WATERnet

Building Capacity for Water Resources Management in Southern Africa

Training Course Report

Water Resources Assessment in Sub-Saharan Africa: Prediction in Ungauged and Data Scarce Basins

21 — 25 January 2008, Dar es Salaam, Tanzania

in collaboration with

TU Delft
Delft University of Technology

Cap-Net

University of Dar es Salaam

Global Water Partnership Southern Africa

University of Botswana

UNESCO-IHE
Institute for Water Education

Water Resources Assessment in Sub-Saharan Africa: Prediction in Ungauged and Data Scarce Basins

Training Course Reader

Authors:
T.A. Bogaard
S. Mkhandi
B.P. Parida
H.C. Winsemius

List of content

Preface

Sub-Saharan countries have limited financial, human and technical resources for developing and maintaining hydrometric networks that can provide data for sustainable water resources assessment (required for planning, design and management). Whereas the needs for hydrological information are increasing, the number of meteorological and hydrological stations in Africa have been declining during the last 30 years. The continuing problems arising from lack of data may impede the effective cooperation on management of shared trans-boundary water resources and drought and flood management with major social, economic and environmental tragedies. It is for these reasons that the conveners developed this training course focusing on hydrological prediction in ungauged, poorly gauged or data scarce environments within southern Africa. The set-up of this course was stimulated by the IAHS Science initiative PUB ('Predictions in Ungauged Basins'; http://pub.iwmi.org, Sivapalan et al., 2003).

The course objective is "to develop a training module to equip water resource managers and planners with the necessary skills and understanding of water resources assessment in case of ungauged or data scarce environments within sub-Saharan Africa". It is also aimed to highlight the problems of lack of reliable data as well as presenting methods to cope with the data shortage such as transferring knowledge between gauged and ungauged basins and the potential of satellite derived information for the hydrological practice.

The course is set up as follows: All participants will follow a one-day refresher course in hydrology in which all participants will get acquainted with the same hydrological vocabulary. Thereafter, an introduction in new data sources is given, together with basic hydrological analysis using the freely available data. The course continues to deepen the participant's knowledge on hydrograph separation and discuss the worth of doing short field campaigns to assess the hydrological behaviour of a catchment. Furthermore, the following two days the participants will work with advanced statistical and regionalisation techniques for water resources assessment to regionalise the hydrological variables.

The course will use and refer to the e-learning basic hydrology course: Vicaire: VIrtual CAmpus In hydrology and water REsources management (http://hydram.epfl.ch/VICAIRE/) as basis. The participants are encouraged to use this information source to improve or refresh their hydrological knowledge.

This reader is set-up as convenient handbook for the course. The reader's content follows the course set-up closely. The reader is assembled from existing course material, from the VICAIRE web-course and partly written specifically for this course. Chapter 1 gives background information for the hydrological processes within the hydrological cycle. Special attention is give to hydrographs and hydrograph separations. The new data sources are described in chapter 2. Chapter 3 gives the statistical analyses and frequency analyses indispensable for water resource management. Chapter 4 continues with regionalisation techniques. The powerpoint hand-outs and exercises and solutions to the exercises are part of the course material but will be handed out during the course.

December 2007
Delft, Dar es Salaam, Garbone
Thom Bogaard, Simon Mkhandi, Bhagabat Parida and Hessel Winsemius

Chapter 1
Basic hydrology, hydrograph analysis and catchment characteristics

Sources:
Vicaire e-learning: http://hydram.epfl.ch/VICAIRE/ §1.1
Land Surface Hydrology Reader, T.A. Bogaard, University Utrecht, 2005 §1.2- §1.3

Recommended reading:
Chapter 11 Brutsaert “Streamflow generation: mechanisms and parametrisation”

Objective: After the first day, all participants have refreshed their basic hydrological knowledge and use the same terminology. Participants are familiar with PUB.

Objective: After the third day, all participants have good understanding of discharge generation processes and skills in hydrograph separation techniques. The participants are familiar with the advantages and disadvantages of this technique. Furthermore, the participants can plan short field campaigns in order to add valuable information about the hydrological behaviour of a study region.

1.1 THE HYDROLOGICAL CYCLE

1.1.1 Definition of Hydrology

Water is vital for all living organisms on Earth. For centuries, people have been investigating where water comes from, where it goes, why some of it is salty and some is fresh, why sometimes there is not enough and sometimes too much. All questions and answers related to water have been grouped together into a discipline. The name of the discipline is hydrology and is formed by two Greek words: "hydro" and "logos" meaning "water" and "science". Hydrology is the science concerned with the occurrence, distribution, movement and properties of all the waters of the Earth. A good understanding of the hydrologic processes is important for the assessment of the water resources, their management and conservation on global and regional scales.

1.1.2 Water Distribution on Earth

The "Blue Planet", as the Earth is called, is easily identified in the solar system due to its distinctive element: water. Oceans and seas cover 71% of the planet's surface. The remaining 29% are land, but water can be found here as well in lakes and rivers, in the soil cover, underground and bound up in the composition of minerals of the Earth's crust and core. The

biosphere contains water and cannot exist without it. Water is held in the atmosphere together with other gases.
Water exists in three states: liquid, solid (ice and snow) and gas (water vapour). Due to the energy supplied by the sun it is in permanent transformation from one state to another, and in constant motion between oceans, land, atmosphere and biosphere.
A reliable assessment of the water storage on Earth considers the amount of water as an average over a long period of time, contained in the hydrosphere. Current estimations weigh up to 1386 10^6 km^3 of water that are divided as shown in Figure 1.1 and Table 1.1.

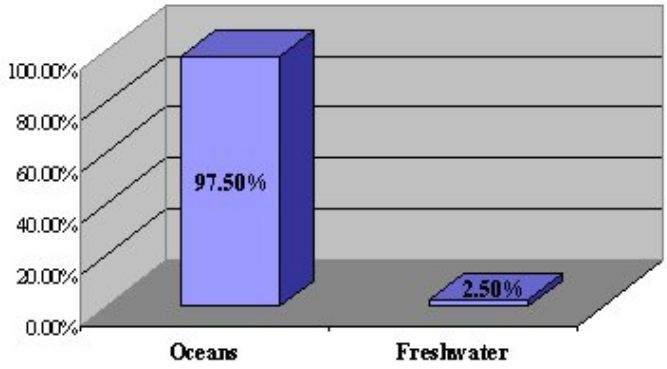


Figure 1.1. Distribution of water on Earth

Table 1.1. Distribution of freshwater on Earth

Component of freshwater	**% of the hydrosphere content**
Glaciers and Permanent Snow Cover	1.74 %
Groundwater	0.75 %
Freshwater Lakes	0.0066 %
Rivers	0.0002 %
Atmosphere	0.0009 %
Biosphere	0.0001 %

Freshwater is only 2.5% from the total, yet most of it is out of human reach. Freshwater usable by humans represents 0.3% of all water on Earth and is drawn from underground, lakes and rivers (Figure 1.2).

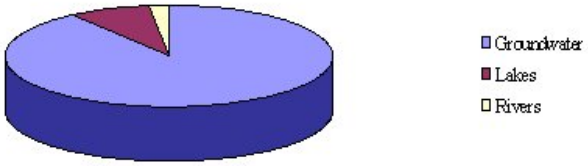


Figure 1.2. Freshwater available for human use (0.3 % of Earth's water)

Groundwater is the second largest storage of freshwater and widely used by humans. People in the arid and semi-arid regions, use groundwater exclusively for all their needs. Still, groundwater is not always within easy reach. The withdrawal of groundwater becomes difficult and expensive when it is confined over 800m depth. The surface water bodies, as lakes and rivers, hold a very small amount of freshwater. Unlike groundwater it is easily accessible, but liable to pollution. At the same time, it is unevenly distributed with regard to continent surfaces and population. For example, 30 % of the world freshwater storage and 6 % of runoff are located in Canada alone.

1.1.3 The Hydrologic Cycle

The total amount of water on Earth is invariable. At the same time water is continuously renewed while circulating between oceans, land and atmosphere. All processes like evaporation, condensation, precipitation, interception, transpiration, infiltration, storage, runoff, groundwater flow, which keep water in motion constitute the hydrologic cycle. Those processes are stimulated by solar energy. The processes evaporation and transpiration are often combined in practice to the term evapotranspiration (ET), but care should be taken with this combined term as the two represent different processes with different time scales. They take place simultaneously and, except for precipitation, continuously. Consider Figure 1.3.

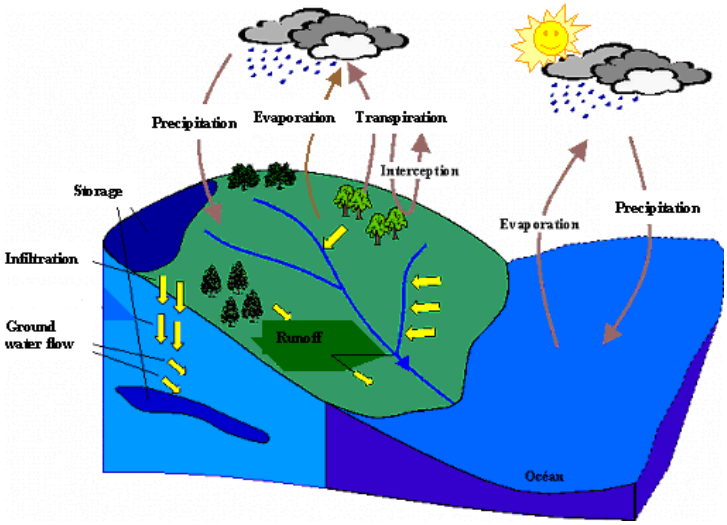


Figure 1.3. The hydrologic cycle (Musy, 2001)

The water volume of each water body in the hydrosphere is fully replenished during the hydrologic cycle, but the time period required varies as is shown in Table 1.2. The period of renewal is defined as the mean discharge (or more general, mean loss) divided by the mean volume of that water body (source).

Table 1.2. Period of renewal of water in the hydrosphere (UNESCO, 2004)

Water body	**Period of renewal**
Polar ice	9700 years
Oceans	2500 years
Mountain glaciers	1600 years
Groundwater	1400 years
Lakes	17 years
Rivers	16 days
Atmospheric moisture	8 days
Water in the biosphere	Several hours

1.1.4 The Water Budget

The water budget represents the inventory of water for a specific water body (or hydrologic region) during a certain time interval. It can be estimated using the continuity equation, which expresses the balance between the inflows, outflows and change of storage in any water body / hydrologic region over a period of time:

$$P - R - G - E - T = \Delta S / \Delta t \qquad (1.1)$$

Where:

- P precipitation, [height / time] or [volume / time]
- R surface runoff, [height / time] or [volume / time]
  - $R = R_{out} - R_{in}$
  - R_{out} = runoff as outflow from the water body / hydrologic region
  - R_{in} = runoff as influx into the water body / hydrologic region
- G groundwater flow, [height / time] or [volume / time]
  - $G = G_{out} - G_{in}$
  - G_{out} = groundwater as outflow from the water body / hydrologic region
  - G_{in} = groundwater as influx into the water body / hydrologic region
- E evaporation, [height / time] or [volume / time]
- T transpiration, [height / time] or [volume / time]
- $\Delta S / \Delta t$ change in storage per time, [height / time] or [volume / time]

Equation 1.1 is the basic equation of hydrology and useably given in [mm/time]. In practice it is successfully applied for local studies when the various hydrologic terms can be properly measured or estimated. Nevertheless estimation is usually rough at larger scales. On a catchment scale, the main storages, S, are the groundwater, soil water, surface water (reservoirs), and vegetation. The stocks are very important for the hydrological responses within a catchment on precipitation events.

Annual water budget of the Earth

Every year approximately 577000 km^3 are transported through the hydrosphere. The annual water budget is displayed on Figure 1.4.

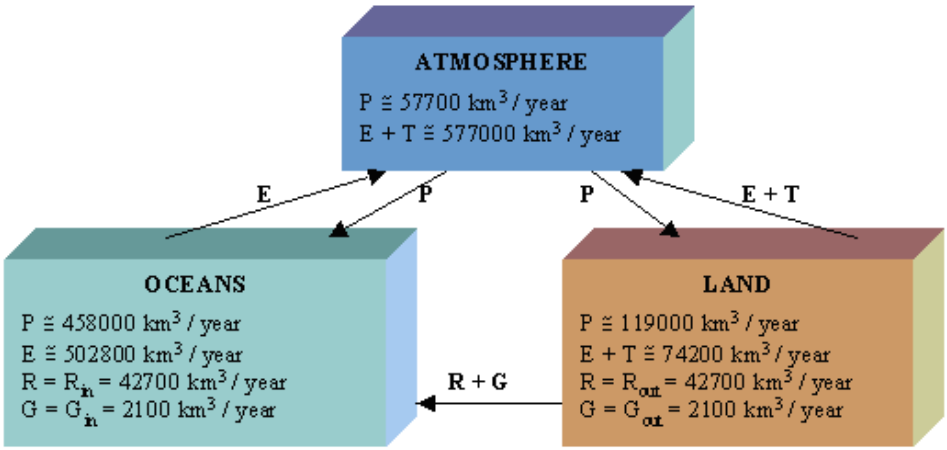


Figure 1.4. Annual water budget of the Earth

Annual budget of some hydrologic regions in Europe and Africa

Table 1.3. Annual water budget of Switzerland (Musy, 2001)

Precipitation			1546 mm / year
Total runoff	Runoff	978 mm / year	1296 mm / year
	Influx into Switzerland	318 mm / year	
Evaporation			484 mm / year

Table 1.4. Annual water budget of Romania (National Institute of Meteorology and Hydrology, Regional Office, Timisoara)

Precipitation	850 mm / year
Runoff	300 mm / year
Evaporation	550 mm / year

Table 1.5. Annual water budget of Bulgaria (Geography of Bulgaria, monograph, Bulgarian Academy of Sciences , 1989)

Precipitation	690 mm / year
Runoff	176 mm / year
Evaporation	514 mm / year

Table 1.6. Annual water budget of Ukraine

Precipitation	625 mm / year
Runoff	86.8 mm / year
Evaporation	538 mm / year

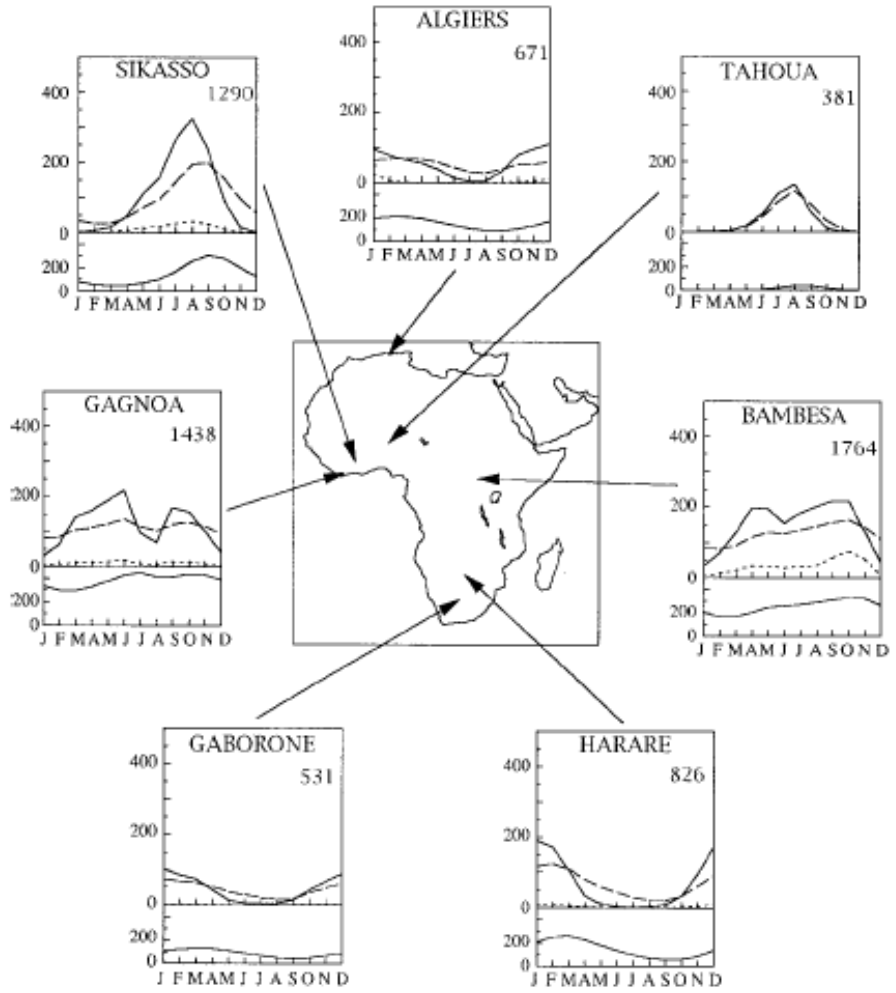


Figure 1.5. Mean water balance at seven diverse African locations. Top diagram: solid line is rainfall, dashed line is evapotranspiration, and dotted line is runoff, all in mm month^{-1}. Bottom diagram: mean monthly exchangeable soil moisture in millimeters. Mean annual rainfall (in mm) is given in the upper right (from Nicholson et al., 1997)

Table 1.7. Water balance parameters for 15 latitudinal zones over Africa. First three columns are the annual runoff ratio, the annual evaporation ratio, and annual evapotranspiration, as calculated by Nicholson et al., 1997. The fourth column is Henning's estimates of evapotranspiration, and the fifth column is Baumgartner and Reichel's estimate of the evaporation ratio (from Nicholson et al., 1997).

	N/P	E/P	E (mm)	E_H (mm)	E/P_{BR}
35 –40 N	10	90	594	355	89
30 –35 N	1	99	215	186	105
25 –30 N	0.0	100	23	55	117
20 –25 N	0.0	100	35	58	97
15 –20 N	0.2	100	244	198	97
10 –15 N	5	95	727	517	93
5–10 N	14	86	1193	773	81
Equator–5 N	13	87	1144	846	74
5 S–equator	12	88	1178	958	77
10 –5 S	9	91	1058	880	84
15 –10 S	12	88	1023	772	86
20 –15 S	5	95	730	715	85
25 –20 S	2	98	446	559	91
30 –25 S	3	97	450	428	93
35 –30 S	3	97	441	423	92

1.1.5 A Brief History of Hydrology

Hydrology has been a subject of investigation and engineering for millennia. For example, in about 4000 B.C. the Nile was dammed to improve agricultural productivity of previously barren lands. Mesopotamian towns were protected from flooding with high earthen walls. In ancient times various hydrologic principles were successfully applied in practice. Early Chinese irrigation and flood control works and Greek and Roman aqueducts are worth mentioning. On the other hand ancient science was based only on logic and intuition, without measurements and observations, and the theories were faulty most of the time.

A famous example of detailed and accurate hydrological measurements from ancient periods, is the Nilometer, that dates as far back as 861 AD and still can be seen on the island of Rodah in central Cairo (figure left side)

Homer (8th century B.C.) believed in the existence of large subterranean reservoirs that supplied rivers, seas, springs and wells. The Roman engineer Marcus Vitruvius (1st century B.C.) developed an early theory of the hydrologic cycle in his treatise 'On Architecture'. According to his theory the rain and snow falling in mountains infiltrated into the ground and later appeared in the lowland as streams and springs. During the Middle Ages, Vitruvius's work was the standard reference book on Hydrology.

One of the first discharge measurements described is by Leonardo da Vinci, measuring flow velocities using floater experiments. He used a floater that was made heavier as displayed in the figure above. The distance was measured using an odometer and he estimated the time by rhythmic chanting (From Chow, 1988, p.15)

In the late 15th century Leonardo da Vinci and Bernard Palissy gave, independently of each other, an accurate explanation of the hydrologic cycle. The theories were based on observations of hydrologic phenomena. In the 17th century the modern science of hydrology was established by Perrault, Mariotte and Halley. Perrault measured the rainfall and runoff in the Seine River and proved that rainfall contributes significantly to river flow. He also measured evaporation and capillarity. Mariotte recorded the velocity of flow in the Seine River and made measurements of the cross section, estimating the discharge. Halley measured evaporation of the Mediterranean Sea.

The Bernoulli piezometer and theorem, the Pitot tube and Chezy's formula are representative achievements of the 18th century. During 19th century experimental hydrology made considerable progress: Darcy's law of flow in porous media and Dupuit-Thiem's well formula were elaborated. Early 20th-century governmental agencies developed their own programs of hydrologic research. Sherman's unit hydrograph, Horton's infiltration theory and Theis's non-equilibrium approach to well hydraulics were based on their analyses and were the results of research programs. After 1950 the progress in sciences and the high-speed digital computers opened new perspectives in hydrology.

1.2 HYDROGRAPHS AND CATCHMENT CHARACTERISTICS

River discharge is the result of a combination of climatic forcing (precipitation and evapotranspiration) and the catchment characteristics, like topography, geology, soil type, land use, soil moisture condition, etc. The shape of the hydrograph is the result of the interaction of all these processes. This chapter describes the hydrograph characteristics. This chapter assumes that the reader has basic knowledge of hydrological (catchment) processes as for example described in ‘Principles of Hydrology’ of R.C. Ward and M. Robinson (1990).

A hydrograph is a graph showing stage, flow, velocity, or other property of water with respect to time (Langbein and Iseri, 1995). A hydrograph is the result of all the rainfall-discharge processes

in a catchment and therefore contains valuable information on catchment processes. It contains direct precipitation, surface runoff and groundwater, it also includes water which flows fast to the outlet and the water that takes long flow paths and thus time to reach the outlet. In combination with rainfall measurements, one can calculate the amount of water that did not reach the outlet after a certain period, because of evapotranspiration, infiltration or temporarily storage within the catchment.

1.2.1 Hydrograph properties

A hydrograph consists of three parts:

1) Rising limb = the rising part of the hydrograph when runoff is increasing
2) Peak = maximum discharge of a flood event
3) Falling limb = the recession curve of the hydrograph when the runoff rate is decreasing.

Preceding and following the flood wave the hydrograph is called the baseflow recession.

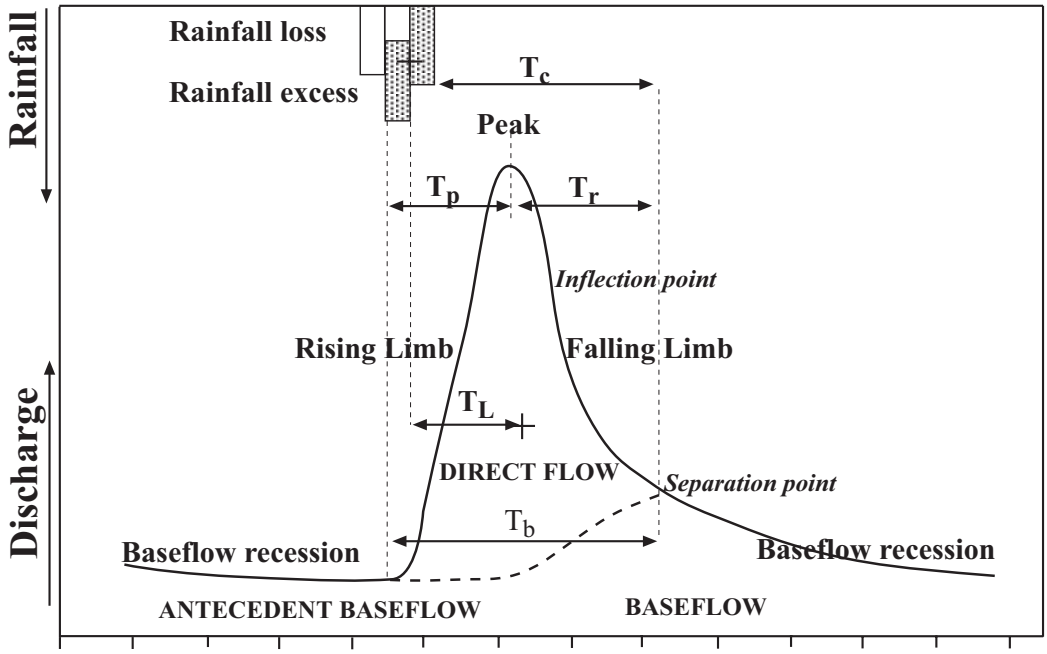


Figure 1.6 Hydrograph properties

A hydrograph shape can be described with the following time properties [second, minute, hour, day, week, month or year](Figure 1.6):

1) Time to peak (t_p) = the time interval between the rainfall excess and peak of the hydrograph.
2) Recession time (t_r) = the time interval from the peak of the hydrograph till the end of the surface runoff.
3) Time base (t_b) = the time interval from the beginning to the end of the surface runoff.
4) Time lag (t_l) = the time between the centre of mass of the precipitation excess and the centre of mass of the discharge.

Note: Sometimes lag time is also defined as the time from beginning (instead of centre of mass) of rainfall to the centre of mass of runoff, or as the time from centre of mass of rainfall to the peak of the runoff.

5) Time of concentration (t_c) = the time required for water to flow from the outermost point on the watershed to the gauging station.

   Some textbooks write that while time of concentration is conceptual the time required for 100 percent of the watershed to contribute (a different definition than ours!), it is sometimes defined as the time from the end of excess rainfall to the inflection point (assuming to separate overland and interflow) on the hydrograph recession limb. The reasoning is that direct runoff ceases at the point of inflection and thus not 100% of the catchment is contributing anymore. Here this reasoning is not followed.

   Here we follow the reasoning that the time of concentration is equal to the time interval from the end of the rainfall excess to the 'separation point' on the recession curve. So we assume that in a catchment the last direct runoff is discharged when the raindrop with the longest travel time reaches the outlet.

Hydrograph records can have different time basis: continuously, hourly, daily or monthly are the most common. Discharge information of large river basins is often reported on a daily time scale, i.e. the arithmetic average (mean) of the discharge from midnight to midnight. In climate studies it is quite common to report discharge as monthly averages. The main disadvantage of discrete time intervals is that you miss information of e.g. floods with a shorter time base then the one used. Main advantage is that it reduces the data load and saves calculation time for modelling studies. It is for example, not necessary (and also sometimes impossible) to do modelling studies to quantify the changes in river runoff of the river Zambesi as result of (enhanced) climate change in the next 100 year while using rainfall and discharge data on hourly time scale.

1.2.2 Stream types

With the above description of hydrographs, it is possible to classify stream types by their hydrograph shapes. Figure 1.7 shows three different types of hydrographs which are prototypes for three catchment systems.

The first is an example of a perennial stream, a slow reacting and continuously flowing river system. A perennial stream can be defined as a stream that flows from source to mouth throughout the year. The catchment has a groundwater storage system which contains enough water through the year to maintain continuous runoff. These systems have hydrographs that generally show slow reaction to rainfall events and gentle recession curves. Often they also have relatively small food peaks compared to the average runoff.

The third graph shows its antagonist, an ephemeral stream. These are streams that flow only in direct response to precipitation, and thus discontinues their flow during dry seasons. Such flow is usually of short duration. Most of the dry washes of more arid regions may be classified as ephemeral streams. These systems have no groundwater system (storage), which is directly connected to the stream flow. The hydrographs show steep rising and falling limbs and dry periods between the rain event.

In-between these two extremes are the intermittent streams. These streams carry water only part of the year, generally in response to periods of heavy runoff either from snowmelt or storms; a

stream or part of a stream that flows only in direct response to precipitation. It receives little or no water from springs or other sources. It is dry for a large part of the year, generally more than three months. Flow is likely to occur for several weeks or months in response to seasonal precipitation, due to groundwater discharge, in contrast to the ephemeral stream that flows but a few hours or days following a single storm. Intermittent streams are also referred to as seasonal streams. These catchments have some but limited groundwater systems that releases its water relatively fast. The recession time are generally shorter then those of the perennial streams.

1.2.3 Catchment characteristics

A catchment is defined as the area which supplies water by surface and subsurface flow from precipitation to a given point in the drainage system. Similar words are drainage basin or area, river basin or watershed. The latter is originally the divide separating one drainage basin (or catchment) from another and in the past has been generally used to convey this meaning. However, over the years, use of the term to signify drainage basin or catchment area has come to predominate, although drainage basin or catchment area is preferred. Drainage divide, or just divide, is used to denote the boundary between one drainage area and another. Used alone, the term "watershed" is ambiguous and should not be used unless the intended meaning is made clear (Langbein and Iseri, 1995). Note that in general it is assumed that the topographic catchment coincides with the groundwater catchment. This does not have to be the case as complex geological structure can transport groundwater underneath catchment divides!
The geomorphology of a catchment like size, slope, stream network influences the drainage characteristic of the catchment which is represented in a hydrograph. Also surface roughness e.g. land use influences the hydrograph. Figure 1.8 is taken from Wanielista (1990), it shows the effects of drainage characteristics on discharge hydrograph. Two characteristics are of importance: reaction characteristic of a flood event and the volume of discharge. Interception, infiltration and storage determine the percentage of rainfall that is discharged from the catchment system. The time that a flood peak arrives after a certain rain event is also an important parameter for a catchment. Steep slopes or concrete river beds will carry rainwater quickly to the outlet.
The larger a catchment is the slower a hydrograph will react on rainfall. Large catchment have a large buffer capacity and the travel time is large (time of concentration). In Chapter 4 catchments model will be discussed which are based on rainfall losses, on the travel time concept and on storage characteristics of a catchment.

