Test

Application

The seasonal Kendall test is an ideal test to use for analysing trends in water quality data. It is not sensitive to many of the complicating factors typically present. It can be used when seasonal cycles are present and may even be used when missing data or tied data are present in the series. The validity of the test does not depend on the data being normally distributed (i.e. it is a non-parametric test). The test computes both the value and the significance of the trend.

For these reasons, the seasonal Kendall slope test is recommended above the more commonly used Linear Regression.

Theory

The test consists of computing the Mann-Kendall test statistic *S* and its variance *VAR(S),* separately for each season. These seasonal statistics are then summed, and a *Z* statistic is computed. The normal distribution is used to test for a statistically significant trend.

The Mann-Kendall statistic *S* is computed for each season with:

$$S_i = \sum_{k=1}^{ni=1} \sum_{l=k+1}^{ni} \text{sign}(x_{il} - x_{ik}) \tag{8.11}$$

where: n = number of years
l, k = years
x = data values
i = season number

If $x_{il} > x_{ik}$ then the sign = 1, if $x_{il} < x_{ik}$ then the sign = -1 else the sign = 0

The variance of *S* is computed for each season as follows:

$$VAR(S_i) = \frac{1}{18}[n_i(n_i-1)(2n_i+5) - \sum_{p=1}^{gi} t_{ip}(t_{ip}-1)(2t_{ip}+5) - \sum_{q=1}^{hi} u_{iq}(u_{iq}-1)(2u_{iq}+5)] + \frac{\sum_{p=1}^{gi} t_{ip}(t_{ip}-1)(t_{ip}-2)\sum_{q=1}^{hi} u_{iq}(u_{iq}-1)(u_{iq}-2)]}{9n_i(n_i-1)(n_i-2)} + \frac{\sum_{p=1}^{gi} t_{ip}(t_{ip}-1)\sum_{q=1}^{hi} u_{iq}(u_{iq}-1)}{2n_i(n_i-1)} \tag{8.12}$$

where: g_i = the number of groups of tied data in season i
t_{ip} = the number of tied data in p^{th} group of season i
h_l = the number of sampling times in season i that contain multiple data
u_{ip} = the number of multiple data in the q^{th} time period of season i

When S_i and $VAR(S_i)$ are computed, they are summed across the K seasons:

$$S' = \sum_{i=1}^{K} S_i \qquad VAR(S') = \sum_{i=1}^{K} VAR(S_i) \tag{8.13}$$

Finally the *Z* statistic is computed from:

$$Z = \frac{(S'-1)}{\sqrt{VAR(S')}} \quad \text{If } S' > 0$$

$$Z = 0 \quad \text{If } S' = 0 \tag{8.14}$$

$$Z = \frac{(S'-1)}{\sqrt{VAR(S')}} \quad \text{If } S' < 0$$

HYMOS considers the following hypothesis:

H_0 : The populations from which the series has been drawn has no trend,

H_1 : The populations has a trend.

Together with the seasonal Kendall test, the seasonal Kendall slope estimator is computed. The N_i individual slope estimates for the i^{th} season is computed form:

$$Q_i = \frac{x_{il} - x_{ik}}{l - k} \tag{8.15}$$

This is done for each season. The individual slopes are ranked and the median values are computed for each season: this is the seasonal Kendall slope estimator N′. A confidence interval around the true slope is obtained by using the normal distribution, and:

$$C_\alpha = z_{1-\alpha/2}\sqrt{VAR(S')}$$

$$M_1 = \frac{(N' - C_\alpha)}{2} \quad \text{and} \quad M_2 = \frac{(N' + C_\alpha)}{2} \tag{8.16}$$

The lower and upper confidence limits are the M_1^{th} largest and the $(M_1+1)^{th}$ largest values of the N' ordered slope estimates.

References:

- R.O. Gilbert, 1987, Statistical Methods for Environmental Pollution Monitoring, John Wiley & Sons Inc..

Example 8.10 – Seasonal Kendall slope estimate

The data set of cadmium concentrations at location STOWA1 (Table 8.20) is again used for illustration. We have already seen from the plotted data (Figure 8.12) that there seems to be a decreasing trend, based on visual inspection of the data. The linear trend estimate also indicated that there is a significant trend.

The Seasonal Kendall Slope Analysis in HYMOS is applied to the data set, for the season definition which can be set by the user. In this example, there are 3 seasons defined (The same seasons as used for the seasonal box and whiskers analysis: Summer, monsoon, winter). Results are given in Table 8.13.

For the defined seasons, the test calculates an overall slope (trend) of -0.00016 per day equivalent to - 0.062 μg Cd/l per year.

Furthermore, the trend is tested for significance at the 5% level. The trend is found to be significant.

```
Statistical Tests on Data Homogeneity and Randomness
====================================================
One Series Test
---------------
Series code : STOWA 1     QCD
Date of first element in series = 19-01-83
Number of data              =  84

Seasonal Kendall Slope Estimator
--------------------------------------------
Number of Seasons           =  3
 Season 1: 28-02 to 31-05   # data 239   Median slope -.00018  Mann-Kendall Statistic -90
 Season 2: 31-05 to 01-11   # data 459   Median slope -.00021  Mann-Kendall Statistic -207
 Season 3: 01-11 to 28-02   # data 357   Median slope -.00008  Mann-Kendall Statistic –31
Overal Median Slope         =  -.00016 (per day)
```

```
Seasonal Kendall Slope Test
-----------------------------------------------
Season 1: Number of tied values (g) 3
   Tied Value (t): 1 = 1     measured 2 times
   Tied Value (t): 2 = 0.5     measured 2 times
   Tied Value (t): 3 = 1.1     measured 2 times
Season 2: Number of tied values (g) 5
   Tied Value (t): 1 = 0.5     measured 3 times
   Tied Value (t): 2 = 0.6     measured 2 times
   Tied Value (t): 3 = 0.73     measured 2 times
   Tied Value (t): 4 = 0.14     measured 2 times
   Tied Value (t): 5 = 0.31     measured 2 times
Season 3: Number of tied values (g) 3
   Tied Value (t): 1 = 1     measured 2 times
   Tied Value (t): 2 = 0.65     measured 2 times
   Tied Value (t): 3 = 1.1     measured 3 times
Season 1: Number of years with multiple data (h) 5
   Period (u): 1984 has 2 data values
   Period (u): 1992 has 2 data values
   Period (u): 1994 has 3 data values
   Period (u): 1995 has 4 data values
   Period (u): 1996 has 3 data values
Season 2: Number of years with multiple data (h) 8
   Period (u): 1984 has 2 data values
   Period (u): 1987 has 3 data values
   Period (u): 1989 has 3 data values
   Period (u): 1990 has 2 data values
   Period (u): 1991 has 2 data values
   Period (u): 1994 has 6 data values
   Period (u): 1995 has 3 data values
   Period (u): 1996 has 5 data values
Season 3: Number of years with multiple data (h) 6
   Period (u): 1987 has 2 data values
   Period (u): 1990 has 2 data values
   Period (u): 1993 has 2 data values
   Period (u): 1994 has 2 data values
   Period (u): 1995 has 2 data values
   Period (u): 1996 has 2 data values
Season 1: VAR(S) = 1412.833
Season 2: VAR(S) = 3736.527
Season 3: VAR(S) = 2550.413

Sum of S      = -328
Sum of VAR(S) = 7699.772

Lower level at: 441.4893 = -.00028
Upper level at: 614.5107 = -.00008

Seasonal Kendall Slope Test Result
-----------------------------------------------
Calculated Z  test = -3.726566
Table         test = 1.960395

Hypothesis: H0: The populations have no trend
            H1: The populations have a trend

Level of significance is 5 Percent
-3.726566 < -1.960395 or -3.726566 > 1.960395

Result:     H0 rejected
```

Table 8.13: Results of HYMOS Kendall Seasonal Slope Analysis

8.7 COMPARING POPULATIONS – STEP TREND

There are some situations where we might expect a sudden change in water quality, for example after installation of a new effluent treatment facility or implementation of a new policy measure. In evaluating the effect of the new facility or policy measure, one can ask the question: 'Has the average value of a water quality parameter changed as compared to the situation before the facility or policy?' If so, we have what is called a *'Step Trend'*, defined as a sudden change in the average concentration of a water quality parameter before and after a specific date (see Figure 8.13).

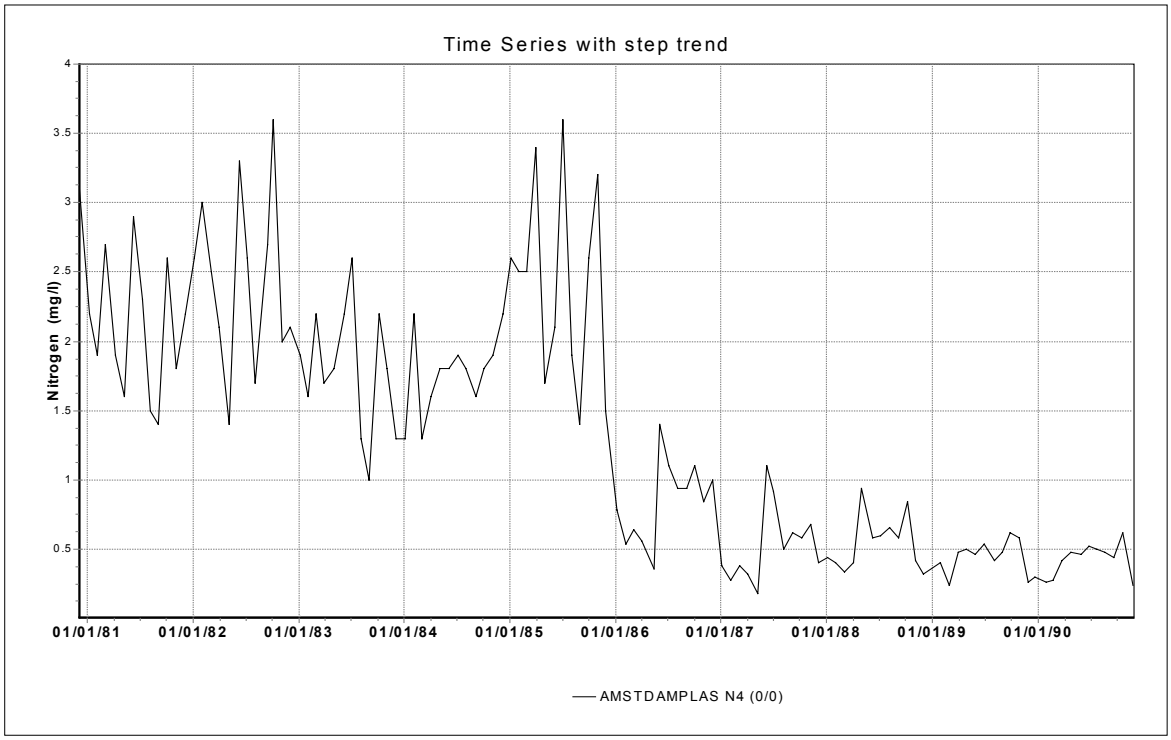


Figure 8.13: Example of water quality data showing a step trend

When analysing data for such a trend, we want to check if the average concentration for a period *after* a given date is different from the average concentration *before* a given date. Two types of tests can be done:

- tests on paired data
- tests on independent data

Furthermore, different statistical tests are available for equidistant or non-equidistant data. An overview of different tests available for Step-Trend analysis in HYMOS is given below in Table 8.12

Test	Data set Requirements			
	Paired data	**Normally distributed (parametric)**	**Equidistant**	**other**
1. Student t-test	No	Yes	Non-Equidist.	• No seasonal influence • Independent values
2. Wilcoxon-W	No	No	Non-Equidist.	-
3. Wilcoxon – Ranked Sum	No	No	Equidistant	• This is non-parametric analog of Student's t-test • No seasonal influence
4. Wilcoxon – Signed Rank	Yes	No	Equidistant	Both series same size, same frequency
5. Wilcoxon- Mann Whitney-U	No	No	Non-Equidist.	--
6. Student t-test (paired data) (Not implemented in HYMOS)	Yes	Yes	-	No missing values

Table 8.14: Overview of statistical tests for comparing populations (applied for detecting step-trend)

8.7.1 PAIRED DATA

By paired data, we mean that two data sets can be linked for some reason such as:

- Samples are collected regularly from 2 locations, upstream and downstream of a pollution source. In this case, each sample concentration from upstream can be paired with a concentration from downstream.
- Water quality samples are regularly sent to 2 different laboratories for analysis. The concentrations measured from Lab 1 can be paired with results from Lab 2, to see if there is any difference in results.
- At one location, 3 years of monthly water quality data *before* installation of a treatment plant are compared to 3 years of monthly water quality data *after* installation of a treatment plant. The data series 'before' can be paired with the data series 'after' (data should be paired month-wise).

The Wilcoxon Signed Rank test is the only statistical test in HYMOS for checking a step trend with paired data.

Wilcoxon Signed Rank Test

Application

The Wilcoxon Signed Rank Test can be used instead of the t-test if the data has a symmetric distribution (though it does not need to be normally distributed). To perform the Wilcoxon Signed Rank test, a series may *not* contain missing values. The data set in HYMOS must be equidistant.

Theory

HYMOS considers the following hypothesis:

H_0: The populations from which the two series have been drawn have the same mean,

H_1: The populations have different means.

The following procedure is followed:

1. Compute the difference for each pair of values.
2. Rank the absolute differences, assign rank 1 to the smallest value, rank two to the second smallest, .., etc. If several data have the same value, assign them the midrank, that is, the average of the ranks that would otherwise be assigned to those data.
3. Each rank is allocated the sign of the corresponding difference.
4. Sum the positive ranks, T+, and the absolute value of the negative ranks, T-.
5. Select the smallest value for T+ and T-.
6. Compute

$$Z_T = \frac{T - \mu_T}{\sigma_T}$$

where: $$\mu_T = \frac{N(N+1)}{4}$$

$$\sigma_T = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

The Z_T statistics under the null-hypothesis is normally distributed. H_0 will not be rejected at an α% significanace level, if:

$$|Z_T| < Z_{1-\alpha/2}$$

Reference:

- R.O. Gilbert, 1987, Statistical Methods for Environmental Pollution Monitoring, John Wiley & Sons Inc..
- Gopal K. Kanji, 1999.100 statistical tests. SAGE Publications, Test 48

Example 8.11 – Wilcoxon Signed Rank

```
Statistical Tests on Data Homogeneity and Randomness
====================================================

Two Series Test
---------------

Series code 1: Amstdamplas Ros
Series code 2: Amstdamplas Ros

Date of first element in series = 05-01-1983
Number of data                  =  36

Wilcoxon Signed Rank Test
-------------------------------------------
Sum of negative ranks     =  0
Sum of positive ranks     =  666
Mean                      =  333
Standard deviation        =  63.65139

Calculated Zb test = -5.231621
Table         test = 1.960395

Hypothesis: H0: The populations have the same mean
            H1: The populations have different means

Level of significance is 5 Percent


5.231621 > 1.960395
Result:      H0 rejected
```

Table 8.15: HYMOS results of Wilcoxon Signed Rank Test

8.7.2 INDEPENDENT DATA

Several tests exist for independent (non-paired) data to evaluate a step trend. Tests on independent data are more easily applied since the data condition of paired data does not exist. Three common tests described below are:

- Student t-test
- Wilcoxon Ranked Sum test
- Wilcoxon W-test

Note: the tests here are presented with the application of analysing the presence of a step trend. The same tests can also be used to compare if 2 separate data sets have the same mean value.

Student t-test.

Application

This is the classic test for testing equality of means between two data sets, and thus can be applied for analysis of a step trend.

The data set is split in two parts, on the basis of when a step break is expected, the so-called 'break-point'. The test checks if the average value of the two parts of the data set are equal. This test is parametric, i.e. assumes a normal distribution of the data. Also the test assumes that the values in both parts of the data set are independent.

The test is described in Annex II, Chapter 6.

Test conditions: Total number of data values (N) ≥10, m≥5 and n≥5

Example 8.12 Student t-test for comparison of means

```
Statistical Tests on Data Homogeneity and Randomness
 ==================================================

 One Series Test
 ---------------

 Series code: Amstdamplas N4

 Date of first element in series= 1980 12  3  0  0
 Number of data                 =  120

 Student t-Test with Welch modification
 --------------------------------------
 Number of data in first set      =        60
 Number of data in second set     =        60
 Test statistic [t]  (abs.value)  =    17.929
 Degrees of freedom               =        83
 Prob(t.le.[t])                   =     1.000
 Mean of first set        (mA)    =     2.133
 St.dev. of first set     (sA)    =      .611
 Mean of second set       (mB)    =      .576
 St.dev. of second set    (sB)    =      .283
 Var. test stat.   (Qi=sA^2/sB^2) =     4.653
 Prob(Q.le.Qi)                    =     1.000
 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 1.989
 Result:     H0 rejected
```

Table 8.16: HYMOS results for Student t-test on comparison of means

Wilcoxon Ranked Sum.

Application

The Wilcoxon Rank Sum Test is a non-parameteric test for independent data sets. The Wilcoxon rank sum test may be used to test for a shift in location between independent series, that is, the measurements from one series tend to be consistently larger (or smaller) than those from the other series. The rank sum test has the advantage that the two data sets need not to be drawn from normal distributions, and the test can handle a moderate number of equal values by treating them as ties. The test assumes, however, that the distributions of the two series are identical in shape (variance), but the distributions need not be symmetric.

Theory

HYMOS considers the following hypothesis:

H_0: The populations from which the two series have been drawn have the same mean,

H_1: The populations have different means.

The following procedure is followed:

1. Combine the two series and rank the m (= n_1 + n_2) data, assign rank 1 to the smallest value, rank two to the next, .., and rank m to the largest value. If several data have the same value, assign them the midrank, that is, the average of the ranks that would otherwise be assigned to those data. When missing values are encountered the function is stopped. The number of values per series must be larger than 10.
2. Sum the ranks assigned to the first series, W_{rs}.
3. If no ties (same values) are present, the test statistic is computed from:

$$Z_{rs} = \frac{W_{rs} - \frac{n_1(m+1)}{2}}{\sqrt{\frac{n_1 n_2 (m+1)}{12}}} \tag{8.18}$$

4. If ties are present, the following function is used:

$$Z_{rs} = \frac{W_{rs} - \frac{n_1(m+1)}{2}}{\sqrt{\frac{n_1 n_2}{12}\left[m+1-\frac{\sum_{j=1}^{g} t_j(t_j^2-1)}{m(m-1)}\right]}} \tag{8.19}$$

where *g* is the number of tied groups and t_j is the number of tied data in the j^{th} group.

5. For an α level two-tailed test, H_0 is rejected $|Z_{rs}| \geq z_{1-\alpha/2}$
6. For a one-tailed α level test, with H_0 stating that the values from series 1 tend to exceed those from series 2, H_0 is rejected if $Z_{rs} \geq z_{1-\alpha}$.
7. For a one-tailed α level test with H_0 stating that the values from series 2 tend to exceed those from series 1, H_0 is rejected if $Z_{rs} \leq z_{\alpha}$.

This test is almost similar to the Mann-Whitney-U test.

Reference:

- R.O. Gilbert, 1987, Statistical Methods for Environmental Pollution Monitoring, John Wiley & Sons Inc.

Wilcoxon W- test

Application

With the Wilcoxon W-test difference in the mean value of two series can be investigated.

Test conditions: $m \geq 10$ and $n \geq 10$

Theory

The Wilcoxon test considers two series A_i,(i=1,m) and B_i,(i=1,n). All values A_i are compared with all B_j, where $w_{i,j}$ is defined by:

$A_i < B_j$: $W_{i,j} = 2$

$A_i = B_j$: $W_{i,j} = 1$ (8.20)

$A_i > B_j$: $W_{i,j} = 0$

The Wilcoxon statistic W is formed by:

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{i,j} \quad (8.21)$$

In case $\mu_A = \mu_B$ the W-statistic is asymptotically *normally* distributed with $N(\mu_W,\sigma_W)$:

$\mu_W = mn$

$\sigma_W^2 = mn(N+1)/3$ (8.22)

where: $N = m+n$

HYMOS considers the following hypothesis:

H_0 : $\mu_A = \mu_B$, and

H_1 : $\mu_A \neq \mu_B$, hence, a two-tailed test is performed

and the absolute value of the following standardised test statistic is computed:

$|u| = |W - \mu_W| / \sigma_W$

Example 8.13 – Wilcoxon W Test for Step trend

```
Statistical Tests on Data Homogeneity and Randomness
 ==================================================


 One Series Test
 ---------------

 Series code: Amstdamplas N4

 Date of first element in series= 1980 12  3  0  0
 Number of data                 =  120

 Wilcoxon-Test on Differences in the Mean
 ----------------------------------------
 Number of data in first set      =        60
 Number of data in second set     =        60
 W-value                          =    37.000
 Mean of W                        =  3600.000
 Standard deviation of W          =   381.051
 Test statistic [u]  (abs.value)  =     9.350
 Prob(u.le.[u])                   =     1.000
 Hypothesis: H0: No difference in the mean of the split samples
             H1: Split samples have different mean values

             A two-tailed test is performed
             Level of significance is  5.00 percent
             Critical value for test statistic 1.960

 Result:     H0 rejected
```

Table 8.17: HYMOS results of Wilcoxon W test for a step trend (non-paired test)

No.	*Conc. (mg/l)*	*No.*	*Conc. (mg/l)*
1.	2.00	29.	4.04
2.	2.10	30.	4.04
3.	2.90	31.	4.04
4.	3.20	32.	4.04
5.	3.40	33.	4.06
6.	3.43	34.	4.08
7.	3.43	35.	4.09
8.	3.50	36.	4.17
9.	3.50	37.	4.17
10.	3.50	38.	4.17
11.	3.58	39.	4.23
12.	3.58	40.	4.23
13.	3.64	41.	4.26
14.	3.64	42.	4.29
15.	3.69	43.	4.32
16.	3.69	44.	4.32
17.	3.71	45.	4.33
18.	3.74	46.	4.33
19.	3.76	47.	4.34
20.	3.76	48.	4.44
21.	3.81	49.	4.45

No.	*Conc. (mg/l)*	*No.*	*Conc. (mg/l)*
22.	3.83	50.	4.46
23.	3.91	51.	4.48
24.	3.91	52.	4.50
25.	3.95	53.	5.00
26.	3.97	54.	5.56
27.	3.99	55.	6.50
28.	4.03		

Table 8.18: Ranked values of Sample Concentrations, sample data for Rosner's test (after Gilbert, 1987; p.190)

No.	*Conc.*	*No.*	*Conc.*
1	1.85	24	6.02
2	2.03	25	6.02
3	2.26	26	6.35
4	3.38	27	6.78
5	3.45	28	7.24
6	3.81	29	7.47
7	3.95	30	7.51
8	3.99	31	7.98
9	4.10	32	8.43
10	4.10	33	8.60
11	4.12	34	9.62
12	4.29	35	11.20
13	4.42	36	11.53
14	4.62	37	12.10
15	4.74	38	12.30
16	4.95	39	14.59
17	5.01	40	14.67
18	5.04	41	14.83
19	5.21	42	14.97
20	5.22	43	15.13
21	5.36	44	19.00
22	5.69	45	22.92
23	5.79	46	23.53

Table 8.19: (Ranked) Data of Nitrate Concentrations at location 'BasicStat' (after Gilbert (1987), p.144)

no	date	value	1-3/31-5 summer	1-6/31-10 monsoon	1-11/28-02 winter
1	19-01-1983	4.00			4
2	19-05-1983	1.00	1		
3	14-09-1983	2.00		2	
4	08-12-1983	0.50			0.5
5	15-03-1984	1.50	1.5		
6	10-05-1984	1.00	1		
7	12-09-1984	2.50		2.5	
8	10-10-1984	4.00		4	
9	16-01-1985	1.50			1.5
10	15-05-1985	0.50	0.5		
11	14-08-1985	0.50		0.5	
12	06-12-1985	1.00			1
13	27-01-1986	2.90			2.9
14	09-04-1986	2.10	2.1		
15	11-08-1986	1.00		1	
16	21-01-1987	1.00			1
17	25-02-1987	2.20			2.2
18	21-05-1987	0.50	0.5		
19	22-06-1987	1.70		1.7	
20	17-08-1987	0.60		0.6	
21	21-09-1987	0.50		0.5	
22	17-11-1987	1.70			1.7
23	13-01-1988	2.79			2.79
24	18-04-1988	2.60	2.6		
25	12-07-1988	2.20		2.2	
26	13-12-1988	1.30			1.3
27	16-03-1989	1.70	1.7		
28	02-06-1989	0.67		0.67	
29	10-08-1989	0.73		0.73	
30	06-10-1989	0.84		0.84	
31	21-11-1989	0.65			0.65
32	07-02-1990	1.60			1.6
33	21-03-1990	1.10	1.1		
34	07-06-1990	0.75		0.75	
35	16-08-1990	0.35		0.35	
36	09-11-1990	0.38			0.38
37	11-03-1991	0.95	0.95		
38	11-06-1991	0.69		0.69	
39	02-09-1991	0.10		0.1	
40	04-12-1991	0.40			0.4
41	25-03-1992	0.97	0.97		
42	21-05-1992	0.80	0.8		
43	11-08-1992	0.42		0.42	
44	10-11-1992	0.65			0.65
45	02-02-1993	1.80			1.8
46	11-05-1993	0.83	0.83		
47	03-08-1993	0.13		0.13	
48	23-11-1993	0.59			0.59

49	14-01-1994	2.70			2.7
50	22-02-1994	1.10			1.1
51	14-03-1994	1.30	1.3		
52	12-04-1994	1.60	1.6		
53	02-05-1994	1.10	1.1		
54	02-06-1994	0.73		0.73	
55	12-07-1994	0.19		0.19	
56	19-08-1994	0.14		0.14	
57	13-09-1994	0.14		0.14	
58	14-09-1994	0.24		0.24	
59	05-10-1994	0.82		0.82	
60	02-11-1994	1.10			1.1
61	20-12-1994	1.10			1.1
62	24-01-1995	2.60			2.6
63	15-02-1995	2.40			2.4
64	22-03-1995	0.99	0.99		
65	11-04-1995	0.94	0.94		
66	09-05-1995	0.25	0.25		
67	11-05-1995	0.36	0.36		
68	20-06-1995	0.60		0.6	
69	10-08-1995	0.05		0.05	
70	17-10-1995	0.31		0.31	
71	15-11-1995	0.76			0.76
72	12-12-1995	0.39			0.39
73	16-01-1996	0.60			0.6
74	27-02-1996	0.70			0.7
75	25-03-1996	0.41	0.41		
76	09-04-1996	0.61	0.61		
77	07-05-1996	0.23	0.23		
78	12-06-1996	0.50		0.5	
79	10-07-1996	0.31		0.31	
80	20-08-1996	0.49		0.49	
81	18-09-1996	0.43		0.43	
82	15-10-1996	0.33		0.33	
83	19-11-1996	0.72			0.72
84	04-12-1996	3.50			3.5
		mean	1.015	0.780	1.470
		max.	2.600	4.000	4.000
		min.	0.230	0.050	0.380
		median	0.970	0.500	1.100
		q1	0.555	0.31	0.65
		q3	1.2	0.7675	2.2
		n	23	32	29

Table 8.20: Data of Cadmium concentrations at locations STOWA1 (Neerbeek) original data (value in ug/L) and values sorted according to seasons (India_3Seasons)

No.	date	Conc.	No.	date	Conc.
1	03/12/80	3.1	61	29/11/85	1.5
2	06/01/81	2.2	62	06/01/86	0.78
3	03/02/81	1.9	63	04/02/86	0.54
4	03/03/81	2.7	64	04/03/86	0.64
5	07/04/81	1.9	65	03/04/86	0.56
6	08/05/81	1.6	66	14/05/86	0.36
7	09/06/81	2.9	67	04/06/86	1.4
8	10/07/81	2.3	68	03/07/86	1.1
9	05/08/81	1.5	69	04/08/86	0.94
10	03/09/81	1.4	70	03/09/86	0.94
11	02/10/81	2.6	71	01/10/86	1.1
12	03/11/81	1.8	72	03/11/86	0.84
13	03/12/81	2.2	73	04/12/86	1
14	05/01/82	2.6	74	05/01/87	0.38
15	02/02/82	3	75	03/02/87	0.28
16	03/03/82	2.5	76	05/03/87	0.38
17	31/03/82	2.1	77	02/04/87	0.32
18	06/05/82	1.4	78	06/05/87	0.18
19	08/06/82	3.3	79	08/06/87	1.1
20	07/07/82	2.6	80	02/07/87	0.92
21	05/08/82	1.7	81	05/08/87	0.5
22	16/09/82	2.7	82	04/09/87	0.62
23	06/10/82	3.6	83	05/10/87	0.58
24	03/11/82	2	84	06/11/87	0.68
25	02/12/82	2.1	85	03/12/87	0.4
26	05/01/83	1.9	86	04/01/88	0.44
27	02/02/83	1.6	87	01/02/88	0.4
28	02/03/83	2.2	88	02/03/88	0.34
29	30/03/83	1.7	89	05/04/88	0.4
30	03/05/83	1.8	90	02/05/88	0.94
31	07/06/83	2.2	91	08/06/88	0.58
32	05/07/83	2.6	92	05/07/88	0.6
33	03/08/83	1.3	93	05/08/88	0.66
34	02/09/83	1	94	05/09/88	0.58
35	05/10/83	2.2	95	06/10/88	0.84
36	02/11/83	1.8	96	02/11/88	0.42
37	02/12/83	1.3	97	01/12/88	0.32
38	04/01/84	1.3	98	30/01/89	0.4
39	03/02/84	2.2	99	27/02/89	0.24
40	02/03/84	1.3	100	30/03/89	0.48
41	04/04/84	1.6	101	28/04/89	0.5
42	04/05/84	1.8	102	31/05/89	0.46
43	05/06/84	1.8	103	29/06/89	0.54
44	05/07/84	1.9	104	03/08/89	0.42
45	03/08/84	1.8	105	30/08/89	0.48

No.	date	Conc.	No.	date	Conc.
46	06/09/84	1.6	106	28/09/89	0.62
47	02/10/84	1.8	107	30/10/89	0.58
48	02/11/84	1.9	108	30/11/89	0.26
49	06/12/84	2.2	109	21/12/89	0.3
50	03/01/85	2.6	110	29/01/90	0.26
51	01/02/85	2.5	111	21/02/90	0.28
52	01/03/85	2.5	112	26/03/90	0.42
53	01/04/85	3.4	113	26/04/90	0.48
54	02/05/85	1.7	114	30/05/90	0.46
55	04/06/85	2.1	115	28/06/90	0.52
56	03/07/85	3.6	116	26/07/90	0.5
57	02/08/85	1.9	117	22/08/90	0.48
58	30/08/85	1.4	118	20/09/90	0.44
59	30/09/85	2.6	119	22/10/90	0.62
60	31/10/85	3.2	120	26/11/90	0.24

*Note: Data are paired; the data set is divided into 2 halves and data are paired month-wise

Data nos. n= 1 to 60 are considered the first half of the data set; (3/12/80 to 31/10/85)

Data nos. n= 61=120 are the second half (29/11/85 to 26/11/90)

Data no. n=1 is paired with n=61 (month 12); n=2 is paired with n=62 (month 1), etc.

Table 8.21: (Paired) data of phosphorus concentrations at location Amstdamplas*

9 REPORTING ON RAINFALL DATA

9.1 GENERAL

Published reports are the primary visible output of the Hydrological Information System. They have several purposes

- to provide information on availability of data for use in planning and design. Rainfall data are used for a variety of purposes and are required at a range of time scales. Real time rainfall data are required for flood forecasting and hydropower and reservoir operation. Summaries of storm rainfall event data are required for assessment of the severity of events at weekly or monthly time scales. Rainfall bulletins for agricultural and irrigation operation are needed at similar time scales. However, the HIS will data at yearly or longer reporting frequency and will not engage in shorter term operation reports. Although the same data may be used for such reports they will not be the direct concern of the HIS.
- to advertise the work of the HIS and its capability and to create interest and awareness amongst potential users.With the availability of data on magnetic media it is conceivable that all requests for data could be met by a direct and specific response to data requests. This in fact is now the practice in many developed countries where there are well established links between data users and data suppliers and annual reports are no longer published in print (although the same information may be provided on the Internet). In India, the availability of rainfall data may not be well known even in related government departments; the annual report of rainfall therefore provides a suitable means of demonstrating the capability of the HIS.
- to provide feedback to data producers and acknowledge the contribution of observers and co-operating agencies. The HIS is an integrated system in which rainfall (and other) data are transferred by stages from the field, to local and regional offices for data entry, processing and

validation. The annual report shows how observations at individual stations are integrated in the network. It provides an encouragement to observers and data processors to ensure that the raw and processed data are reliable.

The HIS provides opportunities for storage, retrieval and reporting on magnetic media and there is now no necessity to publish daily rainfall for all contributing stations. The traditional annual report of daily rainfall is often not the most convenient format of rainfall data for users. For project or design purposes, the user often requires long term records for a single station or a group of stations - i.e., data by station rather than by year. This required the collation of data from a set of annual reports and the keying of the data into the computer for the required analysis. So long as the annual report gives a clear indication of data availability as a basis for user requests, it is now more efficient and cost effective to provide rainfall summary statistics rather than the full daily record.

The HIS thus makes data reporting and use more efficient by:

- reducing the amount of published data and cost of annual reports
- providing statistical summaries in tabular and graphical from which are more accessible and interesting to the user
- avoiding duplication of effort by users in keying in of data by provision on magnetic media

Annual reports are produced with respect to rainfall over the hydrological year from 1 June to 31 May. Since the hydrological year corresponds to a complete cycle of replenishment and depletion, it is appropriate to report on that basis rather than with respect to the calendar year. Such reports incorporate

- a summary of information on the pattern of rainfall over the year in question
- information on the long-term spatial and temporal pattern of rainfall in the region and how the recent year compares with past statistics.

Reports of long term statistics of rainfall will be prepared and published at 5 or 10 years intervals. These will incorporate spatial as well as temporal analysis.

Annual and other reports will be produced at the State Data Processing centre. Annual reports will be produced in draft form within six months from the end of the year covered by the publication and the report published within twelve months.

9.2 YEARLY REPORTS

The annual report provides a summary of the rainfall pattern for the report year in terms of distribution of rainfall in time and space and makes comparisons with long term statistics. Details of the observational network and data availability are included. A summary of the hydrological impact of rainfall is provided with particular reference to floods and droughts. The following are typical contents of the annual report:

(a) Introduction
(b) The Observational Network
  - maps
  - listings
(c) A descriptive account of rainfall occurrence during the report year
(d) Thematic maps of monthly, seasonal and annual rainfall
(e) Graphical and mapped comparisons with average patterns
(f) Basic rainfall statistics
(g) Description and statistical summaries of major storms

(h) Data validation and quality

(i) Bibliography

9.2.1 INTRODUCTION

The report introduction, which may change little from year to year, will describe the administrative organisation of the rainfall network and the steps involved in the collection, data entry, processing, validation, analysis and storage of data. It will list those agencies contributing to the included data. It will describe how the work is linked with other agencies collecting or using rainfall data including the India Meteorological Department and operational departments in hydropower and irrigation. It will describe how additional data may be requested and under what terms and conditions they are supplied.

9.2.2 THE OBSERVATIONAL NETWORK

The salient features of the observational network are summarised in map and tabular form.

The rainfall station map must also show major rivers and basin boundaries and distinguish each site by symbol between daily, autographic and digital recorder and whether rainfall alone is observed or the gauge is sited at a climatological station.

Tabulations of current stations are listed by named basin and sub-basin. Also listed are latitude, longitude, altitude, responsible agency, the full period of observational record and the period of observation which is available in digital format. A similar listing of closed stations, (or a selection of closed stations with long records) may be provided. All additions and closures of stations must be highlighted in the yearly report. Similarly station upgrading and the nature of the upgrading should be reported.

9.2.3 DESCRIPTIVE ACCOUNT OF RAINFALL DURING THE REPORT YEAR.

An account of the rainfall occurrence in the region in the year can be concisely given in the form of a commentary for each month, placed in its meteorological context. Significant stretches of dry or wet periods in the parts of the region under reporting can be highlighted.

9.2.4 MAPS OF MONTHLY, SEASONAL AND YEARLY AREAL RAINFALL

Thematic maps showing spatial distribution of average rainfall over the region for monthly, seasonal or yearly periods provide a convenient summary of the rainfall pattern in space and time. Basin or administrative boundaries may also be shown to illustrate variations between districts or basins. The rainfall may be mapped as the actual value at each station for the specified period or by the drawing of isohyets of equal rainfall over the region. For such interpolations the rainfall is first interpolated on a very fine grid laid over the region using manual or computer-based techniques. Grid point values are then used to draw isohyets at suitable intervals.

9.2.5 GRAPHICAL AND MAPPED COMPARISONS WITH AVERAGE PATTERNS

Maps will also be provided to show relative rainfall - the amount as a percentage of the long term average. The period over which the long term average is taken must be noted.

For a few representative rainfall stations, a graphical comparison of the monthly rainfall amounts for the whole year can be made with the long term average patterns. The actual monthly distribution can be plotted against the long term average for minimum, maximum and average monthly amounts. This kind of plot also makes it easy to comprehend the type of temporal distribution of rainfall.

9.2.6 BASIC STATISTICS FOR VARIOUS DURATION

This forms the core of the report. As noted above the full reporting of daily or hourly data is no longer required though sample tabulations of daily and hourly data may be provided for selected stations to illustrate the format of information available. Instead, summary statistics of monthly rainfall for the report year provide a ready means of making comparisons between stations and between months and will satisfy the needs of general data users.

Again stations are listed by basin and sub-basin order (rather than alphabetical or numerical order). In addition to monthly rainfall totals, the maximum daily amount in the year and the date of its occurrence is noted. Any daily, monthly or annual totals which exceed previous maxima of record are shown in bold type.

For stations with digital or autographic records a similar tabulation is provided by basin giving the maximum observed amount for selected durations including 1 hour, 2, 3, 6, 12 and 24 hours with dates of occurrence.

9.2.7 DESCRIPTION AND STATISTICAL SUMMARIES OF MAJOR STORMS

Major storms which are known to have had an impact on flooding or operation of water resources are described in more detail. Selection of events for description may be made in terms of impact or on an objective basis of areal amount and distribution. For rainfall regimes of arid and semi-arid regions a lower value is adopted whereas for high rainfall regimes a higher threshold value is adopted. Usually, a threshold of about 10% of the seasonal normal rainfall may be taken for the most frequent storm duration over the region. The threshold value also depends upon the size of the catchment area. For smaller catchment a higher threshold and for larger catchments smaller threshold value may be adopted. An average precipitation depth of 50 mm per day over a catchment of medium size (say 10,000 – 15,000 sq. kms.) would be appropriate. The peripheral isohyet for one day storm must be at least 50 mm in the moderate rainfall regime whereas it must be about 10 to 20 mm for arid or semi-arid regions with low seasonal rainfall.

Storms should be described with respect to their meteorological context, centre of concentration, movement across the river basins and also the characteristics of the time distribution of rainfall within the storm.

9.2.8 DATA VALIDATION AND QUALITY

The limitations of data should be made known to users. The validation process not only provides a means of checking the quality of the raw data but also a means of reporting. The number of values corrected or in-filled as a total or a percentage may be noted for individual stations, by basin or by agency. The types of anomaly typically detected by data validation and remedial actions should be described.

9.2.9 BIBLIOGRAPHY

Data users may be interested to know of other sources of rainfall data or of related climatic or hydrological data. The following should be included.

Concurrent annual reports from the HIS of climate or hydrological data

Previous annual rainfall reports (with dates) from the HIS.

Previous annual rainfall reports (with dates) published by each agency and division within the state

Special summary reports of rainfall statistics produced by the HIS or other agencies.

A brief note on the administrative context of previous reports, methods of data compilation, and previous report formats would be helpful.

9.3 PERIODIC REPORTS - LONG TERM STATISTICS

Long term point and areal statistics are important for planning, management and design of water resources systems. They also play an important role in validation and analysis. These statistics must be updated regularly and an interval of 10 years is recommended. The following will be typical contents of such reports.

- Introduction
- Data availability - maps and tabulations
- Descriptive account of annual rainfall since last report
- Thematic maps of mean monthly and seasonal rainfall
- Basic rainfall statistics - monthly and annual means, maxima and minima
  - for the standard climatic normal period (1961-90) where available
  - for the updated decade
  - for the available period of record
- Additional point rainfall statistics for example, daily maximum rainfall, persistence of dry or wet spells during the monsoon, dates of onset or termination of the monsoon.
- Additional areal mean rainfall statistics for administrative or drainage areas for periods of a month or year
- Analysis of temporal variability using moving averages or residual mass curves to identify major wet and dry periods for a number of representative stations.
- Frequency analysis of rainfall data

9.3.1 FREQUENCY ANALYSIS OF RAINFALL DATA

The frequency of occurrence of rainfall of various magnitudes is important in the application of mathematical models for synthesising hydrological data. Estimates of design runoff from small areas are often based on rainfall-runoff relations and rainfall frequency data due to sparse streamflow measurements and limitation in transposing such data among small areas. Generalised estimates of rainfall frequencies for a few durations up to 72 hours and up to a few hundred years are useful if are readily available. Some such maps are available at country level for specified duration of rainfall and frequency of occurrence (or return periods). These maps must be revised after having collected a significant amount of additional data. Standard methods recommended by India Meteorological Department must be followed for the derivation of such maps. Though the primary responsibility for making such maps lies with the India Meteorological Department, it is appropriate to include such maps in the reports with the permission of the IMD.

Information on rainfall frequency is a vital input for planning domestic or industrial water supply, agricultural planning, hydropower and other water use sectors. Inferences on various time intervals such as daily, weekly, ten-daily, fortnightly and monthly are usually required for planning in various sectors.

9.4 PERIODIC REPORTS ON UNUSUAL RAINFALL EVENTS

Special reports should also be prepared on the occurrence of unusual rainfall events. As these will also have unusual hydrological consequences, the reports will normally be combined with reports of the resulting streamflow and flooding within the affected area. The rainfall component of such reports will include the following

- tabulations of hourly or daily point rainfall within the affected area

- isohyetal maps of total storm rainfall
- hyetograph plots of rainfall time distribution based on recording raingauges
- assessment of event return periods for selected durations based on point rainfall
- areal storm rainfall totals over affected basins

10 REPORTING ON CLIMATIC DATA

10.1 GENERAL

Published reports are the primary visible output of the Hydrological Information System. They have several purposes

- to provide information for use in planning, design, operation and evaluation. Evaporation and evapotranspiration data are used for irrigation scheme design, operation and evaluation, for agricultural operations and in flood forecasting models.
- to advertise the work of the HIS and its capability and to create interest and awareness amongst potential users.
- to provide tangible evidence to policy makers of a return on substantial public investment.
- to provide feedback to data producers and acknowledge the contribution of observers and co-operating agencies. The HIS is an integrated system in which evaporation (and other) data are transferred by stages from the field, to local and regional offices for data entry, processing and validation. The annual report shows how observations at individual stations are integrated in the network. It provides an encouragement to observers and data processors to ensure that the raw and processed data are reliable.
- to provide a clear incentive to keep archives up to date and a focus for an annual hydrometric audit

The HIS provides opportunities for storage, retrieval and reporting on magnetic media and there is now no necessity to publish daily records for all contributing stations. Reports are primarily designed to cover a fixed time interval, most commonly the water year. In contrast users most commonly require data as full time series from the beginning to the end of the record. there is thus a degree of incompatibility between user requirements and reporting formats. It is not possible to provide complete records in report form, though these can conveniently be provided on magnetic media from the HIS. The main function of the report therefore with respect to functional use is to inform users of the availability of data in digital and other formats.

Annual reports are produced with respect to evaporation over the hydrological year from 1 June to 31 May. They will generally be combined with annual rainfall reports and may be combined with streamflow.

A broad range of climatic variables is measured at observation station but, for hydrological purposes, the variables are not themselves of direct interest but are used in computing evapotranspiration by theoretical and empirical methods - especially the Penman method. Whilst computed evapotranspiration will be reported, the statistics of climatic variables used in the computation are not required for reporting. Direct measurements of pan evaporation will be included in the report.

10.2 YEARLY REPORTS

The annual report provides a summary of evaporation for the report year in terms of distribution in time and space. It also makes comparisons with long term statistics. Details of the observational network and data availability are included. The following are typical contents of the annual report:

- Introduction
- The Observational Network

  - maps
  - listings
- Basic evaporation statistics
- Annual summaries in graphical form
- Data validation and quality
- Bibliography

10.2.1 INTRODUCTION

The report introduction, which may change little from year to year, will describe the administrative organisation of the climate and evaporation network and the steps involved in the collection, data entry, processing, validation, analysis and storage of data. Standard climatic observational practice for variables required by computation of Penman evapotranspiration will be summarised.

The report will list those agencies contributing to the included data. It will describe how the work is linked with other agencies collecting or using evaporation data including the India Meteorology Department and operational departments in hydropower and irrigation. It will describe how additional data may be requested and under what terms and conditions they are supplied.

10.2.2 THE OBSERVATIONAL NETWORK

The salient features of the observational network are summarised in map and tabular form.

The map of climate stations must also show major rivers and basin boundaries and distinguish each site by symbol between the combination of instruments in use at each station (e.g. automatic weather stations, stations with net radiometer, etc.). Mapped stations must be numbered so that they can be related to information contained in tabular listings.

Tabulations of current stations are listed by named basin and sub-basin. Also listed are latitude, longitude, altitude, responsible agency, the full period of observational record and the period of observation which is available in digital format. A similar listing of closed stations may be provided. All additions and closures of stations must be highlighted in the yearly report. Similarly station upgrading and the nature of the upgrading should be reported.

10.2.3 BASIC EVAPORATION STATISTICS

This forms the core of the report. As noted above the full reporting of daily data is no longer required and the principal output will be monthly statistics of evaporation for each station compared with the average for the period of record. Stations will be ordered by basin and sub-basin - rather than in alphabetical order. Fig. 1 provides an example of such a listing. A typical listing includes:

- For the current year
  - monthly and annual pan evaporation
  - monthly and annual Penman evapotranspiration

- For the previous record
  - mean monthly and annual pan evapotranspiration
  - lowest monthly mean in period
  - highest monthly mean in period
  - various percentile values
  - mean monthly and annual Penman evapotranspiration
  - lowest monthly mean in period
  - highest monthly mean in period

  - various percentile values

- For the station
  - location details and station elevation

Values of evaporation, whether from pan measurements or derived from Penman calculations should be reported to no more than one decimal place (mm). More than one decimal place is beyond the accuracy of measurement and gives a spurious impression of accuracy.

10.2.4 GRAPHICAL AND MAPPED COMPARISONS WITH AVERAGE PATTERNS

Graphical displays often provide the best and most accessible means of illustrating the time series of evaporation during the water year and how this relates to the previous record. The following graphical plots will be presented for a selection of stations.

- Annual histogram plot of monthly evaporation compared with previous mean, maxima and minima
- Map showing annual or seasonal evaporation as a percentage of the long period average.

10.2.5 DATA VALIDATION AND QUALITY

The limitations of the data should be made clear to users. The general limitations of pan evaporation as a measure of open water evaporation should be explained (primarily the difference in heat storage properties of a small metal container and an extensive natural open water surface). In addition, the number of values corrected or infilled as a total or a percentage may be noted for individual stations, by basin or by agency.

10.2.6 BIBLIOGRAPHY

Data users may be interested to know of other sources of evaporation and related climatic and rainfall and streamflow data. The following should be included.

- Concurrent annual reports from the HIS of rainfall or streamflow data
- Previous annual reports incorporating climate and evaporation data (with dates) from the HIS.
- Previous annual reports incorporating climate and evaporation data(with dates) published by each agency and division within the state
- Special summary reports of climate and evaporation statistics produced by the HIS or other agencies.

A brief note on the administrative context of previous reports, methods of data compilation, and previous report formats would be helpful.

10.3 PERIODIC REPORTS - LONG TERM STATISTICS

Long term point and areal statistics are important for planning, management and design of water resources systems. They also play an important role in validation and analysis. These statistics must be updated regularly and an interval of 10 years is recommended. The following will be typical contents of such reports.

- Introduction
- Data availability - maps and tabulations
- Descriptive account of annual measured pan evaporation and computed evapotranspiration since the last report

- Thematic maps of mean monthly and seasonal evaporation
- Basic evaporation statistics - monthly and annual means, various percentiles, maxima and minima
  - for a standard climatic normal period where available
  - for the updated decade
  - for the available period of record

Analysis of periodicity and trend in the evaporation data

11 REPORTING ON STAGE DISCHARGE DATA

11.1 GENERAL

- Published reports are the primary visible output of the Hydrological Information System. The principal reports will be with respect rainfall, climate and streamflow and will cover the water year (from 1 June to 31 May). A limited amount of stage discharge data will be incorporated with reports on streamflow. Reports have several purposes
  - to provide information for use in planning and design. Stage discharge data are not directly used, but they can provide an indication of the reliability of derived streamflow data. Sufficient information should be provided for this purpose.
  - to advertise the work of the HIS and its capability and to create interest and awareness amongst potential users..
  - to provide feedback to data producers and acknowledge the contribution of observers and co-operating agencies. The HIS is an integrated system in which data are transferred by stages from the field, to local and regional offices for data entry, processing and validation. The annual report shows how observations at individual stations are integrated in the network. It provides an encouragement to observers and data processors to ensure that the raw and processed data are reliable.

The HIS provides opportunities for storage, retrieval and reporting on magnetic media and there is now no necessity to publish all available data from contributing stations. There is no necessity to report on every discharge observation made during the year. The parameters of individual stage discharge relationships also need not be reported but must be available on request to users along with time series of discharge.

11.2 LAYOUT OF REPORT

The following table of summary information for each station is recommended as a guide to gauging effort and the reliability of the ratings:

	Current year		Previous record	
	Level	Flow	Level	Flow
Maximum observed				
Maximum gauged				
Minimum observed				
Minimum gauged				
Number of gaugings in the year				
Number of ratings in the year				
Overall standard error of rating (1)				
Overall standard error of rating (2)				
Overall standard error of rating (3)				
Last date of change of rating				

12 REPORTING ON DISCHARGE DATA

12.1 GENERAL

Published reports are the primary visible output of the Hydrological Information System. They have several purposes

- to provide information for use in planning, design, operation and evaluation. The list of potential users of streamflow data is very large. Data are used for:
  - for the design of water resources schemes taking into consideration particularly the risk of drought
  - for flood defence and drainage schemes taking into account the risk of flood discharges
  - for control of water quality considering the volumes of water available for dilution of industrial and domestic effluents
  - for water issues related to fisheries, ecology, recreation and navigation
  - for all the above with respect to education, research, policy making at state, inter-state and international levels.
- to advertise the work of the HIS and its capability and to create interest and awareness amongst potential users.. With the availability of data on magnetic media it is conceivable that all requests for data could be met by a direct and specific response to data requests. This in fact is now the practice in many developed countries where there are well established links between data users and data suppliers and annual reports are no longer published in print (although the same information may be provided on the Internet). In India, the availability of streamflow data may not be well known even in related government departments; the annual report of streamflow therefore provides a suitable means of demonstrating the capability of the HIS.
- to provide tangible evidence to policy makers of a return on substantial public investment
- to provide feedback to data producers and acknowledge the contribution of observers and co-operating agencies. The HIS is an integrated system in which streamflow (and other) data are transferred by stages from the field, to local and regional offices for data entry, processing and validation. The annual report shows how observations at individual stations are integrated in the network. It provides an encouragement to observers and data processors to ensure that the raw and processed data are reliable.
- to provide a clear incentive to keep archives up to date and a focus for an annual hydrometric audit

The HIS provides opportunities for storage, retrieval and reporting on magnetic media and there is now no necessity to publish daily flow records for all contributing stations. Reports are primarily designed to cover a fixed time interval, most commonly the water year. In contrast users most commonly require data as full time series from the beginning to the end of the record. there is thus a degree of incompatibility between user requirements and reporting formats. It is not possible to provide complete records in report form, though these can conveniently be provided on magnetic media from the HIS. The main function of the report therefore with respect to functional use is to inform users of the availability of data in digital and other formats.

The HIS thus makes data reporting and use more efficient by:

- reducing the amount of published data and cost of annual reports
- providing statistical summaries in tabular and graphical form which are more accessible and interesting to the user
- avoiding duplication of effort by users in keying in of data by provision on magnetic media

Annual reports are produced with respect to streamflow over the hydrological year from 1 June to 31 May. Since the hydrological year corresponds to a complete cycle of replenishment and depletion, it is appropriate to report on that basis rather than with respect to the calendar year. Such reports incorporate

- a summary of information on the pattern of streamflow over the year in question
- information on the spatial and temporal pattern of streamflow in the region and how the recent year compares with past statistics.

Reports of long term statistics of streamflow will be prepared and published at 5 or 10 year intervals. These will incorporate spatial as well as temporal analysis.

Annual and other reports will be produced at the State Data Processing Centre. Annual reports will be produced in draft form within six months from the end of the year covered by the publication and the report published within twelve months. Annual streamflow, rainfall and climate data may be presented in a single combined report.

12.2 YEARLY REPORTS

The annual report provides a summary of streamflow for the report year in terms of distribution in time and space. It also makes comparisons with long term statistics. Details of the observational network and data availability are included. The following are typical contents of the annual report:

- Introduction
- The Observational Network
  - maps
  - listings
- A descriptive account of streamflow occurrence during the report year
- Basic streamflow statistics
- Annual summaries in graphical form
- Description and statistical summaries of major floods and droughts
- Data validation and quality
- Bibliography

12.2.1 INTRODUCTION

The report introduction, which may change little from year to year, will describe the administrative organisation of the streamflow network and the steps involved in the collection, data entry, processing, validation, analysis and storage of data. It will list those agencies contributing to the included data. It will describe how the work is linked with other agencies collecting or using streamflow data including the Central Water Commission and operational departments in hydropower and irrigation. It will describe how additional data may be requested and under what terms and conditions they are supplied.

12.2.2 THE OBSERVATIONAL NETWORK

The salient features of the observational network are summarised in map and tabular form.

The map of gauging stations must also show major rivers and basin boundaries and distinguish each site by symbol between operating agency. Mapped stations must be numbered so that they can be related to information contained in tabular listings (Fig. 1).

Tabulations of current stations are listed by named basin and sub-basin. Also listed are latitude, longitude, altitude, responsible agency, the full period of observational record and the period of observation which is available in digital format. A similar listing of closed stations may be provided. All additions and closures of stations must be highlighted in the yearly report. Similarly station upgrading and the nature of the upgrading should be reported.

12.2.3 DESCRIPTIVE ACCOUNT OF STREAMFLOW DURING THE REPORT YEAR.

An account of streamflow occurrence in the region in the year can be concisely given in the form of a commentary for each month, placed in its meteorological context and in relation to the seasonal norms. Especially severe or prolonged periods of high or low flows can be highlighted.

12.2.4 BASIC STREAMFLOW STATISTICS

This forms the core of the report. As noted above the full reporting of daily or hourly data for all stations is no longer required. However for selected major stations a full listing of daily flows will be provided with accompanying statistical information relating to the year in question and with respect to comparisons with the previous gauged record. Stations will be ordered by basin and sub-basin - rather than in alphabetical order. Such a listing includes:

- For the current year
  - the tabulation of daily mean flow for the year
  - the mean, maximum and minimum daily mean flow in each month
  - monthly flows against the frequency curves for different frequencies
  - the maximum instantaneous (peak) flow in each month
  - monthly flow volumes, runoff (mm) and basin rainfall (mm)
  - annual summary statistics

- For the previous record
  - average of monthly means, lowest monthly mean (and year) and highest (and year)
  - annual summary statistics

- For the basin
  - location details, station elevation and catchment area
  - summary description of the gauging station, its controls and limitations
  - summary description of the catchment including principal features of geology and land use
  - summary of artificial factors affecting flow, reservoirs and regulation, abstractions and return flows.

For the remaining stations, abbreviated summary statistics are provided. Fig. 3 provides an example which includes:

- For the current year
  - monthly and annual mean flows
  - monthly and annual maximum flows
  - monthly and annual runoff (mm)
  - monthly and annual basin rainfall

- For the previous record
  - Monthly and annual mean flows
  - Lowest monthly mean in period
  - highest monthly mean in period
  - highest monthly instantaneous flow
  - mean monthly runoff (mm)
  - mean monthly and annual basin rainfall

- For the basin
  - location details, station elevation and catchment area

Values of flow, whether, observed, mean daily or mean monthly should be reported to two decimal places or less. More than two decimal places is beyond the accuracy of measurement and gives a spurious impression of accuracy.

12.2.5 GRAPHICAL AND MAPPED COMPARISONS WITH AVERAGE PATTERNS

Graphical displays often provide the best and most accessible means of illustrating the time series of flow during the water year and how this relates to the previous record. The following graphical plots will be presented for a selection of stations.

- Annual hydrograph plot compared with previous maxima and minima.
- Flow duration curve showing comparison of current year with long term curve.
- Map showing annual runoff as a percentage of the long period average. Note that this is a very generalised map since the value at a gauging station represents an average value over a basin whilst the runoff from different sub-catchments may be quite different in relation to the period norms.

12.2.6 DESCRIPTION AND STATISTICAL SUMMARIES OF MAJOR FLOODS AND DROUGHTS

Major floods which have caused loss of life or serious or widespread damage to property are described in more detail giving details of peak flow and average flow over selected durations for stations within the affected area, and showing how these statistics differ from the previous reported maxima Storms should be described with respect to their meteorological context, the most severely affected areas, and the impact of storm movement across the basin on the resulting flood. The description may be combined with the rainfall report for the storm (Module 13).

Similarly major droughts which have caused serious agricultural impacts or disruption of water supply should be illustrated by comparison of drought flow hydrographs compared with average and previous minima of experience.

12.2.7 DATA VALIDATION AND QUALITY

The limitations of the data should be made clear to users. The accuracy of flow data are dependent primarily on the accuracy of the stage record and on the. the reliability of the stage discharge relationship. With respect to stage the number of values corrected or infilled as a total or a percentage may be noted for individual stations, by basin or by agency. With respect to the reliability of the stage discharge relationship, the number of gaugings in the period and the extent to which the gauging range falls short of the observed range should be reported. A list of reportable quantities is provided in Chapter 10.

12.2.8 BIBLIOGRAPHY

Data users may be interested to know of other sources of streamflow and related climatic and rainfall data. The following should be included.

- Concurrent annual reports from the HIS of rainfall or climate data
- Previous annual streamflow reports (with dates) from the HIS.
- Previous annual streamflow reports (with dates) published by each agency and division within the state
- Special summary reports of streamflow statistics produced by the HIS or other agencies.

A brief note on the administrative context of previous reports, methods of data compilation, and previous report formats would be helpful.

12.3 PERIODIC REPORTS - LONG TERM STATISTICS

Long term point and areal statistics are important for planning, management and design of water resources systems. They also play an important role in validation and analysis. These statistics must be updated regularly and an interval of 10 years is recommended. The following will be typical contents of such reports.

- Introduction
- Data availability - maps and tabulations
- Descriptive account of annual streamflow and runoff since last report
- Thematic maps of mean monthly and seasonal runoff
- Basic streamflow statistics - monthly and annual means, maxima and minima
  - for the standard climatic normal period (1961-90) where available
  - for the updated decade
  - for the available period of record
- Analysis of periodicities and trend in the streamflow data

13 REPORTING ON SEDIMENT TRANSPORT

13.1 GENERAL

The general statements made in Section 12.1 for discharge data are also valid for sediment transport data, hence reference is made to this section.

13.2 YEARLY REPORTS

13.2.1 GENERAL

The annual report provides a summary of sediment loads for the report year in terms of distribution and time. It also makes comparisons with long term statistics. Details of the observational network and data availability are included. The following are typical contents of the annual report with respect to sediment transport:

- Introduction
- Observational network for sediment sampling
  - Network layout and adaptations in report year
  - Monitoring and processing
  - Data collection in report year
- Sediment transport
  - Sediment loads in the report year
  - Sediment loads in comparison to the historical records
- Trends in sediment loads
- Interpretation of various statistics presented in the yearbook on sediment transport

13.2.2 OBSERVATIONAL NETWORK

The standards for presentation of network particulars in tables and graphs should be in line with those used for stream flow. With respect to sediment sampling the monitoring procedures and equipment used are important pieces of information to interpret the sediment loads to be presented in the

yearbook. Hence due attention is to be given in the yearbook on these aspects, particularly when changes have taken place in the procedures and equipment.

Also a clear picture should be sketched of the site conditions and sediment sources. Existence of bank erosion and mining of the river bed upstream of the measuring site, and topographical and land use practices are to be mentioned.

Next, the validation and applied computational procedures used for arriving at the load values for different time intervals (ten-daily, monthly and annually) is to be outlined.

13.2.3 SEDIMENT LOADS

Ten-daily, monthly and annual total suspended sediment loads (in tonnes) for selected stations are presented in the yearbook and for all stations the annual loads are shown. It is noted that the presented data should be based on S-Q relationships derived for the station and valid for that particular year or part of the year. To show the contributions of the coarse, medium and fine fractions to the total load the individual ten-daily, monthly and annual loads or average concentrations (in mg/L or g/L, whichever is appropriate) may also be given, when required.

To show the sediment loads for the current year, in relation with the historical data, the current year is displayed in frequency curves derived from the historical record, based on 10-daily values.

13.2.4 TRENDS

When records of sufficient length are available the long term development of loads/ concentrations for the hydrological year and for the seasons separately could be shown. To the data, information should be presented on possible causes of changes in the S-Q relationships if apparent. In this respect not only land use, bank erosion or mining of the river bed should be mentioned, but also possible effects of changed measuring equipment and/or practices. The latter may be important when investigations show that historical measuring practice (single point measurement in the vertical) has led to biased results.

14 REPORTING ON WATER QUALITY DATA

14.1 INTRODUCTION

Regular reporting on water quality data is expected to take place in the form of an annual yearbook. A yearbook is already made by CWC, and in the future can also be made by the State Surface Water organisations.

The water quality yearbook may be an independent report (e.g. CWC Water Quality Yearbook) or it may be clubbed together with reporting on hydrology and meteorological conditions to create one overall Surface Water Yearbook. If the water quality is clubbed together with hydrology, the reporting may be done for hydrological years (June – May) as opposed to calendar years (January – December).

14.2 GOALS OF WATER QUALITY MONITORING

It has been repeatedly stated that monitoring is a series of steps which follow one another in what is called the ‘Monitoring Cycle’. The cycle begins with the identification of information needs about water quality and, if all goes well, ends with the production of the requested information. This principle is

presented in Figure 14.1 and is discussed in detail in Volume 6 of the Design Manual (Water Quality Sampling).

Because the yearbook is one of the information products of water quality monitoring, it is important that the design and content of the yearbook should respond to the specified information needs and identified goals of water quality monitoring.

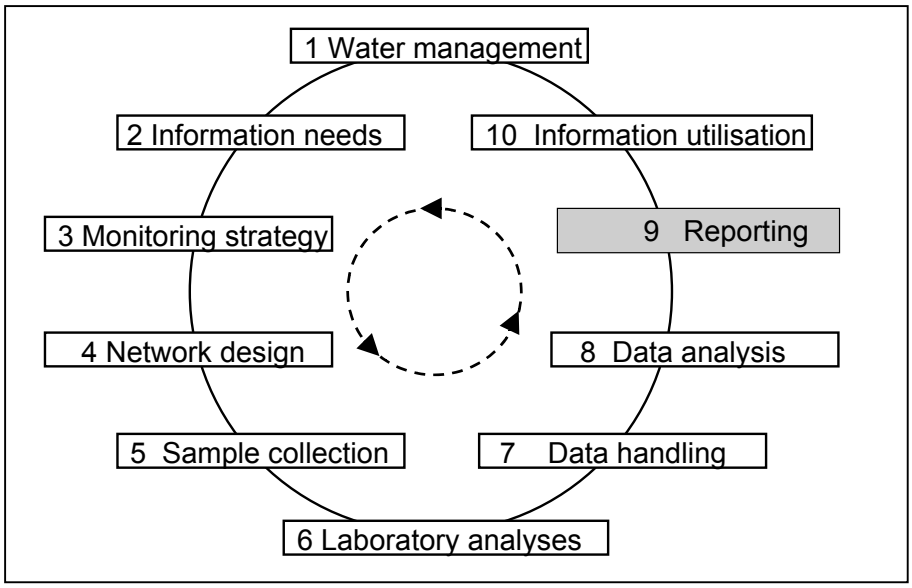


Figure 14.1 Reporting on water quality (with e.g. yearbook) is one of the steps in the monitoring cycle.

Normally water quality monitoring is conducted with one or more of the following 'global objectives' in mind:

a) to build up an overall picture of the aquatic environment thus enabling pollution cause and effect to be judged
b) to provide long-term background data against which future changes can be assessed (i.e. baseline information)
c) to detect trends
d) to provide warnings of potentially deleterious changes
e) to check for compliance or for charging purposes
f) to precisely characterise an effluent or water body (possibly to enable classification to be carried out)
g) to investigate pollution
h) to collect sufficient data to perform in-depth analysis (eg, mathematical modelling) or to allow research to be carried out
i) to assess suitability of water for various uses, such as irrigation

These global objectives can also be considered under three separate categories of sampling, i.e.:

- **Monitoring** - long-term standardised measurements in order to define status or trends (i.e.: a, b and c above)
- **Surveillance** - continuous specific measurements for the purpose of water quality management and operational activities (ie: d and e above)
- **Survey** - a finite duration, intensive programme to measure for a specific purpose (ie: f, g and h above)

These three basic sampling categories can be further split into a number of sample types, each of which have a specific objective. These sample categories, types and their associated objectives are described in Table 14.1. Naming of objectives as routine monitoring, multipurpose monitoring, etc. should be replaced by well defined terms as noted above.

As far as this manual is concerned, it is important to realise that water quality monitoring is a sub-set of the overall hydrological monitoring programme. For this reason the 'Monitoring' category identified above is the most important of the sampling categories as it will enable a complete flow and concentration (and therefore load) profile to be built up for all analytical parameters of interest in all of the catchments within the study area.

Category	Type	Objective
Monitoring	Baseline	Natural Background Concentrations
	Trend	Detection of changes over time due to anthropogenic influences
	Flux	Calculation of load Calculation of mass flux
Surveillance	Water Use	Check that water is fit for use
	Pollution Control	Check effects of discharges Check water quality standards
Survey	Classification	Classification of reach
	Management and Research	Investigation of pollution and need for corrective measures Special Interest Filling in knowledge gaps

Table 14.1: Water Quality Monitoring Objectives for different monitoring categories

14.3 COMPONENTS OF THE WATER QUALITY YEARBOOK

Whether or not water quality is reported independently or together with hydrology, there are several suggested components for the Water Quality yearbook which are presented below.

The yearbook is *not* expected to be a book full of data tables of all measured values. These data are stored in the State Data Centers, and can be obtained if specifically required. Instead, the yearbook should give an overview of the most important water quality issues in the State / river basin via a series of graphs and limited descriptive text. Suggested reporting items are given below. Additional items can be added to the yearbook if they are considered important.

1. Overview Map(s)

A yearbook for a state or should be organised by river basin. For each river basin there should be an overview map showing *all* stations of the following type:

- WQ station CWC
- WQ station state
- WQ station PCB
- Hydrologic station of CWC
- Hydrologic station of the state

Note: that both water quality and quantity stations are to be shown. This is to indicate if river discharge data are also available at (or near) the water quality station. Also, the locations of stations from different monitoring organisations are shown. This is to give an overview of all the water quality data available in the state / river basin. These maps are not expected to change significantly from one year to the next.

2. Time Series Plots

Time series plots are a way of showing water quality conditions. A plot shows results for one station and one parameter, showing all data values for the year (see example in Figure 14.2). The following guidelines should be followed:

- Make plots of selected parameters for a limited no. of stations;
- Water Quality standards (if existing) should be shown on plot.
- Yearly mean values and their confidence interval must be listed. The % of samples violating a specific standard may be shown.

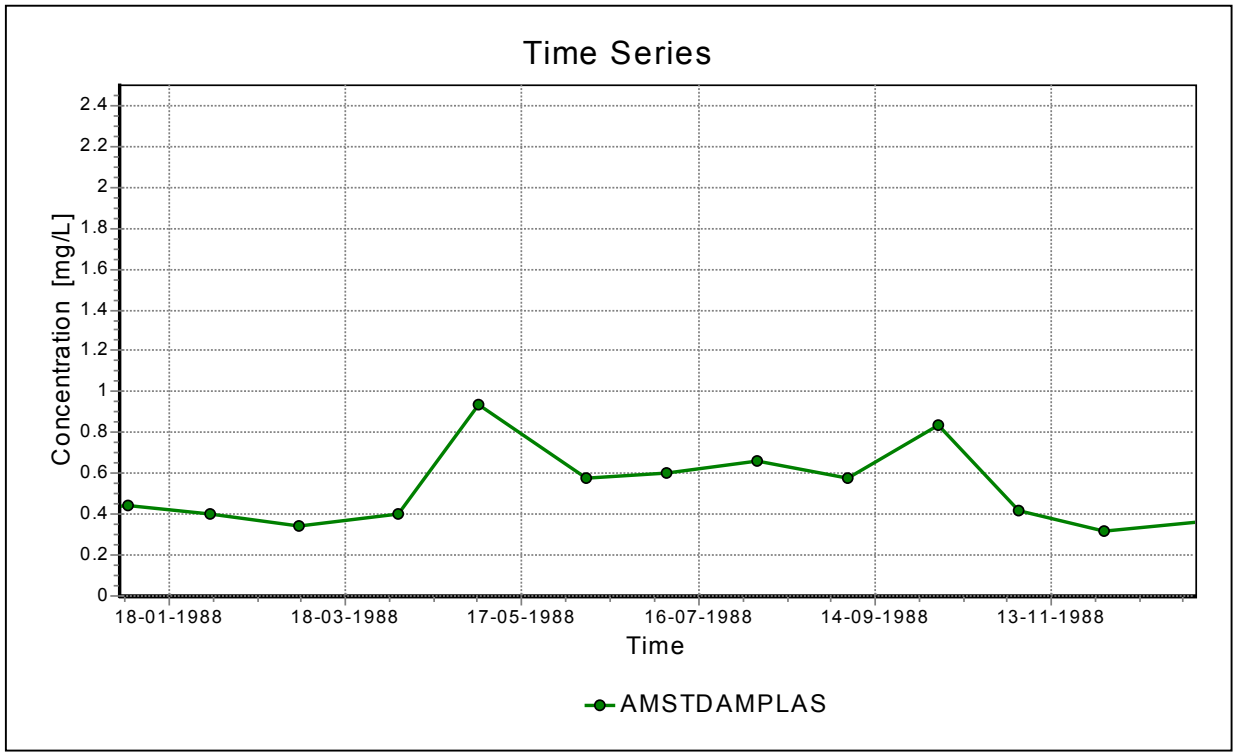


Figure 14.2: Time series plot for the year 1988 for phosphorus concentrations at Amsterdamplas (data in Table 8.21)

Example Analysis: Phosphorous is measured at location Amsterdamplas once a month. Twelve measurements have been made in the year 1988. The time series plot of phosphorus shows that concentrations are fairly constant throughout the year, though the months of May-October have elevated values. May and October have the highest concentrations (>0.9 mg/L).

Time series plots should be made for the selected parameters in Table 14.2 (if data are available). These parameters are good indicators of water quality:

Parameter	Parameter group	WQ standard (target)
Temp	General	None
TDS or EC	General	TDS 500 mg/L , (drinking water std.) EC 2250 umho/cm (for irrigation water)
SAR	Major Ions (indirect)	26 (for irrigation)
DO	General	4 mg/L (min value)
BOD	Organic matter	3 mg/L (target)
TotP & NO_3	Nutrients	 Nitrate 10 mgN/L (drinking water)
selected pollutants	Trace metal or pesticide	

Table 14.2: Suggested water quality parameters to include in water quality yearbook (minimum list)

3. Plotting historical data (multiple years)

The purpose of plotting the data for multiple years is to see how a given year compares with measurements from previous years. Linear trends or step trends may then be apparent. Two main options exist for plotting historical data.

Time Series with historical data

This time series plot will be similar to that under Step 2, but will include the historical data.

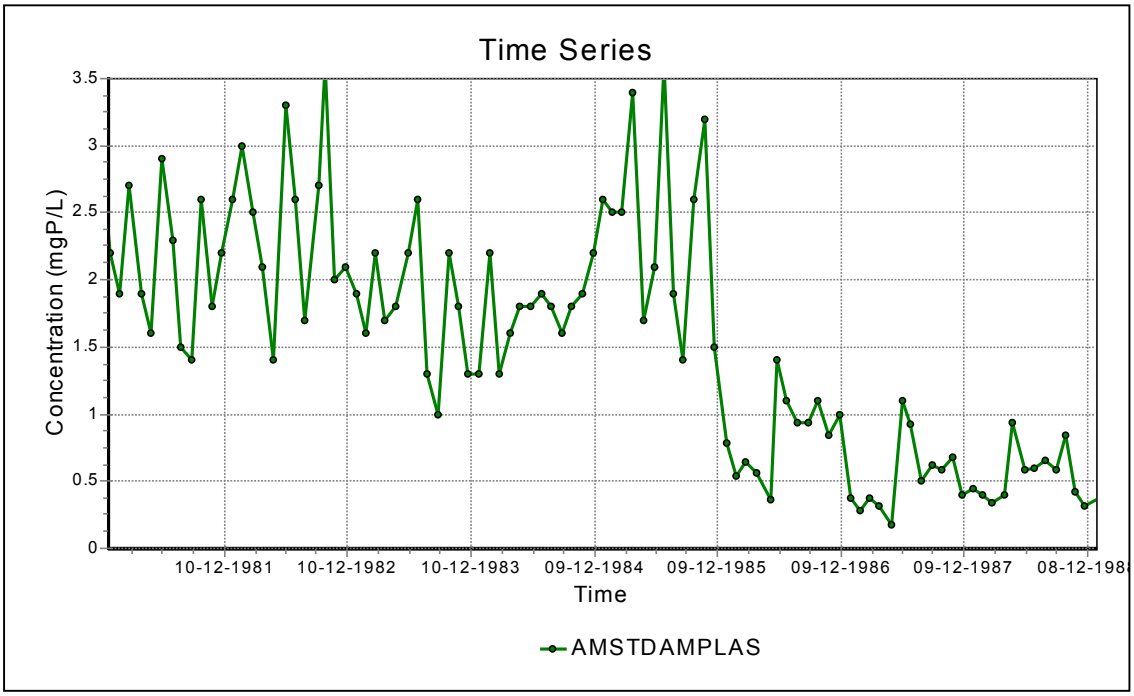


Figure 14.3: Time series plot for the years 1981-88 for phosphorus concentrations at Amsterdamplas (data in Table 8.21).

Annual box and whiskers plot

An (annual) Box-Whisker plot is an ideal way to see the given year compared to previous years. Because the plot shows only the main statistics of each year, (e.g. min, max, and mean) it is easy to see important changes from year to year.

The following general guidelines should be followed for either the historical time series plot or the annual Box-Whisker plot:

- A plot shows results for one station and one parameter, for multiple years (see example in Figure 14.4).
- Make plots of selected parameters for a limited no. of stations.
- Make plots for the same parameters and the same stations as in Step 2.

