Government of the People's Republic of Bangladesh
Ministry of Water Resources

Time Series Data Quality Control Guideline

(For Hydro-meteorological data)

July 2015

WARPO
পানি সম্পদ পরিকল্পনা সংস্থা

Water Resources Planning Organization

Preface

Hydro-meteorological information is very essential for land and water resources planning, development and management in Bangladesh. National Water Resources Database (NWRD) of Water Resources Planning Organization contains valuable and huge amount of time series data that was collected from the Bangladesh Water Development Board, Bangladesh Meteorological Department, Bangladesh Inland Water Transport Authority and other relevant agencies. Surface water level, discharge and cross-section are the main hydrological data available in Bangladesh and similarly meteorological data available are rainfall, temperature, humidity, sunshine, wind speed, ground water level etc. Quality of data is very crucial which refers to the relative accuracy and precision of a particular application of those data in different study or projects.

Over the years, quality assurance for the time series hydro-meteorological data has become increasingly important. In Bangladesh, there are not yet any standard guidelines for processing and quality checking of hydro-meteorological data. Quality guidelines for hydrological data are important particularly for the NWRD, where different type of data from different sources are accumulated and for providing water resources planning and management in Bangladesh. As WARPO is mandated organization, who prepared the Time Series Quality Control Guideline on Hydro-meteorological data to maintain and update National Water Resources Database. The main purpose of this guideline is to setup a national standard for quality assurance of NWRD and the important steps to be followed data quality checking, detecting inconsistencies and recommend necessary steps for data processing and for comparing with international standard.

Although, the guideline focuses on data collection method, procedure, quality assurance & framework, processing factors, etc. responsible for overall data status in Bangladesh, however it also help to analyze the time series, basic statistical anaysis, data validation and exercise of data processing steps.

WARPO has taken necessary steps to review this guideline by the national agencies, universities and research institutes who would be the user of the guideline to ensure their own data quality. Most of comments and suggestions received from the reviewers were incorporated into the document. Some comments that cannot be incorporated into the current version will be incorporated in future update.

This Guideline is a state of procedures or practices that would be useful to the different users. WARPO will be grateful and happy if the guideline will be used to planner, researcher other stakeholder with feedback for further improvement of the document.

List of abbreviations

ADCP	Acoustic Doppler Current Profiler
ANN	Artificial Neural Network
ARFF	Attribute-Relation File Format
ASCII	American Standard Code for Information Interchange
AWLR	Auto Water Level Recorder
BADC	Bangladesh Agricultural Development Corporation
BIWTA	Bangladesh Inland Water Transport Authority
BMD	Bangladesh Meteorological Department
BMDA	Barind Multipurpose Development Authority
BTMA	Bed load Transportmeter Arnhem
BUET	Bangladesh University of Engineering Technology
BWDB	Bangladesh Water Development Board
CDF	Cumulative Distribution Function
DCA	Data Collecting Agencies
DGPS	Digital Global Positioning System
DoE	Department of Environment
DoF	Department of Forest
DPHE	Department of Public Health Engineering
DTW	Deep Tubewell
EGIS	Environmental and Geographic Information System
EMF	Electromagnetic Fields
EV	Extreme Value
GIS	Geographic Information System
GWC	Ground Water Circle
EPA	Environment Protection Agency
FAO	Food and Agriculture Organization
FAP	Flood Action Plan
FFWC	Flood Forecasting and Warning Center
GEV	generalized extreme value
GMT	Greenwich Mean Time
HMV	Hymos Manual Version
H-S	Helley-Smith
ISO	International Standard Organisation
IWM	Institute of Water Modelling
LGED	Local Government Engineering Department
NEMIP	North East Minor Irrigation Project
NWRD	National Water Resources Database
NWMP	National Water Management Plan
PC	Personal Computer
PDB	Point Data Bank
PDF	Probability Density Function
POT	Peaks over threshold
PWD	Public Works Department
PWM	Probability Weighted Moments
QC	Quality Check
RL	Reduced Level
RMSE	Root Mean Square Error
SPARRSO	Space Research and Remote Sensing Organization
WARPO	Water Resources Planning Organization
WMO	World Meteorological Organization

Table of Contents

List of Tables

List of figures

Chapter 1
Introduction

1.1 Background

Hydrological data are an important component for the planning process as the country's land and water resource systems need continuous planning and management. Hydrological data are not only prerequisite for planning, development and management of water resources and the environment, they are also necessary for the scientific study of hydrological processes. With availability of sophisticated methods for data analysis and multi-dimensional problems such as floods of long durations, droughts, drainage congestion, low flow in rivers, and siltation in river beds, emphasis is now being placed on comprehensive and quality data. Data quality refers to the relative accuracy and precision of a particular database. Assessment refers to standardisation and suitability for presentation. The quality of hydrological data depends a lot on field and office practices. These practices encompass the core activities of data collection namely the instrumentation system, data observation program, the central collection for annotation and pre-processing of raw data – inevitably a sound and robust database system for processing, retrieval and dissemination of data to users. Quality guidelines for hydrological data are important particularly for the NWRD, where different data sources are accumulated, and for supporting national water resources planning as well.

1.2 Regulatory background

The hydro-meteorological data quality control guideline states procedures or practices that may be useful to the users to whom they are directed, but are not legal requirements. The guideline represents the organisation's position on a procedure or practice at the time of their issuance. A planner may follow the guideline or may choose to follow alternative procedures. If a planner chooses to use alternate procedures, that person may wish to discuss the matter further with the organization. The guideline does not bind any organization and nor does it create or confer any rights, privileges, immunities, or benefits for or on any person. Where the guideline states a requirement imposed by the National Water Resources Database (NWRD), its force and effect are not changed in any way by users.

1.3 Rationale for the guideline

In Bangladesh, there were no standard guidelines for processing and quality checking of hydrometric and hydro-meteorological data. This data quality control guideline will fill up this gap have a pivotal role in this field. During the development of the NWRD, a huge amount of time series data was collected from the Bangladesh Water Development Board (BWDB), the Bangladesh Meteorological Department (BMD) and other organizations. The NWRD is a database to ensure the qualitative and quality line used for national water resources planning and management. It is a unique databank at the national level, which provides good quality data. The following features are within the scope of the guideline:

- ◊ To document the present data quality and options for quality data collection in the future.
- ◊ To describe data collection procedures followed in different organisations in Bangladesh.
- ◊ To provide some quality checking options for different data and for comparing with international standards.
- ◊ To define the data quality level and the best way of detecting inconsistencies.