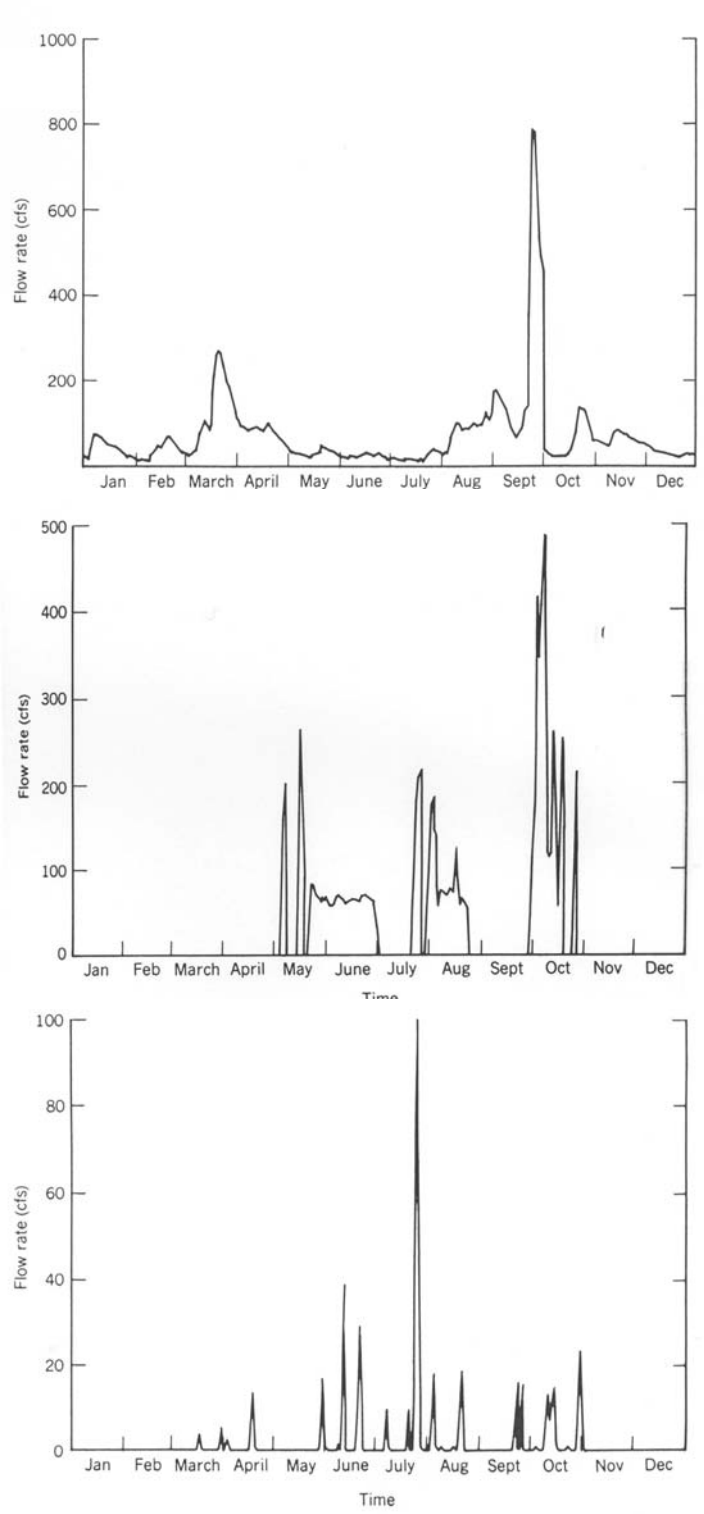


Figure 1.7 Examples of hydrographs of perennial, intermittent and ephemeral streams (From Wanielista, 1990).

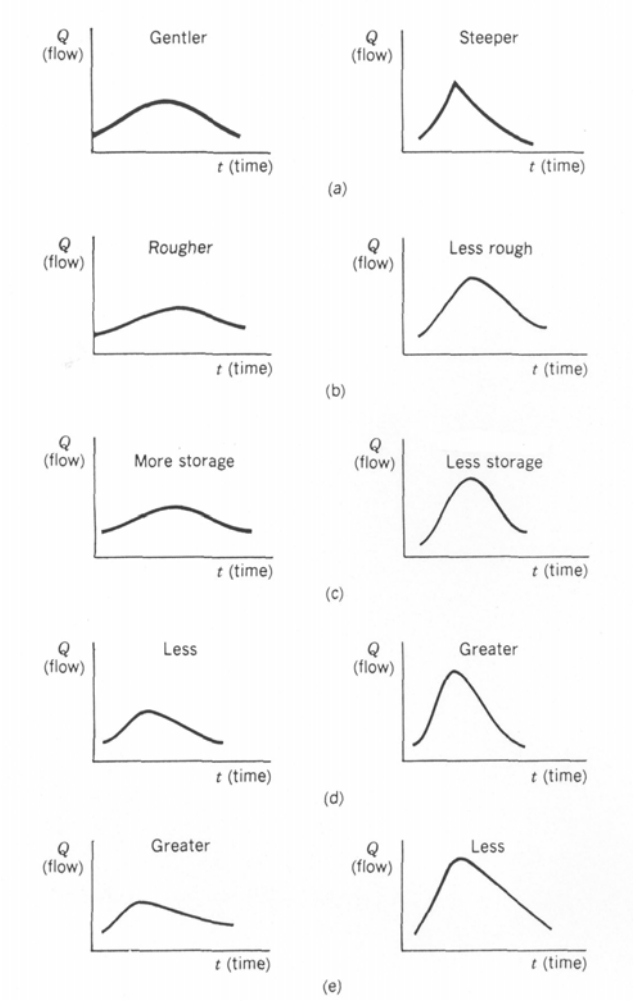


Figure 1.8 The effects of drainage characteristics on discharge hydrograph. a) Catchment with gentle slope versus steep slopes , b) A catchments having relatively rough surface (natural forest) versus a catchment with less surface roughness (urbanised area), c) A catchment having large storage capacity versus less storage capacity, d) shows the influence of direct connected impervious areas, and e) shows the effect of the infiltration volume (From Wanielista, 1990).

It should not be forgotten that also rainfall characteristics have an enormous influence on the hydrograph shape. High rainfall intensities will lead to less infiltration (according to the Horton Overland Flow principle) and thus to higher runoff rates. Also antecedent precipitation will saturate the soil and thus limit the amount of rainfall loss and thus increase runoff. It is important to be aware of the effects of partial rain coverage. Figure 1.9 shows the influence of the size of a rainfall event. As in many practical cases the amount of rain gauges is limited, it is quite difficult to quantify the rainfall coverage and intensity effects in catchment hydrology. Two types of errors can be distinguished. The first is the error associated with the point measurement

(generally in the order of 5-15% for rain and 10-40% for snow). The second error is not an equipment error but the fact that the precipitation is not spatial and temporal homogeneous. Many researchers have shown that the influence of errors in rainfall measurements (equipment as well as spatial and temporal heterogeneity of the rainfall) is much more relevant for hydrologist then the influence of the hydraulic parameterisation of a catchment e.g. infiltration characteristic (e.g. Nandakamur and Mein, 1997). Lastly, it is interesting to remark that also the direction of a storm over a catchment will influence the hydrograph (see Figure 1.10).

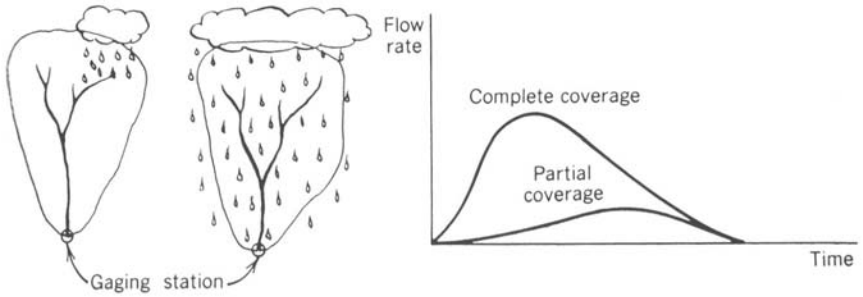


Figure 1.9 The effect of partial coverage of rainfall over a catchment on a hydrograph (From Wanielista, *1990)*

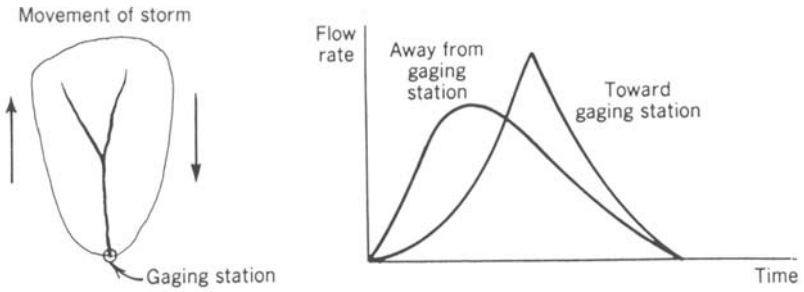


Figure 1.10 The effect of storm direction over a catchment on a hydrograph (From Wanielista, 1990)

1.2.4 Stream networks

For catchment descriptions and modelling it can be useful to have a system to describe and quantify subcatchments and channel networks. The geomorphologist Horton published in 1945 his quantitative study on stream networks. Later his work has been modified by Strahler (1964). The practical hypothesis behind this concept is that catchments can be compared with each other on basis of the size of the catchments, its channel dimensions and streamflow which are all proportional to stream order. One has to take care to investigate a large enough sample.

A catchment is divided into individual streams according to drainage segments. The smallest tributaries that can be distinguished at a certain scale are classified as stream order 1. Where two tributaries of stream order 1 join is the beginning of a stream order 2. If a stream order 1 joins a stream order 2, the downstream order retains 2. So a stream order increases only

when two equal stream orders join. The order of the catchment is designated as the order of the stream drainage at the outlet, which is the highest stream order in the basin (order 4 in Figure 1.11).

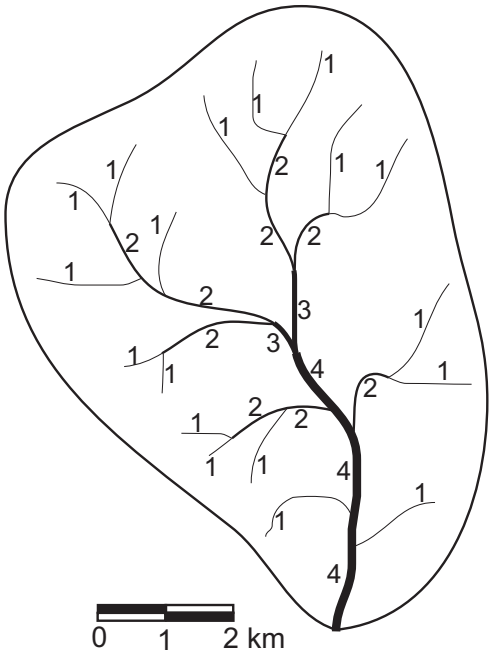


Figure 1.11 Stream order network according to Horton/Strahler

The stream orders and (sub)catchment areas are the basis for quantitative catchment analyses, often performed in a GIS environment with a digital elevation model (DEM) as basis. Horton analysed that the ratio of the number of streams of order N_i to the number N_{i+1} streams is relatively constant from one order to the other. This is the Horton's law of stream orders, or the Bifurcation Ratio R_B.

$$R_B = \frac{N_i}{N_{i+1}}$$

The theoretical minimum value of R_B is 2, but values typical lie in the range of 3-5 (Wanielista, 1990). In Figure 1.11 $R_{B1-2} = 18/6 = 3.0$, $R_{B2-3} = 5/2 = 2.5$ and $R_{B3-4} = 2/1 = 2.0$.

Horton also defined the length ratio R_L, which is the ratio of average order length with the one of one order lower.

$$R_L = \frac{L_{i+1}}{L_i}$$

with L_i being the average length of a stream of order i.
Assuming the following average length for the stream orders in Figure 1.11 : $L_1 = 1.4$ km, $L_2 = 1.9$ km, $L_3 = 0.75$ km and L_4 is 4 km, we obtain $R_{L2-1}=1.9/1.4=1.36$, $R_{L3-2}=0.75/1.9=0.39$ and $R_{L4-3}=4/0.75=5.33$.

Schumm (1956) extended Horton's work with the area ratio.

$$R_A = \frac{A_{i+1}}{A_i}$$

with A_i being the average catchment size of order i. The calculation is similar to that of the length ratio.

Besides that these analyses are quite useful for geomorphological work, they also form the basis for the geomorphological instantaneous unit hydrograph. In that concept the catchment characteristics are translated mathematically into travel times of rainwater to the outlet. By integrating all travel times over a catchment a synthetic hydrograph can be derived.

1.3 HYDROGRAPH SEPARATION: DIRECT FLOW AND BASE FLOW

Besides morphological analysis of a hydrograph and the use of these descriptive characteristics in relating them to catchment characteristics, we can try to relate hydrographs to hydrological processes that play a role in runoff generation. Hydrographs are a combination of water that leaves a catchment long after a rain events ceased, and of water that is discharged from a catchment during and shortly after a rain event.

Rainwater can be transported with all possible velocities after a rain event out of the catchment. Rainwater can be attenuated and delayed because of soil hydrological processes. However, some of the discharge shows a short response on rainfall as it has undergone little or no soil hydrological processes. This is also why the process-based classification is still often used. Fast response is direct runoff and slow response water represents soil and ground water. Intermediate response is interflow water (something like subsurface (storm?) flow). From this insight in catchment response on rainfall events the so-called processes based (or white box) classification originates. The terms used are: surface runoff, interflow, groundwater flow.

A second and stricter classification is the time based (or black box) classification. The hydrograph is separated in water with short response time, and water with a much longer response time. The terms used in this classification are: quick flow and slow flow (delayed flow) or direct flow and base flow. We will use this classification in case we only have hydrograph information. This because half a century of catchment hydrological research (e.g. with tracers) has shown that discharge generation is not restricted to "fast overland flow" and "slow groundwater flow". Fast responses can be generated in the subsurface as well. When analysing hydrographs we should be aware that a response such as a hydrograph does not tell us the underlying generation processes.

Although the former terminology (quick/fast and slow/delayed) are clearly the most strictly related to time, the latter terminology (direct and base) are the ones which are most used by the hydrological community. It is this terminology that we will follow.

The question is how we can separate a hydrograph in direct flow and base flow? All possible separation techniques have some kind of assumptions in it. You have to agree on the method to follow, as there is no overall truth how it should be done. One of the most used assumptions is that base flow behaves like a linear reservoir. By making a graph of the natural

logarithm of the discharge versus time we normally see a straight line in the lower part of the recession limb (see Figure 1.12).

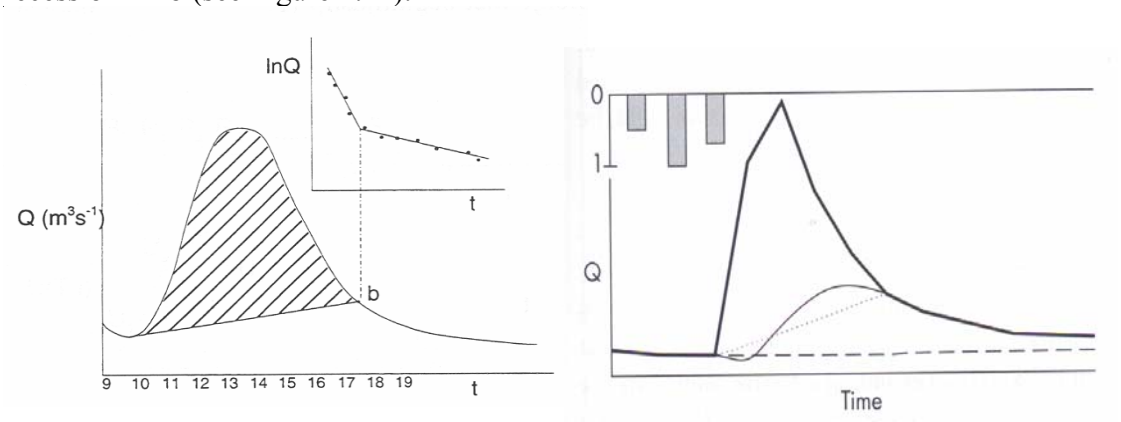


Figure 1.12: Separation of hydrograph in base flow and direct flow components using the normal depletion curve method. The figure on the right shows two other possible separation curve.

We define that part of the hydrograph that discharges according to a linear reservoir model (straight line in ln(Q) - time graph) as base flow. The point on the recession limb of the hydrograph, which is the beginning of the base flow, can be called the separation point (point b in Figure 1.12). Direct flow starts at the moment that the hydrograph rises (rising limb). The separation of the hydrograph in base flow and direct flow is an unknown line between these two points on the hydrograph. Although several kinds of lines are proposed in literature (see Chow, 1988, figure 5.2.3), we will only use the straight line method that connects the beginning of the rising limb with the separation at the recession limb of the hydrograph (see Figure 1.12).

Manual base flow separation methods are labour intensive and are generally not objective; different users given the same data would probably arrive at somewhat different values for base flow. To overcome this lack of objectivity, several automatic base flow separation procedures are proposed. In their series of Low Flow Studies the Institute of Hydrology defined the Base Flow Index (BFI) as a catchment characteristic (IOH, 1980). The base flow is get from a smoothed minimum mean daily flow (5-day increments and the minimum flow during this 5-day period is identified). Straight lines between minimum flow points (called turning points) define the base flow hydrograph. This base flow line is then separated from the total hydrograph to get the direct flow. The BFI is the ratio of the smoothed minimum mean daily flow to the mean daily flow of the total recorded hydrograph and can be considered as a measure of the proportion of the river runoff that derives from natural storage.

For base flow analysis, the research consortium of the FRIEND collaboration defined their own criterion. A base flow recession analysis can be performed on recession curves that start two days after a peak runoff and continue for at least 7 days. In short this criterion simply chooses base flow to start two day after the peak flow.

Bibliography

Al-Wagdany, A.S. and A.R. Rao (1998). Correlation of the velocity parameter of three geomorphological instantaneous unit hydrograph models. Hydrological Processes, Vol.12, pp.651-659.

Beven, K.J. (ed) (1997). Distributed hydrological modelling: Applications of the TOPMODEL concept. John-Wiley & Sons. 348pp

Bos, M.G. (ed.) (1978). Discharge measurement structures. Second edition, ILRI-publication 20, Wageningen, 464 pp.

CHO (1986). Verklarende hydrologische woordenlijst. TNO-CHO. (Dutch-English glossary of hydrological terms)

Chow, V.T. (1959). Open-Channel hydraulics. McGraw-Hill, 680 pp.

Chow, V.T., D.R. Maidment and L.W. Mays (1988). Applied hydrology. McGraw-Hill, 572 pp.

Franchini, M. and P.E. O'Connell (1996). An analysis of the dynamic component of the geomorphologic instantaneous unit hydrograph. Journal of hydrology, Vol.175, pp.407-428.

Gupta, V.K., Waymire, E., Wang, C.T., 1980. A representation of an IUH from geomorphology. Water Resour. Res. 16 (5), 855–862.

Haan, C.T., B.T. Barfield and J.C. Hayes (1993). Design hydrology and sedimentology for small catchments. Academic Press, Inc. San Diego, USA, 588pp.

Hall, M.J., A.F. Zaki and M.M.A. Shanin (2001). Regional analysis using the geomorphoclimatic instantaneous unit hydrograph. Hydrology and Earth system science, Vol.5, No.1, pp.93-102.

Herschy, R.W. (1996). Streamflow measurement. Second edition, Chapman & Hall, 518 pp.

Herschy, R.W. ed. (1998).Hydrometry. Principles and practices. Second edition, Wiley and Sons, 376 pp.

Institute of Hydrology (IOH) (1980). Low Flow Studies. Research Report No.3

Langbein, W. B. and K.T. Iseri (1995). Manual of Hydrology: Part 1. General Surface-Water Techniques. Geological survey water-supply paper 1541-Methods and practices of the Geological Survey HTML version http://water.usgs.gov/wsc.

Linsley (jr), R.K., M.A. Kohler and J.L.H. Paulhus (1988). Hydrology for engineers. McGraw-Hill. 492 pp.

Linsley (jr), R.K., J.B. Franzini, D.L. Freyberg en G. Tchobanoglous (1992). Water-resources engineering. Fourth edition, McGraw-Hill, 841 pp.

Maidment, D.R., (editor) (1993). Handbook of hydrology. McGraw-Hill Inc.

Musy, A. (1998). Hydrologie appliquée, Cours polycopié d'hydrologie générale, Lausanne, Suisse.

Musy, A. (2001). e-drologie. Ecole Polytechnique Fédérale, Lausanne, Suisse.

Nandakamur, N. and R.G. Mein (1997). Uncertainty in rainfall-runoff model simulations and the implications for predicting the hydrologic effects of land-use change. Journal of hydrology, Vol.192, pp.211-232.

Nicholson, S.E., J. Kim, M.B. Ba and A.R. Lare (1997) The Mean Surface Water Balance over Africa and Its Interannual Variability. Journal of Climate, 10, pp.2981-3002

Rinaldo, A., A. Marani, and R. Rigon (1991). Geomorphological dispersion. Water Resources Research, Vol.27, No.4, pp.513-525.

Rodríguez-Iturbe, I., Valdés, J.B., 1979. The geomorphologic structure of hydrologic response. Water Resour. Res. 15 (6), 1409–1420.

Shaw, E. (1994). Hydrology in practice. Third edition, Chapman & Hall, 569 pp.

Snell, J.D., Sivapalan, M., 1994. On geomorphological dispersion in natural catchments and the geomorphological unit hydrograph. Water Resour. Res. 30 (7), 2311–2323.

Newson, M. (1994). Hydrology and the river environment. Clarendon Press, Oxford, 221 pp.

van den Akker (1996). Hydrologie. Collegedictaat TUDelft, 200pp. (in Dutch)

van Rijn, L.C. (1990). Principles of fluid flow and surface waves in rivers, estuaries, seas and oceans. Aqua publications, Amsterdam, 335 pp. + Appendici.
Vicaire e-learning: http://hydram.epfl.ch/VICAIRE
Viessman (jr), W. en G.L. Lewis (1996). Introduction to hydrology. Fourth edition, Harper Collins College Publishers, 760 pp.
Wanielista, M. (1990). Hydrology and water quantity control. John Wiley & Sons, 565 pp.
Ward, A.D. en W.J. Elliot (eds.) (1995). Environmental hydrology. CRC press, 462 pp.
Ward, R.C. en M. Robinson (1990). Principles of hydrology. Fourth edition, McGraw-Hill, 365 pp.
White, A.B., P.Kumar, P.M. Saco, B.L. Rhoads and B.C. Yen (2004). Hydrodynamic and geomorphologic dispersion: scale effects in the Illinois River basin. Journal of Hydrology, Vol.288, pp.237-257.

Chapter 2:
New data sources

Sources	
Winsemius, H.C. Manual hydrological modelling course reader TUDelft	
SOO/STRC WRF EMS web page (http://strc.comet.ucar.edu/wrf/index.htm).	§2.7

Objective: After this day, the participants are familiar with data sources, both global and regional, and both spatial and in situ, both static and dynamic, which are available free of charge.

Preface: during our course we stress the "end products" and their applications and limitations of readily available data sources. In the following chapter we give basic and advanced background information of these data and focus on the way they are collected and processed.

2.1 INTRODUCTION

The availability of data in general is of extreme importance when one wants to tackle a water related issue or problem. Without data, no model or study can be verified and hypotheses will remain unproven. The continuous collection of data to support research and engineering questions has therefore been a major task in many countries.

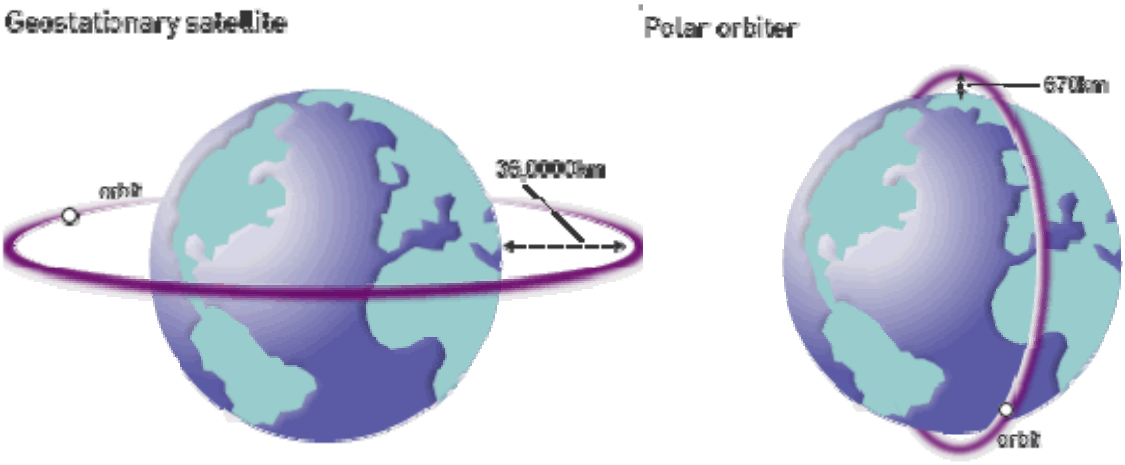


Figure 2.1. A geostationary satellite (left) orbits the earth along the equator on a very high altitude with the same rotation velocity as the earth itself. These satellites can therefore photograph the Earth in a very high frequency but with relatively low spatial resolution. These photos are often used by meteorologists. A polar orbiting satellite (right) orbits the earth over the Poles. It flies much lower than a geostationary satellite, which results in much lower temporal resolution but in a higher spatial resolution.

Unfortunately, many data collection programs all over the world have been abandoned due to socio-economic or political reasons, which resulted in both spatial and temporal data gaps in valuable data time series. Fortunately, many remote sensors were also developed in the last decades and people started to realize that, although many of these sensors were not developed for hydrological purposes, one could actually use them for this. For instance, a geostationary satellite, originally developed to study the movement of clouds and the land surface in high temporal resolution, could be used to make an estimate of rainfall, by calculating the temperature of cloud tops (also referred to as 'cold cloud duration'. Other variables that are of interest for hydrologists include elevation, or elevation-related ones such as slopes and hill slope orientation. Vegetation indices are also very popular amongst earth scientists.

This chapter will deal with some of the available spatial data sources, based on remote sensing. For point data sources, we refer to Appendix .1. We discern in two types of data: static (assumed not to change over time) and dynamic (changing over time) data. Some variables change fast over time, especially fluxes such as rainfall and radiation. Some go very slow, especially states such as vegetation greenness. In this chapter, we will start with the static data sources and then move from slowly changing to fast changing data. Finally, an open-source available weather model is presented, that can be ordered and used by any scientist in the world for weather hind- and forecasting.

2.2 ELEVATION

Most hydrological models start with the analysis of elevation data. Elevation is the prime forcing of lateral flow under gravitational forces and, therefore, determines in which direction water flows, what its potential acceleration due to gravitational potential is, and where possible 'pits' are located. Most hydrologists start with elevation data in order to find the boundaries of a catchment (i.e. find out which points in a certain area theoretically contribute to flow in a catchment).

2.2.1 Acquiring elevation

Is this part necessary? I have the impression it is of topic? Propose to delete.
In small catchments, the most obvious way of collecting information on elevation is with a measurement instrument such as a spirit level in combination with a staff gauge or a total station. A more advanced and quick method is by using satellite positioning, such as Global Positioning System (GPS) or the more enhanced Differential GPS (DGPS). With GPS, one's position may be found by triangulating the distance between the GPS receiver at the ground and several polar orbiting GPS satellites that are contacted through this device. The geographical location and elevation of the observer can thus be found when a minimum of 4 satellites is available. A time dependent error (Selective Availability) has been included for a long time for military purposes and resulted in an error up to 100m. Since 2000, this time-error in the satellite signal has been removed, resulting in an error of max 10. But it can be switch on again easily. With DGPS, a

base station is set up at a known location and elevation within the neighbourhood of the area of investigation that keeps track of the time-variable error and corrects all measurements taken by the receiver that is re-positioned continuously. Points collected can be interpolated to a rectangular grid using a geo-statistical interpolation method and can then be used to derive relevant information for catchment modelling.
A more generic approach that can be used on far larger scales is the use of laser altimetry on board of a satellite. This is a remote sensing approach by which the entire earth's elevation may be observed within a relatively short mission. A result of such a mission is for instance the GTOPO30 database that gives elevation at a resolution of approximately 1 km. For hydrological applications, NASA has corrected some problems that occur especially in relatively flat areas, so as to make sure that the boundaries of large river basins are realistically found when a basin boundary analysis is performed. The result is the HYDRO1k database.
The most recent laser altimetry mission performed is the Shuttle Radar Topography Mission (SRTM), covering the globe with a horizontal resolution of about 90 meters and vertically 1 meter. The result of both interpolated grids from manually gathered elevation points or a laser altimetry mission is often called a 'Digital Elevation Model' (DEM) or 'Digital Terrain Model' (DTM).