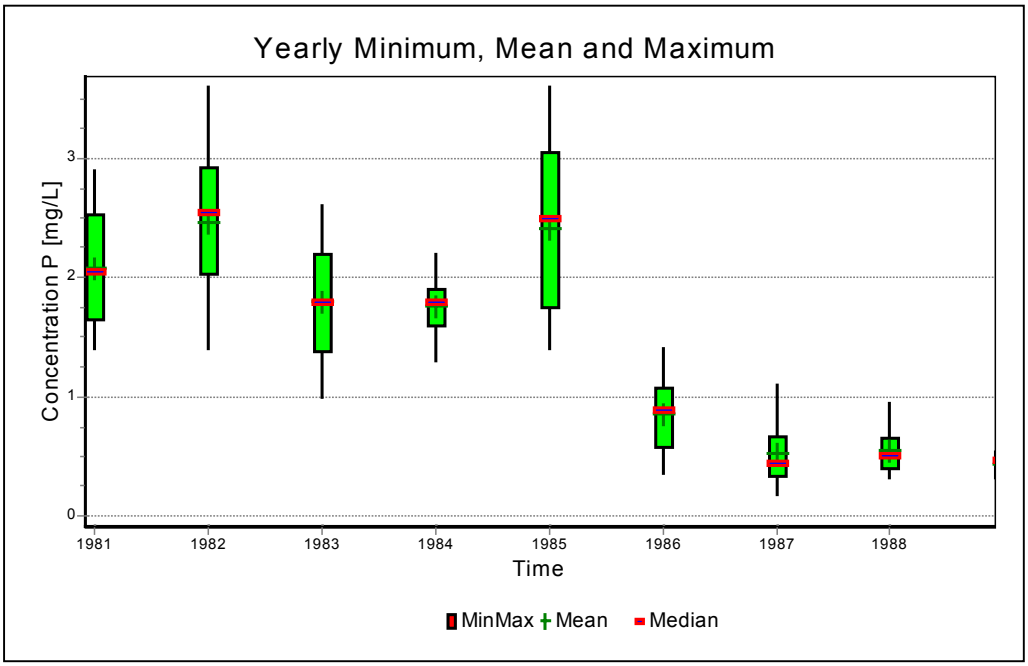


Figure 14.4: Annual box and whiskers plot for phosphorus concentrations at Amsterdamplas for 1981-1988 (data in Table 8.21)

Analysis: From the plots showing the historical data, one can see that 1988 has the lowest measured P concentrations for the period 1981-1988 at location Amsterdamplas. The concentration in 1988 never is above 1 mg/ while in several previous years the mean concentration has been greater than 2.5 mg/L.

4. Comparison of stations for a given year

The purpose of plotting the several different stations for a given year is to see the relative concentrations at different locations. With such a plot it is easy to answer questions such as:

- Which station had the *highest* measured concentration of parameter (x) this year?
- Which station had the *lowest* measured concentration of parameter (x) this year?
- Are the highest concentrations all in the same river basin?

The following guidelines should be followed:

- A plot shows data for multiple stations and one parameter. Data for each station are shown with box-whiskers drawing using data for 1 year (see Figure 14.5).
- This plot can be made for all stations within a river stretch or within a state or basin.

- A separate plot is made for each parameter. All stations where there is available data can be included in the plot. Make plots for the selected parameters as in Step 2 (Table 14.2). Additional parameters may also be selected for plots (if they show relevant information).

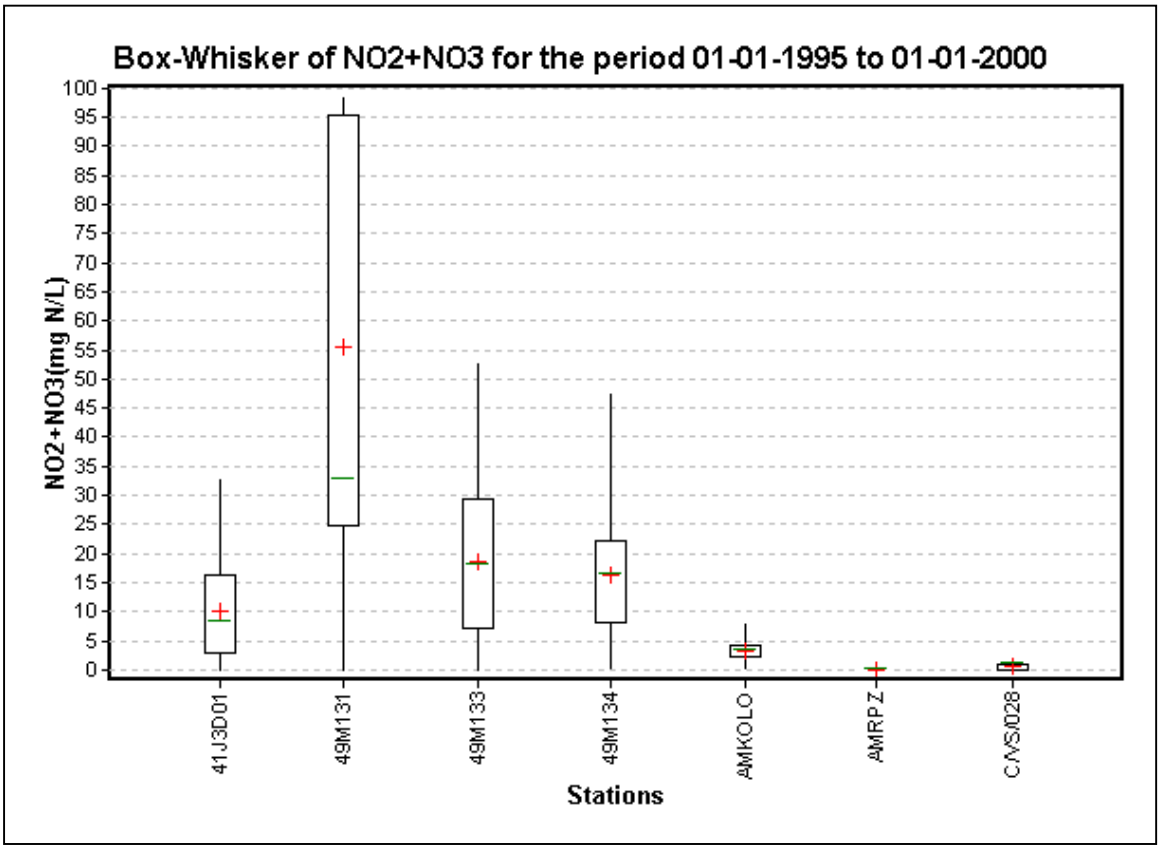


Figure 14.5: Comparison of water quality (concentration P) in 1988 at 6 different locations

5. Mass load calculations

For selected stations (of Flux type) and for selected parameters, e.g. BOD, solids, total Nitrogen and total Phosporus yearly mass loads (expressed in kg/year) should be presented at relevant locations such as state boundary crossings or river flow in to lakes and reservoirs.

6. Water Quality Classification

The Central Pollution Control Board (CPCB) has identified predominant uses of water, calling them designated best use, of different water bodies or stretches of river and also defined water quality criteria for different uses of water. These criteria are given in Table 14.3. Class A is the best water quality while Class E is the lowest quality.

Based on the monitoring data, the water quality is compared with the criteria for the designated best use. If the required water quality parameters have been measured at a locations, then the station can be classified as type A-E. The water quality must meet *all* the criteria in a class order to be in the class. If one of the criteria for a class is *not* measured, then the site cannot be classified in that class; e.g. if coliforms or BOD are not measured at a site, then it is not possible to say if the site is of Class A, B or C since these classes have criteria limits for both coliforms and BOD).

Designated best use	Class	Criteria
Drinking water source without conventional treatment but after disinfection	A	Total coliform organisms MPN/100mL shall be 50 or less. pH between 6.5 and 8.5 Dissolved oxygen 6 mg/L or more Biochemical oxygen demand 2 mg/L or less
Outdoor bathing (organised)	B	Total coliform organisms MPN/100mL shall be 500 or less pH between 6.5 and 8.5 Dissolved oxygen 5 mg/L or more Biochemical oxygen demand 3 mg/L or less
Drinking water source with conventional treatment followed by disinfection	C	Total coliform organisms MPN/ 100mL shall be 5000 or less pH between 6 and 9 Dissolved oxygen 4 mg/L or more Biochemical oxygen demand 3 mg/L or less
Propagation of wild life, fisheries	D	pH between 6.5 and 8.5 Dissolved oxygen 4 mg/L or more Free ammonia (as N) 1.2 mg/L or less
Irrigation, industrial cooling, controlled waste disposal	E	pH between 6.0 and 8.5 Electrical conductivity less than 2250 micro mhos/cm Sodium absorption ratio less than 26 Boron less than 2mg/L

Table 14.3: Primary water quality criteria for various uses of fresh water

An overview of the water quality in a given river basin or state can be given based on this classification system. For example, with the following table format:

Class	No. of Stations in Class
A	3
B	2
C	4
D	4
E	6
NC (Not Classified)*	10
Total no. Stations monitored	29

* due to insufficient parameter information

Table: 14.4: Example (hypothetical) overview of water quality for monitored stations

In the above hypothetical example, 29 stations are monitored in a river basin (or state), of which 10 cannot be classified due to insufficient parameter information (if for example coliforms were not measured). The other 19 stations have been classified as type A-E.

Figure 14.6 shows the application the designated best use criteria to some stations using a colour (see legend) to indicate the fitness for a particular use (drinking and or bathing, wildlife and irrigation) evaluated per parameter

Fitness for use classification according to CPCB system "ABCDE"

1996	Water for Drinking & Bathing (ABC)				Wildlife (D)		Irrigation (E)			
Station	pH	CTM	BOD	DO	pH	DO	pH		B	SAR
AG000C3	8.2	-	0.4	6.4	8.2	6.4	8.2	212	-	0
AG000C7	-	-	-	-	-	-	-	-	-	-
AG000G7	8.3	-	0.5	6.4	8.3	6.4	8.3	339	-	1
AG000J3	8.4	-	0.7	6.5	8.4	6.5	8.4	468	-	1
AGH00C4	8.1	638.7	1.3	6.8	8.1	6.8	8.1	393	-	1
AGH30E2	8.1	3742.8	3.7	6.0	8.1	6.0	8.1	515	-	1
AGH30F6	-	-	-	-	-	-	-	-	-	-
AGH30Q1	8.2	502.8	1.0	7.2	8.2	7.2	8.2	413	-	1
AGH30S9	8.2	166.5	1.7	6.1	8.2	6.1	8.2	410	-	1
AGH32D5	8.2	1502.9	1.2	7.2	8.2	7.2	8.2	429	-	1

Legend

Drinking water (A), Bathing water (B) and Source for drinking water (C)												
pH		Col		BOD		DO						
6	-	50	A	2	A	4	-					
6.5	C	500	B	3	B	5	C					
8.5	A	5000	C		-	6	B					
9	C		-				A					
	-											
Wildlife (D)												
pH		DO										
6.5	-	4	-									
8.5	D		D									
	-											
Irrigation (E)												
pH		EC		B		SAR						
6.5	-	2250	E	2	E	26	E					
8.5	E		-		-							
	-											

Water Quality Yearbook - HYMOS example Report 4 (HP, 2002)

Figure 14.6: River water fitness for various stations (vertical). Three types of use are distinguished here: drinking water preparation and or bathing (classes A,B,C), wildlife (class D) and irrigation (class E). If the water quality data (90-percentile of time period) violate the criteria listed in Table 13.2, the value is coloured according to the legend. Fitness increases from red through orange and yellow to green.

7. Results of Survey Monitoring

Surveillance monitoring will usually not be conducted at the same locations or for the same parameters from year to year. Thus presentation of these results needs to be separate from items 1-4 above. Since surveillance monitoring is conducted for problem issues, it is assumed that the results of these studies is relevant for a yearbook. Surveillance monitoring likely will be for pollution parameters such as (e.g. coliforms, heavy metals, pesticides, organic pollutants, ammonia).

Results of survey monitoring to be presented include:

- target of the study, problem definition, objectives of monitoring
- Map showing location(s) monitored
- Time series plots for 1 year, for the WQ parameters included in the surveillance. The water quality standard must also be shown .
- Comparison of stations for the period of monitoring
- Conclusions of the monitoring programme with respect to the objectives.

15 REFERENCES

Abramowitz, M. and I.A. Stegun (eds.) (1964)
Handbook of Mathematical Functions
National Bureau of Standards, USA, Applied Mathematics Section, Publ. No. 55, Dover, New York

Box, G.E.P. and D.R. Cox (1964)
An analysis of transformations
Journal of RSS B26, pg 211-246

Cunnane, C. (1978)
Unbiased plotting positions – a review
Journal of Hydrology, 37, pg 205-222, Elsevier, Amsterdam

Dingman, L.S. (2002)
Physical Hydrology
Second Edition, Prentice-Hall, Inc. New Jersey, USA

Doorenbos, J. and W.O. Pruitt (1977)
Guidelines for predicting crop water requirements
FAO Irrigation and Drainage Paper 24 2nd ed., Rome

Gilbert, R.O. (1987)
Statistical Methods for Environmental Pollution Monitoring
John Wiley & Sons Inc., New York

Gopal K. K. (1999)
100 statistical tests
SAGE Publications, New Delhi.

Gupta, Shekhar, Vasudev and P. N. Modi (1991)
A regression model for potential evapotranspiration estimation
Journal of Indian Water Resources Society, Vol 11, No 4 pp 30 – 32

Haan, C.T. (1977)
Statistical methods in Hydrology
Iowa Tate University Press

Hald, A. (1952)
Statistical theory with engineering applications
John Wiley, New York

Isaaks, E.H,. and R.M. Srivastava (1989)
Applied Geostatistics
Oxford University Press, New York

Jenkinson, A.F. (1969)
Estimation of maximum floods
Techn. Note No. 98, Chapter 5 pp 183-257: General extreme value distributions, World Meteorological Organisation, Geneva

Klemes, V. (2000)
Tall Tales about Tails of Hydrological Distributions, Part I & II
Journal of Hydrologic Engineering, Volume 5, No 3, July, 2000

Kottegoda, N.T. and R. Rosso (1997)
Statistics, Probability and Reliability for Civil and Environmental Engineers
McGraw-Hill Companies Inc, Civil engineering series, New York

McBean, E.A. and F.A. Rovers (1998)
Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment, Prentice Hall PTR Environmental Management and Engineering Series, Volume 3.

Metcalfe, A.V. (1997)
Statistics in Civil Engineering
Edward Arnold, London

NERC (1975)
Flood Studies Report
National Environmental Research Council, London

Penman, H.L. (1948)
Natural evaporation from open water, bare soil, and grass
Proceedings of the Royal Society of London, Series A, 193: 120-145

Ramasastri, K. S. (1987)
Estimation of evaporation from free water surfaces
Proceedings of National Symposium on Hydrology Roorkee (India), pp II – 16 to II – 27

Reddy, P.J. (1996)
A Text Book of Hydrology
Laxmi Publications (P) Ltd, New Delhi

Reddy, P.J. (1997)
Stochastic Hydrology
Laxmi Publications (P) Ltd, New Delhi

Shaw, E.M. (1988)
Hydrology in Practice
Van Nostrand-Reinhold, London

Subramanya, K. (1994)
Engineering Hydrology
Second Edition, Tata McGraw-Hill Publishing Company, Ltd, New Delhi

Upadhya, D.S. et. al. (1990)
Space correlation structure of rainfall over India.
Mausam 41, 4. pp 523-530.

US Water Resources Council (1976)
Guidelines for Determining Flood Flow Frequency
Washington D.C.

WL|Delft Hydraulics (2002)
HYMOS 4, Processing System for Hydrological Data, User Manual
WL|Delft Hydraulics, Delft, Netherlands

Venkataraman, S. and V. Krishnamurthy (1965)
Studies on the estimation of Pan evaporation from meteorological parameters
Indian Journal of Meteorology and Geophysics, Vol.16 , No.4 pp 585 - 602

World Meteorological Organisation (1966)
Measurement and estimation of evaporation and evapo-transpiration
WMO Technical Note No. 83, Geneva

World Meteorological Organisation (1973)
Comparison between pan and lake evaporation
WMO Technical Note No. 126, Geneva

World Meteorological Organisation (1980)
Manual on Stream Gauging
WMO, Operational Hydrology Report No 13, Geneva

Yevjevich, V. (1972)
Probability and Statistics in Hydrology
Water Resources Publications, Colorado State University, Fort Collins, USA

ANNEXURE I: SPECIMEN FOR SURFACE WATER YEAR BOOK

Surface Water Hydrological Data

<Name of State/Region>

<YYYY>
YEARBOOK

STATE WATER DATA CENTRE <NAME OF STATE/REGION>

Surface Water Hydrological Data
<Name of State>
<YYYY>
YEARBOOK

An account of rainfall, river flows and water quality

FOREWORD

<The yearbook may include a foreword by an officer considered suitable by the agency. This person can typically be the Chief Engineer who has the overall authority and responsibility of the functioning of the HIS in that agency>

…
Chief Engineer

Table of Contents

1 Introduction

<The text given hereunder is not the actual text which must be included as the “Introduction” but could help in appreciating the background with which the new style yearbook is intended to be developed. Also all the examples used in the text, including tables and figures, are only indicative and do not refer to any actual case, basin or state.>

<For briefing the readers adequately on the background, the first of the new style yearbook may include the text giving the evolution of yearbooks in the past and how the contents of the new style would be in line with the user needs and the available technology. This text must also bring up issues of switching over from paper yearbook to the electronic yearbook. And also how the paper yearbook may still be relevant to be produced, however in very less numbers, for and on demand from specific users. One of the benefits, which the water yearbook can still bring for the HIS in the country, is the fact that the target of preparing the yearbook itself makes it mandatory on the part of the system to finalise data in time and produce water yearbooks within the prescribed time frame. It also is a tangible output of the system by which the accomplishment of data observation and data finalisation for the year under consideration can be seen.>

<It is very important in these times to evaluate which medium will be most suited for publishing the yearbooks. The traditional way of bringing out the yearbook as printed documents could now turn towards electronic yearbooks. The electronic yearbooks may still have the same content (one may even afford to include more, in fact, if required) and structure as the hard copy yearbook. However, they are presented to the users in the form of a CD or may even be accessible (in a controlled manner as per the guidelines of the agency) through internet instead of distribution as hard copies. First of all, the system has to ascertain the genuine requirement of the hard copy water yearbooks, as printing yearbooks in large numbers may need a lot of funding. Also, hard copy yearbooks may not always be so effective a medium for the users to get data and information, specially nowadays when most of the information flows digitally. However, there would always be a need to also have a hard copy yearbook, even if it will be required in very few numbers. One of the benefits of the hard copy yearbook is that it serves as an additional paper archive of most of the data. Secondly, for certain users and situations, hard copy yearbooks will be the preferred medium as against the electronic yearbook, which always necessitate availability of a computer. An objective approach could be followed while deciding on the extent upto which the contents of the yearbook has to be printed and distributed as hard copies. For the purpose of archives and as the key reference document for internal use by the agency (like, in the data centres and in design and planning wings), it may be required to print the whole yearbook. Such full copies could be in a limited number. However, the hard copies that are required for distribution to other hydrological data users, may be limited to an abridged version of the whole yearbook. This abridged version may include most of the items of the full yearbook except all the data tables. Data tables for only very few stations could be included, mainly as samples. For such readers, it is assumed that they would not require to refer data for any particular station. They would only be interested to know about the hydro-meteorological and hydrological behaviour in the region in general and that they could preferably use the electronic water year book for any reference whenever needed.>

<The major change in the style of water yearbook is in introducing more graphs and pictures that may enable necessary and adequate comprehension that the reader may like to have about the hydro-meteorology of the region. Pictorial options allow large volumes of data to be summarised in a nutshell. Notwithstanding the fact that these graphical representations could be sufficient for most of the readers and also the fact that most of the data could be readily made available from the organised databases to the requesting users, it would still be essential that the data is also presented and available in well laid-out data tables. As was the case in earlier editions of the yearbooks, it would be desirable that most of the data on daily or larger interval is presented in these tables. Such availability of nicely laid-out tables would enable the data to be presented in an attractive format whenever some reference is required to be made.>

<Another addition to the earlier yearbooks may be by including few interesting articles on some relevant themes of the hydrological regime of the region. It would always be interesting for any reader to get to know about some significant trends in the rainfall or flow or water quality patterns in the region. Floods and droughts continue to haunt people in most of the regions of our country. It could be appropriate to highlight some of the hydrological features of such extreme flood and drought situations. Similarly, awareness about quality of water has grown manifold in the last 1-2 decades due to enormous pressure on water as natural resource and unavailability of good water in sufficient amounts for most of the uses. In such a situation it can be helpful if some alarming water quality situations are highlighted in the form of articles. Besides these articles being informative on one hand, they would many times be eye openers for the policy makers and managers of the water resources systems. At the same time they would make interesting reading for others and also incidentally help in reminding the personnel working for the HIS to understand and appreciate the importance of data being collected and information being derived.>

<The central idea of the yearbook is to review and communicate, to the target readers, what kind of hydro-meteorological and hydrological scenario prevailed in the region during the year under consideration. For this, it is appropriate to first give information about the water resources, drainage system and land use in the region followed by the layout of the monitoring network on the basis of which all the information is obtained and derived. Both, hydro-meteorological and river gauging (including water quality) networks can be shown in the form of maps. Further, hydrological reviews could be given which describe the behaviour of the hydrological processes in the region. Various types of data, viz. rainfall, evaporation, river levels and flows and water quality can be summarized with the help of graphs and data tables. Graphical illustrations showing the process during the year under consideration against the long term pattern could be very effective. Together with the graphical illustrations summarizing the hydrological and hydro-meteorological behaviour, it is worthwhile to tabulate the daily data and the important monthly and yearly statistics along with it. Such tabulations enable easy referencing to any particular data at any point of time, without requiring to interrogate the databases for retrieving the same.>

<Further, it could be valuable for the readers, to include valuable reference material such as bibliography of previous yearbooks or other information sources on the matter, brief notes on procedures followed for observations in the field and notes on the interpretations of the terms used in the yearbook.>

2 Water and Life in <put the name of the region being reported upon>

<The idea is to include few articles in the yearbook so as to make it appealing to various users. Traditionally, the yearbooks have been focusing more on presenting the data tables and less on highlighting some of the extreme events or events which may be of concern or interest to the users in general and designers, planners or managers in particular. Typically, what interests more is something that is different from the average, like severe storms, flooding or drought situations or a changing pattern in rainfall or flows in the region. Similarly, growing concern for water quality warrants bringing up some of the deteriorating WQ situations in the region, for the benefit of the users. Some such articles that are informative and sometimes even can be an eye opener could be written by DPC staff members and included in the yearbook. It is not essential that the yearbook must necessarily contain a specific number of articles. Emphasis would be more towards capturing the readers' attention rather then simply producing the data tables.>

2.1 Article 01 – Extreme rainfall events of 1997

<Such article can include some of the heaviest rain events of the year under consideration. Severe events of particular duration like 1 hr., 3 hr., 6 hr., 12 hr., 1 day, 2 days or 3 days could be identified. A comparison of such identified events with the previously recorded heaviest storms in the region can indicate the importance of such events for derivation of revised heaviest storms for given duration.>

<Illustrations are given in Figure 2.1 and 2.3, showing the spatial distribution of 1 day and 2 days heavy storm over the catchment for the year under consideration. Temporal distribution of this storm as recorded at two stations is also given subsequently in Figure 2.2 and 2.4. Such graphs can help understand the type of temporal distributions of these extreme events. Temporal distribution of previous heaviest 1 day event in the region as recorded at one of the stations is also shown in Figure 2.5 for the purpose of reference.>

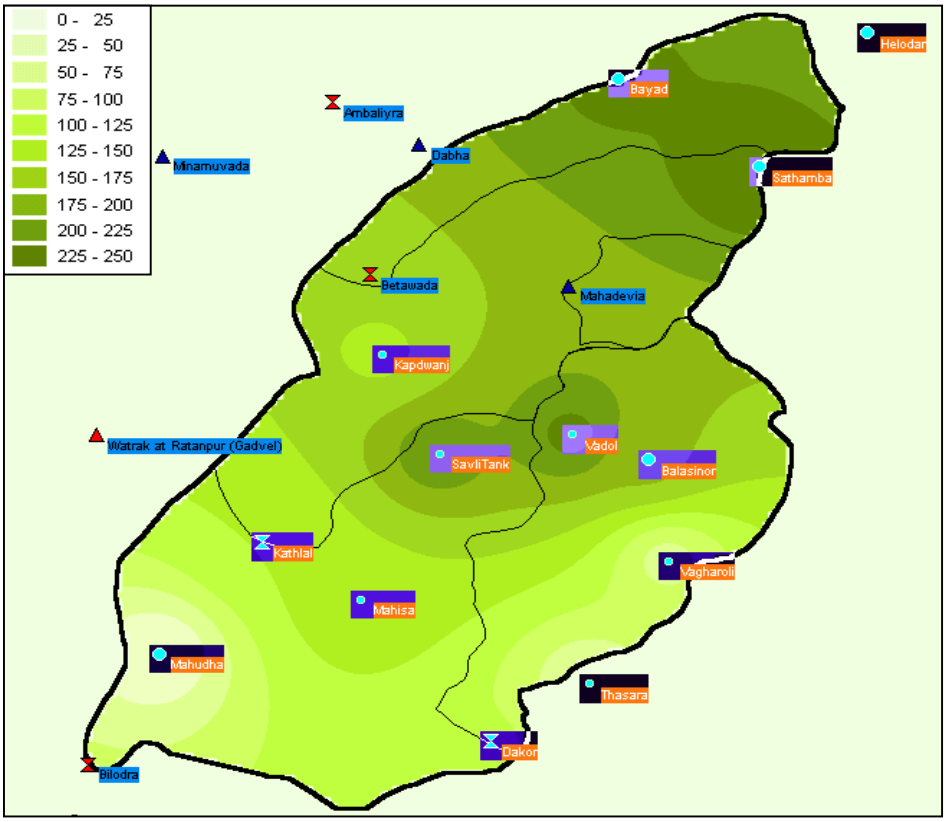


Figure 2.1: Spatial distribution of 1 day storm over the catchment (1 August 1997)

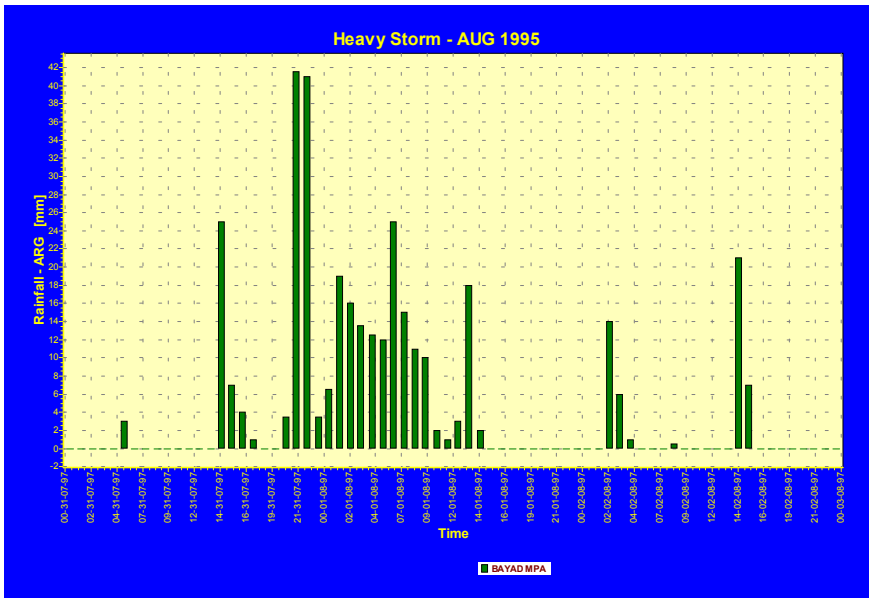


Figure 2.2: Temporal distribution of rainfall on 31 Jul - 1 Aug. 1997 at Station Bayad (290 mm)

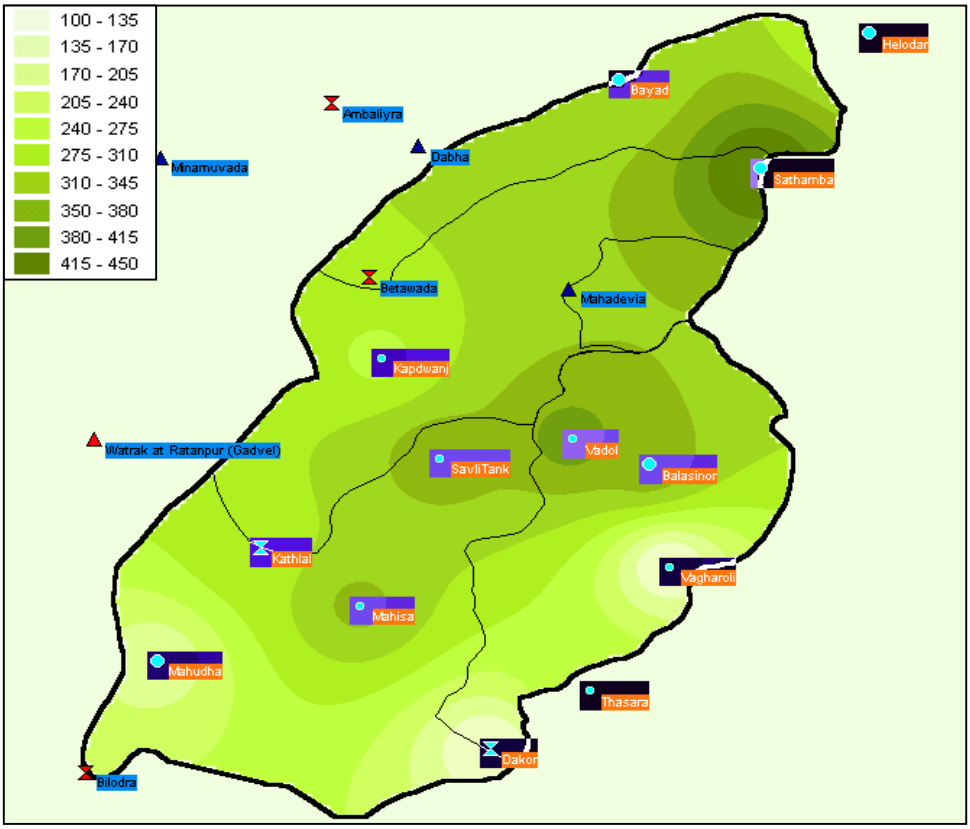


Figure 2.3: Spatial distribution of 2 day storm over the catchment (1-2 August 1997)

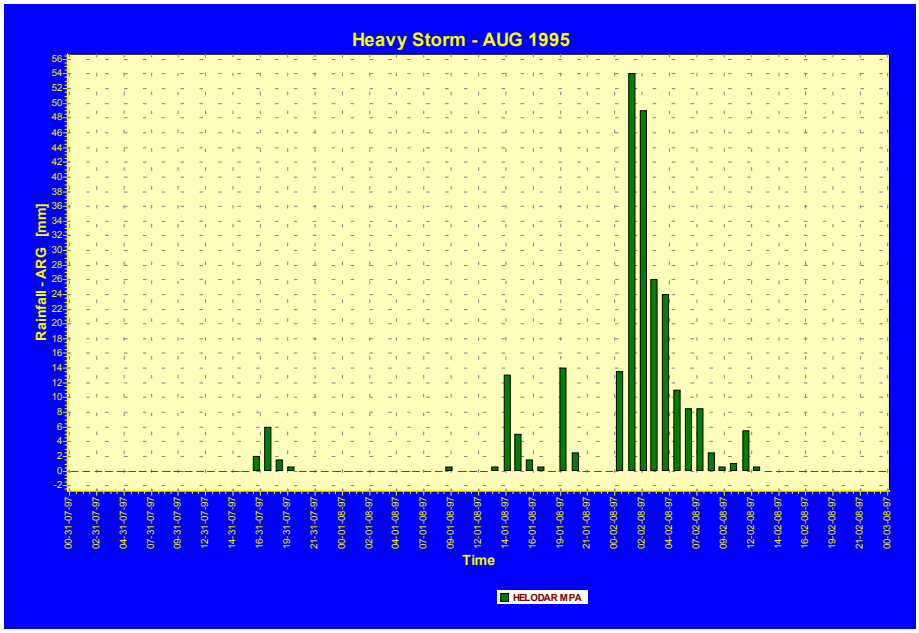


Figure 2.4: Temporal distribution of rainfall on 1-2 August 1997 at Station Helodar (240 mm) (seems to have a 1 day shift in the data)

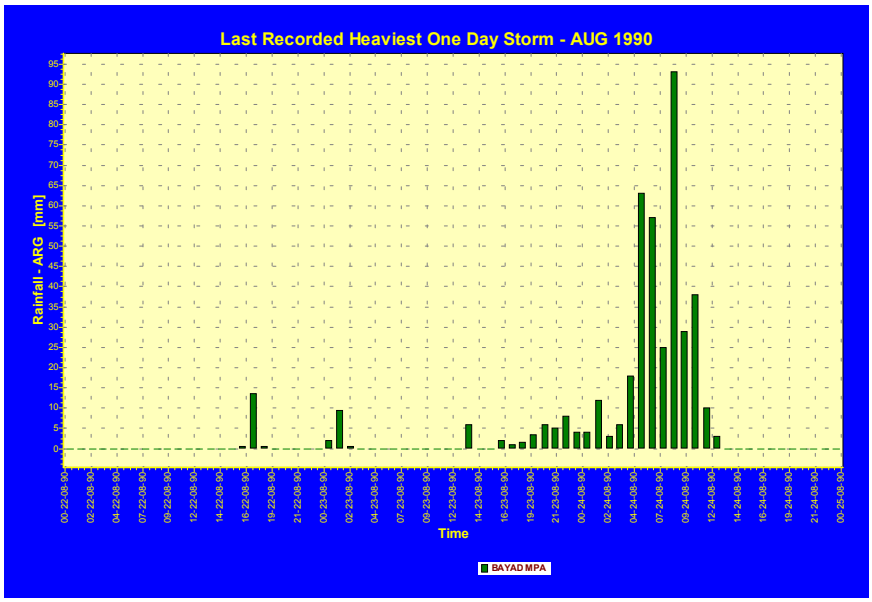


Figure 2.5: Temporal distribution of heaviest 1 day storm recorded in past - 23-24 August 1990 at Station Bayad (398 mm)

2.2 Article 02 – Flooding in the region

<One of the important objectives for continuously monitoring and organising the hydrological data is to be able to mitigate natural disasters as floods and droughts. Inspite of the best efforts, these hazards continue to haunt the societies or settlements. In many cases there still remains a lot to be done in terms of providing further protection from floods and droughts by taking specific structural and non-structural measures. >

<In view of this, it would be appropriate to include some interesting articles on extreme flooding, experienced in parts of the region in the year of reporting. These articles could be supported by some illustrative photographs and hydrographs or other graphs. These articles could create more interest about such extreme events and about hydrological data in general. Incidentally, such articles may provide good reference to such events at a later date to the interested investigators.>

Figure 2.6 Scenes of severe flooding and damage in the region

<Together with giving an account of the severe flooding, with the help of text and photographs, it could be appropriate to include hydrographs at certain locations along the river during the flooding period. Such graphs would give a good view of how long the waters remained above the danger levels and how severe was the flooding. >

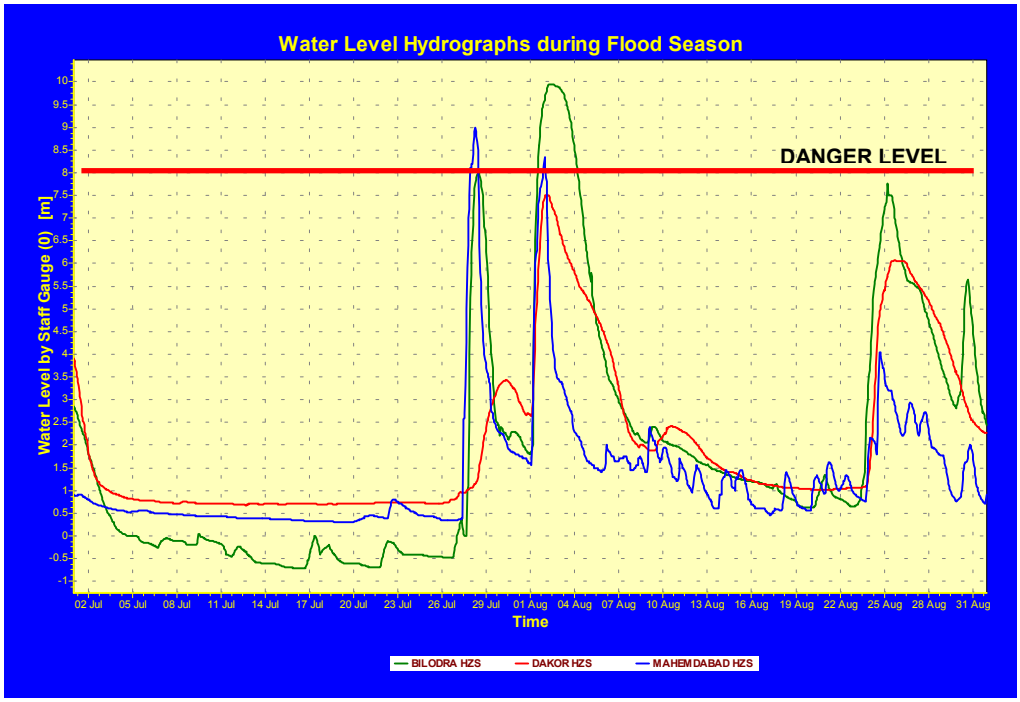


Figure 2.7: Water levels in the rivers at specific locations during flood season

2.3 Article 03 – Drought condition in the region

<Similar to floods, many of the regions in most of the states of the country continue to remain vulnerable to drought conditions. It would be appropriate to bring out some salient features of the droughts passed by, so as to arouse curiosity and generate interest among the data users in general and people responsible for drought mitigation measures in the region in particular.>

<One of the benefits that comes from the preparation of such articles is that the officers engaged in processing the data tries to look at the data closely, learn about the behaviour of the hydrological regime in the region, and becomes familiar with the past important events.>

<Such articles on droughts could be supported by some very illustrative photographs and patterns of the droughts both spatially and temporally and also magnitude wise. Such pictures and graphical patterns are given as examples in Figures 2.8 to 2.11 respectively.>

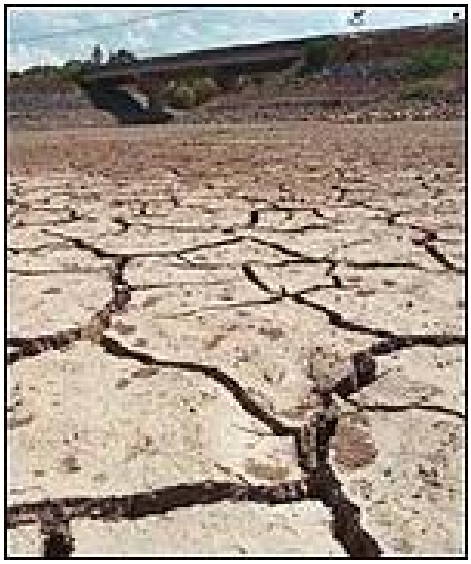


Figure 2.8:
Scenes of severe drought situation in the region

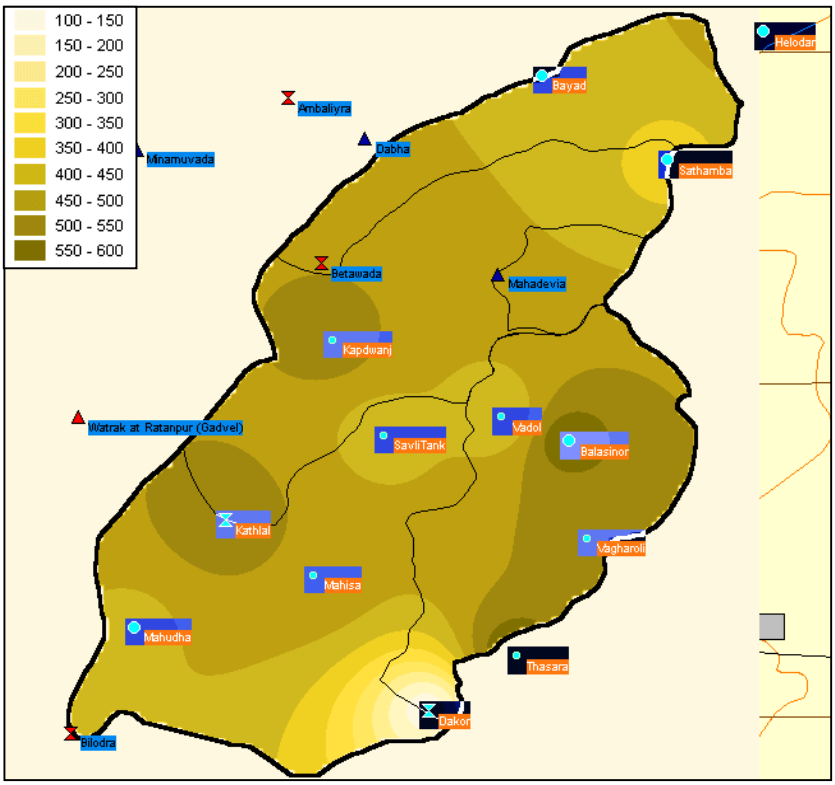


Figure 2.9:
Annual rainfall pattern in the drought stricken region

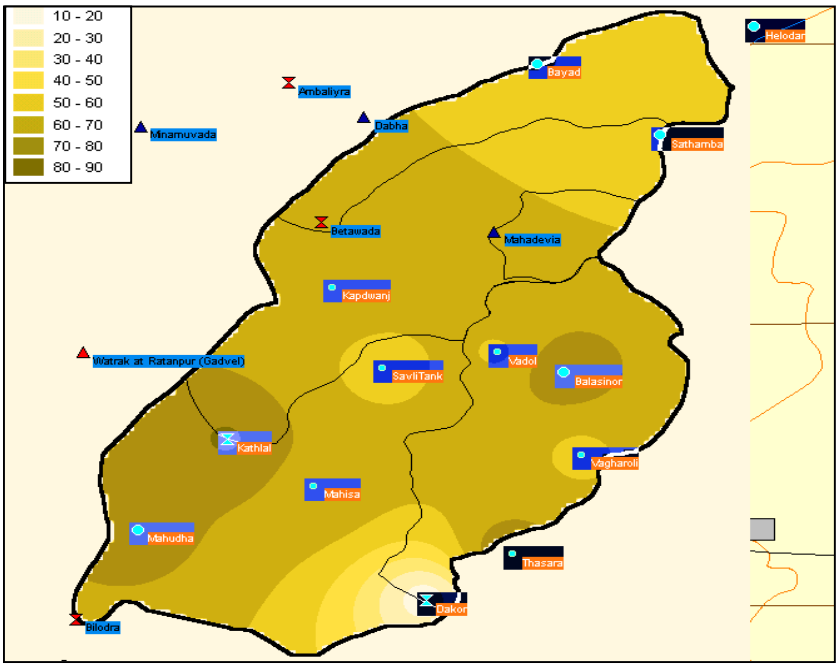


Figure 2.10: Rainfall pattern as percentage of long term annual rainfall

2.4 Article 04 – Trends in rainfall in the region / flows in …<put name of river>

<There is tremendous pressure on water as a resource, specially in our country due to increasing population. At the same time there is growing concern about possible changes in rainfall and runoff regimes in the regions due to global and/or local factors. There is substantial shift in the land use pattern in most of the regions of the country where more and more areas are being brought under cultivation, settlements and industries. In such circumstances the earlier flow patterns are likely to change. One of the basic objectives of hydrological monitoring is to keep abreast with these changing patterns of hydro-meteorological and hydrological factors in the region. In order to highlight any such significant shifts in the patterns, it will be appropriate to include illustrative articles documenting such trends with the help of graphs and interpretations.>

2.5 Article 05 – Water Quality concerns

<As for articles on quantitative aspects of the hydro-meteorological regimes, it would be highly useful and relevant to include articles on water quality aspects. A short example is given hereunder that is having only an indicative value.>

Trends in water quality

A time series plot for BOD (3 years period i.e. from 1996 to 1998), all dates and annual average is plotted as shown in Table 2.1 and Figure 2.11 below. As revealed from the graph BOD values up to 1997 varied between 0.1 and 1.1 mg/L with an average of 0.45 mg/L. The observed increase of the maximum and average value in 1998, 1.4 and 0.67 mg/L respectively are very small compared to the large spread of the data, caused by the sharp decrease in the number of observation in 1998. The data therefore do not indicate a significant increase in BOD at this station.

☞ ***A similar plot as presented in Figure 2.11 may be included for longitudinal analysis of a river or river stretch: different monitoring stations are presented on the horizontal axis of the graph.***

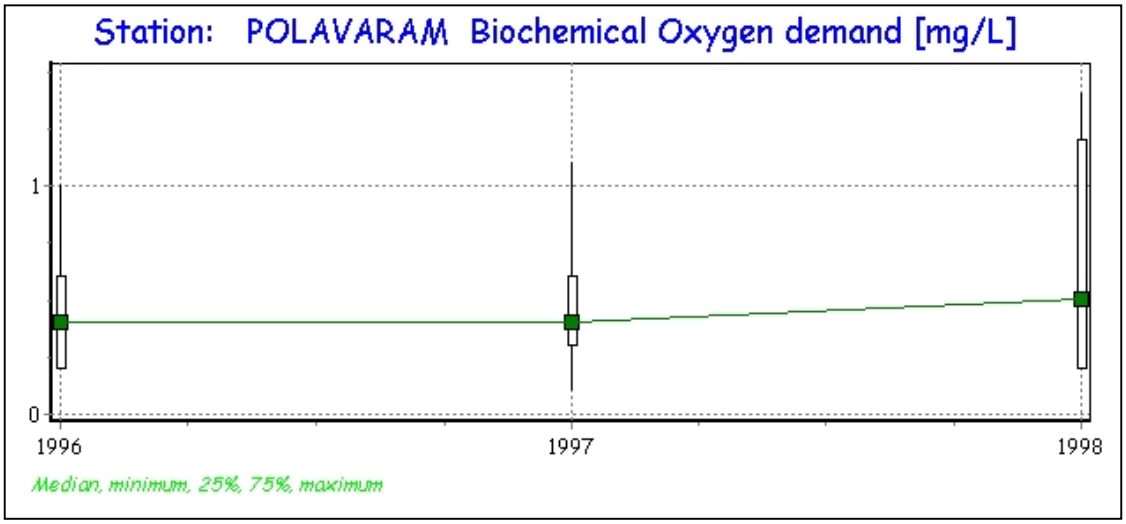


Figure 2.11: Box-whisker graph for BOD at station Polavaram

Year	1995	1996	1997	1998	1999	2000
Max	-	1.000	1.100	1.400	-	-
Mean	-	0.444	0.463	0.667	-	-
Min	-	0.200	0.100	0.200	-	-
Median	-	0.400	0.400	0.500	-	-
10%	-	0.200	0.200	0.200	-	-
25%	-	0.200	0.300	0.200	-	-
75%	-	0.600	0.600	1.200	-	-
90%	-	0.900	0.800	1.400	-	-
No. data	0	32	32	6	0	0

Table 2.1: Yearly time-series for summary statistics for BOD at station Polavaram

3 Hydrological Information System

3.1 Water Resources of …<put the name of the region being reported upon>

<This section would highlight the salient features of the surface water resources available in the region by briefly introducing all major river basins, natural lakes and artificial reservoirs. Together with the appraisal of the drainage and water resource systems available in the region, it would be appropriate to highlight the various types of uses the land is put to. The uses like forest, grazing, agriculture, industry, urban settlement etc. will help in appreciating the tremendous pressure on water in various places of the region. This will indirectly also emphasis the need to better manage the available waters.>

<Much of the information will be available through various maps included in this section, as listed here under. >

<**MAP 01 -** A MAP OF THE PHYSIOGRAPHIC FEATURES – ELEVATION, RIVERS & BASINS, LAKES, RESERVOIRS>

<**MAP 02 -** ANOTHER MAP THAT MAY BE USEFUL IS THE LAND USE MAP OF THE SAME REGION AS GIVEN IN MAP ABOVE >

<ALL MAPS TO BE GIVEN ON FULL PAGES>

< In fact, such a graphical presentation of the water resource systems will provide the background and link for why a good hydrological information system is needed in the region for managing the available water resources appropriately.>

< Together with the information provided by the maps it will be appropriate to include brief description of salient features of various river basins in the state / region. A sample is given hereunder that can further be expanded to include more relevant information in a crisp manner.>

Godavari Basin (only as an example)

The Godavari river basin is one of the 14 major river basins of India having a catchment area of 3,12,812 km^2 which is nearly 10 percent of the total geographical area of the country. It spreads over Maharashtra (48.7%), Madhya Pradesh (20.8%), Andhra Pradesh (23.4%), Orissa (5.7%) and Karnataka (1.4%). The river traverses a distance of 694 km through Maharashtra and 771 km through Andhra Pradesh, totalling 1,465 km, before discharging into the Bay of Bengal.

The major tributaries of the river Godavari are Pravara, Purna, Bindusara, Manjira, Penganga, Wainganga, Wardha, Pranahita, Indravathi, Maner and Sabari.

3.2 Hydro-meteorological and hydrological observation system

3.2.1 General

<The text in this section would include the background on hydro-meteorological and hydrological observations in the area under consideration and how such observation systems have evolved over a period of time. This may also include how different agencies share the whole task and compliment each others networks. A mention of the various field Divisions of the agency covering the entire area may also come in the text.>

<In case of the State agencies, the whole area may be suitably divided into various River Basins/Zones and the reporting in all the subsequent sections should also follow the same grouping. That means all the sections would have sub-sections for each of the River Basins/Zones of the state.>

<The maps and text included in this section gives complete understanding of the three types of networks, viz. hydro-meteorological network (SRG/ARG/FCS), hydrological network (river gauging, reservoirs and lakes) and WQ network in each of the River Basins/Zones. The text must also specify what has been added or removed from these respective networks in the year of reporting. Improvements in terms of equipment and facilities could also be highlighted. Information about the system's coverage in terms of various types of data to be observed together with the respective monitoring frequencies to be maintained will be helpful to the readers in appreciating the scope and extent of the information that can be available from the system. Similarly, an overview of how the data has been scrutinised at various levels for assuring its quality would be beneficial for readers' understanding of the whole mechanism of collection of data in the field to the presentation of data in the water yearbook. Furthermore, a section on how much of the data of the year under reporting fall short of the target would be appropriate. This section would briefly show what information would not be available though it was expected and would thus clear ambiguity about availability of such data. >

<The subsequent sub-sections cover all issues mentioned above for (a) hydro-meteorological, (b) hydrological and (c) water quality observation networks.

3.3 Hydro-meteorological observation system

3.3.1 Network layout and adaptations in reporting year

<The text should refer to the map showing the meteorological network of SRG, ARG and FCS stations. On the basis of number of the stations of various types, a table could be made about the density of observation stations of various types in every river basin or zone.>

<Highlight what improvements have been effected in the year under consideration in terms of new stations, equipment/procedures etc.>

<A comprehensive listing of the observation stations must be provided for the benefit of the user so that any station can be easily referred. A sample of such table in given as Table 3.1. This table must be put in Annex – A, rather than in this main text of the water yearbook. This will avoid imbalance of the text by putting a huge table in between. Also, availability of such table separately in the Annex would be better for accessing it electronically. Highlight what has been changed in the network in terms of addition or removal of stations. This could be given in tabular form giving name, location and the period for which the station remained operational. Any such new station could also be highlighted in the tabular presentation of the list of stations in the network. The station listing in Tbale 3.1 also shows a few key static characteristics along with. It is also important to mention about changes in the frequency of observation introduced in the system, if any, as a result of the systems review in light of emerging user requirements. >

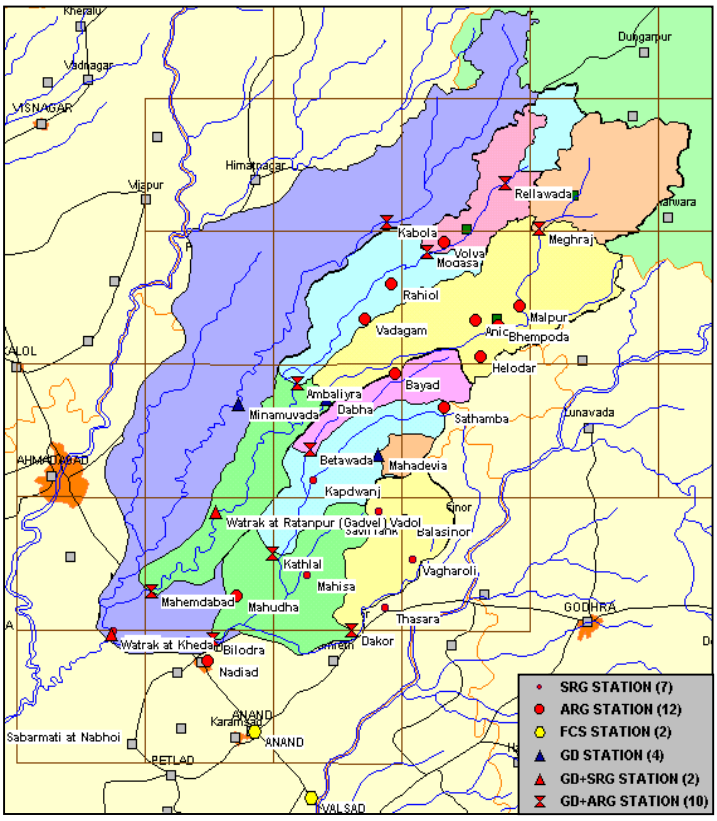


Figure 3.1: Meteorological observation network in ?? River Basin/Zone (full page)

< The map in Figure 4.1 could also simultaneously show agency's jurisdiction together with boundaries of various divisions and locations of SDDPCs/DDPCs/SDPC headquarters and distinct hydrological regions in the state >

Table 3.1: List of meteorological observation stations (Grouped by River and district and sorted by station names)

S.No.	River	District	Station Name	Station Code	Tehsil	Tributary	Lat.	Long.	Alt. (m)	Date of Establi-shment	SW - Type
1	Amba	Raigad	Pali	Pali	Sudhagad	-	183152	731205	22.4	01/06/1972	FCS
2			Payarachiwadi	Payarwadi	Sudhagad	Walki	183300	731600	22.4	31/03/1988	SRG
3			Tuksai	Tuksai	Khalapur	-	184149	731800	43.1	31/03/1988	ARG
4	Bharja	Thane	Gonde(Kd)	Gonde(Kd)	Mokhada	Wagh	195522	732200	400.0	01/06/1986	ARG
5	Daman Ganga	Thane	Jamsar	Jamsar	Jawhar	Wagh	195820	731409	395.0	01/06/1986	SRG
6			Juni Jawhar	Juni Jawhar	Jawhar	Jawhar	195348	731358	400.0	01/08/1977	SRG
7			Khadadi	Khadadi	Jawhar	Wagh	195910	731409	300.0	01/06/1986	ARG
8			Shindyachapada	Shindyapada	Mokhada	Wagh	195817	732234	400.0	01/07/1986	ARG
9	Deoghar	Sindudurg	Baparde	Baparde	Deogad	Local nala	162618	732854	19.6	01/06/1991	ARG
10			Deogad	Deogad	Deogad	-	162219	732328	45.7	01/05/1969	SRG
11			Phondaghat	Phondaghat	Kankawali	Kharada	162154	734757	145.1	01/05/1969	SRG
12			Talere	Talere	Kankawali	Kharada	162712	733916	152.4	01/05/1990	SRG
13	Gad	Sindudurg	Digawale	Digawale	Kankawali	-	161415	735000	91.5	01/05/1969	SRG
14			Golvan	Golvan	Malvan	-	160831	733340	76.2	01/05/1990	SRG
15			Kankawali	Kankawali	Kankawali	-	161642	734244	76.2	01/05/1969	SRG
16			Kasal	Kasal	Kudal	Kasal	161003	734139	45.7	01/05/1990	SRG
17			Nardave	Nardave	Kankawali	-	161201	735207	131.0	01/05/1990	SRG
18			Palsamb	Palsamb	Malvan	-	161407	733251	30.5	01/05/1991	SRG
19			Tarandale	Tarandale	Kankawali	-	161811	734224	76.2	01/05/1990	SRG
20	Godavari	Ahmednagar	Adhala	Adhala	Akola	Adhala	193644	740605	626.0	01/07/1995	ARG
21			Bhagur	Bhagur	Shevgaon	Nani	191929	751226	472.0	01/06/1989	ARG
22			Bhavarwadi	Bhavarwadi	Newasa	-	191638	744917	605.0	01/06/1976	SRG
23			Bodhegaon	Bodhegaon	Shevgaon	-	191816	752752	468.0	01/06/1976	SRG
24			Brahmangaon	Brahmangaon	Kopargaon	-	195742	742617	520.0	01/06/1959	SRG
25			Kopargaon	Kopargaon	Kopargaon	-	195210	742746	499.0	01/06/1989	FCS
26			Kotul	Kotul	Akola	Mula	192603	735808	715.0	01/07/1990	ARG
27			Manjur(Handewadi)	Manjur	Kopargaon	-	195445	741650	530.0	01/06/1959	ARG
28			Mhaladevi(Induri)	Mahaldevi	Akola	Pravara	193250	735537	578.0	01/06/1969	ARG
29			Mungi	Mungi	Shevgaon	-	192421	752648	432.0	01/06/1972	ARG
30			Newasa	Newasa	Newasa	Parvara	193308	745543	471.0	01/01/1978	FCS
31			Padhegaon	Padhegaon	Kopargaon	-	195553	743315	515.0	01/06/1959	ARG
32			Panegaon	Panegaon	Newasa	Mula	192850	744738	483.0	01/06/1988	ARG
33	Godavari	Ahmednagar	Rahata	Rahata	Kopargaon	-	194253	742907	512.0	01/06/1959	SRG
34			Samangaon(Male)	Samangaon M	Shevgaon	Godavari	192026	750626	480.0	01/06/1988	ARG
. . .											
. . .											
. . .											

Note: Shaded rows in the table indicate those stations that are put up newly during the reporting year.

3.3.2 Monitoring and processing

<It would be beneficial to present the salient features of the observation systems being used for various types of stations. Equipment and practices about following type of stations can be outlined, for example:

Hydro-meteorological Observation Network

ARG stations:	Brief note on the equipment available – type of ARGs used in the network and observation practices employed
FCS stations:	Brief note on the type and range of equipment employed – observation practices

< Together with the above note on the data collection plan, it would be good to briefly define the various primary and secondary validations and data processing carried out while finalising the data. Such a background will create greater awareness among the data users about the type of data processing the data has undergone.

3.3.3 Data collection in reporting year

< This sub-section can bring out the accomplishment in terms of percentage of data collection target achieved. Together with such percentages for various data types it would be appropriate to briefly mention about the reasons of the shortfall in collection of data. It may be due to equipment malfunctioning, maintenance issues, availability of required personnel and consumables required for observation.

3.4 Hydrometry and sediment transport

3.4.1 Network layout and adaptations in report year

<The text to refer to the map showing the network of river gauging / reservoir stations (G/GQ/GD/GDS/GDSQ/GDQ).>

<Highlight what improvements have been effected in the year under consideration in terms of new equipment/procedures, etc.>

<Highlight what has been changed in the network in terms of addition or removal of stations. This could be given in tabular form giving name, location and the period for which the station remained operational. Any such new station could also be highlighted in the tabular presentation of the list of stations in the network. It is also important to mention about changes in the frequency of observation.>

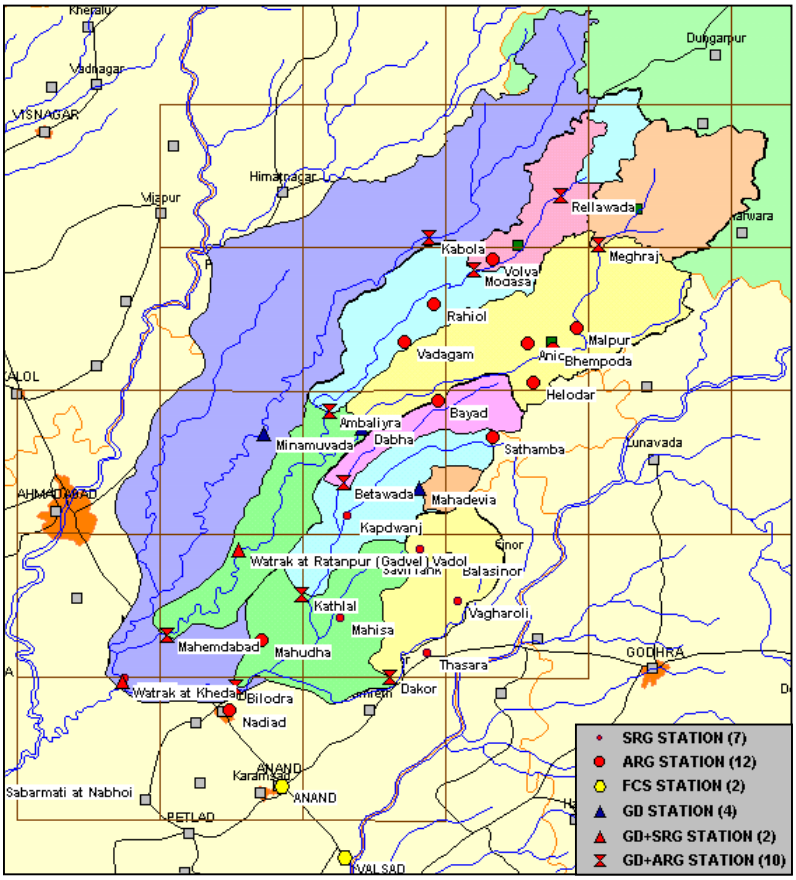


Figure 3.2: Hydrological observation network in ?? River Basin/Zone (full page)

Table 3.2: List of hydrological observation stations (Grouped by River and district and sorted by station names)

S.No.	River	District	Station Name	Station Code	Tehsil	Tributary	Lat.	Long.	Alt. (m)	Date of Establi-shment	SW - Type
1	Amba	Raigad	Burmali	Burmali	Sudhagad	Walki	183129	731330	22.6	01/06/1993	GD
2			Pali	Pali	Sudhagad	-	183152	731205	22.4	01/06/1972	GD
3			Salinde	Salinde	Pen	Nigade	183718	730632	20.0	01/06/1994	GDS
4			Tuksai	Tuksai	Khalapur	-	184149	731800	43.1	31/03/1988	GD
5	Bharja	Ratnagiri	Latwan	Latwan	Mandangad	-	175555	732051	121.9	01/12/1983	GDS
6	Daman Ganga	Nashik	Usthale (Hedpada)	Usthale	Peth	Damaganga	201200	733237	378.0	06/06/1982	G
7		Thane	Gaheli	Gaheli	Jawhar	Wagh	200700	731503	90.0	01/06/1999	GD
8			Khadadi	Khadadi	Jawhar	Wagh	195910	731409	300.0	01/06/1986	GD
9			Khadkhad	Khadkhad	Jawhar	Wagh	195654	731634	30.0	01/06/1987	G
10			Shindyachapada	Shindyapada	Mokhada	Wagh	195817	732234	400.0	01/07/1986	GD
11			Ghonsari(L)	Ghonsari(L)	Kankawali	Kharada	162405	734616	111.8	21/09/1965	GDS
12	Godavari	Ahmednagar	Bhagur	Bhagur	Shevgaon	Nani	191929	751226	472.0	01/06/1989	GDS
13			Kopargaon	Kopargaon	Kopargaon	-	195210	742746	499.0	01/06/1989	GD
14			Mhaladevi(Induri)	Mahaldevi	Akola	Pravara	193250	735537	578.0	01/06/1969	GD
15			Newasa	Newasa	Newasa	Parvara	193308	745543	471.0	01/01/1978	GD
16			Panegaon	Panegaon	Newasa	Mula	192850	744738	483.0	01/06/1988	GD
17	Godavari	Ahmednagar	Samangaon(Male)	Samangaon M	Shevgaon	Godavari	192026	750626	480.0	01/06/1988	G
18			Sangamner(Waghapur)	Sangamner	Sangamner	Pravara	193308	741400	551.0	01/07/1989	GD
19		Amravati	Warud Bagaji	WarudBagaji	Tiwasa	Wardha	205230	781544	274.7	06/08/1987	GD
20		Aurangabad	Chinchkhed Bhavan	Chinchkhed	Sillod	Purna	201622	753843	583.0	01/06/1984	GD
21			Nagamthan	Nagamthan	Vaijapur	-	194344	744726	476.0	01/06/1987	GD
22			Solegaon	Solegaon	Gangapur	Shivna	194215	750558	467.0	01/06/1988	G
23		Bhandara	Chichghat	Chichghat	Bhandara	Wainganga	210448	793406	249.0	18/04/1986	GDS
24			Kardha	Kardha	Bhandara	Wainganga	210840	794019	246.0	01/06/1977	GD
25			Lakhandur	Lakhandur	Lakhandur	Wainganga	204406	795232	222.5	28/06/1988	GD
26			Mahalgaon	Mahalgaon	Tumsar	Wainganga	213033	795422	263.1	01/06/1990	GD
27			Sitekasa	Sitekasa	Bhandara	Wainganga	213104	793514	323.0	01/06/1985	GD
28		Buldhana	Fardapur	Fardapur	Mehkar	Painganga	201030	763354	527.5	01/06/1985	GDS
29			Raheri	Raheri	Sindhkhed Raja	Purna	195900	761654	494.0	01/01/1988	GD
30			Dhaba	Dhaba	Gondpipri	Wardha	193146	793520	162.0	01/07/1988	GD
31			Dindora(Soit)	SoitDindora	Warora	Wardha	201644	784912	203.2	01/06/1988	GD
32	Godavari	Chandrapur	Gadbori	Gadbori	Sindewahi	Wainganga	201731	793534	199.3	01/06/1972	GD
33			Nandgur	Nandgur	Chandrapur	Wainganga	200105	793044	187.6	01/06/1978	GDS
. . .											
. . .											
. . .											

Note: Shaded rows in the table indicate those stations that are put up newly during the reporting year.

3.4.2 Monitoring and processing

<It would be beneficial to present the salient features of the observation systems being used for various types of stations. Equipment and practices about following type of stations can be outlined, for example:>

Hydrological observation Network

Gauge & Discharge stations:	Brief note on available equipment – current meters, water level recorders. Method of observation – single point or multiple point velocities. Mechanism for crossing the river – boats, bridges, cableway (what types)
Sediment stations:	Type of sampler used and analysis procedure employed

< Together with the above note on the data collection plan, it would be good to briefly define the various primary and secondary validations and data processing carried out while finalising the data. Such a background will create greater awareness among the data users about the type of data processing the data has undergone.

3.4.3 Data collection in report year

< This sub-section can bring out the accomplishment in terms of percentage of data collection target achieved. Together with such percentages for various data types it would be appropriate to briefly mention about the reasons of the shortfall in collection of data. It may be due to equipment malfunctioning, maintenance issues, availability of required personnel and consumables required for observation.

3.5 Water quality

3.5.1 Network layout and adaptations in report year

< It is appropriate to indicate which agencies are complementing each others observation network in the region. For example, the water quality of the river Godavari and its tributaries is being monitored by the Central Water Commission (CWC), the Water Resources Departments of Maharashtra and Andhra Pradesh, the Ground Water Survey Division of Madhya Pradesh (SW monitoring as additional activity), and the Central Pollution Control Board (CPCB) through the State Pollution Control Boards (SPCB) of Madhya Pradesh, Maharashtra, and Andhra Pradesh. >

< In Godavari basin, water quality is monitored by CWC at 14 stations out of the 22 hydrological observation stations. Besides, the State Water Resources Department monitors water quality at 13 locations. >

< CPCB has in all 21 Stations (11 on Godavari, 1 on Kalu, 1 on Manjira, 4 on Maner, 1 on Panchaganga, 2 on Ulhas 3 on Wainganga and 1 on Wardha).>

Figure 3.3 shows the monitoring stations of the following type on one map:

- WQ stations of CWC;
- WQ stations of state SW;
- WQ stations of CPCB/SPCBs; and
- Hydrologic discharge stations

<The text to refer to the map showing the network of water quality monitoring stations (GQ/GDQ/GDSQ). Highlighting the various types of WQ stations (baseline, trend, flux, surveillance) would be very useful.>

<Highlight what improvements have been effected in the year under consideration in terms of new equipment/procedures etc.>

<Highlight what has been changed in the network in terms of addition or removal of stations. This could be given in tabular form giving name, location and the period for which the station remained operational. Any such new station could also be highlighted in the tabular presentation of the list of stations in the network. It is also important to mention about changes in the frequency of observation.>

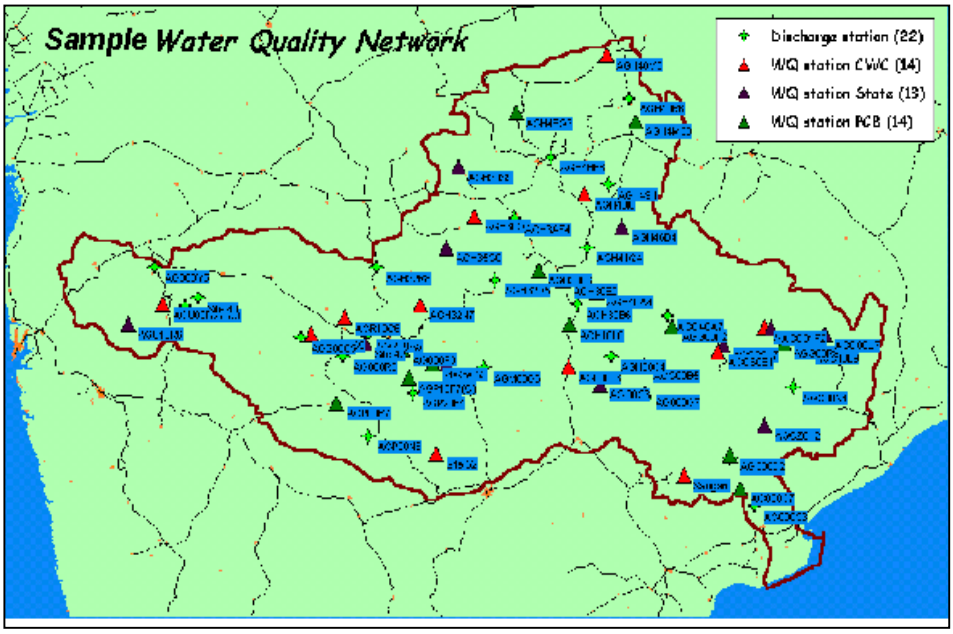


Figure 3.3: Water quality observation network in ?? River Basin/Zone (full page)

Table 3.3 gives the list of WQ stations in the region. Note however that this is a list of example stations only and not the stations shown in Figure 3.3. above.

Table 3.3: List of WQ observation stations (Grouped by River and district and sorted by station names)

S.No.	River	District	Station Name	Station Code	Tehsil	Tributary	Lat.	Long.	Alt. (m)	Date of Establi-shment	SW - Type
1	Amba	Raigad	Pali	Pali	Sudhagad	-	183152	731205	22.4	01/06/1972	Trend
2	Daman Ganga	Thane	Khadadi	Khadadi	Jawhar	Wagh	195910	731409	300.0	01/06/1986	Trend
3			Khadkhad	Khadkhad	Jawhar	Wagh	195654	731634	30.0	01/06/1987	Baseline
4			Shindyachapada	Shindyapada	Mokhada	Wagh	195817	732234	400.0	01/07/1986	Trend
5	Godavari	Ahmednagar	Kopargaon	Kopargaon	Kopargaon	-	195210	742746	499.0	01/06/1989	Trend
6			Newasa	Newasa	Newasa	Parvara	193308	745543	471.0	01/01/1978	Trend
7	Godavari	Bhandara	Kardha	Kardha	Bhandara	Wainganga	210840	794019	246.0	01/06/1977	Trend
8	Godavari	Chandrapur	Wadsa (Chincholi)	WadsaChinch	Bramhapuri	Wainganga	203625	795601	221.2	01/06/1965	Trend
9			Wagholi -Butti	WagholiButi	Saoli	Wainganga	200706	795417	222.1	01/06/1964	Trend
10		Gadchiroli	Bhamragad	Bhamragad	Bhamragad	Indrawati	192500	803510	207.5	20/06/1998	Flux
11			Damrencha	Damrencha	Aheri	Indrawati	191332	792230	175.9	01/03/2000	Baseline
12			Mahagaon	Mahagaon	Aheri	Pranhita	192658	795812	128.3	06/06/1998	Flux
. . .											
. . .											
. . .											

Note: Shaded rows in the table indicate those stations that are put up newly during the reporting year.

3.5.2 Monitoring and processing

< This section will emphasis on the monitoring program and its objectives together with understanding of various types of data processing which is undertaken.

The different organisations involved in water quality monitoring in Godavari river basin have the following objectives for monitoring:

1. To establish *baseline* quality (all agencies);
2. To observe *trend* in water quality over a period of time (all agencies);
3. To calculate the *load* (or *flux*) of water quality constituents of interest (e.g. silt in reservoirs);
4. To prevent and control water pollution (Central and State PCBs);
5. To have surveillance over pollution threats to water quality for sustenance of various beneficial uses, like irrigation (State Irrigation Departments)

Monitoring Frequency

The samples are collected three times in a month by CWC and once a month by CPCB. Sampling dates for CWC are 1st, 10th and 20th of each month in general.

- A general statement on the target frequency of sampling for the respective agencies according to their monitoring objectives may be given here or a table indicating the targeted sampling frequency for each station.
- Elaborate on sampling frequency of irrigation departments
- Describe the sampling programme for seasonal rivers, if applicable.

Analytical Quality Control

Analytical Quality Control (AQC) program is run among various laboratories for ensuring and monitoring standards maintained with respect to analysis performed in the laboratories. The following parameters are covered under the inter-laboratory AQC exercise (carried out once a year) in which the majority of the laboratories take part.

1. Conductivity
2. Total Dissolved Solids
3. Total Hardness
4. Sodium
5. Fluoride
6. Sulphate
7. Nitrate - N
8. Phosphate-P
9. Boron
10. Chloride

- ☞ ***Average accuracy of the participating laboratories for relevant parameters, such as heavy metals, may be given here in addition.***
- ☞ ***Newly introduced parameters in the AQC programme, if any, may be mentioned here.***
- ☞ ***All parameters associated with the laboratory level may be listed here or in an appendix.***

Parameters

The level of the laboratory is an indication of the analytical capacity of the laboratory.

Level I	Laboratory located in the field, generally analysing Temperature, pH, Conductivity, Dissolved Oxygen, colour and odour
Level II	Laboratory has facilities to analyse basic water quality parameters, nutrients, indicators of organic and bacteriological pollution etc.
Level II^{+}	Laboratory has facilities to analyse basic water quality parameters, nutrients, indicators of organic and bacteriological pollution etc. Laboratory is in possession of advanced equipment, such as Atomic Adsorption Spectrophotometer (AAS), Gas Chromatograph (GC), UV-Visible Spectrophoto-meter etc.

Table 3.4: Classification of laboratories involved in monitoring as used by CWC and other HP-Agencies.