- ◊ To analyze and interpret hydrological data series with the goal of providing reliable bases for water management and environmental protection.
- ◊ To recommend necessary steps for data processing in Bangladesh.

While carrying out the above activities, the NWRD needs to establish guidelines for processing and quality assessment of time series data.

1.4 Objectives

The guideline will help users as well as researchers in checking quality and assessing data.

The **short term** objectives of the guideline are:

- ◊ checking data quality and scope of different organisations in Bangladesh
- ◊ providing necessary data and information for planners
- ◊ developing a framework for the best way of data quality checking
- ◊ defining the best statistics for quality checking
- ◊ comparing existing data quality with the international standards
- ◊ suggesting methods of filling and handling missing values

The **long term** objectives of the guideline are:

- ◊ to provide support to micro and macro level planning
- ◊ to establish the best possible accuracy of hydrological data in Bangladesh
- ◊ to disseminate the guidelines at different organisations for maintaining their activities at quality level
- ◊ to introduce a methodology for accuracy maintenance

1.5 Scope of the guideline

The guideline will help to control and assess quality more after data collection than at field level. This report discusses the different methods and instruments but does not give details on how to use the instruments. The user can get information about the method of data collection and data quality assessment from the guideline. The processing steps, which are supposed to be followed for checking the quality of data, have been incorporated in the guideline. The present data quality standards are shown through comparison with the international standards. Different types of data are included in the statistical part to enable users to find out the best possible alternatives for checking data quality.

Chapter 2
Data and Quality

2.1 Introduction

The method used to collect data from the field influences the quality of the data. Different types of data are collected by different methods. The human factor is an important consideration for any field measurement. Properly collected data gives absolute measures of the quality of data. Different types of errors such as random, systematic or non-homogenous errors may be introduced in the observation/sensing, transmitting or recording during measurements. Although it is very difficult to maintain certain standards of reliability and accuracy of time series data, the collection and processing of data are crucial particularly for national databases. The following is a discussion of these difficulties.

2.2 Data types

Hydrological data, which have been classified in hydrometric and hydrometeorological data, have been dealt with in this guideline. In Bangladesh, the main sources of hydrological data are the BWDB, the BMD and the Bangladesh Inland Water Transport Authority (BIWTA). Some project oriented hydrological time series data are also available. More specifically, these two categories of data are known as collection units:

- **Hydrometric data**
    - ◊ Water level
    - ◊ Discharge
    - ◊ Groundwater, and
    - ◊ Sediment.
- **Hydrometeorological data**
    - ◊ Rainfall and
    - ◊ Evaporation (temperature, humidity, wind speed and sunshine).

Upon state, data can be classified as the following types:

- **Time–oriented data**
    - ◊ Equidistant time series (same duration in collection) and
    - ◊ Non-equidistant time series.

Non-equidistant time series can be transformed into equidistant time series. Generally, the non-equidistant series may not fill all equidistant time steps.

- **Relation-oriented data**
    - ◊ Stage-discharge
    - ◊ Rating curve parameters
    - ◊ Relation curve parameters, and
    - ◊ Relation between data series
- **Space-oriented data**
    - ◊ Catchment characteristics
    - ◊ Local origin in geographical coordinates

- ◊ Sub-catchment characteristics–area, boundaries, river slope and length, etc.
- ◊ Plot- graphical display of catchment boundaries
- ◊ Station data: characteristics, log-book and histories
- ◊ Series characteristics, and
- ◊ Geo-hydrological profiles

2.3 Data collection scenarios

2.3.1 Data collecting agencies in Bangladesh

At this level an overview can be given on the overall scenario of the data collection in Bangladesh. In hydrological data collection, any missing observation will be lost forever and can seriously effect development costs of a particular water resources project in the future. As hydrological data series are subject to variation in time and space, they should be observed and measured in accordance with the standard practices and guidelines. In Bangladesh, the key organizations collecting data on hydrology as time series are as follows:

Table 2.1: Sources of hydrometric/hydro-meteorological information in Bangladesh

Types of information	Operating agency	Purpose	Data collection frequency	Unit
Surface Water level	BWDB (Both Tidal and Non-tidal)	Planning, monitoring and designing	As per available document, BWDB collects five data a day and disseminates daily data	Meter (mPWD)
	BIWTA (Tidal)	For navigational and other related infrastructure development	BIWTA collects hourly data	Meter (mPWD, Chart Datum)
Discharge	BWDB (Both Tidal and Non-tidal)	Planning and monitoring	BWDB generally collects data at intervals of one week during the monsoon season (May-November) and fortnightly during the rest of the year. BWDB disseminates daily data from rating curves.	Cubic meter per second (m^3/sec)
Groundwater	BWDB (Level and Quality)	Planning monitoring and designing forecasting purpose, research, academic, and monitoring & commercial application	Weekly continuous	Meter and mPWD
	DPHE (Level and Quality)		Annually and weekly	Meter and mPWD
	BADC (Level and Quality)		Fortnightly to monthly 2*annually continuous	Meter
	BMDA (Level)		Fortnightly	

Types of information	Operating agency	Purpose	Data collection frequency	Unit
Sediment	BWDB	Planning and monitoring	Weekly	ppm
Rainfall	BWDB	Planning and monitoring	Daily	mm/day
	BMD	Forecasting purpose, research, academic, monitoring & commercial application	3 hourly	mm/3hr
Evaporation	BWDB	Planning and monitoring	Daily	mm
	BMD	Forecasting purpose, research, academic, monitoring & commercial application	Daily	mm

Source: BWDB, BMD BIWTA and others

2.3.2 International standard for observations

For certain types of data, standards have been established by international agencies. These provide point of reference against which the density and frequency of observations can be measured. A comparison is given in Table 2.2 between some of the standards used by the World Meteorological Organization (WMO, 1974) and the corresponding level of observations in Bangladesh, which is a generally low-lying plain where aerographic effects are limited.

Table 2.2: Relevant international standards for plains

Data type	Agency	Standard of observation Min^m Area/Station (km^2)*	Available observations Area/Station (km^2)
Water level	BWDB	-	294
Discharge	BWDB	1,750	905
Groundwater level	BWDB	-	110
Sediment	BWDB	-	3,700
Rainfall	BWDB/BMD	575	450
Evaporation	BWDB/BMD	50,000	2,900

**Source: WMO, BWDB and BMD*

Most of the existing network fulfills the WMO criteria for a standard network. The BWDB, the BMD and the BIWTA all hold valuable long time-series data, but few sets exceed 40 years.