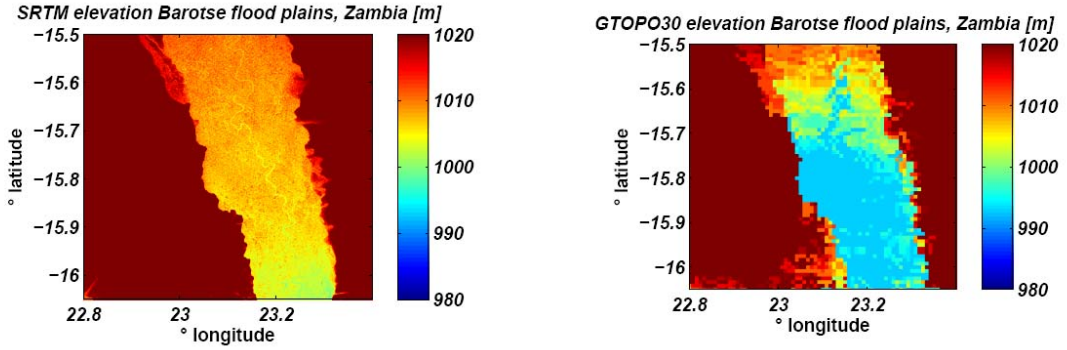


Figure 2.2. SRTM and GTOPO30 compared over a flood plain area in Zambia. Clearly the SRTM data shows more detail than GTOPO30. The Zambezi river is visible in SRTM but not in GTOPO30

2.2.2. Drain directions

As mentioned above, the theoretical lateral flow direction of water can be determined from elevation data. Mostly such a network is referred to as a Local Drain Direction network (LDD). Since most DEMs are of a 2-dimensional rectangular nature, flow directions may be found by looking at a window of 3×3 pixels and determine from the centre pixel, which of the neighbouring pixels is the lowest (has the steepest gradient). This is called the D8 method and has been introduced by O'Callaghan and Mark (1984). Problems, however, may occur when a flat area or a depression is found in the DEM. This can either occur due to the fact that the DEM simply does not show the detail found in the real world, or because the depression is truly a pond or wetland. Usually these areas fill up in due time and flow paths will come in use downstream from them in reality. These locations in the DEM are mostly referred to as pits and must be

removed to obtain an accurate LDD network. This method is available in many Geographical Information System (GIS) packages, for example IDRISI (Clarklabs), ArcGIS (ESRI) and the freely available GIS and dynamic environmental modelling package PCRaster (University Utrecht, http://pcraster.geo.uu.nl).

2.2.3 Catchment derivation

When flow directions are known, the upstream area from a certain point in a headwater can be determined. To find out where the larger streams are located, a Strahler river order network (Strahler, 1964) can be derived from the LDD network. A Strahler order network shows the smallest streams as order 1. Where 2 order 1 streams converge, the downstream resulting stream is of order 2 (see figure 1.11)When 2 second order streams merge, the downstream stream is of order 3, and so on. Mostly we are interested in the upstream area of a certain gauging location. A gauge is usually placed in a higher order stream because it has a large upstream area and thus gives an integrated response to the rainfall in a large area. GIS programs can track from a certain point which pixels in a DEM should contribute to the flow in that certain point, given the flow directions from this DEM.

2.2.4 Slope

Slopes for each cell of a DEM is mostly retrieved from a 3 × 3 cell window, using the elevation of the 8 neighbours of each cell. Slopes can for instance be used when one wants to route calculated discharges over a drainage network, which is often done as a post-processing of hydrological models. The hydrological model gives a runoff estimate per pixel or subcatchment. This runoff estimate is then assumed to flow into a stream instantaneously (or at least within the considered time step). The slopes are then used to compute how the discharge is transported over the river network. It influences the stream velocity.

2.2.5 Aspect

The aspect is the orientation of the slope, mostly with respect to a north-south line. This aspect is sometimes used to compute the influence of hillslope and orientation on energy availability in a certain grid cell (e.g. a south oriented slope in South Africa receives less solar radiation than a North oriented slope due to the sun's geometry with respect to the earth). This influences potential and actual evaporation. For large scale models this influence is usually neglected, especially when the pixel size is so large that hillslopes cannot be distinguished any more.

2.3 LAND COVER

The land cover is a typically slowly changing variable. In most cases, the land cover type is determined through a time series analysis of spectral satellite images. The change over time of the spectral properties of the land surface may give an indication of the land cover type. For instance: if a certain area in Southern Africa (one rainy season) looks like bare soil for a long period, then turns greener and finally quite suddenly becomes bare soil again, one may assume that this is an agricultural area, where crops grow during a relatively short period in the year, and are rapidly harvested in the end of the growing season. When these spectral time series analyses are done within different time frames, one may identify land cover changes. Also land cover is often used to estimate parameters of hydrological models such as root depth or canopy capacity for interception of rainfall.
What is often used for land cover characterization is an unsupervised or supervised classification on a time series of vegetation indices (see below) or other spectral properties. (Un)supervised classification is a GIS operation available in lots of popular GIS packages (e.g. ArcMap, IDRISI, ERDAS Imagine). See for more information on classification methods for instance Chapter 3 in Lillesand et al. (2004).

One of the crucial parts in land cover mapping is ground truthing. An unsupervised classification for instance, discerns several more or less similarly behaving land surfaces over time, but what these land surfaces represent should be checked on the ground.

Pre-processed land cover maps based on satellite information, can be found on the Global Land Cover Characterization (GLCC) webpage (http://edc2.usgs.gov/glcc/glcc.php). You'll notice that they provide several land cover characterizations, some having much more different land cover types than others. Beware that even these should be ground-truthed before use.

2.4 VEGETATION INDICES

Vegetation plays an important role in hydrological and ecological processes. Water balance studies show clearly that in African river basin especially, only little of the rainfall becomes runoff. It simply means that the rest of the rainfall will eventually evaporate and this is for a large part due to the presence of vegetation. When a plant or tree senses that there the conditions to grow are available (i.e. it is unstressed), it will open its stomata and start transpiring water, which in turn stimulates carbon assimilation and thus plant growth. Stress factors, that thus cause a plant to close its stomata, or even shed its leaves are for instance limited availability of solar radiation, non-optimal temperature conditions (too hot or too cold), but one of the main things is limited availability of water. Monitoring of the vegetation conditions is therefore a measure of monitoring drought. For instance, a comparison of a remotely sensed vegetation image of a certain area in the end of a rainy season, with a long term average of many vegetation images of that same season can give an indication of whether this specific season has been a wet or a dry

one. If one does this over an agricultural area, one can even estimate whether crop failure has occurred. One should of course do this type of analysis with care, because if crop species are changed over time the 'greenness' of these other crops may also be different (e.g. maize may look a lot greener than wheat, although it still gives a good yield)

The past decades a number of earth observation systems have been launched that have the ability to observe vegetation states. Some of the best known are the LandSAT mission and the Earth Observation System (EOS), which explores the use of the MODIS sensors on board of the Aqua and Terra polar orbiting satellites. Both are NASA missions. However, the European Space Agency (ESA) also works on earth observation, i.e. the EnviSAT program. Below, two common vegetation products are described.

2.4.1 Normalized Difference Vegetation Index

Normalized Difference Vegetation Index (NDVI) is a mostly remotely sensed retrieved index of the 'greenness' of the observed surface. It is based on the reflective properties of canopy. Active chlorophyl is typically capable of absorbing visual wavelengths, especially wavelengths close to red light, while reflecting near infrared wavelengths. These wavelengths are often separately measured on board of remote sensors (E.g. MODIS, SPOT, LandSAT) and NDVI can therefore be determined with a wide number of sensors. The NDVI is determined as follows:

$$I_{ndv} = \frac{\gamma_{ir} - \gamma_r}{\gamma_{ir} + \gamma_r} \tag{2.1}$$

Where I_{ndv} [-] is the NDVI, γ_{ir} is the infrared spectral band of the sensor [L] and γ_r is the red spectral band of the sensor [L]. The theoretical range of NDVI is between -1 and 1, however, a normal surface typically has an NDVI between 0.1 and 0.8. Water has different reflective properties and has values close to or below 0.

Example:

A very nice application of NDVI for Africa has been built by USAID for the FEWS project web site. Surf for instance to http://www.fews.net/ and click on 'maps, data and remote-sensing → NDVI. You are directed to a page where you can select a certain region in Africa. From here, you can select comparisons of the current situation with long term or short term averages of the same period (i.e. anomalies, see Figure 2.3), which can help you to address food security in a spatial manner.

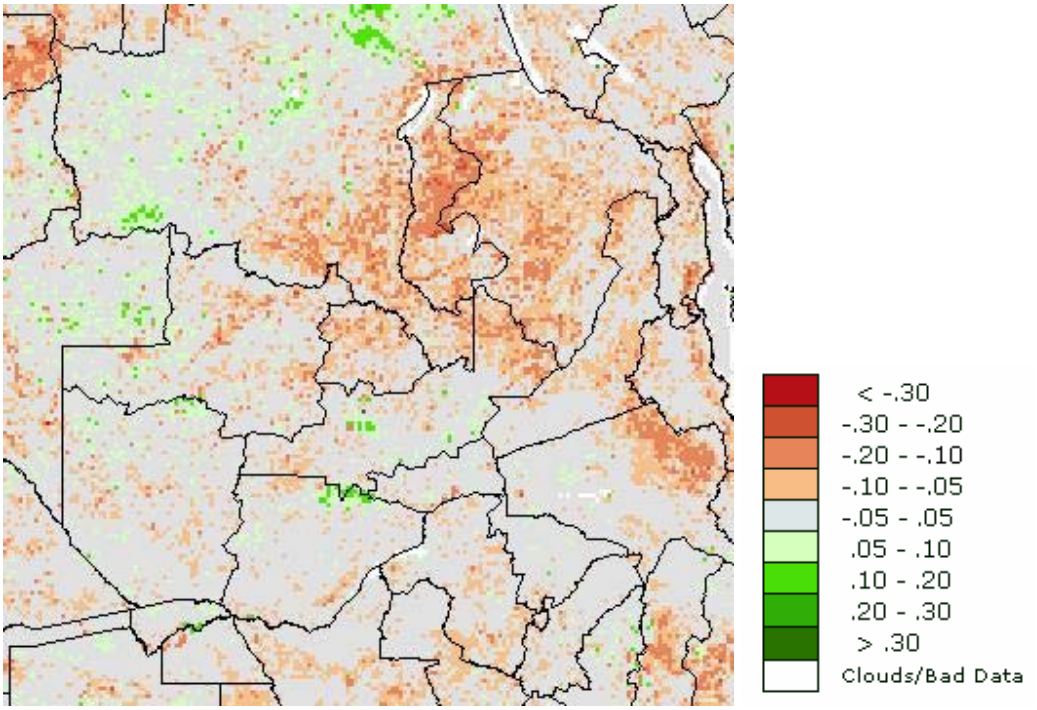


Figure 2.3. An example of NDVI anomaly for Zambia in the second decade of November 2007

NDVI images are often used to empirically derive the Leaf Area Index (LAI). LAI represents the total leaf surface per unit of ground surface. Basically, a large LAI means a lot of leaves and thus large potential for evaporation, if the conditions for it are good. A tropical rainforest for instance normally has a high LAI (>5 m^2/m^2). The problem with LAI derived from NDVI values is that

you cannot observe it directly from space. A canopy may look very green from above given the NDVI, but a grass surface may have the same NDVI as a forest surface, while there LAI is considerably different. Empirical relations between NDVI and LAI should therefore always be treated with care and should never be generalized. Ready-to-use global LAI products are therefore often not reliable. One can better use NDVI and use a NDVI-LAI relation based on expert knowledge of the study area.

2.5 SOLAR AND LONG WAVE RADIATION

Eq. 2.2 also clearly shows, that net radiation is extremely important for evaporation. In most applications, evaporation is considered as a fraction of total incoming radiation. The rest of the net radiation is used for heating up the air near the land surface (sensible heat flux) or heating up of the soil (ground flux). Considering this, it must therefore be very important to know the incoming radiation accurately when it comes to evaporation estimates. An assessment of evaporation can help to estimate or even forecast crop yields (given that the crops have enough water). The term R_n in eq. 2.2 consists of several parts. The general equation is given in eq. 2.2.

$$R_n = R_{s,in} - R_{s,out} + R_{l,in} - R_{l,out} = (1-\alpha)R_{s,in} + R_{l,in} - R_{l,out} \quad (2.2)$$

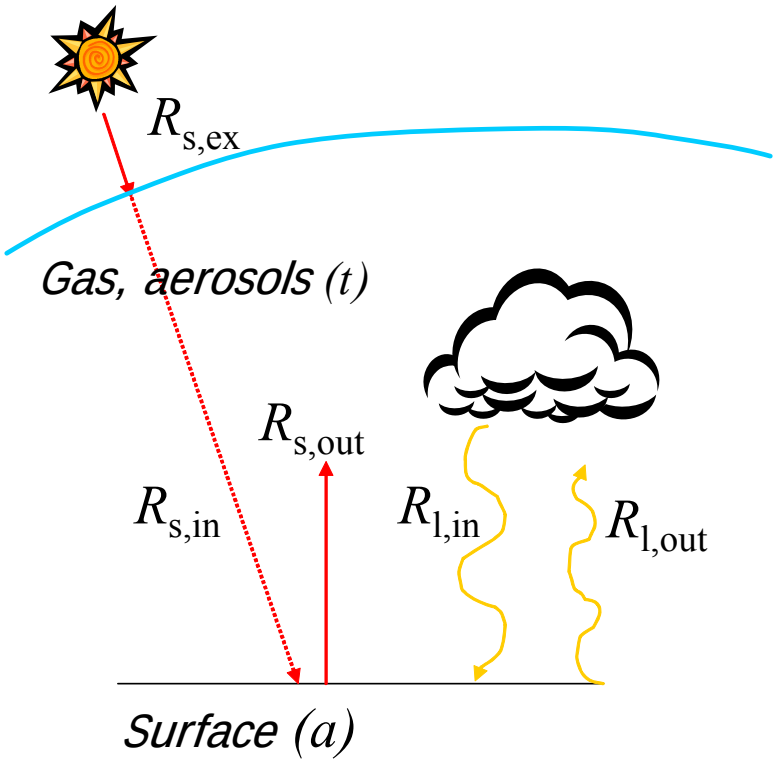


Figure 2.4. A schematic of the energy fluxes on the land.

$R_{s,in}$ is the amount of incoming solar (or short-wave) radiation at the surface. The sun's beams come in at the top of the atmosphere. When the sun is shining with an angle, completely perpendicular to the atmosphere's surface, the amount of energy coming in at the top of the atmosphere is about 1367 W m^{-2}. This is called the solar constant. Within the atmosphere, before the sun's beams hit the Earth's surface, this radiation is scattered, reflected and attenuated in the atmosphere, due to the appearance of clouds, aerosols and of course, the presence of air mass. The effective decrease of solar radiation is often put into one factor, the so-called transmissivity.

If you look at the Earth's surface from above, you may notice that over a forest or open water, it looks very dark. Over a sandy area, it looks relatively bright. An ice surface is not so common in Africa, but is even brighter. You might even consider wearing sunglasses if you have to look at it a long time. What you basically see is that open water and forest, do not reflect much of the incoming solar energy, while sand or ice reflect a lot. This reflected energy is $R_{s,out}$ in eq. 2.3 and is a fraction of the incoming solar radiation. This fraction or reflectivity is called the albedo (α [-]).

Finally, $R_{l,in} - R_{l,out}$ is the net long-wave radiation. $R_{l,in}$ (incoming long waves) is radiation emitted by particles in the air and the air itself. This flux increases with the amount of particles (i.e. when it is cloudy or when there are a lot of aerosols in the air there is a lot of incoming long wave radiation) and with temperature raised to the power 4 of these particles (the so-called Stefan-Boltzmann law). The outgoing long wave radiation is emission of radiation from the Earth's surface and also increases with its absolute temperature raised to the power 4.

These energy fluxes change very fast over time. If a cloud moves over a surface, the presence of the cloud can cause a sudden drop of incoming solar radiation, while emitted long wave radiation may go up. Therefore, geostationary satellites are used to estimate this variability in combination with information from an operational weather model. The satellite observes how clouds move, the weather model shows the 'thickness' of the clouds. Combining the two in a model can give very accurate estimates of solar and long-wave incoming radiation.

The satellite that is really perfect for Africa to do this job is Meteosat Second Generation (MSG), which orbits the earth at subsatellite point 0 longitude, 0 latitude. The Land Surface Analysis Satellite Application Facility (LSA SAF) in Portugal has the job to process the raw satellite data into end-user products. They now make incoming short-wave, long-wave products on a half-hourly basis. Albedo is estimated on daily basis. These products can be downloaded for free (http://landsaf.meteo.pt/) and used for water balance models or estimates of irrigation efficiency.

NOTE: watch out if you want to use these data in hilly irrigated areas. As mentioned before, how much radiation is received by a hill slope, is also dependent on its slope and slope orientation (also called aspect). For instance, in the Netherlands, we get a lot of sunshine from the South (that's why everybody wants their garden faced Southwards in the Netherlands). In South-Africa, this is exactly vice versa. The angle that the Sun makes with respect to the Northern direction is called the solar azimuth angle, a_s. The local surface also has a certain slope β and aspect (or

surface azimuth angle) a. All these angles together give an effective angle that the beams make with the surface, here given as θ. If we multiply the total incoming solar radiation at the top of the atmosphere with the transmissivity τ and the cosine of θ, we get the incoming surface solar radiation (see Figure 2.5. for a schematic of the solar radiation on an inclined surface). Some geometry gives (Chrysoulakis et al., 2004)

$$\cos\theta = \cos\beta\cos z_s + \sin\beta\sin z_s\cos\left(a - a_s\right)$$
$$R_{s,in} = \tau G_s\cos\theta \tag{2.4}$$

This equation can be applied if we know the Sun's geometry (dependent on date, local time, latitude and longitude) and the elevation and the amount of energy loss due to the presence of air, clouds and aerosols,

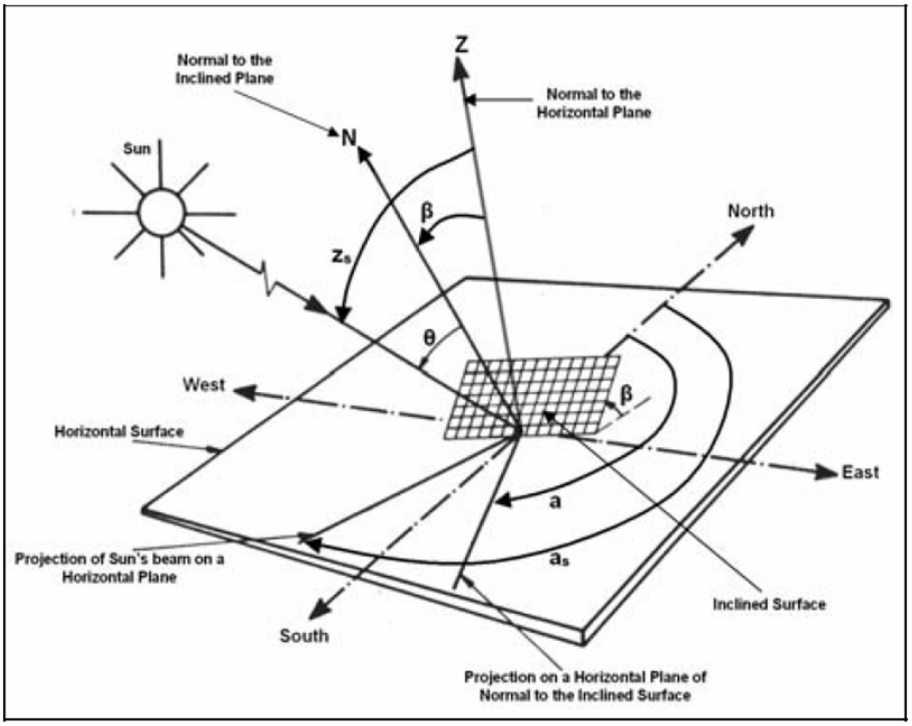


Figure 2.5. The effect of solar geometry and slopes (β) and aspect (a) on the incoming short wave radiation (Chrysoulakis et al., 2004).

2.6 RAINFALL

Rainfall is the prime input to hydrological studies and models. Not only is the spatial and temporal heterogeneity of rainfall very large (especially in tropical areas, where most rainfall has a convective character), it also has a large effect on partitioning of the rainfall over many processes. For instance: a very local rainfall event with high intensity will result in Horton overland flow inclusive erosion. Local rain periods with moderate intensities will result in a large local increase of soil wetness and possibly saturation excess, which might cause a larger amount of it to become runoff. If the same amount of rainfall falls evenly distributed over a large area, the fraction of the area that becomes saturated may be much smaller, resulting in a much lower runoff coefficient for that specific event. This means that not only the spatio-temporally accumulated rainfall is important. Its spatio-temporal distribution is also important.

The maintaining of large precipitation networks is time consuming and expensive. In many remote areas in the world, the meteo-networks are deteriorating and many time series contain a large number of gaps. Remote sensing offers extremely interesting opportunities to not only fill the gaps, but also extend the point measurements into spatio-temporal estimates of rainfall. The first tropical near-real time rainfall mission is the Tropical Rainfall Measuring Mission (TRMM), launched by NASA. This mission is polar orbiting (see Figure 2.1) and covers the whole globe each 3-4 days. The 3 major instruments on board of TRMM satellites are:

- a rain radar, to estimate 3-dimenional cloud characteristics (drop size and distribution);
- a micro-wave imager, to measure surface and cloud temperatures (if the land surface is cold, it is likely that rain has been falling); and
- visible and infrared scanner, to measure cloud-top temperatures (cold cloud duration);

From all 3 instruments a separate estimate of rainfall is modelled, which is interpolated over time by using geostationary satellites to track where the clouds are moving to. All 3 rainfall estimates contain errors. For instance, the cold cloud duration method gives an empirical relation between cloud top temperature and rainfall. If a rainfall event is convective, the cloud top will contain a lot of ice particles and will generally be very cold. A rain event over the ocean is usually not of convective nature and therefore contains fewer ice particles. Advective rainfall (for instance over the ocean) is therefore usually underestimated, while convective events might be overestimated in terms of rainfall intensity. TRMM end-user products therefore do not purely consist of evenly weighted satellite estimates. Operational ground estimates are merged into the rainfall estimates, by appointing weights to the 3 separate products: the estimate that approaches the rainfall station estimate most will get the highest weight. If the nearest rainfall ground station is located very far from where the rainfall estimate is made, the weights become more even and results will become less reliable. It is, however, very hard to make a reliable error estimation. Resulting products are provided ranging from 3-hourly to monthly products on a 25 * 25 km^2 grid. (http://disc.gsfc.nasa.gov/).

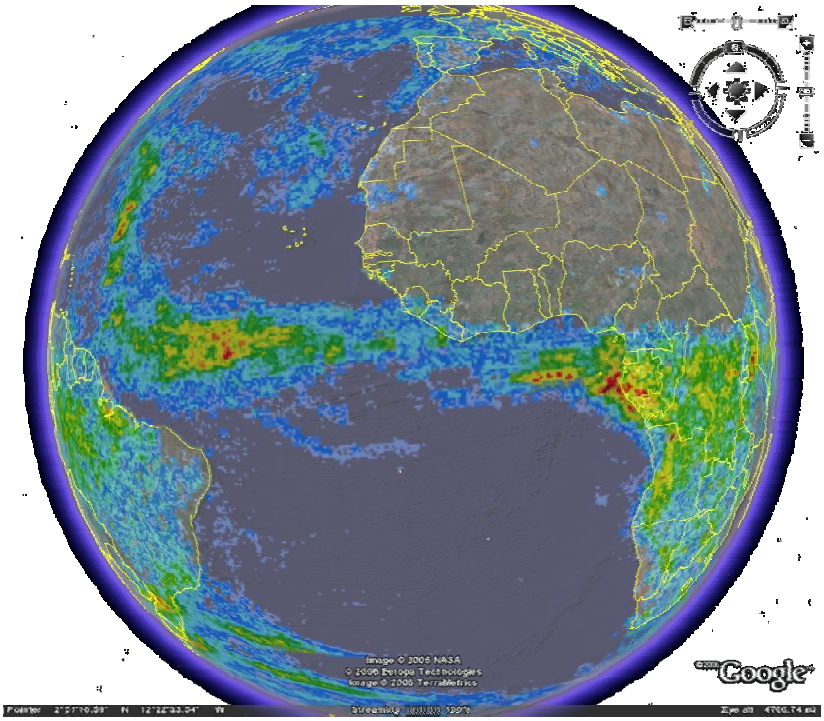


Figure 2.6. Weekly accumulated rainfall estimate from TRMM (background: Google Earth).

Another rainfall estimate that is gaining popularity amongst African researchers is FEWS RFE 2.0 (Herman et al., 1997) generated by the Climate Prediction Center (CPC, NOAA) for the Famine Early Warning System (FEWS) for Africa (see also http://www.fews.net for reports and hazards, and http://www.cpc.ncep.noaa.gov/products/fews/ for end-user data products). Here also 3 satellite-based rainfall estimates are made, one from Special Sensor Microwave Imager (SSM/I), one from geostationary satellite-based cold cloud duration or the 'GOES Precipitation Index' (GPI) and one from AMSU-B, based on measuring brightness temperatures differences due to ice concentration. The products are again merged on a daily basis by determining weights which are dependent on the 'ground-truth', a measurement ground network (Xie and Arkin, 1996).

I would like to emphasize, that, although remote sensing offers opportunities to estimate rainfall, the accuracy of these estimates is all based on weighting procedures that use ground station data as input. Therefore, it will remain extremely important to measure as much as possible rainfall on the ground, in order to have anchor points to tie your satellite based estimates on.

2.7. WRF EMS: AND OPEN-SOURCE WEATHER MODEL

Numerical Weather Prediction (NWP) is a very powerful way to obtain near-future information on a water system. It provides estimates of temperature, wind speed, humidity and of course rainfall. Global weather models exist, that estimate the circulation patterns on a coarse scale (often ~100 km) but over the whole globe (for instance the NCAR and ECMWF models). The problem with these models is that the most interesting output for a hydrologist or water resources manager, rainfall, is usually highly inaccurate. This is mainly due to the fact that rainfall is a relatively small flux that represents a highly non-linear response to a number of often very localized factors, such as orographic uplift, convection and advection. These are all processes that can only be accurately described on a small scale, preserving the needed detail of process description. On the other hand, these large scale models are generally quite good in predicting the large scale circulation patterns. In that case, a regional numerical weather model is created that gets its boundary conditions from a large-scale model, but has a much finer resolution (for instance ~10 km) than the large-scale model. This way of modelling is called 'nesting'. In Figure 2.7, you can see an example of nested models in Western Africa. Model domain 003 has been specifically designed for the Volta river basin.

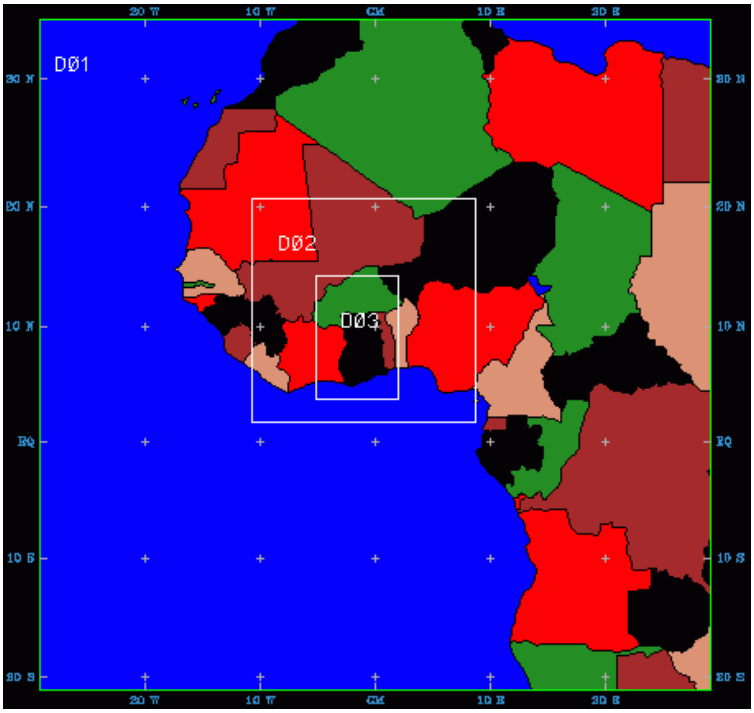


Figure 2.7. An example of nested NWP's. Domain 001 has a spatial resolution of 81x81 km^2 and gets its boundary conditions from the global NCAR model, domain 002 and 003 have a spatial resolution of resp. 27x27 km^2 and 9x9 km^2 and get their boundary conditions from domain 001 and 002 respectively (taken from http://www.glowa-volta.de/research/results_phase1/resultp1_a1.htm)

The effect of nesting is that highly non-linear processes such as cloud formation and rainfall can suddenly be addressed on a much finer scale which will increase the accuracy. Note that where available, it is important to feed the model with as much in-situ data as possible to update the states of the model. Measurements of for instance air pressure and humidity at different levels in the air column can help to adjust the states of the model to a more realistic one, which also increases the accuracy.

This all seems very complicated, but the use of such models in real time for your own use is readily accessible for African hydrologists. The National Weather Service's (NWS) SOO Science and Training Resource Center (STRC) has made a complete, full-physics, NWP package, that incorporates dynamical cores from both the National Center for Atmospheric Research (NCAR) Advanced Research WRF (ARW) and the National Center for Environmental Predictions' (NCEP) Non-hydrostatic Mesoscale Model (NMM-WRF) packages into a single end-to-end forecasting system. It is called Weather Research and Forecasting (WRF) Environmental Modeling System (EMS) and can be freely obtained on DVD by sending an e-mail to mr. Robert Rozumalski (rozumal@ucar.edu). The SOO/STRC WRF EMS is easy to run on most linux workstations; it should be possible for those with limited modeling experience to have the package installed and running in less than 1 hour. Of course you'll need some experience with the linux operating system. Recommended Linux distributions that you may use are OpenSuSe or Fedora (can be downloaded from the internet). If you use Ubuntu (also downloadable from the internet), which is really user friendly, your system settings need some alterations if you want to run this NWP (contact me for assistance at h.c.winsemius@tudelft.nl). The DVD comes with a nice manual that'll help you to set up the software and to make a model for your own domain.