Parameter ID	Parameter Name	Category	Unit	LWL	UWL	Minimum	Maximum	No. of Decimals
FLD Field Determinations								
Colour Code	Colour	Physical	-					
DO	Dissolved oxygen	Chemical	mg/L	0	15	0	30	1
EC_FLD	Electrical Conductivity, Field	Physical	µmho/cm	50	5000	1	10000	0
Odour Code	Odour	Chemical	-					
pH_FLD	pH_Field	Chemical	pH units	5.5	9	2	14	1
Secchi	Secchi Depth	Physical	m	0.01	50	0.005	100	2
Temp	Temperature	Physical	deg C	10	40	0.1	50	1
Laboratory Determinations								
DO_SAT%	Dissolved Oxygen Saturation %	Chemical	%	0	150	0	300	0
EC_GEN	Electrical Conductivity	Physical	µmho/cm	50	5000	1	10000	0
pH_GEN	pH	Chemical	pH units	5.5	9	2	14	1
SS	Solids, Suspended	Physical	mg/L	5	2000	0	3000	0
TDS	Solids, Total Dissolved	Physical	mg/L	50	5000	5	30000	0
TS	Solids, Total	Physical	mg/L	50	5000	10	30000	0
Turb	Turbidity	Physical	NTU	1	2000	0.1	10000	1
NH_3-N	Nitrogen, ammonia	Chemical	mg N/L	0.05	100	0.05	1000	2
NO_2+NO_3	Nitrogen, Total Oxidised	Chemical	mg N/L	0.05	1000	0.05	2000	1
NO_2-N	Nitrogen, Nitrite	Chemical	mgN/L	0	0.5	0	10	1
NO_3-N	Nitrogen, Nitrate	Chemical	mgN/L	0.05	1000	0.01	2000	2
o-PO4-P	Phosphorus, ortho-phosphate	Chemical	mg P/L	0.05	5	0.01	50	3
Org-N	Nitrogen, Organic	Chemical	mgN/L	0.1	200	0.01	1000	1
P-Tot	Phosphorus, total	Chemical	mgP/L	0.01	10	0.001	100	3
$BOD_{3\text{-day, }27\,^{\circ}c}$	Biochemical Oxygen demand	Chemical	mg/L	0.5	200	0.1	5000	1
COD	Chemical Oxygen Demand	Chemical	mg/L	5	5000	1	10000	1
Alk-Phen	Alkalinity, phenolphthalein	Chemical	mg$CaCO_3$/L	0	500	0	3000	1
ALK-TOT	Alkalinity, total	Chemical	mg$CaCO_3$/L	10	1000	5	5000	1

Table 3.5: Example of: Details of parameters analysed by various levels of laboratories

Parameter	Parameter group	WQ standard (target)
Temp	General	none
TDS or EC	General	TDS 500 mg/L , (drinking water std.) EC 2250 umho/cm (for irrigation water)
SAR	Major Ions (indirect)	26 (for irrigation)
DO	General	4 mg/L (min value)
BOD	Organic matter	3 mg/L (target)
TotP & NO_3	Nutrients	Nitrate 10 mgN/L (drinking water)
Selected pollutants	trace metals or pesticides	

Table 3.6: CPCB classification system for quality of water

3.5.3 Data collection in report year

< In this sub-section a brief account of what could be achieved in the year under reporting in terms of data collection is to be given. The short summary can bring out certain typical problems that would have hampered the data collection. An overall account of how much percentage of the target data collection could be achieved together with the number of station-samples taken and analysed in the year would be indicative of the system's performance in terms of collection of data.

In this sub-section an overview of the monitoring related to pollution monitoring is given for the year 1996. In Table 3.7 the content of the database with respect to pollution monitoring is summarised: the number of samples for BOD, Total Coliforms (TC) and Faecal Coliforms (FC) and the first and last sampling date are indicated.

☞ ***The targeted number of samples may be added to Table 4 if one wants to indicate the performance of operating the monitoring network with respect to the activities envisaged in the monitoring programme.***

Station characteristics overview					
Godavari Pollution related Monitoring			**For the year 1996**		
Station ID	**Name**	**Parameter ID**	**First Date**	**Last Date**	**Number of Samples**
AG000C3	Polavaram	BOD	09/03/1996	10/11/1996	32
AG000C3	Polavaram	FC			0
AG000C3	Polavaram	TC			0
AG000C3	Polavaram	DO	07/02/1996	12/11/1996	34
AG000G7	Perur	BOD	23/01/1996	03/11/1996	30
AG000G7	Perur	FC			0
AG000G7	Perur	TC			0
AG000G7	Perur	DO	09/11/1996	01/12/1996	35
AG000J3	Mancherial	BOD	01/02/1996	21/11/1996	35
AG000J3	Mancherial	FC			0
AG000J3	Mancherial	TC			0
AG000J3	Mancherial	DO	24/01/1996	13/12/1996	35
AG000P3	Yelli	BOD	09/03/1996	12/07/1996	15
Water Quality Yearbook – HYMOS Example Report 3 (HP 2002)					

Table 3.7: Contents of the database for pollution related monitoring in 1996

Limitations

At present the infrastructure of the laboratories attached to the Krishna Godavari Basin Organisation of CWC in Hyderabad are meeting the requirements for analysing all the above-mentioned parameters. The laboratories have been upgraded to analyse also the pollution-related parameters.

4 Hydrological review of the year <YYYY>

4.1 Summary

<This section on hydrological review tries to summaries all the aspects of the main components of the hydrological cycle (rainfall, evaporation, runoff and water quality) one-by-one. This presentation of information will be in the form of explanatory notes, graphs and data tables. The Summary in the beginning attempts to highlight the salient features of the hydrological scenario for the year under reporting>.

As an example, a short summary about WQ data is presented hereunder. Note that this is only an indicative text and the actual text is to be based on the region and the year at hand and has to bring out the essence of the WQ regime representing the year.

<The water quality of the Godavari River and its tributaries is being monitored by the Central Water Commission (CWC) at 14 stations and by the Central Pollution Control Board (CPCB) at 21 stations. The State Irrigation Departments of Maharashtra and Andhra Pradesh also started water quality monitoring of the Godavari River. During the reporting period only CWC data are analysed for the purpose of preparation of the specimen Water Quality Yearbook.>

<The monitoring is done three times a month by CWC. >

<The CWC analyses major ions, some inorganics, like phosphates, silicates, ammonia, aluminium and irons, and basic parameters, like temperature, conductivity and pH. Pollution related parameters, like BOD, COD and total and fecal coliforms have recently been introduced. >

<Most of the major ions and inorganics are genarally within the limits of drinking and irrigation standards, whereas coliforms are the major problem in the river basin. Most of the stations monitored do not meet the desired water quality criteria for coliforms and in some cases BOD.>

<With respect to organic pollution, i.e. BOD, COD and DO, the Godavari is worst polluted at Nashik, and Nanded. This is mainly due to discharge of untreated domestic wastewater into the river followed by reduced flow in the river due to water abstraction from the river in the upstream. Similarly, the river is heavily polluted at Ramagundam, Bhadrachalam and Rajamundary towns due to discharges of partially treated / untreated industrial wastewater along with the domestic wastewater.>

4.2 Rainfall

<Rainfall of the year in the region could be characterised by the following figures and tables. A good explanation of the important features which may be inferred from these figures and tables must follow in the sub-section. Different types of figures and tables that could be representing the rainfall in the region could be as follows: >

Figure 4.2.1 Spatial variation of the monthly rainfall (monsoon months) in the region

Figure 4.2.2 Spatial variation of the annual rainfall in the region

Figure 4.2.3 Spatial variation of the rainfall as a percentage of long term annual average rainfall

Figure 4.2.4 Monthly rainfall of the year as seen against monthly frequency curves

Table 4.2.1 Station-wise rainfall data summary

Table 4.2.2 District-wise (or basin/sub-basin-wise, if required) rainfall data summary

Table 4.2.3 Daily Rainfall Data and associated monthly and yearly statistics (This table needs to be given in the main text for only few representative stations and not all the stations in the network. Similar tables for all the stations are however to be included in the Appendices to the yearbook. These appendices need to be printed only in required quantity and not in bulk as mentioned in the introduction. The appendices will be readily available as soft copy in the electronic yearbook form.

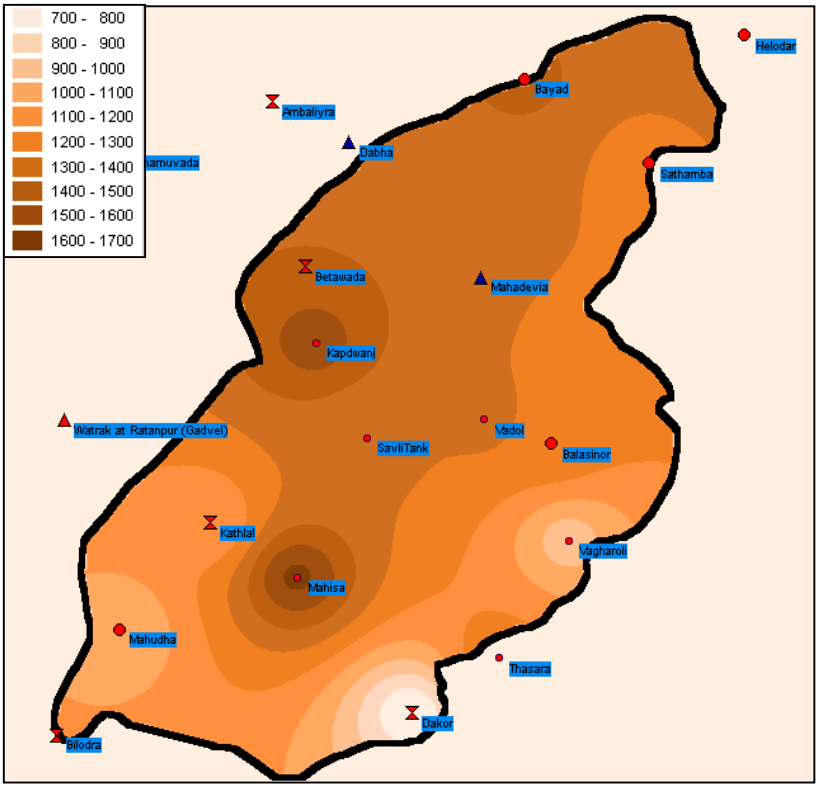


Figure 4.2.2:
Annual Rainfall in 1997

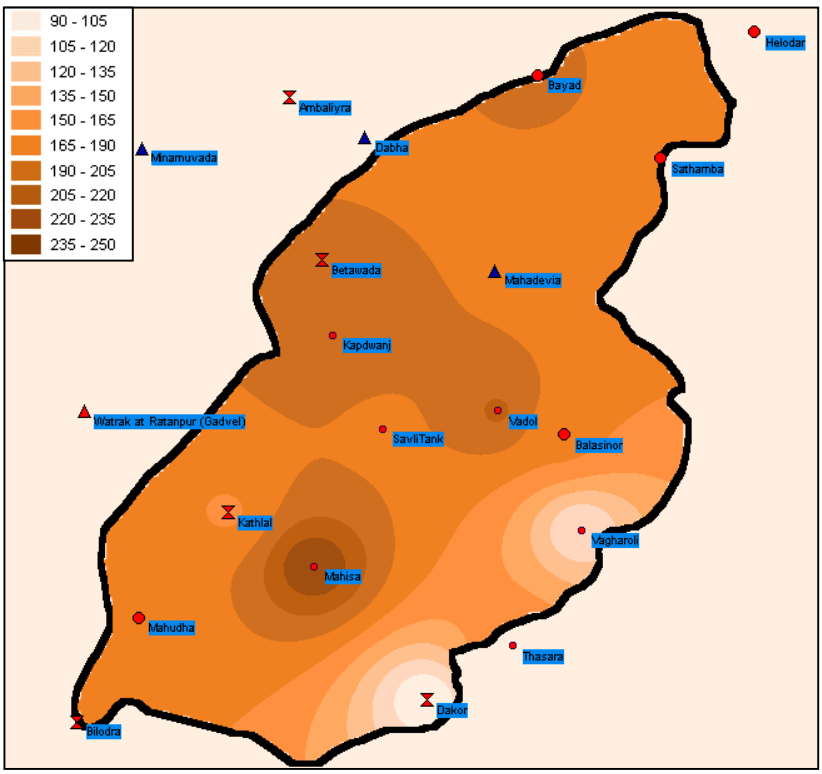


Figure 4.2.3: Annual Rainfall in 1997 as percentage of 1970-2000 average

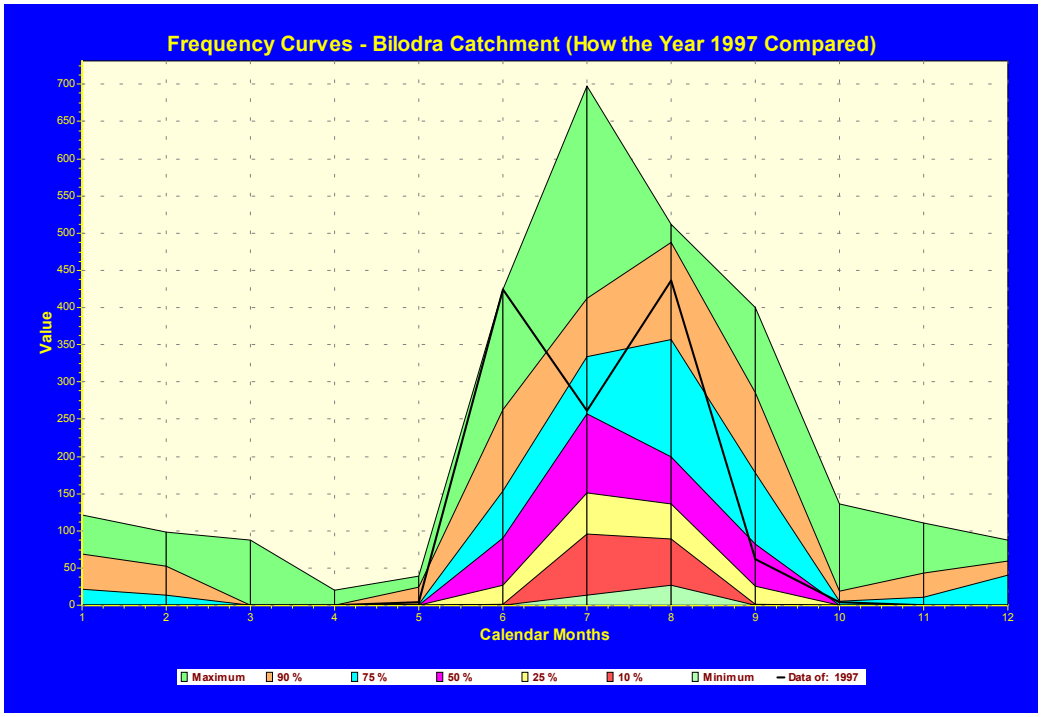


Figure 4.2.4: Monthly rainfall of 1997 as seen against monthly frequency curves (based on 1961-1997 period)

Table 4.2.1: Station-wise Rainfall Data Summary Year - 1997

Station ID	Name	Alt	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year	LTA	Max	Date
Basin A																		
Ambaliyara	Ambaliyara	72	0	0	0	0	0	473	113	279	85	0	0	0	950	900	170.1	25/06/1997
Balasinor	Balasinor	90	0	0	0	0	0	381	226	625	40	0	0	0	1272	750	208.0	01/08/1997
Bayad	Bayad	135	0	0	0	0	0	428	384	526	64	16	0	0	1418	775	257.0	01/08/1997
Dakor	Dakor	55	0	0	0	0	0	315	120	403	0	0	0	0	718	775	127.3	01/08/1997
Helodar	Helodar	180	0	0	0	0	0	338	476	572	3	27	0	0	1416	850	238.0	28/07/1997
Kapadwanj	Kapdwanj	100	0	0	0	0	0	633	197	613	105	0	0	0	1548	800	225.0	28/06/1997
Kathlal	Kathlal	47	0	0	0	0	0	448	108	545	0	0	0	0	1102	675	150.0	01/08/1997
Mahemdabad	Mahemdabad	35	0	0	0	0	0	488	125	425	90	0	0	0	1128	775	130.2	25/06/1997
Mahisa	Mahisa	65	0	0	0	0	0	492	325	642	137	28	0	0	1624	725	200.0	23/08/1997
Mahudha	Mahudha	50	0	0	0	0	0	276	112	500	97	29	0	0	1014	600	132.0	02/08/1997
Sathamba	Sathamba	152	0	0	0	0	0	239	556	468	0	0	0	0	1263	725	275.0	31/07/1997
SavliTank	SavliTank	90	0	0	0	0	0	382	139	687	95	33	0	0	1336	750	245.7	01/08/1997
Thasara	Thasara	65	0	0	0	0	0	428	113	525	59	0	0	0	1125	775	205.0	02/08/1997
Vadol	Vadol	92	0	0	0	0	0	335	233	676	85	0	0	0	1329	675	260.1	01/08/1997
Vagharoli	Vagharoli	72	0	0	0	0	0	473	113	279	85	0	0	0	950	900	170.0	25/06/1997
Basin B																		
…																		…
…																		…
…																		…

In the above table:

Alt : is the altitude of the station in meters

LTA : is the long term average of annual rainfall

Max : is the maximum daily rainfall within the year, date of its occurrence is also given alongwith

Table 4.2.2: District-wise Rainfall Data Summary **Year - 1997**

District Name	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year	LTA
Ahmedabad	5	0	0	0	0	172	113	279	85	0	0	0	950	900
(% of 1970-2000)	100	-	-	-	-	191	113	112	170	0	-	0	105	
Amreli	0	0	0	0	0	187	190	625	40	0	0	0	1272	750
(% of 1975-2000)	0	-	-	-	-	156	136	208	100	0	-	0	170	
Banaskantha	7	0	0	0	0	148	140	526	64	16	0	0	1418	775
(% of 1975-2000)	88	-	-	-	-	197	147	162	107	80	-	0	180	
Bharuch	0	0	0	0	0	176	130	403	105	0	0	0	718	775
(% of 1970-2000)	0	-	-	-	-	161	118	107	233	0	-	0	92	
Bhavnagar	10	0	0	0	0	129	180	572	90	27	0	0	1416	850
(% of 1975-2000)	250	-	-	-	-	103	138	163	164	180	-	0	166	
Dang	0	0	0	0	0	189	108	613	137	0	0	0	1548	600
(% of 1970-2000)	-	-	-	-	-	126	74	189	228	0	-	0	258	
Gandhinagar	14	0	0	0	0	150	90	545	97	0	0	0	1102	725
(% of 1970-2000)	117	-	-	-	-	176	95	136	176	0	-	0	152	
Jamnagar	6	0	0	0	0	120	205	425	95	0	0	0	1128	750
(% of 1975-2000)	60	-	-	-	-	126	205	118	238	0	-	0	150	
Junagadh	0	0	0	0	0	160	80	642	59	28	0	0	1624	775
(% of 1970-2000)	0	-	-	-	-	152	70	201	197	156	-	0	210	
Kheda	0	0	0	0	0	80	200	500	85	29	0	0	1336	900
(% of 1970-2000)	0	-	-	-	-	145	267	161	243	145	-	0	148	
…														
…														
…														

In the above table:

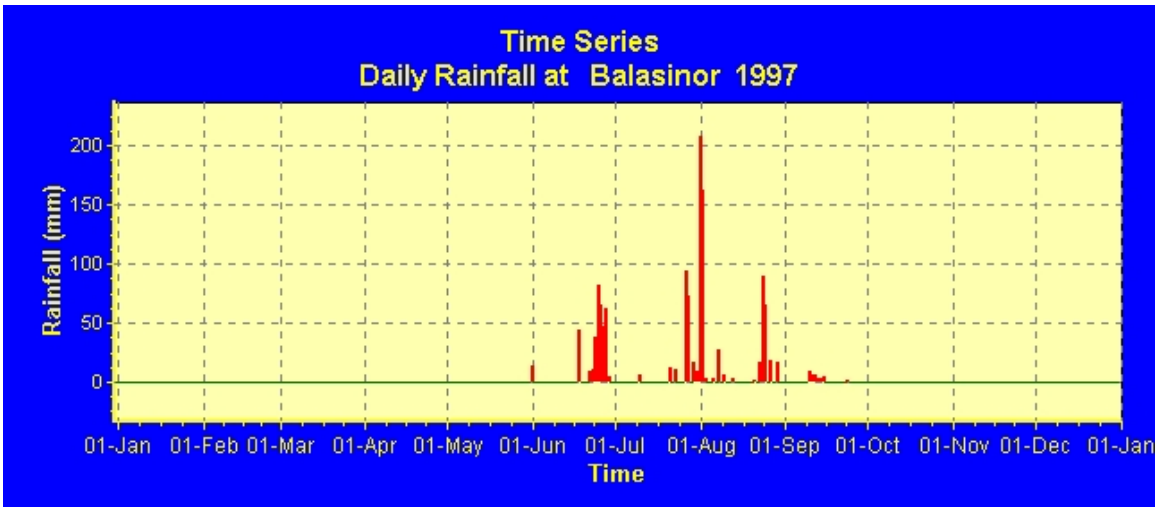
LTA : is the long term average of annual rainfall

Every alternate row is the monthly and yearly rainfall in terms of percentage of long term during the specific month or year as a whole.

Table 4.2.3: Daily Rainfall Data **Year – 1997**

Station Code: Balasinor **District:** Kheda **Independent River :** Sabarmati

Station Name: Balasinor **Units :** mm **Tributary:** Watrak



Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0.0	0.0	0.0	0.0	0.0	14.0	0.0	208.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	1.0	162.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	7.0	6.0	10.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	4.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0	44.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.0	0.0	12.0	2.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	11.0	11.0	17.0	1.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0	38.0	0.0	90.0	2.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0	83.0	0.0	66.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	65.0	0.0	0.0	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0	48.0	94.0	19.0	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.0	63.0	74.0	0.0	0.0	0.0	0.0	0.0
29	0.0		0.0	0.0	0.0	5.0	0.0	17.0	0.0	0.0	0.0	0.0
30	0.0		0.0	0.0	0.0	0.0	17.0	0.0	0.0	0.0	0.0	0.0
31	0.0		0.0		0.0		10.0	0.0		0.0		0.0
Mean	0.0	0.0	0.0	0.0	0.0	12.7	7.3	20.2	1.3	0.0	0.0	0.0
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Max.	0.0	0.0	0.0	0.0	0.0	83.0	94.0	208.0	10.0	0.0	0.0	0.0
Sum	0.0	0.0	0.0	0.0	0.0	381.0	226.0	625.0	40.0	0.0	0.0	0.0

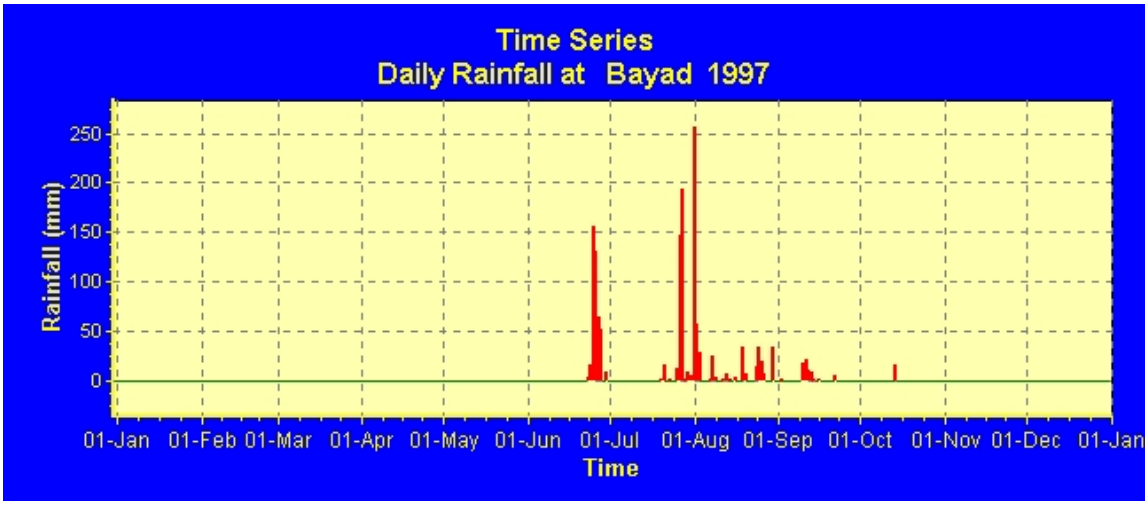
Yearly statistics :

Mean : 3.5 Sum : 1272.0	Minimum : 0.0 Date : 01/01/1997	Maximum : 208.0 Date : 01/08/1997	No. of data : 365 No. of missing data : 0

Table 4.2.3: Daily Rainfall Data **Year - 1997**

Station Code: Bayad **District:** Sabarkantha **Independent River :** Sabarmati

Station Name: Bayad **Units :** mm **Tributary:** Watrak



Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	257.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	57.5	1.5	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.5	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	18.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	10.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	9.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	16.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	1.0	7.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.0	0.0	14.5	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	2.0	1.5	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0	16.0	0.0	13.5	0.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0	156.0	0.0	33.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	130.0	11.0	19.0	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0	64.5	147.5	5.5	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.0	52.0	194.5	0.0	0.0	0.0	0.0	0.0
29	0.0		0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0
30	0.0		0.0	0.0	0.0	7.5	9.0	33.0	0.0	0.0	0.0	0.0
31	0.0		0.0		0.0		4.0	0.0		0.0		0.0
Mean	0.0	0.0	0.0	0.0	0.0	14.3	12.4	17.0	2.2	0.5	0.0	0.0
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Max.	0.0	0.0	0.0	0.0	0.0	156.0	194.5	257.0	20.0	16.0	0.0	0.0
Sum	0.0	0.0	0.0	0.0	0.0	428.0	384.5	525.5	64.5	16.0	0.0	0.0

Yearly statistics :

Mean : 3.9 Sum : 1418.5	Minimum : 0.0 Date : 01/01/1997	Maximum : 257.0 Date : 01/08/1997	No. of data : 365 No. of missing data : 0

4.3 River flows and water levels

<River flows during the year in the region could be characterised by the following figures and tables. A good explanation of the important features which may be inferred from these figures and tables must follow in the sub-sections. Different types of figures and tables that could be representing the runoff in various rivers could be as follows: >

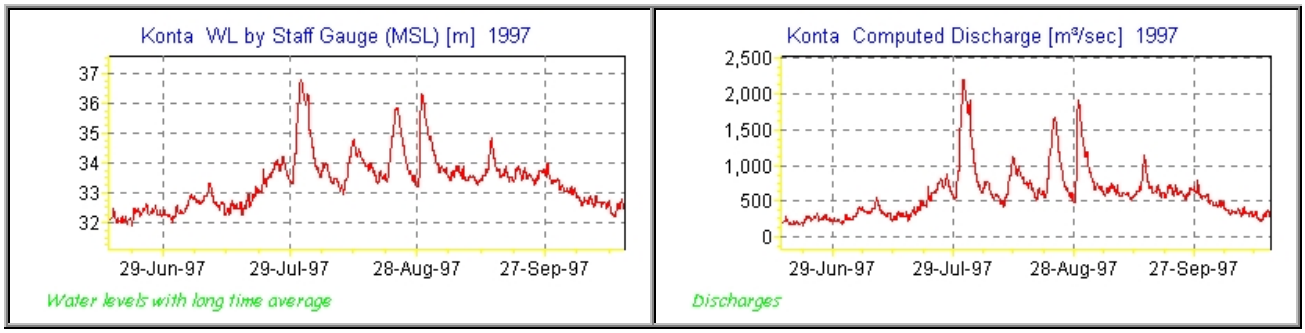
Figure 4.3.1 Water levels and flows observed and available at the least frequency at a few representative stations

Figure 4.3.2 Ten-daily river flows of the year as seen against Ten-daily frequency curves for certain base period. This could be given for few representative river gauging stations

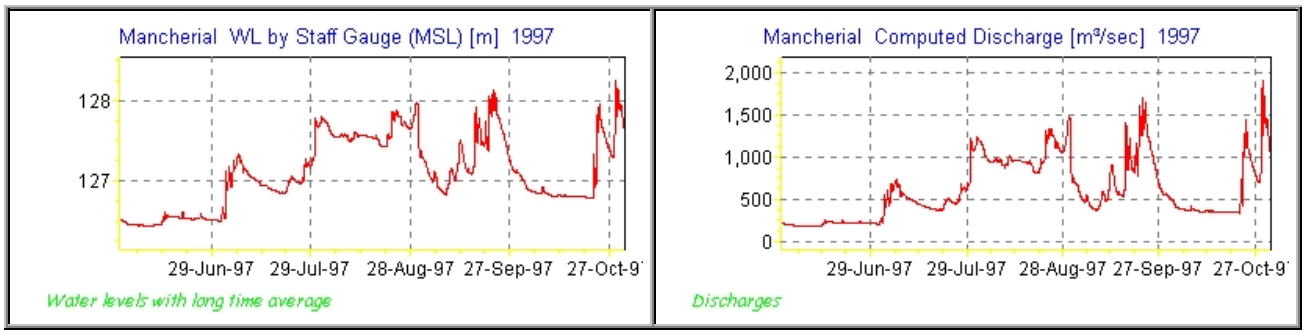
Figure 4.3.3 Flow duration curve for the year under consideration together with the Average duration curve for a certain base period. These curves obtained from daily data could be quite informative of the duration for which certain flow is maintained. This also could be given for few representative stations.

Table 4.3.1 Daily runoff data and associated monthly and yearly statistics. This table needs to be given in the main text for only few representative stations and not all the stations in the network. Similar tables for all the stations are however to be included in the Appendices to the yearbook. These appendices need to be printed only in required quantity and not in bulk as mentioned in the introduction. The appendices will be available as soft copy in the electronic yearbook form.

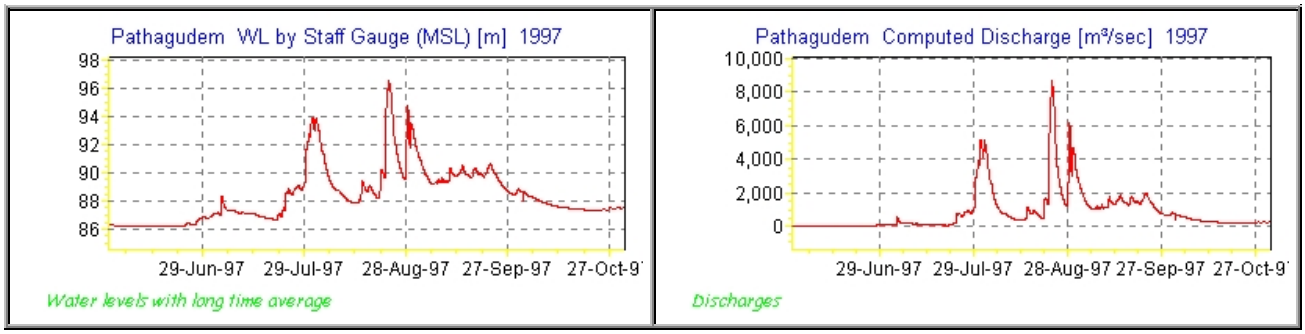
Station: Konta



Station: Mancherial



Station: Pathagudem



Station: Polawaram

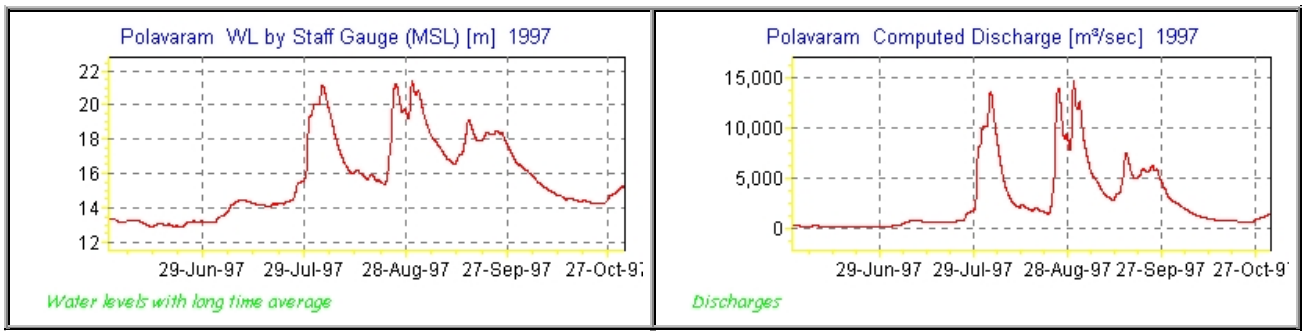


Figure 4.3.1: Hourly water levels and flows during the year at observation stations

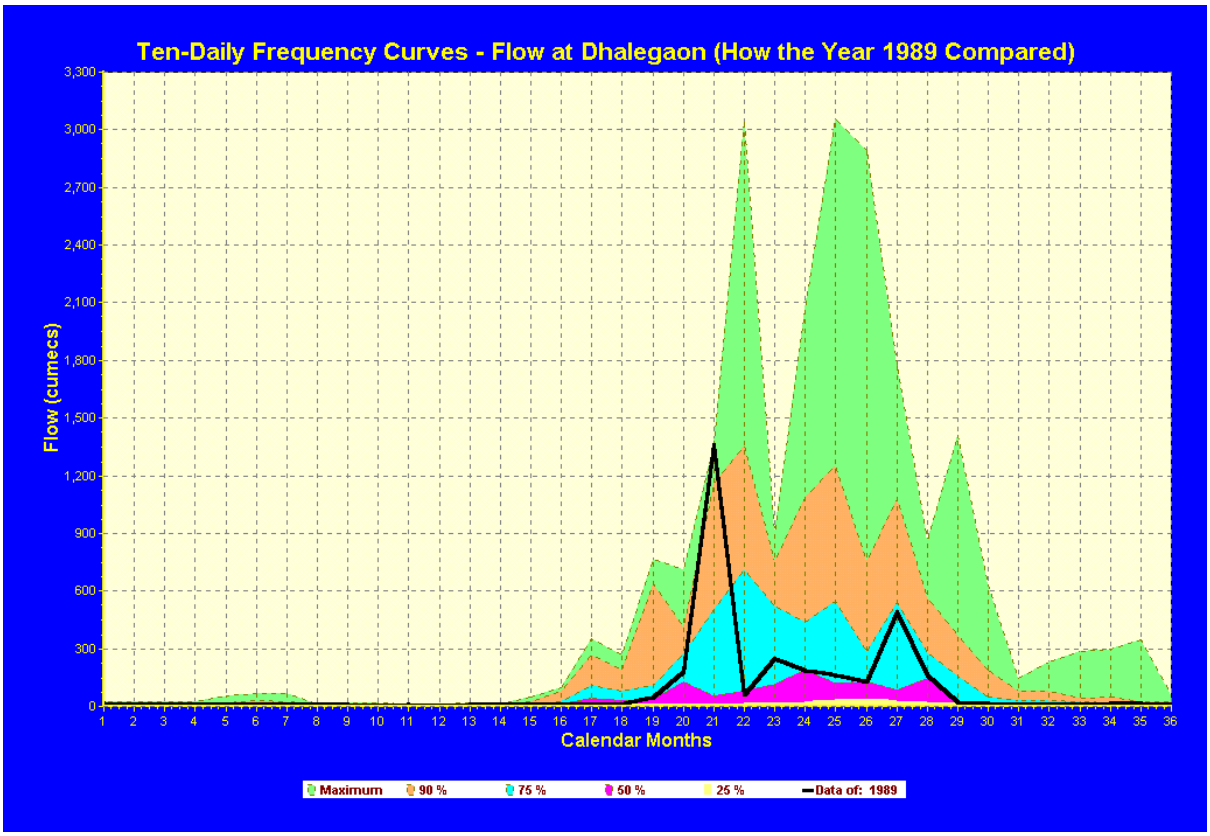


Figure 4.3.2: Ten-daily flows in a river as seen against frequency curves (based on 1970-2000 period)

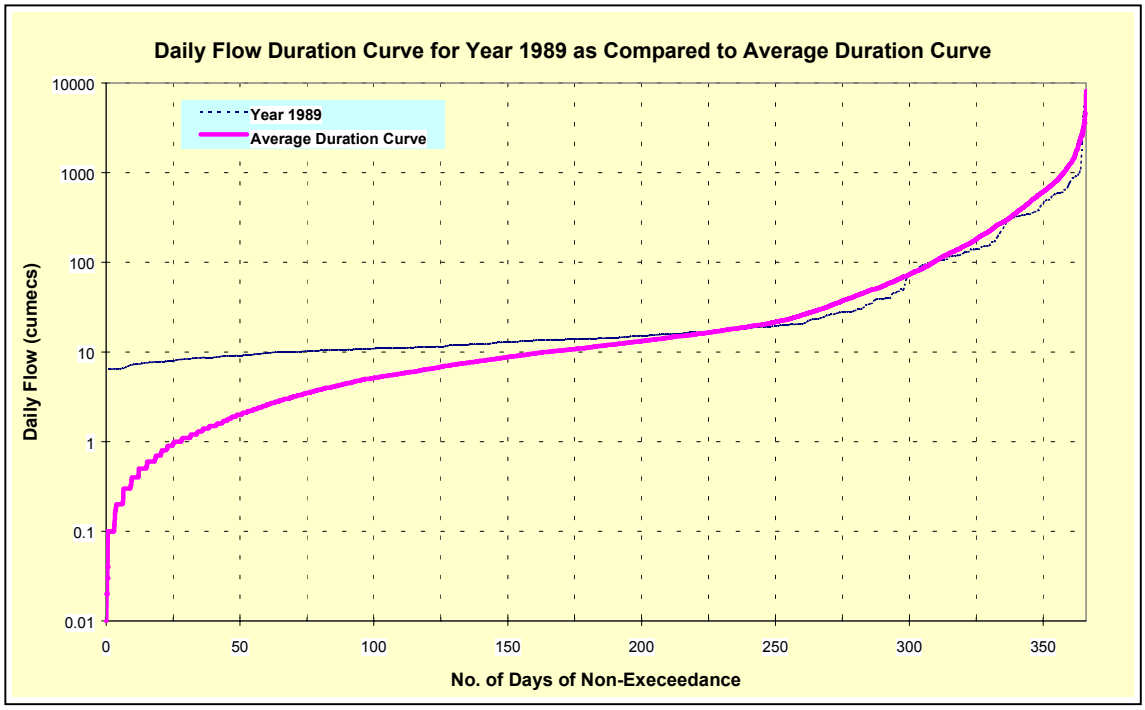
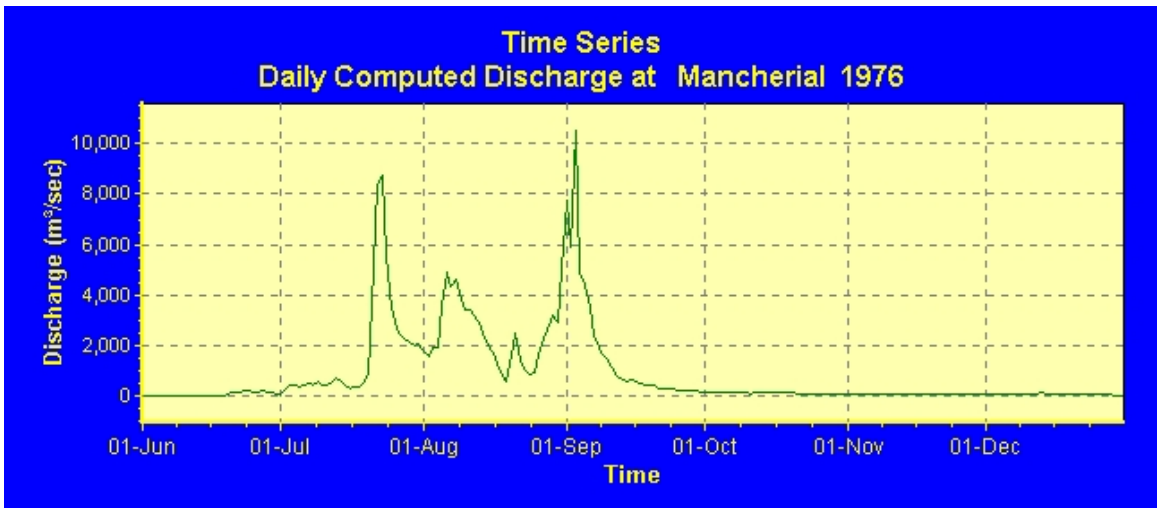


Figure 4.3.3: Average flow duration curve for daily flows in a river (based on 1970-2000 period)

Table 4.3.1 Daily Mean Flow Data **Year - 1976**

Station Code: AG000J3 **District:** Adilabad **Independent River :** Godavari
Station Name: Mancherial **Units :** m³/sec **Tributary:** -



Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	103	69.0	52.1	23.8	7.6	7.6	90.1	1760	7730	154	81.9	48.4
2	102	64.5	54.9	24.7	7.5	6.9	277	1580	5920	160	80.7	48.0
3	102	63.7	56.3	24.1	7.0	6.5	411	1890	10520	155	78.8	41.4
4	100	51.3	56.1	24.5	6.8	5.8	459	1890	4900	141	76.0	47.4
5	84.2	54.2	56.3	24.4	7.8	5.5	391	3560	4410	133	74.3	61.0
6	87.1	51.6	53.1	22.7	7.9	4.9	453	4870	3470	123	73.9	75.9
7	130	50.5	53.0	21.5	7.4	5.2	481	4360	2380	129	80.0	76.2
8	125	56.0	51.5	18.5	6.2	5.2	448	4580	1900	123	74.1	80.6
9	115	56.2	41.8	21.0	7.0	5.5	537	3850	1600	121	83.4	76.9
10	112	59.5	34.7	21.7	6.0	5.5	422	3410	1420	115	72.8	78.0
11	86.0	51.3	34.9	21.0	6.7	6.0	433	3410	1120	110	63.0	80.9
12	72.0	50.5	34.7	17.8	6.0	6.0	528	3200	760	122	62.2	104
13	83.2	50.0	32.5	16.5	6.5	6.6	685	2900	609	132	61.2	115
14	91.0	50.7	36.0	15.3	5.8	6.9	568	2450	600	127	60.0	103
15	94.0	52.0	39.0	14.4	5.7	7.3	441	1970	620	124	57.5	111
16	86.2	48.4	38.5	14.5	6.5	8.1	309	1720	575	120	60.0	83.1
17	85.8	46.1	35.9	14.8	7.9	8.4	347	1300	518	115	56.3	61.5
18	80.0	45.1	35.4	14.5	8.5	7.5	356	763	461	122	56.2	47.0
19	74.2	44.4	34.8	16.8	8.5	7.9	497	607	435	121	53.9	44.5
20	68.8	45.5	32.4	15.6	8.5	120	856	1630	398	117	64.8	55.8
21	87.4	41.1	32.0	13.4	8.7	127	5100	2510	322	105	78.0	42.4
22	83.4	41.0	30.5	12.9	8.6	165	8410	1390	310	101	65.0	52.6
23	69.9	41.0	31.2	14.5	7.0	244	8750	969	274	95.0	61.3	50.1
24	67.2	41.2	31.1	13.0	8.2	227	5620	836	275	90.0	61.8	44.9
25	72.0	39.5	30.9	13.0	6.4	182	3470	946	236	90.0	63.9	44.5
26	70.0	31.7	27.7	12.0	5.4	172	2670	1670	220	91.4	58.4	44.0
27	65.3	32.7	26.4	11.1	5.7	204	2420	2240	220	90.9	55.4	43.7
28	65.9	32.6	27.0	10.1	6.6	173	2190	2660	195	89.8	50.0	42.7
29	66.0	35.0	25.4	9.2	6.3	177	2110	3220	190	87.5	52.9	38.6
30	64.1		24.2	8.4	5.0	103	1970	2940	170	84.0	45.0	38.2
31	66.7		24.0		6.3		2050	5060		85.0		33.8
Min.	64.1	31.7	24.0	8.4	5.0	4.9	90.1	607	170	84.0	45.0	33.8
Max.	130	69.0	56.3	24.7	8.7	244	8750	5060	10500	160	83.4	115
Eff.	31	29	31	30	31	30	31	31	30	31	30	31
Miss.	0	0	0	0	0	0	0	0	0	0	0	0
Mean	85.9	48.1	37.9	16.9	7.0	67.4	1735.8	2459.6	1760.5	115.6	65.4	61.8

Yearly statistics :

Mean : 541.8	Minimum : 4.9 Date : 06/06/1976	Maximum : 10500.0 Date : 03/09/1976	No. of data : 366 No. of missing data : 0

Table 4.3.2 : Daily Mean Flow Data **Year - 1976**

Station Code: AG000S9 **District:** Parbhani **Independent River :** Godavari

Station Name: Dhalegaon **Units :** m³/sec **Tributary:** -

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	2.4	18.6	20.2	2.2	1.1	0.4	247	756	223	19.3	11.7	13.6
2	25.2	19.6	22.9	2.3	0.9	0.4	239	2550	286	57.7	11.2	12.5
3	22.9	20.6	18.9	3.0	0.9	0.5	199	3760	124	50.0	12.0	13.5
4	21.8	18.5	10.9	2.8	0.8	0.5	181	4580	141	64.3	10.3	11.9
5	22.7	19.2	8.0	3.2	0.8	0.5	176	4050	125	44.1	10.3	12.2
6	22.6	17.5	12.2	2.9	0.8	0.8	231	3240	89.8	27.6	10.1	9.8
7	22.0	18.6	14.4	2.6	0.8	1.4	230	2830	66.7	22.2	15.0	12.1
8	20.5	18.6	15.6	2.3	0.6	1.6	242	3290	47.0	21.6	14.7	11.7
9	22.0	17.2	16.4	2.0	0.6	1.5	164	2840	54.1	20.8	13.5	13.4
10	20.8	18.2	17.1	1.8	0.6	1.8	134	2520	49.3	20.0	13.0	15.2
11	20.0	18.6	14.3	2.0	0.5	270	110	2030	71.0	15.4	13.4	12.4
12	20.7	18.8	17.6	2.0	0.6	372	109	1400	55.0	17.3	12.9	11.3
13	21.4	20.0	16.0	2.0	0.5	384	100	1050	44.7	18.0	12.3	11.8
14	20.0	21.3	9.6	1.8	0.5	365	87.1	228	43.3	16.6	12.0	11.9
15	19.6	21.4	8.1	1.6	0.5	332	31.7	100	46.5	16.6	11.3	11.6
16	19.6	20.8	6.0	1.8	0.5	300	28.3	50.8	43.3	17.4	11.1	12.2
17	21.3	19.5	5.0	1.6	0.5	246	669	41.0	44.7	17.0	10.3	14.2
18	20.6	19.7	4.2	1.6	0.5	229	794	36.3	33.8	15.4	9.8	11.8
19	21.1	19.2	3.4	1.4	0.5	191	776	30.5	40.0	16.4	9.8	11.3
20	19.0	20.3	3.6	1.3	0.5	194	1400	30.4	29.9	14.8	11.7	13.0
21	19.4	15.8	3.4	1.2	0.4	161	2380	34.1	28.0	15.9	15.0	15.0
22	20.3	13.4	3.3	1.2	0.4	267	2350	40.0	27.9	15.0	21.8	13.0
23	18.9	10.5	3.0	1.2	0.4	127	1200	28.9	29.4	14.0	30.3	11.2
24	19.6	9.6	2.6	1.3	0.4	132	1220	30.0	33.2	12.0	24.2	14.4
25	19.0	7.9	2.9	1.2	0.3	126	1520	198	31.0	12.5	20.8	11.7
26	19.0	7.9	2.7	1.4	0.4	158	1430	171	58.0	10.4	19.6	10.7
27	21.2	6.8	2.6	1.1	0.4	121	1200	125	22.6	11.0	18.2	11.6
28	20.7	5.2	2.4	1.0	0.4	222	842	361	25.0	10.6	16.0	11.3
29	21.0	14.8	2.2	1.0	0.4	200	686	702	23.9	9.6	15.1	11.8
30	18.8		2.2	1.0	0.4	198	533	438	22.2	11.5	13.8	11.2
31	20.6		2.2		0.5		493	267		14.0		11.1
Min.	18.8	5.2	2.2	1.0	0.3	0.4	28.3	28.9	22.2	9.6	9.8	9.8
Max.	25.2	21.4	22.9	3.2	1.1	384	2380	4580	286	64.3	30.3	15.2
Eff.	31	29	31	30	31	30	31	31	30	31	30	31
Miss.	0	0	0	0	0	0	0	0	0	0	0	0
Mean	20.8	16.5	8.8	1.8	0.6	153.8	646.8	1222.2	65.4	20.9	14.4	12.3

Yearly statistics :

Mean : 184.3	Minimum : 0.3 Date : 25/05/1976	Maximum : 4580.0 Date : 04/08/1976	No. of data : 366 No. of missing data : 0

4.4 Surface water quality

<In order to present the water quantity and quality information in an integrated manner for the user it will be appropriate to include all water quality related summaries, figures and tables as given hereunder, in continuation of the other aspects of the hydrological processes.

On the basis of the sampling and analysis program for any river basin, a summary could be drawn. As an example for an agency that is monitoring only major ions, looking into the data obtained during the reporting period the following inferences are drawn. Hereunder, some inferences for Godawari river are put only as an illustrative example:

- The water of river is generally alkaline in nature mainly due to the presence of bicarbonate. Carbonates were present in the samples of the Godavari river collected from Nashik, Nanded, Mancherial, Bhadrachalam and Pollavaram.
- The conductivity varies between 300 μmhos/cm (Nashik Upstream) to 2000 μmhos/cm (Polavaram).
- Among cations calcium, magnesium and sodium were dominating. Potassium is always low. Maximum calcium (600 mg/L) was observed in Polavaram. Aluminium, iron and ammoniacal nitrogen were present in very small quantity. In general the cations are present in permissible limits of drinking water or irrigation standards of BIS.
- The chloride was the dominant anion followed by bicarbonate and sulphate. The maximum chloride (1000 mg/L) was recorded from Polavaram and bicarbonate (8,000 mg/L) from Polavaram. Silicates were present in significant concentration at all the stations. Nitrates, fluorides and phosphates were present in low concentrations at all the stations throughout the year. In general the anions were within the permissible limit of drinking and irrigation requirements of the Bureau of Indian Standards (BIS).
- Dissolved oxygen was generally at saturation level at all the monitoring stations except the stretch between Nashik to Nanded, where DO was observed as low as zero on many occasions, which may be due to effect of discharge of untreated domestic sewage from Nashik, Aurangabad, Nanded.
- Water quality indices viz. Sodium adsorption ratio (SAR), sodium percentage (%Na) and residual sodium carbonate (RSC) were found within tolerance limits of irrigation standards at all the stations. From salinity classification point of view,the river waters of the basin generally fall under C1S1 and C3S1 classification as per US Salinity diagram.

Overview of water quality for different stations

Station: POLAVARAM

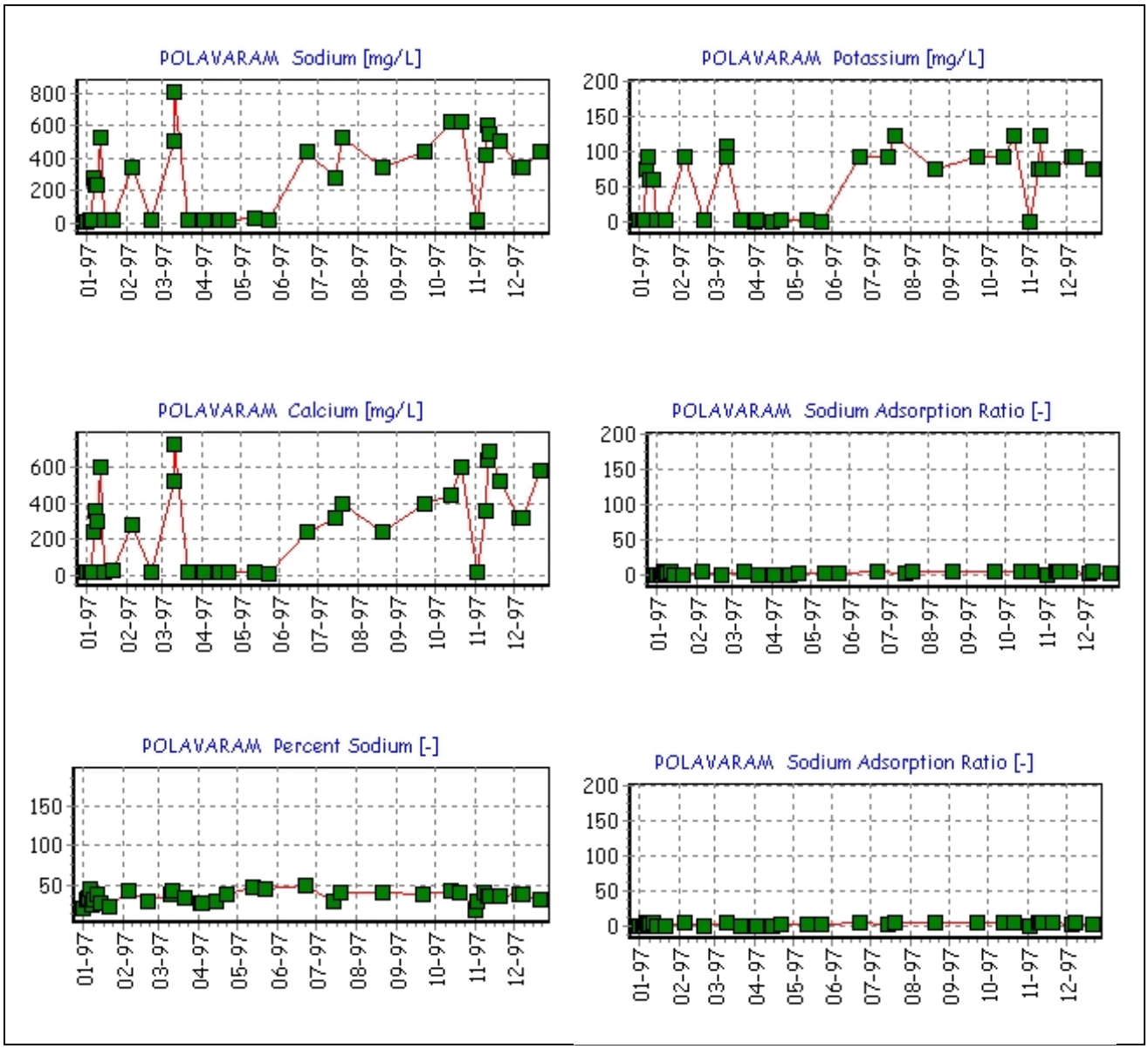


Figure: 4.4.1: Major cations for station Polavaram in the year 1997

Station: POLAVARAM

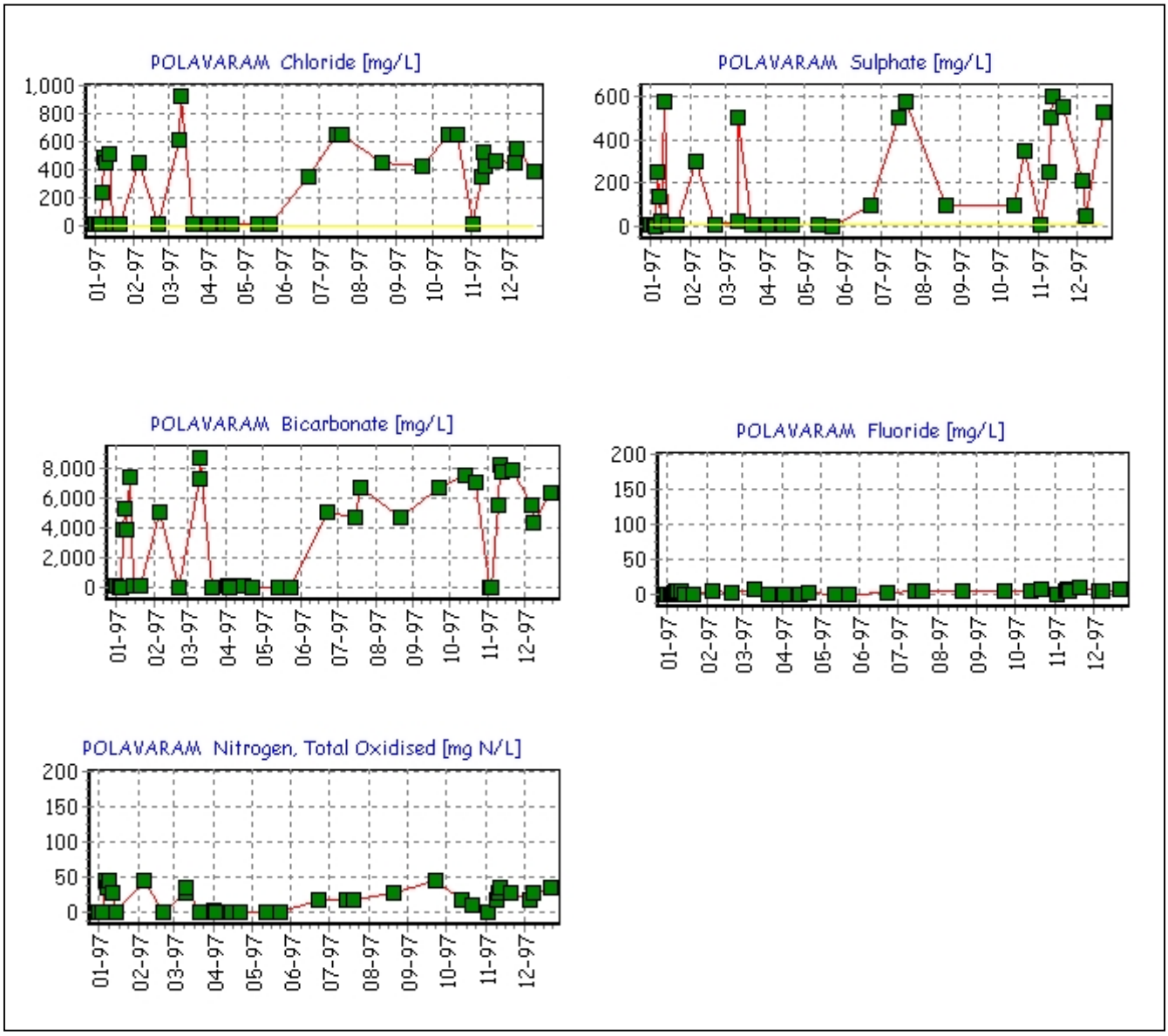


Figure 4.4.2: Major anions for station Polavaram in the year 1997

☞ ***Templates for several parameter groups are available in HYMOS (major cations and anions, general and pollution related parameters), this way Figure 4.4.1 and 4.4.2 can be generated automatically for a desired number of stations best representing a river stretch or a basin.***

Comparing with standards

The CPCB results which contain some of the parameters related to major water quality issues of the river were analysed (Table 7). The main findings are presented as follows:

The Godavari river basin is a relatively clean river in the country.

The major water quality problem in the river basin is mainly due to presence of coliform. A large number of stations are showing coliform values higher than the desired limits identified under “designated best use” criteria of CPCB.

☞ ***The water quality results for the year of interest are compared to water quality standards after summarising the time-series over the year (or season). Normally the 90 percentile value (10% for DO) is used for comparison with standards such as given in Table 4.4.1.***

Fitness for use classification according to CPCB system "ABCDE"

1996	Water for Drinking & Bathing (ABC)				Wildlife (D)		Irrigation (E)			
Station	pH	CTM	BOD	DO	pH	DO	pH		B	SAR
AG000C3	8.2	-	0.4	6.4	8.2	6.4	8.2	212	-	0
AG000C7	-	-	-	-	-	-	-	-	-	-
AG000G7	8.3	-	0.5	6.4	8.3	6.4	8.3	339	-	1
AG000J3	8.4	-	0.7	6.5	8.4	6.5	8.4	468	-	1
AGH00C4	8.1	638.7	1.3	6.8	8.1	6.8	8.1	393	-	1
AGH30E2	8.1	3742.8	3.7	6.0	8.1	6.0	8.1	515	-	1
AGH30F6	-	-	-	-	-	-	-	-	-	-
AGH30Q1	8.2	502.8	1.0	7.2	8.2	7.2	8.2	413	-	1
AGH30S9	8.2	166.5	1.7	6.1	8.2	6.1	8.2	410	-	1
AGH32D5	8.2	1502.9	1.2	7.2	8.2	7.2	8.2	429	-	1

Table 4.4.1: Station-wise water quality problems according to the CPCB classification

Legend

Drinking water (A), Bathing water (B) and Source for drinking water (C)							
pH		Col		BOD		DO	
6	-	50	A	2	A	4	-
6.5	C	500	B	3	B	5	C
8.5	A	5000	C		-	6	B
9	C		-				A
	-						
Wildlife (D)							
pH		DO					
6.5	-	4	-				
8.5	D		D				
	-						
Irrigation (E)							
pH		EC		B		SAR	
6.5	-	2250	E	2	E	26	E
8.5	E		-		-		
	-						

Water Quality Yearbook - HYMOS example Report 4 (HP, 2002)

Trends in water quality

A time series plot for BOD (3years period i.e. from 1996 to 1998), all dates and annual average is plotted as shown in Table 4.4.2 and Figure 4.4.3 below. As reveal from the graph BOD values up to 1997 was vary between 0.1 and 1.1 mg/L with an average of 0.45 mg/L. The observed increase of the maximum and average value in 1998, 1.4 and 0.67 mg/L respectively are very small compared tot the large spread of the data, caused by the sharp decrease in the number of observation in 1998. The data do therefore not indicate an significant increase in BOD at this station.

☞ ***A similar plot as presented in Figure 2 may be included for longitudinal analysis of a river or river stretch: different monitoring stations are presented on the horizontal axis the graph.***

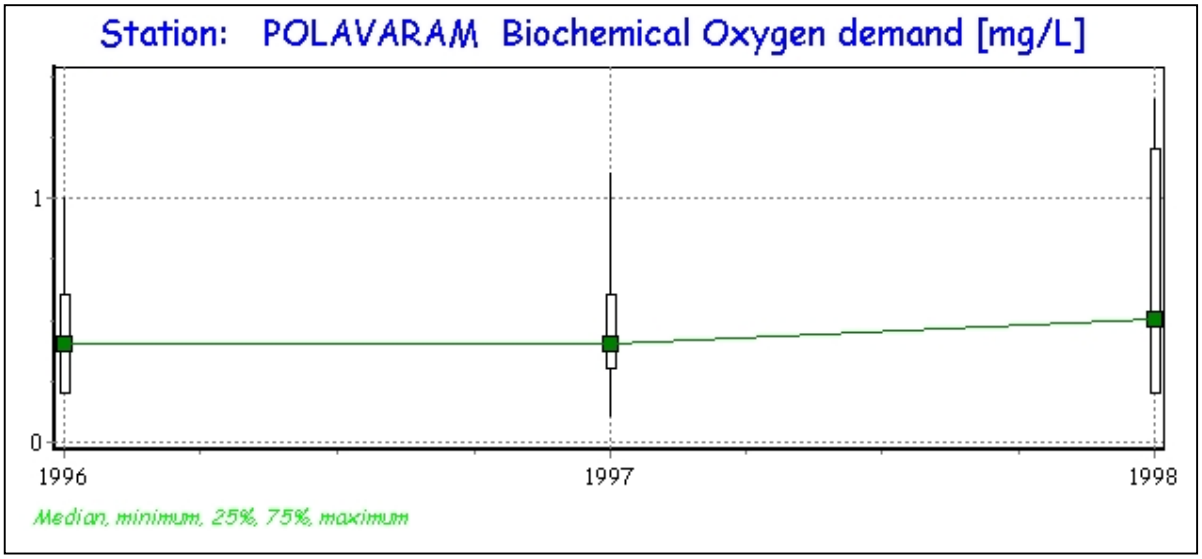


Figure 4.4.3: Box-whisker graph for BOD at station Polavaram

Year	1995	1996	1997	1998	1999	2000
Max	-	1.000	1.100	1.400	-	-
Mean	-	0.444	0.463	0.667	-	-
Min	-	0.200	0.100	0.200	-	-
Median	-	0.400	0.400	0.500	-	-
10%	-	0.200	0.200	0.200	-	-
25%	-	0.200	0.300	0.200	-	-
75%	-	0.600	0.600	1.200	-	-
90%	-	0.900	0.800	1.400	-	-
No. data	0	32	32	6	0	0

Table 4.4.2: Yearly time-series for summary statistics for BOD at station Polavaram

Results of Surveillance Monitoring

Surveillance monitoring will/may not be conducted at the same locations or for the same parameters from year to year. Thus presentation of these results needs to be separate from the above items. Since surveillance monitoring is conducted for problem issues, it is assumed that the results of these studies are relevant for a yearbook. Surveillance monitoring likely will be for pollution parameters such as (e.g. coliforms, heavy metals, pesticides, organic pollutants, ammonia).

- ***Include a map showing location(s) monitored***
- ***Time series plots for 1 year, for the WQ parameters included in the surveillance. The water quality standard must be represented as a horizontal line.***
- ***Comparison of stations for the period of monitoring.***

5 Interpretation of various statistics presented in the yearbook

<It is very important to explain to the readers what the various statistics that are used in the yearbook mean. Some such important terms that need to be explained are as follows;

5.1 Daily rainfall – What time frame does daily rainfall refers to.

5.2 Mean Daily Runoff – How is the mean daily runoff computed.>

6 Options for users for receiving data from the Data Centres

<Now that the dissemination of hydrological data would become very efficient and user-friendly for the data users, it will be useful to give wider publicity to the data retrieval options available to the users. Few aspects which could be put as brief note for this purpose are:

6.1 What major types of hydrological data and information is available in HIS

6.2 What is the extent of data availability in terms of number of stations, length of data on different data types and overall volume of data. This would also incidentally give the

6.3 How a user can request for the data

6.4 What would be the cost of data

6.5 Who would qualify as eligible data users and could get data.>

7 Previous publications of water yearbooks

<It is very important to include a sort of bibliography on the water yearbooks published in the past, highlighting the salient features or the major changes that were introduced from time to time. Such note enlisting what has been published so far by the agency would become a ready reference to anybody seeking to know what type of information is available on hydrological data and how it could be approached for.>

ANNEX II:

STATISTICAL ANALYSIS WITH REFERENCE TO RAINFALL AND DISCHARGE DATA

Table of Contents

1 Introduction

Terminology

A **hydrologic process** is defined as any phenomenon concerning the occurrence and movement of water near the earth's surface continuously changing in time and/or space. If these phenomena are observed at intervals or continuously, **discrete**, respectively, **continuous series** are created, → with time: discrete and continuous time series. One single series element is an **outcome** of the process. A set of outcomes is called a **realisation**, while the set of all possible outcomes is the **ensemble**.

The variation within hydrological processes may be deterministic or stochastic. In a **deterministic process** a definite relation exists between the hydrologic variable and time (or space). The functional equation defines the process for the entire time (or space) of its existence. Each successive observation does **not** represent new information about the process. This, in contrast to a **stochastic process**, which evolves, entirely or in part, according to a random mechanism. It means that future outcomes of the process are not exactly predictable. The hydrologic variable in such cases is called a **stochastic variable**, i.e. a variable whose values are governed by the laws of chance. Its behaviour is mathematically described by probability theory.

The elements, creating a stochastic process, may be **dependent** or **independent**, resulting in **a non-pure random**, respectively, a **pure random** process.

A stochastic process can either be **stationary** or **non-stationary**, i.e. homogeneous or non-homogeneous in time and/or space. Stationary processes are distinguished into **strictly** and **weakly** stationary processes.

A process is said to be **strictly stationary** if all its statistical properties which characterise the process, are unaffected by a change in the origin (time and or space). For a time-process this reads: the joint probability distribution of $x(t_1)$, $x(t_2)$, …, $x(t_n)$ is identical to the joint probability distribution of $x(t_1+\tau)$, $x(t_2+\tau)$, …, $x(t_n+\tau)$ for any n and τ, where τ is a time lag. If instead of the joint probability density function only the first m-moments of that function are independent of time (space) the process is called m^{th} order stationary.

Weak stationarity means that only the lower order moments of the distribution function (order ≤ 2, i.e. the mean and the covariance function) fulfil the property of being independent of time. This is also called stationarity in a **wide sense**. (Note that the terminology stationary/non-stationary is used when dealing with homogeneity or non-homogeneity in **time**).

In practice only a limited set of outcomes, a limited series, is available. Based on this sample set the behaviour of the process is estimated: sample versus population. The elements in a hydrological series may be analysed according to **rank of magnitude** and according to the **sequence of occurrence**. Ranking of elements forms the basis of statistics, the classical **frequency analysis**, thereby ignoring the order of occurrence. In contrast to ranking, the study of the sequence of occurrence presumes that past outcomes of the process may influence the magnitude of the present and the future outcomes. Hence the dependency between successive elements in the series is not ignored: **time series analysis**.

About this module

In this module a review is presented of statistics as applied to hydrology to analyse e.g. rainfall or discharge data. With statistics one describes rather than explains features of hydrological processes. Statements are made based on a sample from the entire population of the hydrological variable of concern. With statistics one describes variables only in probabilistic terms for reasons that the cause and effect relation of the physical process is insufficiently known and also because our description is based on a small part of the entire range of outcomes on the variable.

Statistics provides powerful tools to describe hydrological variables, but one should apply it with care. An important condition the series to be subjected to statistical analysis should fulfil is **stationarity**. To judge whether this condition is fulfilled, knowledge is required of the nature of the hydrological variable(s) of concern. The following components are generally distinguished in hydrological time series, see also Figure 1.1:

- **Deterministic components,** including:
  - Transient component, due to natural or man made changes, which can be a jump, in case of a sudden change in the conditions or a trend, linear or non-linear, due to a gradual change
  - Periodic component, e.g. due to the annual solar cycle
- **Stochastic component:**
  - Stochastic dependent part, where the new value is related to one or more predecessors, e.g. due to storage effects
  - Stochastic independent or random part.

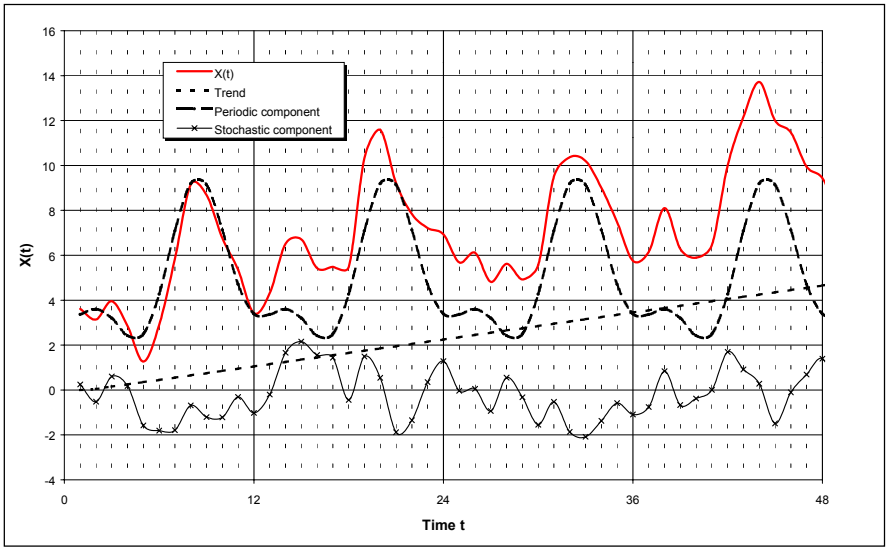


Figure 1.1: Components of a hydrological time series

Figure 1.1 displays a monthly time series, with a clear linear trend and a strong periodic component, repeating itself every year. It will be clear that a series as shown in Figure 1.1 does not fulfil the stationarity condition, the mean value gradually shifts due to the trend. Even with the trend removed the probability distribution changes from month to month due to the existence of the periodic component, again not fulfilling the stationarity condition. If one also eliminates the periodic component in the mean value a process with a stationary mean value is obtained, but still this may not be sufficient as generally also second or higher order moments (variance, covariance, etc.) show periodicity. Therefore, hydrological time series

with time intervals less than a year should not be subjected to statistical analysis. Annual values generally do not have the problem of periodicity (unless spectral analysis shows otherwise due to some over-annual effect) and are fit for statistical analysis, provided that transient components are not present or have been eliminated.

Now, returning to our monthly series, periodicity is avoided if the months are considered separately, that is e.g. if only the values of July of successive years are considered. Similarly, if seasonal series are available, one should consider one season at a time for statistical analysis, i.e. the same season for a number of years.

To illustrate the above considerations monthly rainfall and its statistics of station Chaskman are shown in Figures 1.2 and 1.3. As can be observed from Figure 1.3, there is a strong periodic component in the time series; the mean and standard deviation vary considerably from month to month.

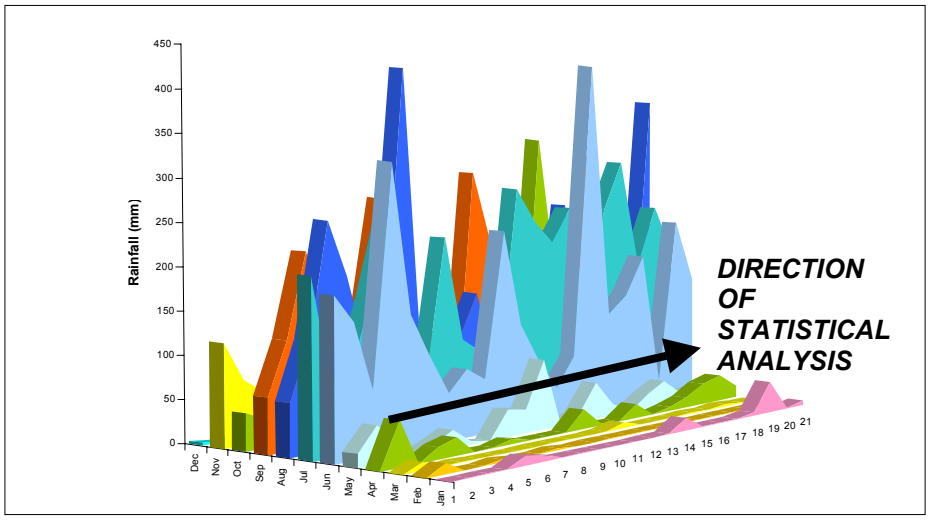


Figure 1.2: 3-D plot of monthly rainfall of station Chaskman

If one would combine the rainfall values of all months one assumes that their probability distribution is the same, which is clearly not so. To fulfil the stationarity condition, statistics is to be applied to each month separately, see Figure 1.2

A series composed of data of a particular month or season in successive years is likely to be serially uncorrelated, unless over-annual effects are existent. Hence, such series will be fully random. Similar observations apply to annual maximum series. It implies that the time sequence of the series considered is unimportant. Above considerations are typical for statistical analysis.