2.3.3 International standards in accuracy

The measurement accuracy of some of the hydrological parameters is described in Table 2.3:

Table 2.3: International standards in accuracy (WMO)

Data type	Specification
River discharge and suspended sediment	5%
Water level	10 mm
Rainfall	0.5 mm
Evaporation	0.1 mm

Source: WMO

2.4 Data collection procedures

2.4.1 Water level data

Different types of gauges

Most of the water level gauging stations are equipped with a gauge and a recorder. In many cases the water level is measured in a stilling well, thereby eliminating strong oscillations. The following table gives an overview of the gauges, which are most commonly used in Bangladesh, indicating also the necessity of a stilling well along with the way of reading.

Table 2.4: Types of gauges

Type of gauge	Stilling well recommended	Way of reading	
		Intermittently directly manual	Continuously recorder-equipped
Staff gauge	--	Yes	--
Float operated gauge	Indispensable	Possible	Yes
Wire-weight gauge	Preferable	Yes	Possible
Pneumatic gauge	--	--	Yes
Flood crest gauges	--	Yes	--

Source: BWDB

Station installation requirements

Installation of a water level measurement station consists of the following:

- ◊ The installation of the water level measurement system (water level sensor(s), Data Collection Platform (DCP), satellite transmitter) and its supporting structure and staff.
- ◊ The recovery and/or installation of a minimum number of benchmarks and a level connection between the benchmarks and water level sensor(s), or tide staff as appropriate.
- ◊ The preparation of all documentation and forms.

Additional field requirements

(a) Generally upon completion of the data acquisition for each gauge installed, the data must be sent all together within a minimum of 30-day unless the data are transmitted via satellite or other quick means.

(b) All water level data from an auto gauge should be downloaded and backed up at least weekly on diskettes/cds whether the gauge data are sent via satellite or not.

(c) For new stations the head office shall be contacted once the location of the gauge has been finalized and the latitude and longitude of the gauge site shall be provided at least three working days prior to actual installation of the gauge in field. The head office will assign a new station number within three available days and inform the hydrographer.

(d) The progress sketch shall show the field sheet, layout, area of hydrography, gauge location, and other information as appropriate. The location of the gauge as shown on the progress sketch, benchmark and station location sketch shall also be verified.

Automatic recording systems

According to the WMO, most of the water level gauges can be adapted for automatic recording except for the staff gauge and the flood crest gauges, which require direct observation. Apart from the pneumatic gauges, recording gauges are usually placed in stilling wells, which dampen fluctuations caused by waves and turbulence. Normally a non-recording gauge is installed next to a recording gauge to enable comparison of direct readings and the recorded stages.

For recording purposes both analogue and digital systems can be used. The analogue system can be made to operate unattended for periods from 2 weeks to 3 months. The analogue system provides a graphical record proportional to the actual rise or fall of the water level with respect to time. The recording is made on recorder paper fixed to a rotation drum, which is moved by a clock. The rate of rotation determines the time base of the hydrograph.

The digital stage-recorder punches code values of stage on paper tape at pre-selected time intervals, for instance every 15 minutes. The punching of a stage requires only a few millimetres of tape; this is wound up by a clock. The two methods of recording are about equal in accuracy, reliability and cost. The advantage of the digital system is that it is more compatible with the use of electronic computers.

Data collection procedure adopted by the BWDB: Surface Water Hydrology (SWH-I) of the BWDB operates in about 100 water level stations in the main rivers and 389 water level stations in other rivers. Water levels are normally measured with wooden staff-gauges five times in a day at 6:00, 9:00, 12:00, 15:00 and 18:00 hours at these stations. Since the seasonal water-level variations exceed by far the reach of a staff gauge, its reach is regularly adjusted by fixing a new gauge close to the previous one. During high flow, the gauge may be shifted to another location. Sometimes it may be shifted up to one kilometer upstream or downstream and datum from a nearby benchmark is carried out. Water-level corrections are carried out at the field offices before the data are transferred to SWH-II for further processing. The water levels received from the field are adjusted for the vertical shifts according to the check leveling.

2.4.2 Discharge data

Measurement procedure

Discharge at a given stage is computed from measurements of velocity and depth at a cross section near the recorder. Velocity is measured at locations or verticals spaced across the cross section using a current meter. The spacing between velocity measurements should be such that not more than 10 percent of total flow is represented by any one vertical. The current meter is suspended on a cable controlled by

a winch or in shallow water and mounted on a measuring rod carried by the gauge. The depth at each vertical is measured using the cable or measuring rod. A current meter contains a rotating element whose speed of rotation is proportional to the water velocity. The vertical-axis that meets with rotating cups (price type meter) is commonly used in the United States. Horizontal-axis rotating screw (or propeller) meters are less prone to pollution by weed. Other methods of measuring include use of floats, a boat equipped with current meter and sonar, a velocity head rod, and ultrasonic equipment.

In water depth measurement a synthetic rope is marked in red at every half meter. Similarly, PVC pipes are used for measuring the water depth marked in red at every centimeter. A uniform section is then closed at the downstream for selecting each regulator. The rope is fastened tightly with wooden pegs and placed on the water surface perpendicular to the direction of the flow. In the wider sections of canals, water depth at every meter is measured with PVC pipes. On the other hand, in narrow sections of canals, depths are measured at every half meter. For wider and narrow canals, the current meter readings are taken at 0.8 and 0.2 depths at the center of each segment of the channel width. Where it is found difficult to place the current meter at 0.8D, measurements are only taken at 0.2D. Measurements are also taken at every half meter in narrow and shallow sections of canals.

Data collection procedure by the BWDB

Discharge is measured at five locations in the main rivers, generally at intervals of one week during the monsoon season (May-November) and fortnightly during the rest of the year. The BWDB uses the velocity-area method for determining discharge. The flow velocities are measured from a survey boat by a non-directional Ott current-meter (propeller type), exposed at 0.2 and 0.8 of depth in the verticals. The measuring time in each point in a vertical is 100 sec. The survey boat is dynamically positioned; i.e. the boat is not anchored. Its location in the transect is determined with a sextant. The suspension cable with the current-meter is used to measure the depth in a vertical. The number of verticals varies according to the actual flow conditions. The rule is that in one vertical no more than 10% of the total flow in a channel is measured. The required number of verticals often becomes very high (at Bahadurabad some 100 verticals) and it takes about two days to complete one measurement of the total discharge. The direction of the flow on the water surface is determined at each measurement point across the river by following the path of a floating bottle. The float positions are measured by sextants.