What is important to note when you are interested in this are the following issues:

- For a nested model, you'll need a powerful work station (for instance holding 2 pentium 4 processors 3Ghz, and at least 1, but preferably 2 or 4 Gb memory).
- Before you even start thinking about setting up such a model. First make sure that you have a decent internet connection available. I was able to retrieve boundary conditions from the internet easily with my home-connection of about 400 kbyte/s but I'd make sure you can manage at least a 100 kbyte/s before you try to set up an NWP model (note of H.C. Winsemius, December 2007).
- The computation time of course depends on the number of raster cells your target area holds. The larger your target area, or the higher resolution you select, the more computation time it will take.
- Never assume that what the model predicts is the absolute truth. Numerical weather predictions are based on extremely non-linear processes, with an incredibly large amount of feedbacks (both positive and negative) which causes the uncertainty to grow rapidly with forecast lead time.
- Comparison of model results with ground data seems attractive but is not easy. You always have to question how representative your ground measurement is with respect to a NWP grid cell.
- If you are an employee of a meteorological department, always use the model together with your traditional way of weather forecasting and keep training your new colleagues

in using the traditional approach as well. In case models fail or are way off the truth, you'll always have a backup.

Bibliography

Chrysoulakis, N., Diamandakis, M. and Prastacos, P., 2004. GIS based estimation and mapping of local level daily irradiation on inclined surfaces, 7th AGILE conference on geographic information science.

Herman, A., Kumar, V.B., Arkin, P.A. and Kousky, J.V., 1997. Objectively determined 10-day African rainfall estimates created for Famine Early Warning Systems, Int. J. Remote Sensing, pp. 2147-2159.

Lillesand, T.M., Kiefer, R.W. and Chipman, J.W., 2004. Remote sensing and image interpretation. John Wiley \& Sons.

Monteith, J.L., 1981. Evaporation and surface temperature, Q. J. R. Meteorol. Soc., pp. 1-27.

O'Callaghan, J.F. and Mark, D.M., 1984. The Extraction of Drainage Networks From Digital Elevation Data, Comput. Vision, Graphics, Image Process., pp. 323-344.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass, Proc. Roy. Soc. London, pp. 120-145.

Sellers, P.J., Mintz, Y., Sud, Y.C. and Dalcher, A., 1986. A simple biosphere model (SiB) for use within general circulation models, J. Atmos. Sci, pp. 505-531.

Strahler, A.N., 1964. Handbook of Applied Hydrology. In: V.T. Chow (Editor). McGraw-Hill, New York, pp. 4-39 4-76.

Van den Hurk, B.J.J.M., Viterbo, P., Beljaars, A.C.M. and Betts, A.K., 2000. Offline validation of the ERA40 surface scheme, TechMemo 295. %(ECMWF, Reading, U.K.

Viterbo, P. and Beljaars, C., 1995. An improved land surface parameterization scheme in the ECMWF model and its validation, J. Climate, pp. 2716-2748.

Xie, P. and Arkin, P.A., 1996. Analyses of global monthly precipitation using gauge observations, satellite estimates and numerical model predictions, J. Clim., pp. 840-858.

Chapter 3
Statistical analysis in water resources assessment

Sources:
Statistical analysis, S. Mkhandi, University of Dar es Salaam §3.1- §3.8
Mamdouch, S. (2002). Hydrology and water resources of Africa, Water science and technology library 41, Kluwer Academic Publishers, Dordrecht, pp.256-258 §3.9
Johnson, C.A. (1999) Data infilling techniques: development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data Unpublished MSc thesis Page 51-63 §3.9

Recommended reading:
USGS: Statistics in water resources (free pdf)

Objective: All participants are, after this day, familiar with the basic statistical techniques and analyses that are used in hydrology (focus on use in Africa). Participants can perform frequency analysis calculation and are aware of (dis)advantages of the techniques. Lastly, the participants will have basic knowledge of data infilling techniques and will be able to apply techniques in their own working environment.

3.1. INTRODUCTION

Practically in dealing with different types of water related projects, the main questions which water resources managers try to answer may be summarized as follows:

- Does the water resources considered give the required amount of water for the project under consideration, eg water supply, irrigation, hydropower etc.
- What is the maximum probable event for which hydraulic structure has to be constructed, e.g., dams, bridges, drainage system, etc.
- What is the general behavior of both quantity and quality with respect to time (on continuous basis) in order to decide size and location of storages required and also treatment.

In order to come up with proper answers to the above questions, hydrologist apply statistical techniques to historical hydrological data to facilitate the understanding of the principles of hydrology as an important step in seeking optimum solutions to various engineering problems.

The fundamental assumptions in the statistical analyses include the following:

- Each observation is independent of previous and subsequent observations. This is reasonable for annual flood maxima.

- The data are free of measurement errors. This should be tested in all cases where the design of major structures is involved, particularly the possibility of systematic errors occurring where peak flows are well in excess of the capacity of the measuring structure.
- That the data are identically distributed i.e., that they can safely be assumed to come from a single parent population, which in turn implies a single type of rainfall producing meteorological phenomenon. This is clearly not the case in areas subject to infrequent tropical cyclones for example. Where this possibility exists, special treatment is required. Analysts should always be aware of the strong possibility that many of the apparent anomalies in statistical analyses arise from the mixture of different meteorological phenomena and different states of antecedent conditions that determine the magnitude of the runoff events.

3.2 BASIC STATISTICS

3.2.1 Introduction

Practicing hydrologists, water resources engineers and other specialists dealing with certain facets in hydrology apply statistics as a tool to analyse and interprent hydrological data to form a basis for making reliable decisions on water development projects. Hydrologists are involved in the collection of hydrological/climatic data in order to solve various problems of water resources engineering. Important data collected include the following:

- Precipitation
- Streamflow
- Evaporation
- Water quality
- Sediment, etc

The longer the period of record, the greater the value of the data. Historical collected hydrological data is used to serve as a base in the design of water related projects, whereby the data is used for projection of future conditions. All water projects have to be planned to meet the future needs.
Collected hydrological/climatic data can be treated as statistical variables and as such statistical techniques are applied to describe the characteristics of the recorded hydrological variables. In this section basic statistical applicable in hydrology are presented.

3.2.2 Definition of basic statistical terms

- *A population* is the whole collection of values under consideration. It may be finite or infinite.
- *A sample* is a set of observed values, more or less representative of the population from which it is drawn.
- *A variable (X)* is the characteristic of a sample, for example the depth of rainfall.
- *A variate (x)* is an individual observation or the value of any variable.

- *A discrete variable* can contain only a finite number of values (or as many values as there are whole numbers), for example the number of rainy days.
- *A continuous variable* can contain all values within a certain range, for example the depth of rainfall.

3.2.3 Descriptive statistics

While there are two main branches of statistics, i.e., descriptive and inferential, in these notes only descriptive statistics is presented. Descriptive statistics involves the organization, summarization, and description of data sets while inferential statistics, on the other hand, is the process of drawing conclusions about an entire population based only on the results obtained from a small sample of that population.
Descriptive statistics is used to describe the characteristics of a statistical distribution and the statistical parameters used for this purpose are defined below.

MEASURES OF LOCATION
The ***arithmetic mean*** is the average most frequently used. It is obtained by adding together all the variates, ∑x, and dividing by the total number of variates N. This is expressed by:

$$\bar{x} = \frac{\sum x}{N} \tag{3.1}$$

The ***median*** is the middle value of the variate which divides the frequencies in a distribution into two equal portions. The arithmetic mean is more commonly used than the median. For skew distributions, however the mean may be misleading. In such cases, the median will provide a better indication, because all variates greater or less than the median always occur half the time.

The ***mode*** is the variate which occurs most frequently. It corresponds to the peak of the frequency curve. For a grouped distribution, the modal class can be defined as the class with the greatest frequency.

MEASURES OF VARIABILITY
The ***mean deviation*** is defined as the mean of the absolute deviations of values from their mean, or:

$$m.d = \frac{\sum \left| x - \bar{x} \right|}{N} \tag{3.2}$$

The ***standard deviation*** is the measure of variability, or spread, which is most adaptable to statistical analysis. It is the square root of the mean-squared deviation of individual observations from their mean, or

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{3.3}$$

which represents the standard deviation of the population. An estimate of this parameter from the sample is denoted by *s* and computed by:

$$s = \sqrt{\frac{\sum \left(x - \bar{x}\right)^2}{N-1}} \tag{3.4}$$

The ***variance*** is the square of the standard deviation which is denoted by σ^2 for the population. The unbiased estimate of the sample variance is $\boldsymbol{s^2}$.

The ***coefficient of variation*** is a measure of spread of the sample in relative terms:

$$C_V = \frac{s}{\bar{x}} \tag{3.5}$$

Quartiles and ***percentiles*** may be considered as measures of spread about the median (the 50-percentile value). With the variates arranged in ascending order of magnitude, the lower quartile is the value at the first quarter of the data series (25 percentile). The upper quartile marks the beginning of the top quarter of the data (75-percentile).

3.2.4 Linear regression

Applying a linear regression to a data set is a very common way to determine the presence or absence of a correlation between the data points. Linear regression applies to bivariate data, for which two variables are related systematically such that one is a fairly constant multiple of the other. Linear regression, also known as the least-squares best-fit line, can be calculated for bivariate data that have a linear appearance when plotted in scatter plots. The best-fit line through these data is defined by the straight-line equation:

$$Y=a+bX \tag{3.6}$$

This equation represents the relationship between an independent variable *X* and dependent variable *Y*. Given a value for the independent variable, use of this equation allows prediction of the dependent variable. The statistical parameters of this equation are the intercept of the line, ***a*** and the slope of the line, ***b***. To calculate the slope of the line, the following equation is used:

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \tag{3.7}$$

Where *n* is the number of measurements.

Once the slope of the line is calculated, the intercept can be calculated using the equation

$$a = \bar{y} - b\bar{x} \tag{3.8}$$

Where $\bar{y}$ is the sample mean of the *y* values sampled, and $\bar{x}$ is the sample mean of the *x* values sampled.

The coefficient of determination r^2 (commonly reported as R^2) is the fraction or percentage of the total variation in the data that is accounted for by the regression equation. For example, a regression equation with a R^2 of 0.8 accounts for 0.8 or 80% of the variability in the data. The 20% that is not accounted for could be due to measurement error or factors not included in the equation. When regression equations are used as predictive tools, we often refer to them as *regression* or *empirical* models.

3.2.5 Correlation coefficient

The *correlation* between the two variables is defined as the level of association between the two variables. The correlation coefficient *r* can be calculated using the following equation:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}.\sqrt{n(\sum y^2) - (\sum y)^2}} \tag{3.9}$$

The correlation coefficient values range from -1.0 to + 1.0, and the following statements hold true:

- A value of r near + 1.0 indicates that, as one variable increases, the other variable increases in a similar trend, and a strong correlation exists.
- A value of r near -1.0 indicates that, as one variable increases, the other decreases, and a strong correlation between the two exists.
- If r is near 0, there is little correlation between the two variables.

3.3. FLOOD AND DROUGHT ANALYSIS

3.3.1 Introduction

Frequency analysis techniques are applied to carry out flood and drought analysis for the purpose of determining peak and low flow magnitudes required in the design of hydraulic structures affected by high and low flows, i.e., spillway, bridges, culverts, water intake, etc. Drought analysis in this particular case refers to analysis of low streamflow records and drought deficiency volumes. Proper determination of design values for the hydraulic structures is important for the safety, economy and proper functioning of such structures. The purpose of the hydrologic design is to estimate the maximum, average or minimum flows, which the structure is expected to handle. The estimate has to be made rather accurately in order that the project functions properly. For example in the design of a spillway, one has to bear in mind that the floodwater to be released through the spillway should

not create flooding downstream. Thus a balance has to be worked out between the economy, efficiency in regard to flood moderation and safety. On the other hand, the design of water intakes structure has to take into consideration the variation of the water level during low flow.
Several methods are available by which the estimate of design flow values can be made. Some of the methods are purely empirical and some are based on statistical analysis of the previous records.

3.3.2 Objective of frequency analysis

The objective of frequency analysis is to estimate a high or low flow corresponding to a specified return period of occurrence. This information is required in the design of hydraulic structures such as dams, bridges, culverts, flood control structures, water supply intakes, irrigation intakes, determination of reservoir storage capacity, etc.
The return period of occurrence (T), also known as recurrence interval, is an expression of the frequency of occurrence of a flow of a given magnitude. Return period is defined as the average time interval between the occurrences of an event equal to or greater than.
The relationship between the flood magnitude (Q_T) and its return period (T) can be estimated from the analysis of observed flood peaks, while the low flow magnitude (q_T) and its return period (T) can be estimated from the analysis of observed low flow values. For the analysis to be valid, the flood peaks and low flow values, under consideration, must be of random magnitude and mutually independent. Given the usual shortness in observed flood records, extrapolation of frequency functions is quite uncertain so that the larger the flood and the greater the hazard it represents, the less reliable is the estimate of the frequency with which the flood is likely to occur. The same applies for the extrapolation to determine the extreme low events.

3.3.3 Data for frequency analysis

The basic input in flood/low flow frequency analysis is a time series of flow data at a location or region of interest. Types of flow data mainly considered in the analysis of flood peaks and low flows are

- Annual Maximum (AMax) series
- Partial Duration (PD) series
- Annual Minimum (AMin) series
- Annual Maximum Drought Volumes (AM-DV) series

ANNUAL MAXIMUM SERIES
The annual maximum (AM) series consists of the peak flow of each year (Fig. 3.1, i.e. blue circles only), and is the most frequent used among the two common series. One of the aspects in favour of the AM series is the reasonable assumption that the data series is not serially correlated, i.e., successive values are independent provided that the 'hydrological year' or 'water year' is carefully chosen. This is an important pre-requisite for the subsequent statistical treatment of data. A disadvantage of AM series is that the second or third, etc, highest events in a particular year may be higher than the maximum event in another year and yet they are totally disregarded.

PARTIAL DURATION SERIES

The disadvantage of disregarding some of the significant high events in any particular year in AM series is remedied in the partial-duration (PD) series sampling method in which all events above a certain base magnitude (Fig. 3.1, i.e., PD series includes both red squares and blue circles) are included in the analysis. Furthermore, when only a limited period of record is available, the use of series over a threshold (POT series) has advantages of including a large number of flood peaks in the analysis. The base is usually selected low enough at least one event in each year is included. It is important that each event that is included in the partial-duration series must be separate and distinct.

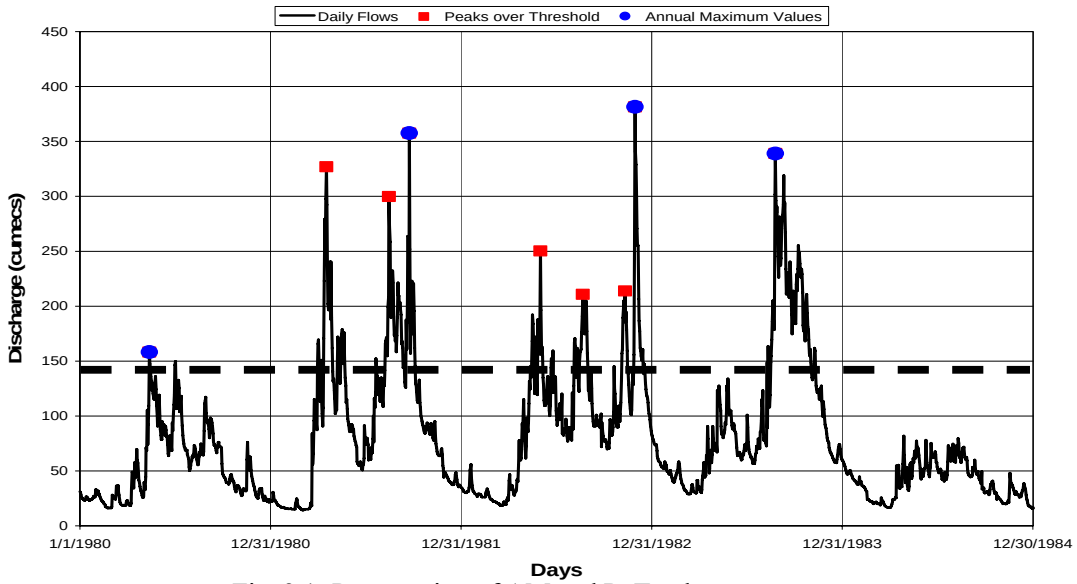


Fig. 3.1 Presentation of AM and PoT values

ANNUAL MINIMUM (AMIN) SERIES

The Annual Minimum flow series is formed by selecting the lowest flow occurring in each year of record. The set of observed annual minima at any gauging station is assumed to be a random statistical sample from the population of all possible annual minima at the site. The resulting series $q_1, q_2 \ldots q_N$ is then assumed to be a random sample of size N from a population in which the q values have a distribution (Figure 3.2). An estimate of such a distribution may be obtained from the sample. An assumption is made about the form of the population distribution function and the validity of the assumption is tested by applying the goodness of fit test. The annual minimum flow of required return period is then estimated from the data with the help of this distribution.

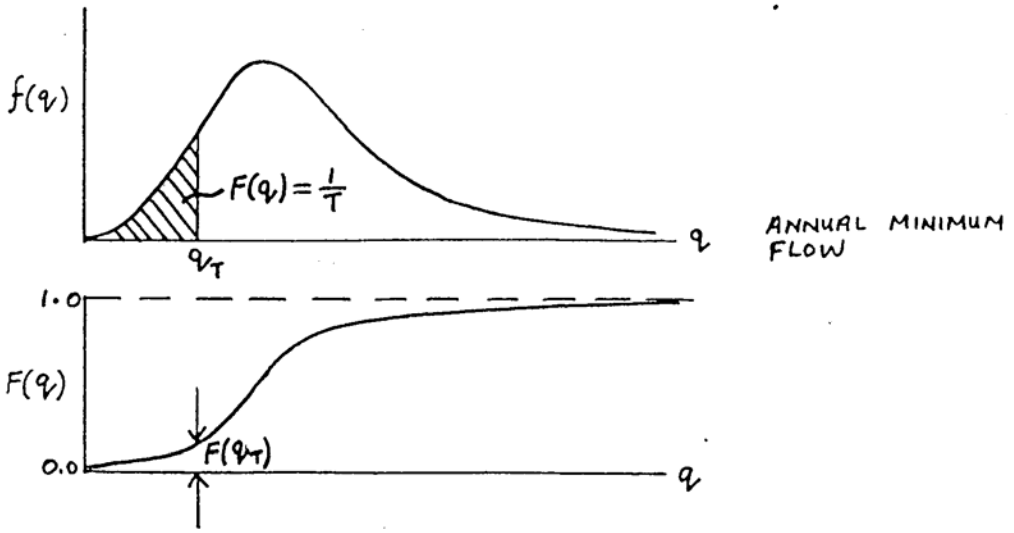


Figure 3.2 Distribution of Annual Minimum flows

Return period, T, in this context is now taken as the reciprocal of the non-exceedence probability. The value q_T corresponding to the design probability of failure 1/T is then obtained from the distribution. If the value of q_T is large in comparison to Q_D, (the demand flow), then the river can be considered to be able to supply the demand satisfactorily. On the other hand if q_T is less than, or of the same order of magnitude as, Q_D, the river alone without some form of regulation could not be considered satisfactory for supplying the demand.

Some indications as to the form of the probability distribution governing the annual minimum flow series can be obtained from the general characteristics of the sample.

- The annual minimum flow series is bounded below by some $q_0 > 0$, since negative flows cannot occur.
- The values of sample skewness of annual minima.

Many authors recommend the use of the Extreme Value Type III distribution, because it has a lower bound and because of the appeal of the extreme value theory in the context of annual minima.

Another possibility is the Extreme Value Type I distribution which has a skewness of 1.14, which is in the middle of the observed range quoted above. The EV1 has no lower bound but the probability of values smaller than μ-2σ occurring is less than 0.1%. Because of this, the lack of a lower bound may not be a serious deterrent to the use of the EV1 distributions.

The lognormal distribution, particularly the three parameter version with a lower bound $q_{0,}$ is another candidate distribution. The three parameter one is quite flexible and can fit a variety of distribution shapes quite well.

ANNUAL MAXIMUM DROUGHT VOLUMES (AM-DV) SERIES

The Annual Maximum drought volumes series is formed by determining the maximum deficiency of river flow to meet the specified water demand in each year of record at a given location. The set of observed annual maximum deficiency volumes at any gauging station is assumed to be a random statistical sample from the population of all possible annual maximum deficiencies at a given site.

In the sketch below (Figure 3.3), suppose Q_D (Demand flow) = $^2/_3$ x (mean flow) and since the river flows can dip close to q_o every few years it is seen that there are volumes of deficiency V_1, V_2, V_3 which must be obtained by storage if Q_D is to be met at all times.

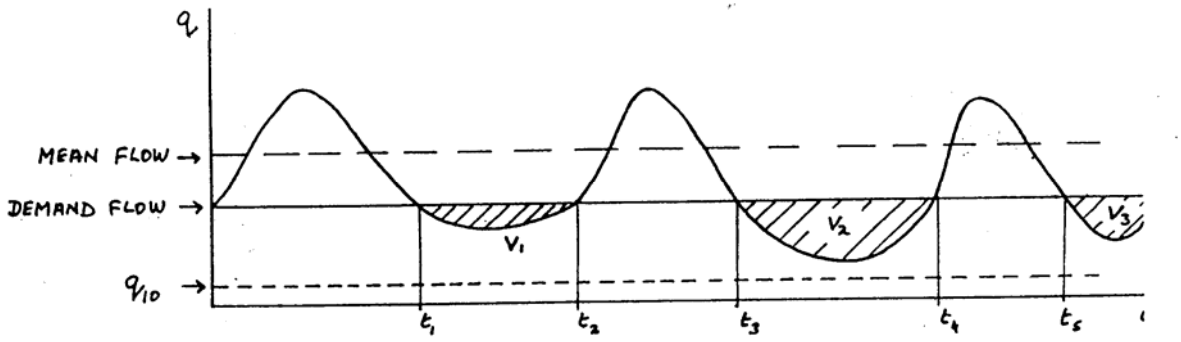


Figure 3.3 Determination of volumes of deficiency

3.4 FREQUENCY ESTIMATION

There are two basic approaches for estimating frequency curves. These are graphical and analytical.

3.4.1 Graphical approach

Graphically, frequencies are evaluated simply by arranging observed values in the order of magnitude and considering that a smooth curve suggested by the array of values is representative of future probabilities.

The plotting position formulae are applied to compute the probability plotting positions for observed events. In other words, plotting position formulae express the relationship between order number of ordered statistics and corresponding average frequency value of that statistic over a large number of samples. Probability plotting of hydrological data requires that individual observations or data points be independent of each other and that the sample data be representative of the population.

Probability plotting positions are used to produce graphical display of annual maximum flood series and serve as estimates of the probability of exceedance of those values. Probability plots allow a visual examination of the adequacy of the fit provided by alternative parametric flood frequency models. They also provide a non-parametric means of forming an estimate of the data's probability distribution. An example of a probability plot is shown in Figure 3.4.

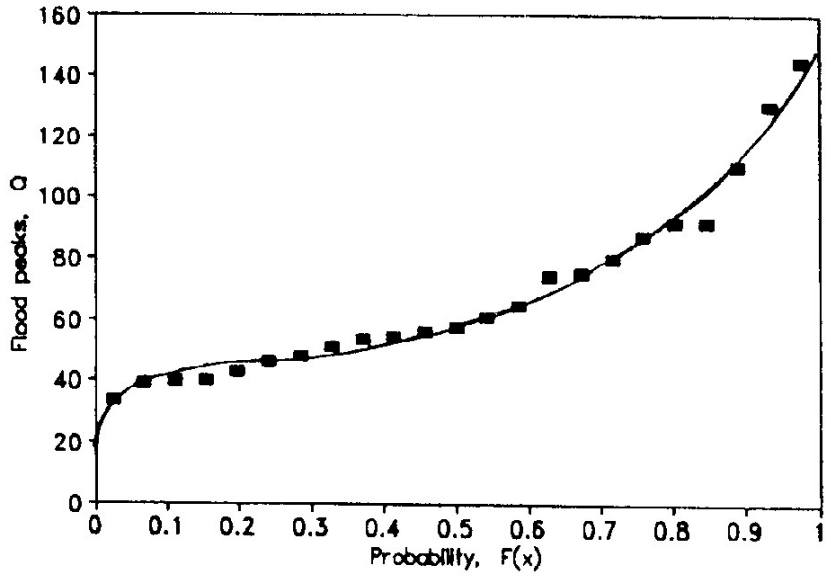


Figure 3.4 Probability plot of flood flows.

The common plotting position relationships used in flood frequency analysis are listed below.

California	F(i) = i/N
Hazen	F(i) = (i-0.5)/N
Weibull	F(i) = i/(N+1)
Gringorten	F(i) = (I-0.44)/(N+0.12)
Blom	F(i) = (i-3/8)/(N+1/4)
Chegodayev	F(i) = (i-0.44)/(N+0.40)

On the basis of comparative study of several plotting position relationships, Cunnane (1978) suggested the use of the unbiased plotting formulae as follows:

- for Normal probability paper

$$F(i)=\frac{\left(i-\frac{3}{8}\right)}{\left(N+\frac{1}{4}\right)} \qquad (3.10)$$

- for Gumbel and Exponential probability paper

$$F(i) = \frac{(i - 0.44)}{(N + 0.12)} \tag{3.11}$$

A Probability paper is a special designed paper which will plot a cumulative distribution function as a straight line. If a set of observed data plot as a straight line on probability paper, the data can be said to be distributed as the distribution corresponding to the probability paper used. The probability plot on Gumbel probability paper is shown on Figure 3.5

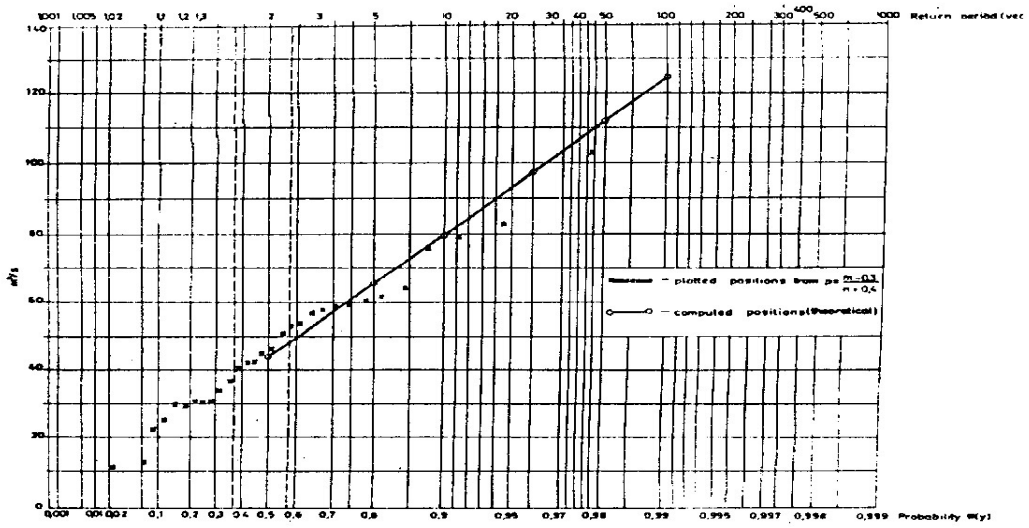


Figure 3.5 Flood frequency curve plotted on Gumbel probability paper.

3.4.2 Analytical Approach

In analytical approach, the concept of theoretical distributions is employed. A distribution is an attribute/reflection of a statistical population. The population consists of elements, each of which has an associated variable *X*. The distribution describes the constitution of the population as seen through its x values, i.e., it tells whether the x values are all bunched together or spread out, whether they are in general very large or very small and whether they are symmetrically disposed on the x axis. The description of the constitution of the population is usually achieved through the description of the statistical properties of the population elements i.e. the mean, standard deviation and skewness.

A distribution can also give the relative frequency in the population in the same way that a histogram gives that information about a sample. These relative frequencies may also be considered as probabilities. The distribution therefore tells the probability, Pr(X<=x),that the X value on an element drawn randomly from the population would be less than a particular value x. Knowing Pr(X<=x) for all x values , the laws of probability may then be used to deduce the probability of any proposition about the behavior of a random sample of X values drawn from the population.