In this module statistics is discussed and the following topics will be dealt with:

- Description of data sets
- Probabilistic concepts
- Discrete and continuous probability distributions
- Estimation of distribution parameters
- Making statistical inference

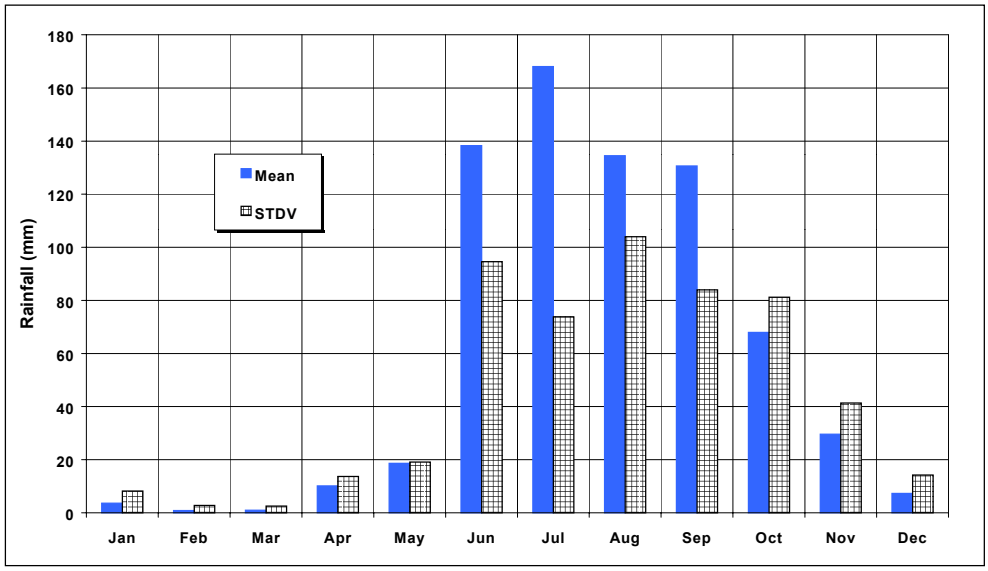


Figure 1.3: Mean and standard deviation of monthly rainfall series of station Chaskman, period 1977 - 1998

2 Description of Datasets

2.1 General

In this sub-section on basic statistics attention will be given to:

- Graphical presentation of data
- Measures of central tendency
- Measures of dispersion
- Measure of asymmetry: skewness
- Measure of peakedness: kurtosis
- Percentiles
- Box plots
- Covariance and correlation coefficient

2.2 Graphical representation

For graphical presentation of the distribution of data the following options are discussed:

- Line diagram or bar chart
- Histogram
- Cumulative relative frequency diagram
- Frequency and duration curves

Note: prior to the presentation of data in whatever frequency oriented graph, it is essential to make a time series plot of the data to make sure that a strong trend or any other type of inhomogeneity, which would invalidate the use of such presentation, does not exist.

Line Diagram or Bar Chart

The occurrences of a **discrete** variate can be classified on a line diagram or a vertical bar chart. In this type of graph, the horizontal axis gives the values of the discrete variable, and the occurrences are represented by the heights of vertical lines. The horizontal spread of these lines and their relative heights indicate the variability and other characteristics of the data. An example is given in Figure 2.1, where the number of occurrences that in one year the monthly rainfall at Chaskman will exceed 100 mm is presented. The period presented refers to the years 1978 – 1997.

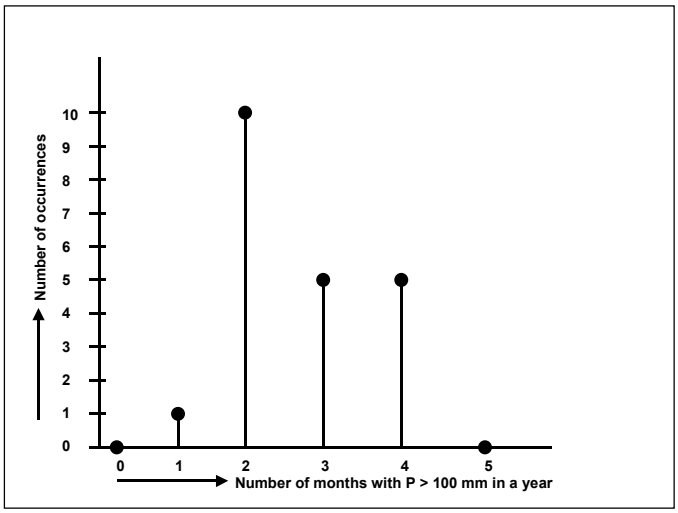


Figure 2.1:
Line diagram of number of months in a year with rainfall sum > 100 mm for period 1978 - 1997

If the number of entries on the horizontal axis would have been small, it means that the variability in the number of months in a year with P > 100 mm is small.

Histogram

If the range of outcomes on the variable is becoming large, then the line diagram is not an appropriate tool anymore to present the distribution of the variable. Grouping of data into classes and displaying the number of occurrences in each class to form a histogram will then provide better insight, see Figure 2.2. By doing so information is lost on the exact values of the variable, but the distribution is made visible. The variability of the data is shown by the horizontal spread of the blocks, and the most common values are found in blocks with the largest areas. Other features such as the symmetry of the data or lack of it are also shown. At least some 25 observations are required to make a histogram.

An important aspect of making a histogram is the selection of the number of classes n_c and of the class limits. The following steps are involved in preparing a histogram:

- The number of classes is determined by one of the following options (see e.g. Kottegoda and Rosso (1997):

$$n_c = \sqrt{N} \tag{2.1}$$

$$n_c = \frac{R\sqrt[3]{n}}{2R_{iq}} \tag{2.2}$$

where: n_c = number of classes

n = number of observations

R = range of observations: $X_{max} – X_{min}$

R_{iq} = interquartile range, defined by: $R_{iq} = M_{up} – M_{low}$

M_{up} = median of highest 50% of the data, i.e. 75% of the data is less

M_{low} = median of lowest 50% of the data, i.e. 25% of the data is less

- To obtain rounded numbers for the class limits convenient lower and upper limits below X_{min} and above X_{max} respectively the lowest and highest value have to be selected.
- Count the occurrences within each class: class frequency
- Present the results in a histogram

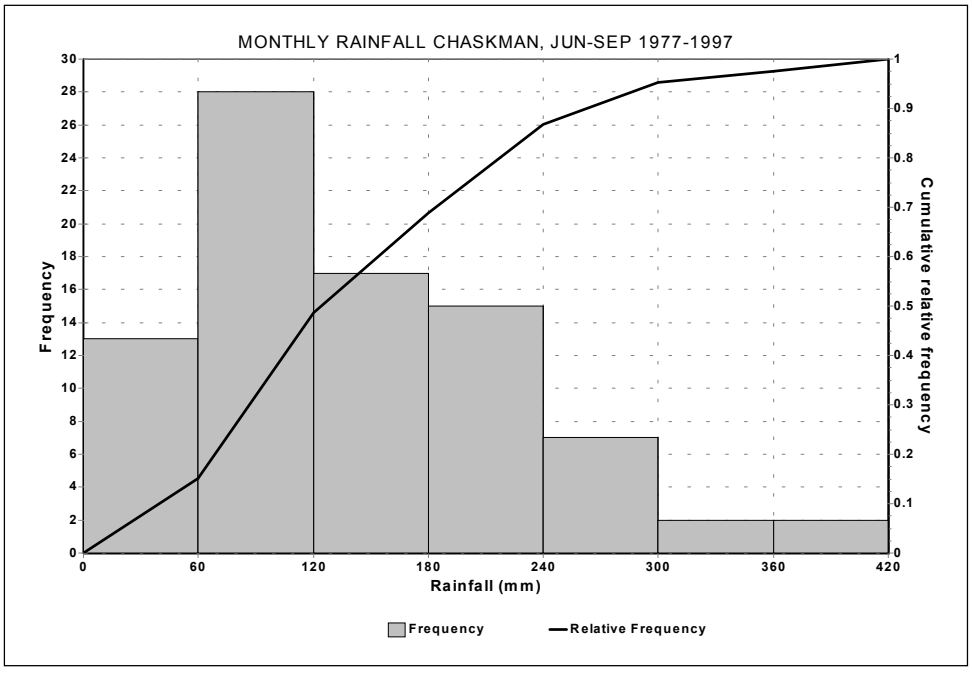


Figure 2.2: Histogram and cumulative relative frequency diagram of monthly rainfall at Chaskman, months June-September, period 1977 – 1997.

The application is shown for monthly rainfall of Chaskman. From Figure 1.2 it is observed, that rainfall in the months June to September behave more or less like a homogeneous group of data. A histogram is made of these monthly values for the years 1977-1997, i.e. 21 years of data. Hence in total the data set comprises 21 x 4 = 84 data points. The data are ranked in ascending order and displayed in Table 2.1

	1	2	3	4	5	6	7	8	9
1	12.1	55.4	*71.8*	92.8	118.1	152.2	196.3	229.0	326.2
2	19.6	55.8	72.2	97.8	124.8	154.4	201.2	234.6	342.6
3	20.8	55.8	74.8	100.2	127.2	158.0	*202.8*	237.2	404.6
4	26.6	61.2	75.4	101.4	128.0	160.2	206.4	258.0	418.7
5	35.4	61.8	75.8	101.4	130.2	161.0	207.0	258.8	
6	37.2	62.8	76.6	103.0	132.8	166.8	221.2	268.2	
7	48.8	64.6	77.4	103.8	136.0	169.2	221.4	268.4	
8	52.4	65.0	77.6	105.2	136.6	172.8	225.7	281.4	
9	52.8	65.6	78.9	105.7	144.0	188.0	227.6	281.8	
10	53.4	69.8	87.2	112.4	148.0	193.4	228.4	282.3	

Table 2.1: June-September monthly rainfall at Chaskman 1977-1997 ranked in ascending order

The values for X_{min} and X_{max} are respectively 12.1 mm and 418.7 mm, hence for the range it follows R = 418.7 – 12.1 = 406.6 mm. Since 84 data points are available 42 data are available in the lowest as well as in the highest group, so the values at positions 21 and 63 in the sorted data vector will give the medians for the lowest and highest 50% of the data M_{low} and M_{up}. These values are respectively 71.8 mm and 202.8 mm, hence the interquartile range is R_{iq} = 202.8 – 71.8 = 131.0 mm. According to (2.1) the number of classes in the histogram should be

$$n_c = \frac{R\sqrt[3]{n}}{2R_{iq}} = \frac{406.6x(84)^{1/3}}{2x131.0} = 6.8 \approx 7$$

Now, with 7 classes, R = 406.6 mm a class interval should be $\geq$ R/7 $\approx$ 58 mm, which is rounded to 60 mm. Using this class-interval and since X_{min} = 12.1 mm and X_{max} = 418.7 mm appropriate overall lower and upper class limits would be 0 mm and 420 mm. The result is displayed in Figure 2.2. The data points in a class are > the lower class limit and $\leq$ the upper class limit, with the exception of the lowest class, where the lowest value may be = lower class limit.

Note that if one uses (2.1) the result would have been $\sqrt{84} \approx 9$ classes, which is a slightly higher value. It follows that the guidelines given in (2.1) and (2.2) are indicative rather than compulsory. In general, at least 5 and at maximum 25 classes are advocated. Equation (2.2) has preference over equation (2.1) as it adapts its number of classes dependent on the peakedness of the distribution. If the histogram is strongly peaked then the inter-quantile range will be small. Consequently, the number of classes will increase, giving a better picture of the peaked zone.

Cumulative Relative Frequency Diagram

By dividing the frequency in each class of the histogram by the total number of data, the relative frequency diagram is obtained. By accumulating the relative frequencies, starting off from the lower limit of the lowest class up to the upper limit of the highest class the cumulative relative frequency diagram is obtained. For the data considered in the above example, the cumulative relative frequency diagram is shown with the histogram in Figure 2.2. The computational procedure is shown in Table 2.2.

Class	LCL	UCL	Freq.	Rel. Freq.	Cum.R. Fr.
1	0	60	13	0.155	0.155
2	60	120	28	0.333	0.488
3	120	180	17	0.202	0.690
4	180	240	15	0.179	0.869
5	240	300	7	0.083	0.952
6	300	360	2	0.024	0.976
7	360	420	2	0.024	**1.000**

Table 2.2:
Computation of cumulative relative frequencies

On the vertical axis of the graph, this line gives the cumulative relative frequencies of values shown on the horizontal axis. Instead of deriving this plot via the histogram, generally it is made by utilising and displaying every item of data distinctly. For this purpose, one ranks the series of size N in ascending order. The cumulative frequency given to the observation at rank *m* then becomes *m*/N, i.e. there are m data points less than or equal to the data point at rank *m.* This is shown in Figure 2.3.

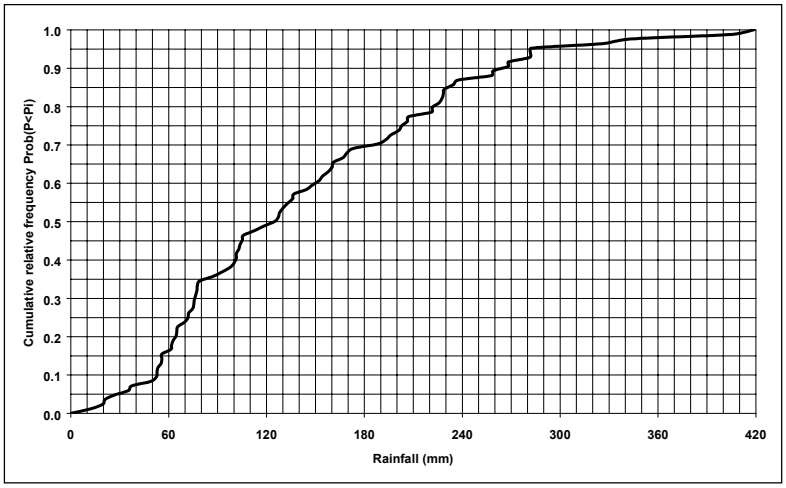


Figure 2.3: Cumulative relative frequency distribution for Chaskman June–September data in the period 1977-1997

In Figure 2.3 the highest ranked data point (*m* = N) gets a cumulative relative frequency of *m*/N = N/N = 1. To describe the distribution of the data in that particular sample series this statement is correct. No observation exceeded the maximum value. However, in statistics one wants to say something about the distribution of data in the population of which the N observations are just one of many possible samples series. The cumulative relative frequency (crf) is then replaced by the non-exceedance probability. A non-exceedance probability of 1 for the maximum observed in the sample series would then imply that all possible outcomes would be less than or equal to that maximum. Unless there is a physical limit to the data such a statement is not justified. The non-exceedance probability of the maximum in the sample series will be less than 1. The non-exceedance probability to be given to the data point with rank *m* can be determined by viewing the series of ranked observations as order statistics: X(1), X(2), X(3), …,,X(*m*),…X(N). The expected value of order statistic X(*m*) depends first of all on the rank of X(m) relative to X(N). Furthermore is the expected value of X(*m*) a function of the probability distribution of the process from which the sample points are drawn. This will be discussed in more detail in Section 4.

Frequency Curves

Considering again the monthly rainfall series of Chaskman, for each month one can make a cumulative frequency distribution. Distinct crf's are the identified, say e.g. 10%, 50% and 90%, for each month. By displaying the rainfall having say a crf = 10% for all months in the year in a graph a frequency curve is obtained. Similarly for other crf's such a curve can be made. This is shown in Figure 2.4.

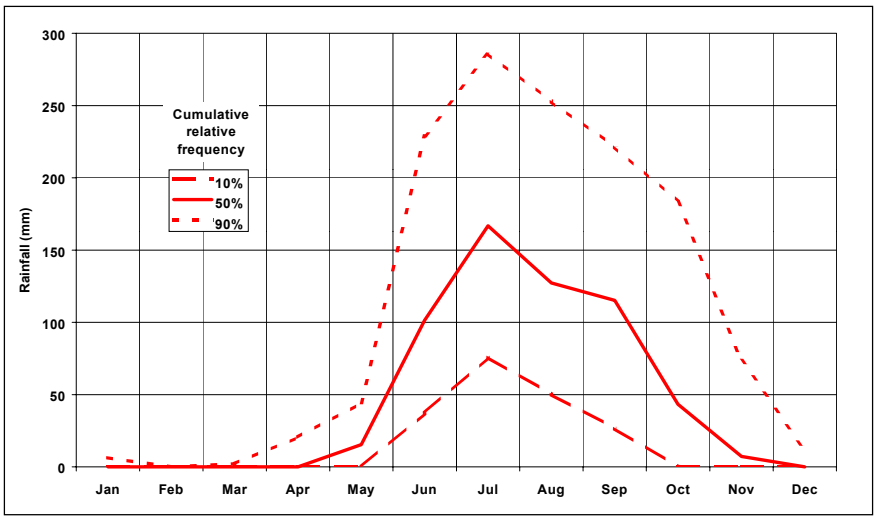


Figure 2.4: Frequency curves of monthly rainfall at station Chaskman, period 1968-1997

The computational procedure to arrive at the frequency curves is presented in the Tables 2.3 and 2.4. In Table 2.3 the actual monthly rainfall for a 30-year period is displayed. Next, the data for each month are put in ascending order, see Table 2.4, with the accompanying crf presented in the first column. The rows with crf = 0.1 (10%), 0.5 (50%) and 0.9 (90%) are highlighted and displayed in Figure 2.4.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
1968	0	0	0	0	0	49.8	144.3	60.5	162.1	43.4	47	0	507.1
1969	0	0	0	0	0	10.5	320.2	267.1	81.2	0	50.2	0	729.2
1970	0	0	0	0	60.8	80	124.9	140.5	30	162.6	0	0	598.8
1971	0	0	0	0	44.4	159.2	85.4	197.8	212.6	12	0	0	711.4
1972	0	0	0	3.2	31.4	46	229.7	38.3	0	0	0	0	348.6
1973	0	0	8.6	0	25	85	312.6	134.6	109.2	101.6	0	0	776.6
1974	0	0	0	0	132.2	72	150.8	175.2	206.2	183.4	0	0	919.8
1975	0	0	0	0	8	123.2	146.2	139.4	191.8	111.6	0	0	720.2
1976	0	0	0	0	0	494.8	323.8	208.6	115.2	3	139.2	0	1284.6
1977	0	0	0	0	16.4	188	207	61.8	64.6	44.2	119.4	3.6	705
1978	0	12.6	10.4	53	43.4	154.4	77.4	127.2	124.8	32.8	73.6	0.2	709.8
1979	0	0	0	0	21.4	75.4	93.8	252.8	221.4	13.4	57	0	735.2
1980	0	0	1	14.2	1.4	325.8	169.2	192.4	136.6	17.8	57.6	11.8	927.8
1981	10.8	2.2	0	20.6	9.2	152.2	258	101.4	160.2	53	7.4	2.2	777.2
1982	6.4	0	0	0	21.2	101.4	71.8	144	132.8	47.8	39.2	0	564.6
1983	0	0	0	0	4.2	55.8	75.8	418.4	268.2	2.8	0	0	825.2
1984	0	1.8	0	5.2	0	78.9	225.45	55.4	104.2	0	7	0	477.95
1985	0	0	0	0	31.6	62.8	105.7	74.8	26.6	91.3	0	0	392.8
1986	0	0	0	0	26.8	229	87.2	97.8	105.2	5.1	3.6	46.6	601.3
1987	0	0	0	0	80.2	118.1	65	148	12.1	89.4	8	11.9	532.7
1988	0	0	0.4	22.2	0	72.2	268.4	53.4	282.3	18.1	0	0	717
1989	0	0	6.4	1.4	9.2	37.2	227.6	61.2	190.7	7.6	0	0	541.3
1990	13.8	0	0	0	33.2	66.6	212	161.4	32.4	195.8	18	3.2	736.4
1991	0	0	0	16.2	12.8	404.6	235.4	50.2	48.6	21	9.4	0.6	798.8
1992	0	0	0	0	10.6	112.4	102	235.2	202.8	13.8	20	0	696.8
1993	0	0	1	3.8	15.8	130.2	226.4	66.6	53.4	304	7.2	31.6	840
1994	3.8	0	0	11.4	26	169	285.8	92.2	85	130.8	40.6	0	844.6
1995	35.2	0	2.2	17	15.4	20.8	157.8	19.8	262	87.8	2.2	0	620.2
1996	0.6	0	0	29.4	10.2	206.4	221.2	55.8	128	217.4	2.4	0	871.4
1997	5.4	0	0	12.4	10	136	166.8	342.6	77.6	66.3	148.7	41	1006.8

Table 2.3: Monthly and annual rainfall at station Chaskman, period 1968-1997

Crf	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
0.033	0	0	0	0	0	10.5	65	19.8	0	0	0	0	348.6
0.067	0	0	0	0	0	20.8	71.8	38.3	12.1	0	0	0	392.8
0.100	**0**	**0**	**0**	**0**	**0**	**37.2**	**75.8**	**50.2**	**26.6**	**0**	**0**	**0**	**478.0**
0.133	0	0	0	0	0	46	77.4	53.4	30	2.8	0	0	507.1
0.167	0	0	0	0	0	49.8	85.4	55.4	32.4	3	0	0	532.7
0.200	0	0	0	0	1.4	55.8	87.2	55.8	48.6	5.1	0	0	541.3
0.233	0	0	0	0	4.2	62.8	93.8	60.5	53.4	7.6	0	0	564.6
0.267	0	0	0	0	8	66.6	102	61.2	64.6	12	0	0	598.8
0.300	0	0	0	0	9.2	72	105.7	61.8	77.6	13.4	0	0	601.3
0.333	0	0	0	0	9.2	72.2	124.9	66.6	81.2	13.8	0	0	620.2
0.367	0	0	0	0	10	75.4	144.3	74.8	85	17.8	2.2	0	696.8

Crf	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
0.400	0	0	0	0	10.2	78.9	146.2	92.2	104.2	18.1	2.4	0	705.0
0.433	0	0	0	0	10.6	80	150.8	97.8	105.2	21	3.6	0	709.8
0.467	0	0	0	0	12.8	85	157.8	101.4	109.2	32.8	7	0	711.4
0.500	**0**	**0**	**0**	**0**	**15.4**	**101.4**	**166.8**	**127.2**	**115.2**	**43.4**	**7.2**	**0**	**717.0**
0.533	0	0	0	0	15.8	112.4	169.2	134.6	124.8	44.2	7.4	0	720.2
0.567	0	0	0	0	16.4	118.1	207	139.4	128	47.8	8	0	729.2
0.600	0	0	0	1.4	21.2	123.2	212	140.5	132.8	53	9.4	0	735.2
0.633	0	0	0	3.2	21.4	130.2	221.2	144	136.6	66.3	18	0	736.4
0.667	0	0	0	3.8	25	136	225.45	148	160.2	87.8	20	0	776.6
0.700	0	0	0	5.2	26	152.2	226.4	161.4	162.1	89.4	39.2	0.2	777.2
0.733	0	0	0	11.4	26.8	154.4	227.6	175.2	190.7	91.3	40.6	0.6	798.8
0.767	0	0	0	12.4	31.4	159.2	229.7	192.4	191.8	101.6	47	2.2	825.2
0.800	0.6	0	0.4	14.2	31.6	169	235.4	197.8	202.8	111.6	50.2	3.2	840.0
0.833	3.8	0	1	16.2	33.2	188	258	208.6	206.2	130.8	57	3.6	844.6
0.867	5.4	0	1	17	43.4	206.4	268.4	235.2	212.6	162.6	57.6	11.8	871.4
0.900	**6.4**	**0**	**2.2**	**20.6**	**44.4**	**229**	**285.8**	**252.8**	**221.4**	**183.4**	**73.6**	**11.9**	**919.8**
0.933	10.8	1.8	6.4	22.2	60.8	325.8	312.6	267.1	262	195.8	119.4	31.6	927.8
0.967	13.8	2.2	8.6	29.4	80.2	404.6	320.2	342.6	268.2	217.4	139.2	41	1006.8
1.000	35.2	12.6	10.4	53	132.2	494.8	323.8	418.4	282.3	304	148.7	46.6	1284.6

Table 2.4: Monthly and annual rainfall at station Chaskman, period 1968-1997 ordered in ascending order per column

By plotting the rainfall of a particular year with the frequency curves one has a proper means to assess how the rainfall in each month in that particular year behaved compared to the long term rainfall in that month. However, the say 10% curve should not be considered as a 10%-wet year. To show this in the last column of Table 2.4 the ranked annual values are presented as well. The rainfall in 10%-wet year amounts 478 mm, whereas the sum of the 10% monthly rainfall amounts add up to 189.8 mm only. Similar conclusions can be drawn for other crf's. This is shown in Figure 2.5 a, b, c.

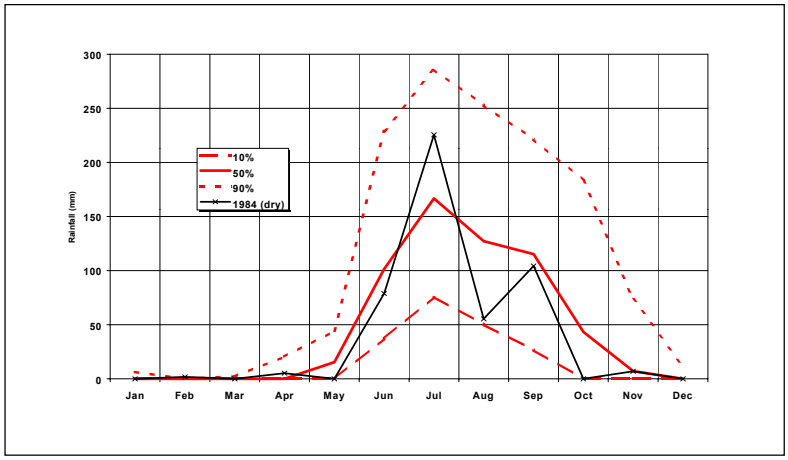


Figure 2.5a

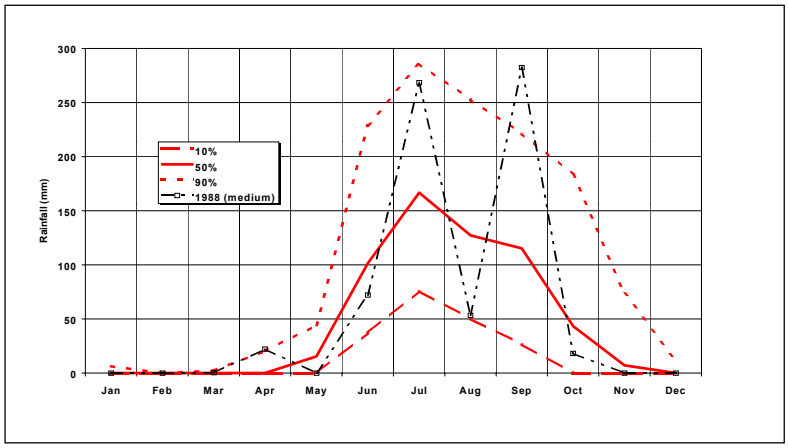


Figure 2.5b

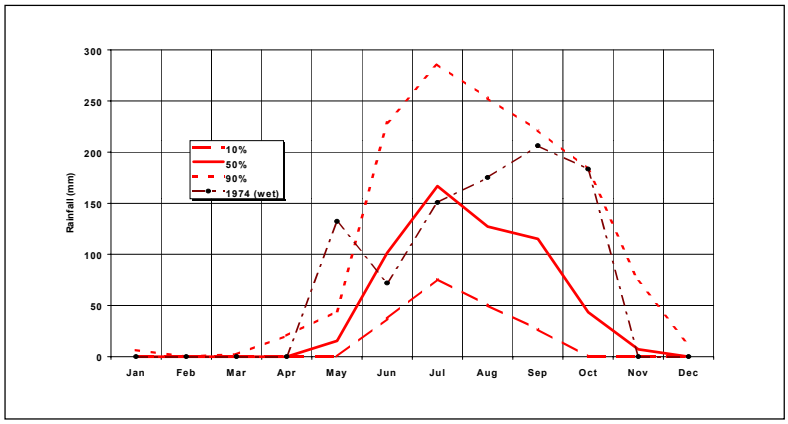


Figure 2.5c

Figure 2.5 a, b, c: Frequency curves of crf = 10, 50 and 90% with 10%, 50% and 90% wet year records.

In the above text frequency curves were discussed for monthly rainfall data. Basically, the technique can be applied to any hydrological variable and the interval may also be day, 10 days, season, etc. Generally, say, we have M observations in a year for N years. Let the observation on the hydrological variable in interval *m* in year *n* be denoted by $X_{m,n}$. Then for *n* = 1, N the X_m's are put in ascending order: $X_{m,k}$, where *k* is the rank of $X_{m,n}$, with *k* running from 1 to N. De crf attributed to $X_{m,n}$ is *k*/N (or *k*/(N+1) or some other estimate for the probability of non-exceedance as discussed earlier). By selecting a specific value for *k* = *k1* corresponding to a required crf the sequence of $X_{m,k1}$ for *m* = 1,M will give us the required frequency curve. In case a required crf, for which a frequency curve is to be made, does not correspond with the k^{th} rank in the sequence of N values, linear interpolation between surrounding values is to be applied.

Duration Curves

For the assessment of water resources, navigational depths, etc. it may be useful to draw duration curves. When dealing with flows in rivers, this type of graphs is known as a flow duration curve. It is in effect a cumulative frequency diagram with specific time scales. On the horizontal axis the percentage of time or the number of days/months per year or season during which the flow is not exceeded may be given. The volume of flow per day/month or flow intensity is given on the vertical axis. (The above convention is the display adopted in HYMOS; others interchange the horizontal and vertical axis.) Similarly, duration curves may be developed for any other type of variable. In Figure 2.6 the duration curve for the monthly rainfall at Chaskman for the period 1968-1997 is presented.

Figure 2.6 tells us that there is no rain during at least four months in a year, and **on average** there is only one month in a year with a monthly total larger than 200 mm. However, from Table 2.3 it can be observed that during 8 years out of 30 the 200 mm threshold was exceeded during two months. So the curve only displays average characteristics. The curve is obtained by multiplying the cumulative relative frequency associated with an observation with the number of intervals one has considered in a year or a season.

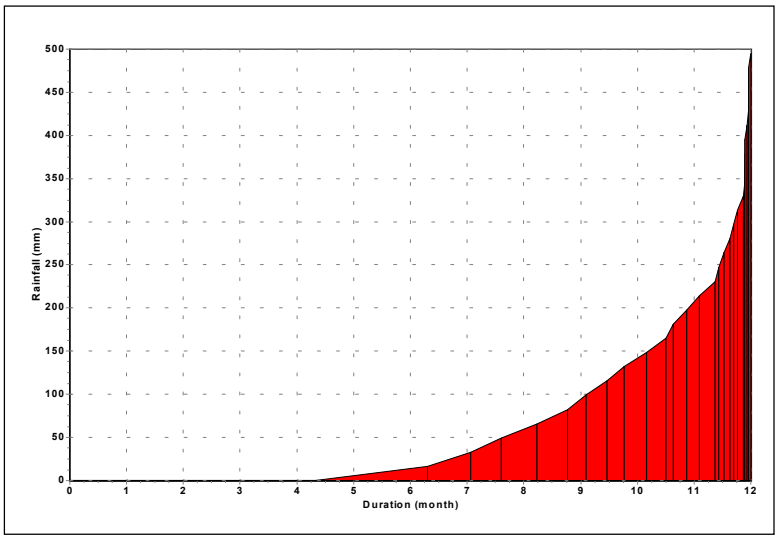


Figure 2.6:
Duration curve of monthly rainfall for station Chaskman

2.3 Measures of Central Tendency

Measures of the central tendency of a series of observations are:

- Mean
- Median
- Mode

Mean

The mean of a sample of size N is defined by

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (2.3)$$

where x_i = individual observed value in the sample

N = sample size i.e. total number of observed values

m = mean of the sample size n.

When dealing with catchment rainfall determined by Thiessen method, the mean is weighted according to the areas enclosed by bisectors around each station. The sum of the weights is 1:

$$m_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad (2.4)$$

Median

The median M of a sample is the middle value of the ranked sample, if N is odd. If N is even it is the average of the two middle values. The cumulative relative frequency of the median is 0.5. For a symmetrical distribution the mean and the median are similar. If the distribution is skewed to the right, then M < m, and when skewed to the left M > m.

Mode

The mode of a sample is the most frequently occurring value and hence corresponds with the value for which the distribution function is maximum. In Figure 2.2 the mode is in the class 60-120 mm and can be estimated as 90 mm.

2.4 Measures of Dispersion

Common measures of dispersion are:

- the range,
- the variance,
- the standard deviation, and
- coefficient of variation.

Range

The range of a sample is the difference between the largest and smallest sample value. Since the sample range is a function of only two of the N sample values it contains no information about the distribution of the data between the minimum and maximum value. The population range of a hydrological variable is in many cases, the interval from 0 to ∞, and as such displays no information about the process.

In hydrology the word ‘range’ is also used to quantify the range of accumulative departures from the mean (also indicated as partial sums). That value has important implications when dealing with water storage. It is a measure for the required storage when the average flow is to be drawn from a reservoir.

Variance

The most common measure of dispersion used in statistical analysis is the variance, estimated by s^2:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2 \qquad (2.5)$$

The reason for using the divisor N-1 instead of N is that it will result in an unbiased estimate for the variance. The units of the variance are the same as the units of x^2.

Standard deviation

The standard deviation s is the root of the variance and provides as such a measure for the dispersion of the data in the sample set in the same dimension as the sample data. It is estimated by:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2} \qquad (2.6)$$

Coefficient of Variation

A dimensionless measure of dispersion is the coefficient of variation C_v defined as the standard deviation divided by the mean:

$$C_v = \frac{s}{m} \tag{2.7}$$

Note that when m = 0 the coefficient of variation C_v becomes undefined; hence for normalised distributions this measure cannot be applied.

From Figure 1.3 it is observed that the coefficient of variation of the monthly rainfall at Chaskman is > 1 for the dry period, but < 1 during the monsoon.

2.5 Measure of Symmetry: Skewness

Distributions of hydrological variables are often skewed, i.e. non-symmetrical. The distributions are generally skewed to the right, like daily rainfall. By aggregation of data, the distribution of the aggregate will approach normality, i.e. will become symmetrical. Positively and negatively skewed distributions and symmetrical distributions are shown in Figure 2.7.

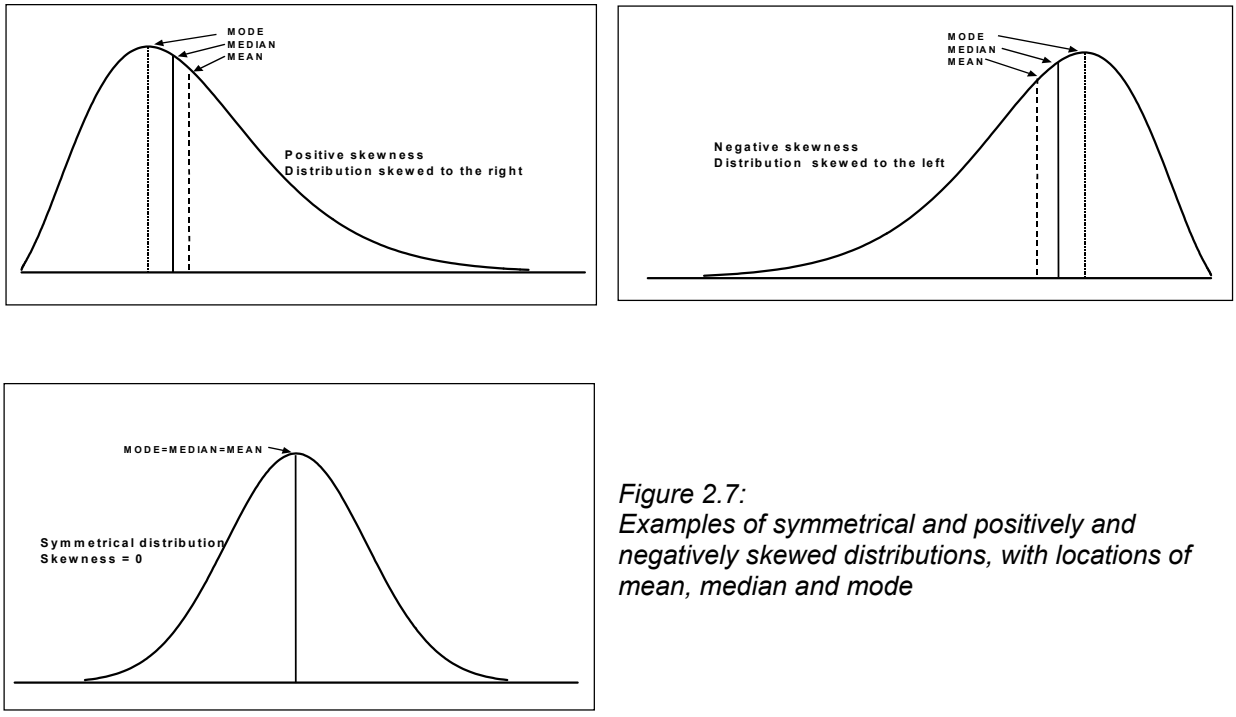


Figure 2.7:
Examples of symmetrical and positively and negatively skewed distributions, with locations of mean, median and mode

The skewness is derived from the third central moment of the distribution, scaled by the standard deviation to the power 3. An unbiased estimate for the coefficient of skewness can be obtained from the following expression:

$$g_1 = \frac{N}{(N-1)(N-2)} \frac{\sum_{i=1}^{N} (x_i - m)^3}{s^3} \tag{2.8}$$

In Figure 2.7 the relative position of the mean, median and mode for symmetrical and positively and negatively skewed distributions is presented.

2.6 Measure of Peakedness: Kurtosis

Kurtosis refers to the extent of peakedness or flatness of a probability distribution in comparison with the normal distribution, see Figure 2.8. The sample estimate for kurtosis is:

$$g_2 = \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \frac{\sum_{i=1}^{N}(x_i - m)^4}{s^4} \qquad (2.9)$$

The kurtosis is seen to be the 4th moment of the distribution about the mean, scaled by the 4th power of the standard deviation. The kurtosis for a normal distribution is 3. The normal distribution is said to be **mesokurtic**. If a distribution has a relatively greater concentration of probability near the mean than does the normal, the kurtosis will be greater than 3 and the distribution is said to be **leptokurtic**. If a distribution has a relatively smaller concentration of a probability near the mean than does the normal, the kurtosis will be less than 3 and the distribution is said to be **platykurtic**.

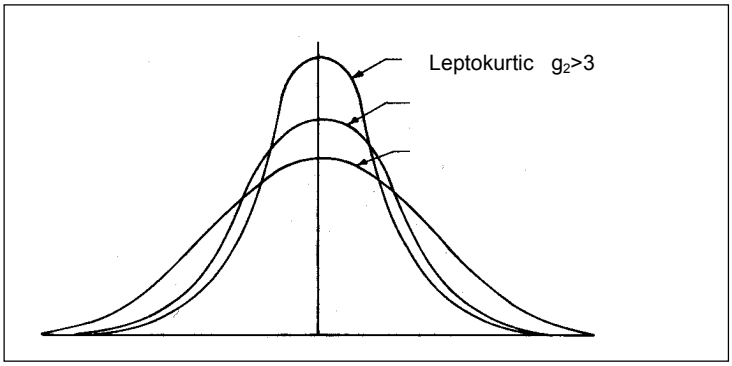


Figure 2.8:
Illustration of Kurtosis

The **coefficient of excess** e is defined as $g_2 - 3$. Therefore for a normal distribution e is 0, for a leptokurtic distribution e is positive and for a platykurtic distribution *e* is negative.

2.7 Quantiles, percentile, deciles and quartiles

The cumulative relative frequency axis of the cumulative relative frequency curve running from 0 to 1 or from 0 to 100% can be split into equal parts. Generally, if the division is in *n* equal parts, one will generate (n-1) **quantiles**. The pth quantile, x_p, is the value that is larger than 100p% of all data. When n = 100, i.e. the division is done in 100 equal parts (percents), then the value of the hydrological variable read from the x-axis corresponding with a crf of p% is called the p^{th} **percentile**. If the frequency axis is divided into 10 equal parts then the corresponding value on the x-axis is called a **decile**. Thus the 10th percentile (also called the first decile) would mean that 10% of the observed values are less than or equal to the percentile. Conversely, the 90th percentile (or 9th decile) would mean that 90% of the observed values are lying below that or 10% of the observed values are lying above that. The median would be the 50th percentile (or fifth decile). Similarly, if the frequency axis is divided in 4 equal parts then one speaks of **quartiles**. The first quartile corresponds with the 25th percentile, i.e. 25% of the values are less or equal than the first quartile; the second quartile is equal to the median and the third quartile equals the 75th percentile.

2.8 Box plot and box and whiskers plot

A **box plot** displays the three **quartiles** of a distribution in the form of a box, see Figure 2.9. If in addition also the **minimum and the maximum** values are displayed by bars extending the box on either side, the plot is called a **box and whiskers plot**. Sometimes also the mean is indicated in the plot. Hence the plot is a 5 or 6 points summary of the actual frequency distribution. Such plots are made for the data in a season or a year or any other selected time interval.

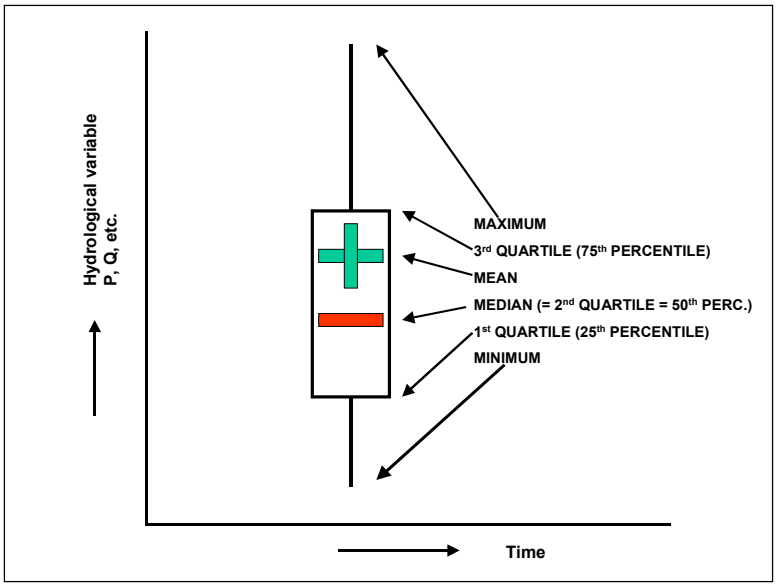


Figure 2.9: Features of a box and whiskers plot

By displaying the box and bars for successive years a quick insight is provided into the variation of the process from year to year. This form is very popular for displaying the behaviour of water quality variables. For that purpose the plot is extended with threshold values on a particular water quality variable.

In Figure 2.10 an example is given of a box and whiskers plot applied to discharge measurements at station Rakshewa in Bhima basin, where the statistics of the measurements from 1995 to 1998 are shown for each year separately.

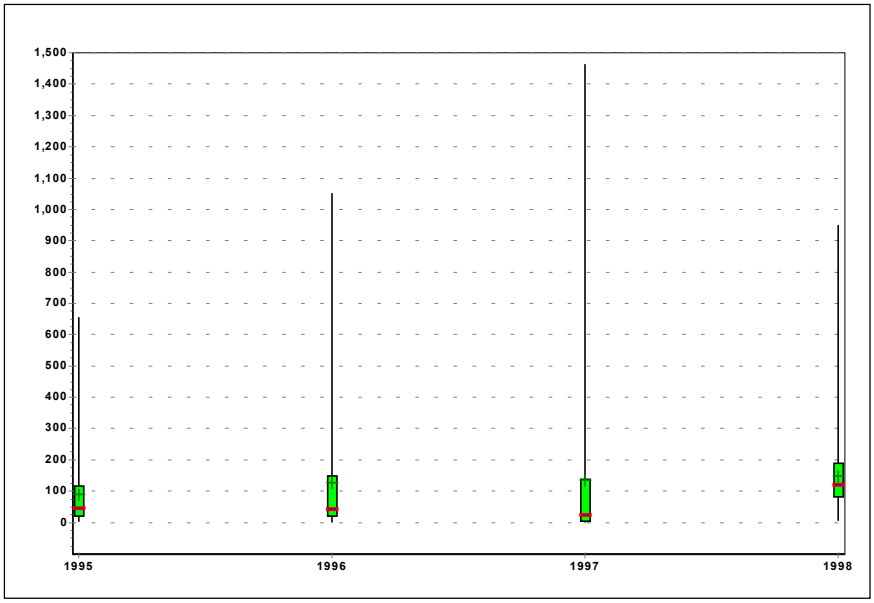


Figure 2.10: Box and whiskers plot of discharge measurements at station Rakshewa in Bhima basin, period 1995 – 1998.

It is clearly observed from the boxes and bars in Figure 2.10 that the distribution of the measured discharges in a year is skewed to the right. Generally, a large number of discharge measurements are available for the very low stages and only a few for the higher stages. Hence the extent of the box (which comprises 50% of the measurements) is very small compared to the range of the data. The mean is seen to be always larger than the median.

2.9 Covariance and Correlation Coefficient

When simultaneous observations on hydrological variables are available then one may be interested in the linear association between the variables. This is expressed by the covariance and correlation coefficient.

If there are N pairs of observations (x_1, y_1), (x_2, y_2), …, (x_N, y_N), of two variables X and Y, the sample covariance is obtained from the following expression:

$$\hat{C}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y) \qquad (2.10)$$

where: m_X, m_Y = sample means of X and Y respectively:

The correlation coefficient r_{XY} is obtained by scaling the covariance by the standard deviations of X and Y:

$$r_{XY} = \frac{\hat{C}_{XY}}{s_X s_Y} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y)}{s_X s_Y} \qquad (2.11)$$

where: s_X , s_Y = sample standard deviations of X and Y.

To get the limits of r_{XY} consider the case that X and Y have a **perfect** linear correlation. Then the relationship between X and Y is given by :

$Y = a + bX$

and hence:

$m_Y = a + bm_X$ and: $s_Y^2 = b^2 s_X^2$ or: $s_Y = |b| s_X$

Substituting above relations in (2.11) gives:

$$r_{XY} = \frac{\frac{1}{N-1}\sum(x_i - m_X)(a + bx_i - (a + bm_X))}{s_X |b| s_X} = \frac{b}{|b|}\frac{\frac{1}{N-1}\sum(x_i - m_X)^2}{s_X^2} = \frac{b}{|b|} \qquad (2.12)$$

If Y increases for increasing X, i.e. they are positively correlated, then $b > 0$ and r_{XY} is seen to be 1. If on the other hand Y decreases when X is increasing, they are negatively correlated; then $b < 0$ and r_{XY} is –1. So r_{XY} is seen to vary between ± 1:

$-1 \le r_{XY} \le 1$.

If there is no linear association between X and Y then r_{XY} is 0. If r_{XY} is 0 it does not mean that X and Y are independent or that there is no association between X and Y. It only means that the linear association is not existing. Still, there may be for example a circular association.

A convenient means to investigate the existence of linear association is by making a XY-scatter plot of the samples. Typical examples of scatter plots are shown in Figure 2.11.

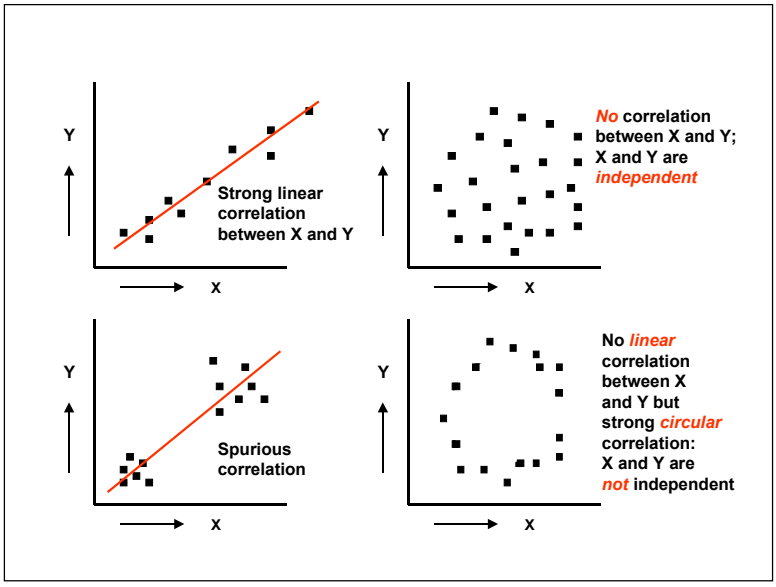


Figure 2.11:
Examples of scatter plots

In some cases the scatter plot may indicate a non-linear type of relationship between the two variables. In such cases some transformation, e.g. a logarithmic, square root, negative reciprocal, or other appropriate transformation to one or both variables may be applied before analysis.

Spurious correlation

The lower left plot in Figure 2.11 gives an example of **spurious** correlation, which is easily obtained in hydrology, when blindly data are being compared. For example if there is a distinct wet and dry period and the discharges of two sites in different regions, but both subjected to monsoonal variation, are plotted in an XY-plot, a situation like the one displayed will occur. In the wet period the data at X and Y may be completely uncorrelated, but simply by the fact of the existence of a dry and wet period, which clusters observations in the low and the high regions, the correlation is seemingly very high. This effect is due to the acceptance of **heterogeneous** data, see also Figure 1.2 and 1.3. By taking the low and high flow values in the same data set, the overall mean value for X and Y will be somewhere between the low and the high values. Hence entries in the wet period on either side will be positive relative to the mean and so will be their products. In the same way, entries in the dry period will both be negative relative to the mean, so their product will be positive as well, ending up into a large positive correlation.

Similarly, wrong conclusions can be drawn by comparing data having the same denominator. If X, Y and Z are **uncorrelated** and X/Z and Y/Z are subjected to correlation analysis, a non-zero correlation will be found (see e.g. Yevjevich, (1972)):

$$r = \frac{C_{v,Z}^2}{\left(C_{v,X}^2 + C_{v,Z}^2\right)^{1/2}\left(C_{v,Y}^2 + C_{v,Z}^2\right)^{1/2}} \qquad (2.13)$$

From (2.13) it is observed, that when all coefficients of variation are equal, it follows that r = 0.5!!!

It indicates that one has to select the sample sets to be subjected to correlation and regression analysis carefully. Common divisors should be avoided. Also, the direction of analysis as indicated in Figure 2.2 is of utmost importance to ensure homogeneous data sets.

3 Fundamental Concepts of Probability

3.1 Axioms and Theorems

Sample and Space Events

The **sample space** denoted by Ω, is defined here as the collection of all possible outcomes of sampling on a hydrological variable.

An **event** is a collection of sample points in the sample space Ω of an experiment. An event can consists of a single sample point called a simple or elementary event, or it can be made up of two or more sample points known as a compound event. An event is (denoted by a capital letter A (or any other letter)) is thus a subset of sample space Ω.

The Null Event, Intersection and Union

Two events A_1 and A_2 are **mutually exclusive** or **disjoint** if the occurrence of A_1 excludes A_2, i.e. none of the points contained in A_1 is contained in A_2. and vice versa.

The **intersection** of the events A_1 and A_2 is that part of the sample space they have in common. This is denoted by $A_1 \cap A_2$ or A_1A_2.

If A_1 and A_2 are mutually exclusive then their intersection constitutes a **null event**: $A_1 \cap A_2 = A_1A_2 = \varnothing$.

The **union** of two events A_1 and A_2 represents their joint occurrences, and it comprises the event containing the entire sample in A_1 and A_2. This is denoted by $A_1 \cup A_2$, or simply $A_1 + A_2$. With the latter notation one has to be careful as the sum of the two has to be corrected for the space in common (i.e. the intersection.

The intersection is equivalent to the "**and**" logical statement, whereas the union equivalent to "**and/or**".

The above definitions have been visualised in Figure 3.1 by means of **Venn diagrams**.

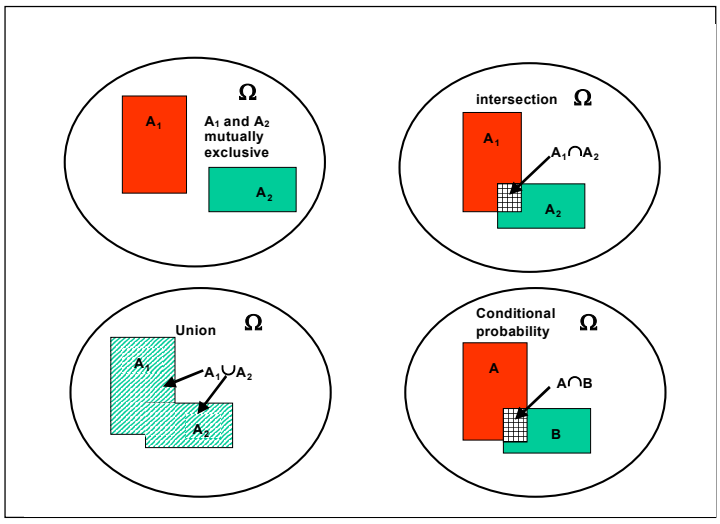


Figure 3.1:
Definition sketch by Venn diagrams

The definitions are illustrated in the following example:

Example 3.1 Events in a sample space.

Sample space and events representing rainy days (i) and total rainfall (p) at a rainfall station during the period 1-10 July are given in Figure 3.2:

The sample space reads: $\Omega \equiv \{(i,p): i = 0, 1, 2,\ldots,10; \text{ and } 0 \leq p\}$

Event $A_1 \equiv \{(i,p): i > 3, \text{ and } p > 50\}$

Event $A_2 \equiv \{(i,p): 3 \leq i < 5, \text{ and } p > 20\}$

Event $A_3 \equiv \{(i,p): 1 \leq i < 3, \text{ and } 2 \leq p < 30\}$

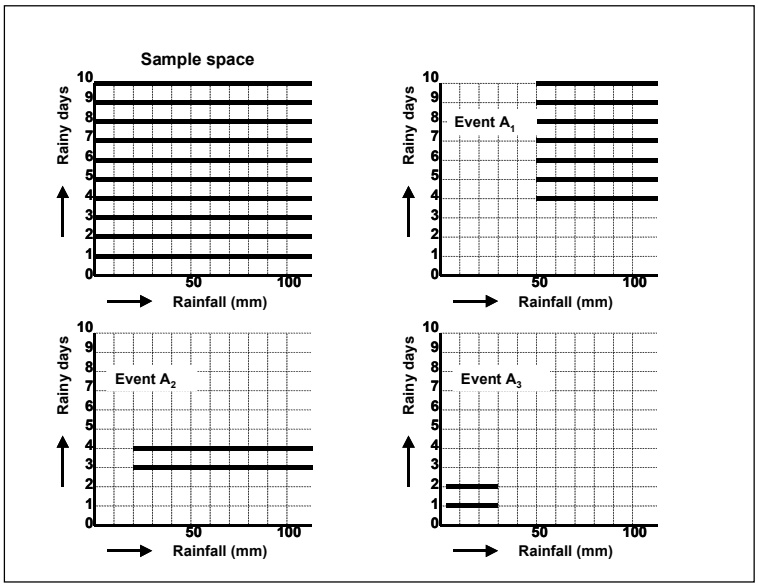


Figure 3.2:
Presentation of sample space Ω events A_1, A_2 and A_3

The union and intersection of A_1 and A_2 and of A_2 and A_3 are presented in Figure 3.3.

Event $A_1 + A_2 \equiv \{(i,p): 3 \leq i < 5, \text{ and } p > 20; i \geq 5, \text{ and } p > 50\}$

Event $A_1A_2 \equiv \{(i,p): i = 4 \text{ and } p > 50\}$

Event $A_2 + A_3 \equiv \{(i,p): 1 \leq i < 3, \text{ and } 2 \leq p < 30; 3 \leq i < 5, \text{ and } p > 20\}$

Event $A_2A_3 = \varnothing$, since A_2 and A_3 are disjoint, having no points in common.

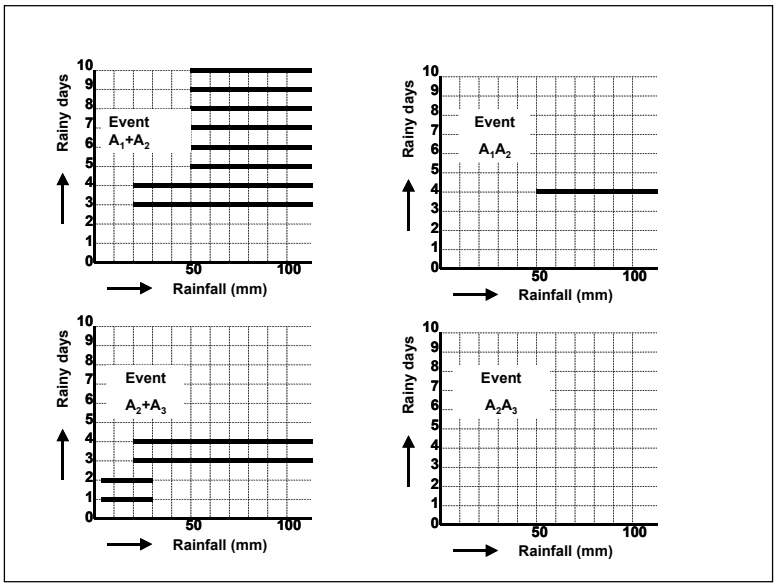


Figure 3.3:
Unions and intersections of A_1 and A_2 and of A_2 and A_3

Probability axioms and theorems

Using these definitions the following axioms and theorems are discussed dealing with the probability of an event or several events in the sample space.

Definition of probability

If a random events occurs a large number of times N, of which N_A times the event A happens, then the probability of the occurrence of event A is:

$$P(A) = \lim_{N\to\infty} \frac{N_A}{N} \qquad (3.1)$$

Hence, if A is any event in a sample space Ω, then:

$$0 \le P(A) \le 1 \qquad (3.2)$$

The event in the sample space not contained in A is the complement of A, denoted by A^C:

$$P(A^C) = 1 - P(A) \qquad (3.3)$$

If A is a certain event then:

$$P(A) = 1 \qquad (3.4)$$

Probability of the union of events

For any set of arbitrary events A_1 and A_2 the probability of the **union** of the events, i.e. the probability of event A_1 **and/or** A_2 is:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad (3.5)$$

The last term is the intersection of A_1 and A_2, i.e. the part in the sample space they have in common. So, if A_1 and A_2 have no outcomes in common, i.e if they are **mutually exclusive**, then the intersection of the two events is a null event and then (3.5) reduces to:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \tag{3.6}$$

For three joint events it generally follows:

$$P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1\,A_2) - P(A_1\,A_3) - P(A_2\,A_3) + P(A_1\,A_2\,A_3) \tag{3.7}$$

For any set of arbitrary events $A_1, A_2, \ldots, A_m$ the probability of the union becomes a complicated expression, (see e.g. Suhir, 1997), but if the events $A_1, A_2, \ldots, A_m$ have no outcomes or elements in common, i.e. if they are mutually exclusive, then the union of the events have the probability:

$$P\left(\sum_{j=1}^{m} A_j\right) = \sum_{j=1}^{m} P(A_j) \tag{3.8}$$

Hence, the probability of the intersection is seen to have vanished as it constitutes a null event for mutually exclusive events.

Conditional probability

The **conditional** probability P(B|A) gives the probability of event B given that A has occurred. Here A serves as a new (reduced) sample space (see Figure 3.1) and P(B|A) is that fraction of P(A) which corresponds to $A \cap B$, hence:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{3.9}$$

Denoting $P(A \cap B) \equiv P(AB)$ it follows:

$$P(AB) = P(B|A) \,.\, P(A) \tag{3.10}$$

Independence

If A and B are **independent** events, i.e. the occurrence of B is not affected by the occurrence of A, then:

$$P(B|A) = P(B) \tag{3.11}$$

and hence:

$$P(AB) = P(B) \,.\, P(A) \tag{3.12}$$

It states that if the events A and B are independent, the probability of the occurrence of event A **and** B equals the product of the **marginal** probabilities of the individual events.

Total probability

Consider an event B in Ω with $P(B) \neq 0$ and the **mutually exclusive** events A_1, A_2, …, A_m, which are **collectively exhaustive**, i.e. $A_1 + A_2 + \ldots + A_m = \Omega$. Then the events BA_1, BA_2, …, BA_m are also mutually exclusive and $BA_1 + BA_2 + \ldots + BA_m = B(A_1 + A_2 + \ldots + A_m) = B\Omega = B$. Hence:

$$P(B) = \sum_{j=1}^{m} P(BA_j) = \sum_{j=1}^{m} P(B \mid A_j).P(A_j) \qquad (3.13)$$

This is called the theorem of **total probability**, which is visualised in Figure 3.4.

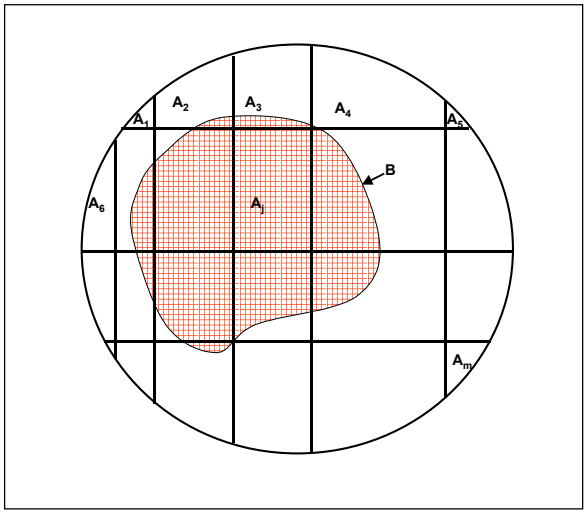


Figure 3.4:
Concept of total probability

Bayes theorem

Observe now the following conditional probability:

$$P(A_i \mid B) = \frac{P(BA_i)}{P(B)}$$

The numerator reads according to (3.10) $P(BA_i) = P(B|A_i) \cdot P(A_i)$. The denominator is given by (3.13). It then follows for $P(A_i|B)$, Bayes rule:

$$P(A_i \mid B) = \frac{P(B \mid A_i).P(A_i)}{\sum_{j=1}^{m} P(B \mid A_j).P(A_j)} \qquad (3.14)$$

Bayes rule provides a method to update the probabilities about the true state of a system (A), by sampling (B) in stages. The probabilities $P(A_i)$'s on the right hand side of (3.14) are the probabilities about the state of the system before the sample is taken (**prior** probabilities). After each sampling the **prior** probabilities $P(A_i)$'s are updated, by replacing them with the **posterior** probability (= left hand side of the equation), found through the outcome of the sampling: B. The conditional probabilities $P(B|A_j)$ represent basically the quality of the sampling method or equipment: the probability of getting a particular sample B given that the true state of the system is A_i. Bayes rule can therefore be interpreted as follows:

$$P(\text{state}_{\text{posterior}} \mid \text{sample}) = \frac{P(\text{sample} \mid \text{state}).P(\text{state}_{\text{prior}})}{\sum_{\text{all states}} P(\text{sample} \mid \text{state}).P(\text{state}_{\text{prior}})} \qquad (3.15)$$

To illustrate the above axioms and theorems the following examples are given.

Example 3.2 Annual monthly maximum rainfall

The annual monthly maximum rainfall for station Chaskman is presented in Table 3.2 and Figure 3.5.

Year	P_{max} (mm)	Year	P_{max} (mm)	Year	P_{max} (mm)
1968	162.1	1978	154.4	1988	282.3
1969	320.2	1979	252.8	1989	227.6
1970	162.6	1980	325.8	1990	212.0
1971	212.6	1981	258.0	1991	404.6
1972	229.7	1982	144.0	1992	235.2
1973	312.6	1983	418.4	1993	304.0
1974	206.2	1984	225.5	1994	285.8
1975	191.8	1985	105.7	1995	262.0
1976	494.8	1986	229.0	1996	221.2
1977	207.0	1987	148.0	1997	342.6

Table 3.1:
Annual monthly maximum rainfall for Chaksman, period 1968-1997

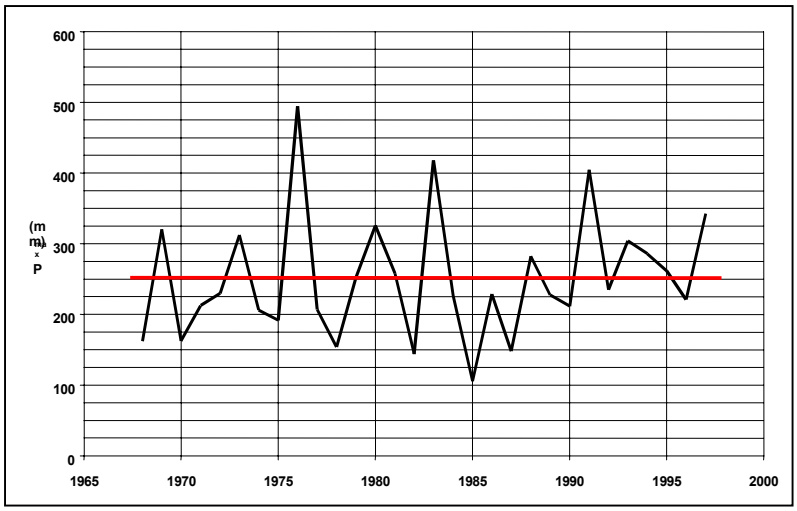


Figure 3.5:
Annual monthly maximum rainfall for Chaksman, period 1968-1997

From the table and figure it is observed that a monthly maximum > 260 mm has occurred 11 times in a period of 30 years, hence P_{max} > 260 mm = 11/30 = 0.367 in any one year. Assuming that the elements of the annual monthly maximum series are independent, it follows that the probability of having two annual maximum values in sequence > 260 mm = 0.367 x 0.367 = 0.135. From the series one observes that this event happened only 2 times in 30 years, that is 2 out of 29, i.e. having a probability of 2/29 = 0.069. If event A is the occurrence that P_{max} > 260 mm and B is the event that P_{max} > 260 mm in a second successive year then: $P(B|A) = P(A \cap B)/P(A)=(2/29)/(11/30) = 0.19$.

Example 3.3 Daily rainfall Balasinor (Gujarat)

Based on daily rainfall data of station Balasinor for the month of July in the period 1961 to 1970, the following probabilities have been determined:

Probability of a rainy day following a rainy day = 0.34

Probability of a rainy day following a dry day = 0.17

Probability of a dry day following a rainy day = 0.16

Probability of a dry day following a dry day = 0.33

Given that a particular day is dry, what is the probability the next two days are (1) dry and (2) wet?

(1) Call event A = dry day 1 after a dry day and event B = dry day 2 after a dry day. Hence required is $P(A \cap B) = P(B|A) \cdot P(A)$. The probability of having a dry day after a dry day is P(A) = 0.33 and the probability of a dry day given that the previous day was dry P(B|A) = 0.33. So, $P(A \cap B) = P(B|A) \cdot P(A) = 0.33 \cdot 0.33 = 0.11$.

(2) Call event A = wet day 1 after a dry day and event B = wet day 2 after a dry day. Now we require again $P(A \cap B) = P(B|A) \cdot P(A)$. The probability of a wet day after a dry day is P(A) = 0.17 and the probability of a wet day given that the previous day was also wet = P(B|A) = 0.34. Hence, $P(A \cap B) = P(B|A). P(A) = 0.34 \cdot 0.17 = 0.06$. This probability is seen to be about half the probability of having two dry days in a row after a dry day. This is due to the fact that for Balasinor the probability of having a wet day followed by a dry day or vice versa is about half the probability of having two wet or two dry days sequentially.

Example 3.4 Prior and posterior probabilities, using Bayes rule

In a basin for a considerable period of time rainfall was measured using a dense network. Based on these values for the month July the following classification is used for the basin rainfall.

Class	Rainfall (mm)	Probability
Dry	$P < 50$	$P[A_1] = 0.15$
Moderate	$50 \le P < 200$	$P[A_2] = 0.50$
Wet	$200 \le P < 400$	$P[A_3] = 0.30$
Extremely wet	$P \ge 400$	$P[A_4] = 0.05$

Table 3.2: Rainfall classes and probability.

The probabilities presented in Table 3.2 refer to prior probabilities. Furthermore, from the historical record it has been deduced that the percentage of gauges, which gave a rainfall amount in a certain class given that the basin rainfall felt in a certain class is given in Table 3.3.

Basin rainfall	Percentage of gauges			
	P < 50	50 ≤ P < 200	200 ≤ P < 400	P ≥ 400
P < 50	80	15	5	0
50 ≤ P < 200	25	65	8	2
200 ≤ P < 400	5	20	60	15
P ≥ 400	0	10	25	65

Table 3.3: Conditional probabilities for gauge value given the basin rainfall

Note that the conditional probabilities in the **rows** add up to 100%.

For a particular year a gauge gives a rainfall amount for July of 230 mm. Given that sample value of 230 mm, what is the class of the basin rainfall in July for that year.

Note that the point rainfall falls in class III. The posterior probability of the actual basin rainfall in July of that year becomes:

$$P[A_i \mid \text{sample } 1] = \frac{P[\text{sample } 1 \mid A_i].P[A_i]}{\sum_{i=1}^{4} P[\text{sample } 1 \mid A_i].P[A_i]}$$

The denominator becomes:

$$\sum_{i=1}^{4} [\text{sample1} \mid A_i] = 0.05x0.15 + 0.20x0.50 + 0.60x0.30 + 0.15x0.05 = 0.295$$

The denominator expresses the probability of getting sample 1 when the prior probabilities are as given in Table 3.2, which is of course very low.

Hence,

$$P[A_1 \mid sample1] = \frac{0.05x0.15}{0.295} = 0.025$$
$$P[A_2 \mid sample1] = \frac{0.20x0.50}{0.295} = 0.340$$
$$P[A_3 \mid sample1] = \frac{0.60x0.30}{0.295} = 0.610$$
$$P[A_4 \mid sample1] = \frac{0.15x0.05}{0.295} = 0.025$$

Note that the sum of posterior probabilities adds up to 1.

Now, for the same month in the same year from another gauge a rainfall of 280 mm is obtained. Based on this second sample the posterior probability of the actual July basin rainfall in that particular year can be obtained by using the above posterior probabilities as revised prior probabilities for the July rainfall:

$$\sum_{i=1}^{4}[sample\ 2 \mid A_i] = 0.05x0.025 + 0.20x0.340 + 0.60x0.610 + 0.15x0.025 = 0.439$$

$$P[A_1 \mid sample2] = \frac{0.05x0.025}{0.439} = 0.003$$
$$P[A_2 \mid sample2] = \frac{0.20x0.340}{0.439} = 0.155$$
$$P[A_3 \mid sample2] = \frac{0.60x0.610}{0.439} = 0.834$$
$$P[A_4 \mid sample2] = \frac{0.15x0.025}{0.439} = 0.008$$

Note that the denominator has increased from 0.240 to 0.478.

Again note that the posterior probabilities add up to 1. From the above it is seen how the probability on the state of July rainfall changes with the two sample values:

Class	Prior probability	After sample 1	After sample 2
I	0.15	0.025	0.003
II	0.50	0.340	0.155
III	0.30	0.610	0.834
IV	0.05	0.025	0.008

Table 3.4: Updating of state probabilities by sampling

Given the two samples, the probability that the rainfall in July for that year is of class III has increased from 0.30 to 0.834.

Question: What will be the change in the last column of Table 3.4 if the third sample gives a value of 180 mm?

3.2 Frequency distributions

3.2.1 Univariate distributions

Discrete random variables

Formally, given a data set x_1, x_2, …., x_N of a stochatic variable X, the **probability mass function** (**pmf**) $p_X(x)$ expresses:

$$p_X(x) = P(X = x) \tag{3.16}$$

and the **cumulative distribution function** (**cdf**) $F_X(x)$ gives the probability of occurrence $X \leq x$:

$$F_X(X) = P(X \leq x) = \sum_{all x_i \leq x} p x(x_i) \quad \text{and} \quad \sum_{all x_i} p x(x_i) = 1 \tag{3.17}$$

Continuous random variables

In terms of continuous random variables, the continuous equivalent of the pmf is the **probability density function** (**pdf**), $f_X(x)$. The probability that X takes on values in the interval (x, x + dx) then reads $f_X(x).dx$:

$$f_X(x).dx = P(x \leq X < x + dx) \tag{3.18}$$

The **cumulative probability density function** (**cdf**) $F_X(x)$ is now defined as:

$$F_x(x) = P(X \leq x) = \int_{-\infty}^{x} f_x(y)dy \quad \text{and} \quad \int_{-\infty}^{\infty} f_x(y)dy = 1 \tag{3.19}$$

The functions are displayed in Figure 3.6.

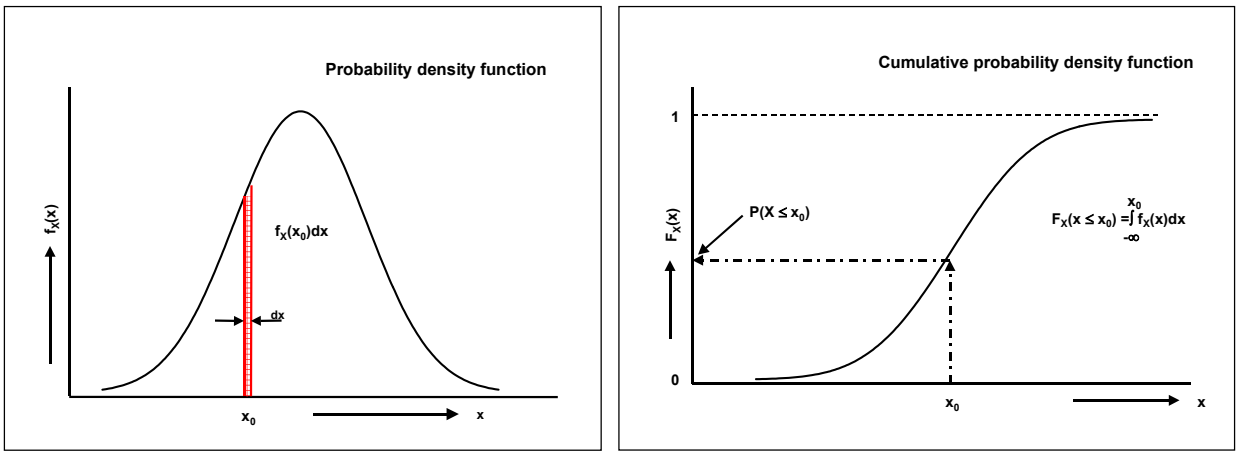


Figure 3.6: Probability density cumulative probability density function

$F_X(x)$ has the following properties:

- $F_X(-\infty) = 0$
- If $x_1 < x_2$ then $F_X(x_1) < F_X(x_2)$ ($F_X(x)$ is monotonous increasing)

- $\lim F_X(x + h) = F_X(x)$ for $h \downarrow 0$ ($F_X(x)$ is right continuous)

For the pdf it follows:

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{3.20}$$

Example 3.5 Exponential pdf and cdf

The exponential pdf reads:

$f_X(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$

Hence, the exponential cdf becomes with (3.19):

$$F_X(x) = \int_0^x \lambda\exp(-\lambda z)dz = -\exp(-\lambda z)\Big|_0^x = 1 - \exp(-\lambda x)$$

The exponential pdf and cdf for $\lambda = 0.2$ is shown in Figure 3.7. For example ($P(X \leq 7) = F_x(7) = 1 - \exp(-0.2 \times 7) = 0.75$ as shown in Figure 3.7.

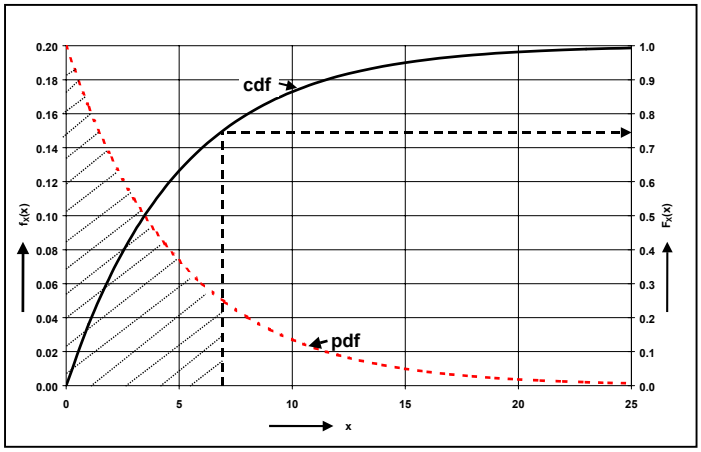


Figure 3.7:
Exponential pdf and cdf for λ=0.2

3.2.2 Features of distributions

In Chapter 2 some features of relative distribution functions were discussed. Here in a similar fashion this will be done for the pdf and the cdf. The following features of distributions are discussed:

- parameters
- return period
- mathematical expectation
- moments

Parameters

The distribution functions commonly used in hydrology are not specified uniquely by the functional form; the parameters together with the functional form describe the distribution. The parameters determine the **location**, **scale** and **shape** of the distribution.