2.4.3 Groundwater

The origin of groundwater is through infiltration, influent streams, and seepage from reservoirs, artificial recharge, condensation, seepage from oceans, water trapped in sedimentary rock, and juvenile water. Groundwater occurs in two zones:

- Saturated zones, and
- Unsaturated zones

Groundwater level monitoring agencies in Bangladesh

The water level in any well usually does not remain constant, but changes in response to several factors. Rainfall distribution and the amount may affect groundwater recharge and discharge, and subsequently may affect the water level in area wells. Also, wells that are hydraulically connected to a stream may show fluctuations in the water level as the stream level changes. In some cases, depending upon the hydraulic properties of the geologic formation, the intense pumping of a well, or number of wells, the water level in some nearby wells may be lowered. Groundwater levels are being monitored by the following agencies:

- Bangladesh Water Development Board (BWDB)

- Bangladesh Agricultural Development Corporation (BADC)
- Department of Public Health Engineering (DPHE)
- Barind Multipurpose Development Authority (BMDA)

Table 2.5: Available groundwater level data

Agency	Station type	Frequency
BWDB	Piezometric AWLRs	weekly continuous
BADC	DTWs AWLRs	fortnightly to monthly 2*annually continuous
DPHE	Rural and Urban	Annually and weekly
BMDA	DTWs	fortnightly

Source: BWDB, DPHE, BADC and BMDA

Groundwater level measurement BWDB

The BWDB measures groundwater level data every Monday at 9:00 a.m in a week throughout the country. Wells are set to a suitable position and measurements are taken from the measuring point at the top of the parapet level. Tapes or ropes are scaled for carrying out measurements. The height of the measuring point is noted with respect of the Reduced Level (R.L.). Besides 20 automatic groundwater level recorders are used for measuring water level.

Standard measurement procedure (U. S. EPA Environmental Response Team)

Monitoring of water table depth is carried out using manual measurements, automatic water-level recorders, or pressure transducers. A survey mark should be placed on top of the riser pipe or casing as a reference point for groundwater level measurements. If the lip of the riser pipe is not flat, the reference point may be located on the grout apron or the top of the outer protective casing (if present). The measurement reference point should be documented in the site logbook and on the groundwater level data form, if used. All field personnel must be made aware of the measurement reference point being used in order to ensure the collection of comparable data. Before measurements are made, water levels in piezometers and monitor wells should be allowed to stabilize for a minimum of 24 hours after well construction and development. In low yield situations, recovery of water levels to equilibrium may take longer. All measurements should be made to an accuracy of 3.05 mm. Water level measuring equipment must be decontaminated and, in general, measurements should proceed from the least to the most contaminated wells. The well should be opened and the headspace monitored with appropriate air-monitoring instruments to determine the presence of volatile organic compounds. For electrical sounders the device should be lowered into the well until the water surface is reached as indicated by a tone or meter deflection. The distance should be recorded from the water surface to the reference point. Measurement with a chalked tape will necessitate lowering of the tape below the water level and holding a convenient foot marker at the reference point. Both the water levels should be recorded as indicated on the chalked tape section and the depth mark held at the reference point. The depth to water is the difference between the two readings. The measuring device should then be removed, the riser pipe cap replaced, and the equipment decontaminated as necessary. Note that if a separate phase is present, an oil/water indicator probe is required for measuring product thickness and water level.

2.4.4 Sediment data

Methods of sediment measurement

In addition to the flow recordings, the routine gauging comprised determination of suspended sediment transport and bed load transport as well as sediment grain-size distributions, vertical distribution of sediment concentration and grain-size/settling velocity distribution. The methods and analyses included in the routine monitoring are listed below:

Table 2.6: Methods of different samplings

Method	Instrument
Suspended sediment sampling	Pump bottle Integrated bottle Optical turbidity ADCP back-scatter
Near-bed sediment sampling	Dune tracking with echo-sounder and side-scan sonar Helly-Smith trap sampler
Bed material sampling	US BM-54, Scoop sampler Van Veen Grab sampler

Source: Special Report 12, FAP24

Data collection by BWDB

The BWDB measures suspended sediment in a few selected stations in the main rivers. In these rivers suspended sediment is considered to be more important than bed load transport. Bed load transport is not measured regularly due to difficulties in measuring it and its presumed relative low importance. Therefore, the attention is focused on suspended sediment transport measurements. The BWDB has selected the Binckley silt sampler made by Kelvin Hughes as the standard instrument for suspended sediment sampling.

After entrapping a quantity of water and suspended sediment the Binckley sampler is raised to the surface, opened and emptied into the elutriator, thus separation of the sand and silt/clay fractions of the samples is achieved. Instantaneous suspended sediment point samples are collected at each alternate vertical assigned for flow velocity measurement. In this system samples are taken at one point at the relative depth of 0.6m if the local water depth is less than 1 meter. On the other hand, samples are taken at two points at the relative depths of 0.2 and 0.8 if the local depth is more. At relatively high flow velocities a special method is developed earlier in order to avoid a large drag angle of the sampler.

The sample of river bed is collected with a Kolb Bed Sampler. All the samples are stored in a plastic bag and sent to the River Research Institute (RRI) for analysis.

2.4.5 Rainfall

Measuring equipment

Recording gauges are of three different types:

- ◊ weighing type
- ◊ tilting bucket type, and
- ◊ float type

The weighing type observes precipitation directly when it falls (including snow) by recording the weight of the reservoir e.g. by pen on a chart. With the float type rain is collected in a float chamber and the vertical movement of the float is recorded by pen on a chart. Both types have to be emptied manually

or by automatic means. The tilting or tipping bucket type is a very simple recording rain gauge, but less accurate (only registration when bucket is full, losses during tipping and due to evaporation).

Storage gauges for daily rainfall measurement are observed at a fixed time each morning. Recording gauges may be equipped with charts that have to be replaced daily, weekly or monthly, depending on the clockwork. The rainfall is usually recorded cumulatively (mass curve) from which the hyetograph (a plot of the rainfall with time) is easily derived. The tipping bucket uses an electronic counter or magnetic tape to register the counts (each count corresponds to 1 mm, for instance, per 15 minutes time interval).

International standards

- ◊ Manual measurements are taken at a fixed time – 8.00 am daily from 203 mm (8 inch) diameter rain gauge.
- ◊ Self-recording Hattori type weekly or long-term recorders are used to collect rainfall data.
- ◊ 0.5 mm tipping bucket coupled with locally manufactured data loggers are used to collect rainfall data.
- ◊ Real time rainfall data for flood forecasting purpose are available from telemetric stations.