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

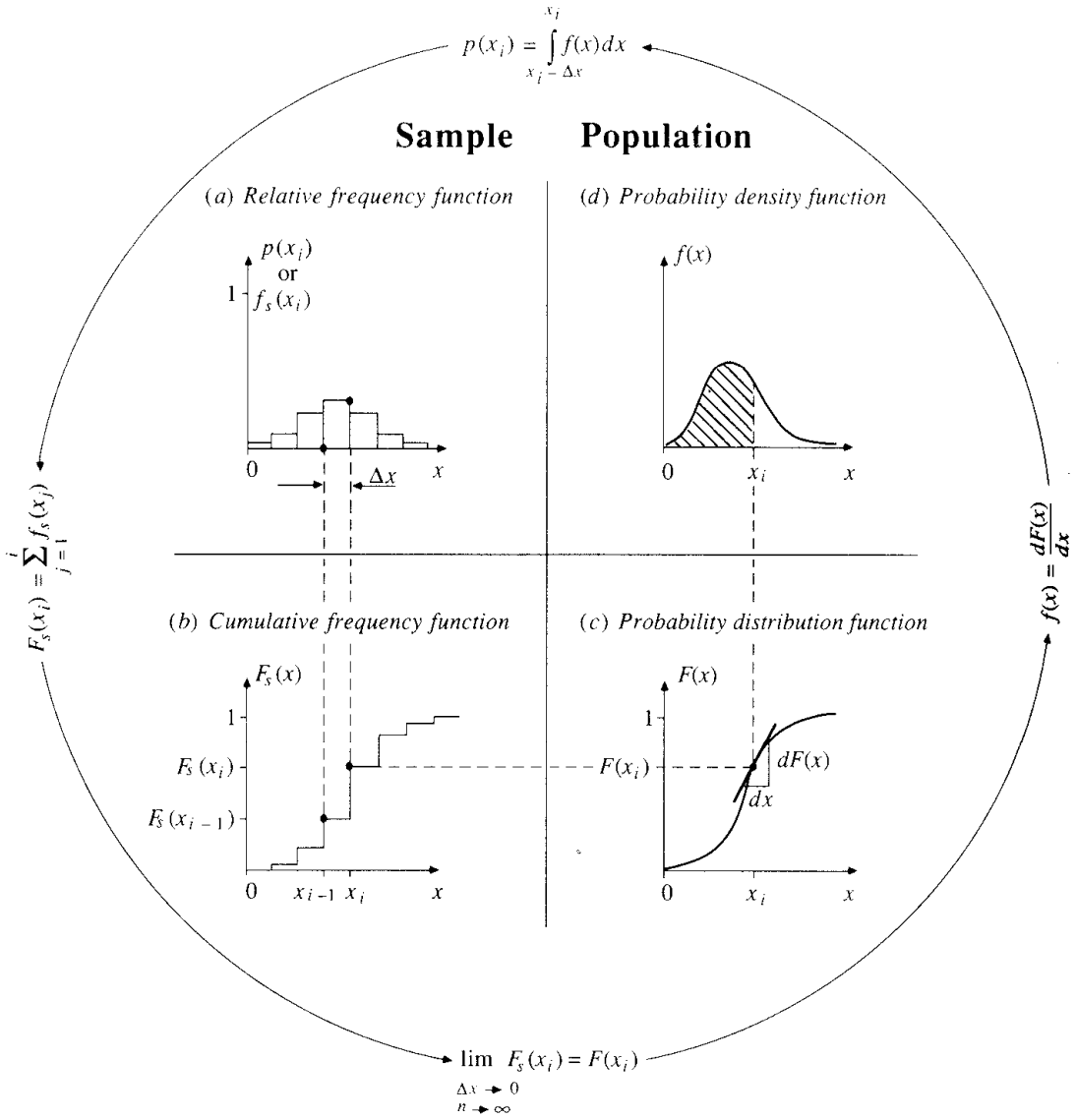


Figure 3.6 Relation between relative and cumulative frequency function and probability distribution and density functions (fFrom V.T. Chow, 1988)

DEFINITIONS OF FUNCTIONS DESCRIBING A DISTRIBUTION

Probability Density Function (pdf)

This is the function that gives the probability of occurrence of an event. The PDF is constructed so that the area under its curve bounded by the x-axis is equal to 1 when computed over the range of x for which f(x) is defined (Figure 3.6d).

Cumulative Distribution Function (cdf)
This is the function that gives the probability of occurrence of all the events that are equal to or less than a specified event (Figure 3.6c).

Non-Exceedance Probability
This is the probability expressed by the cumulative distribution function (F(x)), i.e., it is the probability of occurrence of events that are equal to or less than a specified event, i.e., Pr(X≤x).

Exceedance Probability
This is the complement of non-exceedance probability or CDF. This is the probability of occurrence of all events that are equal to or greater than a specified event, i.e.,[1- Pr(X≤x)]. F

STATISTICAL DISTRIBUTIONS
Many statistical distributions have been proposed for flood frequency analysis todate. Various distributions have been proposed because of their ability to model different shapes of histogram of flood peaks. Typical shapes of histograms of flood peaks are shown in Figure 7.

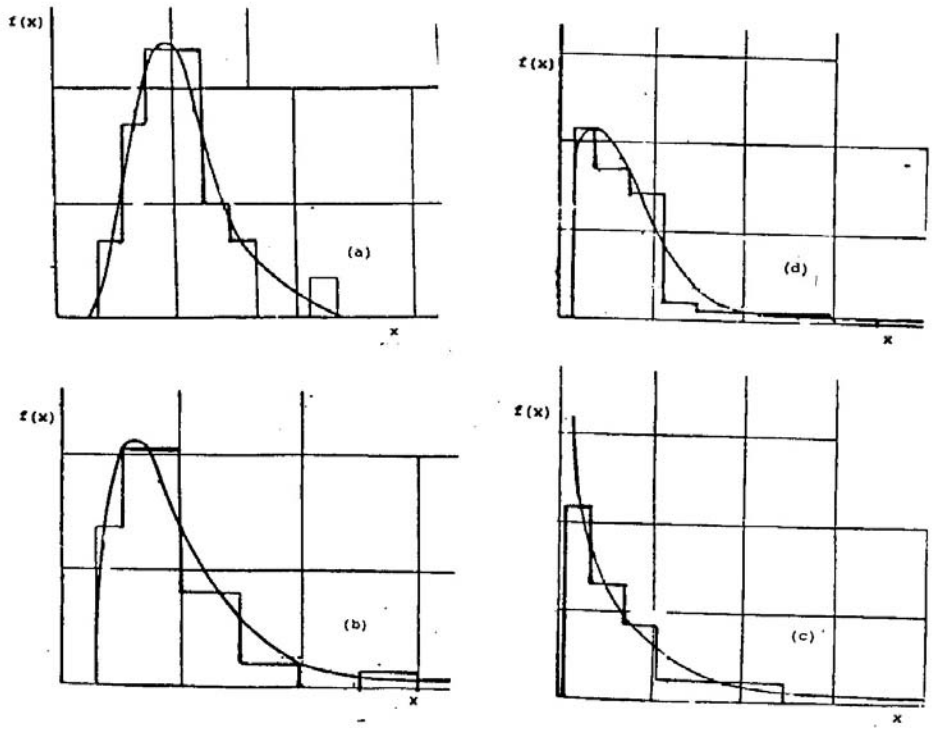


Figure 3.7 Typical shapes of histogram of flood peaks

The statistical distributions provide the essential basic formulae to model a peak and low flow data aimed at deriving frequency curves required to determine flood low flow magnitudes.
Many statistical distributions have been proposed to date. In this section only three commonly used distributions are described. These are:

- ***Extreme value type 1 (EV1)*** (Gumbel,1941)
- ***Lognormal (LN)*** (Hazen,1914)
- ***Log Pearson type 3 (LP3)*** (USWRC,1967)

METHODS OF PARAMETER ESTIMATION

Fitting a distribution to data sets provides a compact and smoothed representation of the frequency distribution revealed by the available data, and leads to a systematic procedure for extrapolation to frequencies beyond the range of the data set. When flood data from a given site is assumed to follow a certain distribution, the next step is to estimate the parameters of that distribution so that the required flood magnitudes can be calculated with the fitted model. Several general approaches are vailable for estimating the parameters of a distribution. The most commonly used estimation methods are the following:

Method of moments (MOM)

The method of moments is one of the most commonly and simply used method for estimating the parameters of a statistical distribution. In most cases the first three central moments , i.e., the mean, variance and skewness are adequate to estimate the required parameters.

Maximum likelihood (ML)

Parameters estimated by maximum likelihood are determined by maximizing the sample log likelihood function. The unknown parameters may be obtained by setting each of the partial derivatives with respect to each parameter equal to zero and solving equations simultaneously. These equations unfortunately do not often take a simple closed form and the numerical solutions have to be used. Sometimes there is a failure to obtain proper solutions to the equations due to complexity of log likelihood functions, particularly when sample size is small or when the distribution has more than two parameters.

Probability weighted moments (PWM)

This is a statistical estimation procedure based on the calculation of probability weighted moments (PWM). This method is simple, unbiased and stable.

Least square method (LSM)

The method of least squares uses, in effect, regression analysis to fit a straight line to the data on a probability plot. The plotting positions have to be expressed as standardized variate values, y, rather than probability values. If unbiased plotting positions are used this method of estimation is unbiased and quite efficient.

APPLICATION OF STATISTICAL DISTRIBUTIONS

The Method of Moments (MoM) is the only method described here to determine the parameters of the statistical distributions. The other techniques are not convenient to use in manual calculations.

Extreme Value type 1 (EV1)

The extreme type I (EV1) distribution is also known as Gumbel distribution. The c.d.f., F(x) of Gumbel distribution is defined as

$$F(x) = e^{-e^{-\left(\frac{x-u}{\alpha}\right)}} \tag{3.12}$$

where μ = Location parameter
α = Scale parameter

Estimation of parameters (MOM)
Mean: $\mu = u + 0.57728\alpha$
St. Deviation: $s = 1.28\alpha$
Skewness: $g = 1.14$
The parameters u and α of the distribution are determined from the sample data.

The flood magnitude X_T of return period T is estimated from the expression

$$X_T = u + \alpha y_T \tag{3.13}$$

where y_T is a the EV1 reduced variate corresponding to a given return period T. The values of y_T are provided in Table 3.1 for selected values of return period.

Table 3.1 Frequency factors for use in the Log-Normal and EV1 distributions.

Return Period, T	Reduced variate, z_T (for log-normal distribution)	Reduced variate, y_T (for EV1 distribution)
2	0.00	0.37
5	0.84	1.50
10	1.28	2.25
25	1.75	3.20
50	2.05	3.90
100	2.33	4.60
1000	3.09	6.90

Lognormal Distribution

The probability density function (p.d.f.), f(x) and the distribution function (d.f.), F(x) of the lognormal distribution are defined as

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log_e x - \mu_y}{\sigma_y}\right)^2} \tag{3.14}$$

$$F(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \int_0^X e^{-\frac{1}{2}\left(\frac{\log_e x - \mu_y}{\sigma_y}\right)^2} dx \qquad (3.15)$$

where $y = \log_e x$

μ_y = Location parameter of y series

σ_y= Scale parameter of y series

Estimation of parameters (MOM)

Mean = $E(y) = \mu_y$

Variance = $E(y-E(y))^2 = \sigma_y^2$

Application of the Log-Normal distribution involves transforming annual floods to logarithmic values ($y = \log_e x$) and then computing the following statistics:

Mean =μ_y

St. deviation =σ_y

The flood magnitude X_T of return period T is estimated from the expression

$$y_T = \mu_y + \sigma_y . Z_T \qquad (3.16)$$

where Z_T is a standardised reduced variate corresponding to a given return period T. The values of Z_T for some selected return periods are provided in Table 3.1. The values of Z_T for any given non-exceedance probability can be determined from Appendix 2, tables 2, which gives the area under the normal curve. The untransformed value of the flood magnitude $X_T = 10^{yT}$.

Log Pearson type III Distribution

The probability density function (p.d.f.), f(x) and the distribution function (d.f.), F(x) of the Log-Pearson type III distribution are defined as

$$f(x) = \frac{(\log_e x - x_0)^{\gamma - 1} e^{-\left(\frac{\log_e x - x_0}{\beta}\right)}}{\beta^{\gamma} \Gamma \gamma} \qquad (3.17)$$

$$F(x) = \int_{x_0}^{x} \frac{(\log_e x - x_0)^{\gamma - 1} e^{-\left(\frac{\log_e x - x_0}{\beta}\right)}}{\beta^{\gamma} \Gamma \gamma} dx \qquad (3.18)$$

Where x_0 = Location parameter

β = Scale parameter

γ = shape parameter

Estimation of parameters (MOM)

Mean , $\mu = x_0 + \beta\gamma$

Variance, $\sigma^2 = \beta^2\gamma$

Skewness, $g = 2/\sqrt{\gamma}$

Application of the Log-Pearson type 3 involves transforming annual floods to logarithmic values ($z = \log_{10}X$) and then computing the following statistics:

Mean = μ_z
St. deviation = σ_z
Skewness = g

The flood magnitude z_T of return period T is estimated from the expression

$$z_T = \mu_{z+K_T}.\sigma_z \qquad (3.19)$$

where K_T = f(T,g), a function of both recurrence interval and skewness. These values are provided in Appendix 2, table 3. The untransformed value of the flood magnitude, $X_T = 10^{Z_T}$.

3.5 RISK OF FAILURE

- The decision of what return period is appropriate for the design of a particular project is not solely a hydrological problem. Constraints usually considered include: Economic, Political & Environmental
- Improvement on the safety or reliability of the scheme has implications on the costs of the project
- It is however important to consider the probability, or risk, of the design flood being exceeded during the expected life of the project

Example:

Suppose the design event has a return period of T years and the projected life period of the project is L years

- The corresponding annual probability of exceedance, ***p*** is given as ***p*** = 1/T
- The probability of non-occurrence in any one year is ***q*** = (1-1/T)
- The probability of non-occurrence in L years is $q = (1-1/T)^L$
- The probability that the flood magnitude Q will occur at least once in the L years is ***r*** =$1-(1-1/T)^L$
- The value of ***r*** is the risk of failure, the flood Q will be exceeded in the L years of the projected life of the project

3.6 FLOW DURATION CURVES

One of the most important characteristics of a stream is a flow duration. A flow duration curve answers provides answers to the following frequently asked questions by hydrologists.

- Does the stream have equitable flow, or does it flow at high and low extremes?
- Does the stream have permanent flow?
- What are the low-flow and high-flow characteristics (is it flashy)?
- How much of the time (proportion of time) does it flow at various discharges?
- If the stream is to furnish power, provide water, provide deep water for transportation, transport sewage effluent away or dilute pollution, how often does this happen?

A flow duration curve shows graphically the relationship between any given discharge and the percentage of time that discharge is equalled or exceeded. Flow duration curves are widely used in the design of water projects as well as hydrological studies. For example, flow duration curves are applied in the following areas:

- Determination of diversion flow
- Investigation of low flows
  - Values of 90%, 95%, 96% & 99% have been used as measures of stream's low flow potential
- Reservoir capacity analysis
- Hydropower studies
  - The 90% value is used as a measure of 'run-of the river ' hydropower potential
- Measure of groundwater contribution to streamflow (90% value)
- Tool for comparing drainage basic characteristics, particularly the effect of geology on low flows
- Water quality studies
  - Flow duration curve is used to indicate the % of time that various levels of water pollution will occur following the introduction of a pollutant of given volume and strength into a stream

Construction of flow duration curve

The construction of flow duration curve involves the following steps

- All measured daily flows are grouped into class intervals
- The total number of occurrences above the lowest limit of each class interval are then expressed as percentage of the total number of days in the record
- A duration curve is obtained by plotting the computed percentage against the lower limit of the intervals

Figure 3.8 shows constructed duration curves for a number of rivers in Ruvu River basin. It can be noted that for some of the streams the curves are "flatter" and for some are "flashy".

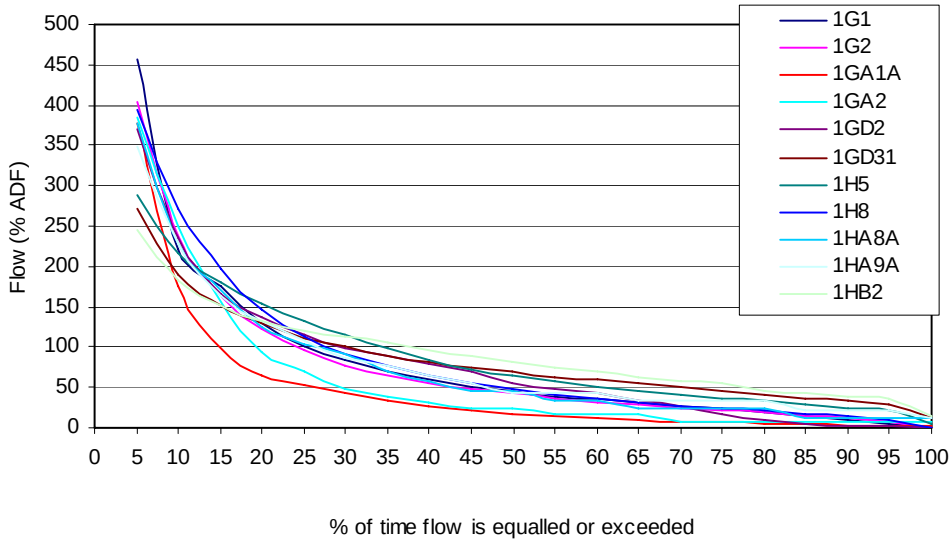


Figure 3.8 Flow duration curves for Ruvu sub-catchments in Tanzania

3.7 MASS CURVE ANALYSIS

A mass curve procedure (Rippl Diagram) is one of the earliest methods for estimating the size of storage required to meet a given draft. A mass curve is constructed by plotting the accumulative monthly or yearly flows against time. The slope of the mass curve at any time is a measure of the inflow rate at that time. Demand curves are straight lines having a slope equal to the demand rate which is uniform. Demand lines drawn tangent to the high points of the mass curve represent rates of withdrawal from the reservoir. Assuming the reservoir to be full whenever a demand line intersects the mass curve, the maximum departure between the demand line and the mass curve represents the reservoir capacity required to satisfy the demand. It should be noted that the accumulative inflows have to be adjusted for evaporation loss and required releases for downstream users. If the demand is not uniform, the demand line becomes a curve. It is essential, however that the demand line for non-uniform demand coincide chronologically with the mass curve.

Steps to determine the reservoir storage:

- For the proposed location, construct a cumulative curve of monthly streamflow. Determine the slope of the cumulative draft appropriate for the graphical scales adopted.
- On the mass curve diagram, superimpose the cumulative draft lines such that it is tangential to the mass inflow curve as shown in Fig. 7.
- Measure the largest intercept between the inflow curve and the draft line. From Fig. 9, the intercept C2 is greater than C1, and therefore the design capacity would be taken as C2 (=150,000 m^3). From the figure, it can be seen that the reservoir is full at A, begins to empty from A to B, and refills from B to C. From C to D, the reservoir spills, and again, the content falls until it just empties at E. The critical drawdown period, D to E, for this example is approximately 28 months.

In the mass curve procedure, two important assumptions are made:

- The reservoir is full at the beginning of the critical drawdown period.
- As the analysis utilizes historical streamflow data, it is implicit that future sequences of inflow will not contain a more severe drought than the historical sequence.

The procedure exhibits two important attributes namely

- It is simple and widely understood
- Because the analysis used historical data, seasonality, autocorrelation, and other flow parameters are taken into account.

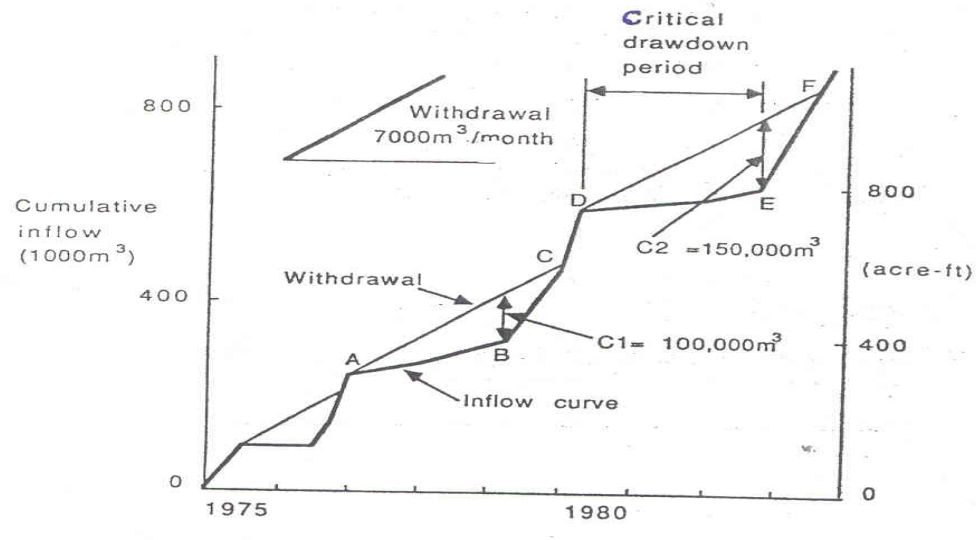


Figure 3.9 Reservoir capacity-yield estimation by mass curve procedure (From Maidment, 1992)

3.8 APPLICATIONS

This section presents examples of application of statistical techniques in water resources planning and design by posing questions.

Example 1 Design of drainage structure

In connection with the design of a culvert along the Pugu - Kisarawe road at a location called Kisimati. The Annual Maximum discharge data for the stream crossing the road at Kisimati are available at a gauging station located close to the site where the culvent is going to be constructed. The data are summarized in the table below

Year	**Peak Discharge (m^3/sec)**
1995	14

1996	17
1997	23
1998	27
1999	9
2000	6
2001	8
2002	10
2003	19
2004	12

Required:

- Determine the design flood with frequency of occurrence of 25 years under the assumption that the observed data follows Gumbel Distribution
- Assuming the life period for the culvert to be constructed is 20 years. Estimate the probability of non-occurrence of the 25 year flood during the life period of the structure ?

Example 2 Determination of peak flows

Given a sample of Annual Maximum series (Q) has the following statistics:

Mean of Log Q = 2.190
Std Deviation of Log Q = 0.164
Skewness of log Q = .17

Determine the peak flow of return period equal to 50 years under each of the assumptions:

- -parameter distribution
- Log-Pearson distribution

Example 3 Assessment of water availability for water supply planning

Given that daily streamflow records from one of the rivers in Tanzania for a period from 1/10/2005 to 20/1/2006 has been clustered according to low magnitudes as shown in the Table below:

Class interval of flow magnitudes(m^3/sec)	**No of occurrences of flow records**
0-3.00	5
3.01-6.00	35
6.01-9.00	36
9.01-12.00	15
12.01-15.00	13
15.01-18.00	5
18.01-21.00	2
21.01-24.00	1

Using the data given in the above table, construct a flow duration curve and estimate the flow that is available 70% of the time.

Example 4 Determination of reservoir storage capacity
Given below are monthly flows for a certain river in Tanzania

Month	Flow (^{3}/sec)	Month	Flow (^{3}/sec)
Jan 2001	32.5	Jan 2002	7.5
Feb 2001	32.0	Feb 2002	14.0
Mar 2001	7.0	Mar 2002	27.0
Apr 2001	1.0	Apr 2002	10.8
May 2001	0.5	May 2002	4.0
Jun 2001	0	Jun 2002	1.4
Jul 2001	0	Jul 2002	0
Aug 2001	0	Aug 2002	0
Sep 2001	0	Sep 2002	0
Oct. 2001	0	Oct. 2002	6.6
Nov. 2001	2.0	Nov. 2002	2.1
Dec 2001	7.5	Dec 2002	24.1

Determine the reservoir storage capacity for a constant demand rate of 4.5×10^6 m^3 per month.

3.9 DATA INFILLING TECHNIQUES

3.9.1 Overview of Infilling Missing Environmental Data

The goal of any infilling technique is the production of a complete data set which may then be analyzed using complete data inferential methods (Little and Rubin, 1987). This, in turn, generally leads to better estimates of the mean and variation of the flow and pollutant loading variables. For example, it may be useful to apply data generation techniques to synthesize or generate hydrological data in cases where: (1) there are gaps in the series of observed data; (2) the observation period is short; and (3) data are not available at the site of interest (e.g., potential dam site) but in the neighbouring region. (Sokolov et al., 1952).

There are, however, methods available for excluding missing data events by using a variant of the seasonal Kendall estimator for trend and slope, or by treating missing data as an additive outlier (Alvo and Cabillo, 1994; Gilbert, 1987; Hirsch et al, 1982; Liu and Chen, 1991). In general, these methods provide less power in drawing conclusions from the data as data are either removed or excluded from the statistical calculations.

A variety of methods exist in the literature for infilling missing environmental data, ranging from the simple to the complex. Infilling methods are identified for data sets generated from: (1) streamflow monitoring; (2) streamflow monitoring during periods of ice effect; (3) atmospheric aerosol sampling; (4) estuarine monitoring and modelling; (5) rainfall sampling; (6) groundwater monitoring and modelling; (7) solar radiation and relative humidity monitoring; (8) snowcover area-runoff monitoring and modelling: and (9) home injury incident reporting (Bryant and Shih,

1989; Conn, 1988; Dey, 1984; Engqvist, 1990;Heidam, 1982; Hirsch, 1982; Melcher and Walker, 1990; Shih and Cheng. 1989; Snijders, 1986). Methods for infilling missing ground-level ozone and other criteria air pollutant data arc also given in the literature (Batterman, 1992; Glen et at., 1996, Hemphill, 1988).

Missing data may also be addressed by specifically designed computer software. One such computer software program (SOLAS, Version 1.0) allows for several methods of infilling missing data; (1) Mean value infilling; (2) Hot-deck infilling: (3) Last value carried forward; and (4) multiple infilling (Statistical Solutions, 1997). Multiple infilling is different from other infilling methods in that this method uses more than one value for each missing data entry to estimate the characteristics of the complete data set. Multiple passes through the data are required until the program converges on a "new" data set. This program is not utilized in this study due to its limited menu of infilling methods.

3.9.2 Stationary or Non-Stationary Determination

A stationary time series is a series whose properties do not change with respect to time. Environmental time series occurring in practice are usually non-stationary and may be divided into three classes

1) Those which exhibit stationary properties over a long period of lime;
2) Those which are approximately stationary over short periods of time; and
3) Those which exhibit non-stationary properties, that is, their properties are continuously changing with time (Fuller and Tsokos, 1971).

Reasons for non-stationary behaviour in time series may include natural causes (e.g., temperature, seasonal effects), or they may be anthropogenic (e.g. urbanization of a watershed, new restrictions on point source discharges). If periodicity or trends are present, then a time series analysis method (e.g., autoregressive moving average [ARMA] methods) should be utilized to generate a synthetic series of data (Shin and Cheng, 1989). If there is no correlation between the data and time then standard stationary time series methods may be employed.

Some infilling methods, (e.g., interpolation of missing data using data from adjacent stations or the station itself), are not always applicable for data that exhibit seasonal cyclic patterns and spatial variation within a given region (Shili and Cheng, 1989). Therefore, it is important to determine whether a data set is stationary or non-stationary before performing any infilling methods. For example, Hill (1986) observes that NO_3-N concentrations in streams vary seasonally, and as a result uses separate flow-concentration relationships, one for each season, to account for the seasonal variation.

3.9.3. Description of Several Missing Environmental Data Infilling Techniques

The principal infilling methods for environmental data sets are:

- mean value infilling
- interpolation equations

- hot-deck infilling
- intra-station interpolation
- regression infilling
- composite methods
- multiple infilling methods
- spline function and forecasting models

(Gyau-Boakye and Schultz, 1994; Little and Rubin, 1987: Mizumura, 1985; Reinelt and Grimvall, 1992: Sokolov et al., 1982).

MEAN VALUE INFILLING

Mean value infilling uses means from known values as a substitute for missing values. Mean value infilling is a relatively simple method for infilling missing data. Mean values may be sub-divided into classes (i.e., seasons, months) to account for seasonality or other non-stationary time series variations. Median or winsorized values may also be substituted for missing values to dampen the effect of outliers. The '"winsorization" method can be utilized to estimate the mean and standard deviation of a symmetric distribution even though the data set has a few missing or unreliable values at either or both ends of the ordered data set (Gilbert, 1987). Roughly speaking, winsorized means are more robust than straight means, and more sensitive than medians (Snijders, 1986). Shih and Cheng (1989) use this method for infilling missing solar radiation data because they conclude that this method is easy to use and provides accuracy equal to the standard error of the mean, They recommend using mean value infilling when regression techniques are not applicable.

The major problem with mean value infilling is that standard methods of analysis treat all points, even infilled points, as if they have equal precision. Mean value infilling is a smoothing process and may result in underestimating the inherent variability of the data (Galpin, 1990).

INTERPOLATION EQUATION

The interpolation equation can be expressed as

$$R_i = Q_o \pm \frac{1}{N}(Q_N - Q_o) \tag{3.20}$$

where R is the computed runoff for the ith day Q_0, is the observed runoff at the beginning of the gap Q_N, is the observed runoff at the end of the gap, and N is the number of missing data +1 day.

HOT-DECK INFILLING

In contrast with mean value infilling, hot-deck infilling involves substituting individual values drawn from similar responding units. An example of hot-deck infilling for a watershed monitoring program is to replace a missing data entry in one station with a value from another station with similar watershed characteristics for the same time period. This is the method recommended by Rantz et al. (1982) for the infilling of missing flow data for a stream in a period

of fluctuating discharge. Shih and Cheng (1989) use a regional average from other nearby stations to infill missing solar radiation data for a single station.
Various environmental data collection programs, including watershed monitoring programs, utilize variants of the simple hot-deck infilling method. Kottegoda and Elgy (1979) evaluate two variants of hot-deck infilling using monthly flow data sets from England and Wales. The variants include: (1) the weighted average model, in which the parameters are inversely related to the distances between the stations with, complete and incomplete records; and (2) the modified weighted average model, where biases in the mean and variance are corrected.

Bryant and Shih (1989) use a variation of the hot-deck infilling method for estimating lost groundwater data, instead of infilling values directly from nearby groundwater monitoring wells, missing piezometric data from base wells are estimated by observing the amount of change in the data from reference piezometers. The average change in reference station data is calculated and added to or subtracted from the last known datum taken at the base station. Hemphill et al., (1983) use a variant of the hot-deck infilling method for infilling missing ground-level ozone data. Ozone data sets with missing data are infilled based on a weighted difference function that gives more emphasis to differences during the mid-afternoon, when ozone values are generally the highest for the day.

The construction of a daily hydroclimatological data set for 1,009 stations across the continental United States from 1948 through 1988 uses another variant of the hot-deck infilling method (Wallis et al., 1991). Missing data are infilled using a neighbouring station value multiplied by the ratio of the long-term monthly mean at the target site to the long-term mean at the neighbouring (or reference) site.