Return period

The cdf gives the non-exceedance probability $P(X \leq x)$. Hence, the exceedance probability follows from: $P(X > x) = 1 - F_X(x)$ is. Its reciproke is called the return period. So if T is the return period and x_T is its corresponding quantile, then:

$$T = \frac{1}{P(X > x_T)} = \frac{1}{1 - P(X \leq x_T)} = \frac{1}{1 - F_X(x_T)} \tag{3.21}$$

Note that in the above the notation for the quantile x_T or x(T) is used. Others use the notation x_p for quantile where $p = F_x(x_p)$, i.e. non-exceedance probability.

Mathematical expectation

If X is any continuous random variable with pdf $f_X(x)$, and if g(X) is any real-valued function, defined for all real x for which $f_X(x)$ is not zero, then the **mathematical expectation** of the function g(X) reads:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx \tag{3.22}$$

Moments

If one chooses $g(X) = X^k$, where k = 1, 2, …. Then the k^{th} moment of X **about the origin** is defined by:

$$\mu_k' = E[X^k] = \int_{-\infty}^{+\infty} x^k f_X(x) dx \tag{3.23}$$

Note that an (') is used to indicate moments about the origin. Of special interest is the first moment about the origin, i.e. the mean:

$$\mu_1' = \mu_x = E[X] = \int_{-\infty}^{+\infty} x f_x(x) dx \tag{3.24}$$

If instead of the origin, the moment is taken around the mean, then the central moment follows (μ_k). Note that the accent (') is omitted here to denote a central moment. The second central moment is the variance:

$$\mu_2 = \mathrm{Var}(X) = E[(X - E[X])^2] = E[(X - \mu_X)^2] = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f_X(x) dx \tag{3.25}$$

With the above one defines:

- the standard deviation σ_X, which expresses the spread around the mean in the same dimension as the original variate:

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\mu_2} \tag{3.26}$$

- the coefficient of variation C_v:

$$C_v = \frac{\sqrt{\mu_2}}{\mu_1'} = \frac{\sigma_X}{\mu_X} \tag{3.27}$$

- the skewness coefficient $\gamma_{1,x}$ of the distribution is defined by:

$$\gamma_{1,X} = \frac{\mu_3}{\sigma_X^3} = \frac{1}{\sigma_X^3} \int_{-\infty}^{+\infty} (x - \mu_X)^3 f_X(x) dx \tag{3.28}$$

- the peakedness of the distribution, expressed by the kurtosis $\gamma_{2,X}$:

$$\gamma_{2,X} = \frac{\mu_4}{\sigma_X^4} = \frac{1}{\sigma_X^4}\int_{-\infty}^{+\infty}(x-\mu_X)^4 f_X(x)dx \qquad (3.29)$$

The parameter μ_X is a **location** parameter, σ_X a **scale** parameter, while $\gamma_{1,X}$ and $\gamma_{2,X}$ are **shape** parameters. The central moments μ_k are related to the moments about the origin μ_k' as follows:

$$\begin{aligned}\mu_1 &= 0\\ \mu_2 &= \mu_2' - (\mu_1')^2\\ \mu_3 &= \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3\\ \mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4\end{aligned} \qquad (3.30)$$

Example 3.7 Moments of the exponential distribution

Since the exponential pdf reads:

$f_X(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$

its first moment about the origin is:

$$\mu_1' = \mu_X = \int_0^\infty x\lambda\exp(-\lambda x)dx = \lambda\left\{\frac{\exp(-\lambda x)}{-\lambda}\left(x+\frac{1}{\lambda}\right)\right\}\Bigg|_0^\infty = \lambda\left\{0 - \frac{1}{-\lambda}\left(0+\frac{1}{\lambda}\right)\right\} = \frac{1}{\lambda}$$

It shows that the parameter λ is the reciproke of the mean value. The exponential distribution is well suited to model inter-arrival times, for example of flood occurrences. Then x has the dimension of time, and λ 1/time. If a flood of say 1,000 m^3/s is on average exceeded once every 5 years, and the exponential distribution applies, then μ_X = 5 years and hence λ = 1/5 = 0.2.

In extension to the above derivation, one can easily show, that the k^{th} order moments about the origin of the exponential distribution read:

$$\mu_k' = \frac{k!}{\lambda^k} \text{ hence : } \mu_1' = \frac{1}{\lambda};\ \mu_2' = \frac{2}{\lambda^2};\ \mu_3' = \frac{6}{\lambda^3};\ \mu_4' = \frac{24}{\lambda^4}$$

Then from (3.30) it follows for the central moments:

$$\mu_2 = \sigma_X^2 = \mu_2' - (\mu_1')^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

$$\mu_3 = \mu_3' - 3\mu_2'.\mu_1' + 2(\mu_1')^3 = \frac{6}{\lambda^3} - 3\frac{2}{\lambda^2}.\frac{1}{\lambda} + 2\left(\frac{1}{\lambda}\right)^3 = \frac{2}{\lambda^3}$$

$$\mu_4 = \mu_4' - 4\mu_3'.\mu_1' + 6\mu_2'.(\mu_1')^2 - 3(\mu_1')^4 = \frac{24}{\lambda^4} - 4\frac{6}{\lambda^3}.\frac{1}{\lambda} + 6\frac{2}{\lambda^2}.\left(\frac{1}{\lambda}\right)^2 - 3\left(\frac{1}{\lambda}\right)^4 = \frac{9}{\lambda^4}$$

And for the standard deviation, skewness and kutosis with (3.26), (3.28) and (3.29):

$$\sigma_X = \frac{1}{\lambda}$$

$$\gamma_{1,X} = \frac{\mu_3}{\sigma_X^3} = \frac{2/\lambda^3}{1/\lambda^3} = 2$$

$$\gamma_{2,X} = \frac{\mu_4}{\sigma_X^4} = \frac{9/\lambda^4}{1/\lambda^4} = 9$$

It is observed from the above that for the exponential distribution the mean and the standard deviation are the same. The distribution has a fixed positive skewness and a kurtosis of 9, which implies that the probability density of an exponential distribution is more closely concentrated around the mean than for a normal distribution.

3.2.3 Multivariate distribution functions

Occasionally, statistics about the joint occurrence of stochastic variables is of concern. In this subsection we discuss:

- Joint cdf and pdf
- Marginal cdf and pdf
- Conditional distribution function
- Moments
- Covariance and correlation

Joint distributions

The probability of joint events (i.e. intersections in the sample space) is given by the joint k-dimensional cdf $F_{X1,\ X2,\ \ldots,\ Xk}(x_1, x_2, \ldots, x_k)$.

In case of two stochastic variables X and Y the joint 2-dimensional cdf $F_{XY}(x,y)$ reads:

$$F_{XY}(x,y) = P(X \le x \cap Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(s,t)\,ds\,dt \tag{3.31}$$

where $f_{XY}(x,y)$ is the **joint 2-dimensional pdf**:

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} \tag{3.32}$$

Marginal distributions

The **marginal cdf** $F_X(x)$ of X only, gives the non-exceedance probability of X irrespective of the value of Y, hence

$$F_X(x) = P(X \le x \cap -\infty < Y < \infty) = F_{XY}(x,\infty) \tag{3.33}$$

and similarly the **marginal pdf** $f_X(x)$ reads:

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx}F_{XY}(x,\infty) = \int_{-\infty}^{\infty} f_{XY}(x,t)\,dt \tag{3.34}$$

Conditional distribution

Analogous to (3.5) the **conditional distribution function** can be defined:

$$F_{X|Y}(x,y) = P(X \le x \mid Y \le y) = \frac{F_{XY}(x,y)}{F_Y(y)} \tag{3.35}$$

and the conditional pdf:

$$f_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{3.36}$$

Independent variables

Equivalently to (3.8), if X and Y are **independent** stochastic variables, the distribution function can be written as:

$$F_{XY} = P(X \le x \cap Y \le y) = P(X \le x).P(Y \le y) = F_X(x).F_Y(y) \tag{3.37}$$

and similarly for the density function:

$$f_{XY}(x,y) = f_X(x).f_Y(y) \tag{3.38}$$

Moments

In addition to the moments for univariate distributions the moments for bivariate distributions are defined as follows:

$$\mu'_{k,m} = E\left[X^k Y^m\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m f_{XY}(x,y)dxdy \tag{3.39}$$

Covariance and correlation function

Of special interest is the central moment expressing the linear dependency between X and Y, i.e. the **covariance**:

$$C_{XY} = E\left[(X - E[X])(Y - E[Y])\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{XY}(x,y)dxdy \tag{3.40}$$

Note that if X is independent of Y, then with (3.38) it follows:

$$C_{XY} = \int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx \int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy = 0 \tag{3.41}$$

As discussed in Chapter 2, a standardised representation of the covariance is given by the correlation coefficient ρ_{XY}:

$$\rho_{XY} = \frac{C_{XY}}{\sqrt{C_{XX}C_{YY}}} = \frac{C_{XY}}{\sigma_X\sigma_Y} \tag{3.42}$$

In Chapter 2 it was shown that ρ_{XY} varies between +1 (positive correlation) and –1 (negative correlation). If X and Y are independent, then with (3.41) it follows $\rho_{XY} = 0$.

Example 3.6: Bivariate exponential and normal distribution

Assume that storm duration and intensity, (X and Y), are both distributed according to an exponential distribution (see Kottegoda and Rosso, 1997):

$$F_X(x) = 1 - \exp(-ax), \quad x \geq 0;\ a > 0 \qquad F_Y(y) = 1 - \exp(-by), \quad y \geq 0;\ b > 0 \tag{3.43}$$

Their **joint cdf** given as a bivariate exponential distribution reads:

$$F_{XY}(x,y) = 1 - \exp(-ax) - \exp(-by) + \exp(-ax - by - cxy)$$
$$\text{with}: x,y \geq 0;\ a > 0, b > 0 \text{ and } 0 \leq c \leq ab \tag{3.44}$$

Hence, with (3.32), the **joint pdf** becomes:

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = \frac{\partial}{\partial y}\left(\frac{\partial F_{XY}(x,y)}{\partial x}\right) = \tag{3.45}$$
$$= \frac{\partial}{\partial y}\{a\exp(-ax) - (a+cy)\exp(-ax-by-cxy)\} =$$
$$= \{(a+cy)(b+cx) - c\}\exp(-ax-by-cxy)$$

The joint exponential probability density function with a = 0.05 h^{-1}, b = 0.4 h/mm and c = 0.01 $mm^{-1}$is shown in Figure 3.8.

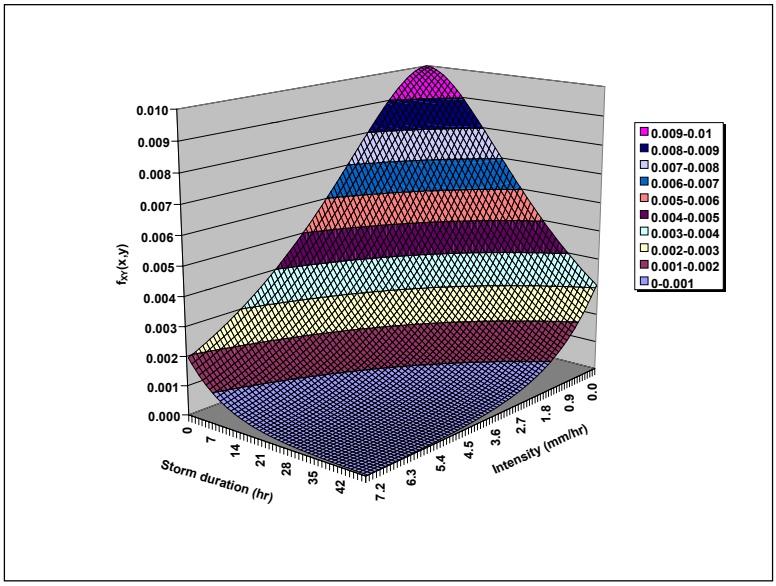


Figure 3.8: Joint probability density function of storm duration and rainfall intensity

The **conditional pdf** of storm intensity given rain duration is:

$$f_{Y|X}(x,y) = \frac{f_{XY}(x,y)}{f_X(x)} \frac{\{(a+cy)(b+cx)-c\}\exp(-ax-by-cxy)}{a\exp(-ax)} =$$
$$= \frac{\{(a+cy)(b+cx)-c\}}{a}\exp(-y(b+cx)) \tag{3.46}$$

The **conditional cdf** of a storm of given duration not exceeding a certain intensity reads:

$$F_{Y|X}(x,y) = \int_0^y f_{Y|X}(x,t)dt = \int_0^y \frac{\{a+ct)(b+cx)-c\}}{a}\exp(-t(b+cx))dt =$$

$$= 1 - \frac{a+cy}{a}\exp(-(b+cx)y) \qquad (3.47)$$

With a = 0.05 h^{-1}, b = 0.4 h/mm and c = 0.01 mm^{-1}, the conditional probability that a storm lasting 8 hours will exceed an average intensity of 4 mm/h becomes:

$$P(Y > 4 \mid X = 8) = 1 - F_{Y|X}(8,4) =$$

$$= 1 - 1 + \frac{0.05 + 0.01x4}{0.05}\exp(-(0.4 + 0.01x8)4) =$$

$$= 0.26$$

The **marginal distributions** follow from:

$$\left.\begin{aligned} f_X(x) &= \int_0^\infty f_{XY}(x,y)dy = \int_0^\infty \{(a+cy)(b+cx)-c\}\exp(-ax-by-cxy)dy = a\exp(-ax) \\ f_Y(y) &= \int_0^\infty f_{XY}(x,y)dx = \int_0^\infty \{(a+cy)(b+cx)-c\}\exp(-ax-by-cxy)dx = b\exp(-by) \end{aligned}\right\} \qquad (3.49)$$

If X and Y are **independent**, then c = 0 and it follows from (3.45):

$$f_{XY}(x,y) = ab\exp(-ax-by) = a\exp(-ax).b\exp(-by) = f_X(x).f_Y(y) \qquad (3.50)$$

Other examples of joint probability density functions are given in Figures 3.9 and 3.10, with the effect of correlation. In Figure 3.9 the joint standard normal pdf is given when the variables are independent, whereas in Figure 3.10 the variables are positively correlated (ρ = 0.8)

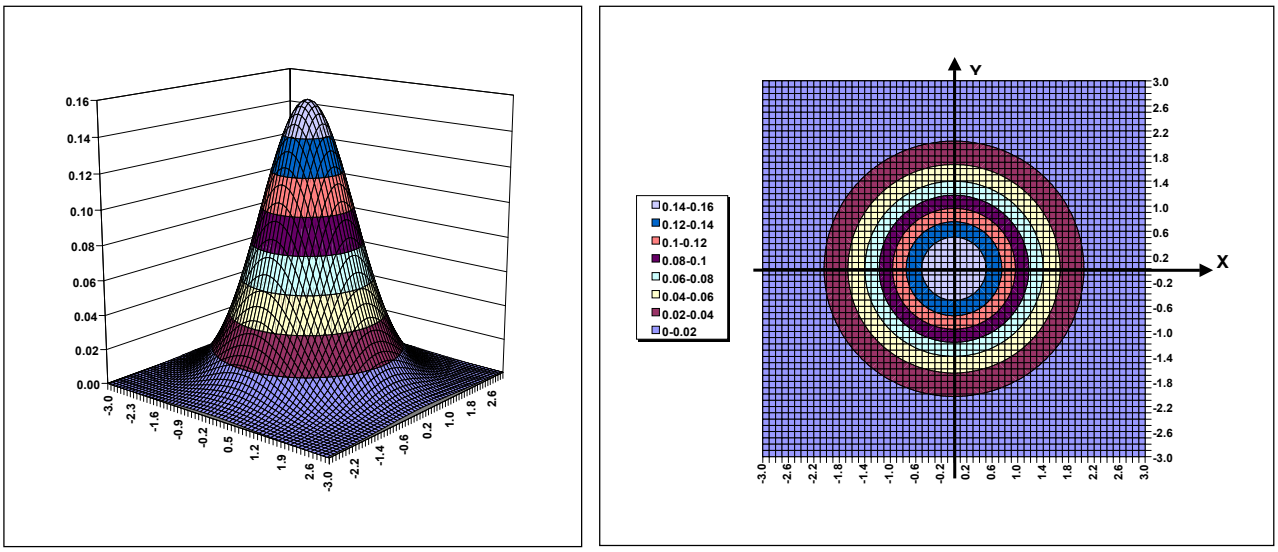


Figure 3.9: Bivariate standard normal distribution (ρ = 0)

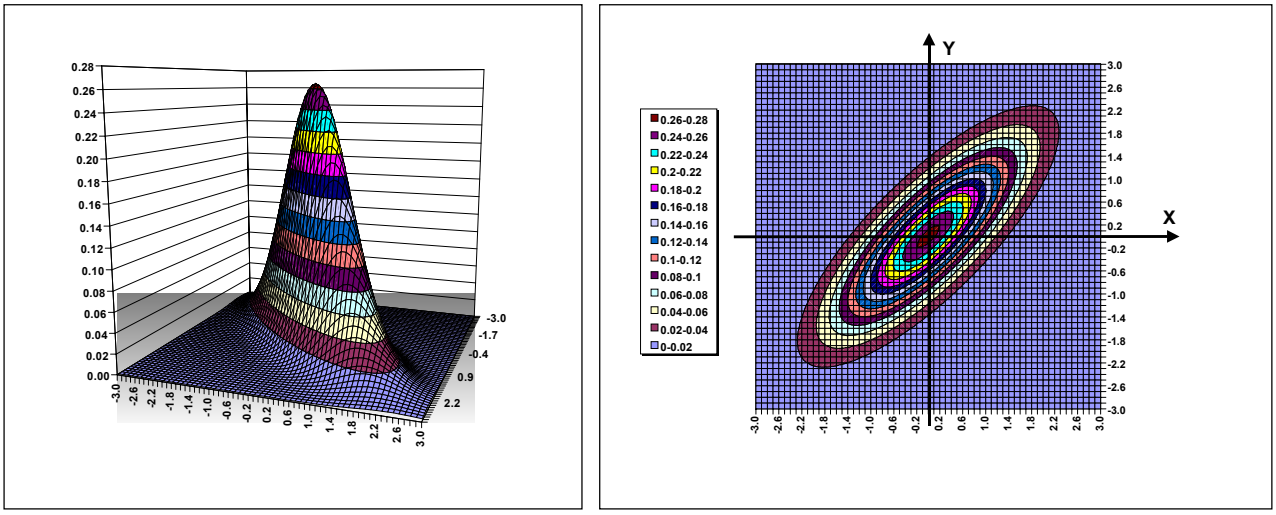


Figure 3.10: Bivariate standard normal distribution (ρ = 08)

The effect of correlation on the probability density function is clearly observed from the density contours in the right hand side representations of the joint pdf's.

3.2.4 Moment generating function

In some cases the moments as discussed before, cannot be computed in a simple manner. Then, often, use can be made of an auxiliary function, called the **moment generating function** G(s), which is the expectation of exp(sX): G(s) = E[exp(sX)]. In case of a continuous distribution:

$$G(s) = E[\exp(sX)] = \int_{-\infty}^{\infty} \exp(sx) f_X(x)dx \tag{3.50}$$

Assuming that differentiation under the integral sign is permitted one obtains:

$$\frac{d^k G(s)}{ds^k} = \int_{-\infty}^{\infty} x^k \exp(sx) f_X(x)dx \tag{3.51}$$

For s = 0 it follows: exp(sx) = 1, and the right hand side of (3.51) is seen to equal the k^{th} moment about the origin:

$$E[X^k] = G^{(k)}(0) \text{ where : } G^{(k)}(0) = \left.\frac{d^k G}{ds^k}\right|_{s=0} \tag{3.52}$$

Of course this method can only be applied to distributions for which the integral exists. Similar to the one-dimensional case, the moment generating function for bivariate distributions is defined by:

$$H(s,t) = E[\exp(sx + ty] = \int\int \exp((sx + ty) f_{XY}(x,y)dxdy \tag{3.53}$$

of which by partial differentiation to s and t the moments are found.

Example 3.7: Moment generating function for exponential distribution

The moment generating function for an exponential distribution and the k-th moments are according to (3.50) and (3.52):

$$G(s) = \int_0^\infty \exp(sx)\lambda \exp(-\lambda x)dx = \frac{\lambda}{\lambda - s}$$

$$E[X] = \left.\frac{dG}{ds}\right|_{s=0} = \left.\frac{\lambda}{(\lambda - s)^2}\right|_{s=0} = \frac{1}{\lambda}$$

$$E[X^2] = \left.\frac{d^2G}{ds^2}\right|_{s=0} = \left.\frac{2\lambda}{(\lambda - s)^3}\right|_{s=0} = \frac{2}{\lambda^2}$$

$$E[X^3] = \left.\frac{d^3G}{ds^3}\right|_{s=0} = \left.\frac{2x3\lambda}{(\lambda - s)^4}\right|_{s=0} = \frac{2x3}{\lambda^3} = \frac{6}{\lambda^3}$$

.....

$$E[X^k] = \left.\frac{d^kG}{ds^k}\right|_{s=0} = \left.\frac{2x3x...xk\lambda}{(\lambda - s)^{k+1}}\right|_{s=0} = \frac{k!}{\lambda^k} \qquad (3.54)$$

3.2.5 Derived distributions

Consider the variables X and Y and their one to one relationship Y = h(X). Let the pdf of X be $f_X(x)$, then what is the pdf of Y? For this, consider Figure 3.11. It is observed that the probability that X falls in the interval x, x + dx equals the probability that Y falls in the interval y, y + dy. Hence,

$$f_Y(y)dy = f_X(x)dx \qquad (3.55)$$

Since fY(y) cannot be negative, it follows:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \text{ or } f_Y(y) = |J| f_X(x) \qquad (3.56)$$

where the first derivative is called the **Jacobian** of the transformation, denoted by J.

In a similar manner bivariate distributions can be transformed.

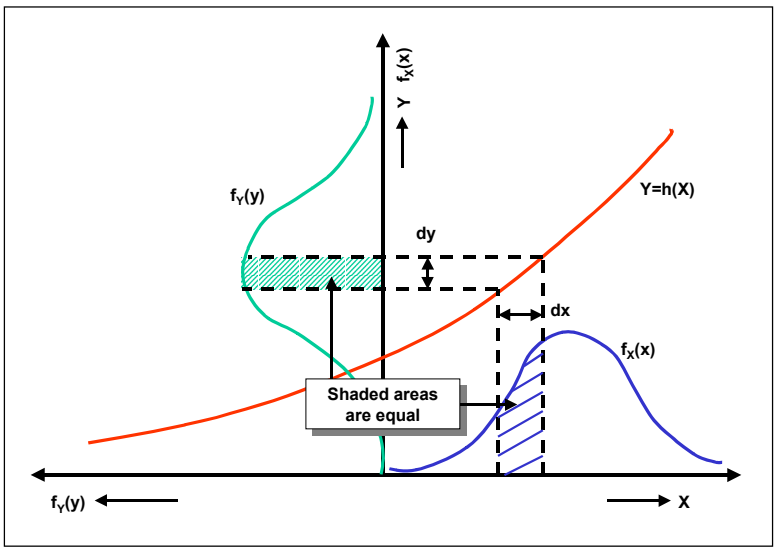


Figure 3.11: Definition sketch for derived distributions

Example 3.8: Transformation of normal to lognormal pdf

A variable Y is said to have a logarithmic normal or shortly log-normal distribution if its logarithm is normally distributed, hence ln(Y) = X. So:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_X}{\sigma_X} \right)^2 \right) \quad -\infty < X < \infty$$

$$X = \ln(Y)$$

$$\left| \frac{dx}{dy} \right| = \frac{1}{y}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi} y \sigma_{\ln(Y)}} \exp\left(-\frac{1}{2}\left(\frac{\ln y - \mu_{\ln Y}}{\sigma_{\ln Y}} \right)^2 \right) \quad 0 < Y < \infty$$

3.2.6 Transformation of stochastic variables

Consider the function Z = a + bX + cY, where X, Y and Z are stochastic variables and a, b and c are coefficients. Then for the mean and the variance of Z it follows:

$$E[Z] = E[a + bX + cY] = a + bE[X] + cE[Y] \tag{3.57}$$

$$E[(Z - E[Z])^2] = E[(a + bX + cY - a - bE[X] - cE[Y])^2] =$$

$$= E[b^2(X - E[X])^2 + c^2(Y - E[Y])^2 + 2bc(X - E[X])(Y - E[Y])] =$$

$$= b^2E[(X - E[X])^2] + c^2E[(Y - E[Y])^2] + 2bcE[(X - E[X])(Y - E[Y])]$$

or:

$$Var(Z) = b^2Var(X) + c^2Var(Y) + 2bcCov(X,Y) \tag{3.58}$$

Equations (3.57) and (3.58) are easily extendible for any linear function Z of n-random variables:

$$Z = \sum_{i=1}^{n} a_i X_i \tag{3.59}$$

$$E[Z] = \sum_{i=1}^{n} a_i E[X_i] = \sum_{i=1}^{n} a_i \mu_i$$

$$Var(Z) = Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i X_j) \tag{3.60}$$

Or in matrix notation by considering the vectors:

$$[\mathbf{a}] = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ a_n \end{bmatrix} \quad [\mathbf{X}] = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ . \\ . \\ X_n \end{bmatrix} \tag{3.61}$$

$$E[\mathbf{Z}] = E([\mathbf{a}]^T[\mathbf{X}]) = [\mathbf{a}]^T E([\mathbf{X}]) = [\mathbf{a}]^T[\mu]$$
$$\text{where : } E([\mathbf{X}]) = [\mu]$$

$$\begin{aligned} &Var(\mathbf{Z}) = Var([\mathbf{a}]^T[\mathbf{X}]) = E([\mathbf{a}]^T([\mathbf{X}]-[\mu])([\mathbf{X}]-[\mu])^T[\mathbf{a}]) = [\mathbf{a}]^T[\mathbf{V}][\mathbf{a}] \\ &\text{where : } [\mathbf{V}] = E(([\mathbf{X}]-[\mu])([\mathbf{X}]-[\mu])^T) \end{aligned} \tag{3.62}$$

The matrix [**V**] contains the following elements:

$$[\mathbf{V}] = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \ldots\ldots\ldots\ldots & Cov(X_1,X_n) \\ Cov(X_1,X_2) & Var(X_2) & \ldots\ldots\ldots\ldots & Cov(X_2,X_n) \\ . & & & \\ . & & & \\ . & & & \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \ldots\ldots\ldots\ldots & Var(X_n) \end{bmatrix} \tag{3.63}$$

This matrix is seen to be symmetric, since $Cov(X_i,X_j) = Cov(X_j,X_i)$. This implies $[\mathbf{V}] = [\mathbf{V}]^T$. Furthermore, since the variance of a random variable is always positive, so is $Var([\mathbf{a}]^T[\mathbf{X}])$.

Taylor's series expansion

For non-linear relationships it is generally difficult to derive the moments of the dependent variable. In such cases with the aid of Taylor's series expansion approximate expressions for the mean and the variance can be obtained. If Z = g(X,Y), then (see e.g. Kottegoda and Rosso (1997)):

$$\left.\begin{aligned} E[Z] &\approx g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}Var(X) + \frac{1}{2}\frac{\partial^2 g}{\partial y^2}Var(Y) + \frac{\partial^2 g}{\partial x \partial y}Cov(X,Y) \\ Var(Z) &\approx \left(\frac{\partial g}{\partial x}\right)^2 Var(X) + \left(\frac{\partial g}{\partial y}\right)^2 Var(Y) + 2\left(\frac{\partial g}{\partial x}\frac{\partial g}{\partial y}\right)Cov(X,Y) \end{aligned}\right\} \tag{3.64}$$

Above expressions are easily extendable to more variables. Often the variables in g(..) can be considered to be independent, i.e. Cov(..) = 0. Then (3.64) reduces to:

$$\left.\begin{aligned} E[Z] &\approx g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}Var(X) + \frac{1}{2}\frac{\partial^2 g}{\partial y^2}Var(Y) \\ Var(Z) &\approx \left(\frac{\partial g}{\partial x}\right)^2 Var(X) + \left(\frac{\partial g}{\partial y}\right)^2 Var(Y) \end{aligned}\right\} \tag{3.65}$$

Example 3.9

Given a function Z = X/Y, where X and Y are independent. Required are the mean and the variance of Z.

Use is made of equation (3.65). The coefficients read:

$$\left.\begin{aligned}\frac{\partial g}{\partial x} &= \frac{1}{y}\frac{\partial^2 g}{\partial x^2} = 0\\ \frac{\partial g}{\partial y} &= -\frac{x}{y^2}\frac{\partial^2 g}{\partial y^2} = \frac{2x}{y^3}\end{aligned}\right\} \qquad (3.66)$$

Hence:

$$\left.\begin{aligned}E[Z] &\approx \frac{\mu_X}{\mu_Y} + \frac{\mu_X}{\mu_Y^3}\sigma_Y^2 = \frac{\mu_X}{\mu_Y}\left(1 + C_{vY}^2\right)\\ Var(Z) &\approx \left(\frac{1}{\mu_Y}\right)^2\sigma_X^2 + \left(\frac{\mu_X}{\mu_Y^2}\right)^2\sigma_Y^2 = \left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Y^2}{\mu_Y^2}\right) = \left(\frac{\mu_X}{\mu_Y}\right)^2\left(C_{vX}^2 + C_{vY}^2\right)\end{aligned}\right\} \qquad (3.67)$$

Example 3.10: Joint cumulative distribution function

The joint pdf of X and Y reads:

$$f_{XY}(x,y) = \exp(-x - y/2) \quad \text{for:} \quad x > 0, y > 0$$
$$f_{XY}(x,y) = 0 \quad \text{for:} \quad x \le 0, y \le 0$$

Q: determine the probability that 2<X<5 and 1<Y<7

A: the requested probability is obtained from:

$$P(2 < X < 5 \cap 1 < Y < 7) = \int_2^5\int_1^7 f_{XY}(x,y)dxdy = \int_2^5 \exp(-x)dx\int_1^7 \exp(-y/2)dy = (-\exp(-x)\Big|_2^5)(-2\exp(-y/2)\Big|_1^7) =$$
$$= (-0.0067 - (-0.1353))(-0.0302 - (-0.6065)) = 0.0741$$

Example 3.11: Marginal distributions and independence (from: Reddy, 1997)

Given is the joint pdf of the variables X and Y:

$$f_{XY}(x,y) = \frac{2}{3}(x + 2y) \quad \text{for:} \quad 0 < x < 1, 0 < y < 1$$
$$f_{XY}(x,y) = 0 \quad \text{for:} \quad x \le 0, x \ge 1, y \le 0, y \ge 1$$

Q: a. find the marginal distributions of X and Y and

b. are X and Y independent?

A: a. the marginal distributions are obtained from:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_0^1 \frac{2}{3}(x+2y)dy = \frac{2}{3}(xy + 2\frac{y^2}{2})\Big|_0^1 = \frac{2}{3}\{(x+1)-(0)\} = \frac{2}{3}(x+1) \quad \text{for}: \quad 0 < x < 1$$

$$f_X(x) = 0 \quad \text{for}: \quad x \le 0, x \ge 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx = \int_0^1 \frac{2}{3}(x+2y)dx = \frac{2}{3}(\frac{x^2}{2} + 2xy)\Big|_0^1 = \frac{2}{3}\{(\frac{1}{2}+2y)-(0)\} = \frac{1}{3}(1+4y) \text{ for}: \quad 0 < y < 1$$

$$f_Y(y) = 0 \quad \text{for}: \quad y \le 0, y \ge 1$$

b. if X and Y are independent, then their conditional distributions should be equal to their marginal distributions. Hence is $f_{X|Y}(x,y) = f_X(x)$ or is $f_{Y|X}(x,y) = f_Y(y)$?

$$f_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{\frac{2}{3}(x+y)}{\frac{1}{3}(1+4y)} = 2\frac{(x+y)}{(1+4y)} \neq \frac{2}{3}(x+1)$$

So: $f_{X|Y}(x,y) \neq f_X(x)$, i.e. X and Y are not independent. A similar answer would of course have been obtained while examining $f_{Y|X}(x,y)$ relative to $f_Y(y)$.

Example 3.12: Joint pdf and independence (adapted from: Reddy, 1997)

Given are two variables X and Y who's marginal distributions read:

$$f_X(x) = 2a\ \exp(-bx) \quad \text{for}: \quad 0 \le x < \infty$$
$$f_Y(y) = 2a\ \exp(-by) \quad \text{for}: \quad 0 \le y < \infty$$

Q: a. find the joint pdf of X and Y if X and Y are independent

b. find the probability that X is always larger than Y

A: a. If X and Y are independent then their joint pdf is the product of their marginal distributions:

$$f_{XY}(x,y) = f_X(x).f_Y(y) = 4a^2\ \exp(-b(x+y)$$

b. the probability that X is always larger than Y can be obtained from the answer under a:

$$P(0 \le X < \infty \cap 0 \le Y < x) = \int_0^{\infty}\int_0^{x} f_{XY}(x,y)dydx = 4a^2 \int_0^{\infty} \exp(-bx)\left\{\int_0^{x} \exp(-by)dy\right\}dx =$$

$$4a^2 \int_0^{\infty} \exp(-bx)\left\{\frac{\exp(-bx)}{-b}\Big|_0^{x}\right\}dx = 4a^2 \int_0^{\infty} \exp(-bx)\left\{(-\frac{1}{b}\exp(-bx)) - (-\frac{1}{b})\right\}dx =$$

$$\frac{4a^2}{b} \int_0^{\infty} \exp(-bx)\{1-\exp(-bx)\}dx = \frac{4a^2}{b}\left\{\int_0^{\infty} \exp(-bx)dx - \int_0^{\infty} \exp(-2bx)dx\right\} =$$

$$\frac{4a^2}{b}\left\{\frac{\exp(-bx)}{-b}\Big|_0^{\infty} - \frac{\exp(-2bx)}{-2b}\Big|_0^{\infty}\right\} = \frac{4a^2}{b}\left\{\left(0-\frac{1}{-b}\right)-\left(0-\frac{1}{-2b}\right)\right\} = \frac{4a^2}{b}\left(\frac{1}{b}-\frac{1}{2b}\right) = 2\left(\frac{a}{b}\right)^2$$

4 Theoretical Distribution Functions

4.1 General

A number of theoretical (analytical) frequency distributions has been developed to model or represent the relative frequency distributions found in practice. In this chapter a summary is given of the distribution functions commonly used in hydrology and included in HYMOS.

A distinction is made between:

- Discrete distributions, and
- Continuous distributions.

A discrete distribution is used to model a random variable that has integer-valued outcomes, like the number of times an event occurs (successes) out of a number of trials. In contrast to this are the continuous distributions where the random variable is real-valued.

The **discrete** distributions (Section 4.2), which will be discussed, include:

- Binomial distribution
- Poisson distribution

The **continuous** distribution models comprise:

- Uniform distribution (Section 4.3),
- Distributions related to the normal distribution(Section 4.4), including:
  - Normal distribution
  - Log-normal distribution
  - Box-Cox transformations to normality
- Distributions related to Gamma or Pearson distribution, (Section 4.5), including:
  - Exponential distribution
  - Gamma distribution
  - Pearson Type 3 distribution or 3 parameter gamma distribution
  - Log-Pearson Type 3 distribution
  - Weibull distribution
  - Rayleigh distribution
- Distributions for extreme values(Section 4.6), including:
  - Generalised Extreme Value distributions, including the EV-1, EV-2 and EV-3 distributions for largest and smallest value
  - Generalised Pareto distributions, including Pareto Type 1, 2 and 3 distributions
- Sampling distributions(Section 4.7),:
  - Chi-square distribution
  - Student's t-distribution
  - Fisher F-distribution

It is stressed here that none of the theoretical distributions do have a physical background. They do not explain the physical phenomenon behind a population, but rather describe the behaviour of its frequency distribution. In this sub-section a short description of the various distributions is given.

Binomial distribution

The binomial distribution applies to a series of Bernoulli trials. In a Bernoulli trial there are two possible outcomes, that is an event occurs or does not occur. If the event occurs one speaks of a success (probability p) and if it does not occur it is a failure (probability 1 - p). If the probability of a success in each trial is constant, then the binomial distribution gives the distribution of the number successes in a series of independent trials. For example, the trial outcome could be that the water level in the river exceeds the crest of the embankment in a year and the other possible outcome that it does not. Let's call the event of an exceedance (how unfortunate for the designers) a "success". If the climatic conditions and the drainage characteristics in the basin do not vary one can assume that the success probability is constant from year to year. Knowing this success probability, then the Bernoulli distribution can be used to determine the probability of having exactly 0, 1, 2,…, or $\leq$1, $\leq$2, $\leq$…exceedances ("successes") during the next say 75 years (or any other number of years = number of trials). The distribution is therefore of extreme importance in risk analysis.

Poisson distribution

The Poisson distribution is a limiting case of the binomial distribution when the number of trials becomes large and the probability of success small, but their product finite. The distribution describes the number of occurrences of an event (a success) in a period of time (or space). Occurrences in a period of time (space) form a Poisson process if they are random, independent, and occur at some constant average rate. Essential is that the time (space) interval between the last occurrence and the next one is independent of past occurrences; a Poisson process, therefore, is **memory-less**.

Uniform distribution

The uniform distribution describes a random variable having equal probability density in a given interval. The distribution is particularly of importance for data generation, where the non-exceedance probability is a random variable with constant probability density in the interval 0,1.

Normal distribution

The normal distribution has a bell shaped probability density function, which is an appropriate model for a random variable being the sum of a large number of smaller components. Apart from being used as a sampling distribution or error model, the distribution applies particularly to the modelling of the frequency of aggregated data like monthly and annual rainfall or runoff. Direct application to model hydrological measurements is limited in view of its range from $-\infty$ to $+\infty$.

Lognormal distribution

If Y = ln X has normal distribution, then X is said to have a 2-parameter lognormal distribution. In view of its definition and with reference to the normal distribution, X can be seen as the product of a large number of small components. Its range from 0 to $+\infty$ is more appropriate to model hydrological series, whereas the logarithmic transformation reduces the positive skewness often found in hydrological data sets. Its applicability in hydrology is further enhanced by introducing a shift parameter x_0 to X to allow a data range from x_0 to $+\infty$. Then, if $Y = \ln(X - x_0)$ has normal distribution it follows that X has a 3-parameter lognormal distribution

Box-Cox transformation

The Box-Cox transformation is a suitable, effective two-parameter transformation to data sets to normality. Such transformations may be desired in view of the extensive tabulation of the normal distribution.

Exponential distribution

The time interval between occurrences of events in a Poisson process or inter-arrival time is described by the exponential distribution, where the distribution parameter represents the average occurrence rate of the events.

Gamma distribution

The distribution of the time until the γth occurrence in a Poisson process has a gamma distribution. In view of the definition of the exponential distribution the gamma distribution models the sum of γ independent, identical exponentially distributed random variables. Note that γ may be a non-integer positive value. The gamma distribution is capable of modelling skewed hydrological data series as well as the lognormal distribution is capable of. The gamma distribution has a zero lower bound and is therefore not applicable to phenomena with a non-zero lower bound, unless a shift parameter is introduced.

Pearson Type 3 or 3-parameter gamma distribution

The gamma distribution with a shift parameter to increase the flexibility on the lower bound is called the Pearson Type 3 distribution. Sometimes it is also called 3-parameter gamma distribution, though in literature the name gamma distribution is generally related to the 2-parameter case. The distribution can take on variety of shapes like the 3-parameter lognormal distribution and is therefore often used to model the distribution of hydrological variables. A large number of distributions are related to the Pearson Type 3 distribution. For this, consider the standard incomplete gamma function ratio:

$$F(z) = \frac{1}{\Gamma(\gamma)} \int_0^z s^{\gamma-1} \exp(-s) ds \text{ where : } Z = \left(\frac{X - x_0}{\beta} \right)^k$$

Note that the distribution reduces to an exponential function when $\gamma = 1$. In the above distribution x0 = location parameter, β = scale parameter and γ and k are shape parameters. The following distributions are included:

- k = 1, γ = 1: exponential distribution
- k = 1, x_0 = 0: gamma distribution
- k = 1, x_0 = 0, β = 2, $\gamma = \nu/2$: chi-squared distribution
- k = 1: 3-parameter gamma or Pearson Type 3 distribution
- k = 1, $Z = (\ln(X - x_0) - y_0)/\beta)^k$: log-Pearson Type 3 distribution
- k = -1: Pearson Type 5 distribution
- k = 2, γ = 1: Rayleigh distribution
- k = 2, γ = 3/2: Maxwell distribution
- γ = 1: Weibull distribution

Log-Pearson Type 3 distribution

If $X = \ln(Y - y_0)$ has a Pearson Type 3 distribution, then Y follows a log-Pearson Type 3 distribution. The distribution is often used to model annual maximum floods when the skewness is high.

Weibull distribution

The Weibull distribution is a special type of exponential or Pearson Type 3 distribution. The Weibull distribution is often used to model distributions of annual minimum values and as such it equals the Extreme Value Type III distribution for smallest values.

Rayleigh distribution

The Rayleigh distribution is a special case of the Weibull distribution. By comparison with the definition of the chi-squared distribution it is observed that a random variable is Rayleigh distributed if it is the root of the sum of two squared normal random variables. The distribution is often used to model distributions of maximum wind speed but also for annual maximum flows, if the skewness is limited.

Generalised Extreme Value distributions

Three types of Extreme Value distributions have been developed as asymptotic distributions for the largest or the smallest values. It depends on the parent distribution which type applies. The distributions are often called Fisher-Tippett Type I, II and III or shortly EV-1, EV-2 and EV-3 distributions for largest and smallest value. EV-1 for largest is known as the Gumbel distribution, EV-2 for largest as Fréchet distribution and EV-3 for smallest value as Weibull or Goodrich distribution. Above models apply typically to annual maximum or minimum series. Despite the fact that these distributions have particularly been derived for extreme values, it does not mean that one of the types always applies. Often the lognormal, Pearson and log-Pearson Type 3, Weibull or Rayleigh distributions may provide a good fit.

Generalised Pareto distributions

The Pareto distributions are particularly suited to model the distribution of partial duration series or annual exceedance series. The Extreme Value distributions for the annual maximum value can be shown to be related to the Pareto distributions with an appropriate model for the number of exceedances. Consequently as for the Extreme Value distributions also for the generalised Pareto distributions three types are distinguished: Pareto Type 1, 2 and 3 distributions.

Sampling distributions

An estimate is thought of as a single value from the imaginary distribution of all possible estimates, called the sampling distribution. Sampling distributions are introduced to be able to give the likely range of the true value of a parameter for which an estimate is made.

Chi-squared distribution

The sum of ν squared normally distributed random variables has a chi-squared distribution, where ν is the number of degrees of freedom. The distribution is a special case of the gamma distribution. The distribution is used to describe the sampling distribution of the variance; also, it finds application in goodness of fit tests for frequency distributions.

Student's t-distribution

The sampling distribution of many statistics is approximately standard normal if the statistic is scaled by its standard deviation. If the latter is replaced by its sample estimate with ν degrees of freedom then the sampling distribution of the statistic becomes a Student's t-distribution with the same number of degrees of freedom. When the number of degrees of freedom is sufficiently large, the Student distribution can be replaced by the normal distribution. The t variable is the ratio of a normal and the root of a chi-distributed variable divided by the number of degrees of freedom.

Fisher F-distribution

The ratio of two chi-squared variables divided by their degrees of freedom has a Fisher F-distribution. The distribution is used in significance tests on difference between variances of two series.

4.2 Discrete distribution functions

4.2.1 Binomial distribution

Distribution and cumulative distribution function

A **Bernoulli trial** is defined as a trial with only two possible outcomes: a **success** or a **failure**, with constant probability p and (1-p) respectively. The outcomes of a series of such trials are independent. Let X be the random variable for the number of successes out of n trials. Its probability distribution $p_X(x)$ is then given by the **binomial distribution**:

$$p_X(x) \equiv P(X = x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \text{ with : } x = 0,1,2,...,n \text{ and : } \binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{4.1}$$

The cdf reads:

$$F_X(x) \equiv P(X \le x; n; p) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

Moment related distribution parameters

The mean, variance and skewness are given by:

$$\begin{aligned} \mu_X &= np \\ \sigma_X^2 &= np(1-p) \end{aligned} \tag{4.3a}$$

$$\begin{aligned} \gamma_{1,X} &= \frac{(1-2p)}{\sqrt{np(1-p)}} \\ \gamma_{2,X} &= 3 + \frac{1-6p(1-p)}{np(1-p)} \end{aligned} \tag{4.3b}$$

From the skewness it is observed that only for p = 0.5 a symmetrical distribution function is obtained. For p < 0.5 the distribution is skewed to the right and for p > 0.5 skewed to the left. A few examples are given in Figure 4.1

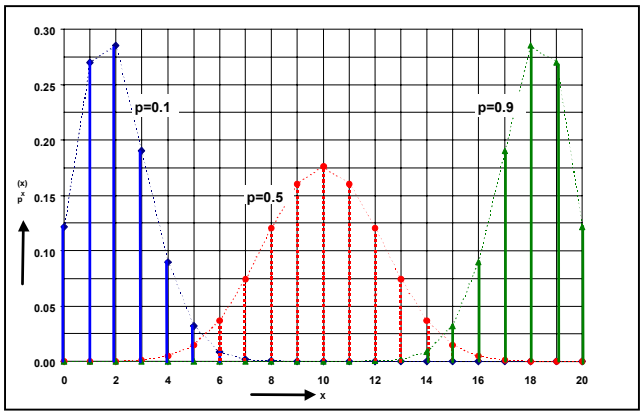


Figure 4.1:
Binomial distributions for n = 20 and p = 0.1, 0.5 and 0.9

From (4.3b) and Figure 4.2 it is observed that for large n, the skewness $\gamma_{1,X}$ gradually tends to 0 and the kurtosis $\gamma_{2,X}$ becomes close to 3. Then, the distribution approaches the **normal** distribution with same mean and variance (see Subsection 4.3.2).

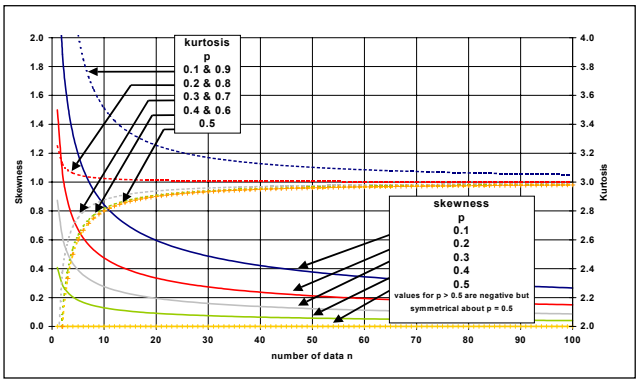


Figure 4.2:
Skewness and kurtosis of binomial distribution as function of n and p

Example 4.1 Number of rainy days in a week

Let the probability of a rainy day in a particular week in the year be 0.3, then:

- what is the probability of having exactly 4 rainy days in that week, and
- what is the probability of having at least 4 rainy days in that week?

Assuming that the occurrence of rainy days are independent, then the random variable X being the number of rainy days in that week follows a binomial distribution with n = 7 and p = 0.3. From (4.1) it then follows:

$$p_X(x) \equiv P(X = 4;7,0.3) = \binom{7}{4} 0.3^4 (1-0.3)^{7-4} = \frac{7!}{4!\,x3!} 0.3^4 0.7^3 = 0.097$$

Note that this is different from the probability of having 4 successive rainy days, which probability is 0.3 x 0.3 x 0.3 x 0.3 = 0.008, which is of course much less.

The probability of having at least 4 rainy days in that week of the year should be larger than 0.097, because also the probabilities of having 5, 6 or 7 days of rain should be included. The solution is obtained from (4.2):

$$F_X(X \geq 4) = 1 - F_X(X \leq 3) = 1 - \sum_{k=0}^{3} \binom{7}{k} 0.3^k (1-0.3)^{7-k} = 1 - (0.082 + 0.247 + 0.318 + 0.226) = 0.127$$

From the above it is observed that in case n and X are big numbers the elaboration of the sum will require some effort. In such cases the normal approximation is a better less cumbersome approach.

Related distributions

If the number of trials n = 1 then the binomial distribution is called **Bernoulli distribution** with mean p and variance p(1-p). The **geometric distribution** describes the probability that the first success takes place on the N^{th} trial. This distribution can be derived from (4.1) by noting that the N^{th} trial is preceded by (N – 1) trials without success, followed by a successful one. The probability of having first (N-1) failures is $(1-p)^{N-1}$ (from (3.12) or (4.1) with n = N-1 and x = 0) and the successful one has probability p, hence the probability of the first success in the N^{th} trial is $p(1-p)^{N-1}$ for N = 1,2,3,… In a similar manner the distribution function for the **negative binomial distribution** can be derived. This distribution describes the probability that the k^{th} exceedance takes place in the N^{th} trial. Hence, the N^{th} trial was preceded by (k-1) successes in (N-1) trials, which is given by (4.1) (with: n = N-1 and x = k-1), followed by a success with probability p.

4.2.2 Risk and return period

Consider a series of annual maximum discharges Q_{max} (t): t = 1,…, n. If a discharge Q_d is exceeded during these n years k-times then Q_d has in any one year an average probability of being exceeded of p_E = k/n and the average interval between the exceedances is n/k = $1/p_E$. The latter is called the **return period** T = $1/p_E$, as discussed in Sub-section 3.2.2, equation (3.21).

More generally, instead of Q_{max}, if we denote the random variable by Q, then the relation between $F_Q(q)$, T and p is:

$$F_Q(q) = P(Q \le q) = 1 - P(Q > q) = 1 - p_E = 1 - \frac{1}{T} \tag{4.4}$$

If one states that an embankment has been designed for a discharge with a return period of T years it means that **on average** only **once** during T years the river will overtop the embankment. But **each** year there is a probability p = 1/T that the river overtops the embankment. Consequently, each year the probability that the river does not overtop the embankment is $(1 - p_E) = F_Q(q)$. Since the outcomes in any one-year are independent, the probability of not being exceeded in N consecutive years is given by:

$$P(\text{no exceedances of q in N years}) = (F_Q(q))^N = (1-p_E)^N = \left(1 - \frac{1}{T}\right)^N \tag{4.5}$$

Note that this result is directly obtained from (4.1) with the number of successes x = 0. If q is the design level (storm, flow, stage, etc.), then the probability that this level q will be exceeded one or more times during the lifetime N of a structure (i.e. the probability of one or more failures), is simply the complement of the probability of no failures in N years. The probability of failure is called the **risk** r, hence:

$$r = 1 - (F_Q(q))^N = 1 - (1-p_E)^N = 1 - \left(1 - \frac{1}{T}\right)^N \tag{4.6}$$

It is noted that the above definition of risk is basically incomplete. The consequence of failure should also be taken into account. Risk is therefore often defined as the probability of failure times the consequence of failure.

Example 4.2 Risk of failure

A culvert has been designed to convey a discharge with a return period of 100 years. The lifetime of the structure is 50 years. What is the probability of failure during the lifetime of the structure?

$$r = 1 - \left(1 - \frac{1}{100}\right)^{50} = 1 - 0.605 = 0.395 \approx 40\%$$

Example 4.3 Return period and risk

To be 90% sure that a design discharge is not exceeded in an 80-year period, what should be the return period of the design discharge?

If we want to be 90% sure, then we take a risk of failure of 10%. From (4.6) it follows:

$$T = \frac{1}{1-(1-r)^{1/N}} = \frac{1}{1-(1-0.10)^{1/80}} = 760 \text{ years}$$

Hence for an event with an average return period of 760 years there is a 10% chance that in a period of 80 years such an event will happen.

4.2.3 Poisson distribution

Distribution and cumulative distribution function

If in the binomial distribution n becomes large and p very small, then (4.1) can be approximated by the **Poisson distribution**. Let the average number of successes in a series of n Bernoulli trials be ν = np, then the distribution of the number of successes X in n trials, with probability of occurrence in each trial of p, becomes, see also Figure 4.3:

$$p_X(x) \equiv P(X = x;\nu) = \frac{\nu^x \exp(-\nu)}{x!} \text{ for : } x = 0,1,2,....,n \qquad (4.7)$$

The cdf of the Poisson distribution reads:

$$F_X(x) \equiv P(X \le x;\nu) = \sum_{k=0}^{x} \frac{\nu^k \exp(-\nu)}{k!} \qquad (4.8)$$

Moment related distribution parameters

The mean, variance, skewness and kurtosis are:

$$\mu_X = \nu$$
$$\sigma_X^2 = \nu \qquad (4.9a)$$

$$\gamma_{1,X} = \frac{1}{\sqrt{\nu}}$$
$$\gamma_{2,X} = 3 + \frac{1}{\nu} \qquad (4.9b)$$

For $\nu \to \infty$ the skewness becomes 0 and the kurtosis 3, and the Poisson distribution converges to a normal pdf.

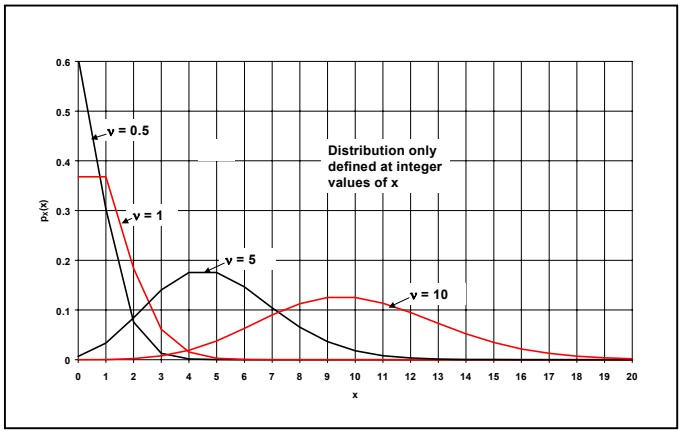


Figure 4.3:
Poisson distribution for different values of ν

Example 4.4: Drought

From a statistical analysis it was deduced that the monsoon rainfall at a location falls below 200 mm on average once in 100 years. What is the probability that the monsoon rainfall will fall below 200 mm less than twice in a 75-year period?

In this case n = 75 and the 'success' probability (falling below 200 mm) p = 1/100 = 0.01, hence n is large and p is small, which fulfils the condition for the applicability of the Poisson distribution. With ν = np = 75 x 0.01 = 0.75 it follows from (4.8):

$$F_X(x) = P(X \le 1;0.75) = \sum_{k=0}^{1} \frac{0.75^k \exp(-0.75)}{k!} =$$

$$= \left(\frac{0.75^0}{0!} + \frac{0.75^1}{1!} \right) \exp(-0.75) = (1 + 0.75)\exp(-0.75) = 0.8266$$

With the binomial cdf (4.2) we would have obtained:

$$F_X(x) = P(X \le 1\,;\,75\,;\,0.01) = \sum_{k=0}^{1} \binom{75}{k} 0.01^k (1-0.01)^{75-k} =$$

$$= 1\text{x}1\text{x}0.99^{75} + 75\text{x}0.01\text{x}0.99^{74} = 0.4706 + 0.3565 = 0.8271$$

The result is seen to differ by < 0.1%, hence the Poisson distribution is a simple practical alternative to the binomial distribution under the conditions of large n and small p.

Poisson and exponential distribution

The Poisson distribution forms the basis for the exponential distribution. For this ν is considered as the average number of arrivals or happenings in a time period t. The arrival rate is denoted by $\lambda = \nu/t$. Consider the time between arrivals as a random variable T_a, and its probability distribution is $P(T_a \le t) = 1 - P(T_a > t)$, where $P(T_a > t)$ represents the probability of **no** occurrences or arrivals (i.e. no successes) in a period t and is according to (4.7) given:

$$P(T_a > t) = P(X = 0;\lambda t) = \frac{(\lambda t)^0 \exp(-\lambda t)}{0!} = \exp(-\lambda t) \qquad (4.10)$$

Hence the cumulative probability distribution of the time between arrivals becomes with (4.10):

$$F_{T_a}(t) = P(T_a \le t) = 1 - \exp(-\lambda t) \qquad (4.11)$$

It shows that the **waiting time** between successive events of a Poisson process follows an **exponential distribution**. Instead of time, the Poisson process can also be defined for space, length, etc. Essential for a Poisson process is that the "period" can be divided in subintervals Δt so small, that the probability of an arrival in Δt tends to $\lambda\Delta t$, while the probability of more than one arrival in Δt is zero and an occurrence in one subinterval is independent of the occurrence in any other, (Kottegoda and Rosso, 1997). This makes the process memory-less.

Example 4.2 continued **Risk of failure**

The average waiting time for the design event was 100 years. The structure will fail in the 50 year period, if the waiting time between the design events is less or equal to 50 years, which was defined as risk. From (4.11) with $\lambda = 1/T = 1/100$ and $t = N = 50$ we obtain:

$$r = F(T_a \le 50) = 1 - \exp(-\frac{1}{100}50) = 1 - 0.607 = 0.393$$

This result is seen to be close to the outcome of (4.6), which was r = 0.395.

4.3 Uniform distribution

Probability density and cumulative frequency distribution

The uniform or rectangular distribution describes the probability distribution of a random variable X, which has equal non-zero density in an interval 'ab' and zero density outside. Since the area under the pdf should equal 1, the pdf of X is given by:

$$\left.\begin{aligned} f_X(x) &= \frac{1}{b-a} \text{ for : } a \le x \le b \\ f_X(x) &= 0 \text{ for : } x < a \,;\, x > b \end{aligned}\right\} \qquad (4.12)$$

The cdf of the uniform distribution reads:

$$\left.\begin{aligned} F_X(x) &= 0 \quad \text{for : } x < a \\ F_X(x) &= \int_a^x \frac{1}{b-a} ds = \frac{x-a}{b-a} \quad \text{for : } a \le x \le b \\ F_X(x) &= 1 \quad \text{for : } x > b \end{aligned}\right\} \qquad (4.13)$$

The pdf and cdf of the uniform distribution are shown in Figure 4.4.

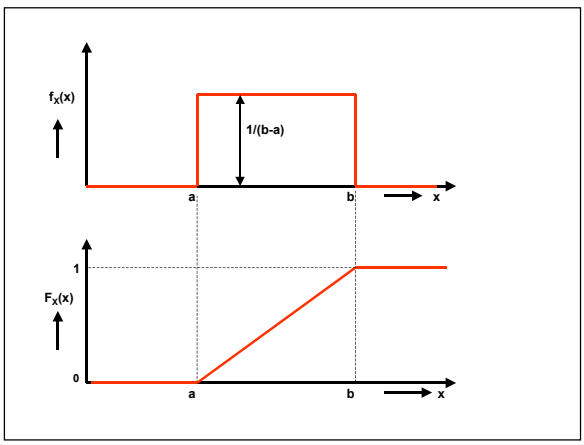


Figure 4.4:
Pdf and cdf of uniform distribution

Moment related distribution parameters

The mean and the variance simply follow from the definition of the moments:

$$\left.\begin{aligned} \mu_X &= \frac{a+b}{2} \\ \sigma_X^2 &= \frac{(b-a)^2}{12} \end{aligned}\right\} \qquad (4.14)$$

The uniform distribution is of particular importance for data generation, where with a = 0 and b = 1 the density function provides a means to generate the non-exceedance probabilities. It provides also a means to assess the error in measurements due to limitations in the scale. If the scale interval is c, it implies that an indicated value is ± ½ c and the standard deviation of the measurement error is $\sigma = \sqrt{(c^2/12)} \approx 0.3c$.

4.4 Normal distribution related distributions

4.4.1 Normal Distribution

Four conditions are necessary for a random variable to have a **normal** or **Gaussian distribution** (Yevjevich, 1972):

- A very large number of causative factors affect the outcome
- Each factor taken separately has a relatively small influence on the outcome
- The effect of each factor is independent of the effect of all other factors
- The effect of various factors on the outcome is additive.

Probability density and cumulative frequency distribution

The pdf and cdf of the normal distribution read:

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \quad \text{with} \quad -\infty < x < \infty, \quad -\infty < \mu_X < \infty \quad \text{and} \quad \sigma_X > 0 \qquad (4.15)$$

$$F_X(x) \equiv P[X \le x] = \int_{-\infty}^{x} \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{s-\mu_X}{\sigma_X}\right)^2\right) ds \qquad (4.16)$$

where: x = normal random variable

μ_X, σ_X = parameters of the distribution, respectively the mean and the standard deviation of X.

The pdf and cdf are displayed in Figure 4.5.

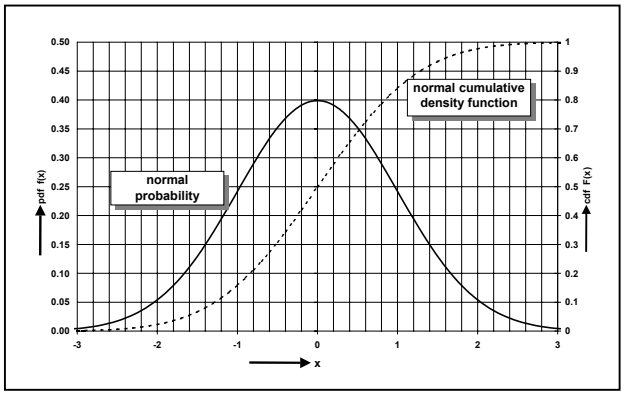


Figure 4.5:
Normal probability density and cumulative density functions for μ = 0 and σ = 1

The normal pdf is seen to be a bell-shaped symmetric distribution, fully defined by the two parameters μ_X and σ_X. The coefficient $(\sigma_X\sqrt{(2\pi)})^{-1}$ in Equation (4.15) is introduced to ensure that the area under the pdf-curve equals unity, because the integral:

$$\int_{-\infty}^{\infty} \exp(-ax^2)dx = 2\int_{0}^{\infty} \exp(-ax^2)dx = \sqrt{\frac{\pi}{a}}$$

With a = $1/(2\sigma_X^2)$ the integral becomes $\sigma_X\sqrt{(2\pi)}$, so dividing the integral by the same makes the area under the pdf equal to 1.

The notation $N(\mu_X, \sigma_X^2)$ is a shorthand for the normal distribution. The normal pdf for different values of μ_X and of σ_X are shown in Figures 4.6 and 4.7. Clearly, μ_X is a **location** parameter; it shifts the distribution along the x-axis, but does not change the shape or scale of the distribution as is shown in Figure 4.6. The parameter σ_X is a **scale** parameter; it stretches or reduces the scale of the horizontal axis, see Figure 4.7, but it has no effect on the shape of the distribution.

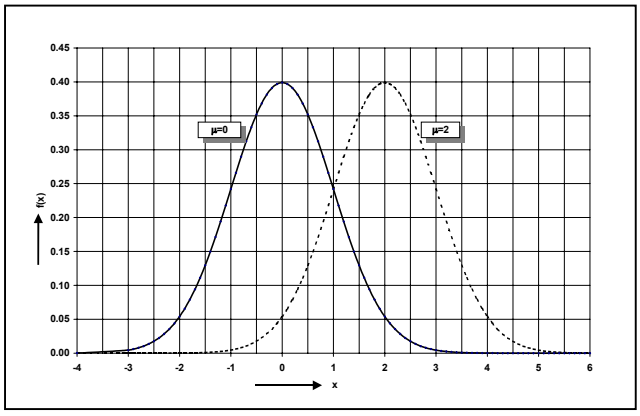


Figure 4.6:
Normal probability density functions for different values of μ_x' (σ_x=1)

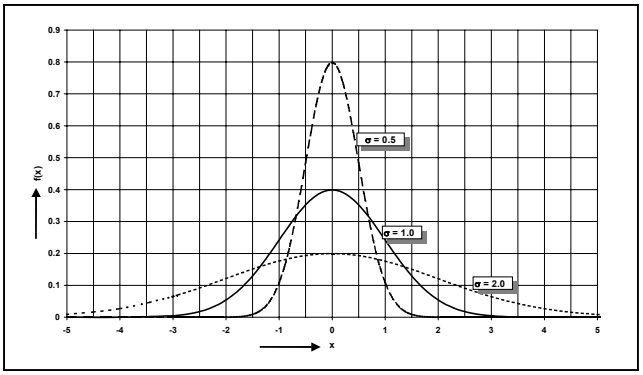


Figure 4.7:
Normal probability density functions for different values of σ_X, (μ_X = 0).

Moment related parameters of the distribution

The characteristics of the distribution are as follows:

Mean = median = mode: μ_X (4.17a)

Variance: σ_X^2 (4.17b)

Standard deviation: σ_X (4.17c)

Coefficient of variation: $C_{v,X} = \sigma_X/\mu_X$ (4.17d)

Skewness: $\gamma_{1,X} = 0$ (4.17e)

Kurtosis: $\gamma_{2,X} = 3$ (4.17f)

Standard normal distribution

The location and scale parameters μ_X and σ_X are used to define the **standard normal variate** or **reduced variate** Z:

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{4.18}$$

It is observed that Z = X for μ_X =0 and σ_X = 1, hence Z is an N(0,1) variate with pdf and cdf respectively:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) \tag{4.19}$$

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{1}{2}s^2)ds = \frac{1}{2} + \frac{1}{2}\mathrm{erf}(\frac{z}{\sqrt{2}}) \tag{4.20}$$

Equations (4.19) and (4.20) describe the **standard normal** probability density and cumulative density function, see Figure 4.5. From (4.18) it follows:

$$dz = \frac{1}{\sigma_X} dx$$

Substitution of this expression in (4.20) with (4.18) results in equation (4.16) and by taking the derivative with respect to X one obtains (4.15). The procedures used in HYMOS to solve (4.20) given Z and to calculate the inverse (i.e. the value of Z given $F_Z(z)$) are presented in Annex 4.1.

The standard normal distribution is generally tabulated in statistical textbooks. Such tables generally only address the positive arguments. To apply these tables for negative arguments as well, note that because of the symmetry of the pdf it follows:

$f_Z(-z) = f_Z(z)$ (4.21)

and

$F_Z(-z) = 1 - F_Z(z)$ (4.22)

Quantiles

Values of x_T and z_T for which $F_X(x_T) = F_Z(z_T) = 1 - 1/T$ are related by (4.18) and by its inverse:

$$x_T = \mu_X + \sigma_X z_T \qquad (4.23)$$

Z_T is obtained as the inverse of the standard normal distribution.

Example 4.5 Tables of the normal distribution

For z = 2, $f_Z(2) = 0.0540$, hence $f_Z(-2) = 0.0540$

For z = 1.96 $F_Z(1.96) = 0.9750$,

Hence: $F_Z(-1.96) = 1 - 0.9750 = 0.0250$

It implies that the area under the pdf between z = -1.96 and z = 1.96 (see Figure 4.8) amounts 0.9750 – 0.0250 = 0.95 or 95%.

Given that the mean of a random variable is 100 and its standard deviation is 50, the quantile for T = 100 is derived as follows:

For T = 100, $F_Z(z) = 1 - 1/100 = 0.99$. From the table of the normal distribution this non-exceedance probability corresponds with a reduced variate $z_T = 2.33$. Hence, using (4.23):

$x_T = \mu_X + \sigma_X z_T = 100 + 50 \times 2.33 = 216.5$

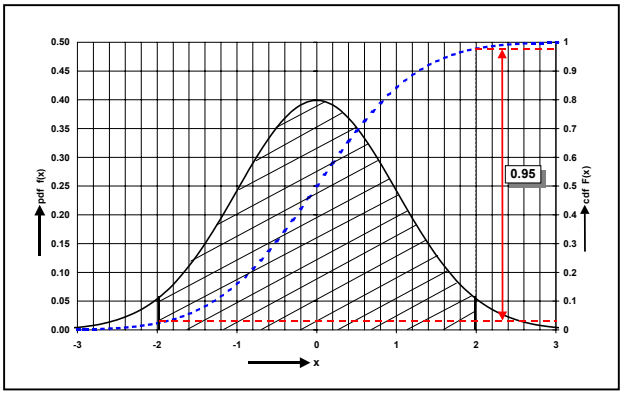


Figure 4.8:
Use of symmetry of standard normal pdf around 0 to find non-exceedance probabilities

Some Properties of the Normal Distribution

1. A linear transformation Y = a + bX of an $N(\mu_X, \sigma_X^2)$ random variable X makes Y an $N(a + b\mu_X, b^2\sigma_X^2)$ random variable.

2. If S_n is the sum of n independent and identically distributed random variables X_i each having a mean μ_X and variance σ_X^2, then in the limit as n approaches infinity, the distribution of S_n approaches a normal distribution with mean $n\mu_X$ and variance $n\sigma_X^2$.

3. Combining 1 and 2, for the mean X_m of X_i it follows, using the statement under 1 with a = 0 and b = 1/n, that X_m tends to have an $N(\mu_X, \sigma_X^2/n)$ distribution as n approaches infinity:

$$m_x = \frac{1}{n}\sum_{i=1} x_i = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n \text{ so: } E[m_x] = \frac{1}{n}E[E_{x_i}] = \frac{1}{n}\ E[\sum x_i] = \frac{1}{n}.n\mu_x = \mu_x$$

$$Var(m_x) = \frac{1}{n^2}Var(x_i) + \frac{1}{n^2}Var(x_i) + \cdots + \frac{1}{n^2}Var(x_n) = \frac{1}{n^2}\sum Var(x_i) = \frac{1}{n^2}\cdot n\,Var(x) = \frac{\sigma_x^2}{n}$$

If X_i is from an $N(\mu_X, \sigma_X^2)$ population, then the result for the sum and the mean holds regardless of the sample size n. The **Central Limit Theorem**, though, states that **irrespective of the distribution of X_i** the sum S_n and the mean X_m will tend to normality asymptotically. According to Haan (1979) if interest is in the main bulk of the distribution of S_n or X_m then n as small as 5 or 6 will suffice for approximate normality, whereas larger n is required for the tails of the distribution of S_n or X_m. It can also be shown that even if the X_i's have different means and variances the distribution of S_n will tend to be normal for large n with $N(\Sigma\mu_{Xi}; \Sigma\sigma_{Xi}^2)$, provided that each X_i has a negligible effect on the distribution of S_n, i.e. there are no few dominating X_i's.

An important outcome of the Central Limit Theorem is that if a hydrological variable is the outcome of n independent effects and n is relatively large, the distribution of the variable is approximately normal.

Application in hydrology

The normal distribution function is generally appropriate to fit annual rainfall and annual runoff series, whereas quite often also monthly rainfall series can be modelled by the normal distribution. The distribution also plays an important role in modelling random errors in measurements.

4.4.2 Lognormal Distribution

Definition

In the previous section it was reasoned, that the addition of a large number of small random effects will tend to make the distribution of the aggregate approximately normal. Similarly, a phenomenon, which arises from the **multiplicative** effect of a large number of uncorrelated factors, the distribution tends to be lognormal (or logarithmic normal); that is, the logarithm of the variable becomes normally distributed (because if $X = X_1X_2X_3\ldots$ Then $\ln(X) = \ln(X_1) + \ln(X_2) + \ln(X_3) + \ldots$).

Let X be a random variable such that $X - x_0 > 0$ and define

$$Y = \ln(X - x_0) \tag{4.23}$$

If Y has a normal distribution $N(\mu_Y, \sigma_Y^2)$, then X is said to have a 3-parameter log-normal distribution $LN(x_0, \mu_Y, \sigma_Y)$ or shortly LN-3. If x_0 is zero (or given) then the distribution of X is called a 2-parameter log-normal distribution $LN(\mu_Y, \sigma_Y)$ or LN-2.

Probability density and cumulative frequency distribution

The pdf of the normal random variable Y is given by:

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right) \tag{4.24}$$

The pdf of X is obtained from the general transformation relation (3.56):

$$f_X(x) = f_Y(y)\left|\frac{dy}{dx}\right|$$

Since $Y = \ln(X - x_0)$ so: $|dy/dx| = 1/(X - x_0)$ for $X > x_0$, it follows from (4.24) for the pdf of X:

$$f_X(x) = \frac{1}{(x - x_0)\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x - x_0) - \mu_Y}{\sigma_Y}\right)^2\right) \quad \text{with:} \quad x > x_0 \tag{4.25}$$

Equation (4.25) is the LN-3 pdf. The LN-2 pdf follows from (4.25) with $x_0 = 0$:

$$f_X(x) = \frac{1}{x\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu_Y}{\sigma_Y}\right)^2\right) \quad \text{with:} \quad x > 0 \tag{4.26}$$

To appreciate the parameters of the distribution, note the relation between the moment related parameters of the distribution and the parameters x_0, μ_Y and σ_Y:

Moment related parameters

$$\begin{aligned}
&\text{Mean:} && \mu_X = x_0 + \exp(\mu_Y + \tfrac{1}{2}\sigma_Y^2) \\
&\text{Median:} && M_X = x_0 + \exp(\mu_Y) \\
&\text{Mode:} && m_X = x_0 + \exp(\mu_Y - \sigma_Y^2)
\end{aligned} \tag{4.27a}$$

$$\begin{aligned}
&\text{Variance:} && \sigma_X^2 = \left(\exp(\mu_Y + \tfrac{1}{2}\sigma_Y^2)\right)^2\left(\exp(\sigma_Y^2) - 1\right) \\
&\text{Stdv:} && \sigma_X = \exp(\mu_Y + \tfrac{1}{2}\sigma_Y^2)\sqrt{\exp(\sigma_Y^2) - 1} \\
&\text{Parameter } \eta\text{:} && \eta = \frac{\sigma_X}{\mu_X - x_0} = \sqrt{\exp(\sigma_Y^2) - 1} \\
&\text{Skewness:} && \gamma_{1,X} = \eta^3 + 3\eta \\
&\text{Kurtosis:} && \gamma_{2,X} = 3 + 16\eta^2 + 15\eta^4 + 6\eta^6 + \eta^8
\end{aligned} \tag{4.27b}$$

It is observed from the above equations that the first moment parameters are dependent on x_0, μ_Y and σ_Y. The variance depends on μ_Y and σ_Y, whereas the skewness and kurtosis are only dependent on σ_Y. This is also illustrated in the Figures 4.9 to 4.11. Clearly, x_0 is a ***location*** parameter (see Figure 4.9); it shifts only the distribution function, whereas μ_Y is a ***scale*** parameter, as the latter does not affect the skewness (see Figure 4.10). The parameter σ_Y is a ***shape*** parameter, since it affects the shape of the pdf as is deduced from (4.27) and Figure 4.11).