Methods of measurement in Bangladesh

Two principal devices are used by the BWDB for measuring rainfall. These are: (i) Non-recording rain gauge in which total precipitation for the preceding period, usually 24 hours, is determined by direct measurement of water retained in the gauge. (ii) Recording rain gauge, which makes automatic record of rainfall and graphic chart, is readily analyzed for the time distribution of precipitation. On the other hand, the BMD uses recording gauges in all stations and uses microprocessor technology to telemeter electrical signals from tipping-bucket gauges. BMD keeps records at 00, 03, 06, 09, 12, 15, 18, 21 GMT.

Wind turbulence affects the catch of rainfall. Tests have shown that rain gauges installed on the roof of a building may catch substantially less rainfall as a result of turbulence (10-20%). Wind is probably the most important factor in rain-gauge accuracy. Updrafts resulting from air moving up and round the instrument reduce the rainfall catchment. To reduce the effects of wind, rain gauges can be provided with windshields. Moreover, obstacles should be kept far from rain gauges (distance at least twice the height of such an object) and the gauge heights should be minimized (e.g. ground-level rain gauge with screen to prevent splashing).

2.4.6 Evaporation

Measuring equipment

The following instrument are used for evaporation measurement-

- ◊ Atmometers
- ◊ Evaporation pans
- ◊ Lysimeters

International standards

Manual measurements are taken at a fixed time – 8.00 a.m. daily from the Class-A Pan used by the U.S. Weather Bureau.

Measurement procedures

Direct measurement of evaporation or evapotranspiration from extensive water or land surfaces has not been achieved yet. There are several indirect methods, which are discussed below:

- Evaporation estimated by calculation
    - ◊ Energy-budget method
    - ◊ Mass-transfer method
    - ◊ Water budget method
    - ◊ Empirical formulae
    - ◊ Evaporation measurement using pans
- Actual evaporation
    - ◊ FAO Penman-Monteith method for estimating evaporation
    - ◊ Catchment water balance method
- For evapotranspiration calculation
    - ◊ Methods based on temperature (Thornwaite's heat index)
    - ◊ Radiation method
    - ◊ Penman method

Data collection by BWDB & BMD

The BWDB mostly uses the steel sheet Evaporation pan, which is almost similar to the Class-A Land pan used by the U.S. Weather Bureau. The pan used by the BWDB is of 48 inch diameter and 17 inches high. The water level is maintained 7 inches below the rim. The BWDB keeps records in millimetre per day while the BMD, which uses automatic recorders and telemetry to transmit data, keeps records at 00, 03, 06, 09, 12, 15, 18, 21 GMT.

2.5 Quality assurance

Hydrological time series data are sequences of numeric values that collectively represent fluctuations in their states. As the means of sensing these fluctuations are often approximations or prone to interference by other forces of nature, there are risks of errors, which are small or large, random or systematic, obvious or subtle. As the sensing is usually remote in collecting the data, it can be difficult for data collectors to detect and quantify potential errors.

Data collection and processing therefore need as many checks and fail-safe components as are reasonable to include. Data collectors, with knowledge of what and how they are measuring, can build in a good level of checking. Over the years this has improved as knowledge and techniques have progressed.

Hydrometric quality assurance (QA) begins with concise, useable (documented) procedures and traceably calibrated measuring equipment, and goes through sophisticated checks and statistics on the end product – in this case the data. QA systems are often modeled on recognized systems with a series of integrated elements.

QA involves provision of instruction manuals, training, internal auditing, calibration of equipment, monitoring non-conforming products and maintaining various records aimed at verifying and recording information on quality. The quality system encompasses many elements, including tracking the batches of data collected, range and sensibility checks, sensor calibration, tracking of data editing and archiving, and a structured data review and checking process.

2.5.1 Elements of data quality

For the processing of time series hydrological data the key elements that will be considered for grading the quality levels are as follows:

- ◊ Conformance to expectations: fulfilling arbitrary thresholds
- ◊ Following established procedures: as with geodetic standards
- ◊ Fitness for use: Truth in labeling (distinct roles of producer and consumer)

2.5.2 Guidance on quality control procedures

Quality control procedure should be based on the following criteria:

- ◊ All data should be subject to quality control and validation procedures.
- ◊ Procedures should aim at detecting errors as close to the source as possible.
- ◊ Original data values should be recoverable after quality control has been performed.

2.5.3 Definition of quality levels

As a guideline for data processing, NWRD data management will apply the following data levels.

Level 0: Original, direct instrumental reading. Collected from the field using manual gauges and stored in hard copy.

Level 1: Raw data in digital format. This may be available through automatic gauges or converting hard copies to digital form

Level 2: Data are referred for cross-checking.

Level 3: Level 2 data have been tested for significance.

Level 4: Level 3 data having gaps filled with the missing values.

Level 5: Homogeneous data fields derived by analysis or modeling techniques.

2.6 Present data quality scenario of Bangladesh

Quality, which is practiced during data collection, analysis and other different stages, is meant as existing quality. In Bangladesh the major sources of qualitative hydrometric and hydro-meteorological data are the BWDB and the BMD respectively. In most cases, obvious errors are found in the data. Some of the findings from processing NWRD data can be mentioned here, as these would lead to finding the best processing methods for this guideline.

Data quality is a very important consideration for any kind of analysis. Prior to undertaking the sophisticated analysis methods now available to water resource planners and managers, input data need to be thoroughly checked and institutional reforms made for ensuring that agencies responsible for data collection are also accountable for the quality of data they collect and disseminate. WARPO has started this process, and it is clear that much can be done to improve the data, but it takes time and considerable effort. The simple collection, generation and storing of data could not fulfill the requirements of water sector planners, as almost all data are in raw format and need at least a primary or obvious error check. This has brought forth the issue of quality data. In this context, the NWRD has started to check the obvious errors of time series and spatial data. Quality check of data is a long-term process, as it requires the involvement of data collecting agencies and field verification. So the process is on going. Initial quality checking of time series hydrological data layers has been done in the NWRD in consultation with NWMP expatriate and WARPO planners.

2.6.1 Findings of NWRD

To check the obvious errors of water level, discharge, groundwater, sediment, rainfall and evaporation data three main characteristics of time series data have been considered as criteria of variability check. These are:

Punching error

If this type of data is plotted against time it would follow a more or less systematic variability shape from where marked deviation within one time interval is quite impossible. These deviations are clear when plotted graphically. Therefore applying judgment, the sharp rise and fall that breaks the trend of the particular hydrograph remarkably have been considered as punching errors. It has been assumed that these errors have intruded while punching data or keeping records. As these records have been consequently deleted, they appear as missing data.