The drainage area ratio method, utilized for estimating lost streamflow data, is another variant of hot-deck infilling. Because of its simplicity, this method is widely employed by hydrologists. This method estimates lost streamflow data by multiplying the streamflow data of the reference station against the ratio of the target and reference watershed drainage areas. For example, OWML currently uses this method for estimating missing flow data for the gauging station near Clifton, Virginia (ST40). The reference station is upstream near Manassas, Virginia (ST45). Missing data for ST40 is given by the following relationship:

Estimated ST40 Flow = (ST45 Flow)*(Drainage Area of ST40/Drainage Area of ST45) (3.21)

Shields and Sanders (1986) use the drainage area ratio method for estimating streamflow data lost because of malfunctioning equipment. Hirsch (1979) and Parrett and Johnson (1994) use this method as a means for extending streamflow records against base station records. The performance of the drainage area method may improve with an increased similarity of the two watersheds (e.g., morphology, land use, imperviousness, drainage area, etc.).

INTRO-STATION INTERPOLATION

Intra-station interpolation replaces missing data by use of interpolation on related known data collected at the same station. For example, if the date of a missing data event is known, a linear interpolation may be performed using: (1) the date of the missing data event; (2) the date and

value of the previously observed sampling event; and (3) the date and value of the following observed sampling event. This method is employed by Cluis and Boucher (19S2) in their study of weekly sampled water quality parameters. Melcher and Walker (1990) use this method for estimating streamflow during periods of ice-effect. Intra-station interpolation is also the method recommended by Rantz et al. (1982) for the infilling of missing flow data for a stream in a period of low or medium flow recession.
A variation of intra-station linear interpolation is the method of "Last Value Carried Forward" (Little and Rubin, 1987). This method is easy to implement and simply replaces the missing value with the last known recorded value for that class. For example, if the Total Phosphorus (TP) concentration for a particular stormflow event is missing, then the replacement value is the previously recorded stormflow TP concentration measured at the same station.

REGRESSION INFILLING
Various types of regression relationships are useful for infilling missing data. Reference variables may be of the same type (e.g., flow vs. flow) or different (e.g., flow vs. concentration, precipitation vs. concentration, land use vs. concentration). In general, regression infilling replaces the missing values with predicted values based on a regression of reference values.

Standard linear regression methods or polynomial regression methods are often utilized to predict values for missing data. However, regression relationships that apply the method of least squares are very sensitive to imperfect data quality, especially to outliers (Snijders, 1986). Moreover, the least squares method, using either untransformed values or log-transformed values, produces less variability in the synthesized data than is present in the real data series, i.e. variance reduction (Hirsch, 1979)

It is also important to correct for transformational bias when regression analyses use the logarithms of data. The bias is introduced in the re-transformation from "log space." Where regression estimates are derived, to "real space," generally the realm of interest (Cohn et al, 1989). This is particularly important for estimating watershed loads from monitoring data.

For example, one method of estimating watershed loads is to use a regression relationship based on streamflow and pollutant concentration data. To estimate the mass load of a pollutant, streamflow and pollutant concentration means for short time periods are regressed and summed to estimate the total mass over a longer period. No data transformations permit this approach (i.e., the analyst is summing the means). However, if a log-transformation is used, summing the mass over the re-transformed values results in summing the median. This results in an estimate that is biased low for me total mass (EPA, 1997d). Various methods are available for correcting the re-transformational bias (Cohn et al., 1989; Gilbert, 1987; Koch and Smillie, 1986).

Flow-Flow Relationships
Hydrologists often utilize linear regression relationships for extending the records of streamflow. These relationships use the records of a reference stream-gauging station which cover the period of interest.

Simple and multiple regression models are used extensively for establishing relationships between two or more variables. The established relationship can be used to estimate a missing value of one of the variables given the corresponding value (s) of the other variable (s). An example of *regression model without including rainfall* is:

$$Q_A = \alpha + \beta_1 Q_B + \beta_2 Q_C + \cdots + \beta_m Q_N + \varepsilon \tag{3.22}$$

where Q_A, Q_B, Q_C, .. Q_N are the discharges at stations *A, B, C,...,* and *N* and α is constant which can be taken as representing baseflow and channel storage, and β_1, β_2,… β_m are partial regression coefficients. Likewise, an example of *regression model with rainfall included* is:

$$Q_{A,t} = \alpha + \beta_1 Q_{B,t} + \cdots + \beta_m Q_{N,t} + \lambda 1_1 P_t + \lambda_2 P_{t-1} + \cdots + \lambda_r P_{t-r} + \varepsilon \tag{3.23}$$

where λ_1, λ_2,.. and λ_r, are additional model parameters to be associated with current and antecedent rainfall.
Some runoff models incorporate only rainfall in current and earlier time periods P_t, P_{t-1}, P_{t-r}. These models are either linear or non-linear. The basic equation describing a non-linear model can be written as:

$$Q(t) = \int_0^t h_1(\tau) P(t-\tau) d\tau + \int_0^t \int_0^t h_2(\tau_1, \tau_2) P(t-\tau_1) P(t-\tau_2) d\tau_1 d\tau_2 + \cdots + \cdots +$$

$$\int_0^t \int_0^t h_r(\tau_1, \tau_2, \cdots, \tau_r) P(t-\tau_1) \cdots P(t-\tau_r) d\tau_1, \cdots, d\tau_r, + \cdots \tag{3.24}$$

Another example: For water management this method is used to estimate daily flows for the two stream-gauging stations. The stream-gauging station near Aden, Virginia (ST25) uses the reference station ST45, while the stream-gauging station near Bristow, Virginia (ST30) uses the reference station ST25. The linear regression relationships are:

ST25 Daily Flow = (0.79861) (ST45 Daily Flow) + (-9.83473) (R^2 = 0.739)
ST30 Daily Flow = (0.40794) (ST25 Daily Flow) + (0) (R^2 = 0.625)

The coefficient of determination (R^2) is a measure of the linearity of the regression relationship (Neter et al., 1985). Values of R^2 close to 1 indicate a near perfect linear relationship. However, as previously stated, the use of linear regression for estimation of missing values may cause a reduction in the variation of the values at the short-record station. An underestimation of the variance is seen in the linear or log-linear regression technique and may result in an

underestimation of hydrologic extremes (Hirsch, 1982). There are equations available to adjust the mean and standard deviation of a short-record station with a long-record station (Sokolov et al, 1982).
The above-mentioned methods can be used for filling gaps in hydrologic data series of both the high-flow and low-flow seasons.

Flow-Concentration Relationships
Regression relationships are also utilized to relate flow data against concentration data of a water quality chemical parameter (e.g., TP, NH3-N) (Smith et al., 1982; Stack and Belt, 1989). Smith et al (1982) present the most comprehensive study of this method by using 303 USGS National Stream Quality Accounting Network (NASQAN) stations across the continental United States. For each station, a regression function is selected from among eleven possible relationships on the basis of the R2 value. The study identifies that the relationship between discharge and concentration may be expressed as a flow-adjustment equation of the form: C - a + b* f(Q); where, "a" and "b" are derived constants, "C" is the estimated concentration, ”Q" is the instantaneous discharge, and "f(O)" may have one of the following forms:

Functional Form	Name
f(Q) = Q	linear
f(Q) = InQ	log
f(Q)= l/(1+BQ)	hyperbolic (where B is positive)
f(Q)= 1/Q	inverse

The study analyzes pollutant concentration data, estimated from flow, for detection of trends and magnitude of any trends using the procedures outlined by Gilbert (1987) for the Seasonal Kendall test and the Seasonal Kendall Slope estimator. These two statistical procedures do not assume an underlying distribution for either the flow or pollutant concentration data. Under two different significance criteria (a = 0.10 and a = 0.05. two-tailed), significant trends are observed at far more NASQAN stations than would be expected by chance alone (Smith et al., 1982).
The Ratio Estimator Method is another method applicable for infilling missing chemical data (Leitch, 1998). This method provides a more precise measure of missing chemical data as the infilled data are flow-weighted. The Ratio Estimator Method is well-suited for use when ample flow data exist with only limited chemical concentration data. The method is further refined if two separate analyses are performed on the flow data: one using baseflow data and the other using stormflow data. Leitch (1998) provides a description of the method with applications of the model in the literature.

Precipitation Data Relationships
Precipitation data are generally easier to collect than stormflow data. Some areas of the world have a very limited amount of recorded stormflow data, yet rainfall data may be more available (Sokolov et al., 1982). Consequently, one may obtain a relatively long sequence of precipitation data and use it in a model to simulate a corresponding sequence of stormflow. This simulated sequence may be utilized to make inferences about the statistical properties of actual stormflow (Troutman, 1985).

Often it is possible to develop relationships with climatic and basin characteristics to define selected hydrologic variables. There are a variety of methods for the conversion of precipitation data to stormflow data. These techniques range from simple empirical formulae that employ a runoff coefficient to computer models that attempt to simulate various components of the hydrologic cycle. One of the most widely utilized methods is the unit hydrograph method. This method requires the determination of parameters that govern loss (or infiltration) rates, the precipitation-runoff transformation function (i.e., unit hydrograph), and baseflow (Sokolov el al., 1982). EPA (I998c) presents a summary of infiltration models.
One relatively simple method relating precipitation data against streamflow data is the Rational Method (Adams, 1998). The method is very straightforward and converts rainfall over an area into a runoff rate. It may be expressed as:

Q = (C)(A)(I)(43,560 ft2/acre)(l hr/3600sec)(l ft/12 in)
where: Q = runoff rate (ft^3/s)
C = a dimensionless runoff coefficient ranging from 0 to 1.0 (1,0 is totally impervious to infiltration)
I = rainfall intensity (in/hr) [which may be plotted using a hyetograph]
A = Area (acres)

Discharge (ft^3) associated with a particular storm may be estimated by multiplying the above calculated runoff rate with the time of concentration (t_c). The time of concentration parameter is estimated from the hydrograph of the storm and is equivalent to the time it takes for the runoff from the farthest hydrological point of the watershed to reach the watershed outlet (Adams, 1998). This method may be utilized to estimate lost flow data if reliable information is available for drainage area, percentage of imperviousness, and rainfall intensity parameters.
Leitch (1998) describes an extension of the Rational Method. This method, called the Simple Method, is based on data from the Nationwide Urban Runoff Program (NURP) study. The model is a variant of the Rational Method as the three factors utilized in the Rational Method are incorporated into the Simple Method along with several other factors (e.g., a factor that corrects for storms that produce no stormflow, a factor for the flow-weighted mean concentration of phosphorus in stormflow) to estimate phosphorus loads. This model is a deterministic model and its precision is described as being within the true load by one order of magnitude (Leitch, 1998). The method was developed primarily for use on sites with drainage areas less than one square mile. Moreover, the Simple Method only predicts loads associated with stormflow (i.e., baseflow loadings are not incorporated into this model).

Recession equation
The recession equation may take different forms, one of which can be written as:

$$Q_{t+\Delta t} = Q_t K^{\Delta t} \tag{3.25}$$

Where $Q_{\Delta t}$, is the discharge at time t+Δt. Q_t is the discharge at time *t* and *K* is the recession constant. The recession limbs in the continuous sections of the hydrographs are separated for each of the test catchments. Applying Eq. (3.25) to the selected calibration hydrographs, an

average value of *K* can be obtained for the basin in question through optimisation using the test parameter denoted by *f* given by:

$$f = \Sigma_1^n (Q_{t+1} - Q_{t+1}^*)^2 \tag{3.26}$$

Recursive models

The recursive models are based on the input, storage and output of a river basin system. The models are based on the continuity and storage equations combined with a storage-runoff relationship. The latter can be the linear, log-linear or non-linear storage-runoff relationship. The equations of the model assuming the log-linear storage –runoff to hold can be expressed as:

$$Q_{t+\Delta t} = Q_t \left[\frac{B_t}{Q_t} (1 - e^{-\Delta t / K}) + e^{-\Delta t / K} \right] \tag{3.27}$$

where B_t>0 and

$$Q_{t+\Delta t} = Q_t e^{-\Delta t / K} \tag{3.28}$$

Where B_t =< 0 and

$$B_t = (P_t - E_t) \tag{3.29}$$

Multivariate Relationships

Other physical parameters may be utilized to develop relationships for estimating pollutant mean concentrations or flow data. Multivariate methods are available for estimating streamflow and pollutant loads based on easily measured physical, land-use, and climatic characteristics for a given area of study. For example, Charbenau and Barrett (1998) note that the single most important variable for predicting TSS loads (and other positively stormflow correlated parameters) is stormflow volume. Therefore, multivariate relationships based on stormflow volume and other factors may be used for estimating TSS. Driver and Tasker (1988) and Driver and Troutman (1989) use the following parameters to estimate stormflow volumes and the associated loads for various chemical constituents:

Physical and land-use characteristics

- Total contributing drainage area (DA), in square miles
- Impervious area (IA), as a percent of total contributing drainage area
- Industrial land use (LUT), as a percent of total contributing drainage area
- Commercial land use (LUC), as a percent of total contributing drainage area
- Residential land use (LUR), as a percent of total contributing drainage area
- Nonurban land use (LUK), as a percent of total contributing drainage area
- Population density (PD), in people per square mile

Climatic characteristics

- Total storm rainfall (TRN), in inches
- Duration of each storm (DRN), in minutes
- Maximum 24-hour precipitation intensity that has a 2-year recurrence interval (INT), in inches
- Mean annual rainfall (MAR), in inches
- Mean annual nitrogen load in precipitation (MNL), in pounds of nitrogen per acre
- Mean minimum January temperature (MJT), in degrees Fahrenheit.
- The regression equation is of the form:

$Y = [a_0 * X_1^{a1} * X_2^{a2} ... X_n^{an}] * BCF$ where:

Y = estimated storm-runoff load or volume

a_c, $a_,$, a_2, a_n = regression coefficients

$X_,$, X_2,..., X_n = physical, land use, or climatic characteristics

n = number of physical, land use, or climatic characteristics in the regression model

BCF = bias correction factor (used for re-transformation back from "log-space" to "real-space")

Driver and Tasker (1988) and Driver and Troutman (1989) divide the United States into three regions and present regression relationships for a variety of chemical constituents (e.g., TP, TN, TKN. Flow) for each region. The following is an example of the application of this method.

A city planner from Reno, Nevada, is trying to estimate a storm-runoff load for TN for storms (TRN) that averaged 0.5 inch in a particular drainage area (DA) of 0.1 square mile, which has 5 percent industrial land use (LUI), 10 percent commercial land use (LUC), and 15 percent nonurban land use (LUN). The city planner would use the TN I model in the study (as Reno, Nevada, is grouped into Region I). Using the above equation, adding the appropriate constants to the land-use variables, and using the mean annual rainfall (MAR) of 7.20 inches for Reno, Nevada, the storm-runoff load is calculated as follows:

TN I = 1,132 * (0.5)(0.798) * (0.1)(0.960) * (5+l)(0.462) *(10+ 1)(0.260) * (15+2)(-0.194) * (7.20)(0.951) * 1.139

TN 1 =31 pounds

If the median response of the variable instead of the mean response is desired, the BCF of 1.139 would not be applied to the model.

Therefore, this method may be utilized with a measured or estimated rainfalJ data to predict missing stormflow and pollutant loadings. Driver and Tasker (1988) observe that their regression relationships are more accurate for predicting soluble constituent concentrations (e.g., dissolved solids, total ammonia plus organic nitrogen as nitrogen, TN) than suspended solids concentrations (e.g., TSS, TP).

COMPOSITE METHODS

Composite methods are defined in this study as infilling models which combine elements from various different other infilling models. Examples of various composite methods are given in the literature (Alley and Bums, 1983; Hirsch, 1982; Mimikou and Rao, 1982; Wallis et al, 1991).

Alley and Bums (1983) use simple linear regression with the variation of using data from among different base stations in the region for infilling missing data. This method differs from traditional approaches as different stations may be selected as the base station at different times. This approach also provides a decision rule for using only flow values from the same month or all flow values in developing the extension equation for estimating a particular missing value.

Mimikou and Rao (1982) utilize a three part model for synthesizing daily flows and flood hydrographs at a site with missing data. The parts of the model include: (I) an inflow-outflow transfer function model; (2) a Kalman filter for the lateral flow contribution of the drainage basin; and (3) a second order autoregressive model for the noise component The model preserves the historical mean, skewness, kurtosis, lag-one autocorrelation, and inflow-outflow cross correlation coefficients at the 90 percent confidence level, and the variance at the 95 percent confidence level. Mimikou and Rao (1982) report the model as very efficient in extending daily flow records, estimating missing daily data and flow hydrographs, and for simulating daily reservoir inflows in real time.

Wallis et al. (1991) combine a variant of the hot-deck infilling method, described above, with the mean value infilling method. In this study, the closest three stations are identified, and the missing days are estimated from the closest station with data. If the closest neighboring station does not have data then the data from the next nearest neighboring station is used. If none of the three closest stations have data, then the infilling value is the long-term mean for the appropriate base station and month.

Hirsch (1982) uses a composite method that combines a standard linear equation element and a relationship that uses a normal independent random variable with zero mean and unit variance.

MULTIPLE INFILLING METHODS

As previously stated, standard variance formulae systematically underestimate the variances when single datum-entry infilling methods infill missing data. This is true even if the method generating the single datum-entry infilling values is correct. Multiple infilling methods generate more than one value for a missing datum. Repeated passes are made through the data until convergence is reached (Little and Rubin, 1987). As stated previously, statistical software is available to aid in the use of this infilling method (Statistical Solutions, 1997).

SPLINE FUNCTIONS AND FORECASTING MODELS

Data series may also be fitted exactly, or may be approximated, by spline functions. These functions are not restricted to equally spaced data points, and may be fitted across missing data gaps. Once a spline has been fitted to a series, estimates may be obtained at any desired point (Mizumura, 1985). Since environmental data sets are particularly susceptible to outliers, some limited smoothing may be sensible. This indicates that cubic splines (interpolating splines) or least square splines (approximating splines) may be appropriate (Galpin and Basson, 1990).

Lettenmaier (1980) uses an extension of the Box-Jenkins forecasting technique for estimating missing data. In general, this method and other forecasting methods (e.g., ARMA methods) use data before and after the missing data gap to build an estimate of the missing data. These models, however, can be complex and difficult to implement. Various examples of forecasting models for missing data infilling are in the literature (Carlson et al.. 1970; Fuller and Tsokos, 1971; Galpin

and Bassoon, 1990; Lettenmaier, 1980; Little and Rubin, 1987; McKerchar and Delleur, 1974; McMichael and Hunter, 1972).

Autoregressive models

Generated runoff sequences using these models basically depend on statistical properties, including frequency analysis, of the available historical runoff records. A simple model (Markov) relating the discharge at a certain time t, Q_t and the discharge at earlier time intervals, Q_{t-1} can be expressed by:

$$Q_t = \alpha_o + \alpha_1 Q_{t-1} + \alpha_2 Q_{t-2} + \ldots \alpha_n Q_{t-n} \tag{3.30}$$

where α_0 is a constant, and α_1 , α_2, … and α_n are model parameters to be estimated from the historical record (e.g. Yevjevich, 1972, and Kottegoda & Elgy, 1977).

Extended autoregressive models

This class of models, as a matter of fact, is a special class of the general autoregressive models. In the extended models historical rainfall records are included. The runoff-generating algorithm is

$$Q_t = \beta_o P_t + \beta_1 P_{t-1} + \ldots + \beta_n P_{t-n} + \varepsilon \tag{3.31}$$

where β_1, β_2, …and β_n are model parameters, and P_t is the precipitation at the current time, and P_{t-1},…P_{t-n} are the antecedent precipitations corresponding to 1,..., n time intervals, respectively.

Bibliography

Adams, Thomas R., 1998. Storm Water Facility Design: Calculating the First Flush, Pollution Engineering, Vol. 30, No. 13.

Alley, William M., Alan W. Burns, 1983. Mixed-Station Extension of Monthly Streamflow Records, Journal of Hydraulic Engineering, Vol. 109, No. 10.

Alvo, Mayer, Paul Cabilio, 1994. Rank Test of Trend when Data are Incomplete, Environmetrics, Vol. 5.

American Public Health Association, American Water Works Association, and Water Pollution Control Federation, 1981. Standard Methods for the Examination of Water and Wastewater (15th ed.), American Public Health Association, Washington, DC.

American Society for Quality Control, 1994. Specifications and Guidelines for Quality Systems for Environmental Data Collection and Environmental Technology Programs, ANSI/ASQC E4-1994, Milwaukee, WI.

Angle, J.S., V.A. Bandel, D.B. Beegle, D.R. Bouldin, H.L. Brodie, G.W. Hawkins, L.E. Lanyon, J.R. Miller, W.S. Reid, W.F. Ritter, C.B. Sperow, and R.A. Weismiller, 1986. Best Management Practices for Nutrient Uses in the Chesapeake Basin, Bulletin 308, Extension Services of the Chesapeake Basin, University of Maryland, College Park, MD.

Ator, Scott W., Joel D. Blomquist, John W. Brakebill, Janet M. Denis, Matthew J. Ferrari, Cherie V. Miller, and Humbert Zappia, 1998. Water Quality in the Potomac River Basin, Maryland, Pennsylvania, Virginia, West Virginia, and the District of Columbia, 1992–96, U.S. Geologic Survey Circular 1166, Reston, VA.

Batterman, Stuart A, 1992. Optimal Estimators for Ambient Air Quality Levels, Atmospheric Environment, Vol. 26A, No. 1.
Brankov, Elvira, S. Trivikrama Rao, P. Steven Porter, 1999. Identifying Pollution Source Regions Using Multiply Censored Data, Environmental Science and Technology, Vol. 33, No. 13.
Bryant, C.T., S.F. Shih, 1989. Site vs. Regional Analyses of the Groundwater Flow in Lake Apopka, Florida, Soil and Crop Science Society of Florida Proceedings, Vol. 50.
Canning, Kathie, 1988. Federal Report Sets Direction for TMDL Program, Pollution Engineering, Vol. 30, No. 10.
Carlson, Robert F., A.J.A. MacCormick, Donald G. Watts, 1970. Application of Linear Random Models to Four Annual Streamflow Series, Water Resources Research, Vol. 6, No. 4.
Carpenter, D.H., 1985. Cost Effectiveness of the Federal Stream-Gaging Program in Virginia, U.S. Geologic Survey Water-Resources Investigations Report 85-4345, Towson, MD.
Carpenter, Joanne, 1999. Personal Communication, Occoquan Watershed Monitoring Laboratory, Department of Civil and Environmental Engineering, Virginia Tech, Manassas, VA.
Charbeneau, Randall J., Michael E. Barrett, 1998. Evaluation of Methods for Estimating Stormwater Pollutant Loads, Water Environment Research, Vol. 70, No. 7.
Childress, Carolyn J. Oblinger, Thomas H. Chaney, Donna N. Myers, J. Michael Norris, Janet Hren, 1989. Water-Quality Data-Collection Activities in Colorado and Ohio: Phase II -Evaluation of 1984 Field and Laboratory Quality-Assurance Practices, U.S. Geologic Survey Water-Supply Paper 2295, Reston, VA.
Christensen, Harriet H.; Johnson, Darryl R.; Brookes, Martha H. 1992. Vandalism: research, prevention, and social policy, Gen. Tech. Rep. PNW-GTR-293, U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station Portland, OR.
Clarke, J. U., 1998. Evaluation of Censored Data Methods to Allow Statistical Comparisons among Very Small Samples with Below Detection Limit Observations, Environmental Science and Technology, Vol. 32, No. 1.
Cluis, Daniel A., Pierre Boucher, 1982. Persistence Estimation from a Time-Series Containing Occasional Missing Data, Developments in Water Science, No. 17. Elsevier Scientific Publishing Co., New York, NY.
Cohn, Timothy A., Lewis L. DeLong, Edward J. Gilroy, Robert M. Hirsch, Deborah K. Wells, 1989. Estimating Constituent Loads, Water Resources Research, Vol. 25, No. 5.
Cole, Stephen, 1998. Reclaimed Wastewater Continues Flow Toward Tap, Environmental Science and Technology, Vol. 32, No. 21.
Conn, Judith M., Kung-Jong Lui, Daniel L. McGee, 1989. A Model-Based Approach to the Imputation of Missing Data: Home Injury Incidences, Statistics in Medicine, Vol. 8, No. 3.
Cooke, Dennis G., Robert E. Carlson, 1989. Reservoir Management for Water Quality and THM Precursor Control, American Water Works Research Foundation, Denver, CO.
Cunnane, C., 1978: Unbiased plotting positions – a review. J. Hydrol. 379/4), 205-222.
Dey, B., D.C. Goswami, 1984. Evaluating a Model of Snow Cover Area versus Runoff Against a Concurrent Flow Correlation Model in the Western Himalayas, Nordic Hydrology Vol. 15, No. 2.
Driver, Nancy E., Brent M. Troutman, 1989. Regression Models for Estimating Urban Storm-runoff Quality and Quantity in the United States, Journal of Hydrology, Vol.109.
Driver, Nancy E., Gary D. Tasker, 1988. Techniques for Estimation of Storm-Runoff Loads, Volumes, and Selected Constituent Concentrations in Urban Watersheds in the United States, U.S. Geologic Survey Water-Supply Paper 2363, Reston, VA.
Eckenfelder, W. Wesley, 1989. Industrial Water Pollution Control, McGraw-Hill, Inc., New York, NY.
Engqvist, A., 1990. Accuracy in Material Budget Estimates with Regard to Temporal and Spatial Resolution of Monitored Factors, Estuarine Coastal and Shelf Science, Vol. 30, No. 3.
Fairfax County Water Authority, 1999. Fairfax County Water Authority Web Site, www.fcwa.org, Fairfax County, VA.

Fontaine, R.A., M.E. Moss, J.A. Smith, W.O. Thomas, Jr., 1984. Cost Effectiveness of the Stream-Gaging Program in Maine- A Prototype for Nationwide Implementation, U.S. Geologic Survey Water-Supply Paper 2244, Reston, VA.

Fuller Jr., F.C., Chris P. Tsokos, 1971. Time Series Analysis of Water Pollution Data, Biometrics, Vol. 27.

Galpin, J.S., B. Basson, 1990. Some aspects of analyzing irregularly space time dependent data, South African Journal of Science, Vol. 86, No. 7-10.

Gibbons, Robert D., 1994. Statistical Methods for Groundwater Monitoring, John Wiley & Sons, New York, NY.

Gilbert, Richard O., 1987. Statistical Methods of Environmental Pollution Monitoring, Van Nostrand Reinhold, New York, NY.

Glen, W. Graham, Michael P. Zelenka, Richard C. Graham, 1996. Relating Meteorological Variables and Trends in Motor Vehicle Emissions to Monthly Urban Carbon Monoxide Concentrations, Atmospheric Environment, Vol. 30, No. 24.

Gumbel, E.J., 1941: The return periodof flood flows, Annals of Mathematical Statistics, 12(2), 163-190.

Gyau-Boakye, P., G.A. Schultz, 1994. Filling in runoff time series in West Africa, Hydrologic Sciences Journal, Vol. 39, No. 6.

Harcum, J.B., Jim C. Loftis, Robert C. Ward, 1992. Selecting Trend Test for Water Quality Series with Serial Correlation and Missing Values, Water Resources Bulletin, Vol. 28, No. 3.

Harrison, Jack, Paul Wetherbee, Ann Quenzer, Bob Beduhn, 1999. Developing Tools to Tackle Non-Point Source Pollution, Pollution Engineering, Vol. 31, No.5.

Hazen, A., 1932: Flood flows. John Wiley, New York.

Heidam, Neils Z., 1982. Atmospheric Aerosol Factor Models, Mass and Missing Data, Atmospheric Environment, Vol. 16, No. 8.

Helsel, Dennis R., 1990. Less than Obvious: Statistical Treatment of Data Below the Detection Limit, Environmental Science and Technology, Vol. 24, No. 12.

Hemphill, M.W., James P. Gise, Bruce A. Broberg, 1988. Statewide ozone trends -- Texas, 81st Annual APCA Annual Meeting & Exhibition, Dallas, TX.

Hill, A.R., 1986. Stream Nitrate-N Loads in Relation to Variations in Annual and Seasonal Runoff Regimes, Water Resources Bulletin, Vol. 22, No. 5.

Hirsch, Robert M., 1979. An Evaluation of Some Record Reconstruction Techniques, Water Resources Research, Vol. 15, No. 6.

Hirsch, Robert M., 1982. A Comparison of Four Streamflow Record Extension Techniques, Water Resources Research, Vol. 18, No.4.

Hirsch, Robert M., James R. Slack, and Richard A. Smith, 1982. Techniques of Trend Analysis for Monthly Water Quality Data, Water Resources Research, Vol. 18, No. 1.

Huth, Steve, 1999. The Second Shift: A Florida Water Reclamation Project Benefits Local Government and Businesses Alike, Environmental Protection, Vol. 10, No.7, Stevens Publishing, Dallas, TX.

Jarrell, Wesley M., 1999. Getting Started with TMDLs, YSI Incorporated, Yellow Springs, OH.

Jeffcoat, Hillary H., 1987. Cost Effectiveness of the U.S. Geological Survey Stream-Gaging Program in Alabama, U.S. Geologic Survey Water-Resources Investigations Report 86-4336, Tuscaloosa, AL.

Johnson, C.A. (1999) Data infilling techniques: development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data Unpublished MSc thesis of Virginia Polytechnic Institute and State University.

Koch, Roy W., Gary M. Smillie, 1986. Bias in Hydrologic Prediction Using Log-Transformed Regression Models, Water Resources Bulletin, Vol. 22, No. 5.

Kottegoda, N.T., J. Elgy, 1979. Infilling Missing Data, Modeling Hydrologic Processes, Edited by H.J. Morel-Seytoux et al., Proceedings Fort Collings 3rd International Hydrologic Symposium on Theoretical and Applied Hydrology, Colorado State University, Fort Collins, CO.