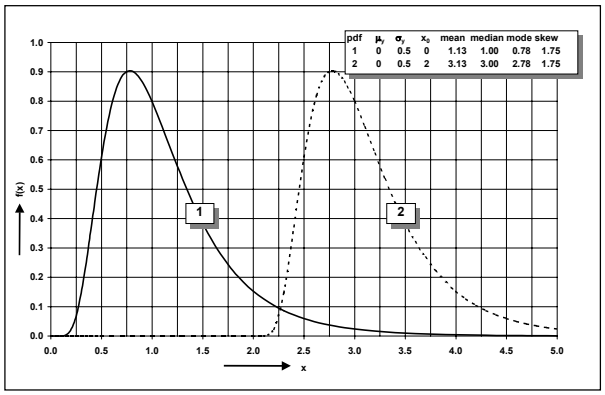


Figure 4.9:
Effect of location parameter x_0 on lognormal distribution

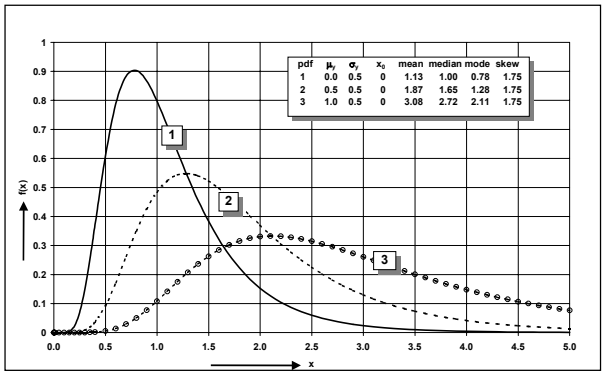


Figure 4.10:
Effect of scale parameter μ_y on lognormal distribution

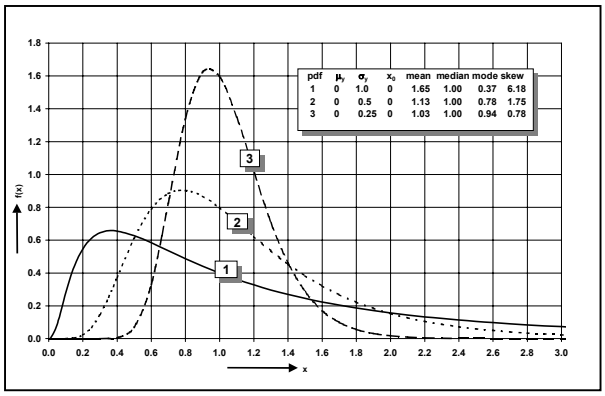


Figure 4.11:
Effect of is a shape parameter σ_y on lognormal distribution

Equation (4.27a) shows that for a lognormal distribution the following inequality holds:

x_0 < mode < median < mean

From (4.27b) it is observed that $\eta > 0$ hence $\gamma_1 > 0$ and $\gamma_2 > 3$; so the skewness is always positive and since the kurtosis is greater than 3 the lognormal distribution has a relatively greater concentration of probability near the mean than a normal distribution. The relation between γ_1 and η is displayed in Figure 4.12.To cope with negative skewness and distributions of smallest values, the sign of X or (X-x_0) has to be changed, see Sub-section 4.3.13.

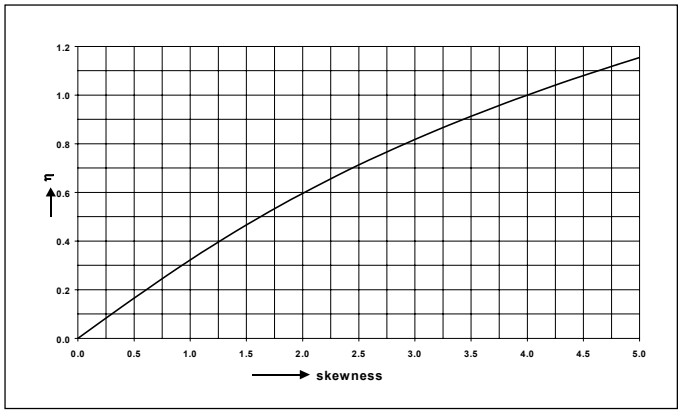


Figure 4.12:
η as function of skewness γ_1

Distribution parameters expressed in moment related parameters

The distinction between LN-2 and LN-3 is important. From equation (4.27) it is observed that when $x_0 = 0$ the parameters μ_Y and σ_Y are fully determined by the first two moments μ_X and σ_X which then also determine the skewness and kurtosis through their fixed relation with the coefficient of variation η.

For LN-2 the following inverse relations can be derived:

$$\mu_Y = \ln(\mu_X) - \frac{1}{2}\ln\left(\left(\frac{\sigma_X}{\mu_X}\right)^2 + 1\right) = \ln(\mu_X) - \frac{1}{2}\ln(C_{v,X}^2 + 1) \tag{4.28}$$

$$\sigma_Y = \sqrt{\ln\left(\left(\frac{\sigma_X}{\mu_X}\right)^2 + 1\right)} = \sqrt{\ln(C_{v,X}^2 + 1)} \tag{4.29}$$

The mean and the coefficient of variation of X are seen to describe the LN-2 pdf.

For **LN-3** the inverse relations are more complex as the starting point is the cubic equation in η relating η and $\gamma_{1,X}$, from (4.27b):

$$\eta^3 + 3\eta - \gamma_{1,X} = 0 \tag{4.30}$$

The parameters of the LN-3 distribution can be expressed in η (i.e. $\gamma_{1,X}$), μ_X and σ_X :

$$\eta = \left(\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2}\right)^2}\right)^{1/3} - \left(-\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2}\right)^2}\right)^{1/3} \tag{4.31}$$

The parameters of the LN-3 distribution can be expressed in η (i.e. $\gamma_{1,X}$), μ_X and σ_X:

$$x_0 = \mu_X - \frac{\sigma_X}{\eta} \tag{4.32}$$

$$\sigma_Y = \ln(\eta^2 + 1) \tag{4.33}$$

$$\mu_Y = \ln(\mu_X - x_0) - \frac{1}{2}\sigma_Y^2 = \frac{1}{2}\ln\left(\frac{\sigma_X^2}{\eta^2(\eta^2 + 1)}\right) \tag{4.34}$$

If the parameters would be determined according to equations (4.32) to (4.34) one observes that the **shape** parameter σ_Y is solely determined by the skewness, the **scale** parameter μ_Y by the variance and the skewness and the **location** parameter x_0 by the first three moments.

Moment generating function

The expressions presented in (4.27a/b) can be derived by observing that:

$$E[(X-x_0)^k] = \int_{-\infty}^{\infty} (x-x_0)^k f_X(x)dx = \int_{-\infty}^{\infty} \exp(ky)\frac{1}{\sigma_Y\sqrt{2\pi}}\exp(-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 dy = E[\exp(kY)] =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(ky - \frac{1}{2}\frac{(y-\mu_y)^2}{\sigma_y}\right)dy$$

$$\text{with : } u = \frac{y-\mu_Y}{\sigma_Y} - k\sigma_Y \text{ it follows :}$$

$$u^2 = \left(\frac{y-\mu_y}{\sigma_y} - k\sigma_y\right)^2 = \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2k(y-\mu_y) + k^2\sigma_y^2$$

$$-\frac{1}{2}u^2 = -\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2 + ky - k\mu_y - \frac{1}{2}k^2\sigma_y^2 \quad \text{or :} \quad ky - \frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2 = k\mu_y + \frac{1}{2}k^2\sigma_y^2 - \frac{1}{2}u^2$$

Hence, the power of the exponential can be replaced by:

$$k\mu_y + \frac{1}{2}k^2\sigma_y^2 - \frac{1}{2}u^2 \quad \text{and with :} \quad du = \frac{1}{\sigma_y}dy \text{ or : } dy = \sigma_y du \text{ one gets :}$$

$$E[\exp(kY)] = \exp(k\mu_Y + \frac{1}{2}k^2\sigma_Y^2)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)du$$

The last integral is seen to be 1, hence it follows for $E[(X-x_0)^k] = E[\exp(kY)]$:

$$E[(X-x_0)^k] = \exp(k\mu_Y + \frac{1}{2}k^2\sigma_Y^2) \tag{4.35}$$

Quantiles

The non-exceedance probability of the lognormally distributed variable X is derived through the standard normal distribution by inserting the standard normal variate Z derived as follows:

$$Z = \frac{Y-\mu_Y}{\sigma_Y} \text{ and : } Y = \ln(X-x_0) \tag{4.36}$$

The computation of the standard normal distribution is presented in Annex A4.1 or is obtained from tables in statistical textbooks.

The reverse, given a return period T or non-exceedance probability p, the quantile x_T or x_p is obtained from the standard normal distribution presented in Annex A4.2 or from tables through the standard normal deviate Z as follows:

$$x_T = x_0 + \exp(\mu_Y + z_T\sigma_Y) \tag{4.37}$$

Example 4.6 Lognormal distribution

Given is a LN-3 distributed variate X with mean 20, standard deviation 6 and skewness 1.5. Derive:

- the quantile for T=10.
- Return period of x = 35

To solve the first problem use is made of equation (4.37). The reduced variate z_T is obtained as the inverse of the standard normal distribution for a non-exceedance probability of $F_Z(z_T)$ =1-1/10 = 0.9. From the tables of the standard normal distribution one obtains:

$z_T = 1.282$

Next application of (4.37) requires values for the parameters x_0, σ_Y and μ_Y. These are determined using equations (4.31) to (4.34). The parameter η as a function of the skewness follows from (4.31), which gives with $\gamma_{1,X} = 1.5$:

$$\eta = \left(\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2} \right)^2} \right)^{1/3} - \left(-\frac{\gamma_{1,X}}{2} + \sqrt{1 + \left(\frac{\gamma_{1,X}}{2} \right)^2} \right)^{1/3} = 1.260 - 0.794 = 0.466$$

Then for x_0, σ_Y and μ_Y it follows from (4.32) to (4.34) respectively:

$$x_0 = \mu_X - \frac{\sigma_X}{\eta} = 20 - \frac{6}{0.466} = 7.130$$
$$\sigma_Y^2 = \ln(\eta^2 + 1) = 0.197 \text{ so : } \sigma_Y = 0.444$$
$$\mu_Y = \ln(\mu_X - x_0) - \frac{1}{2}\sigma_Y^2 = 2.456$$

Hence with (4.37) one obtains for the quantile x_T:

$$x_T = x_0 + \exp(\mu_Y + z_T \sigma_Y) = 7.13 + \exp(2.456 + 1.282 \text{x} 0.444) = 7.13 + 20.60 = 27.7$$

To solve the second problem, use is made of equation (4.36). The normal variate y is derived from the LN-3 variate x = 35 and x_0:

$$y = \ln(x - x_0) = \ln(35 - 7.130) = 3.328$$
$$z = \frac{y - \mu_Y}{\sigma_Y} = \frac{3.328 - 2.456}{0.444} = 1.963$$

Since z is a standard normal variate, the non-exceedance probability attached to Z is found from the standard normal distribution:

$$F_Z(z) = P(Z \le 1.963) = 0.975$$
$$P(Z > 1.963) = 1 - 0.975 = 0.025$$
$$T = \frac{1}{P(Z > 1.963)} = \frac{1}{0.025} = 40$$

Application in hydrology

The lognormal distribution function finds wide application in hydrology. It is generally appropriate to fit monthly rainfall and runoff series, whereas quite often also annual maximum discharge series can be modelled by the lognormal distribution.

4.4.3 Box-Cox transformation

Transformation equations

Box and Cox (1964) describe a general transformation of the following form:

$$Y = \frac{(X - x_0)^{\lambda} - 1}{\lambda} \text{for} : \lambda \neq 0 \qquad (4.38)$$
$$Y = \ln(X - x_0) \text{for} : \lambda = 0$$

The transformed variable Y has, by approximation, a normal distribution $N(\mu_Y, \sigma_Y)$. The transformation is seen to have two parameters, a **location** or shift parameter x_0 and the **power** and **scale** parameter λ.

The **reduced variate** Z, defined by:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \qquad (4.39)$$

with Y defined by (4.38) has a standard normal distribution. Once x_0 and λ are known, with the inverse of (4.39) and (4.38) the quantiles can be derived from the standard normal distribution.

Quantiles

For a particular return period T it follows for **quantile** x_T:

$$\left.\begin{aligned} x_T &= x_0 + \left(1 + \lambda(\mu_Y + z_T\sigma_Y)\right)^{1/\lambda} \quad \text{for} : \lambda \neq 0 \\ x_T &= x_0 + \exp(\mu_Y + z_T\sigma_Y) \quad \text{for} : \lambda = 0 \end{aligned}\right\} \qquad (4.40)$$

It is noted that for very extreme values this transformation should not be used in view of the normality by approximation. In HYMOS flexibility is added by considering $|X-x_0|$ instead of $(X-x_0)$.

Application of the transformation shows that it returns a transformed series Y with a skewness close to zero and a kurtosis near 3.

Example 4.7 Box-Cox transformation

An example of its application is given below for annual maximum rainfall for Denee (Belgium), period 1882-1993.

Statistics before Box-Cox transformation	
Number of data	112
Mean	37.0
Standard deviation	11.8
Skewness	1.23
Kurtosis	4.56
Statistics after Box-Cox transformation with x_0 = 15.0 and λ = 0.142	
Number of data	112
Mean	3.70
Standard deviation	0.81
Skewness	0.00
Kurtosis	3.05

Table 4.1: Results of Box-Cox transformation on annual maximum rainfall

From the result it is observed that the skewness and kurtosis of the transformed variable are indeed close to 0 and 3. On the other hand λ is seen to be very small. It implies that the normal variates will be raised to a very high power to arrive at the quantiles, which is rather unfortunate. In such a case a lognormal distribution would be more appropriate.

4.5 Gamma or Pearson related distributions

4.5.1 Exponential distribution

Probability density and cumulative frequency distribution

In Sub-section 4.2.2 the exponential distribution was derived from the Poisson distribution. The exponential distribution models the distribution of the waiting time between successive events of a Poisson process. The exponential distribution is a special case of the gamma or Pearson Type 3 distribution (see next sub-sections). The general form of the exponential distribution is given by:

$$f_X(x) = \frac{1}{\beta}\exp(-\frac{x - x_0}{\beta}) \quad \text{for : } x > x_0 \tag{4.41}$$

and the cdf reads:

$$F_X(x) = \frac{1}{\beta}\int_{x_0}^{x} \exp(-\frac{s - x_0}{\beta})ds = 1 - \exp(-\frac{x - x_0}{\beta}) \tag{4.42}$$

The distribution is seen to have 2 parameters x_0 and β and will therefore be denoted by **E-2**. With x_0 = 0 it reduces to 1-parameter exponential distribution **E-1**.

Standardised distribution

Introducing the **reduced variate** Z:

$$Z = \frac{X - x_0}{\beta} \tag{4.43}$$

it is observed that Z = X if x_0 = 0 and β = 1, hence the **standardised exponential pdf** becomes:

$$f_Z(z) = \exp(-z) \tag{4.44}$$

and the **standardised exponential cdf** is given by:

$$F_Z(z) = 1 - \exp(-z) \tag{4.45}$$

Replacing Z in (4.45) by (4.43) equation (4.42) is seen to be obtained, and differentiating the cdf with respect to X gives pdf (4.41).

Moment related distribution parameters

The moment related parameters are given by:

$$\left.\begin{aligned} \mu_X &= x_0 + \beta \\ \sigma_X^2 &= \beta^2 \\ \gamma_{1,X} &= 2 \end{aligned}\right\} \tag{4.46}$$

It is observed that the distribution parameter x_0 is a **location** parameter as it affects only the first moment of the distribution. The parameter β is a **scale** parameter as it scales variate X. The skewness of the distribution is fixed. The distribution is shown in Figure 4.13.

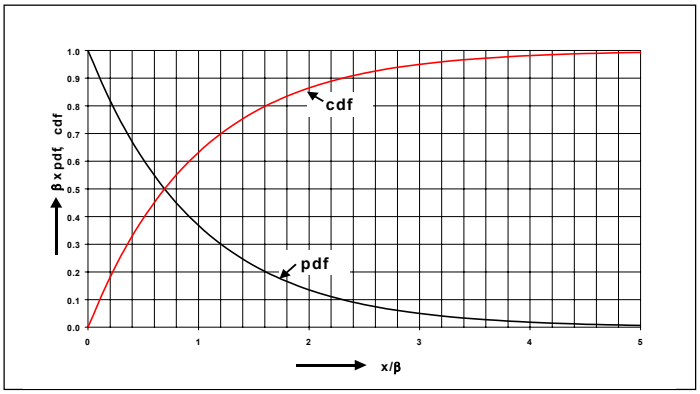


Figure 4.13:
Exponential distribution as function of the reduced variate
$(x-x_0)$

From (4.46) it follows for the mean, variance and skewness of the standardised gamma function (x_0 = 0, β = 1) respectively 1, 1 and 2.

Distribution parameters expressed in moment related parameters

From (4.46) it follows for the distribution parameters as function of the moments:

$$\beta = \sigma_X \tag{4.47}$$

$$x_0 = \mu_X - \sigma_X \tag{4.48}$$

If x_0 = 0 the distribution reduces to 1-parameter exponential distribution **E-1**. Then the mean and the standard deviation are seen to be identical. Note also that with x_0 = 0 and λ = 1/β substituted in (4.42) equation (4.11) is obtained.

Quantiles

The values of X and Z for which $F_X(x) = F_Z(z)$ are related by (4.43). Using the inverse the quantiles x_T are obtained from the reduced variate z_T for a specified return period T:

$$x_T = x_0 + \beta z_T \tag{4.49}$$

The quantile x_T can also directly be obtained from the first two moments and T:

$$x_T = \mu_X + \sigma_X\left(\ln(T) - 1\right) \tag{4.50}$$

Example 4.8: Exponential distribution

A variate X is exponentially distributed with mean 50 and standard deviation 20. Determine:

- the value of X, which corresponds with a non-exceedance probability of 0.95.
- the probability that $50 \leq X \leq 75$.

Note that since $\mu_X \neq \sigma_X$ the exponential distribution is E-2.The non-exceedance probability implies an exceedance probability of 1 – 0.95 = 0.05, hence the return period T is 1/0.05 = 20. From (4.50) the variate value for this return period becomes:

$X_T = 50 + 20 \times \{\ln(20) - 1\} = 50 + 20 \times (3.0 - 1) = 90$

To solve the second problem equation (4.42) is used, which requires the parameters x_0 and β to be available. From (4.47) one gets $\beta = \sigma_X = 20$ and from (4.48) $x_0 = \mu_X - \beta = 30$, hence:

$$P\{50 \leq X \leq 75\} = F_X(75) - F_X(50) =$$
$$= 1 - \exp(-\frac{75-30}{20}) - (1 - \exp(-\frac{50-30}{20}) = 0.895 - 0.632 = 0.263$$

Application in hydrology

The exponential distribution finds wide application. In engineering one applies the distribution to model time to failure, inter-arrival time, etc. In hydrology the distribution is a.o. applied to model time between flood peaks exceeding a threshold value. Furthermore, the distribution models a process, where the outcomes are independent of past occurrences, i.e. the process is **memory-less**.

4.5.2 Gamma distribution

Definition

The distribution of the sum of k exponentially distributed random variables each with parameter β (equation (4.41) with $x_0 = 0$) results in a gamma distribution with parameter k and β. The gamma distribution describes the waiting time till the k^{th} exceedance and is readily derived from the Poisson distribution (like the exponential) by multiplying the probability of having (k-1) arrivals till t, described by equation (4.7), and the arrival rate (λ=1/β) at t, leading to the Erlang distribution. Since k does not need to be an integer it is replaced by the positive real γ, and a gamma distribution with two parameters γ and β is obtained, shortly denoted by G-2.

Probability density and distribution function

The **gamma** pdf has the following form:

$$f_X(x) = \frac{\left(\frac{x}{\beta}\right)^{\gamma-1} \exp(-\left(\frac{x}{\beta}\right))}{\beta\Gamma(\gamma)} \text{ with : } x > 0; \beta > 0; \gamma > 0 \qquad (4.51)$$

and the cdf reads:

$$F_X(x) = \frac{1}{\beta\Gamma(\gamma)} \int_0^x \left(\frac{s}{\beta}\right)^{\gamma-1} \exp(-\left(\frac{s}{\beta}\right)) ds \text{ for : } \beta > 0; \gamma > 0 \qquad (4.52)$$

Standardised gamma distribution

Introducing the **reduced gamma variate** Z, defined by:

$$Z = \frac{X}{\beta} \qquad (4.53)$$

it is observed that Z = X for β = 1 and the pdf and cdf of the **standardised gamma distribution** then read:

$$f_Z(z) = \frac{z^{\gamma-1}\exp(-z)}{\Gamma(\gamma)} \tag{4.54}$$

$$F_Z(z) = \frac{1}{\Gamma(\gamma)}\int_0^z t^{\gamma-1}\exp(-t)dt \tag{4.55}$$

Note that by substituting (4.53) in (4.55) and with dx = β dz equation (4.52) is obtained, and by differentiating the cdf with respect to X the pdf equation (4.51) follows.

Gamma function

Equation (4.55) is called the **incomplete gamma function ratio**. The **complete** (standard) gamma function $\Gamma(\gamma)$, needed to get area = 1 under the pdf curve, is defined by:

$$\Gamma(\gamma) = \int_0^\infty t^{\gamma-1}\exp(-t)dt \tag{4.56}$$

The **gamma function** provides a continuous alternative for discrete factorials. The function has the following properties:

$$\Gamma(n+1) = n! \tag{4.57}$$

And hence:

$$\Gamma(n+1) = n\Gamma(n) \text{ for : } n = 0,1,2,... \text{ with : } 0! = 1 \tag{4.58}$$

Furthermore:

$$\begin{aligned} &\Gamma(0) = \infty \\ &\Gamma(1/2) = \sqrt{\pi} \\ &\Gamma(1) = \Gamma(2) = 1 \\ &0.88560 \le \Gamma(y) \le 1 \text{ for : } 1 \le \gamma \le 2 \end{aligned} \tag{4.59}$$

The gamma function is tabulated for values of γ: $1 \le \gamma \le 2$. In HYMOS the complete gamma function is computed in two steps:

- first γ is reduced to a value between 1 and 2 using the recursive equation (4.58):

$\Gamma(\gamma - 1) = \Gamma(\gamma)/\gamma$ for $\gamma < 1$ or: $\Gamma(\gamma + 1) = \gamma\Gamma(\gamma)$ for $\gamma > 2$, and then

- secondly, a third order interpolation procedure is used to obtain a value from the basic gamma function table.

Example 4.9 Gamma function

Derive the gamma function values for γ = 3.2 and 0.6.

Procedure:

γ = 3.2, then $\Gamma(3.2) = 2.2\Gamma(2.2) = 1.2x2.2\Gamma(1.2) = 1.2x2.2x0.9182 = 2.424$

γ = 0.6, then $\Gamma(0.6) = \Gamma(1.6)/0.6 = 0.8935/0.6 = 1.489$

Note that the values for $\Gamma(1.2)$ and $\Gamma(1.6)$ are obtained from the basic gamma function table.

The computational procedure for the **incomplete** gamma function as used in HYMOS is presented in Annex A4.3 and A4.4 for its inverse.

Moment related parameters of the distribution

The mean, mode, variance, skewness and kurtosis of the gamma distribution read:

$$\left.\begin{aligned} \mu_X &= \beta\gamma \\ m_X &= \beta(\gamma - 1) \\ \sigma_X^2 &= \beta^2\gamma \end{aligned}\right\} \qquad (4.60a)$$

$$\left.\begin{aligned} \gamma_{1,X} &= \frac{2}{\sqrt{\gamma}} \\ \gamma_{2,X} &= \frac{3(\gamma+2)}{\gamma} \end{aligned}\right\} \qquad (4.60b)$$

From (4.53) it is observed that β is a **scale** parameter and from (4.60b) γ is a **shape** parameter. This is also seen from Figures 4.14 to 4.16. Comparison of (4.60a) with (4.46) with $x_0 = 0$ shows that the mean and the variance of the gamma distribution is indeed γ-times the mean and the variance of the exponential distribution. This supports the statement that the gamma distribution is the distribution of the sum of γ exponentially distributed random variables. Note that for large γ the skewness tends to zero and kurtosis to 3 and hence the gamma distribution approaches the normal distribution. Note that the mode $m_X > 0$ for $\gamma > 1$ and the distribution is single peaked. If $\gamma \leq 1$ the pdf has a reversed J-shape.

From (4.60a) it is also observed that with β = 1 the mean and the variance of the standardised gamma distribution are both equal to γ; the skewness and kurtosis are as in (4.60b).

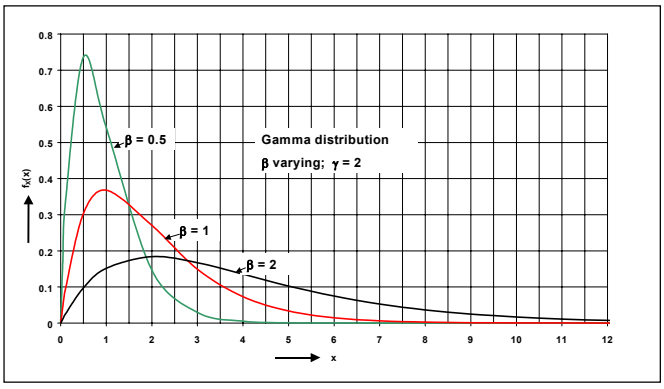


Figure 4.14:
Gamma distribution effect of scale parameter β

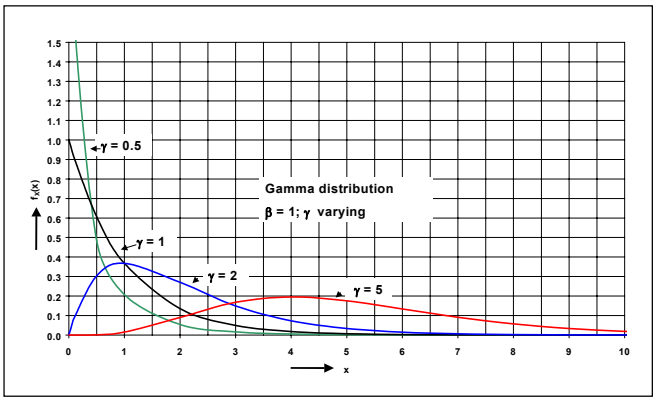


Figure 4.15:
Gamma distribution effect of shape parameter γ

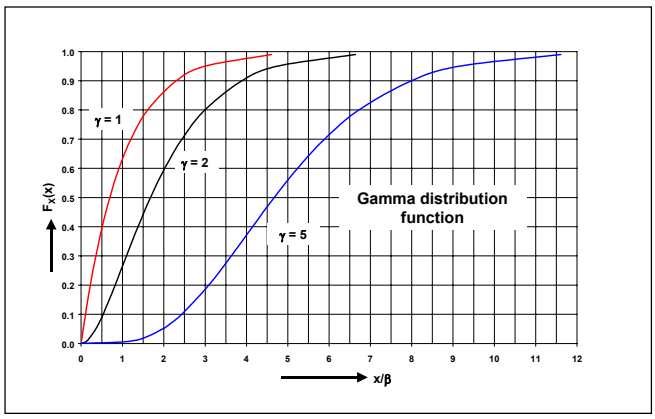


Figure 4.16:
Gamma cdf's

Distribution parameters expressed in moment related parameters

From (4.60a) it follows for the gamma parameters β and γ:

$$\beta = \frac{\sigma_X^2}{\mu_X} \tag{4.61}$$

$$\gamma = \left(\frac{\mu_X}{\sigma_X}\right)^2 = \frac{1}{C_{v,X}^2} \tag{4.62}$$

Hence, by the mean and the standard deviation the distribution parameters are fully determined. From a comparison of (4.62) with (4.60b) it is observed that for the gamma distribution there is a fixed relation between the coefficient of variation and the skewness. It follows:

$$\gamma_{1,X} = 2C_{v,X} \tag{4.63}$$

It implies that from a simple comparison of the coefficient of variation with the skewness a first impression can be obtained about the suitability of the 2-parameter gamma distribution to model the observed frequency distribution. As will be shown in the next sub-section more flexibility is obtained by adding a location parameter to the distribution.

Quantiles of the gamma distribution

The quantiles x_T of the gamma distribution are derived from the inverse of the standard incomplete gamma function and the reduced variate z_T:

$$x_T = \beta z_T \tag{4.64}$$

The required parameters γ for the standard incomplete gamma function and β to transform the standardised variate z_T into x_T can be obtained from equations (4.61) and (4.62) or some other parameter estimation method.

4.5.3 Chi-squared and gamma distribution

Probability density and cumulative distribution function

By putting $\beta = 2$ and $\gamma = \nu/2$ the gamma distribution becomes the Chi-squared distribution:

$$f_X(x) = \frac{1}{2\Gamma(\nu/2)}\left(\frac{x}{2}\right)^{\nu/2-1} \exp(-\frac{x}{2}) \text{ for : } x \geq 0, \nu > 0 \tag{4.65}$$

$$F_X(x) = \frac{1}{2\Gamma(\nu/2)}\int_0^x \left(\frac{s}{2}\right)^{\nu/2-1} \exp(-\frac{s}{2})ds \tag{4.66}$$

The parameter ν is the number of degrees of freedom. The chi-square distribution is the distribution of the sum of ν squared normally distributed random variables N(0, 1) and find wide application in variance testing and goodness of fit testing of observed to theoretical distributions. It also follows, that the sum of 2 squared standard normal variables has an exponential distribution.

4.5.4 Pearson type 3 distribution

Probability density and cumulative distribution function

By introducing a **location** parameter x_0 in the gamma distribution, discussed in the previous sub-section, a **Pearson type 3** distribution is obtained, shortly denoted by **P-3**. This distribution is sometimes also called a **3-parameter gamma** distribution or **G-3**. Its pdf has the following form:

$$f_X(x) = \frac{\left(\frac{x-x_0}{\beta}\right)^{\gamma-1} \exp\left(-\left(\frac{x-x_0}{\beta}\right)\right)}{\beta\Gamma(\gamma)} \text{ with : } x > x_0 \text{ ; } \beta > 0 \text{ ; } \gamma > 0 \tag{4.67}$$

and the cdf reads:

$$F_X(x) = \frac{1}{\beta\Gamma(\gamma)}\int_{x_0}^x \left(\frac{s-x_0}{\beta}\right)^{\gamma-1} \exp\left(-\left(\frac{s-x_0}{\beta}\right)\right)ds \text{ for : } \beta > 0 \text{ ; } \gamma > 0 \tag{4.68}$$

The reduced Pearson Type 3 variate Z, is defined by:

$$Z = \frac{X - x_0}{\beta} \tag{4.69}$$

It is observed that Z = X for $x_0 = 0$ and $\beta = 1$. Introducing this into (4.67) and (4.68) leads to the standardised gamma distributions presented in equations (4.54) and (4.55).

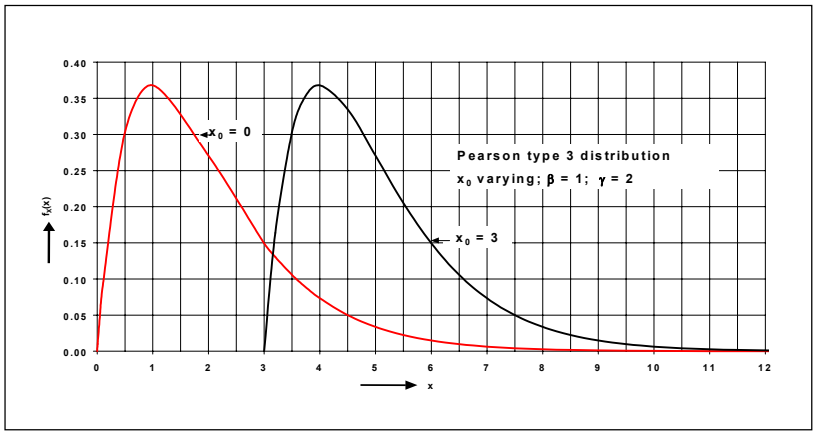
Moment related parameters of the distribution

The mean, mode, variance, skewness and kurtosis of the P-3 distribution read:

$$\left.\begin{aligned} \mu_X &= x_0 + \beta\gamma \\ m_X &= x_0 + \beta(\gamma - 1) \\ \sigma_X^2 &= \beta^2\gamma \end{aligned}\right\} \qquad (4.70a)$$

$$\left.\begin{aligned} \gamma_{1,X} &= \frac{2}{\sqrt{\gamma}} \\ \gamma_{2,X} &= \frac{3(\gamma+2)}{\gamma} \end{aligned}\right\} \qquad (4.70b)$$

It is observed that x_0 is a **location** parameter as it affects only the first moment of the distribution about the origin. This is also seen from Figures 4.17. As for the (2-parameter) gamma distribution β is a **scale** parameter and γ is a **shape** parameter. Also, for large γ the distribution becomes normal. The mode of the distribution is at $x_0+\beta(\gamma-1)$, for $\gamma > 1$ and the distribution is unimodal. For $\gamma \leq 1$ the distribution is J-shaped similar to the gamma distribution, with its maximum at x_0.



Distribution parameters expressed in moment related parameters

The parameters of the Pearson Type 3 distribution can be expressed in the mean, standard deviation and skewness as follows:

$$\gamma = \left(\frac{2}{\gamma_{1,X}}\right)^2 \qquad (4.71)$$

$$\beta = \frac{1}{2}\sigma_X\gamma_{1,X} \qquad (4.72)$$

$$x_0 = \mu_X - 2\frac{\sigma_X}{\gamma_{1,X}} \qquad (4.73)$$

From the last expression it is observed that:

$$\gamma_{1,X} = 2\left(\frac{\sigma_X}{\mu_X - x_0}\right) \qquad (4.74)$$

The term within brackets can be seen as an adjusted coefficient of variation, and then the similarity with Equation (4.63) is observed.

Moment generating function

The moments of the distribution are easily obtained from the moment generating function:

$$G(s) = E[\exp((sx)] = \int_{x_0}^{\infty} \exp(sx) \frac{\left(\frac{x - x_0}{\beta}\right)^{\gamma-1} \exp\left(-\left(\frac{x - x_0}{\beta}\right)\right)}{\beta\Gamma(\gamma)} dx \tag{4.75}$$

Or introducing the reduced variate $Z = (x-x_0)/\beta$, and $dx = \beta\, dz$:

$$G(s) = \exp(sx_0) \int_0^{\infty} \frac{z^{\gamma-1} \exp(-z(1-s\beta)}{\Gamma(\gamma)} dz$$

Introducing further: $u = z(1-s\beta)$, or $z = u/(1-s\beta)$ and $dz = 1/(1-s\beta)du$, it follows:

$$\begin{aligned} G(s) &= \exp(sx_0)(1-s\beta)^{-\gamma} \int_0^{\infty} \frac{u^{\gamma-1} \exp(-u)}{\Gamma(\gamma)} du = \\ &= \exp(sx_0)(1-s\beta)^{-\gamma} \end{aligned} \tag{4.76}$$

By taking the derivatives of G(s) with respect to s at s = 0 the moments about the origin can be obtained:

$$\frac{dG(0)}{ds} = \mu_1' = \exp(sx_0)\{x_0(1-s\beta)^{-\gamma} + \beta\gamma(1-s\beta)^{-(\gamma+1)}\}\Big|_{s=0}$$

$$\text{so: } \mu_1' = \mu_X = x_0 + \beta\gamma$$

Since for the computation of the central moments the location parameter is of no importance, the moment generating function can be simplified with $x_0 = 0$ to:

$$\left.\begin{aligned} &G(s) = (1-s\beta)^{-\lambda} \\ &\text{hence :} \\ &\frac{dG(0)}{ds} = \beta\gamma(1-s\beta)^{-(\gamma+1)}\Big|_{s=0} \rightarrow \mu_1'\Big|_{x_0=0} = \beta\gamma \\ &\frac{d^2G(0)}{ds^2} = \beta^2\gamma(\gamma+1)(1-s\beta)^{-(\gamma+2)}\Big|_{s=0} \rightarrow \mu_2' = \beta^2\gamma(\gamma+1) \\ &\text{etc.} \end{aligned}\right\} \tag{4.78}$$

Using equation (3.30) the central moments can be derived from the above moments about the origin.

Quantiles

The quantile x_T of the gamma distribution follows from the inverse of the standard incomplete gamma function z_T and (4.67):

$$x_T = x_0 + \beta z_T \tag{4.79}$$

Example 4.10: Gamma distribution

The mean, standard deviation and skewness of a P-3 variate are respectively 50, 20 and 1.2. Required is the variate value at a return period of 100.

First, the parameters of the P-3 distribution are determined from (4.71) – (4.73). It follows:

$$\gamma = \left(\frac{2}{\gamma_{1,X}}\right)^2 = \left(\frac{2}{1.2}\right)^2 = 2.78$$

$$\beta = \frac{\sigma_X}{\sqrt{\gamma}} = \frac{\gamma_{1,X}\sigma_X}{2} = \frac{1.2x20}{2} = 12$$

$$x_0 = \mu_X - \beta\gamma = 50 - 12x2.78 = 16.67$$

From the standard incomplete gamma function with γ = 2.78 it follows that $z_T = z_{100}$ = 8.03. Then from (4.77) it follows for $x_T = x_{100}$:

$$x_T = x_0 + \beta z_T = 16.67 + 12x8.03 = 113$$

Note that the standardised gamma variate can also be obtained from the tables of the chi-squared distribution for distinct non-exceedance probabilities. Since $\gamma = \nu/2$ it follows $\nu = 2\gamma = 2 \times 2.78 = 5.56$. From the χ^2 - tables one gets for T = 100 or p = 0.99 a χ^2 – value by interpolation between ν = 5 and ν = 6 of 16.052. For the chi-squared distribution β = 2, so: $\chi_T^2 = \beta z_T$ or z_T or $z_T = \chi_T^2/\beta$ = 16.052/2 = 8.03. The values can of course also directly be obtained via the "Statistical Tables" option in HYMOS under "Analysis".

Related distributions

For specific choices of the parameters x_0, β and γ, a number of distribution functions are included in the Pearson Type 3 or 3-parameter gamma distribution, see Tables 4.2 and 4.3.

The moment related parameters of these distributions are summarised in Table 4.3. By considering the logarithm of the variate or by raising the reduced variate Z of (4.69) to a power k further distributions like Weibull and Rayleigh distributions can be defined as presented in Sub-section 4.1 Those are discussed in the next sub-sections.

Pearson Type 3 or 3-parameter gamma (x_0,β,γ)	γ = 1: exponential	x_0 = 0: 1-par. exponential
		$x_0 \neq 0$: 2-par. exponential
	x_0 = 0: gamma	β = 1: 1-par gamma
		$\beta \neq 1$: 2-par gamma
		β = 2, $\gamma = \nu/2$: chi-squared

Table 4.2: Summary of related distributions

distribution	mean	mode	Variance	Skewness	kurtosis	Standardised variate z
1-par. exponential	β	-	β^2	2	9	$z=x/\beta$
2-par. exponential	$x_0+\beta$	-	β^2	2	9	$z=(x-x_0)/\beta$
1-par. gamma	γ	γ-1, γ>1	γ	$2/\sqrt{\gamma}$	$3(\gamma+2)/\gamma$	z=x
2-par. gamma	$\beta\gamma$	$\beta(\gamma-1)$	$\beta^2\gamma$	$2/\sqrt{\gamma}$	$3(\gamma+2)/\gamma$	$z=x/\beta$
3-par. Gamma or P-3	$x_0+\beta\gamma$	$x_0+\beta(\gamma-1)$	$\beta^2\gamma$	$2/\sqrt{\gamma}$	$3(\gamma+2)/\gamma$	$z=(x-x_0)/\beta$
Chi-squared	ν	ν-2, ν>2	2ν	$2^{3/2}/\sqrt{\nu}$	$3(\nu+4)/\nu$	z=x/2

Table 4.3: Moment related parameters of the exponential and gamma family of distributions

4.5.5 Log-Pearson Type 3 distribution

Probability density function

When $Y = \ln(X - x_0)$ follows a Pearson Type 3 distribution then $(X - x_0)$ is log-Pearson Type 3 distributed. Its pdf is given by:

$$f_X(x) = \frac{1}{\beta(x-x_0)\Gamma(\gamma)}\left(\frac{\ln(x-x_0)-y_0}{\beta}\right)^{\gamma-1}\exp\left(-\left(\frac{\ln(x-x_0)-y_0}{\beta}\right)\right) \text{ for : } \ln(x-x_0) > y_0 \qquad (4.82)$$

The log-Pearson Type 3distribution finds application in hydrology particularly for strongly positively skewed annual flood peaks. The skewness is reduced by a logarithmic transformation, to arrive at a Pearson type III distribution. In the USA the log-Pearson type III is the standard for modelling annual maximum floods (Water Resources Council, 1976). All relations presented in the previous sub-section are valid for $\ln(X-x_0)$.

Quantiles of LP-3

The quantiles x_T of the LP-3 distribution are obtained from the inverse of the standard incomplete gamma function leading to z_T and (4.81):

$$x_T = x_0 + \exp(y_0 + \beta z_T) \qquad (4.81)$$

4.5.6 Weibull distribution

Probability density and cumulative distribution function

With γ = 1 equation (4.55) reduces to:

$$F(z) = \int_0^z \exp(-s)ds \text{ with : } z = \left(\frac{x-x_0}{\beta}\right)^k \text{ so : } dz = \frac{k}{\beta}\left(\frac{x-x_0}{\beta}\right)^{k-1} dx$$

it follows for the pdf and cdf of the Weibull distribution:

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{k}{\beta}\left(\frac{x - x_0}{\beta}\right)^{k-1} \exp\left(-\left(\frac{x - x_0}{\beta}\right)^k\right) \quad \text{for : } x \geq x_0, k > 0, \beta > 0 \tag{4.82}$$

$$F_X(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta}\right)^k\right) \tag{4.83}$$

Note that for k = 1 the Weibull distribution reduces to an exponential distribution.

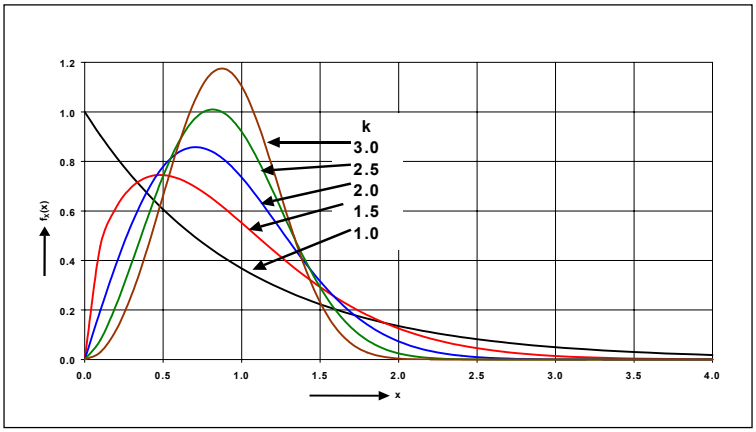
Moment related parameters of the distribution

The mean, mode, variance and skewness of the Weibull distribution read:

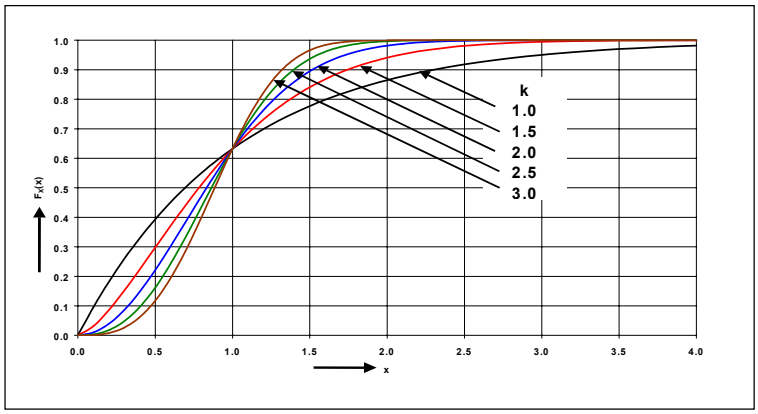
$$\left.\begin{aligned} \mu_X &= x_0 + \beta\Gamma\left(1 + \frac{1}{k}\right) \\ M_X &= x_0 + \beta(\ln 2)^{1/k} \\ m_X &= x_0 + \beta\left(\frac{k-1}{k}\right)^{1/k} \\ \sigma_X^2 &= \beta^2\left\{\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right\} \end{aligned}\right\} \tag{4.84a}$$

$$\gamma_{1,X} = \frac{\Gamma(1 + \frac{3}{k}) - 3\Gamma(1 + \frac{2}{k})\Gamma(1 + \frac{1}{k}) + 2\Gamma^3(1 + \frac{1}{k})}{\left(\Gamma(1 + \frac{2}{k}) - \Gamma^2(1 + \frac{1}{k})\right)^{3/2}} \tag{4.84b}$$

The distribution is seen to have 3 parameters: x_0 is a **location** parameter, β a **scale** parameter and k is a **shape** parameter. For k > 1 the pdf is seen to be unimodal, see also Figure 4.19.



(a)



(b)

Figure 4.19 (a) and 4.19 (b): Weibull distribution for various values of k(x_0 = 0 and β = 1)

The expression for the skewness as a function of k is rather complicated and has therefore been visualised in Figure 4.20. From the Figure it is observed that for k < 1 the skewness increases rapidly to very high values. In practice the region 1< k < 3 is mostly of interest. Note that for k > 3.5 the skewness becomes slightly negative.

Note also that above expressions for the mean, variance and skewness can easily be derived from the moment generating function. For x_0 = 0 the r^{th} moment about the origin becomes:

$$\mu_r' = \beta^r \Gamma(1+\frac{r}{k}) \qquad (4.85)$$

Subsequently, equation (3.30) is used to obtain the central moments. For the mean x_0 has to be added.

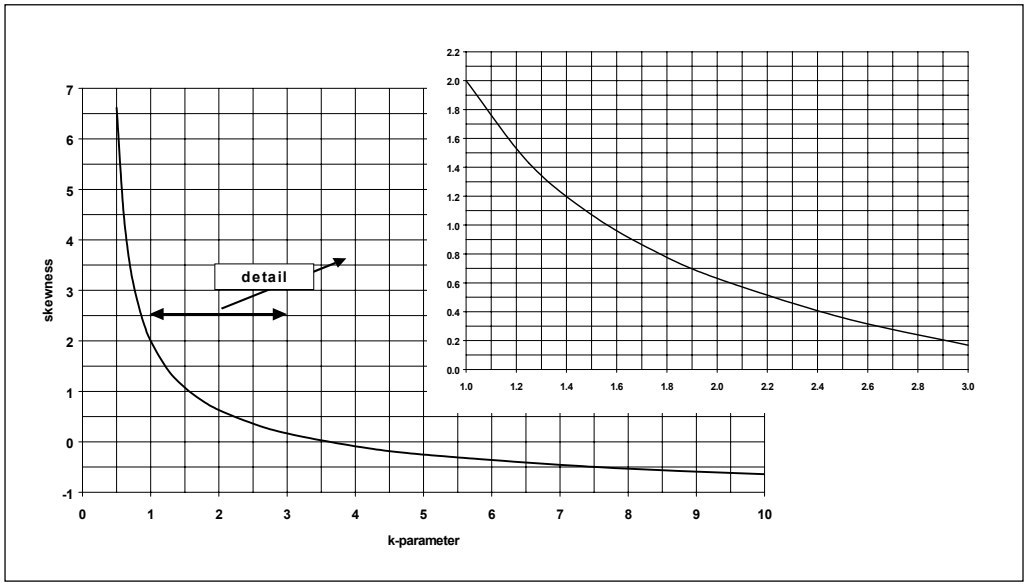


Figure 4.20: Skewness of W-3 as function of k

Quantiles of W-3

From (4.83) the quantile of the Weibull distribution is easily determined. For a given return period T it follows for x_T:

$$x_T = x_0 + \beta(\ln T)^{1/k} \tag{4.86}$$

From (4.86) it is observed that for given x_0, β and T values x_T decreases with increasing k.

The Weibull distribution is often used to model the frequency distribution of wind speed and flow extremes (minimum and maximum). It is one of the asymptotic distributions of the general extreme value theory, to be discussed in the next sub-section.

4.5.7 Rayleigh distribution

Probability density and cumulative distribution function

From the Weibull distribution with k = 2 the Rayleigh distribution is obtained. Its pdf and cdf read:

$$f_X(x) = \frac{2}{\beta}\left(\frac{x - x_0}{\beta}\right)\exp\left(-\left(\frac{x - x_0}{\beta}\right)^2\right) \tag{4.87}$$

$$F_X(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta}\right)^2\right) \tag{4.88}$$

Moment related parameters of the distribution

From (4.84) the mean, mode, variance and skewness are given by:

$$\left.\begin{aligned} \mu_X &= x_0 + \Gamma(1.5)\beta = x_0 + 0.88623\beta \\ m_X &= x_0 + \frac{1}{2}\sqrt{2}\beta = x_0 + 0.70711\beta \\ \sigma_X^2 &= (1 - \Gamma^2(1.5))\beta = 0.21460\beta^2 \\ \gamma_{1,X} &= \frac{\Gamma(1.5)\left\{2\Gamma^2(1.5) - 1.5\right\}}{\left\{1 - \Gamma^2(1.5)\right\}^{3/2}} = 0.631 \end{aligned}\right\} \tag{4.89}$$

The distribution is seen to have **location** parameter x_0 and a **scale** parameter β. The skewness of the distribution is fixed. The pdf and cdf of the Rayleigh distribution are shown in Figure 4.21.

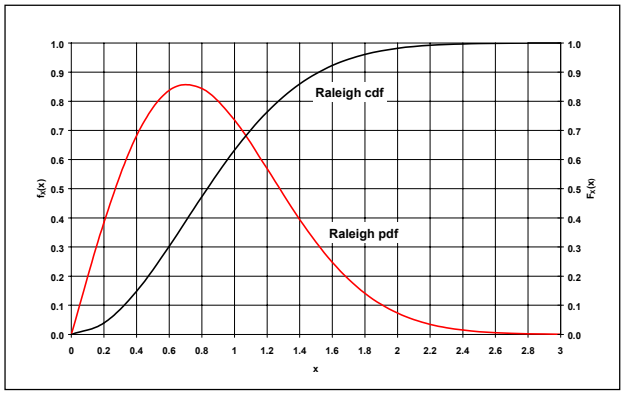


Figure 4.21:
Rayleigh distribution

The distribution parameters are easily related to the mean and standard deviation of the Rayleigh variate X:

$$\beta = 2.15866\sigma_X \tag{4.90}$$

$$x_0 = \mu_X - 1.91307\sigma_X \tag{4.91}$$

Quantiles of R-2

The quantiles x_T of the Rayleigh distribution for a return period T follow from (4.88):

$$x_T = x_0 + \beta\sqrt{\ln T} \tag{4.92}$$

The Rayleigh distribution is suitable to model frequency distributions of wind speed and of annual flood peaks in particular.

4.6 Extreme value distributions

4.6.1 Introduction

A number of distribution functions are available specially suited to model frequency distributions of extreme values, i.e. either largest values or smallest values. These can be divided in two groups:

General extreme value distributions GEV, or EV-1, EV-2 and EV-3, and

1. Generalised Pareto distributions, also with 3 types, P-1, P-2 and P-3.

The GEV distributions and the generalised Pareto distributions are related. The first group is generally applicable to annual maximum or annual minimum series, whereas the Pareto distributions are often used to model exceedance series, i.e. peaks exceeding a threshold value. Though any of the distributions may be applied to any of the series of extremes. There is however a distinct difference in the interpretation of the return period between extremes in a fixed interval and extremes exceeding a threshold, though both methods are related.

It is noted that instead of the extreme value distributions also the distributions dealt with in the previous sections may be applied to model the distribution to extremes.

Note further that statistical distributions are generally used far beyond the observed frequency range. It is noted, though, that the use of statistical distributions for extrapolation purposes is strongly limited by physical features and limitations in sources and basins, neither included in the distribution or in the data used to fit the distribution. The main difficulty is with the assumption of the independent identically distributed random variable ('iidrv') and the invariability of the distribution with time. In this respect, you are strongly advised to read the paper by V. Klemes entitled: 'Tall tales about tails of hydrological distributions' in Journal of Hydrologic Engineering, Vol 5, No 3, July 2000, pages 227 – 239. As an example consider the routing of a design storm through a channel reach. The design storms for different return periods are determined using the procedures proposed by NERC (1975). The design storms are routed through a channel reach with an inbank capacity of 350 m^3/s. Beyond that discharge level part of the flow is transferred through the floodplain. The exceedance of the inbank capacity occurs on average once in 30 years. Two types of flood plains are considered: a narrow one and a wide one. The effect of the two types of flood plains on the behaviour of the distribution function of the flood peaks, observed at the downstream end of the reach, is shown in Figure 4.22.

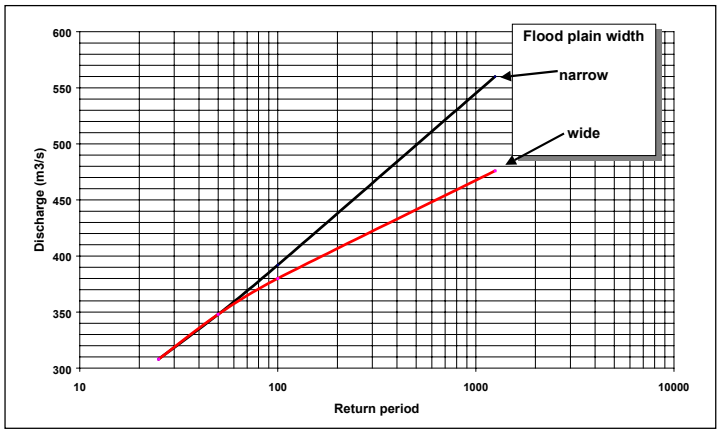


Figure 4.22:
Extreme value distribution of routed design storms

From Figure 4.22 it is observed that the frequency distribution is strongly affected by physical features of the river, which affect discharges of various magnitudes differently. It implies that data points gathered for the more frequent extreme events may include no information for the rare extreme events. Hence the validity of extrapolation beyond the measured range, no matter how scientific and/or complex the mathematical expressions may be, remains highly questionable. It should always be verified whether physical limitations and behaviour under very wet or very dry conditions may affect the extreme events. Blind application of extreme value distributions is always wrong.

The use of confidence bands about the frequency distribution will not help you much, as those are based on the assumption that the used distribution is applicable to the considered case. If the distribution is not applicable, the confidence limits will give a completely false picture of the uncertainty in the extreme value for a particular return period. Also, the use of goodness of fit tests will not help you in this respect and may lead you to an unjustified believe in the applicability of the distribution.

4.6.2 General extreme value distributions

The general extreme value distributions are applicable to series with a fixed interval like annual maximum or annual minimum series; i.e. one value per interval. Consider the extreme values (largest X_{max} and smallest X_{min}) of a sample of size n. Hence, $X_{max} = \max(X_1, X_2, \ldots, X_n)$ and let the X_i's be **independent** and **identically distributed**, then:

$$F_{X\max}(x) = P\left(X_1 \le x \cap X_2 \le x \cap \ldots\ldots \cap X_n \le x\right) = \prod_{i=1}^{n} F_{X_i}(x) = \left(F_X(x)\right)^n \qquad (4.95)$$

Note that the third expression stems from the independence of the X_i's, whereas the fourth expression is due to the identical distribution of the X_i's. The pdf of X_{max} reads:

$$f_{X\max}(x) = n(F_X(x))^{n-1} f_X(x) \tag{4.96}$$

Similarly for X_{min} = min(X_1, X_2, …., X_n) it follows under the same assumptions of independence and identical distribution:

$$F_{X\min}(x) = 1 - P(X_1 > x \cap X_2 > x \cap ..... \cap X_n > x) = 1 - \prod_{i=1}^{n}(1 - F_{X_i}(x)) = 1 - (1 - F_X(x))^n \tag{4.97}$$

and the pdf of X_{min}:

$$f_{X\min}(x) = n(1 - F_X(x))^{n-1} f_X(x) \tag{4.98}$$

Above expressions for X_{max} and X_{min} show that their distributions depend on sample size and the parent distribution from which the sample is taken. However, it can be shown, that full details about the parent distribution are not required to arrive at the distribution of extremes. For large n and limited assumptions about the parent distributions three types of asymptotic distributions for extreme values have been developed:

1. **Type I**: parent distribution is unbounded in the direction of the extreme and all moments of the distribution exist (exponential type distributions), like
   - Largest: normal, lognormal, exponential, gamma, Weibull
   - Smallest: normal
2. **Type II**: parent distribution is unbounded in the direction of the extreme but not all moments exist (Pareto type distributions):
   - Largest: Cauchy, Pareto, log-gamma, Student's t
   - Smallest: Cauchy distribution
3. **Type III**: parent distribution is bounded in the direction of the extreme (limited distributions):
   - Largest: beta
   - Smallest: beta, lognormal, gamma, exponential.

The above types of extreme value distributions are often indicated as Fisher-Tippett Type I, II and III distributions or shortly as EV-1, EV-2 and EV-3 respectively.

Asymptotic distributions for X_{max}

The distributions for X_{max} of the 3 distinguished types have the following forms:

- **Type I distribution, largest value,** for $-\infty < x < \infty$ and $\beta > 0$:

$$F_{X\max}(x) = \exp(-\exp(-\left(\frac{x - x_0}{\beta}\right)) \tag{4.99}$$

- **Type II distribution, largest value,** for $x \geq x0$, $k < 0$ and $\beta > 0$

$$F_{X\max}(x) = \exp(-\left(\frac{x - x_0}{\beta}\right)^{1/k}) \tag{4.100}$$

- **Type III distribution, largest value,** for $x \leq x_0$, $k > 0$ and $\beta > 0$

$$F_{X\max}(x) = \exp(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}) \tag{4.101}$$

It is observed that the forms of the Type II and Type III distributions are similar, apart from sign differences and location of boundaries relative to the variable. All above asymptotic distributions for the largest value can be represented by the following general form of the extreme value distribution or shortly GEV distribution (Jenkinson, 1969):

$$F_{Xmax}(x) = \exp\left(-\left(1-k\left(\frac{x-b}{a}\right)\right)^{1/k}\right) \qquad (4.102)$$

Dependent on the sign of k the following cases are distinguished:

- k = 0: extreme value distribution Type I, EV-1
- k < 0: extreme value distribution Type II, EV-2
- k > 0: extreme value distribution Type III, EV-3

To arrive at the Type I distribution from (4.102) consider the Taylor series expansion of the argument of the exponential function in the limit for $k \rightarrow 0$:

$$\lim_{k\rightarrow 0}\left(1-k\left(\frac{x-b}{a}\right)\right)^{1/k} = \exp\left(-\frac{x-b}{a}\right)$$

Hence, for k = 0 with b = x_0 and a = β equation (4.99) is obtained from (4.102). Equivalently, with b + a/k = x_0 and $\pm$a/k = β equations (4.100) and (4.101) for the Type II and Type III distributions follow from (4.102). The GEV-form is sometimes used in literature on extreme value distributions to describe the Type II and Type III distributions, like in the Flood Studies Report (NERC, 1975). The different type of distributions for X_{max} are presented in Figure 4.23. It is observed that there is an upper limit to X_{max} in case of EV-3.

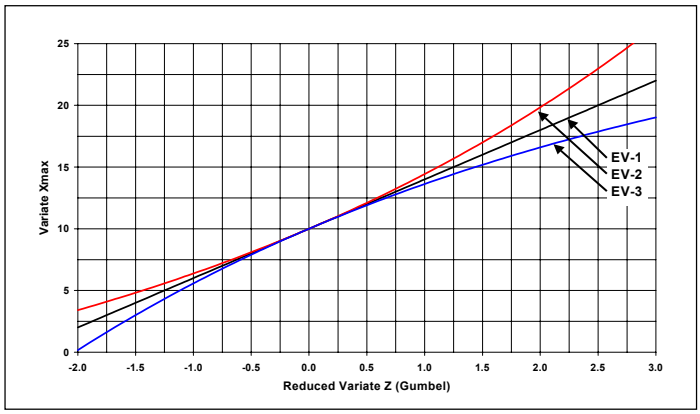


Figure 4.23:
Presentation of EV-1, EV-2 and EV-3 as function of reduced EV-1 variate

As shown in Figure 4.24, there is a distinct difference in the skewness of the X_{max} series suitable to be modelled by one of the EV-distributions. EV-1 has a fixed skewness (= 1.14), whereas EV-2 has a skewness > 1.14 and EV-3 a skewness < 1.14. Hence, a simple investigation of the skewness of a series of X_{max} will give a first indication of the suitability of a distribution.

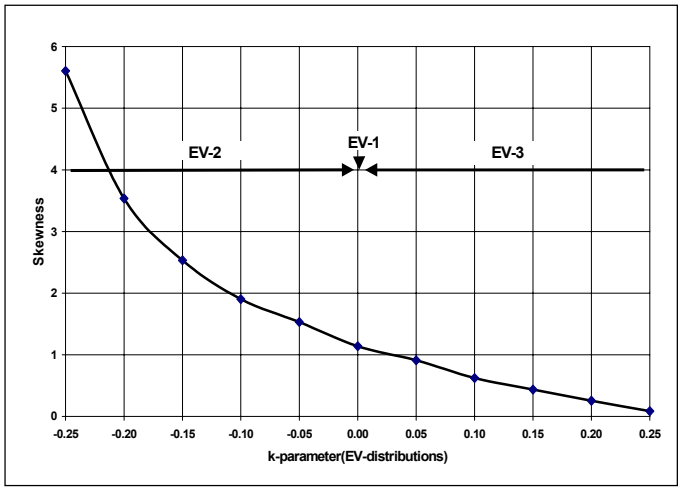


Figure 4.24:
Skewness as function of EV-parameter k

Asymptotic distributions for X_{min}

From the principle of symmetry (see e.g. Kottegoda and Rosso, 1997), the asymptotic distributions for the smallest value can be derived from the distribution of the largest value by **reversing the sign** and taking the **complementary probabilities**. Let X denote a variate with pdf $f_X(x)$ and X* a variate whose pdf is the mirror image of $f_X(x)$, it then follows: $f_X(x) = f_{X^*}(-x)$ and therefore: $1 - F_X(x) = F_{X^*}(-x)$. So for the distributions of X_{min} as a function of those of X_{max} it follows:

$$F_{X\min}(x) = 1 - F_{X\max}(-x) \tag{4.103}$$

Hence, the asymptotic distributions of X_{min} for the 3 distinguished types read:

- **Type I distribution, smallest value,** for $-\infty < x < \infty$ and $\beta > 0$:

$$F_{X\min}(x) = 1 - \exp(-\exp\left(\frac{x - x_0}{\beta}\right) \tag{4.104}$$

- **Type II distribution, smallest value,** for $x \le x_0$, $k < 0$ and $\beta > 0$:

$$F_{X\min}(x) = 1 - \exp\left(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.105}$$

- **Type III distribution, smallest value,** for $x \ge x_0$, $k > 0$ and $\beta > 0$

$$F_{X\min}(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \tag{4.106}$$

In hydrology, particularly Type I for largest value and Type III for smallest value are frequently used. In the next sub-sections all types are discussed.

4.6.3 Extreme value Type 1 or Gumbel distribution

EV-1 for largest value

The Extreme Value Type I distribution for the largest value was given by equation (4.99):

$$F_{X\max}(x) = \exp\left\{-\exp\left(-\left(\frac{x - x_0}{\beta}\right)\right)\right\} \text{ for : } -\infty < x < \infty \text{ and } \beta > 0 \tag{4.99}$$

The pdf is obtained by differentiating (4.99) with respect to x and reads:

$$f_{Xmax}(x) = \frac{1}{\beta}\exp\left\{-\left(\frac{x-x_0}{\beta}\right) - \exp\left(-\left(\frac{x-x_0}{\beta}\right)\right)\right\} \qquad (4.107)$$

In view of the form, equation (4.99) is called the **double exponential** distribution or in honour to its promoter the **Gumbel** distribution. Introducing the reduced or standardised variate Z, defined by:

$$Z = \frac{X_{max} - x_0}{\beta} \qquad (4.108)$$

The standardised Gumbel distribution is obtained by observing that Z = X for x_0 = 0 and β = 1:

$$F_Z(z) = \exp(-\exp(-z)) \qquad (4.109)$$

$$f_Z(z) = \exp(-z - \exp(-z)) \qquad (4.110)$$

The standardised pdf and cdf are shown in Figure 4.25

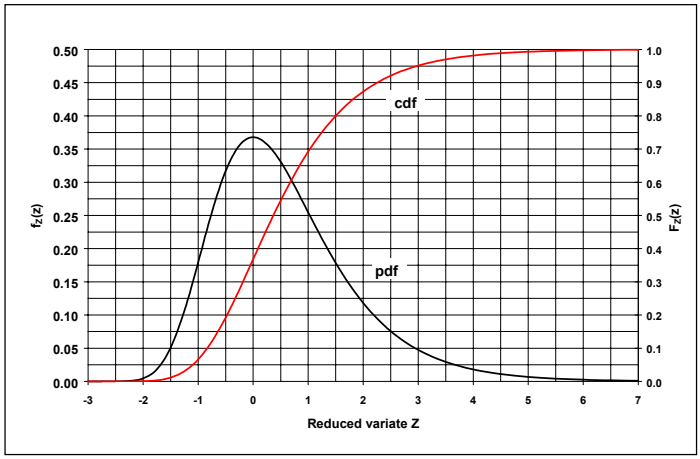


Figure 4.25:
Standardised Gumbel pdf and cdf

The moment related parameters of the distribution, the mean, median, mode, variance skewness and kurtosis are given by:

$$\left.\begin{aligned} \mu_{Xmax} &= x_0 + \gamma_E\beta = x_0 + 0.5772\beta \\ M_{Xmax} &= x_0 + \beta\ln(\ln 2) = x_0 + 0.3665\beta \\ m_{Xmax} &= x_0 \end{aligned}\right\} \qquad (4.111a)$$

$$\left.\begin{aligned} \sigma^2_{Xmax} &= \frac{\pi^2\beta^2}{6} \\ \gamma_{1,Xmax} &\approx 1.1396 \\ \gamma_{2,Xmax} &= 5.4 \end{aligned}\right\} \qquad (4.111b)$$

The constant γ_E = 0.577216 is called Euler's constant and can be read from mathematical tables. The parameter x_0 is seen to be a **location** parameter and β is a **scale** parameter. The skewness is fixed at 1.14 and the kurtosis is > 3, hence the pdf is more peaked than the normal distribution.

The moments of the distribution and its related parameters can be obtained from the moment generating function:

$$G_{Xmax}(s) = \exp(x_0 s)\Gamma(1-\beta s) \tag{4.112}$$

More easily the moment related parameters for the Gumbel distribution can be obtained from the cumulants κ_n of the distribution (see e.g. Abramowitz and Stegun, 1970):

$$\left.\begin{aligned} &\kappa_1 = x_{0+}\gamma_E\beta \\ &\kappa_n = \beta^n\Gamma(n)\varsigma(n) \\ &\text{where : } \varsigma(n) = \sum_{r=1}^{\infty}\frac{1}{r^n} \text{ specifically : } \varsigma(2) = \frac{\pi^2}{6};\ \varsigma(4) = \frac{\pi^4}{90} \end{aligned}\right\} \tag{4.113}$$

The function $\zeta(n)$ is the Riemann Zeta Function and is tabulated in mathematical tables. The relation between the cumulants and the moments are:

$$\kappa_1 = \mu_1';\ \kappa_2 = \mu_2;\kappa_3 = \mu_3;\ \kappa_4 = \mu_4 - 3\mu_2^2 \tag{4.114}$$

Hence:

$$\sigma^2 = \kappa_2 = \beta^2\Gamma(2)\varsigma(2) = \beta^2\frac{\pi^2}{6}$$

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{\beta^3\Gamma(3)\varsigma(3)}{\left(\beta^2\varsigma(2)\right)^{3/2}} = \frac{\Gamma(3)\varsigma(3)}{\left(\frac{\pi^2}{6}\right)^{3/2}} = \frac{2x1.20205}{2.10971} = 1.139541$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} + 3 = \frac{\beta^4\Gamma(4)\varsigma(4)}{\left(\beta^2\Gamma(2)\varsigma(2)\right)^2} + 3 = \frac{\Gamma(4)\frac{\pi^4}{90}}{\left(\Gamma(2)\frac{\pi^2}{6}\right)^2} + 3 = \frac{2x3x36}{90} + 3 = \frac{12}{5} + 3 = 5.4$$

Distribution parameters expressed in moment related parameters

From (4.111) the following relations between x_0, β and μ and σ are obtained:

$$\beta = \frac{\sqrt{6}}{\pi}\sigma \tag{4.115}$$

$$x_0 = \mu - \gamma_E\frac{\sqrt{6}}{\pi}\sigma = \mu - 0.45\sigma \tag{4.116}$$

Quantiles of EV-1 for X_{max}

The value for X_{max} for a specified return period T, $x_{max}(T)$, can be derived from (4.108) and (4.109):

$$x_{max}(T) = x_0 - \beta\ln\left(\ln\left(\frac{T}{T-1}\right)\right) = \mu_{Xmax} - \sigma_{Xmax}\frac{\sqrt{6}}{\pi}\left\{\gamma_E + \ln\left(\ln\left(\frac{T}{T-1}\right)\right)\right\} \tag{4.117}$$

In some textbooks the quantiles are determined with the aid of a frequency factor K(T):

$$x_{max}(T) = \mu_{Xmax} + K(T)\sigma_{Xmax} \tag{4.118}$$

Hence:

$$K(T) = -\frac{\sqrt{6}}{\pi}\left\{\gamma_E + \ln\left(\ln\left(\frac{T}{T-1}\right)\right)\right\} \tag{4.119}$$

Values for K(T) for selected return periods are presented in table below:

T	K(T)	T	K(T)
2	-0.1643	100	3.1367
5	0.7195	250	3.8535
10	1.3046	500	4.3947
25	2.0438	1000	4.9355
50	2.5923	1250	5.1096

From (4.118) it is observed that if to a given set of extremes some very low values are added the quantile for high return periods may increase!! This stems from the fact that though μ_{xmax} may reduce some what, σ_{xmax} will increase, since the overall variance increases. Because for large T, K(T) becomes large, it follows that $x_{max}(T)$ may be larger than before. This is a “lever” effect.

Application of EV-1 for largest value

The Gumbel distribution appears to be a suitable model for annual maximum rainfall and runoff in a number of cases, though many a times it does not apply. A first rapid indication about the applicability of the Gumbel distribution can be obtained from the skewness of the data set of maximum values. If this deviates substantially from 1.14, the distribution is not suitable to model the extremes.

EV-1 for smallest value

The cdf of the EV-1 distribution for the smallest value is given by (4.104):

$$F_{Xmin}(x) = 1 - \exp\left(-\exp\left(\frac{x - x_0}{\beta}\right)\right) \tag{4.104}$$

and the pdf then reads:

$$f_{Xmin}(x) = \frac{1}{\beta}\exp\left\{\left(\frac{x - x_0}{\beta}\right) - \exp\left(\frac{x - x_0}{\beta}\right)\right\} \tag{4.120}$$

Introducing the reduced variate Z defined by:

$$Z = \frac{X_{min} - x_0}{\beta} \tag{4.121}$$

then the standardised cdf and pdf read:

$$F_Z(z) = 1 - \exp(-\exp(z)) \tag{4.122}$$

$$f_Z(z) = \exp(z - \exp(z)) \tag{4.123}$$

The standardised distribution is shown in Figure 4.26. From this figure it is observed that the pdf for the smallest value is the mirror image of the pdf of the largest value around z = 0.

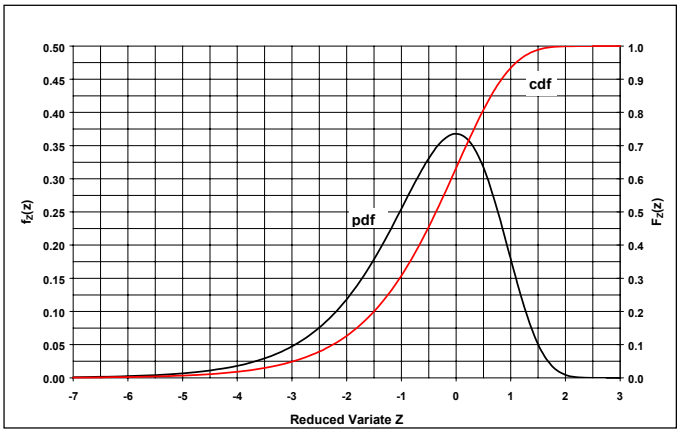


Figure 4.26:
Standardise EV-1 pdf and cdf for smallest value

The moment related parameters of the distribution, the mean, median, mode, variance skewness and kurtosis are given by:

$$\left.\begin{aligned} \mu_{X\min} &= x_0 - \gamma_E\beta = x_0 - 0.5772\beta \\ M_{X\min} &= x_0 - \beta\ln(\ln 2) = x_0 - 0.3665\beta \\ m_{X\min} &= x_0 \end{aligned}\right\} \qquad (4.124a)$$

$$\left.\begin{aligned} \sigma^2_{X\min} &= \frac{\pi^2\beta^2}{6} \\ \gamma_{1,X\min} &\approx -1.1396 \\ \gamma_{2,X\min} &= 5.4 \end{aligned}\right\} \qquad (4.124b)$$

Comparing these results with (4.111) it is observed that, apart from some changes in sign, the components of the above formulae are similar. For the distribution parameters expressed in the moment related parameters it now follows:

$$\beta = \frac{\sqrt{6}}{\pi}\sigma \qquad (4.125)$$

$$x_0 = \mu + \gamma_E\frac{\sqrt{6}}{\pi}\sigma = \mu + 0.45\sigma \qquad (4.126)$$

Quantiles of EV-1 for X_{min}

In case of the smallest value we are interested in non-exceedance probability of X_{min}. Let this non-exceedance probability pbe denoted by p then the value of X_{min} for a specified non-exceedance probability p can be derived from (4.121) and (4.122):

$$x_{\min}(p) = x_0 + \beta\ln(-\ln(1-p)) = \mu_{X\min} + \sigma_{X\min}\frac{\sqrt{6}}{\pi}\{\gamma_E + \ln(-\ln(1-p))\} \qquad (4.125)$$

Example 4.11 EV-1 for smallest value

Annual minimum flow series of a river have a mean and standard deviation of 500 m^3/s and 200 m^3/s. Assuming that the frequency distribution of the minimum flows is EV-1, what is the probability of zero flow?

The problem can be solved by equation (4.104), which requires values for x_0 and β. From (4.125) and (4.126) it follows for x_0 and β:

$$\beta = \frac{\sqrt{6}}{\pi}\sigma = 0.7797 \text{x} 200 = 155.9$$
$$x_0 = \mu + 0.45\sigma = 500 + 0.45 \text{x} 200 = 590.0$$

Substituting the parameter values in equation (4.104) gives:

Hence, on average once every 45 years the river will run dry according to the EV-1 distribution

$$F_{X\min}(0) = 1 - \exp\left(-\exp\left(\frac{0 - x_0}{\beta}\right)\right) = 1 - \exp\left(-\exp\left(-\frac{590.0}{155.9}\right)\right) = 1 - 0.9775 = 0.0225 \approx \frac{1}{45}$$

4.6.4 Extreme value Type 2 or Fréchet distribution

EV-2 for largest value

The cdf of the Extreme Value Type II distribution for largest value for is given by (4.100):

$$F_{X\max}(x) = \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \text{ for : } x \geq x_0 \text{ ; } k < 0 \text{ ; } \beta > 0 \qquad (4.100)$$

The pdf is obtained by differentiation:

$$f_{X\max}(x) = -\frac{1}{k\beta}\left(\frac{x - x_0}{\beta}\right)^{1/k-1} \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \qquad (4.126)$$

Introducing the reduced variate Z according to (4.108), the following standardised forms are obtained for the cdf and the pdf:

$$F_Z(z) = \exp(-z^{1/k}) \qquad (4.127)$$

$$f_Z(z) = -\frac{1}{k} z^{1/k-1} \exp(-z^{1/k}) \qquad (4.128)$$

In Figures 4.27 and 4.28 the pdf and cdf of the EV-2 distribution are presented for different values of k.