For discharge data this punching error is more difficult than for water level as discharge is depended on both water level and velocity of flow. Therefore error can be from three parameters i.e., water level, measuring or recording of velocity and from punching data. In summary, ISO (1983) identifies the following error types in measuring average flow velocity:

- ◊ Instrumental errors e.g., if the current meter behavior deviates from the calibration curve or application of incorrect flow angles relative to the normal to the transect.
- ◊ Type I error: exposure time of the local point velocity.
- ◊ Type II error: number of points in the vertical.
- ◊ Type III error: number of verticals in the cross-section.

Furthermore, in the compilation of discharge data from the flow velocity measurements additional errors are introduced due to random and systematic errors in the water level. To check the discharge data, sometimes water level data of the same time have been checked and, applying judgment, anomalous data have been marked. The example below shows the old data with punching errors and the data after being updated by the NWRD. Both old and updated time series data are available in the NWRD.

The phenomenon of obvious error described above is not true for all types of meteorological data. For example, to some extent evaporation data can be checked in this way, but for rainfall data the obvious error check is difficult with visual inspection or plots. Therefore to make the data free from obvious error, the NWRD has set some criteria and have deleted those erroneous data. The criteria applied are summarized below:

- ◊ All negative values have been deleted.
- ◊ Very high/low values have been marked but not deleted.

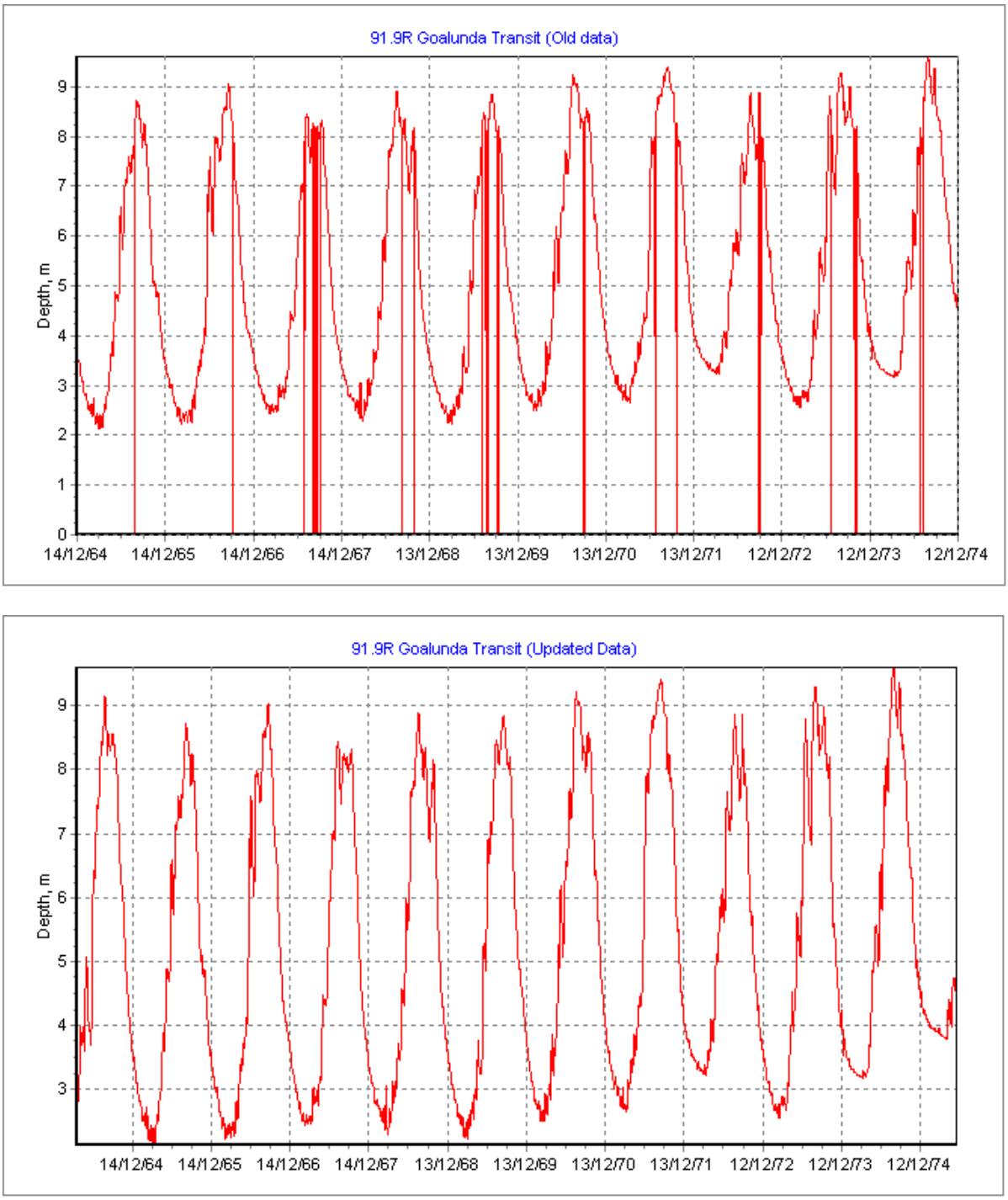


Figure 2.1: Water level data at Goalanda Transit, (a)Data with a lot of undershoots breaking the trend (b) A smooth hydrograph after removal of undershoots

Change in Trend

Water level is measured by data collecting agencies with automatic or manual gauges. Therefore, every time the gauge is shifted its position with the standard datum should also be read. However, if the reduced level is not read properly, gauge shift is reflected on the time series plot as 'change of trend'. These trend changes in discharge and water level data have been marked by the NWRD as obvious

errors for some selected stations and need to be corrected in consultation with the data collecting agencies. It can be mentioned here that changes in trend can occur for other reasons like construction of structures that can alter the flow or depth of rivers. Therefore, these types of errors should be treated carefully. During the FAP24 study, initiatives were taken to check zero values of gauges. About 50 of the BWDB gauges were checked for BM corrections and the results were published in a report (FAP-24). The variations were found in the range of –0.3m to +0.4m. The NWRD needs to find out whether the corrections of FAP24 have been incorporated in the BWDB data when transferred to the NWRD. Discussion should also be held with the BWDB for correcting the discharge and water level data.

For rainfall and evaporation, any kind of obstacle causing precipitation to fall in the gauges or evaporation to occur may lead to changes in the trend (FAP-2).