Krug, William R., Warren A. Gebert, and David J. Graczyk, 1990. Preparation of Average Annual Runoff Map of the United States, 1951-80, U.S. Geologic Survey Open-File Report 87-535, Denver, CO.
Laufer, Susan Marie, 1986. Nutrient Dynamics in the Lake Manassas (Virginia) Watershed, M.S. Thesis, Virginia Tech, Blacksburg, VA.
Leitch, Katherine McArthur, 1998. Estimating Tributary Phosphorus Loads Using Flow-Weighted Composite Storm Sampling, M.S. Thesis, Virginia Tech, Blacksburg, VA.
Lettenmaier, Dennis P., 1978. Design Considerations for Ambient Stream Quality Monitoring, Water Resources Bulletin, Vol. 14, No.4.
Lettenmaier, Dennis P., 1980. Intervention Analysis with Missing Data, Water Resources Research, Vol. 16, No. 1.
Little, Roderick J.A., Donald B. Rubin, 1987. Statistical Analysis with Missing Data, John Wiley & Sons, New York, NY.
Loague, Keith, Dennis L. Corwin, Timothy R. Ellsworth, 1998. The Challenge of Predicting Nonpoint Source Pollution, Environmental Science and Technology, Vol. 32, No. 5.
Lui, Lon-Mu, Chung Chen, 1991. Recent Developments of Time Series Analysis in Environmental Impact Studies, Journal of Science and Health, Part A, Vol. 26, No. 7.
McKerchar, A.I., J.W. Delleur, 1974. Application of Seasonal Parametric Linear Stochastic Models to Monthly Flow Data, Water Resources Research, Vol. 10, No. 2.
McMichael, Francis Clay, J. Stuart Hunter, 1972. Stochastic Modeling of Temperature and Flow in Rivers, Water Resources Research, Vol. 8, No. 1.
Mamdouch, S. (2002). Hydrology and water resources of Africa, Water science and technology library 41, Kluwer Academic Publishers, Dordrecht, pp.256-258
Martenson, I.V., 1982. Problems of Reliability of Mechanical Equipment of Hydraulic Structures, Hydrotechnical Construction, Vol. 16, No. 2.
Melcher, N.B., J.F. Walker, Evaluation of Selected Methods for Determining Streamflow During Periods of Ice Effect, U.S. Geologic Survey Open-File Report 90-554, Denver, CO.
Mimikou, A., A. Ramachandra Rao, 1982. Rainfall-Runoff Model for Daily Flow Synthesis, Developments in Water Science, No. 17, Elsevier Scientific Publishing Co., NY.
Mizumura, Kazumasa, 1985. Estimation of Hydraulic Data by Spline Functions, Journal of Hydraulic Engineering, Vol. 111, No. 9.
Neter, John, William Wasserman, Michael H. Kutner, 1985. Applied Linear Statistical Models, Irwin Publishing Company, Homewood, IL.
Occoquan Watershed Monitoring Laboratory, 1998. An Updated Water Quality Assessment for the Occoquan Reservoir and Tributary Watershed (1973-1997), Department of Civil and Environmental Engineering, Virginia Tech, Manassas, VA.
Occoquan Watershed Monitoring Laboratory, 1997. Occoquan Watershed Monitoring Laboratory Standard Operating Procedures Part II: Field, Department of Civil and Environmental Engineering, Virginia Tech, Manassas, VA.
Occoquan Watershed Monitoring Laboratory, 1994. A Water Quality Assessment for the Occoquan Reservoir (1972-1993), Department of Civil and Environmental Engineering, Virginia Tech, Manassas, VA.
Parrett, Charles, Dave R. Johnson, 1994. Estimates of Monthly Streamflow Characteristics and Dominant-Discharge Hydrographs for Selected Sites in the Lower Missouri and Little Missouri River Basins in Montana, U.S. Geologic Survey Water-Resources Investigation Report 94-4098, Denver, CO.
Pontius, Frederick W., 1990. Water Quality and Treatment (4th Edition), American Water Works Association, Washington, DC.
Post, Harry, 1999. Personal Communication, Occoquan Watershed Monitoring Laboratory, Department of Civil and Environmental Engineering, Virginia Tech, Manassas, VA.

Randall, C.W., T.J. Grizzard, R.C. Hoehn, 1978. The Effect of Upstream Control Measures on a Water Supply Reservoir, Journal of the Water Pollution Control Federation, Vol. 46.
Rantz, et al., 1982. Measurement and Computation of Streamflow, U.S. Geologic Survey Water- Supply Paper 2175, Reston, VA.
Reinelt, Lorin E., Anders Grimvall, 1992. Estimation of Nonpoint Source Loadings with Data Obtained from Limited Sampling Programs, Environmental Monitoring and Assessment, Vol. 21, No.3.
Shields, F. Douglas Jr, Thomas G. Sanders, 1986. Water Quality Effects of Excavation and Diversion, Journal of Environmental Engineering, Vol. 112, No. 2.
Shih, S.F., K.S. Cheng, 1989. Generation of Synthetic and Missing Climatic Data for Puerto Rico, Water Resources Bulletin, Vol. 25, No. 4.
Shih, G., W. Abtew, J. Obeysekera, 1994. Accuracy of Nutrient Runoff Load Calculations Using Time-Composite Sampling, Transactions of the American Society of Agricultural Engineers, Vol. 37, No. 2.
Shirmohammadi, A., K.S. Yoon, W.L. Magette, 1997. Water Quality in Mixed Land-Use Watershed - Piedmont Region in Maryland, American Society of Agricultural Engineers, Vol. 40, No. 6.
Smith, Richard A., Robert M. Hirsch, James R. Slack, 1982. A Study of Trends in Total Phosphorus Measurements at NASQAN Stations, U.S. Geologic Survey Water-Supply Paper 2190, Reston, VA.
Snijders, Tom A., 1986. Interstation Correlations and Nonstationarity of Burkina Faso Rainfall, Journal of Climate and Applied Meteorology, Vol. 25, No. 4.
Sokolov, A.A., B.S. Eichert, J. Kindler, G.A. Schultz, 1982. Methods of Hydrological Computations for Water Projects, United Nations Educational Scientific and Cultural Organization, Paris, France.
Stack, William P., Kenneth T. Belt, 1989. The Selection of Appropriate Flow Averaging Periods in Evaluating Pollutant Loadings Using the Flow Interval Method, Lake and Reservoir Management, Vol. 5, No. 2.
Statistical Solutions, 1997. SOLAS for Missing Data Analysis, Manual for Version 1.0, Stonehill Corporate Center, Suite 104, 999 Broadway, Saugus, MA 01906.
Troutman, Brent M., 1985. Errors and Parameter Estimation in Precipitation-Runoff Modeling: 1. Theory, Water Resources Research, Vol. 21, No. 8.
U.S. Environmental Protection Agency, 1982. Handbook for Sampling and Sample Preservation of Water and Wastewater, EPA-600-4-82-029, Washington, DC.
U.S. Environmental Protection Agency, 1990. Monitoring Lake and Reservoir Restoration, EPA-440-4-90-007, Washington, DC.
U.S. Environmental Protection Agency, 1997a. Section 319 Success Stories: Volume II -Highlights of State and Tribal Nonpoint Source Programs, EPA-841-R-97-001, Washington, DC.
U.S. Environmental Protection Agency, 1997b. Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls, EPA-841-B-96-004, Washington, DC.
U.S. Environmental Protection Agency, 1997c. Compendium of Tools for Watershed Assessment and TMDL Development, EPA-841-B-97-006, Washington, DC.
U.S. Environmental Protection Agency, 1997d. Linear Regression for Nonpoint Source Pollution
Analyses, EPA-841-B-97-007, Washington, DC.
U.S. Environmental Protection Agency. 1998a, Clean Water Action Plan: Restoring And Protecting America's Waters, www.cleanwater.gov, Washington, DC.
U.S. Environmental Protection Agency. 1998b. Report of the Federal Advisory Committee on the Total Maximum Daily Load (TMDL) Program, EPA 100-R-98-006,Washington, DC.
U.S. Environmental Protection Agency. 1998c. Estimation of Infiltration Rate in the Vadose Zone: Volume I & II, EPA-600-R-97-128 A&B, Washington, DC.
U.S. General Accounting Office, 1999. Water Quality: Federal Role in Addressing and Contributing to Nonpoint Source Pollution, GAO/RCED-99-45, Washington, DC.
U.S. Geologic Survey, 1998. A New Evaluation of the USGS Streamgaging Network, A Report to Congress, November 30, 1998, Washington, DC.

U.S. Geologic Survey, 1991. Water Resources Data, Florida, Water Year 1991, Vol. 2A. South Florida Surface Water, Reston, VA.

USWRC, 1967: A uniform technique for determining flood flow frequencies, Bulletin 15, Hydrology Committee, Water resources Council, Washington.

Virginia State Water Control Board, 1971. A Policy for Waste Treatment and Water Quality Management in the Occoquan Watershed, Virginia State Water Control Board, Richmond, VA.

Wahl, Kenneth L., Wilbert O. Thomas, Jr., and Robert M. Hirsch, 1995. The Stream-Gaging Program of the U.S. Geologic Survey, U.S. Geologic Survey Circular 1123, Reston, VA.

Wallis, J.R., D.P. Lettenmaier, E.F. Wood, 1991. A Daily Hydroclimatological Data Set for the Continental United States, Water Resources Research, Vol. 27, No. 7.

Ward, Richard C., Jim C. Loftis, Knud S. Nielsen, R. Dennis Anderson, 1979. Statistical Evaluation of Sampling Frequencies in Monitoring Networks, Journal of the Water Pollution Control Federation, Vol. 51, No. 9.

Werblow, Steve, 1999. TMDL Rules Drive Water Monitoring, WaterWorld, PennWellPublishing, Tulsa, OK, Vol. 15, No. 1.

Whipple, William, Neil S. Grigg, Thomas Grizzard, Clifford W. Randall, Robert P. Shubinski, L.

Scott Tucker, 1983, Stormwater Management in Urbanizing Areas, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Wyka, Theodore A., 1995. Operational Assessment of the Occoquan Watershed Monitoring Laboratory, M.S. Thesis, Virginia Tech, Falls Church, VA.

Chapter 4:
Regionalisation techniques for water resources assessment

Source:
BP Parida Regionalisation techniques for water resources assessment

Recommended reading:
Regional Frequency Analysis: An Approach Based On L-Moments, J.R.M. Hosking and J.R.Wallis, Cambridge University Press, Cambridge.(1997)
Handbook of Hydrology,D.R. Maidment, McGraw Hill Inc., New York., (1993)
Flood Studies Report, Vol.1, Natural Environmental Research Council, UK (1975)

Objective: After this day, all participants understand the concept of regionalisation and its use in prediction of flood quantiles and/or drought characteristics for ungauged basins and be able to apply the methods on real life data.

4.1 GENERAL INTRODUCTION

Extreme environmental events, such as floods, droughts, rain storms, and high winds, have severe consequences for human society. How frequently an event of a given magnitude may be expected to occur is of great importance. Planning for weather-related emergencies, design of civil engineering structures, reservoir management, pollution control, and insurance risk calculations, all rely on knowledge of the frequency of these extreme events. Estimation of these frequencies can be done easily using the observed records of annual maximum or annual minimum flows and the procedures given in Chapter 3, but is often difficult because extreme events are by definition rare and data records are often short.
This therefore would require knowledge of some robust procedures through which augmentation of data could be done and on which appropriate estimation techniques could be used such that more meaningful and results could be obtained for decision making.

4.2 UNGAUGED BASINS?

A basin where no flows are recorded is generally known as an ungauged basin. In such situations, the basic problem of finding relationship between flood or drought magnitudes and return period at an ungauged site thus becomes complex or difficult. The most desperate situation arises when no records are available anywhere in the basin.
There are also other categories of ungauged basins e.g. sites on a river which is gauged different locations upstream or downstream or gauged on some tributaries rather than the main river. Even sites which have only few years of record or have half hazard records has to be dealt with in some aspects, as if it were an ungauged basin because information contained in it has to be augmented.

4.3 REGIONAL FREQUENCY ANALYSIS

It has been observed that at-site frequency analysis methods mostly use long records failing which the flood statistics such as Coefficient of Variation (Cv), Coefficient of Skewness (Cs) and Coefficient of Kurtosis (Ck) obtained from small samples can have large standard errors and downward bias, thereby yielding unreliable flood quantiles. However, a large numbers of rivers and streams in developing countries have either typically small length of record or even none.

Regional frequency analysis resolves such problems by 'trading space for time' through which data from several sites are used in estimating event frequencies at any one site. In other words, the main objective of the Regional Frequency Analysis (RFA) is to augment the available record at a gauging site by way of transferring information from adjacent basins within the region.

Regional analysis therefore consists of analyzing the hydrometric records of all gauged sites in a region, summarizing each record by one or two statistical values calculated from it and then finding relationship between these statistic values and numerically expressed basin characteristics. The mean, Xb, of the annual maximum flood series is nearly always used as one of these statistics. Relations between Xb and basin characteristics are dealt with separately. The other statistics commonly used is the growth factor X_T/Xb which can be plotted graphically against return period T to obtain a regional growth curve. Since the plot of T is not linear in the x-axis, it is usual to use the Gumbel reduced variate (y) equivalent of T as given in Table 3.1, instead to keep the scale of x-axis linear. Atypical plot of growth curve for two different regions of Botswana is given in Fig. 4.1 below.

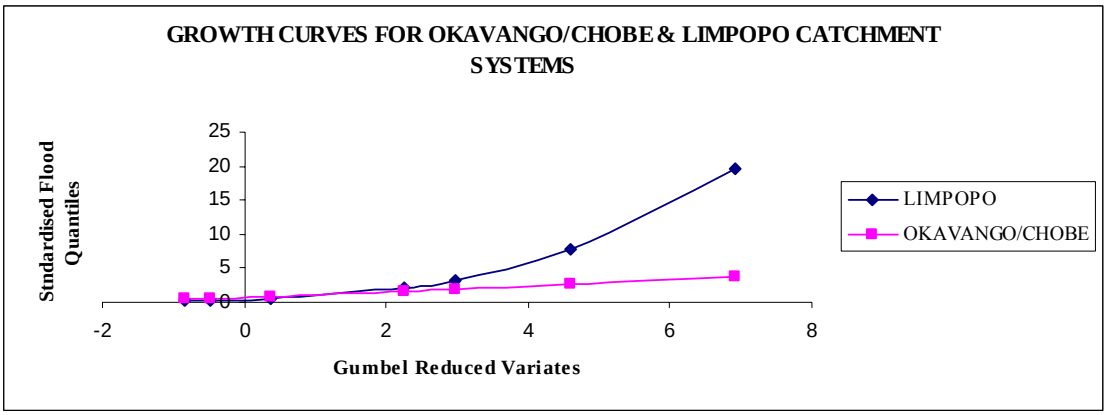


Fig. 4.1 Typical plot of growth curves for the Limpopo and Okavango/Chobe Regions of Botswana

Please observe: The growth curve for Limpopo catchment systems which is regarded as one of the homogeneous regions, shows a sharply rising growth curve compared to the growth curve for Okavango/Chobe region. The former indicates that the flood quantiles at higher return periods are considerably higher, whereas for the Okavango/Chobe region the flood quantiles change marginally even at higher return periods even when T=100 years (or Gumbel y = 4.6).

It is obvious that, success of RFA would primarily depend upon

(i) identification of a homogeneous region;
(ii) use of appropriate method of aggregation of data (through identification of a common distribution); and
(iii) disaggregation of the regional information to the site of interest.

It has been observed that common meteorological phenomena in a region bring out similar flood responses at different gauging sites. Magnitude of these flood responses, however, gets modified depending upon the physical, geo-morphological and man-made changes in the basin. As a result no two sites are likely to get exactly similar flood responses and flood distribution. Further, the generally accepted procedure of smoothening of flood data using linear interpolation across the region is not possible due to temporal and spatial variation of flood quantiles. It is therefore presently not possible to define homogeneity of the region in form of say similar flood magnitudes or magnitudes defined through an average value plus a random component. However, considering the extent of a region, which is of the order of hundreds of square kilometres, it may be possible to identify some dimensionless statistical parameters to group flood series. The most common of these parameters are Cv or Cs observed at different gauging sites. The spatial variation of these statistical features may represent the behaviour of the region. Also, the regional skew maps and regional Cv may be used to obtain a regional smoothened value of these parameters in terms of averages (Nash and Shaw, 1966 and USWRC, 1979)

Other forms of regional smoothening can be achieved by non-dimensionalising flood quantiles by dividing the quantile estimates by an index flood such as the mean annual flood suggested by Darlymple (1960) which suggest that average of flood ratios ($X_{10,i}/Xb_i$) evaluated at all i stations could be considered as an index of homogeneity. Although there are many forms, basically the concept of regional homogeneity, which is a form of regional-smoothening, has more or less remained around the averaged value of dimensionless statistical parameters, across a geographical boundary.

Though geographical regions can assist in assigning an ungauged site to a region, doubts often are raised on such a consideration because of a likely high cross-correlation between the gauged stations which may result in loss of precision in regional flood quantile estimates. Also, as had been said earlier, values of Cs computed at stations with small record length, can often provide misleading values unrepresentative of the population. To this effect, Lettenmaier (1985) and Lettenmaier et al., (1987) have also shown that approaches which assume regional homogeneity in moments higher than order one (such as Cv) are sensitive to record length. Though many studies have advocated use of clustering techniques in various forms, decisions through them may not entirely eliminate subjective decisions though it greatly facilitates interpretation of a data set.

Of late, it has been shown that use of L-Moments can greatly overcome the estimation of statistical characteristics of gauged flood records even if they were small in sample size or were possibly infested with outliers. Considering these, Hosking and Wallis (1997) have suggested a method of determining the Homogeneity Index through simulation of 500 regions with identical

sample size of each station, to compute H-statistics, hence to identify the state of heterogeneity of the chosen region. As per Hosking and Wallis (1997), if the value of H was less than or equal to 1 then the region is said to be connoted as 'acceptably homogeneous', or as 'possibly homogeneous' if the H value was between 1 and 2. The H value greater than 2 is termed as 'definitely heterogeneous'.

However, it should also be remembered that, increased partitioning to achieve more homogeneity may not be preferable and has been shown that slight heterogeneity may not negate the overall results and also the overall philosophy of regionalization (Lettenmaier et al., 1985; Cunnane, 1988). In view of these, a region which shows H values up to 2 can always be accepted as a homogeneous region, while regions with H value greater than 2 can be termed as heterogeneous. The detailed procedure for computation of H-values has been dealt later in this Chapter.

4.4 METHODS FOR AGGREGATION OF DATA

Once a homogeneous region is identified, next step in RFA is to aggregate the data available at all gauging stations. In other words, the objective is to represent the flood behaviour of the region through a common relationship called the growth curve or through a common set of parameters. This is possible only when the data from all sites are more or less of the same magnitude, which therefore necessitates the scaling/standardisation of record at individual gauging stations.
Cunnane (1988) has brought out an exhaustive list of RFA methods some of which are given below:

i. Station Year Method
ii. Index Flood Method
iii. Method using Linear Multiple Regression Technique

Besides these, some other methods such as suggested by Natural Environmental Research Council (NERC) U.K., United States Water Resources Council (USWRC) and use of standardised Probability Weighted Moments are also in wide use, particularly the later one whose concept has been used in the Method of L-Moments which will be discussed later. The method by NERC though is an innovative method to take into account some historical observations in the growth curve, it has some of subjective ness particularly while making the graphical fit and may therefore yield unreliable quantiles at higher T. The method is greatly influenced by small sample effect and may yield biased quantiles even under favourable conditions. Similarly, as the method suggested by USWRC greatly depends upon the use of a country wide reliable skew map may pose problems for the developing countries where data are scarce.

4.4.1. Station year Method:

The method is based on pooling of standardised data [(XS)j = Xj/Xb] from M stations with N(.) years of record at each station (.) i.e. [$(XS)_{ij}$, i = 1,…., N_j ; j = 1,…., M] of a total length NL to form one sample which is used to develop XS~ T relationship using a suitable distribution.

The main shortcoming of this method is that, it does not take account inter-site dependence of data; hence it can be used for stations with low inter-site dependence. In other words, the cross-correlation between the flood values at any two sites should not be significantly high (for example, it should not be more than say 0.6).

4.4.2. Index-Flood Method

This method suggested by Darlymple (1960) is a classical and an extensively used method. This method uses equal lengths of unregulated record of N_j years from all available sites in the region. An assumed EV-1 (Gumbel) distribution is fitted to these data at each site from which index flood, in this case $X_{2.33}$, is computed. Ration of quantiles at T=10 years with that of the index flood is used to test regional homogeneity. For the stations passing homogeneity test, flood quantiles at other return periods are computed and scaled using the index flood. Median of all such scaled values are at different T are joined by a smoothly guided curve to obtain the Regional Growth Curve.

Some shortcomings of this method are:

i. The method is based on homogeneity test at a 10-year (T) return period level.
ii. The growth curve (XS ~T) relationship is normally considered valid for return period up to three times the length of record i.e. if the data base period is say 15-16 years, then quantile estimates beyond T=50 years may not be reliable.

4.4.3. Method using Linear Multiple Regression Technique:

A third approach which is the most attractive has been attempted by Nash and Shaw (1966). In this it is assumed that the entire statistical properties of the annual maximum series can be described by say a two parameter distribution such as EV-1 (Gumbel), which in turn can be defined by its first two moments, such as mean and standard deviation. The magnitude of mean is well known to be related to basin area, but standard deviation is also numerically dependent on basin area. Hence, mean and standard deviation are highly correlated and two separate relations between them and basin characteristics could not be regarded as determining two quantities independently from basin characteristics. Therefore, they used the dimensionless standard deviation i.e. Cv = mean / std. deviation, as the second variable. In some cases, instead of Cv, the basin characteristics could even be related to the standardised quantiles ($XS_T = X_T/Xb$) at specified recurrence intervals.

$$Xb = c.\ A^a\ .\ S^s\ .\ R^r\ .\ SF^f \qquad ()$$

or

$$XS_T = c.\ A^a\ .\ S^s\ .\ R^r\ .\ SF^f \qquad (4.2)$$

where, A is Basin area in km^2;. SF is stream frequency (Number of stream junctions divided by A); S is channel slope in parts per thousand, R is the annual rainfall in mm. and a,s,r,f,c represent the coefficients (determined by regression).

First, the equations relating the above explained parameters separately to the basin characteristics are made linear by taking logs of the variables. Then a multiple regression is undertaken using the logs of basin characteristics as independent variables and the logs of mean or coefficient of variation as the dependent variable. The final equation with the appropriate choice of the basin characteristics is decided on the factor which gives high values of coefficient of determination and minimum standard error of the coefficients (for e.g. c, a, s, r and f in the above equations.).

Success of this method will depend when the number of stations used is large.

4.5 THE METHOD OF L-MOMENTS

L-moments are a recent development within statistics. They form the basis of an elegant mathematical theory in their own right, and can be used to facilitate the estimation process in regional frequency analysis. *L*-moment methods are demonstrably superior to those that have been used previously, and are now being adopted by major organizations worldwide.

L-moments are summary statistics for probability distributions and data samples. They are analogous to ordinary moments -- they provide measures of location, dispersion, skewness, kurtosis, and other aspects of the shape of probability distributions or data samples -- but are computed from linear combinations of the ordered data values (hence the prefix *L*).
L-moments have the following theoretical advantages over ordinary moments:

- For *L*-moments of a probability distribution to be meaningful, we require only that the distribution have finite mean; no higher-order moments need be finite [Hosking (1990), Theorem 1].
- For standard errors of *L*-moments to be finite, we require only that the distribution have finite variance; no higher-order moments need be finite [Hosking, 1990, Theorem 3].
- Although moment ratios can be arbitrarily large, sample moment ratios have algebraic bounds; sample *L*-moment ratios can take any values that the corresponding population quantities can [Hosking, 1990, page 115].

In addition, the following properties hold in a wide range of practical situations:

- Asymptotic approximations to sampling distributions are better for *L*-moments than for ordinary moments [Hosking, 1990, Figure 4].
- *L*-moments are less sensitive to outlying data values [Royston, (1992), Figure 7; Vogel and Fennessey, (1993), Figures 3 and 4].
- *L*-moments provide better identification of the parent distribution that generated a particular data sample [Hosking, 1990, Figure 6].

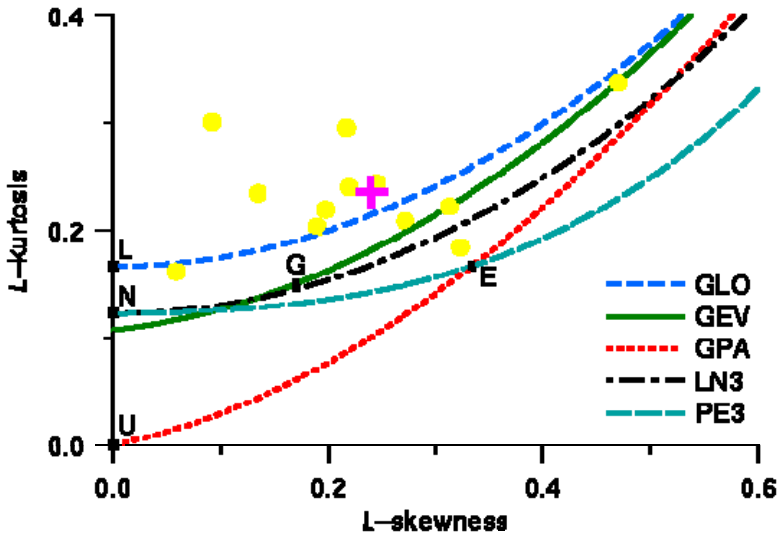


Fig. 4.2 Shows the L-Moment diagram with type curves from different 3-P distributions (Hosking and Wallis, 1997)

4.5.1 L-moments for data samples

Probability weighted moments, defined by Greenwood et al. (1979), are precursors of *L*-moments. Sample probability weighted moments, computed from data values X_1, X_2, ... X_n, arranged in increasing order, are given by

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

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

(4.3)

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

$$\ell_1 = b_0 ,$$
$$\ell_2 = 2b_1 - b_0 ,$$
$$\ell_3 = 6b_2 - 6b_1 + b_0 ,$$
$$\ell_4 = 20b_3 - 30b_2 + 12b_1 - b_0$$

(4.4)

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

The first *L*-moment is the sample mean, a measure of location. The second *L*-moment is (a multiple of) Gini's mean difference statistic, a measure of the dispersion of the data values about their mean. By dividing the higher-order *L*-moments by the dispersion measure, we obtain the *L*-moment ratios,

$$t_r = \ell_r / \ell_2 . \tag{4.5}$$

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

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

$$t = \ell_2 / \ell_1 . \tag{4.6}$$

It takes values between 0 and 1.

4.5.2. L-moments for probability distributions

For a probability distribution with cumulative distribution function $F(x)$, probability weighted moments are defined by

$$\beta_r = \int x \{F(x)\}^r \, dF(x), \qquad r = 0, 1, 2, \ldots . \tag{4.7}$$

L-moments are defined in terms of probability weighted moments, analogously to the sample *L*-moments:

$$\begin{aligned} \lambda_1 &= \beta_0 , \\ \lambda_2 &= 2\beta_1 - \beta_0 , \\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0 , \\ \lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 . \end{aligned} \tag{4.8}$$

L-moment ratios are defined by

$$\tau_r = \lambda_r / \lambda_2 . \tag{4.9}$$

For Example:

Normal distribution with mean 0 and variance 1 will have the first two *L* –moments and the *L*-skewness and *L*-kurtosis respectively as:

$$\lambda_1 = 0, \quad \lambda_2 = 1/\sqrt{\pi}, \quad \tau_3 = 0, \quad \tau_4 \approx 0.123.$$

The L-moment ratio diagram can be used to compare the L-skewness--L-kurtosis relations of different distributions and data samples.

4.5.3 Approximations for use in constructing L-moment ratio diagrams

To construct an *L*-moment ratio diagram it is convenient to have simple explicit expressions for `tau_4` (*L*-kurtosis) in terms of `tau_3` (*L*-skewness) for some commonly used probability distribtions. Polynomial approximations of the form (Hosking, 1991; Hosking and Wallis, 1997)

$$tau\_4 = A0 + A1 * (tau\_3) + A2 * (tau\_3)^2 + A3*(tau\_3)^3 + ... + A8 * (tau\_3)^8 \qquad (4.10)$$

have been obtained, and the coefficients are given in the table below. "Overall lower bound" is the lower bound on `tau_4` for all distributions [Hosking, *J. R. Statist. Soc. B*, 1990, eq.(2.7)]. For given `tau_3`, the approximations yield values of `tau_4` that are accurate to within 0.0005 provided that `tau_3` is in the range -0.9 to +0.9, except that for the generalized extreme-value distribution 0.0005 accuracy is attained only when `tau_3` is between -0.6 and +0.9,

Table 4.2 Coefficients for construction of type curves for different distributions in L-Moments diagram

	Generalized logistic	Generalized extreme-value	Generalized Pareto	Lognormal	Pearson type III	Overall lower bound
A0	0.16667	0.10701	0.	0.12282	0.12240	-0.25
A1	.	0.11090	0.20196	.	.	.
A2	0.83333	0.84838	0.95924	0.77518	0.30115	1.25
A3	.	-0.06669	-0.20096	.	.	.
A4	.	0.00567	0.04061	0.12279	0.95812	.
A5	.	-0.04208	.	.	.	.
A6	.	0.03673	.	-0.13638	-0.57488	.
A7	.	.	.	.	.	.
A8	.	.	.	0.11368	0.19383	.