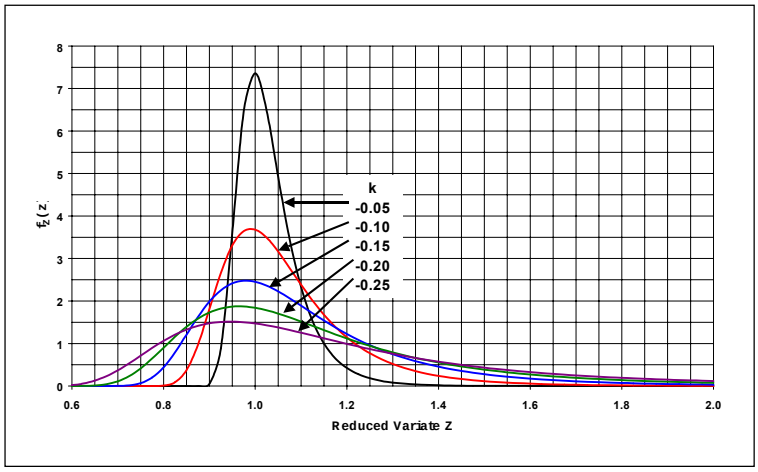


Figure 4.27:
Pdf of EV-2 distribution for different k values

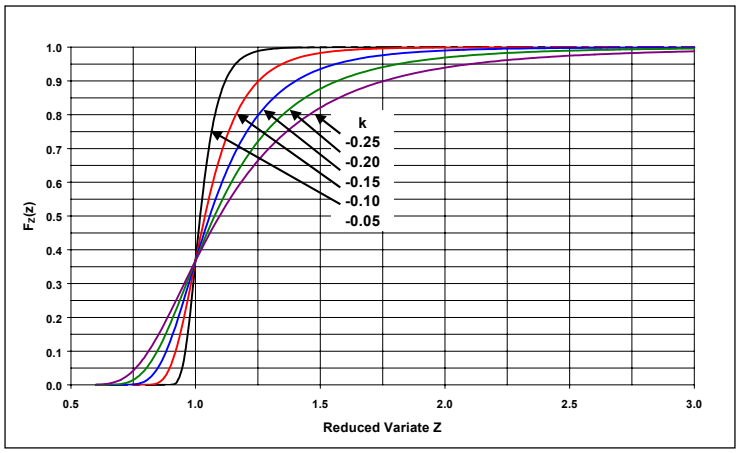


Figure 4.28:
Cdf of EV-2 distribution for different k-values

The moment related parameters of the distribution read:

$$\left.\begin{aligned} \mu_{Xmax} &= x_0 + \beta\Gamma(1+k) \\ M_{Xmax} &= x_0 + \beta(\ln 2)^k \\ m_{Xmax} &= x_0 + \beta(1-k)^k \end{aligned}\right\} \qquad (4.129a)$$

$$\left.\begin{aligned} \sigma^2_{Xmax} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,Xmax} &= \frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.129b)$$

Above expressions show that x_0 is a **location** parameter, β a **scale** parameter and k a **shape** parameter as the latter is the sole parameter affecting skewness. From the above figures it is observed that the skewness decreases with increasing k.

The moment related parameters (4.129 a and b) can easily be derived from the following expression for the r^{th} moment about the origin in case $x_0 = 0$ substituted in (3.30):

$$\mu_r' = \beta^r\Gamma(1+rk) \qquad (4.130)$$

From (4.129) it is observed that the distribution parameters cannot analytically be expressed in the moments of the distribution; an iterative procedure is required for this.

Quantiles of EV-2 for X_{max}

The quantile $x_{max}(T)$ for a given return period T follows from (4.100):

$$x_{max}(T) = x_0 + \beta\left(\ln\left(\frac{T}{T-1}\right)\right)^k \qquad (4.131)$$

Fréchet and log-Gumbel distributions

EV-2 for the largest value is also indicated as **Fréchet** distribution or **log-Gumbel** distribution. With respect to the latter it can be shown that if $(x_{max}-x_0)$ has a EV-2 distribution, its logarithm $Y= \ln(x_{max}-x_0)$ has a Gumbel distribution with parameters *a* and *b*, as follows:

$$F_Y(y) = \exp\left\{-\exp\left(-\left(\frac{y-b}{a}\right)\right)\right\}$$

$$F_{Xmax}(x) = \exp\left\{-\exp\left(-\left(\frac{\ln(x-x_0)-b}{a}\right)\right)\right\}$$

Since:

$$\exp\left(-\left(\frac{\ln(x-x_0)}{a}\right)\right) = (x-x_0)^{-1/a}$$

it follows:

$$F_{Xmax}(x) = \exp\left\{-(x-x_0)^{-1/a}e^{b/a}\right\} = \exp\left\{-\left(\frac{x-x_0}{e^b}\right)^{-1/a}\right\}$$

It is observed that above expression equals (4.100) for

$$\left.\begin{aligned} a &= -k \\ b &= \ln(\beta) \end{aligned}\right\} \qquad (4.132)$$

EV-2 for smallest value

The Extreme Value Type II distribution for the smallest value is given by (4.105)

$$F_{Xmin}(x) = 1-\exp\left(-\left(-\frac{x-x_0}{\beta}\right)^{1/k}\right) \text{ for : } x \le x_0 \text{ ; } k < 0 \text{ ; } \beta > 0 \qquad (4.105)$$

The pdf can be derived by taking the derivative of (4.105) with respect to x:

$$f_{Xmin}(x) = -\frac{1}{k\beta}\left(-\frac{x-x_0}{\beta}\right)^{1/k-1}\exp\left(-\left(-\frac{x-x_0}{\beta}\right)^{1/k}\right) \qquad (4.133)$$

The moment related parameters of the distribution can easily be obtained from (4.129a and b) knowing that the pdf is the mirror image of the pdf for the largest value:

$$\left.\begin{aligned} \mu_{X\min} &= x_0 - \beta\Gamma(1+k) \\ M_{X\min} &= x_0 - \beta(\ln 2)^k \\ m_{X\min} &= x_0 - \beta(1-k)^k \end{aligned}\right\} \qquad (4.134a)$$

$$\left.\begin{aligned} \sigma^2_{X\min} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,X\min} &= -\frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.134b)$$

It appears that the EV-2 for the smallest value finds little application in hydrology and will therefore not be discussed any further.

4.6.5 Extreme value Type 3 distribution

EV-3 for largest value

The Extreme Value Type III distribution for largest value is given by (4.101) and is defined for $x \le x_0$, $k > 0$ and $\beta > 0$

$$F_{X\max}(x) = \exp\left(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}\right) \qquad (4.101)$$

The pdf reads:

$$f_{X\max}(x) = \frac{1}{k\beta}\left(-\frac{x - x_0}{\beta}\right)^{1/k-1} \exp\left(-\left(-\frac{x - x_0}{\beta}\right)^{1/k}\right) \qquad (4.135)$$

The mean, median, mode, variance and skewness are given by:

$$\left.\begin{aligned} \mu_{X\max} &= x_0 - \beta\Gamma(1+k) \\ M_{X\max} &= x_0 - \beta(\ln 2)^k \\ m_{X\max} &= x_0 - \beta(1-k)^k \end{aligned}\right\} \qquad (4.136a)$$

$$\left.\begin{aligned} \sigma^2_{X\max} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,X\max} &= -\frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.136b)$$

Note that these expressions are similar to those of the smallest value modelled as EV-2. Above moment related parameters are easily obtained from the r^{th} moment of $(x_0 - X_{max})$ which can shown to be:

$$E\left[(x_0 - X_{max})^r\right] = \beta^r\Gamma(1+rk) \qquad (4.137)$$

To simplify the computation, note that for the higher moments x_0 can be omitted, so for $r > 1$ one can put $x_0 = 0$ and use (3.30). Equation (4.137) then simplifies to:

$$\mu_r' = (-1)^r \beta^r \Gamma(1+rk)$$

So:

$$\mu_2' = \beta^2 \Gamma(1+2k)$$
$$\mu_3' = -\beta^3 \Gamma(1+3k)$$

The fact that X_{max} is bounded by x_0 makes that EV-3 is seldom used in hydrology for modelling the distribution of X_{max}. Its application only make sense, if there is a physical reason that limits X_{max} to x_0.

EV-3 for smallest value

The extreme value Type III distribution for the smallest value, for $x \geq x_0$, $k > 0$ and $\beta > 0$, has the following form:

$$F_{Xmin}(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \qquad (4.106)$$

and the pdf reads:

$$f_{Xmin}(x) = \frac{1}{k\beta}\left(\frac{x - x_0}{\beta}\right)^{1/k-1} \exp\left(-\left(\frac{x - x_0}{\beta}\right)^{1/k}\right) \qquad (4.138)$$

In above equations, x_0 is a location parameter, β a scale parameter and k a shape parameter.

This distribution is seen to be identical to the **Weibull** distribution, equation (4.84) and (4.85), by putting 1/k = k*, where k* is the shape parameter of the Weibull distribution. Hence reference is made to Sub-section 4.3.11 for further elaboration of this distribution. Above distribution is also called **Goodrich** distribution.

The moment related parameters according to the above definition are shown here, as it corresponds to the parameter definition adopted in HYMOS. The mean, median, mode, variance and skewness read:

$$\left.\begin{aligned} \mu_{Xmin} &= x_0 + \beta\Gamma(1+k) \\ M_{Xmin} &= x_0 + \beta(\ln 2)^k \\ m_{Xmin} &= x_0 + \beta(1-k)^k \end{aligned}\right\} \qquad (4.139a)$$

$$\left.\begin{aligned} \sigma^2_{Xmin} &= \beta^2\left\{\Gamma(1+2k) - \Gamma^2(1+k)\right\} \\ \gamma_{1,Xmin} &= \frac{\Gamma(1+3k) - 3\Gamma(1+k)\Gamma(1+2k) + 2\Gamma^3(1+k)}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{3/2}} \end{aligned}\right\} \qquad (4.139b)$$

The location parameter x_0 is seen to be the lower bound of the distribution. Often, the parent distribution will have a lower bound equal to zero and so will have the EV-3 for the smallest value. Above form with x_0 is therefore often indicated as the **shifted** Weibull distribution.

In literature the shifted Weibull distribution is often presented as:

$$F_{Xmin}(x) = 1 - \exp\left(-\left(\frac{x-b}{a-b}\right)^c\right) \quad \text{with : } x > b \,;\, a > b \qquad (4.140)$$

where the resemblance with the above parameter definition is seen for: $x_0 = b$, $\beta = a - b$ and $k = 1/c$.

Quantiles of EV-3 for X_{min}

Since one is dealing with the smallest value, interest is in the non-exceedance probability of X_{min}. If this non-exceedance probability is denoted by p then the value of x_{min} for a specified non-exceedance probability p can be derived from (4.106):

$$x_{min}(p) = x_0 + \beta\{-\ln(1-p)\}^k \qquad (4.141)$$

Example 4.11 (continued.) EV-3 for smallest value.

Annual minimum flow series of a river have a mean and standard deviation of 500 m^3/s and 200 m^3/s. Assuming that the frequency distribution of the minimum flows is EV-3, with x_0 = 0, what low flow value will not be exceeded on average once in 100 years?

The non-exceedance probability q = 0.01. To apply (4.141) k and β have to be known. The parameters k and β are obtained as follows. Note that for x_0, the coefficient of variation becomes:

$$C^2_{v,Xmin} = \left(\frac{\sigma_{Xmin}}{\mu_{Xmin}}\right)^2 = \frac{\Gamma(1+2k)}{\Gamma^2(1+k)} - 1 = \left(\frac{200}{500}\right)^2 = 0.16$$

From above equation it is observed that the coefficient of variation is only a function of k when x_0 = 0. By iteration one finds k = 0.37. From (4.139b) it follows for β:

$$\beta = \frac{\sigma_{Xmin}}{\left(\Gamma(1+2k) - \Gamma^2(1+k)\right)^{1/2}} = \frac{200}{0.3549} = 564$$

With β = 564 and k = 0.37 one finds with (4.141) for the 100 year low flow:

According to the EV-1 distribution for the smallest value, which was applied to the same series in Sub-section 4.4.3, Q = 103 m^3/s has a return period of about 23 years. It follows that the two distributions lead to very different results. In practice, the EV-3 for smallest value finds widest application.

$$X_{min}(0.01) = 0 + 564x\{-\ln(1-0.01)\}^{0.37} = 564x0.182 = 103 m^3/s$$

4.6.6 Generalised Pareto distribution

For modelling frequency distributions of extremes, particularly of partial duration series, the Pareto distribution is often used.The cdf of the generalised Pareto distribution has the following form:

$$F_X(x) = 1 - \left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta} \qquad (4.142)$$

Like for the Extreme Value distributions as discussed in the previous sub-sections, three types of Pareto distributions are distinguished, which are directly related to EV-1, 2 and 3 (see next sub-chapter):

- **Type I distribution, P-1:**

$$F_X(x) = 1 - \exp\left(-\left(\frac{x - x_0}{\sigma}\right)\right) \text{ for : } x_0 \le x < \infty \text{ when : } \theta = 0 \qquad (4.143)$$

- **Type II distribution, P-2:**

$$F_X(x) = 1 - \left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta} \quad \text{for : } x_0 \le x < \infty \text{ when : } \theta < 0 \tag{4.144}$$

- **Type III distribution, P-3:**

$$F_X(x) = 1 - \left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta} \quad \text{for : } x_0 \le x \le x_0 + \frac{\sigma}{\theta} \text{ when : } \theta > 0 \tag{4.145}$$

The pdf's of the Pareto distributions are respectively with the validity range as defined for the cdf's above, for **P-1**:

$$f_X(x) = \frac{1}{\sigma}\exp\left(-\left(\frac{x - x_0}{\sigma}\right)\right) \tag{4.146}$$

and for **P-2** and **P-3**:

$$f_X(x) = \frac{1}{\sigma}\left(1 - \theta\left(\frac{x - x_0}{\sigma}\right)\right)^{1/\theta - 1} \tag{4.147}$$

Note that the P-1 distribution results as a special for θ = 0 from P-2 or P-3 similar to the EV-1 distribution resulting from GEV, see Sub-section 4.4.2. In the above distributions, x_0 is a **location** parameter, σ is a **scale** parameter and θ is a **shape** parameter. The mean, variance, skewness and kurtosis of the distributions are given by:

$$\left.\begin{aligned} \mu_X &= x_0 + \frac{\sigma}{1+\theta} \\ \sigma_X^2 &= \frac{\sigma^2}{(1+\theta^2)(1+2\theta)} \\ \gamma_{1,X} &= \frac{2(1-\theta)\sqrt{1+2\theta}}{1+3\theta} \\ \gamma_{2,X} &= \frac{3(1+2\theta)(3-\theta+2\theta^2)}{(1+3\theta)(1+4\theta)} \end{aligned}\right\} \tag{4.148}$$

Above expressions can be derived by noticing (Metcalfe, 1997):

$$E\left[(1 - \theta\frac{X}{\sigma})^r\right] = \frac{1}{1+r\theta} \text{ for : } \theta > -\frac{1}{r} \tag{4.149}$$

For θ < -1/r the r^{th} moment does not exist.

The generalised Pareto distribution in a standardised form (x_0 = 0 and σ = 1) for various values of θ are given in Figures 4.29 and 4.30.

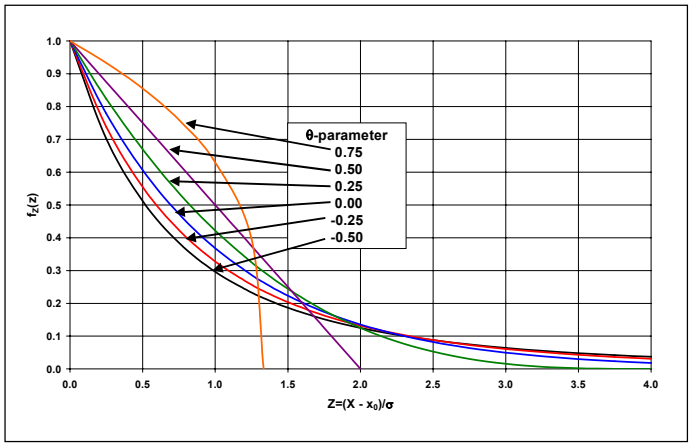


Figure 4.29:
Pdf of Pareto distribution for various values of shape parameter

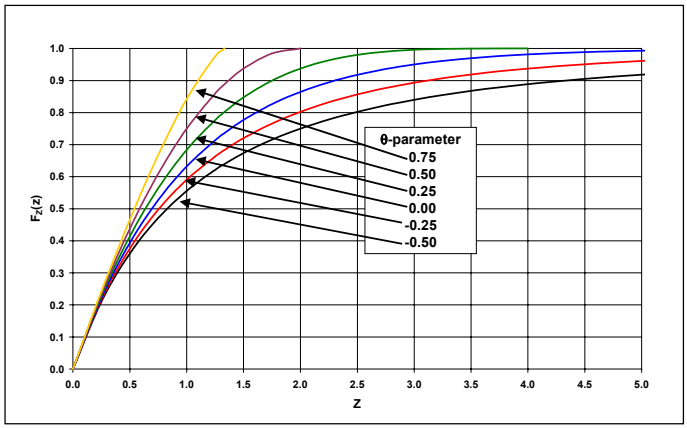


Figure 4.30:
Cdf of Pareto distribution for various of shape parameter

Quantiles

The quantiles, referring to a return period of T years, follow from (4.143) to (1.145) and read:

- For Type I distribution P-1:

$$x_T = x_0 + \sigma \ln T \qquad (4.150)$$

- For Type II and III distributions, P-2, P-3:

$$x_T = x_0 + \frac{\sigma}{\theta}(1 - T^{-\theta}) \qquad (4.151)$$

Note that above two expressions should not directly be applied to exceedance series unless the number of data points coincide with the number of years, see next sub-section.

4.6.7 Relation between maximum and exceedance series

The GEV distributions are applicable to series with a fixed interval, e.g. a year: series of the largest or smallest value of a variable each year, like annual maximum or minimum flows. If one considers largest values, such a series is called an **annual maximum series.** Similarly, **annual minimum series** can be defined.

In contrast to this, one can also consider series of extreme values above or below a certain threshold value, i.e. the maximum value between an upcrossing and a downcrossing or the minimum between a downcrossing and an upcrossing, see Figure 4.31.

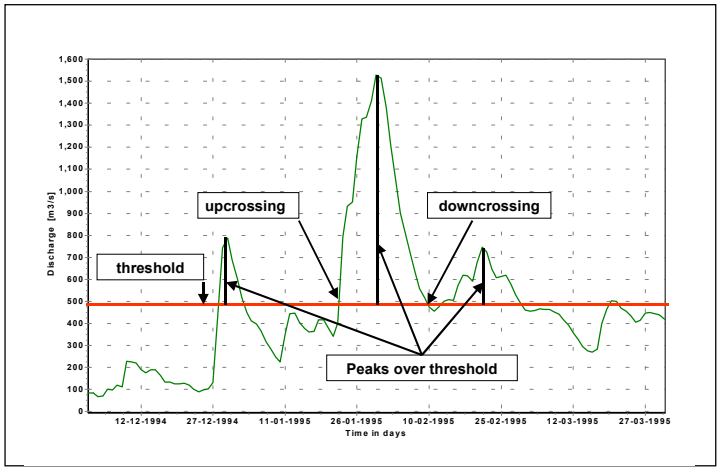


Figure 4.31:
Definition of partial duration or peaks over threshold series

The series resulting from exceedance of a base or threshold value x_0 thereby considering only the maximum between an upcrossing and a downcrossing is called a partial duration series (PDS) or peaks over threshold series, POT-series. The statistics may be developed for the exceedance of the value relative to the base only or for the value as from zero. The latter approach will be followed here. In a similar manner partial duration series for non-exceedance of a threshold value can be defined. When considering largest values, if the threshold is chosen such that the number of exceedances N of the threshold value equals the number of years n, the series is called annual exceedance series. So, if there are n years of data, in the annual exceedance series the n largest independent peaks out of $N \geq n$ are considered. To arrive at independent peaks, there should be sufficient time between successive peaks. The physics of the process determines what is a sufficient time interval between peaks to be independent; for flood peaks a hydrograph analysis should be carried out. The generalised Pareto distribution is particularly suited to model the exceedance series.

Note that there is a distinct difference between annual maximum and annual exceedance series. In an annual maximum series, for each year the maximum value is taken, no matter how low the value is compared to the rest of the series. Therefore, the maximum in a particular year may be less than the second or the third largest in another year, which values are considered in the annual exceedance series if the ranking so permits. Hence the lowest ranked annual maximums are less than (or at the most equal to) the tail values of the ranked annual exceedance series values.

The procedure to arrive at the annual exceedance series via a partial duration series and its comparison with the annual maximum series is shown in the following figures, from a record of station Chooz on Meuse river in northern France (data 1968-1997). The original discharge series is shown in Figure 4.32. Next a threshold level of 400 m^3/s has been assumed. The maximum values between each upcrossing and the next downcrossing are considered. In this particular case, peaks which are distanced $\geq$ 14 days apart are expected to be independent and are included in the partial duration series, shown in Figure 4.33. This results in 72 peaks. Since there are 30 years of record, the partial duration series has to be reduced to the 30 largest values. For this the series values are ranked in descending order and the first 30 values are taken to form the annual exceedance series. The threshold value for the annual exceedance series appears to 620 m^3/s. The annual exceedance series is shown in Figure 4.34. It is observed that some years do not contribute to the series, as their peak values were less than 620 m^3/s, whereas other years contribute with 2 or some even with 3 peaks. The annual maximum series is presented in Figure 4.35, together with the threshold for the annual exceedance series. It is observed that indeed for a number of years that threshold level was not reached. A comparison of the two series is depicted in Figure 4.36.

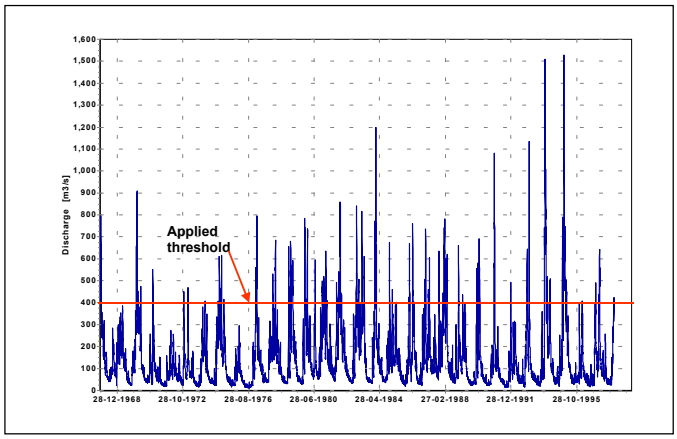


Figure 4.32:
Discharge series of station chooz on Meuse river with applied threshold Q = 400 m^3/s

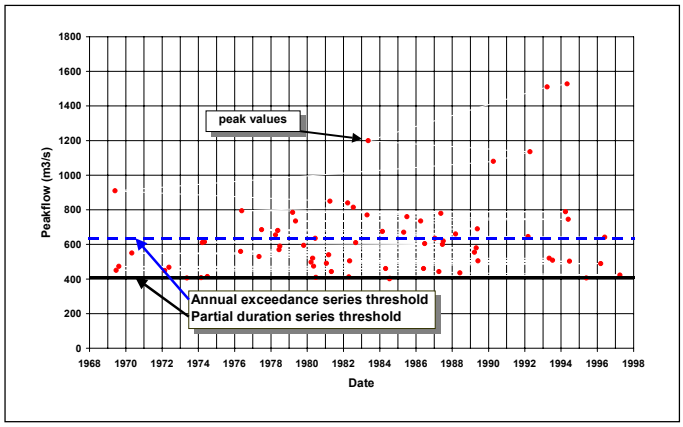


Figure 4.33:
Partial duration series of peaks over 400 m^3/s

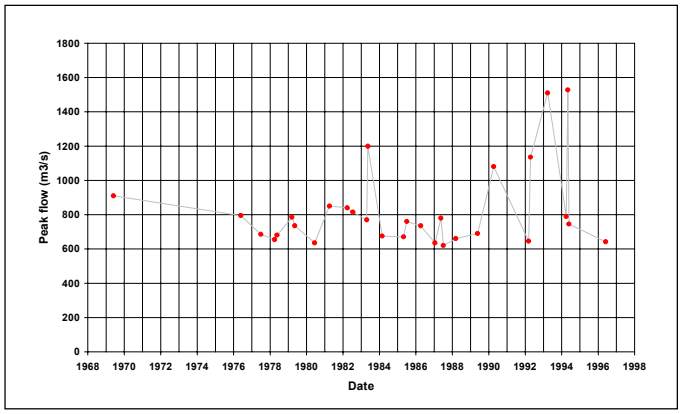


Figure 4.34:
Annual exceedance series $Q \geq 620\ m^3/s$

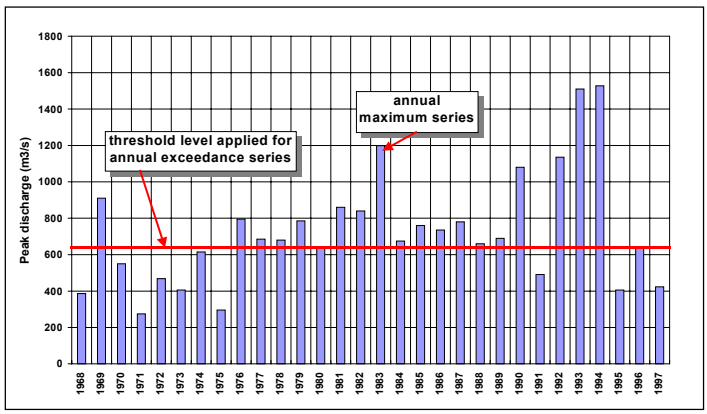


Figure 4.35:
Annual maximum series

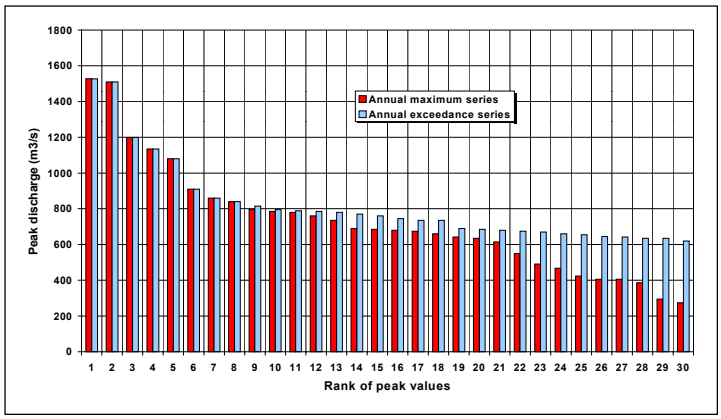


Figure 4.36:
Comparison of annual maximum series and annual exceedance series

From Figure 4.36 it is observed that the largest values in both series are the same, but the lower tail is quite different. It follows that the annual maximum series will produce lower extremes for low return periods, say up to T = 5 or T = 10 years return period.

Conditional exceedance probabilities

It is noted that straightforward application of fitting a frequency distribution to a partial duration or peak over threshold series (i.e. an **exceedance** series) involves a conditional distribution, i.e. the probability of an exceedance given that a threshold level x_0 has been exceeded. Let this distribution of peaks over a threshold x_0 be denoted by $F_{POT}(x)$. If there are N_e exceedances of x_0 during N_y years, then the average number of exceedances of x_0 in one year is $\lambda = N_e / N_y$, and the average number of peaks $X>x|x>x_0$ per year becomes $\lambda(1- F_{POT}(x))$. The average number of peaks $X>x|x>x_0$ in T years then is $\lambda T(1- F_{POT}(x))$. To arrive at the T year flood the average number of peaks in T year should be 1, i.e.

$$\lambda T\left(1-F_{POT}(x)\right)=1$$

or:

$$F_{POT}(x)=1-\frac{1}{\lambda T} \qquad (4.152)$$

Substitution of a suitable model for $F_{POT}(x)$ in (4.152) like the P-1 distribution gives for the quantile x_T:

$$x_T = x_0 + \sigma \ln(\lambda T) \qquad (4.153)$$

It is observed that (4.153) is identical to (4.150) for λ = 1, i.e. when the number of exceedances is equal to the number of years and then the peak over threshold series becomes the annual exceedance series.

From exceedances to maximum

Consider again the distribution of the peaks over threshold: $F_{POT}(x)$. The number of exceedances N of the threshold in a fixed time period is a random variable, having a certain probability mass function $p_N(n)$. It can be shown (see e.g. Kottegoda and Rosso, 1997) that the cdf of X_{max} (i.e. the largest of the exceedances) can be derived from the conditional frequency distribution $F_{POT}(x)$ and $p_N(n)$ as follows:

$$F_{Xmax}(x)=\sum_{n=0}^{\infty}\left\{F_{POT}(x)\right\}^n p_N(n) \qquad (4.154)$$

If $p_N(n)$, i.e. the number of exceedances, is modelled by a **Poisson distribution**, which is equivalent to stating that the intervals between exceedances is exponentially distributed, then (4.154) simplifies to:

$$F_{Xmax}(x) = \exp\{-\lambda(1-F_{POT}(x))\} \tag{4.155}$$

where: λ = average number of exceedances (e.g. per year).

Equation (4.155) gives a relation between the **conditional exceedance** distribution $F_{POT}(x)$ and the **unconditional (annual) maximum** distribution. If **annual exceedance** series are considered (i.e. on average one exceedance per year: $\lambda = 1$) with distribution function $F_{AE}(x)$ it follows from (4.155):

$$F_{Xmax}(x) = \exp\{-(1-F_{AE}(x)\} \tag{4.156}$$

Equation (4.156) gives the relation between the annual maximum distribution $F_{xmax}(x)=F_{AM}(x)$ and the frequency distribution of the annual exceedance series $F_{AE}(x)$. For the relation between the return period of the annual exceedance series T_{AE} and the annual maximum series T_{AM} it follows:

$$1-\frac{1}{T_{AM}} = \exp\left(-\frac{1}{T_{AE}}\right) \text{ or : } T_{AM} = \left\{1-\exp\left(-\frac{1}{T_{AE}}\right)\right\}^{-1} \tag{4.157}$$

Equivalently

$$T_E = \left\{\ln\left(\frac{T_M}{T_M-1}\right)\right\}^{-1} \tag{4.158}$$

From Pareto to GEV

If one substitutes in equation (4.156) for the distribution of the exceedances $F_{AE}(x)$ the generalised Pareto distribution as discussed in the previous sub-section, then the distribution of X_{max} will be a GEV distribution with the same **shape** parameter. The cdf of the generalised Pareto distribution was given by (4.142):

$$F_X(x) = 1-\left(1-\theta\left(\frac{x-x_0}{\sigma}\right)\right)^{1/\theta} \tag{4.142}$$

Substitution in (4.156) gives:

$$F_{Xmax}(x) = \exp(-\lambda\left\{1-\left[1-\left(1-\theta\left(\frac{x-x_0}{\sigma}\right)\right)^{1/\theta}\right]\right\}) = \exp(-\left(\frac{\frac{\sigma}{\theta}-(x-x_0)}{\frac{\lambda^{-\theta}\sigma}{\theta}}\right)^{1/\theta})$$

To prove the resemblance with the GEV distribution given by equation (4.102), note that:

$$F_{Xmax}(x) = \exp(-\left(1-k\left(\frac{x-b}{a}\right)\right)^{1/k}) = \exp(-\left(\frac{\frac{a}{k}-(x-b)}{\frac{a}{k}}\right)^{1/k}) \tag{4.102}$$

It follows that (4.142) and (4.102) are equivalent if:

$$\left.\begin{aligned} k &= \theta \\ a &= \lambda^{-\theta}\sigma \\ b &= x_0 + \frac{\sigma}{\theta}(1-\lambda^{-\theta}) \end{aligned}\right\} \qquad (4.159)$$

It shows that the generalised Pareto distribution and the GEV distribution are directly related, provided that the number of exceedances per fixed period of time can be modelled by a Poisson distribution.

Example 12: Annual exceedances and annual maxima

As an example consider the exceedances shown above for Chooz on Meuse river. Since there are 72 exceedances in 30 years, the average number of exceedances per year is 72/30 = 2.4, hence λ = 2.4. The comparison of the Poisson distribution with the observed distribution of exceedances N is presented in Figure 4.37.

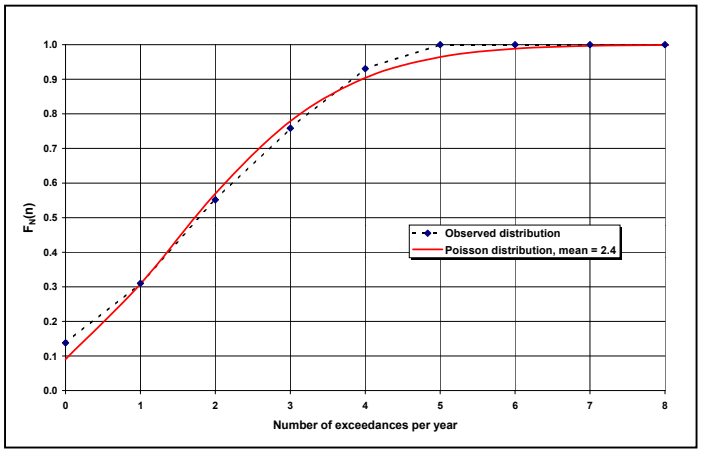


Figure 4.37:
Modelling of number of Q = 400 m^3/s threshold; Meuse river at Chooz

From Figure 4.37 it is observed that in the example case the Poisson distribution is a suitable model for the frequency distribution of the number of exceedances per year.

Summing up

To model the distributions of exceedances, apart from Pareto type distributions, basically any other distribution may be used, provided a proper fit is obtained. Then equation (4.155) or (4.156) is used to compute from such a fit the return period referring to the annual maximum value, consistent with annual maximum series. It follows:

$$T_{AM}(x) = \frac{1}{1-\exp\{-\lambda(1-F_{POT}(x))\}} = \frac{1}{1-\exp\{-(1-F_{AE}(x))\}} \qquad (4.160)$$

Example 12 (continued)

To show the procedure let's follow the Meuse example presented above. The average number of exceedances per year was λ = 2.4. The exceedances are fitted by an exponential distribution. The average discharge of the recorded peak flows exceeding 400 m^3/s is 232.5 m^3/s, hence x_0 = 400 and β = 233, see Sub-section 4.5.1 Hence $F_{POT}(x)$ reads:

$$F_{POT}(x) = 1 - \exp\left(-\left(\frac{x-400}{233}\right)\right) \tag{4.161}$$

The fit of the exponential distribution to the observed frequencies is shown in Figure 4.38.

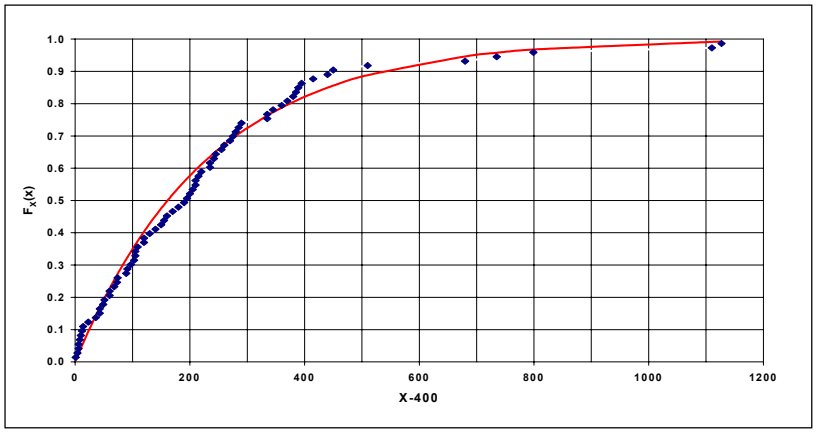


Figure 4.38: Fir of exponential distribution to Meuse flow at Chooz exceeding threshold of 400 m^3/s

From equation (4.155) the cdf of the annual flood discharge then reads:

$$F_{Xmax}(x) = \exp(-\lambda(1-F_X(x))) = \exp(-2.4(1-(1-\exp(-\left(\frac{x-400}{233}\right))) = \\ = \exp(-\exp(-\left(\frac{x-604}{233}\right))) \tag{4.157}$$

The distribution of the annual maximum is seen to have a Gumbel distribution, and for the return period it follows:

$$T(x) = \frac{1}{1-\exp(-\exp(-\left(\frac{x-604}{233}\right)))} \tag{4.158}$$

If the procedure is carried out by applying the Gumbel distribution on annual maximum series for the same period, the parameter values are instead of 604 and 233, respectively 591 and 238. A comparison between both approaches is shown in Figure 4.39. It is observed that both procedures give very similar results (differences <1% for 2<T≤100).

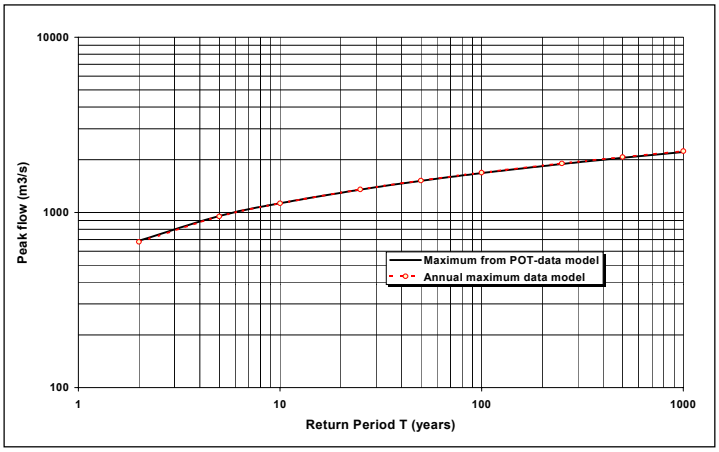


Figure 4.39: Flow extremes as function of return period derived from POT-series transferred to maximum and directly from annual maximum series.

From (4.158) it follows for the quantile x_T:

$$x_T = 604 - 233\ln(\ln\left(\frac{T}{T-1}\right)) \qquad (4.159)$$

According to the conditional distribution it follows from (4.153) and (4.161) with $\lambda = 2.4$ for the quantile x_T:

$$x_T = 400 + 233\ln(2.4T) \qquad (4.160)$$

A comparison between both approaches is seen in Figure 4.40:

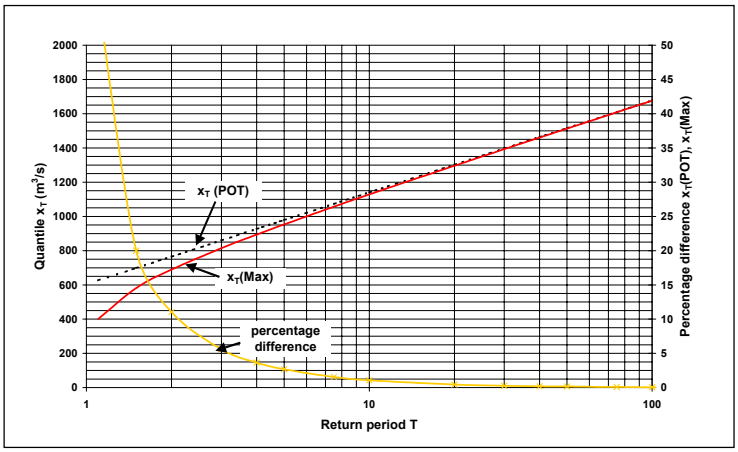


Figure 4.40: Quantiles according to POT and Maximum, both from exceedance series

From Figure (4.40) it is observed that there is a distinct difference between the two approaches for return periods up to about 3 (diff > 5%), at a return periods of 10 the difference in only 1% and reduces thereafter to insignificant differences.

4.7 Sampling distributions

4.7.1 General

A distribution parameter can be estimated from a particular sample in a number of ways. The rule or method used to estimate a parameter is called an **estimator**; the value that the estimator gets, when applied, is called an **estimate**. An estimate of a distribution parameter of a particular series will assume a number of values dependent on the sample taken from the entire population. It is a random variable itself with a particular frequency distribution. Hence, one can only speak about the true value of a parameter in probabilistic terms. Consequently, also the quantiles computed from the frequency distributions are random variables with a particular distribution. Many of the estimated distribution parameters and quantiles are asymptotically normally distributed. This implies that for large sample sizes N the estimate and the standard error fully describe the probability distribution of the statistic. For small sample sizes the sampling distributions may, however, deviate significantly from normality. In addition to the normal distribution important sampling distributions are the Chi-square distribution, the Student-t distribution and the Fisher F-distribution. The normal distribution was described in detail in Sub-section 4.4.1. The latter 3 distributions will be described in the next sub-sections.

4.7.2 Chi-squared distribution

Let Z_1, Z_2, Z_3, …, Z_ν be ν independent standard normal random variables, then the Chi-squared variable χ_ν^2 with ν degrees of freedom is defined as:

$$\chi_\nu^2 = Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_\nu^2 \qquad (4.161)$$

The number of degrees of freedom ν represents the number of independent or 'free' squares entering into the expression. The pdf and cdf are given by (4.65) and (4.66) respectively, which with *X* replaced by χ^2 read:

$$f_{\chi^2}(x) = \frac{1}{2\Gamma(\nu/2)}\left(\frac{x}{2}\right)^{\nu/2-1} \exp(-\frac{x}{2}) \text{ for : } x \geq 0\,,\ \nu > 0 \qquad (4.162)$$

$$F_{\chi^2}(x) = \frac{1}{2\Gamma(\nu/2)}\int_0^x\left(\frac{s}{2}\right)^{\nu/2-1} \exp(-\frac{s}{2})ds \qquad (4.163)$$

The χ^2 -distribution is a particular case of the gamma distribution by putting β = 2 and γ = ν/2 in equations (4.51) and (4.52).The function $f_{\chi 2}$ (x) for different degrees of freedom is depicted in Figure 4.41.

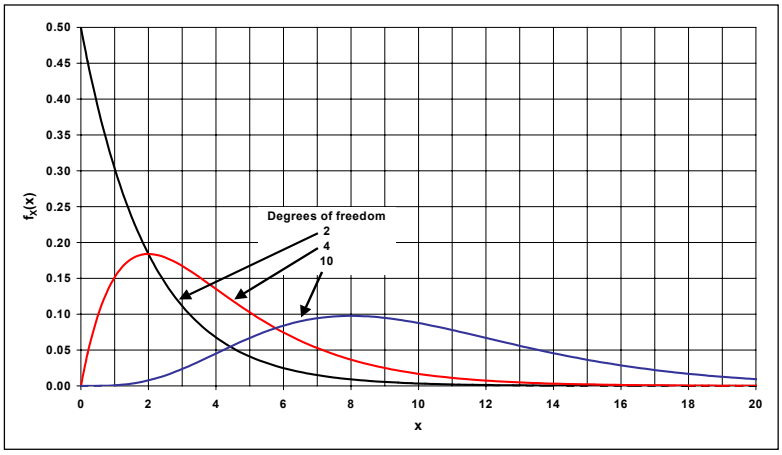


Figure 4.41:
χ_ν^2 -probability density function for ν = 2, 4 and 10 degrees of freedom

Moment related parameters of the distribution

The mean, mode, variance, skewness and kurtosis of the distribution of χ_ν^2 are:

$$\left.\begin{aligned} \mu_{\chi_\nu^2} &= \nu \\ m_{\chi_\nu^2} &= \nu - 2 \text{ for : } \nu \geq 2 \\ \sigma^2_{\chi_\nu^2} &= 2\nu \end{aligned}\right\} \qquad (4.164a)$$

$$\left.\begin{aligned} \gamma_{1,\chi_\nu^2} &= \sqrt{\frac{8}{\nu}} \\ \gamma_{2,\chi_\nu^2} &= 3\left(\frac{\nu+4}{\nu}\right) \end{aligned}\right\} \qquad (4.164b)$$

From (164b) it is observed that for large ν the skewness tends to 0 and the kurtosis becomes 3, and the χ^2 -distribution approaches the normal distribution, with N(ν, 2ν).

It is noted that the addition theorem is valid for the χ^2 –distribution. This implies that a new variable formed by $\chi_\nu^2 = \chi_{\nu 1}^2 + \chi_{\nu 2}^2$ has $\nu = \nu_1 + \nu_2$ degrees of freedom as is simple seen from (4.161). The χ^2 –

distribution is often used for making statistical inference about the variance. An unbiased estimator for the variance reads, see (2.5), with the mean estimated by (2.3):

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)^2 \tag{2.5}$$

The sum term can be written as follows:

$$\sum_{i=1}^{N}(x_i - m_X)^2 = \sum_{i=1}^{N}\{(x_i - \mu_X) - (m_X - \mu_X)\}^2 = \sum_{i=1}^{N}(x_i - \mu_X)^2 - N(m_X - \mu_X)^2 \tag{4.165}$$

When the first term of the last right-hand part is divided by σ_X, then one gets a sum of N standard normal variates; if one divides the second part by the standard deviation of the mean, which is $\sigma_X/\sqrt{N}$ then one standard normal variate is obtained. Hence it follows:

$$\sum_{i=1}^{N}(x_i - m_X)^2 = \sigma_X^2\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\right)^2 - \sigma_X^2\left(\frac{m_X - \mu_X}{\sigma_X/\sqrt{N}}\right)^2 = \sigma_X^2(\chi_N^2 - \chi_1^2) = \sigma_X^2\chi_{N-1}^2 \tag{4.166}$$

Substitution of (4.166) into (2.5) gives:

$$\frac{(N-1)s_X^2}{\sigma_X^2} = \chi_{N-1}^2 \text{ or } :\frac{\nu s_X^2}{\sigma_X^2} = \chi_\nu^2 \text{ with: } \nu = N-1 \tag{4.167}$$

Hence the random variable $\nu s_X^2/\sigma_X^2$ has a χ^2 –distribution with ν = N-1 degrees of freedom. So, the distribution can be used to make statistical inference about the variance. The χ^2 –distribution is also used for statistical tests on the goodness of fit of a theoretical distribution function to an observed one. This will be discussed in Chapter 6.

4.7.3 Student t distribution

The Student t-distribution results from a combination of a normal and a chi-square random variable. Let Y and Z be independent random variables, such that Y has a χ_ν^2 -distribution and Z a standard normal distribution then the variable T_ν is the Student t variable with ν degrees of freedom when defined by:

$$T_\nu = \frac{Z}{\sqrt{Y/\nu}} \tag{4.168}$$

The probability density function of T_ν it follows:

$$f_T(t) = \frac{\Gamma\{(\nu+1)/2\}}{\Gamma(\nu/2)}\frac{1}{\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \tag{4.169}$$

The function $f_T(t)$ for different degrees of freedom is shown in Figure 4.42.

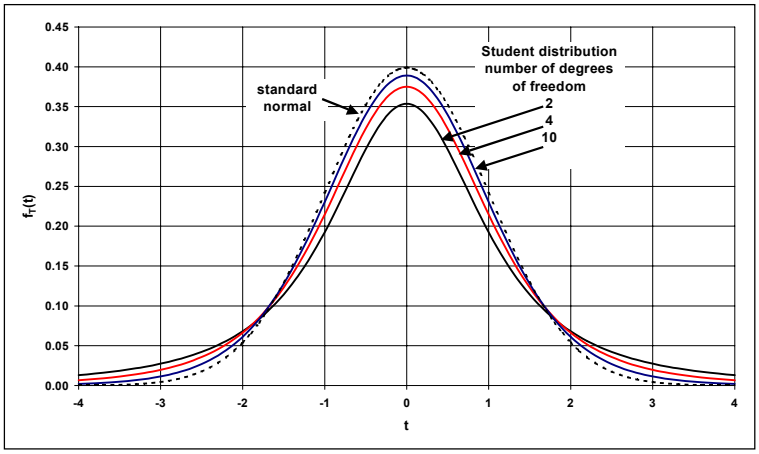


Figure 4.42:
Student t-distribution for ν = 2, 4 and 10 degrees of freedom

Moment related parameters of the distribution

The mean and the variance of the variable T_ν are respectively:

$$\left.\begin{aligned} \mu_T &= 0 \quad \text{for} : \nu > 1 \\ \sigma_T^2 &= \frac{\nu}{\nu - 2} \quad \text{for} : \nu > 2 \end{aligned}\right\} \tag{4.170}$$

The Student t-distribution approaches a standard normal distribution when the number of degrees of freedom becomes large. From (4.170) it is observed that the standard deviation is slightly larger than 1 particularly for small ν. Hence, the dispersion about the mean is somewhat larger than in the standard normal case.

The sampling distribution of the sample mean when the standard deviation is estimated by (2.5) can shown to be a t-distribution as follows. Consider the random variable:

$$\frac{m_X - \mu_X}{s_X / \sqrt{N}} = \left(\frac{m_X - \mu_X}{\sigma_X / \sqrt{N}}\right)\frac{\sigma_X}{s_X} = \left(\frac{m_X - \mu_X}{\sigma_X / \sqrt{N}}\right)\frac{1}{\sqrt{\frac{\chi_\nu^2}{\nu}}} \quad \text{with} : \nu = N - 1 \tag{4.171}$$

The first part of the last term is a standard normal variate, whereas the second part, which followed from (4.167), is the root of a χ^2-variate with ν = N-1 divided by ν. Hence the expression is a T_ν – variate with ν = N-1 degrees of freedom:

$$\frac{m_X - \mu_X}{s_X / \sqrt{N}} = T_\nu \quad \text{with} : \nu = N - 1 \tag{4.172}$$

It will be shown in the next sub-section that the t-distribution is related to the Fisher F-distribution.

4.7.4 Fisher's F-distribution

Let X and Y de independent random variables, both distributed as χ^2 with respectively ν_1 and ν_2 degrees of freedom, then the random variable F defined by:

$$F = \frac{X / \nu_1}{Y / \nu_2} \tag{4.173}$$

has a so called F-distribution, which probability density function reads:

$$h_F(f) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{f^{(\nu_1/2)-1}}{(1+\frac{\nu_1}{\nu_2}f)^{(\nu_1+\nu_2)/2}} \qquad (4.174)$$

With the definition of the beta function $B(\alpha,\beta)$:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad (4.175)$$

equation (4.174) may also be written as:

$$h_F(f) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B(\frac{\nu_1}{2},\frac{\nu_2}{2})} \frac{f^{(\nu_1/2)-1}}{(1+\frac{\nu_1}{\nu_2}f)^{(\nu_1+\nu_2)/2}} \qquad (4.176)$$

The pdf is shown in Figure 4.43:

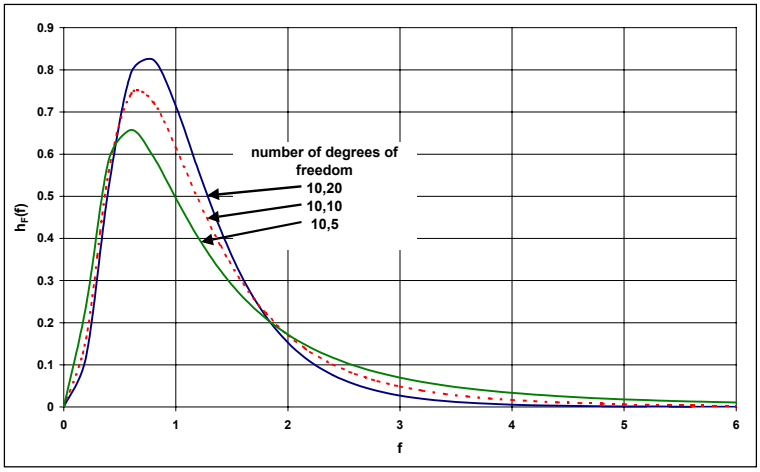


Figure 4.43:
Fisher F-probability density function for various degrees of freedom

The F-distribution is also called the variance-ratio distribution as from the definition of the F-variable (4.173) combined with (4.167) can be observed. Hence, if we consider m respectively n observations from two standard normal random variables Z_1 and Z_2 with variances σ_1^2 and σ_2^2 estimated according to (2.5) by s_1^2 and s_2^2 then the ratio:

$$\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = F_{m-1,n-1} \qquad (4.177)$$

has an F-distribution with (m-1,n-1) degrees of freedom. The F-distribution is thus particularly suited for variance ratio tests. From a comparison of (4.173) with (4.167) it is observed that the root of an F-variate with $(1,\nu)$ degrees of freedom has a Student t-distribution

5 Estimation of Statistical Parameters

5.1 General

To apply the theoretical distribution functions dealt with in the previous chapter the following steps are required:

1. Investigate the homogeneity of the data series, subjected to frequency analysis
2. Estimate the parameters of the postulated theoretical frequency distribution
3. Test the goodness of fit of the theoretical to the observed frequency distribution

In this chapter the second step will be dealt with. The objective of representing the observed frequency distribution by a theoretical one is to increase its mathematical tractability, and to facilitate extrapolation. The procedure in itself is no more than curve fitting. It involves the estimation of the parameters of a theoretical distribution function based on a sample from the population. It implies that the sample values of the parameters are stochastic variables themselves with a frequency distribution, called the sampling distribution as discussed in Chapter 4. The parameters can be estimated in various ways including:

1. Graphical method, and
2. Analytical methods, like:
    - Method of moments
    - Maximum likelihood method
    - Method of least squares
    - Mixed moment-maximum likelihood method, etc.

Estimation error

The parameters estimated with the above methods differ. To compare the quality of different estimators of a parameter, some measure of accuracy is required. The following measures are in use:

- mean square error and root mean square error
- error variance and standard error
- bias
- efficiency
- consistency

Mean square error

A measure for the quality of an estimator is the **mean square error,** mse. It is defined by:

$$\text{mse} = E[(\phi - \Phi)^2] \tag{5.1}$$

where ϕ is an estimator for Φ.

Hence, the mse is the average of the squared differences between the sample value and the **true** value. Equation (5.1) can be expanded to the following expression:

$$\text{mse} = E[(\phi - E[\phi])^2] + E[(E[\phi] - \Phi)^2] \tag{5.2}$$

Since:

$$E\left[(\phi - E[\phi])^2\right] = \sigma_\phi^2 \tag{5.3}$$

and:

$$E[(E[\phi] - \Phi)^2] = b_\phi^2 \tag{5.4}$$

it follows that:

$$\text{mse} = \sigma_\phi^2 + b_\phi^2 \tag{5.5}$$

The mean square error is seen to be the sum of two parts:

- the first term is the **variance** of ϕ, equation (5.3), i.e. the average of the squared differences between the sample value and the **expected** mean value of ϕ based on the sample values, which represents the **random** portion of the error, and
- the second term of (5.5) is the square of the **bias** of ϕ, equation (5.4), describing the systematic deviation of expected mean value of ϕ from its true value Φ, i.e. the **systematic** portion of the error.

Note that if the bias in ϕ is zero, then mse = σ_ϕ^2. Hence, for **unbiased** estimators, i.e. if systematic errors are absent, the mean square error and the variance are equivalent. If mse(ϕ_1) < mse (ϕ_2) then ϕ_1 is said to be **more efficient** than ϕ_2 with respect to Φ.

Root mean square error

Instead of using the mse it is customary to work with its square root to arrive at an error measure, which is expressed in the same units as Φ, leading to the **root mean square (rms) error**:

$$\text{rmse} = \sqrt{E[(\phi - \Phi)^2]} = \sqrt{\sigma_\phi^2 + b_\phi^2} \tag{5.6}$$

Standard error

When discussing the frequency distribution of statistics like of the mean or the standard deviation, for the standard deviation σ_ϕ the term **standard error** is used, e.g. standard error of the mean and standard error of the standard deviation, etc.

$$\sigma_\phi = \sqrt{E[(\phi - E[\phi])^2]} \tag{5.7}$$

In Table 5.1, a summary of unbiased estimators for moment parameters is given, together with their standard error. With respect to the latter it is assumed that the sample elements are **serially uncorrelated**. If the sample elements are serially correlated a so-called **effective number of data** N_{eff} has to be applied in the expressions for the standard error in Table 5.1

Consistency

If the probability that ϕ approaches Φ becomes unity if the sample becomes large then the estimator is said to be consistent or asymptotically unbiased:

$$\lim_{n\to\infty} \Pr ob(|\varphi - \Phi| > \varepsilon) = 0 \text{ for any } \varepsilon \tag{5.8}$$

To meet this requirement it is sufficient to have a zero mean square error in the limit for $n\to\infty$.

5.2 Graphical estimation

In graphical estimations, the variate under consideration is regarded as a function of the standardised or reduced variate with known distribution. With a properly chosen probability scale a linear relationship can be obtained between the variate and the reduced variate representing the transformed probability of non-exceedance. Consider for this the Gumbel distribution. From (4.108) it follows:

$$x = x_0 + \beta z \tag{5.9}$$

According to the Gumbel distribution the reduced variate z is related to the non-exceedance probability by:

$$z = -\ln(-\ln(F_X(x)) \tag{5.10}$$

To arrive at an estimate for x_0 and β we plot the ranked observations x_i against z_i by estimating the non-exceedance probability of x_i, i.e. F_i. The latter is called the plotting position of x_i, i.e. the probability to be assigned to each data point to be plotted on probability paper. Basically, appropriate plotting positions depend on the distribution function one wants to fit the observed distribution function to. A number of plotting positions has been proposed, which is summarised in Table 5.4.To arrive at an unbiased plotting position for the Gumbel distribution Gringorten's plotting position has to be applied, which reads:

$$F_i = \frac{i - 0.44}{N + 0.12} \tag{5.11}$$

This non-exceedance frequency is transformed into the reduced variate z_i by using (5.10). If the data x_i are from a Gumbel distribution then the plot of x_i versus z_i will produce approximately a straight line. The slope of the line gives an estimate for the parameter β and the intercept is x_0. Hence the steps involved are as follows:

1. Rank the observations in ascending order, i = 1 is the smallest and i = N the largest
2. Compute the non-exceedance frequency F_i of x_i using (5.11)
3. Transform F_i into z_i using equation (5.10)
4. Plot x_i versus z_i and draw a straight line through the points
5. Estimate the slope of the line and the intercept at z=0 to get estimates for β and the intercept is x_0

The same steps apply to other frequency distributions, though with different plotting positions.

Parameter	Estimator	Standard error	Remarks
Mean	$m_Y = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\sigma_{m_Y} = \frac{\sigma_Y}{\sqrt{N}}$	The sampling distribution of m_Y is very nearly normal for N>30, even when the population is non-normal. In practice σ_Y is not known and is estimated by s_Y. Then the sampling distribution of m_Y has a Student distribution, with N-1 degrees of freedom
Variance	$s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2$	$\sigma_{s_Y^2} = \sqrt{\frac{2}{N}}\sigma_Y^2$	Expression applies if the distribution of Y is approximately normal. The sampling distribution of s_Y^2 is nearly normal for N>100. For small N the distribution of s_Y^2 is chi-square (χ^2), with N-1 degrees of freedom
Standard deviation	$s_Y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2}$	$\sigma_{s_Y} = \frac{\sigma_Y}{\sqrt{2N}}$	The remarks made for the standard error of the variance apply here as well
Coefficient of variation	$\hat{CV}_Y = \frac{s_Y}{m_Y}$ Sample value of CV_Y limited to: $CV_Y < \sqrt{(N-1)}$	$\sigma_{\hat{CV}} = \frac{\frac{\sigma_Y}{\mu_Y}}{\sqrt{2N}}\sqrt{1+2\left(\frac{\sigma_Y}{\mu_Y}\right)^2}$	This result holds if Y being normally or nearly normally distributed and N>100.
Covariance	$\hat{C}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y)$		
Correlation coefficient	$r_{XY} = \frac{C_{XY}}{s_X s_Y}$	$\sigma_W = \frac{1}{\sqrt{N-3}}$ where $W = \frac{1}{2}\ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right)$	Rather then the standard error of r_{XY} the standard error of the transformed variable W is considered. The quantity W is approximately normally distributed for N>25.
Lag one auto-correlation coefficient	$r_{YY}(1) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(y_i - m_Y)(y_{i+1} - m_Y)}{s_Y^2}$	as for r_{XY} above	
Skewness	$g_Y = \frac{\frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}(y_i - m_Y)^3}{s_Y^3}$ Skewness limited to: $g_Y < \frac{N-2}{\sqrt{N-1}}$	$\sigma_{g_Y} = \sqrt{\frac{6}{N}}$	A reasonably reliable estimate of the skewness requires a large sample size. Standard error applies if Y is normally distributed.
Quantiles	1. first rank the sample values in ascending order: $y_{(i)} < y_{(i+1)}$ 2. next assign to each ranked value a non-exceedance probability i/(N+1) 3. then interpolate between the probabilities to arrive at the quantile value $\hat{y}_p$ of the required non-exceedance level	$\sigma_{\hat{y}_p} = \frac{1}{f_Y(y_p)}\sqrt{\frac{p(1-p)}{N}}$ $\sigma_{\hat{y}_p} = \frac{\beta}{\sqrt{N}}\sigma_Y$	The denominator is derived from the pdf of Y. If Y is normally distributed then the standard error of the quantile is determined by the second expression. The coefficient β depends on the non-exceedance probability p. For various values of p the value of β can be obtained from Table 5.2.

Table 5.1: Estimators of sample parameters with their standard error

p	0.5	0.4/ 0.6	0.3/ 0.7	0.25/0.75	0.2/ 0.8	0.15/0.85	0.1/0.9	0.05/0.95
β	1.253	1.268	1.318	1.362	1.428	1.531	1.709	2.114

Table 5.2: $\beta(p)$ for computation of σ of quantiles if Y is normally distributed

Example 5.1: Graphical estimation of distribution parameters

Above procedure is shown for annual maximum river flows of the Meuse river at Chooz for the period 1968-1997 presented in Example 4.12. In Table 5.3 the peak flows are presented in Column 2. In Column 4 the ranked discharges are presented in ascending order. Subsequently the non-exceedance frequency F_i of x_i is presented in Column 5, derived from equation (5.11), whereas in the last column the reduced variate z_i referring to the non-exceedance frequency F_i.

Year	Q_{max}	Rank	x_i	Freq	z_i	Year	Q_{max}	Rank	x_i	Freq	z_i
1	2	3	4	5	6	1	2	3	4	5	6
1968	386	1	274	0.019	-1.383	1983	1199	16	685	0.517	0.415
1969	910	2	295	0.052	-1.085	1984	675	17	690	0.550	0.514
1970	550	3	386	0.085	-0.902	1985	760	18	735	0.583	0.617
1971	274	4	406	0.118	-0.759	1986	735	19	760	0.616	0.725
1972	468	5	406	0.151	-0.635	1987	780	20	780	0.649	0.840
1973	406	6	423	0.185	-0.524	1988	660	21	785	0.683	0.963
1974	615	7	468	0.218	-0.421	1989	690	22	795	0.716	1.096
1975	295	8	491	0.251	-0.324	1990	1080	23	840	0.749	1.241
1976	795	9	550	0.284	-0.230	1991	491	24	860	0.782	1.404
1977	685	10	615	0.317	-0.138	1992	1135	25	910	0.815	1.589
1978	680	11	635	0.351	-0.047	1993	1510	26	1080	0.849	1.807
1979	785	12	642	0.384	0.043	1994	1527	27	1135	0.882	2.073
1980	635	13	660	0.417	0.134	1995	406	28	1199	0.915	2.421
1981	860	14	675	0.450	0.226	1996	642	29	1510	0.948	2.934
1982	840	15	680	0.483	0.319	1997	423	30	1527	0.981	3.976

Table 5.3: Annual maximum river flows of Meuse river at Chooz, period 1968-1997

The Columns 6 and 4 are plotted in Figure 5.1. It is observed that the points are located on a straight line, which indicates that the Gumbel distribution is applicable to data set of annual maximum riverflows in this case. The slope of the line is estimated at 1200/4.85 = 247 and the intercept at z = 0 is about 590 m^3/s, which are the estimates for β and x_0 respectively.

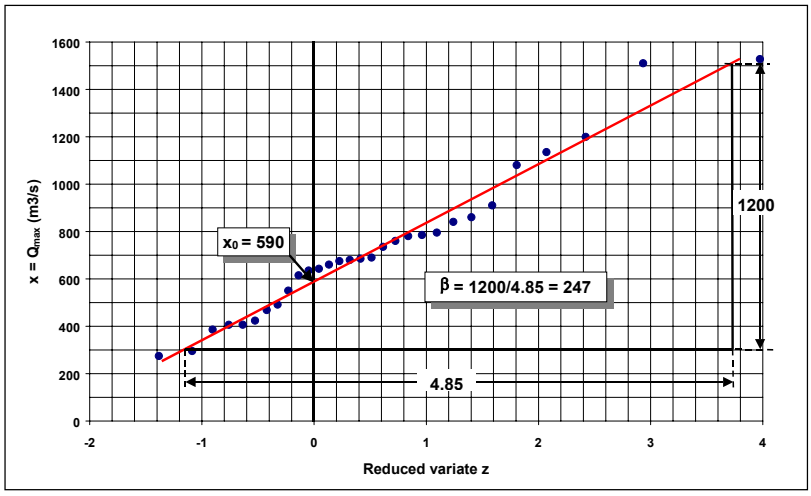


Figure 5.1: Application of graphical estimation method to annual maximum river flows of Meuse river at Chooz, period 1968-1997

In Chapter 4 Example 4.12 the parameters were estimated using the maximum likelihood method (MLM), which gave estimates for β and x_0 respectively of 238 and 591 m^3/s. For a 100 year return period flood (T = 100 years, i.e. $F_X(x) = 1 - 1/100 = 0.99$ or $z = -\ln(-\ln(0.99)) = 4.60$) the quantile $x_{T=100}$ becomes with the two methods using (5.9):

Graphical method: $x_{100} = 590 + 247 \times 4.60 = 1726\ m^3/s$

MLM: $x_{100} = 591 + 238 \times 4.60 = 1686\ m^3/s$

It is observed that the difference between the methods in this case is very small.

There is in the graphical method, however, a strong subjective element. Different analysts may obtain different results. This method is therefore not suitable for final design calculations. Plotting of the observed frequency distribution with the fitted one is extremely important. Before accepting a theoretical frequency distribution to be applicable to an observed frequency distribution inspection of the frequency plot is a must. Such a comparison gives you a visual impression about the goodness of fit particularly at the lower and upper end of the curve, something a statistical test does not give. In this respect it is of importance to apply the appropriate plotting position for each of the frequency distributions to arrive at an unbiased plotting position.

Plotting positions

Defining the plotting position for each data point, when put in ascending order, by:

$$F_i = \frac{i - b}{N - 2b + 1} \qquad (5.12)$$

where: F_i = non-exceedance frequency of x_i ranked in ascending order

i = i^{th} element in ranked sequence in ascending order

N = number of data in series

b = parameter dependent on type of distribution

Cunnane (1978) investigated various plotting positions that can be derived from (5.12) by assuming an appropriate value for b. Two criteria were used:

- unbiasedness, which implies that for a large number of equally sized samples the average of the plotted points for each i will fall on the theoretical line
- minimum variance, i.e. the variance of the plotted point about the theoretical line is minimum.

It appears that the often-used Weibull plotting position with b = 0 gives a biased result, plotting the largest values at a too low return period. Some of his results and those of NERC (1975) are summarised in Table 5.4.

Name of formula	b	distribution	remarks
Hazen	0.5	-	For i = N: T = 2N
Weibull	0	-	biased
Blom	3/8	N, LN-2, LN-3, G-2 for large γ	LP-3: for γ_1>0 b>3/8 and γ_1<0 b<3/8
Chegodayev	0.3	various	Overall compromise
Gringorten	0.44	EV-1, E-1, E-2, G-2	
NERC	2/5	G-2, P-3	Compromise plotting position
Tukey	1/3	-	

Table 5.4: Plotting position formula (Cunnane, 1978; NERC, 1975)

In HYMOS the parameter b can be set to the requirement; the default value is b = 0.3.

5.3 Parameter estimation by method of moments

The method of moments makes use of the fact that if all the moments of a frequency distribution are known, then everything about the distribution is known. As many moments as there are parameters are needed to define the distribution. The frequency distributions discussed in Chapter 4 contain at maximum four parameters, hence the first four moments, generally represented by the mean, variance, skewness and kurtosis, are at maximum required to specify the distribution and to derive the distribution parameters. Most distributions, however, need only one, two or three parameters to be estimated. It is to be understood that the higher the order of the moment the larger the standard error will be.

In HYMOS the unbiased estimators for the mean, variance, skewness and kurtosis as presented by equations (2.3), (2.5) or (2.6), (2.8) and (2.9) are used, see also Table 5.1. Substitution of the required moments in the relations between the distribution parameters and the moments will provide the moment estimators:

- Normal distribution: the two parameters are the mean and the standard deviation, which follow from (2.3) and (2.6) immediately
- LN-2: equations (2.3) and (2.6) substituted in (4.28) and (4.29)
- LN-3: equations (2.3), (2.6) and (2.8) substituted in (4.31) to (4.34)
- G-2: equations (2.3) and (2.6) substituted in (4.61) and (4.62)
- P-3: equations (2.3), (2.6) and (2.8) substituted in (4.71) to (4.73)
- EV-1:equations (2.3) and (2.6) substituted in (4.115) and (4.116)

For all other distributions the method of moments is not applied in HYMOS.

Biased-unbiased

From (2.5) it is observed that the variance is estimated from:

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)^2 \qquad (2.5)$$

The denominator (N-1) is introduced to obtain an unbiased estimator. A straightforward estimator for the variance would have been:

$$\hat{s}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)^2 \qquad (5.13)$$

The expected value of this estimator, in case the x_i's are *independent*, is:

$$E[\hat{s}_X^2] = \frac{1}{N}\sum_{i=1}^{N}E[(x_i - m_X)^2] = E[\left((x_i - \mu_X) - (m_X - \mu_X)\right)^2] = \sigma_X^2 - \frac{\sigma_X^2}{N} = \left(\frac{N-1}{N}\right)\sigma_X^2 \qquad (5.14)$$

From equation (5.14) it is observed that although the estimator is consistent, it is biased. Hence, to get an unbiased estimator for σ_X^2 the moment estimator should be multiplied by N/(N-1), which leads to (2.5)

Remark

The method of moments provides a simple procedure to estimate distribution parameters. For small sample sizes, say N < 30, the sample moments may differ substantially from the population values. Particularly if third order moments are being used to estimate the parameters, the quality of the

parameters will be poor if the sample size is small. In such cases single outliers will have a strong effect on the parameter estimates.

Probability weighted moments and L-moments

The above method of moments is called Product Moments. The negative effects the use of higher moments have on the parameter estimation is eliminated by making use of L-moments, which are linear functions of **probability weighted moments** (PWM's). Probability weighted moments are generally defined by:

$$M_{p,r,s} = E\left[X^p \{F_X(x)\}^r \{1-F_X(x)\}^s\right] \tag{5.15}$$

By choosing p=1 and s=0 in (5.15) one obtains the r^{th} PWM, which reads:

$$\beta_r = E\left[X\{F_X(x)\}^r\right] = \int_{-\infty}^{\infty} x\{F_X(x)\}^r f_X(x)dx \tag{5.16}$$

Comparing this expression with the definition of moments in (3.23) it is observed that instead of raising the variable to a power ≥ 1 now the cdf is raised to a power ≥ 1. Since the latter has values < 1, it is observed that these moments are much less sensitive for outliers, which in the case of product moments strongly affect the moments and hence the parameters to be estimated.

L-moments are developed for order statistics. Let the X_I's be independent random variables out of a eries of sample of size N, which are put in ascending order:

$X_{1:N}$ $X_{2:N}$ <….<$X_{N:N}$

then $X_{i:N}$ is the i^{th} largest in a random sample of N, and is known as the **i^{th} order statistic**. L-moments are used to characterize the distribution of order statistics. In practice the first four L-moments are of importance:

$$\begin{aligned}
\lambda_1 &= E[X] \\
\lambda_2 &= \frac{1}{2}\{E[X_{2:2}] - E[X_{1:2}]\} \\
\lambda_3 &= \frac{1}{3}\{E[X_{3:3}] - 2E[X_{2:3}] + E[X_{1:3}]\} \\
\lambda_4 &= \frac{1}{4}\{E[X_{4:4}] - 3E[X_{3:4}] + 3E[X_{2:4}] - E[X_{1:4}]\}
\end{aligned} \tag{5.17}$$

The first moment is seen to be the mean, the second a measure of the spread or scale of the distribution, the third a measure of asymmetry and the fourth a measure of peakedness. Dimensionless analogues to the skewness and kurtosis are (Metcalfe, 1997):

$$\begin{aligned}
L-\text{skewness}: \tau_3 &= \frac{\lambda_3}{\lambda_2} \text{ with}: -1 < \tau_3 < 1 \\
L-\text{kurtosis}: \tau_4 &= \frac{\lambda_4}{\lambda_2} \text{ with}: \frac{1}{4}(5\tau_3^2 - 1) \leq \tau_4 < 1
\end{aligned} \tag{5.18}$$

The relation between the L-moments and parameters of a large number of distributions are presented in a number of statistical textbooks. For some distributions they are given below (taken from Dingman, 2002):

- Uniform distribution

$$\begin{aligned}\lambda_1 &= \frac{a+b}{2}\\ \lambda_2 &= \frac{b-a}{6}\\ \tau_3 &= 0\\ \tau_4 &= 0\end{aligned} \qquad (5.19)$$

- Normal distribution

$$\begin{aligned}\lambda_1 &= \mu_X\\ \lambda_2 &= \frac{\sigma_X}{\sqrt{\pi}}\\ \tau_3 &= 0\\ \tau_4 &= 0.1226\end{aligned} \qquad (5.20)$$

- Gumbel

$$\begin{aligned}\lambda_1 &= x_0 + 0.5772\,\beta\\ \lambda_2 &= 0.6931\,\beta\\ \tau_3 &= 0.1699\\ \tau_4 &= 0.1504\end{aligned} \qquad (5.21)$$

So to estimate the parameters of a distribution estimates of L-moments are required. From (5.17) it is observed that to estimate the L-moments all possible combinations of samples of size 2, 3 and 4 have to be selected to arrive at the expected value of the various order statistics. This is a rather cumbersome exercise. However, the L-moments can be related to the probability weighted moments as follows:

$$\begin{aligned}\lambda_1 &= \beta_0\\ \lambda_2 &= 2\beta_1 - \beta_0\\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0\\ \lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0\end{aligned} \qquad (5.22)$$

The sample estimates of the probability weighted moments follow from the ordered set of data:

$$\begin{aligned}b_0 &= \frac{1}{N}\sum_{i=1}^{N} x_{i:N} = m_X\\ b_1 &= \frac{1}{N(N-1)}\sum_{i=2}^{N}(i-1)x_{i:N}\\ b_2 &= \frac{1}{N(N-1)(N-2}\sum_{i=3}^{N}(i-1)(i-2)x_{i:N}\\ b_3 &= \frac{1}{N(N-1)(N-2)(N-3)}\sum_{i=4}^{N}(i-1)(i-2)(i-3)x_{i:N}\end{aligned} \qquad (5.23)$$

Example 5.1: continued

The L-moments method is applied to the annual maximum river flows of Meuse river at Chooz. The computation of the probability weighted moments is presented in Table 5.5. Note that first the data are ordered. The ordered series is presented in Column 2. In Column 3 the numerical values of $(i - 1)x_{i:N}$ is presented, which is the sum term in the derivation of b_1; similarly the columns 4 and 5 contain the sum-terms for the derivation of b_2 and b_3. The values in the columns are summed and subsequently divided by N, N(N-1), N(N-1)(N-2) and N(N-1)(N-2)(N-3) respectively to arrive at the estimates for the probability weighted moments b_0, b_1, b_2 and b_3, according to equation (5.23).