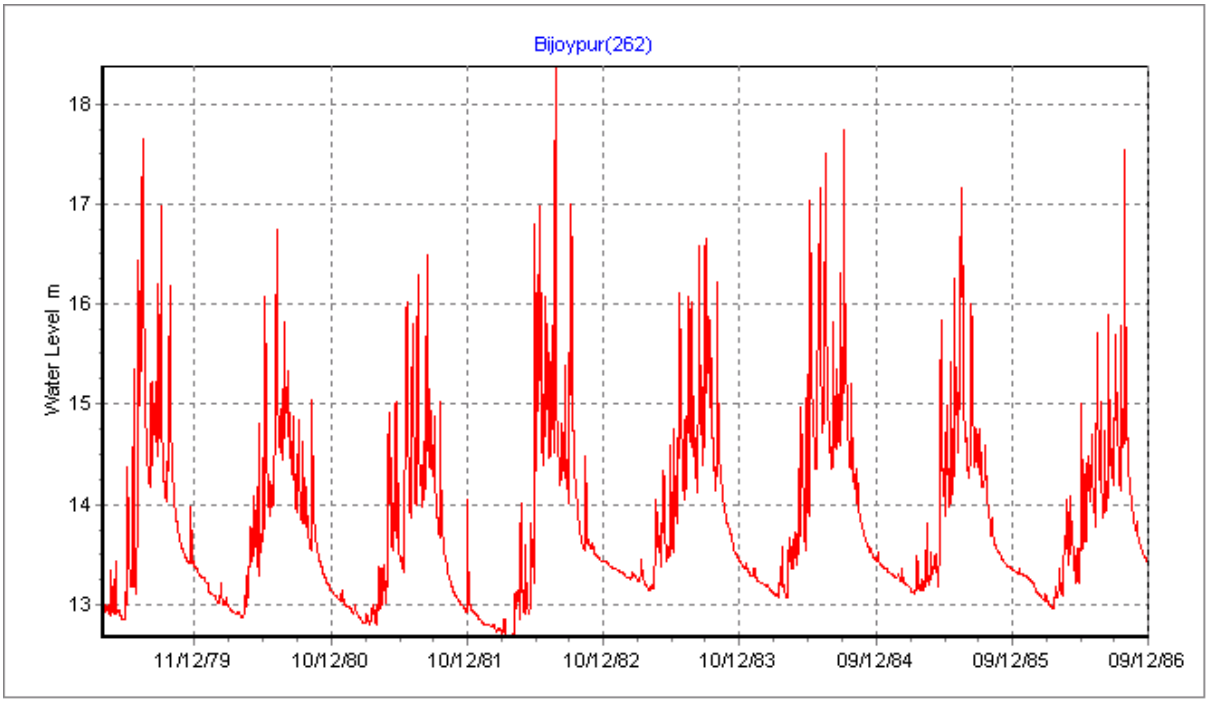


Figure 2.2: Water level data at Bijoypur showing change in trend

Missing Data

Time series data are collected at regular time intervals. When data is not collected or available for some period for some reason, the gap should be filled for analysis purposes. Therefore NWRD has noted the ranges of missing data that should be filled with proper methodology in the future. Most of the data layers are not complete and on an average 10-20% of the data are missing.

These were three main considerations for obvious error check of rainfall, evaporation, water level and discharge data in the NWRD. The data are not still error free. For example, in some cases data of same stations are found to be different when collected from different sources.

2.7 Quality framework

Table 2.7: Quality framework

Data	Task	Existing Checking			
		Bangladesh	**Reference**	**International**	**Recommend by**
Water level	Measurement procedures	Analogue recorder Digital recorder	BWDB	Analogue recorder Digital recorder	WMO
	Methods of measurement	Staff gauge Float operated gauge Pneumatic gauge	BWDB	Staff gauge Float operated gauge Pneumatic gauge	WMO
	Validation	Double mass Least square method Moving average	BWDB	Double mass Frequency analysis Moving average	WMO
	Filling missing value	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	BWDB	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	WMO
	Compilation	Frequency distribution T-test/Wilcoxon test F-test	BWDB	Frequency distribution T-test/Wilcoxon test F-test	WMO
	Goodness of fit test	Chi-square test Kolmogorv-Smirnov test	BWDB	Chi-square test Kolmogorv-Smirnov test	WMO
Discharge	Measurement procedure	Direct measurement	BWDB	-	-
	Methods of measurement	Velocity area method ADCP-EMF moving boat	BWDB	-	-
	Trend analysis	Double mass Frequency analysis Least square method Moving average	BWDB	Double mass Gamma family Moving average	WMO
	Filling missing value	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	BWDB	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	WMO
	Compilation	Frequency distribution T-test/Wilcoxon test F-test	BWDB	Frequency distribution T-test/Wilcoxon test F-test	WMO
	Goodness of fit test	Chi-square test Kolmogorv-Smirnov test	BWDB	Chi-square test Kolmogorv-Smirnov test	WMO
	Methods of measurement	Tape or rope scaled Automatic groundwater level recorder	BWDB	Automatic groundwater level recorder	WMO

Data	Task	Existing Checking			
		Bangladesh	**Reference**	**International**	**Recommend by**
Groundwater	Trend analysis	Double mass Frequency analysis Moving average	BWDB	Double mass Gamma family Moving average	WMO
	Filling missing value	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	BWDB	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	WMO
	Compilation	Frequency distribution T-test/Wilcoxon test F-test	BWDB	Frequency distribution T-test/Wilcoxon test F-test	WMO
	Goodness of fit test	Chi-square test Kolmogorv-Smirnov test	BWDB	Chi-square test Kolmogorv-Smirnov test	WMO
Rainfall	Methods of measurement	Weighing type Tilting bucket type Float type	BWDB	Weighing type Tilting bucket type Float type	WMO
	Trend analysis	Double mass Frequency analysis Moving average	BWDB	Double mass Gamma family Moving average	WMO
	Filling missing value	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	BWDB	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	WMO
	Compilation	Frequency distribution T-test/Wilcoxon test F-test	BWDB	Frequency distribution T-test/Wilcoxon test F-test	WMO
	Goodness of fit test	Chi-square test Kolmogorv-Smirnov test	BWDB	Chi-square test Kolmogorv-Smirnov test	WMO
Evaporation	Methods of measurement	Energy-budget method Mass-transfer method Water budget method Empirical formulae Evaporation using pans	BWDB	Energy-budget method Mass-transfer method Water budget method Empirical formulae Evaporation using pans	WMO
	Trend analysis	Double mass Least square method Moving average	BWDB	Double mass Gamma family Moving average	WMO

Data	Task	Existing Checking			
		Bangladesh	**Reference**	**International**	**Recommend by**
	Filling missing value	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	BWDB	Double mass Moving average Correlation & regression Normal ratio Arithmetic average National Weather Service	WMO
	Compilation	Normality test T-test/Wilcoxon test F-test	BWDB	Normality test Significant test Goodness to fit	WMO
	Goodness of fit test	Chi-square test Kolmogorv-Smirnov test		Chi-square test Kolmogorv-Sm. test	WMO

2.8 Recommendations for data collection

During data collection any missing value causes serious effect in development projects. Hydrological data series are subject to variations in time and space. Therefore, the missing values could be filled following standard guidelines and practice.