The approximations are not intended for detailed analytical calculations -- for that purpose, use the routines in the LMOMENTS software package -- but they are sufficiently accurate for use in plotting theoretical *L*-moment relationships on an *L*-moment ratio diagram

4.6 PROCEDURE FOR INDEX-FLOOD METHOD

The data sets of annual maximum floods for each gauging station within a region are first screened and stations having less than five years and those which are on regulated rivers are excluded.

Of the remaining stations, those having missing years have values filled in from neighbouring streams by regression methods and if possible all the data sets used after being filled in or extended have the same length. The filled in values are not used explicitly any further but they serve to ensure that values have their proper rank within the common base period.

Each set of data are then plotted on a Gumbel plot, a smooth curve drawn through the points and the ordinates of X read off at T= 2.33. This graphical estimate of X is preferred in this case to the arithmetic mean, because the effect of outliers is reduced while the correction to a common base period may help to remove other deficiencies.

Secondly a homogeneity test is carried out based on station estimates of X_{10}. Stations whose X_{10} values plot outside the pair of control curves on a special plot, are regarded as not belonging to the homogeneous group and are from then on excluded.

Table 4.2 Upper and lower values of Gumbel y_T for plotting upper and lower bounds of the control curve

Sample Size	Lower y_T Limit	Upper y_T Limit
5	-0.59	5.09
10	0.25	4.25
20	0.83	3.67
50	1.35	3.15
100	1.52	2.88

Thirdly, for each station which remains, the ratios X_T/Xb for T=2, 5, 10, 50 are obtained. For each return period, the median value of X_T/Xb is taken and plotted against T on Gumbel paper (or even on a plain graph paper using Gumbel y) and a smooth curve is drawn though the points.

Finally, a relationship between Xb and drainage area A is obtained graphically on a log-log plot, or Xb is related to area and other basin characteristics by multiple (logarithmic) regression.

4.7 PROCEDURE FOR USING MULTIPLE LINEAR REGRESSION TECHNIQUE WITH BASIN CHARACTERISTICS

Basin Characteristics
The physical characteristics of a basin may be grouped loosely under a number of general headings

(i) Size and Shape
(ii) Density and distribution of streams
(iii) Overland and channel slope
(iv) Basin storage
(v) Soil /Geology
(vi) Rainfall / Climate

Items (i) and (vi) above determine the scale of the hydrological process while (ii)-(v) modify it in several ways. Definition of numerical measures or indices which are completely effective in

explaining variations in flood response of different basins may not be possible. Although many indices have been described by various authors, in any one application at most six or seven can be of value.

The indices mentioned above are:

(i) Basin area, A in km^2 ; Main stream Length, L in km.

(ii) Stream Frequency, SF (Number of stream junctions divided by A); Another measure could be Drainage Density = Total length of streams/A

(iii) Channel slope or S_{1085} in parts per thousand

(iv) Basin Storage is represented by either (a) the proportion of basin area occupied by lakes, ponds and swamps

(v) Soils and geology are difficult to express in a single numerical figure for flood predicting purposes. But some kind of runoff coefficients can be assigned to a typical type of land use to separate between them. It could account soil type, slope and drainage, typical depth to water table, underlying geology and other known drainage characteristics.

(vi) The most frequently used indicator of climate and rainfall for flood studies is that of average annual rainfall.

For example, if it assumed that the mean annual flood Xb is related to basin characteristics such as Area (km^2) , mean overland slope S (parts per 10000) and average annual rainfall over the basin R (mm), then the model in its non-linear form can be written as:

$$Xb = c. A^a. R^r .S^s \qquad (4.11)$$

The constants or coefficients c, a, r and s in the above equation can be determined using regression analysis. Taking logarithms of both the sides of the equation, the model can be written in its linear form as:

$$\text{Log (Xb)} = \text{Log c} + a\ \text{Log (A)} + r\ \text{Log (R)} + s.\ \text{Log (S)} \qquad (4.12)$$

Using the technique of multiple regression, several sets of regression can be performed thus:

Log Xb on Log A
Log Xb on Log R
Log Xb on Log S
Log Xb on Log A and Log S
Log Xb on Log A and Log R
Log Xb on Log R and Log S
Log Xb on Log A , Log S and Log R

Based on these, a correlation table between the logs of variables (Basin Characteristics) is then constructed which will then be used to adjudge the true independentness of the variables, hence their possibility for inclusion in the regression equation.

Out of the several plausible regression alternatives as listed above, the one which yields the highest co-efficient of determination, has minimum standard error of its coefficients and has a F –statistics greater than the critical value, is chosen. Some times with smaller number of stations, one is inclined to reasonably accept the relationship or the model, even if two of the aforesaid criteria are satisfied.

For example using data from 42 basins in Botswana and South Africa, Faruharson et al, (1992) developed the following relationship between Mean Annual Flow (MAF) and the Area and Mean Annual Precipitation (MAP) as

$MAF = 106.7 . Area^{0.374} MAP^{-0.396}$

4.8 PROCEDURE FOR USING L-MOMENT TECHNIQUE

First the observed annual flood series at each of the gauging site is ranked in the ascending order of magnitude. Then either using the Equations (4.3) and (4.4) together or the Equations (4.13-4.16) (Maidment, 1993) on the observed ranked data, computation of four Linear moments (L-Moments) λ_1, λ_2 , λ_3 and λ_4 are carried out. These are then used to compute three dimensionless L-moments as given in the Eqns (4.17)-(4.19) below.

$\lambda_1 = E[x]$ (4.13)

$\lambda_2 = (1/2) E[x_{(1:2)} - x_{(2:2)}]$ (4.14)

$\lambda_3 = (1/3) E[x_{(1:3)} - 2x_{(2:3)} + x_{(3:3)}]$ (4.15)

$\lambda_4 = (1/4) E[x_{(1:4)} - 3x_{(2:4)} - 3x_{(3:4)} - x_{(4:4)}]$ (4.16)

where, λ_1, λ_2 , λ_3 and λ_4 respectively represent the parameters related to location, scale, shape and peakedness. The connotations (1:2) and (2:2) in the Eqn. (4.14), means the first and second large value respectively in a sample of size two drawn from the entire data observed at a station. In a similar manner appropriate connotations can be used in other equations.

Then the dimensionless L-moments are,

L-Coeff. Of Variation (L-Cv), $\tau_2 = \lambda_2/\lambda_1$ (4.17)

L-Coeff. Of Skewness (L-Sk), $\tau_3 = \lambda_3/\lambda_2$ (4.18)

L-Coeff. Of Kurtosis (L-Ku), $\tau_4 = \lambda_4/\lambda_2$ (4.19)

Again, using the aforesaid L-Moment/Index Flood procedure, discordant sites whose at-site L-moments are markedly different from other sites if any, are identified and removed from analyses. Then through a simulation exercise (Hosking, 1994; Hosking and Wallis, 1997), the sites are tested for their homogeneity (H). This can be established either through the use of either

L-Cv or (L-Cv/L-Kutosis) or (L-Skew/L-Kurtosis) values to represent V in Eqn.(4.20) below. For example, if Cv is considered on its own then H can be computed from:

$$H = (V - \mu_V) / \sigma_V \qquad (4.20)$$

where,
V = weighted (standard deviation) of τ_2 values
μ_V, σ_V =the mean and standard deviation of N_{sim} values (=500) of V

As has been said earlier, depending on the values of H, i.e. either less than 1, or between 1 and 2, or greater than 2, a region containing a group of stations, can be connoted as, 'acceptably homogeneous', 'possibly homogeneous', or 'definitely heterogeneous'.

Finally, the goodness-of fit measure is computed from:

$$Z^{Dist} = \frac{\tau_4^{Dist} - \bar{t}_4 + \beta_4}{\sigma_4} \qquad (4.21)$$

where, $\bar{t}_4$ = average L-Kurt value computed from the data of the region
β_4 = bias in L-Kurt values ,t_4, computed from the original data.
τ_4^{Dist} = average of L-Kurt value computed from simulation of a fitted distribution.
σ_4 = standard deviation of L-Kurt values obtained from simulated data.

All distributions whose absolute Z value is less than or close to 1.64 in the L-Moment diagram qualified as a possible candidate. Since, the exercise involves a good amount of computation, it can be achieved through use of the L-Moments software developed by IBM Research, USA.

4.9 REGIONALISATION OF LOW FLOW /DROUGHT CHARACTERISTICS

While dealing with low flow studies, generally the annual minimum series (§3.3) or the flow duration curves (§ 3.6) are used to estimate Low Flow characteristics of a stream such as: flow that will be exceeded at a specified recurrence interval or D- day flow that will be equalled or exceeded at specified percent of times. While trying to estimate such quantities for the ungauged basins it is imperative to use the multiple linear regression technique first for establishing relationship between such quantities and the basin characteristics of all gauged basins in the region and then to use the regional equation to estimate the low flow characteristics at the ungauged site.

However, when considering the drought characteristics such as the annual maximum deficit volumes (Figure 3.3) or annual maximum deficit durations, the method of L-Moments can be undertaken to obtain reliable estimates of such characteristics at specified recurrence interval for the identified region and transfer to the ungauged site in a similar manner as done with the annual maximum flow data.

4.10 CONCLUSION

In conclusion it can be said that regionalisation procedure such as the index-flood procedure with the L-Moments being the robust can reliably be used for prediction of flood or drought quantiles in both gauged and ungauged basins.

None-the-less, where a large number of gauged basins are available, the method of multiple linear regression technique can be used to relate such quantities with the basin characteristics such that the regional equation could be used for prediction of the identified quantities in ungauged basins with the use of basin characteristics of the ungauged one.

Bibliography

Cunnane, C. (1988), Methods and Merits of Regional flood Frequency Analysis*, J. Hydrology*, 100, 269-290.

Darlymple, T., 1960, Flood Frequency Methods, U.S. Geological survey, Water Supply Paper 1543A, 11-51.

Farquharson, F.a.K., Meigh, J.R. and Sutcliffe, J.V. (1992). Regional Flood Frequency Analysis in Semi-Arid Areas, *J.of Hydrology,* **138**, 487-501.

Greenwood, J.A., Landwehr J.M., Matalas, N.C. and Wallis, J.R.M., 1979. 'Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form', *Water Resour. Res*., **15(3)**, 1049-1064.

Hosking, J.R.M.(1986) The Theory of Probability Weighted Moments. IBM Research Report No. RC 12210, IBM Research Div., NY.

Hosking, J.R.M. (1990). 'L-Moments: Analysis and estimation of distribution using linear combination of order statistics', *J. R. Stat. Soc. – B*, **52 (1),** 105-124.

Hosking, J.R.M. (!991) "Approximations for use in constructing *L*-moment ratio diagrams", *Research Report RC 16635*, IBM Research Division, Yorktown Heights, N.Y..

Hosking, J.R.M. (1994) The 4-Parameter Kappa distribution, Research Report No. RC 13412, IBM Research Div., NY.

Hosking, J.R.M. and Wallis, J.R. (1997) Regional Frequency Analysis: An Approach Based On L-Moments, Cambridge University Press, Cambridge.

Lettenmaier, D.P. (1985), Regionalisation in Flood Frequency Analysis: Is it the Answer ?, US-China Bilateral Symposium on the Analysis of extra-Ordinary Flood Events, China, pp-25.

Lettenmaier, D.P., Wallis, J.R. and Wood, E.F. , (1987), Effect of Regional Heterogeneity on Flood Frequency Estimation, *Water Resources Research*, **23**(2), 313-323.

Maidment, D.R. (1993) Handbook of Hydrology, McGraw Hill Inc., New York.

Nash, J.E. and Shaw, B.L. (1966), Flood Frequency as a Function of catchment Characteristics, Proc. Symp. on River Flood Hydrology, Instt. Of Civil engineers, London, 191-198.

Natural Environmental Research Council (1975), Flood Studies Report, Vol.1, NERC, UK.

Parida, B.P., Kachroo, R.K. and Srestha, D.B. (1998) Regional Flood Frequency Analysis Of Mahi-Sabarmati Basin (Sub-Zone 3a) Using Index Flood Procedure With L-Moments, Jr. Water Resources Management, 124, 1-12.

Royston, P., (1992) Which measures of Skewness and Kurtosis are best? *Statistics in Medicine*, **11**, 333-343.

US Water Resources Council, 1976, Guidelines for Determining Flood Flow Frequency, Bulletin 17, Hydrology Committee, Water Resources Council, Washington D.C. (Also revised versions, Bulletin 17A, 1977; Bulettin !7B, 1979, 1981, 1982)

Vogel, R.M. and Fennessey, N.M.,(1993), L-Moment Diagrams Should Replace Product Moment Diagrams, *Water Resources Research*, **29, 1745-1752.**

WMO (1989), Statistical Distributions for Flood Frequency Analysis: Operational Hydrology Report No. 33, WMO 718, Geneva,Switzerland,pp.73.

Appendix 1

Global databases:

General
International Glossary of Hydrology (Multilanguage):
http://webworld.unesco.org/water/ihp/db/glossary/glu/HINDEN.HTM
Direct link to french hydrology glossary:
http://webworld.unesco.org/water/ihp/db/glossary/glu/indexdic.htm
Free online Research Journal access
www.oaresciences.org

e-learning, publications and Wiki adresses
Vicaire: **VI**rtual **CA**mpus **I**n hydrology and water **RE**sources management
http://hydram.epfl.ch/VICAIRE/ Basic and applied hydrology for engineers

Mamdouh Shahin, 2002. Hydrology and Water Resources of Africa, ***Water Science and*** *Technology Library V. 41* Kluwer Academic Press, Boston. eBook ISBN: 9780306480652 ISBN: 9781402008665. Visit it via: http://ebooks.springerlink.com/

USGS website has many free publications, models and data
http://www.usgs.gov/

Free on-line encyclopedia:
http://en.wikipedia.org/wiki/Main_Page

Water knowledge Wiki startpage:
http://water.wikia.com/wiki/Main_Page
Experiments and information about equipment
http://www.experimental-hydrology.net/wiki/
Hydro Wiki (student hydrology exchange platform)
http://www.hydrowiki.psu.edu/wiki/

Runoff
Global Runoff Data Centre: http://grdc.bafg.de/
(under WMO guidance)
See guideline and stationlist: 20070807_GRDC_Stations.xls
Overview of internet River Data sources: River_data_Online.xls
http://grdc.bafg.de/servlet/is/3422/River_data_Online.xls?command=downloadContent&filename=River_data_Online.xls

The Global River Discharge Database (RivDIS v1.1): http://www.rivdis.sr.unh.edu/
Is part of GHAAS: Global Hydrological Archive and Analysis System

River Discharge Database: http://www.sage.wisc.edu/riverdata/
Center for Sustainability and the Global Environment (SAGE), hosted at Nelson institute for environmental studies, University of Wisconsin-Madison
This site contains a compilation of **monthly mean river discharge** data for over 3500 sites worldwide. The data sources are RivDis2.0, the United States Geological Survey, Brazilian National Department of Water and Electrical Energy, and HYDAT-Environment Canada. The period of record for each station is variable, from 3 years to greater than 100. All data is in m3/s.

Meteorological time series
Climate explorer of the Royal Dutch Meteorological Institution: http://climexp.knmi.nl/
This database offers an overview of a number of existing databases. You can directly look at data such as precipitation and temperature from stations all over the world and also download the data in ASCII format. There are also a number of gridded or modeled data sources available such as reanalysis data and CRU data (see below).

Global Historical Climate Network (GHCN v. 2): ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v2
Here you can find historical monthly station data of precipitation, minimum and maximum temperature for stations all over the world. 5x5 degree gridded data are also provided.

Climate Research Unit (CRU): http://www.cru.uea.ac.uk/
These provide monthly climate grids from the past century, based on station data (at 0.1 or 0.5 degrees). The quality is largely dependent on the regional availability of stations within a period. Data is given in an ascii format. More information is given on the webpage.

LandSAF MSG: http://landsaf.meteo.pt/
LandSAF provides temporally varying land surface characteristics derived from Meteosat Second Generation. Fast varying variables are for instance incoming short / long wave radiation and land surface temperature. Daily varying variables are e.g. land surface albedo and vegetation indices. Register first to have access to the data (HDF5 format).

FEWS RFE 2.0 rainfall: information is found on
http://earlywarning.usgs.gov/CentralAmerica/dekrferm.php
Downloading data can be done on ftp://ftp.cpc.ncep.noaa.gov/. Login as anonymous. More information about the rainfall product can be found in Xie et al (2000)
http://daac.ornl.gov/safari2k/climate_meteorology/FEWS_precip/comp/FEWS_companion.pdf

FEWS end-user products: www.fews.net

Tropical Rainfall Measuring Mission (TRMM): several products (including near-real time rainfall) and descriptions can be found on http://trmm.gsfc.nasa.gov/.
Easier bulk downloads of monthly adjusted products can be done on ftp://disc2.nascom.nasa.gov/ (login: anonymous). Bulk downloads on near-real time basis can be done on ftp://trmmopen.gsfc.nasa.gov/ (login: anonymous). Note that the near-real time products are found to deviate substantially from the monthly corrected products.

Static data:
HYDRO1k database: http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html
1 x 1 km^2 elevation for the whole World. Suitable for large scale modeling and pre-projected to Lambert Azimuthal Equal Area projection.

USGS Seamless data distribution: http://seamless.usgs.gov/
This is really the place to retrieve elevation data on high resolution. Other datasets that are available for Africa are the 10-daily FEWS estimates of NDVI and RFE 2.0 rainfall (both on 0.1 deg. Resolution)

Global Land Cover Characterization: http://edcsns17.cr.usgs.gov/glcc/
Land cover database. It's a bit old but can still be used for large scale applications.

Land surface data:
SPOT vegetation: http://free.vgt.vito.be/
Offers free 10-daily 1x1 km^2 NDVI data. VGTExtract is a tool that helps clipping and cropping to designated regions, after the data has been downloaden. It is crucial to remove clouded pixels (designated in the quality flags) before using it. You can consider filling these with long-term averaged NDVI values from other sources or do a time-series analysis, for instance, keeping the NDVI value of the previous time step when it is clouded. Register and log in to download the data.

Remotely sensed landcover information and DEM
http://www.landcover.org/index.shtml
The GLCF is a center for land cover science with a focus on research using remotely sensed satellite data and products to assess land cover change for local to global systems.
• GLCF FAQs
• UMD MODIS Research
• GOFC-GOLD
• Landsat GeoCover
• SRTM DEM GeoTIFFs
• Rapid Response
• IUCN Protected Areas

Digital Soil map of the world and derived soil parameters from FAO
http://www.fao.org/ag/agl/agll/dsmw.htm

De Fao internet catalogue!
http://www.fao.org/icatalog/search/result.asp?subcat_id=206

Appendix 2

Table 2 (a) Areas under the normal curve for negative Z values

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table 2(b) Areas under the normal curve for positive Z values

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6024	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7424	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table 3(a) Frequency factors for use in Pearson type III and Log-Pearon type III distributions for positive skew

						Recurrence Interval					
Skew	**1.0101**	**1.0526**	**1.1111**	**1.25**	**2**	**5**	**10**	**25**	**50**	**100**	**200**
						Percent Chance					
	99	**95**	**90**	**80**	**50**	**20**	**10**	**4**	**2**	**1**	**0.5**
3.0	-0.667	-0.665	-0.660	-0.636	-0.396	0.420	1.180	2.278	3.152	4.051	4.970
2.9	-0.690	-0.688	-0.631	-0.651	-0.390	0.440	1.195	2.277	3.134	4.013	4.909
2.8	-0.714	-0.711	-0.702	-0.666	-0.384	0.460	1.210	2.275	3.114	3.973	4.847
2.7	-0.740	-0.736	-0.724	-0.681	-0.376	0.479	1.224	2.272	3.093	3.932	4.783
2.6	-0.769	-0.762	-0.747	-0.696	-0.368	0.499	1.238	2.267	3.071	3.889	4.718
2.5	-0.799	-0.790	-0.771	-0.711	-0.360	0.518	1.250	2.262	3.048	3.845	4.652
2.4	-0.832	-0.819	-0.795	-0.725	-0.351	0.537	1.262	2.256	3.023	3.800	4.584
2.3	-0.867	-0.850	-0.819	-0.739	-0.341	0.555	1.274	2.248	2.997	3.753	4.515
2.2	-0.905	-0.882	-0.844	-0.752	-0.330	0.574	1.284	2.240	2.970	3.705	4.444
2.1	-0.946	-0.914	-0.869	-0.765	-0.319	0.592	1.294	2.230	2.942	3.656	4.372
2.0	-0.990	-0.949	-0.895	-0.777	-0.307	0.609	1.302	2.219	2.912	3.605	4.298
1.9	-1.037	-0.984	-0.920	-0.788	-0.294	0.627	1.310	2.207	2.881	3.553	4.223
1.8	-1.087	-1.020	-0.945	-0.799	-0.282	0.643	1.318	2.193	2.848	3.499	4.147
1.7	-1.140	-1.056	-0.970	-0.808	-0.268	0.660	1.324	2.179	2.815	3.444	4.009
1.6	-1.197	-1.093	-0.994	-0.817	-0.254	0.675	1.329	2.163	2.780	3.388	3.990
1.5	-1.256	-1.131	-1.018	-0.825	-0.240	0.690	1.333	2.146	2.743	3.330	3.910
1.4	-1.316	-1.168	-1.041	-0.832	-0.225	0.705	1.337	2.128	2.706	3.271	3.828
1.3	-1.383	-1.206	-1.064	-0.838	-0.210	0.719	1.339	2.108	2.666	3.211	3.745
1.2	-1.449	-1.243	-1.086	-0.844	-0.195	0.732	1.340	2.087	2.626	3.149	3.661
1.1	-1.518	-1.280	-1.107	-0.848	-0.180	0.745	1.341	2.066	2.585	3.087	3.575
1.0	-1.588	-1.317	-1.128	-0.852	-0.164	0.758	1.340	2.043	2.542	3.022	3.489
0.9	-1.660	-1.353	-1.147	-0.854	-0.148	0.769	1.339	2.018	2.498	2.957	3.401
0.8	-1.733	-1.388	-1.166	-0.856	-0.132	0.780	1.336	1.993	2.453	2.891	3.312
0.7	-1.800	-1.423	-1.183	-0.857	-0.116	0.790	1.333	1.967	2.407	2.824	3.223
0.6	-1.880	-1.458	-1.200	-0.857	-0.099	0.800	1.328	1.939	2.359	2.755	3.132
0.5	-1.955	-1.491	-1.216	-0.856	-0.083	0.808	1.323	1.910	2.311	2.686	3.041
0.4	-2.029	-1.524	-1.231	-0.855	-0.066	0.816	1.317	1.880	2.261	2.615	2.949
0.3	-2.104	-1.555	-1.245	-0.853	-0.050	0.824	1.309	1.849	2.211	2.544	2.856
0.2	-2.178	-1.586	-1.258	-0.850	-0.033	0.830	1.301	1.818	2.159	2.472	2.763
0.1	-2.252	-1.616	-1.270	-0.846	-0.017	0.836	1.292	1.785	2.107	2.400	2.670
0.0	-2.326	-1.645	-1.282	-0.842	0.000	0.842	1.282	1.751	2.054	2.326	2.576

Table 3 (b) Frequency factors for use in Pearson type III AND Log-Pearon type III distributions for negative skew

Skew	1.0101	1.0526	1.1111	1.25	2	5	10	25	50	100	200
						Percent Chase					
	99	95	90	80	50	20	10	4	2	1	0.5
0.0	-2.326	-1.645	-1.282	-0.842	0.000	0.842	1.282	1.751	2.054	2.326	2.576
-0.1	-2.400	-1.673	-1.292	-0.836	0.017	0.846	1.270	1.716	2.000	2.252	2.432
-0.2	-2.472	-1.700	-1.301	-0.830	0.033	0.850	1.258	1.680	1.945	2.178	2.388
-0.3	-2.544	-1.726	-1.309	-0.824	0.050	0.853	1.245	1.643	1.890	2.104	2.294
-0.4	-2.615	-1.750	-1.317	-0.816	0.066	0.855	1.231	1.606	1.834	2.029	2.201
-0.5	-2.686	-1.774	-1.323	-0.808	0.083	0.856	1.216	1.567	1.777	1.955	2.103
-0.6	-2.755	-1.797	-1.328	-0.800	0.099	0.857	1.200	1.528	1.720	1.880	2.016
-0.7	-2.824	-1.819	-1.333	-0.790	0.116	0.857	1.183	1.488	1.663	1.880	2.016
-0.8	-2.891	-1.839	-1.336	-0.780	0.132	0.856	1.166	1.448	1.606	1.733	1.837
-0.9	-2.957	-1.858	-1.339	-0.769	0.148	0.854	1.147	1.407	1.549	1.660	1.749
-1.0	-3.022	-1.877	-1.340	-0.758	0.164	0.852	1.128	1.306	1.492	1.588	1.664
-1.1	-3.087	-1.894	-1.341	-0.745	0.180	0.848	1.107	1.324	1.435	1.518	1.581
-1.2	-3.149	-1.910	-1.340	-0.732	0.195	0.844	1.086	1.282	1.329	1.449	1.501
-1.3	-3.211	-1.925	-1.339	-0.719	0.210	0.838	1.064	1.240	1.324	1.383	1.424
-1.4	-3.271	-1.938	-1.337	-0.705	0.225	0.832	1.041	1.198	1.270	1.318	1.351
-1.5	-3.330	-1.951	-1.333	-0.690	0.240	0.825	1.018	1.157	1.217	1.256	1.202
-1.6	-3.388	-1.962	-1.329	-0.675	0.254	0.817	0.994	1.116	1.166	1.197	1.216
-1.7	-3.444	-1.972	-1.324	-0.660	0.268	0.808	0.970	1.075	1.116	1.140	1.155
-1.8	-3.499	-1.981	-1.318	-0.643	0.282	0.799	0.945	1.035	1.069	1.087	1.097
-1.9	-3.553	-1.989	-1.310	-0.627	0.294	0.788	0.920	0.996	1.023	1.037	1.044
-2.0	-3.605	-1.996	-1.302	-0.609	0.307	0.777	0.895	0.959	0.980	0.990	0.995
-2.1	-3.656	-2.001	-1.294	-0.592	0.319	0.765	0.669	0.923	0.939	0.946	0.949
-2.2	-3.705	-2.006	-1.284	-0.574	0.330	0.752	0.844	0.888	0.900	0.905	0.907
-2.3	-3.753	-2.009	-1.274	-0.555	0.341	0.739	0.819	0.855	0.864	0.867	0.869
-2.4	-3.800	-2.011	-1.262	-0.537	0.351	0.725	0.795	0.823	0.830	0.832	0.833
-2.5	-3.845	-2.012	-1.250	-0.518	0.360	0.711	0.771	0.791	0.798	0.799	0.800
-2.6	-3.889	-2.013	-1.238	-0.499	0.368	0.696	0.747	0.764	0.768	0.769	0.769
-2.7	-3.932	-2.012	-1.224	-0.479	0.376	0.681	0.724	0.738	0.740	0.740	0.741
-2.8	-3.973	-2.010	-1.210	-0.460	0.384	0.666	0.702	0.712	0.714	0.714	0.714
-2.9	-4.013	-2.007	-1.195	-0.440	0.390	0.651	0.681	0.683	0.689	0.690	0.690
-3.0	-4.051	-2.003	-1.180	-0.420	0.396	0.636	0.660	0.666	0.666	0.667	0.667

Appendix 3 The blind men and the Elephant

by John Godfrey Saxe (1816-1887)

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"

The Second, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

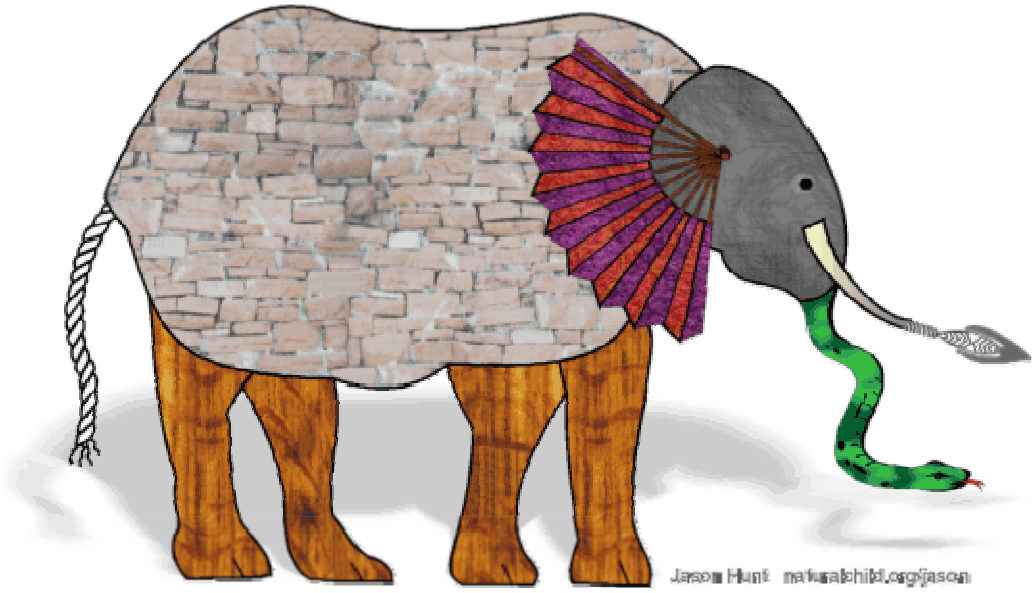
The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"

The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
"Tis clear enough the Elephant
Is very like a tree!"

The Fifth, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
Is very like a fan!"

The Sixth no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!



So oft in theologic wars,
The disputants, I ween,
Rail on in utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!