Rank	Q-max	C-b1	C-b2	C-b3
1	274			
2	295	295		
3	386	772	772	
4	406	1217	2435	2435
5	406	1624	4872	9744
6	423	2117	8467	25402
7	468	2808	14040	56160
8	491	3437	20622	103110
9	550	4400	30800	184800
10	615	5535	44280	309960
11	635	6350	57150	457200
12	642	7066	70656	635908
13	660	7920	87120	871200
14	675	8775	105300	1158300
15	680	9520	123760	1485120
16	685	10275	143850	1870050
17	690	11040	165600	2318400
18	735	12495	199920	2998800
19	760	13680	232560	3720960
20	780	14820	266760	4534920
21	785	15700	298300	5369400
22	795	16695	333900	6344100
23	840	18480	388080	7761600
24	860	19780	435160	9138360
25	910	21840	502320	11051040
26	1080	27000	648000	14904000
27	1135	29511	737776	17706624
28	1199	32373	841698	21042450
29	1510	42270	1141295	29673680
30	1527	44295	1240273	33487375
Sum	21898	392090	8145767	177221098
Parameters	b_0	b_1	b_2	b_3
	729.92	450.68	334.39	269.45

Table 5.5: Annual maximum river flows of Meuse river at Chooz, period 1968-1997

From the probability weighted moments one can derive the L-moments, with the aid of equation (5.22) as follows. If the estimates for λ are indicated by L then:

$L_1 = b_0 = 729.92$

$L_2 = 2b_1 – b_0 = 2x450.68 – 729.92 = 171.44$

$L_3 = 6b_2 – 6b_1 + b_0 = 6x334.39 – 6x450.68 + 729.92 = 32.18$

$L_4 = 20b_3 - 30b_2 + 12b_1 – b_0 = 20x269.45 – 30x334.39 + 12x450.68 – 729.92 = 35.54$

The parameters of the Gumbel distribution can be obtained through equation (5.21):

$$\hat{\beta} = \frac{L_2}{0.6931} = \frac{171.44}{0.6931} = 247$$

$$\hat{x}_0 = L_1 - 0.5772\hat{\beta} = 729.92 - 0.5772x247.35 = 587$$

$$\hat{\tau}_3 = \frac{L_3}{L_2} = \frac{32.18}{171.44} = 0.19$$

$$\hat{\tau}_4 = \frac{L_4}{L_2} = \frac{35.54}{171.44} = 0.21$$

With the product moment method one obtains for the two parameters respectively 244 and 589 and with the MLM-method 238 and 591. Hence the 100-year flood derived with the various methods becomes:

Product moments: 589 + 244x4.6 = 1711 m^3/s

L-moments: 587 + 247x4.6 = 1723 m^3/s

MLM-method: 591 + 238x4.6 = 1686 m^3/s

The 100-year flood values are seen to be very close to each other. The values for the L-skewness and L-kurtosis of 0.19 and 0.21, respectively, are close to their theoretical values of 0.17 and 0.15 for the Gumbel distribution, which shows that the distribution is an appropriate model for the data set. Charts have been designed where L-skewness and L-kurtosis are plotted against each other for various distributions to guide the selection of a distribution, see also Figure 5.2.

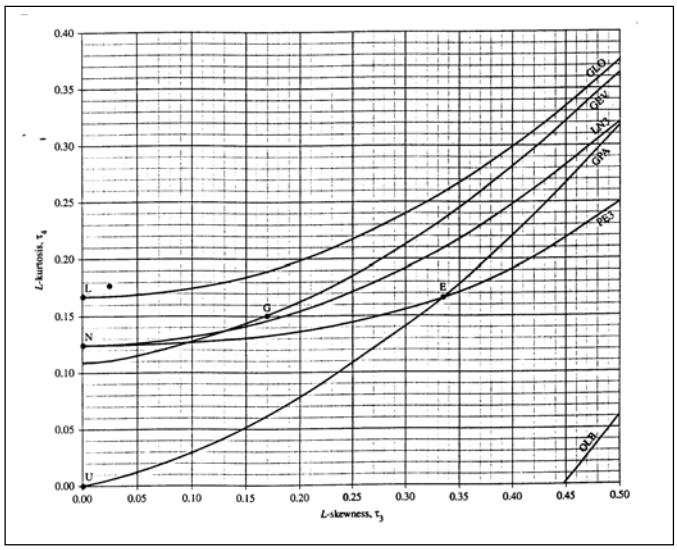


Figure 5.2:
L-moment diagram

(source: Dingman, 2002)

Note

By definition of the probability weighted moments and by close observation of Table 5.5 it is noticed that in the estimation of the probability weighted moments larger weight is given to the higher ranked values in the data set. Hence, the method is biased towards the larger values, particularly when more than 2 parameters have to be estimated. So, though the method is less sensitive to outliers than the product moment method, its application also has its drawbacks.

5.4 Parameter estimation by maximum likelihood method

The Maximum Likelihood method (MLM) was developed by R.A. Fisher in 1922. It is based on the idea that the best estimators for a (set of) parameter(s) are those, which give the greatest probability that precisely the sample series is obtained with the set of parameters. Let X be a random variable with pdf $f_X(x)$, with parameters α_1, α_2, …, α_k. The sample taken out of X is x_i, i=1, 2, …, N. Making the basic assumption that the sample values are **independent and identically distributed**, then with the parameter set α the probability that the random variable will fall in the interval including x_i is $f_X(x_i|\alpha)dx$. So, the joint probability of the occurrence of the sample set x_i, i=1, 2, …, N is, in view of their independence, equal to the product:

$$f_X(x_1 \mid \alpha)dx.f_X(x_2 \mid \alpha)dx.....f_X(x_N \mid \alpha)dx = \left(\prod_{i=1}^{N} f_X(x_i \mid \alpha)\right)dx^N$$

Since all dx are equal, maximising the joint probability simply implies the maximisation of the product:

$$L(x \mid \alpha) = \prod_{i=1}^{n} f_X(x_i \mid \alpha) \tag{5.24}$$

L is called the likelihood function. Then the best set of parameters α are those which maximise L. Hence the estimators for the parameters α_1, α_2, …, α_k are found from:

$$\frac{\partial L(x \mid \alpha)}{\partial \alpha_i} = 0 \text{ for : } i = 1, 2, 3, ..., k \tag{5.25}$$

The estimators obtained in this way are called Maximum Likelihood estimators. Instead of using the likelihood function itself it is usually more convenient to maximise its logarithm in view of the many distributions of the exponential type. Therefore instead of (5.25) the log-likelihood function lnL is usually maximised:

$$\frac{\partial \ln L(x \mid \alpha)}{\partial \alpha_i} = 0 \text{ for : } i = 1, 2, 3, ..., k \tag{5.26}$$

This has the advantage of replacing the products by sum-terms.

Application to lognormal distribution

The procedure will be shown for getting estimators for the lognormal-2 distribution, LN-2.

From (4.26) the likelihood function for a sample x_i, i=1, 2, …, N reads:

$$L(x \mid \mu_Y, \sigma_Y) = \prod_{i=1}^{N} \frac{1}{x_i \sigma_Y \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x_i - \mu_Y}{\sigma_Y}\right)^2\right) \tag{5.27}$$

Hence, the log-likelihood function reads:

$$\ln L(x \mid \mu_Y, \sigma_Y) = -\sum_{i=1}^{N}(\ln x_i) - \frac{N}{2}\ln \sigma_Y^2 - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma_Y^2}\sum_{i=1}^{N}(\ln x_i - \mu_Y)^2 \tag{5.28}$$

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu_Y} &= -\frac{1}{2\sigma_Y^2}\sum_{i=1}^{N} 2(\ln(x_i - \mu_Y)(-1) = \frac{1}{\sigma_Y^2}\left(\sum_{i=1}^{N}(\ln x_i) - N\mu_Y\right) = 0 \\ \frac{\partial \ln L}{\partial \sigma_Y^2} &= -\frac{N}{2}\frac{1}{\sigma_Y^2} - \frac{1}{2}\left(\sum_{i=1}^{N}(\ln x_i - \mu_Y\right)^2\left(-\frac{1}{(\sigma_Y^2)^2}\right) = 0 \end{aligned} \tag{5.29}$$

From above equations the MLM estimators for μ_Y and σ_Y^2 become respectively:

$$\hat{\mu}_Y = \frac{1}{N}\sum_{i=1}^{N} \ln x_i \qquad (5.30)$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\sum_{i=1}^{N} (\ln x_i - \hat{\mu}_Y)^2 \qquad (5.31)$$

From (5.30) and (5.31) it is observed that the MLM estimators are equivalent to the first moment about the origin and the second moment about the mean of ln(x). In a similar manner for the 3-parameter lognormal distribution the estimators for the distribution parameters can be derived, however, at the expense of more complicated equations. As is discussed in Sub-section 5.6 mixed moment-maximum likelihood estimators are preferred when a third parameter (generally the shift or location parameter) is to be estimated particularly when the sample sizes are small.

For the other distribution functions the MLM procedure can also easily be developed along the same lines as discussed for the lognormal distribution, though their solutions are sometimes cumbersome. Reference is made to the HYMOS manual for a description of the formulas used.

5.5 Parameter estimation by method of least squares

The graphical estimation procedure explained in Subsection 5.3 by drawing a line through the data points of the variable x and the reduced variate z can also be done applying linear regression, with z the independent variable and x the dependent variable. The parameters then follow from a minimisation of the sum of squared differences. Such a procedure does not suffer from subjectivity as the graphical method does. The procedures for regression analysis are dealt with in detail in Module 37.

Example 5.1: continued

The annual maximum flows presented in column 4 of Table 5.3 are regressed against the reduced variate z shown in column 6. From linear regression the following estimates for the parameters are obtained (with standard error): x_0 = 589 ± 10.8 and β = 250 ± 8.0, values which are very close to those obtained from the graphical method.

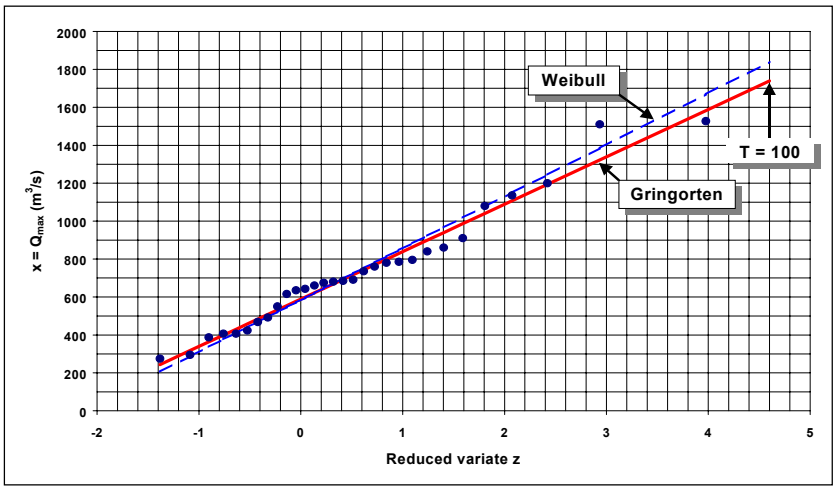


Figure 5.3: Fitting annual maximum flows by regression on reduced variate

If instead of Gringorten's plotting position Weibull's plotting position would have been used, the result would have been: x_0 = 584 ± 11.3 and β = 273 ± 9.2. The T=100 year floods from these procedures would have been for:

- Gringorten: x_{100} = 1739 m^3/s and
- Weibull: x_{100} = 1840 m^3/s

The difference with the MLM estimate are respectively: 3% and 9%. It is observed that the Weibull procedure leads to considerably higher quantile values. This is due to the fact that this method assigns a relatively low return period to the largest values. As a result, the slope of the regression line (i.e. β) will be larger, and so will be the quantiles.

5.6 Parameter estimation by mixed moment-maximum likelihood method

For frequency distributions with a location parameter often the MLM method performs poorly particularly when the sample series is small, like for LN-3 and P-3. In such cases estimating one parameter from a moment relation and the rest with the MLM procedure provides much better parameter estimators, as can be shown by means of Monte Carlo simulations.

The procedure will be shown for LN-3. For given location parameters the MLM estimators for μ_Y and σ_Y^2 become similar to (5.30) and (5.31) with x replaced by $x-x_0$ respectively:

$$\hat{\mu}_Y = \frac{1}{N}\sum_{i=1}^{N} \ln(x_i - x_0) \tag{5.32}$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\ln(x_i - x_0) - \hat{\mu}_Y\right)^2 \tag{5.33}$$

Next, the first moment relation for the lognormal distribution is taken, (4.27a), to arrive at a value for x_0:

$$x_0 = \hat{\mu}_X - \exp\left(\hat{\mu}_Y + \frac{\hat{\sigma}_Y^2}{2}\right) \tag{5.34}$$

The location parameter x_0 is solved iteratively from a modified form of (5.34) as follows:

$$g(x_0) = \hat{\mu}_Y + \frac{\hat{\sigma}_Y^2}{2} - \ln(\hat{\mu}_X - x_0) = 0 \tag{5.35}$$

For each value of x_0 the parameters μ_Y and σ_Y^2 are estimated by (5.32) and (5.33). Given an initial estimate of x_0, an improved estimate is obtained by means of the Newton-Raphson method:

$$x_{0,new} = x_{0,old} - \frac{g(x_{0,old})}{g'(x_{0,old})} \tag{5.36}$$

Since μ_Y and σ_Y^2 are also a function of x_0 it follows for the derivative $g'(x_{0,old})$:

$$g'(x_0) = \frac{dg}{dx_0} = (\hat{\mu}_Y - 1)\frac{1}{N}\sum_{i=1}^{N}(x_i - x_0)^{-1} + (\hat{\mu}_X - x_0)^{-1} \tag{5.37}$$

To speed up the computations, in HYMOS the expected value of $g'(x_{0,old})$ is calculated rather than computing $g'(x_{0,old})$ for each x_0:

$$E[g'(x_0)] = \frac{(\hat{\sigma}_Y^2 - 1)\exp(\hat{\sigma}_Y^2) + 1}{\hat{\mu}_X - x_0} \tag{5.38}$$

By substitution of (5.37) in (5.36) it follows for the improved estimate of x_0:

$$x_{0,new} = x_{0,old} - \frac{\left(\hat{\mu}_Y + \frac{\hat{\sigma}_Y^2}{2} - \ln(\hat{\mu}_X - x_0)\right).(\hat{\mu}_X - x_0)}{1 + (\hat{\sigma}_Y^2 - 1).\exp(\hat{\sigma}_Y^2)} \tag{5.39}$$

The iteration is continued till:

$$\left|x_{0,new} - x_{0,old}\right| < \varepsilon \text{ with : } \varepsilon = \frac{\mu_X - x_{min}}{1000} \tag{5.40}$$

The initial value of x_0 is taken as:

$$x_0 = x_{min} - 0.1(\mu_X - x_{min}) \tag{5.41}$$

Similar to this mixture of moment and MLM procedures, HYMOS provides mixed moment MLM estimators for the Pearson Type distributions. Reference is made to the HYMOS manual for the details.

5.7 Censoring of data

In some cases one wants to eliminate data from frequency analysis either at the upper end or at the lower end. Eliminating data from the frequency analysis at the upper end is called **right censoring** and eliminating data at the lower end is called **left censoring**. This is illustrated in Figure 5.3.

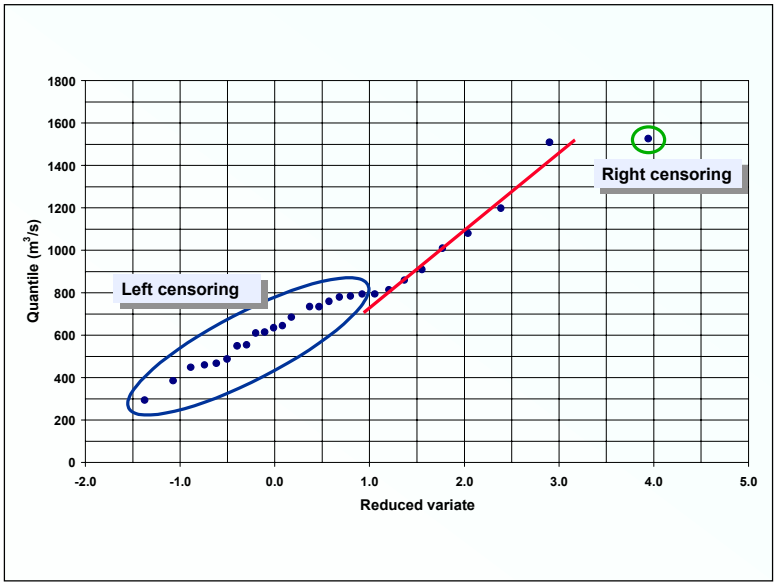


Figure 5.4:
Left and right censoring

With censoring, the **relative frequencies** attached to the remaining data is left **unchanged**. Hence, one performs frequency analysis on a reduced data set, but with frequency information from the **original** set. So the procedure is not the same as simply eliminating data from the data set and working with a reduced set, where the relative frequencies are determined based on the reduced series.

Right censoring may be required when there is evidence that the highest or a few of the highest values are unreliable (poorly measured extremes) or do have a return period which is believed to be much higher than one would expect based on the ordered data set. Left censoring may be required if the lower part of the ordered data set is not representative for the physics of the phenomena, which govern the higher part. Then, if one wants to extrapolate based on the higher values, the lower part can be censored, thereby leaving the relative frequencies of the higher ones intact. This procedure is often applied for analysis of river flow extremes, where the flow extremes refer to situations when the river stays inbank for the low peaks (lower part) and enters the flood plain with strong attenuation of the flood peaks (higher part). In such case the lower part will be steeper than the higher part (opposite to what is shown in Figure 5.3 !!).

In HYMOS censoring is possible for the Gumbel distribution. Great care is needed in applying censoring: there should be clear evidence that censoring is required.

5.8 Quantile uncertainty and confidence limits

Quantile uncertainty

The estimates for the distribution parameters involve estimation errors, and hence the same applies for the quantiles derived from it. The parameter uncertainties have to be translated to the uncertainty in the estimate of the quantile. The estimation error is used to draw the confidence limits about the estimated quantiles to indicate the likely range of the true value of the quantile. The procedure to derive the confidence limits will be illustrated for the quantile of a normally distributed random variable. From (4.23) the quantile x_p is given by:

$$x_p = \mu_X + \sigma_X.z_p \tag{5.42}$$

where: z_p = standard normal deviate corresponding to a non-exceedance probability p. The quantile is estimated by:

$$x_p = m_X + s_X.z_p \tag{5.43}$$

The parameters m and s are estimated by (2.3) and (2.6) respectively. The estimation variance of the quantile follows from:

$$var(x_{e,p}) = var\,(m_X + s_X.z_p) = var(m_X) + z_P^2\ var\,(s_X) + 2\ cov(m_X,s_X) \tag{5.44}$$

Since var $(m_X) = \sigma_X{}^2/N$, var $(s_X) \approx \sigma_X^2/(2N)$ (see Table 5.1) and for a normally distributed variable $cov(m_X, s_X) = 0$, the variance of x_p becomes approximately:

$$var(x_p) \approx var(m_x) + z_p^2\ var(s_x) = \frac{\sigma_x^2}{N}\left(1+\frac{1}{2}z_p^2\right) \tag{5.45}$$

Hence with σ_X replaced by s_X, the standard error of the quantile follows from:

$$s_{x_p} \approx s_x \sqrt{\frac{1}{N}\left(1+\frac{1}{2}z_p^2\right)} \tag{5.46}$$

The 100(1-α)% confidence limits for x_p then read:

$$x_{p,LCL} = x_p - z_{1-\alpha/2}s_X\sqrt{\frac{1}{N}\left(1+\frac{1}{2}z_p^2\right)} \quad x_{p,UCL} = x_p + z_{1-\alpha/2}s_X\sqrt{\frac{1}{N}\left(1+\frac{1}{2}z_p^2\right)} \tag{5.47}$$

The confidence limits express that the true quantile x_p falls within the interval $x_{p,LCL}$ and $x_{p,UCL}$ with a confidence of 100(1-α)%. The quantity 100(1-α)% is the **confidence level** and α is the **significance level**. From the limits shown in (5.47) it is observed that the confidence band about the quantile increases with z_p, i.e. the further away from the mean of the distribution the larger the uncertainty of the quantile becomes. Also the effect of the number of data is apparent from (5.47); a small number of data results in a large uncertainty for the quantile.

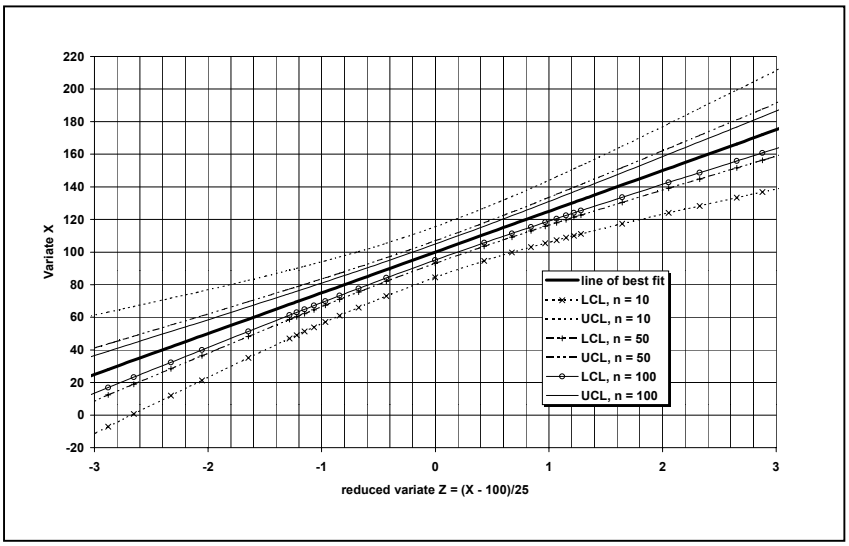


Figure 5.5: Fit by normal distribution (m_X= 100, s_X = 25) with 95% confidence limits for different length of data series

Uncertainty in the probability of the quantile

In the above we were looking at the standard error of the quantile for a given non-exceedance probability. One can also look at the uncertainty in the non-exceedance probability for a fixed value of x_p. From (5.42) it follows:

$$z_p = \frac{x_p - \mu_X}{\sigma_X} \tag{5.48}$$

Hence, the standard error of the reduced variate z_p becomes:

$$\sigma_{z_p} = \frac{\sigma_{x_p}}{\sigma_X} \text{ estimated by } s_{z_p} = \frac{s_{x_p}}{s_X} \tag{5.49}$$

The reduced variate z_p is approximately normally distributed with $N(z_p, \sigma_{zp})$. Hence, the confidence interval for p at a significance level α becomes $P_{LCL} = F_N(z_p - z_{1-\alpha/2}.\sigma_{zp})$ and $P_{UCL} = F_N(z_p - z_{1-\alpha/2}.\sigma_{zp})$, where F_N is the standard normal distribution function. The standard error σ_p of p for fixed x_p then becomes:

$$\sigma_p \approx f_N(z_p)\sigma_{z_p} \approx \left(\frac{1}{\sqrt{2\pi}}\exp(-\frac{z_p^2}{2})\right)s_{z_p} = \left(\frac{1}{\sqrt{2\pi}}\exp(-\frac{z_p^2}{2})\right)\frac{s_{x_p}}{s_X} \tag{5.50}$$

Example 5.2: Annual rainfall Vagharoli

Annual rainfall of station Vagharoli for the period 1978 –1997 is considered. After having tested the homogeneity of the series, the observed frequency distribution was fitted by the normal distribution, which should be applicable on basis of the conditions needed for a Gaussian distribution.

The result with HYMOS is presented in the table below. In the result first the basic statistics are presented. From the skewness and kurtosis being close to 0 and 3 respectively it is observed that the data are approximately normally distributed.

In the next part of the result a summary is presented of the ranked observations, including:

- In the 1st column the year number as from 1978 onward is presented for each ranked observation; e.g. the first row has year number 10 which means that this represents the value of year (1978 – 1) +10 i.e. 1987. The observation for the year 1978 is seen to be ranked as one but highest value.
- The 2nd column shows the ranked observations.
- The 3rd column gives the non-exceedance probability of the observation according to the observed frequency distribution, using the plotting position most appropriate for the normal distribution. According to Table 5.4, Blom's formula gives an unbiased plotting position for the normal distribution. For the first row (rank 1) the following value will then be obtained:

$$F_i = \frac{i - 3/8}{N + 1/4} \text{ becomes : } F_1 = \frac{1 - 3/8}{20 + 1/4} = \frac{0.625}{20.25} = 0.0309$$

- The 4th column gives the theoretical non-exceedance probability accepting the normal distribution with mean m = 877.3 and standard deviation 357.5. The reduced variate then reads:

$$z = \frac{x - m_X}{s_X} = \frac{x - 877.3}{357.5}$$

For the lowest ranked value (on the first row) it then follows:

$$z_1 = \frac{x_1 - 877}{357} = \frac{232 - 877.3}{357.5} = -1.805$$

From tables of the normal distribution one reads for z = 1.805 a non-exceedance probability of p = 0.9645. Hence the non-exceedance probability for z = -1.805 is in view of the symmetry of the normal distribution p_1 = 1 - 0.9645 = 0.0355. Using HYMOS it is not necessary to consult a statistical textbooks for the table of the normal distribution as it is included in the software under the option 'Statistical Tables'.

- The 5th column gives the return period, which is derived from the non-exceedance probability by:

$$T = \frac{1}{1 - F(x)} \text{ hence : } T_{x_1} = \frac{1}{1 - p_1} = \frac{1}{1 - 0.0355} = 1.037 \approx 1.04$$

The 6th column presents the standard error of the quantile x_p, derived from (5.46). Since we are discussing here observations, hence, there is no statistical uncertainty in it as such (apart from measurement errors). But the standard error mentioned here refers to the standard error one would have obtained for a quantile with the same value as the observation when derived from the normal distribution. It is a necessary step to derive the uncertainty in the non-exceedance probability presented in column 7. For the first row e.g. it then follows with (5.46):

$$s_{x_p} = s_X \sqrt{\frac{1}{N}\left(1 + \frac{z_p^2}{2}\right)} \text{ hence : } s_{x_{0.0355}} = 357 \sqrt{\frac{1}{20}\left(1 + \frac{(-1.805)^2}{2}\right)} = 357 \text{x} 0.363 = 129.6$$

$$s_p \approx \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_p^2}{2}\right)\right) \frac{s_{x_p}}{s_x} \text{ hence for the first row :}$$

$$s_{p_1} \approx \left(\frac{1}{\sqrt{2 \text{x} 3.14}} \exp\frac{(-1.805)^2}{2}\right) \frac{129.6}{357.5} = 0.0283$$

- The standard error of the non-exceedance probability follows from (5.50):

In the third part of the results the output from the goodness of fit tests are presented. This will be discussed in the next chapter.

In the last part of the results for distinct return periods and non-exceedance probabilities the quantiles are presented with their standard error and 100(1-α) = 95% confidence limits, which are also shown in the plot of the observed distribution fitted by the normal one in Figure 5.4. The values are obtained as follows:

- The 1st column presents the return period
- The 2nd column gives the non-exceedance probability associated with the return period in column 1
- In the 3rd column the quantile is given, which is derived from (5.43) for the reduced variate corresponding with the non-exceedance probability; this is derived from the inverse of the standard normal distribution. E.g. for T=100, p = 0.99, z_p =2.33 and the quantile follows from:
- $x_p = m_X + s_X z_p$ hence : $x_p = 877.3 + 357.5 \text{x} 2.33 = 1709.0$ mm
- In the 4th column the standard error of x_p is given which is obtained from (5.46)., e.g. for the T= 100 year event:

$$s_{x_p} \approx s_X \sqrt{\frac{1}{N}\left(1 + \frac{z_p^2}{2}\right)} = 357 \quad . \quad 5\sqrt{\frac{1}{20}\left(1 + \frac{(2.33)^2}{2}\right)} = 153 \quad . \quad 9 \text{ mm}$$

- In the 5th and 6th column the lower and upper confidence limits for the quantile are given, which are derived from (5.47) in case of 95% confidence limits. In case e.g. 90% limits are used (hence α = 0.10 instead of 0.05) then in equation (5.47) the value 1.96 (p=1-α/2 =0.975) has to be replaced by 1.64 (p=1-α/2=0.95), values which can be obtained from the tables of the normal distribution or from the Statistical Tables option in HYMOS. It follows for the 100 year event:

$$x_{p,LCL} = x_p - 1.96 s_{x_p} = 1709 - 1.96 \text{x} 153.9 = 1407.3 \text{ mm}$$
$$x_{p,UCL} = x_p + 1.96 s_{x_p} = 1709 + 1.96 \text{x} 153.9 = 2010.7 \text{ mm}$$

Results by HYMOS:

```
Annual rainfall Vagharoli

Period 1978 - 1997

 Fitting the normal distribution function

 Number of data     =    20
 Mean               =    877.283
 Standard deviation =    357.474
 Skewness           =      -.088
 Kurtosis           =      2.617

   Nr./year observation obs.freq. theor.freq.p theo.ret-per.  st.dev.xp st.dev.p
     10     232.000       .0309       .0355        1.04    129.6295       .0283
      5     267.000       .0802       .0439        1.05    125.3182       .0325
      9     505.000       .1296       .1488        1.17     99.2686       .0644
     18     525.000       .1790       .1622        1.19     97.4253       .0669
     15     606.000       .2284       .2240        1.29     90.7089       .0759
     14     628.000       .2778       .2428        1.32     89.1161       .0780
      7     649.580       .3272       .2621        1.36     87.6599       .0799
```

```
   Nr./year observation obs.freq. theor.freq.p theo.ret-per.  st.dev.xp st.dev.p
      4     722.000       .3765       .3320        1.50     83.6122       .0849
     11     849.400       .4259       .4689        1.88     80.0545       .0891
      3     892.000       .4753       .5164        2.07     79.9673       .0892
     16     924.000       .5247       .5520        2.23     80.2727       .0888
     20     950.000       .5741       .5806        2.38     80.7532       .0883
     19    1050.000       .6235       .6855        3.18     84.4622       .0839
      6    1110.000       .6728       .7425        3.88     87.9885       .0795
     12    1167.684       .7222       .7917        4.80     92.1776       .0740
      8    1173.000       .7716       .7959        4.90     92.5994       .0734
     13    1174.000       .8210       .7967        4.92     92.6794       .0733
      2    1197.000       .8704       .8144        5.39     94.5736       .0708
      1    1347.000       .9198       .9056       10.59    109.1187       .0513
     17    1577.000       .9691       .9748       39.76    136.5096       .0224

Results of Binomial goodness of fit test
 variate dn = max(|Fobs-Fest|)/sd=      .7833 at Fest= .7917
 prob. of exceedance P(DN>dn)    =      .4335
 number of observations          =    20

Results of Kolmogorov-Smirnov test
 variate dn = max(|Fobs-Fest|)   =      .0925
 prob. of exceedance P(DN>dn)    =      .9955

 Results of Chi-Square test
 variate = chi-square            =     1.2000
 prob. of exceedance of variate =      .2733
 number of classes               =     4
 number of observations          =    20
 degrees of freedom              =     1

  Values for distinct return periods
  Return per.  prob(xi<x) p     value x  st. dev. x   confidence intervals
                                                        lower       upper
            2      .50000     877.283      79.934     720.582    1033.985
            5      .80000    1178.082      93.013     995.740    1360.424
           10      .90000    1335.468     107.878    1123.984    1546.952
           25      .96000    1503.247     127.221    1253.844    1752.650
           50      .98000    1611.602     140.961    1335.263    1887.941
          100      .99000    1709.048     153.900    1407.343    2010.753
          250      .99600    1825.469     169.899    1492.399    2158.539
          500      .99800    1906.275     181.273    1550.908    2261.643
         1000      .99900    1982.065     192.101    1605.471    2358.660
         1250      .99920    2005.533     195.482    1622.312    2388.754
         2500      .99960    2075.895     205.685    1672.672    2479.118
         5000      .99980    2142.841     215.477    1720.421    2565.260
        10000      .99990    2206.758     224.893    1765.878    2647.638
```

The fit of the normal distribution to the observed frequency distribution is shown in Figure 5.6. The Blom plotting position has been used to assign non-exceedance frequencies to the ranked observations.

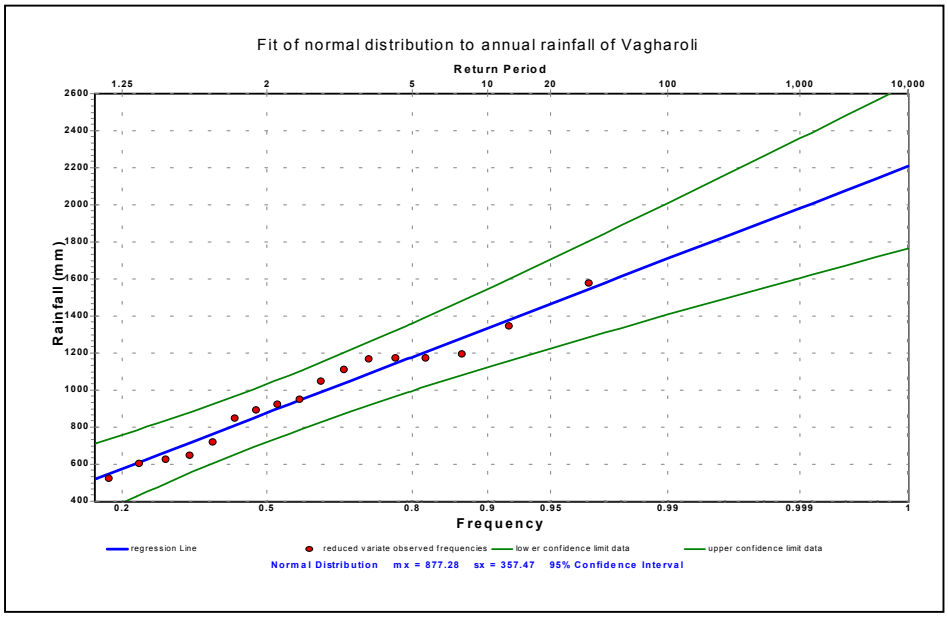


Figure 5.6: Fit of normal distribution to annual rainfall at Vaharoli, period 1978-1997

6 Hypothesis Testing

6.1 General

To apply the theoretical distribution functions dealt with in Chapter 5 the following steps are required:

1. Investigate the homogeneity of the data series, subjected to frequency analysis
2. Estimate the parameters of the postulated theoretical frequency distribution
3. Test the goodness of fit of the theoretical to the observed frequency distribution

In this chapter attention will be given to series homogeneity tests and goodness of fit tests. First an overview is given of the principles of hypothesis testing.

6.2 Principles

A statistical hypothesis is an assumption about the distribution of a statistical parameter. The assumption is stated in the **null-hypothesis H_0** and is tested against one or more alternatives formulated in the **alternative hypothesis H_1**. For easy reference the parameter under investigation is usually presented as a ***standardised*** variate, called ***test statistic***. Under the null-hypothesis the test statistic has some standardised sampling distribution, e.g. a standard normal, a Student t-distribution, etc. as discussed in Chapter 4. For the null-hypothesis to be true the value of the test statistic should be within the acceptance region of the sampling distribution of the parameter under the null-hypothesis. If the test statistic does not lie in the acceptance region, the null-hypothesis is rejected and the alternative is assumed to be true. Some risk, however, is involved that we make the wrong decision about the test:

- **Type I error**, i.e. rejecting H_0 when it is true, and
- **Type II error**, i.e. accepting H_0 when it is false.

The probability of making a Type I error is equal to the significance level of the test α. When a test is performed at a 0.05 or 5% level of significance it means that there is about 5% chance that the null-hypothesis will be rejected when it should have been accepted. This probability represents the critical

region at the extreme end(s) of the sampling distribution under H0. Note, however, the smaller the significance level is taken, the larger becomes the risk of making Type II error and the less is the discriminative power of the test.

Choosing the significance level α

Consider the following hypothesis. Let Φ denote the parameter under investigation and let:

H_0: $\Phi = \Phi_0$, and

H_1: $\Phi = \Phi_1$, with $\Phi_1 > \Phi_0$

The estimate of Φ is ϕ. The hypothesis is tested by means of a one-tailed test. The decision rule of acceptance is stated as follows:

Accept H_0 if: $\phi \leq c$

Reject H_0 and accept H_1 if: $\phi > c$

where c is a constant, for the time being chosen arbitrarily between Φ_0 and Φ_1. To specify c the relative positions of the pdf's of ϕ are considered $f_0(\phi|H_0)$ and $f_1((\phi|H_1)$ are, see Figure 6.1.

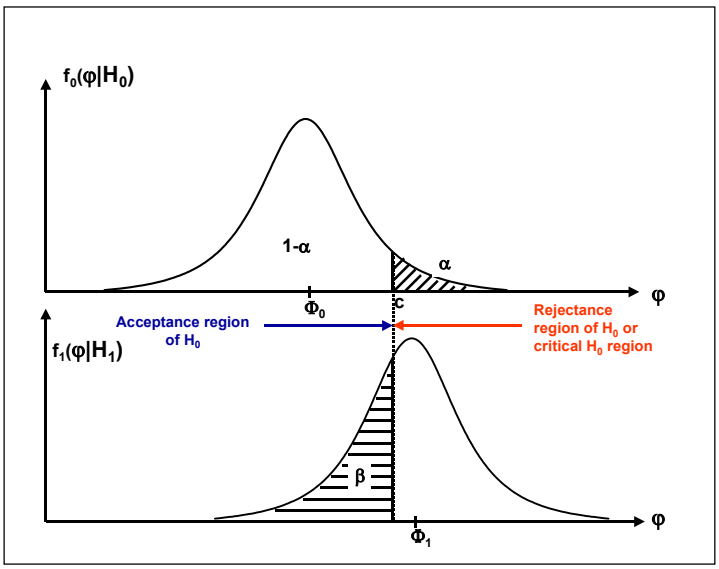


Figure 6.1:
Definition sketch for hypothesis testing

The region $\phi \leq c$ is called the **acceptance region** for H_0 and, reversely, the region $\phi > c$ is called the **rejectance or critical H_0 region**. If H_0 is true and $\phi \leq c$, then the right decision is made. However, if H_0 is true and $\phi > c$ then the wrong decision is made, i.e. an error of Type I. Formally:

$$P(\text{error of Type I}) = P(\phi > c \mid H_0 \text{ is true}) = \int_c^{\infty} f_0(\phi \mid H_0) d\phi = \alpha \qquad (6.1)$$

On the other hand, if H_1 is true and $\phi \leq c$, or equivalently, accepting H_0 when it is false, then a Type II error is made. It has a probability of occurrence defined by:

$$P(\text{error of Type II}) = P(\phi \leq c \mid H_1 \text{ is true}) = \int_{-\infty}^{c} f_1(\phi \mid H_1) d\phi = \beta \qquad (6.2)$$

In production processes, the risk associated with Type I errors is called the **producer's risk** and the Type II risk the **consumer's risk**. Now basically c has to be chosen such that the total loss associated with making errors of Type I and of Type II are minimised. Hence, if L_α and L_β are the losses associated with errors of Type I and Type II respectively, and L is the total loss, with:

$$L = \alpha(c)\, L_\alpha + \beta(c) L_\beta \qquad (6.3)$$

Then c follows from the minimum of L. In practice, however, the loss functions L_α and L_β are usually unknown and the **significance level** α is chosen **arbitrarily** small like 0.1 or 0.05. From Figure 6.1 it is observed that a low value of α implies a very high value of β. The test then is seen to have a very low **discriminative power**; the likelihood of accepting H_0, when it is false, is becoming very large. By definition, the **power of a test** = 1 - β, i.e. the complement of β and it expresses the probability of rejecting H_0 when it is false, or the probability of avoiding Type II errors. In this case:

$$1-\beta = \int_c^\infty f_1(\phi \mid H_1)\, d\phi \qquad (6.4)$$

If the test is two-sided with acceptance region for H_0: $d \le \phi \le c$, the power of the test is given by:

$$1-\beta = \int_{-\infty}^{d} f_1(\phi \mid H_1)\, d\phi + \int_c^\infty f_1(\phi \mid H_1)\, d\phi \qquad (6.5)$$

If the alternative is not a single number, but can take on different values, then β becomes a function of ϕ. This function $\beta(\phi)$ is called the **operating characteristic** (OC) of the test and its curve the OC-curve. Similarly, $\eta(\phi) = 1-\beta(\phi)$ is called the **power function** of the test.

In summary: Type I and Type II errors in testing a hypothesis $\Phi = \Phi_0$ against an alternative $\Phi = \Phi_1$ read:

		Test hypothesis H_0: $\Phi = \Phi_0$	
		Accepted	**Rejected**
True state	$\Phi = \Phi_0$	**Correct decision** P = 1 - α	**Type I error** P = α
	$\Phi = \Phi_1$	**Type II error** P = β	**Correct decision** P = 1 - β

Table 6.1: Overview of hypothesis test results

Test procedure

Generally, the following procedure is used in making statistical tests (Haan, 1977):

1. Formulate the hypothesis to be tested
2. Formulate an alternative hypothesis
3. Determine a test statistic
4. Determine the distribution of the test statistic
5. Collect data needed to calculate the test statistic

6. Determine if the calculated value of the test statistic falls in the rejection region of the distribution of the test statistic.

Depending on the type of alternative hypothesis H_1 one- or two-tailed tests are considered. This is explained by the following example. To test the significance of serial correlation the value of the serial correlation coefficient r is considered. The null-hypothesis reads H_0: $\rho = 0$ against one of the following alternatives:

1. $H_1 : \rho > 0$, i.e. a right-sided test
2. $H_1 : \rho < 0$, i.e. a left-sided test
4. $H_1 : \rho \neq 0$, i.e. a two-sided test

The serial correlation coefficient is estimated from:

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(x_i - m_X)(x_{i+1} - m_X)}{\frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)^2} \qquad (6.6)$$

The test statistic to measure the significance of r is:

$$T_r = r\sqrt{\frac{N-3}{1-r^2}} \qquad (6.7)$$

Under the null-hypothesis the test statistic T_r has a Student t-distribution with ν = N-3 degrees of freedom. Let the tests be performed at a significance level α, then H_0 will not be rejected in:

1. a right-sided test, if: $T_r \leq t_{\nu,1-\alpha}$
2. a left-sided test, if: $T_r \geq t_{\nu,\alpha}$
3. a two-sided test, if: $t_{\nu,\alpha/2} \leq T_r \leq t_{\nu,1-\alpha/2}$

Since the Student distribution is symmetrical the last expression may be replaced by:

$$|T_r| \leq t_{\nu,1-\alpha/2} \qquad (6.8)$$

The latter condition is investigated when testing randomness of a series. The various options are displayed in Figure 6.2.

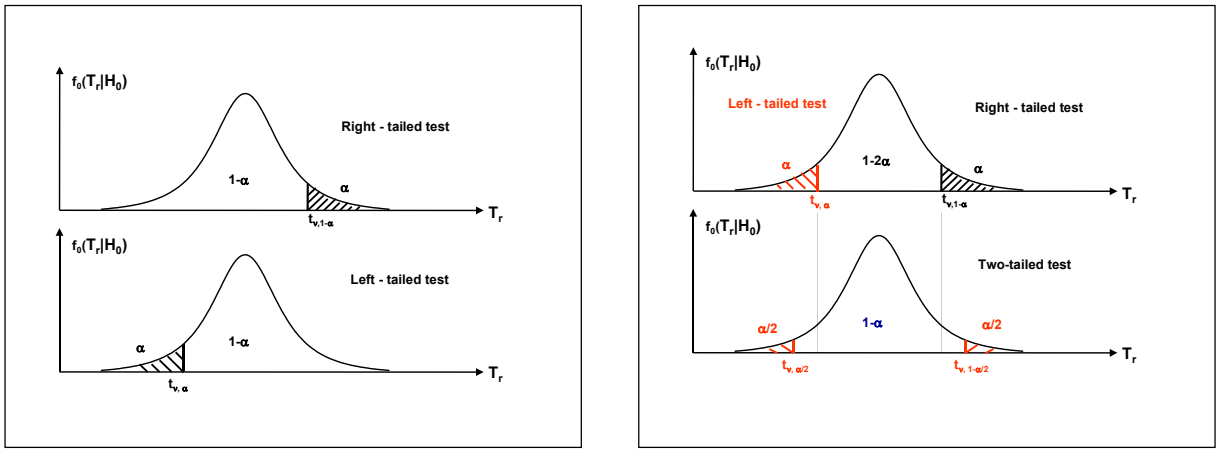


Figure 6.2: Right-tailed, left-tailed and two-tailed tests

From Figure 6.2 it is observed that for the same significance level the critical values differ in a one-tailed or a two-tailed test.

6.3 Investigating homogeneity

Prior to fitting of theoretical distributions to observed ones, the sample series should fulfil the following conditions:

stationarity: i.e. the properties or characteristics of the series do not vary with time;
homogeneity: i.e. all elements of a series belong to the same population;
randomness: i.e. series elements are independent.

The first two conditions are transparent and obvious. Violating the last one, while the series were tested homogeneous, means that the effective number of data is reduced and hence the power of the tests and the quality of the estimates. Lack of randomness may, however, have several causes; in case of a trend there will also be serial correlation.

HYMOS includes numerous statistical tests to investigate the stationarity, homogeneity or randomness. A number of them are **parametric** tests, which assume that the sample is taken from a population with an approximately normal distribution. **Non-parametric** or distribution-free do not set conditions to the distribution of the sample. Generally, this freedom affects the discriminative power of the test negatively.

Tests included in HYMOS suitable for series inspection prior to frequency analysis comprise a.o.:

On randomness:

1. **Median run test**: a test for randomness by calculating the number of runs above and below the median;
2. **Turning point test**: a test for randomness by calculating the number of turning points;
3. **Difference sign test**: a test for randomness by calculating the number of positive and negative differences;

On correlation and trend:

1. **Spearman rank correlation test**: the Spearman rank correlation coefficient is computed to test serial correlation or significance of a trend;
2. **Spearman rank trend test**
3. **Arithmetic serial correlation coefficient**: a test for serial correlation;
4. **Linear trend test**: a test on significance of linear trend by statistical inference on slope of trend line;

On homogeneity:

1. **Wilcoxon-Mann-Whitney U-test**: a test to investigate whether two series are from the same population;
2. **Student t-test**: a test on difference in the mean between two series;
3. **Wilcoxon W-test**: a test on difference in the mean between two series;
4. **Rescaled adjusted range test**: a test for series homogeneity by the rescaled adjusted range.

From each group an example will be given.

Difference sign test

The difference-sign test counts the number of positive differences n_p and of negative differences n_n between successive values of series x_i,(i = 1,N): $x_{(i+1)} - x_{(i)}$. Let the maximum of the two be given by N_{ds}:

$$N_{ds} = \text{Max}(n_p, n_n) \tag{6.9}$$

For an independent stationary series of length N_{eff} (N_{eff} = N - zero differences) the number of negative or positive differences is asymptotically *normally* distributed with $N(\mu_{ds},\sigma_{ds})$:

$$\left.\begin{aligned} \mu_{ds} &= \frac{1}{2}(N_{eff} - 1) \\ \sigma_{ds}^2 &= \frac{1}{12}(N_{eff} + 1) \end{aligned}\right\} \tag{6.10}$$

The following hypothesis is considered:

H_0: series x_i is random, and

H_1: series is not random, with no direction for the deviation of randomness; hence, a two-tailed test is performed

The following standardised test statistic is considered:

$$|n_{ds}| = \frac{|N_{ds} - \mu_{ds}|}{\sigma_{ds}} \tag{6.11}$$

The null-hypothesis will not be rejected at a α level of significance if:

$$|n_{ds}| < z_{1-\alpha/2} \tag{6.12}$$

where $z1-\alpha/2$ is the standard normal deviate with $F(z < z1-\alpha/2) = 1-\alpha/2$. A requirement is that the sample size has to be $N \geq 10$.

Linear trend test

The slope of the trend line of series x_i, (i=1,N) with time or sequence is investigated. The linear trend equation reads:

$$x_i = b_1 + b_2 i + \varepsilon_i \quad \text{with}: \varepsilon_i \approx N(0, \sigma_\varepsilon) \tag{6.13}$$

The trend parameters are given by:

$$\left.\begin{aligned} b_2 &= \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(i - m_i)}{\sigma_i^2} \quad \text{with}: m_i = \frac{N+1}{2} \quad \text{and}: \sigma_i^2 = \frac{1}{12}N(N+1) \\ b_1 &= m_X - b_2 m_i \end{aligned}\right. \tag{6.14}$$

where: m_X = mean of x_i, i = 1, N

The following hypothesis is made:

H_0: no trend, i.e. the slope of the trend line should be zero: $\mu b2 = 0$, and

H_1: significant trend, i.e. $\mu b2 \neq 0$, hence a two-tailed test is performed

The absolute value of the following standardised test statistic is computed:

$$|T_t| = \frac{|b_2|}{s_{b_2}} \quad \text{with}: s_{b_2}^2 = \frac{1}{N-1}\frac{\sigma_n^2}{\sigma_i^2} \quad \text{and}: \sigma_n^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(x_i - (b_1 + b_2 i)\right)^2 \tag{6.15}$$

Under the null-hypothesis of no trend, the test statistic Tt has a Student t-distribution with ν=N-2 degrees of freedom for $N \geq 10$. The null-hypothesis of zero trend will not be rejected at a significance level α, if:

$$|T_t| < t_{\nu,1-\alpha/2} \tag{6.16}$$

where $t\nu,1\text{-}\alpha/2$ is the Student-t variate defined by: $F(t<t\nu,1\text{-}\alpha/2) = 1\text{-}\alpha/2$

Student t-test and Fisher F-test

A good indicator for stationarity and homogeneity of a series is the behaviour of the mean value, for which the t-test is appropriate. With the Student t-test differences in mean values of two series yi,(i=1,m) and zi,(i=1,n) are investigated. In this case of frequency analysis the test is used as a split-sample test as it will be applied to the data from the same data set xi, I = 1, N. The series X is split in two parts Y and Z. The series Y and Z are chosen such that the first m represent Y and the last N-m are represented by Z. Let m_Y and m_Z denote the sample values of population means of Y and Z: μ_Y and μ_Z.

The following hypothesis is now tested:

H_0: $\mu_Y = \mu_Z$, and

H_1: $\mu_Y \neq \mu_Z$, hence a two-tailed test is performed

The absolute value of the following standardised test statistic is therefore investigated:

$$|T_S| = \frac{|m_Y - m_Z|}{s_{YZ}} \tag{6.17}$$

Under the null-hypothesis of equal population means the test statistic T_S has a *Student t*-distribution with $\nu = m+n-2$ degrees of freedom for $N = m + n > 10$. The null-hypothesis $\mu_Y = \mu_Z$ will not be rejected at a significance level α, if:

$$|T_S| < t_{\nu,1-\alpha/2} \tag{6.18}$$

where $t_{\nu,1-\alpha/2}$ is the Student-t variate defined by: $F(t<t_{\nu,1-\alpha/2}) = 1\text{-}\alpha/2$

The way the standard deviation s_{YZ} is computed depends on whether the series Y and Z have the same population variance. For this a Fisher F-test is performed on the ratio of the variances.

The following hypothesis is made:

H_0: $\sigma_Y^2 = \sigma_Z^2$, and

H_1: $\sigma_Y^2 \neq \sigma_Z^2$, by putting the largest one on top a one-tailed test is performed.

Following test statistic is considered:

$$F_S = \frac{s_Y^2}{s_Z^2} \text{ if : } s_Y^2 > s_Z^2 \text{ else : } F_S = \frac{s_Z^2}{s_Y^2} \tag{6.19}$$

Under the null-hypothesis the test statistic FS has a Fisher F-distribution with (m-1, n-1) degrees of freedom if sY2 > sZ2, otherwise the number of degrees of freedom is (n-1, m-1). The null-hypothesis σY2 = σZ2 will not be rejected at a significance level α, if:

$$F_S < f_{m-1,n-1,1-\alpha} \tag{6.20}$$

where fm-1,n-1,1-α is the Fisher-F variate defined by: F(f < fm-1,n-1,1-α) = 1-α.

For fitting distributions to the sample series X it is essential that the hypothesis on the mean and the variance are both not rejected. If one of the hypotheses is rejected, the series should not be applied.

The outcome of the variance test determines in which way the standard deviation s_{YZ} is being estimated (Hald, 1952). The standard deviation s_{YZ} is computed from:

1. in case of equal variances:

$$s_{YZ} = \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\frac{(m-1)s_Y^2 + (n-1)s_Z^2}{m+n-2}} \tag{6.21}$$

2. in case of unequal variances:

$$s_{YZ} = \sqrt{\frac{s_Y^2}{m} + \frac{s_Z^2}{n}} \text{ and : } \nu = \left(\frac{\psi^2}{m-1} + \frac{(1-\psi)^2}{n-1}\right)^{-1} \text{ and : } \psi = \frac{\frac{s_Y^2}{m}}{\frac{s_Y^2}{m} + \frac{s_Z^2}{n}} \tag{6.22}$$

Practically, it implies that in the latter case the number of degrees of freedom ν becomes less than in the equal variance case, so the discriminative power of the test diminishes somewhat. With respect to the sample size it is noted that the following conditions apply: N≥10, m≥5 and n≥5.

Example 5.2: continued: Annual rainfall Vagharoli.

The above-discussed tests have been applied to the annual rainfall series of Vagharoli available for the period 1978-1997. In the split-sample test on the mean and the variance the series have been split in equal parts. It is noted though, that in practice one should first inspect the time series plot of the series to determine where the boundary between the two parts is to be put. The time series of the annual rainfall is shown in Figure 6.3.

Results of tests

Difference Sign Test

Number of difference signs (=N_{ds})= 11
Mean of N_{ds} = 9.500
Standard deviation of N_{ds} = 1.323
Test statistic [n_{ds}](abs.value) = 1.134
Prob(n_{ds}.le. $n_{ds,obs}$ = .872
Hypothesis: H_0: Series is random
H_1: Series is not random
A two-tailed test is performed
Level of significance is α 5.00 percent
Critical value for test statistic $z_{1-\alpha/2}$ = 1.960
Result: H_0 not rejected

Test for Significance of Linear Trend

Intercept parameter (=b_1) = 871.612
Slope parameter (=b_2) = .5401E+00
St.dev. of b2 (=s_{b2}) = .1424E+02
St.dev. of residual (=s_e) = .3673E+03
Test statistic [T_t] (abs.value) = .038
Degrees of freedom □ = 18
Prob(T_t.le $T_{t,obs}$ = .515
Hypothesis: H_0: Series is random
H_1: Series is not random
A two-tailed test is performed
Level of significance α is 5.00 percent
Critical value for test statistic $t_{□,1-\alpha/2}$ = 2.101

Result: H_0 not rejected

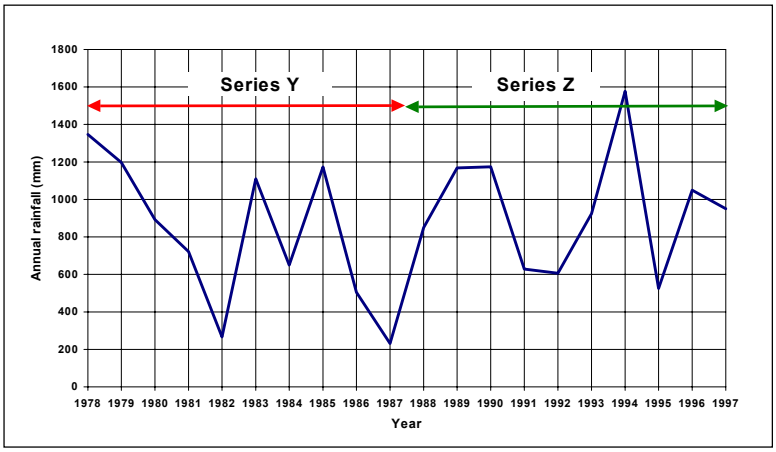


Figure 6.3:
Annual rainfall at Vagharoli, period 1978-1997, with division for split sample test

Student t-Test with Welch modification

Number of data in first set		=	10
Number of data in second set		=	10
Test statistic [T_S] (abs.value)		=	.842
Degrees of freedom		=	18
Prob(t.<.[T_S])		=	.795
Mean of first set	(m_Y)	=	809.458
St.dev. of first set	(s_Y)	=	397.501
Mean of second set	(m_Z)	=	945.108
St.dev. of second set	(s_Z)	=	318.659
Var. test stat. $F_S = s_Y^2/s_Z^2$		=	1.556
Prob($F \leq F_S$)		=	.740

Hypothesis: H_0: Series is homogeneous
H_1: Series is not homogeneous
A two-tailed test is performed
Level of significance is α = 5.00 percent
Critical value for test statistic mean $t_{\square,1-\alpha/2}$ = 2.101
Critical value for test statistic variance $F_{m-1,n-1,1-\alpha}$ = 3.18

Result: H_0 not rejected

6.4 Goodness of fit tests

To investigate the goodness of fit of theoretical frequency distribution to the observed one three tests are discussed, which are standard output in the results of frequency analysis when using HYMOS, viz:

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

Chi-square goodness of fit test

The hypothesis is that F(x) is the distribution function of a population from which the sample x_i, i = 1,…,N is taken. The hypothesis is tested by comparing the actual to the theoretical number of occurrences within given class intervals. The following procedure is followed in the test

First, the data set is divided in k class intervals such that each class contains at least 5 values. The class limits are selected such that all classes have equal probability $p_j = 1/k = F(z_j)-F(z_{j-1})$. For example if there are 5 classes, the upper class limits will be derived from the variate corresponding with the non-exceedance frequencies p = 0.20, 0.40, 0.60, 0.80 and 1.0. The interval j contains all x_i with: $U_c(j-1) < x_i \leq U_c(j)$, where $U_c(j)$ is the upper class limit of class j, see Figure 6.4. The number of sample values falling in class j is denoted by b_j.

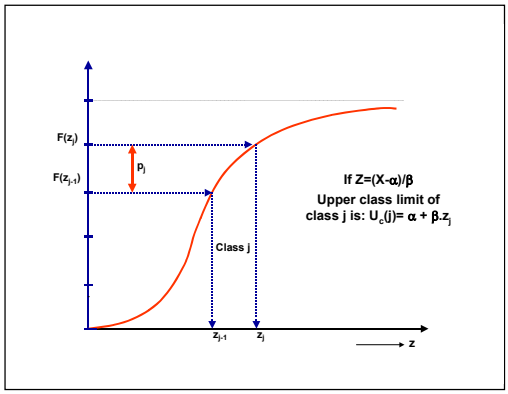


Figure 6.4:
Definition sketch for class selection in Chi-square goodness of fit test

Next, the number of values expected in class j according to the theoretical distribution is determined, which number is denoted by e_j. The theoretical number of values expected in any class is N/k, since all classes have equal probability.

The following test statistic is considered:

$$\chi_c^2 = \sum_{j=1}^{k} \frac{(b_j - e_j)^2}{e_j} \qquad (6.23)$$

This test statistic has under the assumption of the null-hypothesis a chi-squared distribution with ν = k-1-m degrees of freedom, where k = number of classes and m = number of parameters in the theoretical distribution. Because of the choice of equal probabilities (6.23) can be simplified as follows:

$$\chi_c^2 = \sum_{j=1}^{k} \frac{(b_j - N/k)^2}{N/k} = \frac{k}{N} \sum_{j=1}^{k} b_j^2 - N \qquad (6.24)$$

The null-hypothesis will not be rejected at a significance level α if:

$$\chi_c^2 < \chi_{\nu,1-\alpha}^2 \quad \text{with : } \nu = k - 1 - m \qquad (6.25)$$

The following number of class intervals k given N are suggested, see Table 6.2

N	k	N	k	N	k
20-29	5	100-199	13	800-999	27
30-39	7	200-399	16	1000-1499	30
40-49	9	400-599	20	1500-1999	35
50-99	10	600-799	24	≥ 2000	39

Table 6.2: Recommended number of class intervals for Ch-square goodness of fit test

Example 5.2: continued:

Annual rainfall Vagharoli. It is investigated if the null-hypothesis that the sample series of annual rainfall fits to the normal distribution. It is observed from the results in Chapter 5 that HYMOS has selected 4 class intervals, hence k = 4 and the upper class levels are obtained at non-exceedance probabilities 0.25, 0.50, 0.75 and 1.00. The reduced variates for these probabilities can be obtained from tables of the normal distribution or with the Statistical Tables option in HYMOS. The reduced variates are respectively –0.674, 0.000, 0.674 and ∞, hence with mean = 877 and standard deviation = 357 the class limits become 877-0.674x357, 877, 877+0.674x357 and ∞, i.e. 636, 877, 1118 and ∞.

The number of occurrences in each class is subsequently easily obtained from the ranked rainfall values presented in Chapter 5, Example 5.2. The results are presented in Table 6.3

Non-exc. probability of upper class limits	Reduced variate of upper class limits	Class intervals expressed in mm	Number of occurrences b_j	b_j^2
0.25	-0.67	0- 636	6	36
0.50	0.00	637-877	3	9
0.75	0.67	878-1118	5	25
1.00	∞	1119-∞	6	36
			sum	106

Table 6.3: Number of occurrences in classes

From Table 6.3 it follows for the test statistic (6.24):

$$\chi_c^2 = \frac{4}{20} x106 - 20 = 1.2$$

The critical value at a 5% significance level, according to the chi-squared distribution for ν = 4-1-2 = 1 degrees of freedom, is 3.84. Hence the computed value is less than the critical value. Consequently, the null-hypothesis is not rejected at the assumed significance level, as can be observed from the HYMOS results as well.

Kolmogorov-Smirnov test

In the Kolmogorov-Smirnov test the differences between the theoretical and observed frequency distribution is analysed and when the difference at a particular non-exceedance frequency exceeds a critical limit then the null-hypothesis that the sample is from the assumed theoretical distribution is rejected.

Let the observed frequency distribution be denoted by $S_N(x)$ and is defined by:

$$S_N(x) = \begin{cases} 0 & \text{for : } x < x_1 \\ \frac{i}{N} & \text{for : } x_i \le x < x_{i+1} \\ 1 & \text{for : } x_N \le x \end{cases} \qquad (6.26)$$

where x1 and xN are respectively the smallest and largest elements of the sample. Now, at each observed value xi, I = 1,N the difference between F(x), i.e. the theoretical distribution, and SN(x) is determined. The difference has two values as SN(x) changes at each step. If these two differences are denoted by ∂i+ and ∂i-, (see Figure 6.5) then the test statistic DN is developed as follows:

$$\begin{aligned} &\partial_i^+ = \frac{i}{N} - F(x) \quad \text{and:} \quad \partial_i^- = F(x) - \frac{(i-1)}{N} \\ &d_i = \text{Max}(\partial_i^+, \partial_i^-) \\ &D_N = \text{Max}(d_1, d_2, ..., d_N) \end{aligned} \qquad (6.27)$$

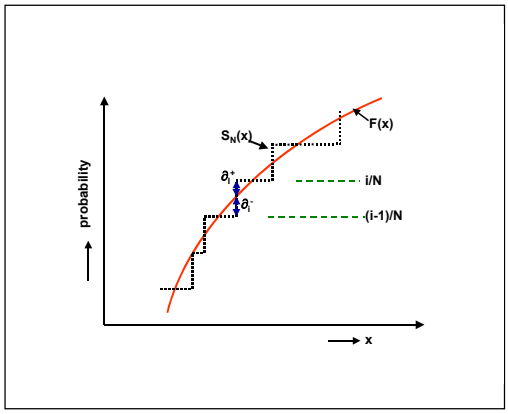


Figure 6.5:
Definition sketch Kolmogorov-Smirnov test

(adapted from NERC, 1975)

The null hypothesis is not rejected at a significance level α if D_N does not exceed the critical values Δ read from Kolmogorov-Smirnov's table:

$$D_N < \Delta_\alpha \qquad (6.28)$$

Critical values at the 10, 5 and 1% significance level for $N \geq 35$ are respectively $1.22/\sqrt{N}$, $1.36/\sqrt{N}$, and $1.63/\sqrt{N}$.

Example 5.2: continued: annual rainfall Vagharoli.

The results of the application of the Kolmogorov-Smirnov test to the annual rainfall series of Vagharoli are presented in the table below.

It is observed from Table 6.4 that the test statistic $D_N = 0.0925$. According to the Statistical Tables of the Kolmogorov-Smirnov test the critical value at a 5% confidence level for N = 20 is: $\Delta_5 = 0.29$. Hence, the observed D_N is less than the critical value, so the null hypothesis that the observations are drawn from a normal distribution with mean 877 mm and standard deviation 357 mm is not rejected.

Year nr	Rainfall	Blom	i/N	(i-1)/N	F(x)	d+	d-	max(d+,d-)
10	232	0.031	0.05	0.00	0.0355	0.0145	0.0355	0.0355
5	267	0.080	0.10	0.05	0.0439	0.0561	-0.0061	0.0561
9	505	0.130	0.15	0.10	0.1488	0.0012	0.0488	0.0488
18	525	0.179	0.20	0.15	0.1622	0.0378	0.0122	0.0378
15	606	0.228	0.25	0.20	0.2240	0.0260	0.0240	0.0260
14	628	0.278	0.30	0.25	0.2428	0.0572	0.0072	0.0572
7	650	0.327	0.35	0.30	0.2621	0.0879	0.0379	0.0879
4	722	0.377	0.40	0.35	0.3320	0.0680	-0.0180	0.0680
11	849	0.426	0.45	0.40	0.4689	-0.0189	0.0689	0.0689
3	892	0.475	0.50	0.45	0.5164	-0.0164	0.0664	0.0664
16	924	0.525	0.55	0.50	0.5520	-0.0020	0.0520	0.0520
20	950	0.574	0.60	0.55	0.5806	0.0194	0.0306	0.0306
19	1050	0.624	0.65	0.60	0.6855	-0.0355	0.0855	0.0855
6	1110	0.673	0.70	0.65	0.7425	-0.0425	0.0925	0.0925
12	1168	0.722	0.75	0.70	0.7917	-0.0417	0.0917	0.0917
8	1173	0.772	0.80	0.75	0.7959	0.0041	0.0459	0.0459
13	1174	0.821	0.85	0.80	0.7967	0.0533	-0.0033	0.0533
2	1197	0.870	0.90	0.85	0.8144	0.0856	-0.0356	0.0856
1	1347	0.920	0.95	0.90	0.9056	0.0444	0.0056	0.0444

Year nr	Rainfall	Blom	i/N	(i-1)/N	F(x)	d+	d-	max(d+,d-)
17	1577	0.969	1.00	0.95	0.9748	0.0252	0.0248	0.0252
							Max	0.0925

Table 6.4: Kolmogorov-Smirnov test on annual rainfall

Binomial goodness of fit test

A third goodness of fit test is based on the fact that, when the observed and the theoretical distribution functions, respectively F1(x) and F2(x), are from the same distribution, then the standardised variate DB, defined by:

$$D_B = \frac{|F_1(x) - F_2(x)|}{s_B} \quad \text{with:} \quad s_B = \sqrt{\frac{F_2(x)(1-F_2(x))}{N}} \tag{6.29}$$

is approximately normally distributed with N(0,1). Hence, the null-hypothesis is not rejected at a α % significance level if:

$$D_B < z_{1-\alpha/2} \tag{6.30}$$

The test is used in the range where:

$$N\ F2(x)\{1-\ F2(x)\} > 1 \tag{6.31}$$

This criterion generally means that the tails of the frequency distribution are not subjected to the test.

Example 5.2 continued: annual rainfall Vagharoli. The results of the test are displayed in Table 6.5

Nr./year	observation	$F_1(x)$	$F_2(x)$	s_B	D_B	criterion
10	232	0.0343	0.0355	0.0414	0.0290	0.6848
5	267	0.0833	0.0439	0.0458	0.8601	0.8395
9	505	0.1324	0.1488	0.0796	0.2061	2.5332
18	525	0.1814	0.1622	0.0824	0.2329	2.7178
15	606	0.2304	0.2240	0.0932	0.0686	3.4765
14	628	0.2794	0.2428	0.0959	0.3817	3.6770
7	650	0.3284	0.2621	0.0983	0.6742	3.8681
4	722	0.3775	0.3320	0.1053	0.4321	4.4355
11	849	0.4265	0.4689	0.1116	0.3800	4.9807
3	892	0.4755	0.5164	0.1117	0.3660	4.9946
16	924	0.5245	0.5520	0.1112	0.2473	4.9459
20	950	0.5735	0.5806	0.1103	0.0643	4.8701
19	1050	0.6225	0.6855	0.1038	0.6068	4.3118
6	1110	0.6716	0.7425	0.0978	0.7251	3.8239
12	1168	0.7206	0.7917	0.0908	0.7830	3.2982
8	1173	0.7696	0.7959	0.0901	0.2918	3.2489
13	1174	0.8186	0.7967	0.0900	0.2434	3.2394
2	1197	0.8676	0.8144	0.0869	0.6120	3.0231
1	1347	0.9167	0.9056	0.0654	0.1698	1.7098
17	1577	0.9657	0.9748	0.0350	0.2597	0.4913
				Max	**0.7830**	

Table 6.5: Results of binomial goodness of fit test, annual rainfall Vagharoli

In HYMOS, the observed non-exceedance frequency distribution $F_1(x)$ is obtained from Chegodayev plotting position, see Table 5.4. From Table 6.5 it is observed that the maximum value for $D_B = 0.8601$ at a non-exceedance frequency = 0.0439. However, criterion (6.31), which is presented in the last column, is not fulfilled for that non-exceedance frequency (criterion is less than 1). For the range of data for which this criterion is fulfilled, the maximum value for $D_B = 0.7830$ at $F_2(x) = 0.7917$. The critical value for D_B at a 5% confidence level is 1.96, hence, according to (6.30), the null-hypothesis that both $F_1(x)$ and $F_2(x)$ are from the same distribution is not rejected.

ANNEX 4.1 Standard normal distribution

The standard normal distribution function reads:

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-s^2)ds = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right) \qquad \text{(A4.1.1)}$$

The following approximation is used in HYMOS to solve $F_Z(z)$ for a given value of the standard normal variate z:

$$F = \exp\left(-\frac{z^2}{2}\right)((((a_1T + a_2)T + a_3)T + a_4)T + a_5)T \text{ with } T = \frac{1}{1 + b|z|}$$

$$\text{For } z \le 0 : F_Z(z) = F \qquad \text{(A4.1.2)}$$

$$\text{For } z > 0 : F_Z(z) = 1 - F$$

The coefficients in (A4.2) read:

a_1 = 0.530702715
a_2 = -0.726576014
a_3 = 0.71070687
a_4 = -0.142248368
a_5 = 0.127414796
b = 0.2316419

The absolute error in above approximation is < 7.5 x 10^{-8}.

The equation in a slightly different form can be found in Ambramowitz et al (1970) equation 26.2.17

ANNEX 4.2 Inverse of the standard normal distribution

The standard normal distribution function is given by (A4.1.1). The inverse of the standard normal distribution is found from:

$$
\begin{aligned}
&y = T - \frac{a_1 + a_2 T + a_3 T^2}{1 + a_4 T + a_5 T^2 + a_6 T^3} \\
&\text{for} F_Z(z) < 0.5 : z = -y \\
&\text{for} F_Z(z) \geq 0.5 : z = y \\
&\text{with} : T = \sqrt{-2 \ln P} \\
&\text{where} : P = F_Z(z) \text{for} F_Z(z) \leq 0.5 \\
&\text{and} P = 1 - F_Z(z) \text{for} F_Z(z) > 0.5
\end{aligned}
\qquad \text{(A4.2.1)}
$$

The coefficients in (A4.2.1) read:

a_1 = 2.515517

a_2 = 0.802853

a_3 = 0.010328

a_4 = 1.432788

a_5 = 0.189269

a_6 = 0.001308

The absolute error in above approximation is < 4.5×10^{-4}.

The equation can be found in Ambramowitz et al (1970) equation 26.2.23.

ANNEX 4.3 Incomplete gamma function

The incomplete gamma function is defined by:

$$F_Z(z) = \frac{1}{\Gamma(\gamma)} \int_0^z t^{\gamma-1} \exp(-t)dt \qquad \text{(A4.3.1)}$$

To determine the non-exceedance probability for any value of z > 0 the following procedure is used. Three options are considered dependent on the value of γ and z:

- If $\gamma \geq 500$: then the Wilson-Hilverty transformation:

$$y = 3\sqrt{\gamma}\left[\left(\frac{z}{\gamma}\right)^{1/3} - 1 + \frac{1}{9\gamma}\right] \qquad \text{(A4.3.2)}$$

  The variable y has a standard normal distribution.

- If $z \leq \gamma$ or $z \leq 1$ a rapidly converging series development is used:

$$F_Z(z) = \exp(-z)z^{\gamma} \sum_{j=1}^{\infty} \frac{z^{j-1}}{\Gamma(\gamma + j)} \qquad \text{(A4.3.3)}$$

  The algorithm is taken to have converged when the summation S fulfils:

$$\frac{S_n - S_{n-1}}{S_n} \leq 10^{-6}$$

- If $z > \gamma$ and $z > 1$ a rapidly converging continued fraction development is used:

$$F_Z(z) = 1 - \frac{\exp(-z)z^{\gamma}}{\Gamma(\gamma)} \left(\cfrac{1}{z + \cfrac{1-\gamma}{1 + \cfrac{1}{z + \cfrac{2-\gamma}{1 + \cfrac{2}{z + \cfrac{3-\gamma}{1 + \cfrac{3}{z + \ldots.}}}}}}} \right)$$

  or shortly written as:

$$F_Z(z) = 1 - \frac{\exp(-z)z^{\gamma}}{\Gamma(\gamma)} \left(\frac{1}{z+} \frac{1-\gamma}{1+} \frac{1}{z+} \frac{2-\gamma}{1+} \frac{2}{z+} \frac{3-\gamma}{1+} \frac{3}{z+} \ldots\ldots\ldots \right) \qquad \text{(A4.3.4)}$$

The continued fraction S can be rewritten as:

$$S = \frac{1}{z}\left(1 + \frac{\gamma - 1}{(2 - \gamma + z) +} \frac{\gamma - 2}{(4 - \gamma + z) +} \frac{2(\gamma - 3)}{(6 - \gamma + z) +} .....\right)$$

The n^{th} convergent of S reads:

$$S_n = \frac{A_n}{B_n} = \frac{1}{z}\left(1 + \frac{a_1}{b_{1+}} \frac{a_2}{b_2 +} \frac{a_3}{b_3 +} ..... \frac{a_n}{b_n}\right) \qquad \text{(A4.3.5)}$$

which is calculated using recursively:

$A_0 = 1$ $B_0 = z$

$A_1 = z + 1$ $B_1 = z(2 - \gamma + z)$

$a_j = (j - 1)(\gamma - j)$ $b_j = 2j - g + z$

$A_j = b_jA_{j-1} + a_jA_{j-2}$ $B_j = b_jB_{j-1} + a_jB_{j-2}$ for: j = 2, ……, n

The iteration is taken to have converged when:

$$\frac{S_n - S_{n-1}}{S_n} \leq 10^{-6}$$

ANNEX 4.4 Inverse of incomplete gamma function

The above procedure is also used to arrive at the inverse of the incomplete gamma function. For this the routine to compute the incomplete gamma function is seeded with a variate $z = 2^k$, for k = 1, 2, …, 50. The function returns the non-exceedance probability $F_Z(z)$ for each z.

Let the required exceedance probability be denoted by P. If for a particular value of $z = 2^k$ the function return be an $F_Z(z) > P$, then the computation is stopped and an interpolation is made between $z = 2^{k-1}$ and 2^k such that $F_Z(z) - P = 0$. The interpolation is repeated to arrive at a required accuracy.