- ◊ As much as possible, data should be collected in a consistent manner
- ◊ Well-documented quality control should be applied to the collected data
- ◊ Data should be archived under appropriate conditions to maintain integrity.

2.8.1 Water level

Depending on the types of error to be expected in gauge readings due to the observation and operation practice, the following systematic checking procedure is recommended:

For each discharge station to be checked at least two adjacent water-level stations have to be selected for comparison. Annual tables should be made for the data to carry out a first check on variations in the estimated maximum rates of rise or fall.

Water-level time series for each discharge station should be plotted for every year together with the time series of two comparison stations. This gives a first visual impression of possible erroneous/shift in data.

2.8.2 Discharge data

The following are recommendations on qualitative data collection (FAP-24):

- ◊ Accurate gauging of hydraulic parameters is critical at primary gauging stations. If this is not possible, serious consideration should be given to relocating the stations.
- ◊ Discharge measurements should be made bi-weekly, weekly, or fortnightly, depending on the different phases of the hydro-graph.
- ◊ A DGPS positioning system should be pursued. This technique is developing quickly and provides a better accuracy and easier operation as compared with sextant or Decca positioning.
- ◊ Although propeller-type current meters are found to be sustainable, other instruments are developing quickly and should be considered for future implementation
- ◊ The traditional acoustic-type echo sounder should be maintained for depth measurements.

- ◊ Although manual staff gauges have been found to be sustainable, other devices that provide better accuracy should be considered.
- ◊ The velocity-area discharge estimation method, although found generally suitable, needs improved instrumentation and a potential means for optimization of the procedure has been identified. Also, a cost-effective application of the ADCP-EMF moving boat method can be sustainable under similar circumstances.

2.8.3 Groundwater

Execution of field level data measurements involves the following tasks:

- ◊ Activating reference stations and checking positional quality.
- ◊ Determining or checking accuracy e.g. transect perpendicularity at discharge transects.
- ◊ Ensuring station function.
- ◊ Preliminarily checking instruments.
- ◊ Actually executing work and logging data.
- ◊ On-line data processing, quality checking and recording on PC hard disks.
- ◊ Storing data and making two back-up tapes.
- ◊ Maintaining a record of executed observations in the log-book.

2.8.4 Sediment

The recommendations (FAP24) for sediment transport measurements are summarized as follows:

1. The human factor is an important consideration for any field measurement. A sophisticated and high-tech survey can fail if the equipment and method are not operated properly. Therefore, human resources development through training should be a fundamental step towards executing an effective survey.

2. A survey vessel, with its facilities for handling survey operations, is a basic requirement, which determines the failure or success of sediment measurements. Therefore, careful consideration should be given when selecting the vessel.

3. Utmost care should be given to measuring sediment transport parameters accurately at the primary gauging stations. Relocation of certain stations may be considered.

4. Different phases of the hydrograph show that it is necessary for suspended sediment measurements to be taken either be-weekly, weekly or fortnightly. Near-bed sediment measurements should be taken similarly unless a suitable relation is established with total sediment transport for all hydraulic conditions.

5. For an accurate sampling of suspended sediment, a sufficient recording length is required to obtain appropriate time averaged values. This means that either a time integrated sampler should be used, applying sufficient exposure time, or an instantaneous sampler should be used for obtaining a series of repeated measurements. This implies that the Binckley instantaneous sampler currently in use with the BWDB should preferably be used for time series to improve the accuracy or be replaced. A depth-integrated sampler such as a collapsible-bag can be used for details. A pump-method is the best alternative. Use of the Delft-bottle as a transport method, either as point-integrated or depth-integrated, is attractive but the loss of finer fractions should be quantified.

6. Some exercise carried out during the low-flow and high-flow seasons show that with an exposure time of 100s which was followed by FAP24, a relative error of 4 to 36% is made. As often suggested in literature, the error reduces to 2 to 30% with 300 s of

exposure time. Based on this, a 100s exposure time (integration time) appears to be a reasonable choice for a time-integrated sampler.

7. As accurate sampling of near-bed sediment is difficult and expensive, such sampling should be made to establish a suitable relationship with the total or suspended sediment transport if possible. If such a relationship does not exist it should be investigated whether a shift in location improves the situation. For routine estimates, near-bed sampling is not recommended.

8. For special near bed sediment measurement the Delft Bottle or pump sampling is recommended. For the bed load sampling a trap sampler like the BTMA or H-S is useful, especially in the mid flow range.

9. When a multi-point suspended sediment sampling is made over a vertical, at least 6-points over the vertical should be sampled with at least 1-point near to the bed (0.5m above the bed). The Straub or Chinese methods appear to give unacceptably huge errors.

10. For selecting the numbers and locations have suspended sediment-sampling verticals. A u2/h or a suspended sediment distribution over the river width should be used. The verticals have to be placed in such a way that the schematised concentration profile over the verticals fits best with the concentration distribution over the channel width. To effectively follow this method a reconnaissance survey should be made beforehand.

11. When large bedforms are present, it is not clear as to where a near-bed sampling vertical should be located. In such cases, a mid-point location on the stoss-face may provide reasonable average results. Such methods have been applied (Zaire, Jamuna) but not on a routine basis as they are time consuming.

2.9 Data entry & editing

As data entry is the primary concern of data collection it must be done carefully. Entry of data is followed by:

- ◊ Reading data from file - with full description of data.
- ◊ Electronic media- current meter data-record (date, no. of observation, gauge level, water level, discharge, width, wetted perimeter cross-sectional area and carefully gradient or fall) are given.

Proper correction of data is crucial for data processing. To get qualitative data, it is necessary to go through each record. Editing of data allows the examination of each mistake that would significantly facilitate further preparation. Data entry should be checked to:

- ◊ examine missing data;
- ◊ find out errors in preparation; and
- ◊ find out accessibility of input or record of data.

While keeping records some important information should also be kept with real data. These are:

2.9.1 Station /Series definition

The code name and location of the station where the data originated from are essential for specifying station/series database. Before the entry and editing of data the following terms must be kept in mind:

- ◊ station code and station data

- ◊ station log-book
- ◊ station history
- ◊ series characteristics

More specifically, these records can be elaborated as follows:

Station code and station data

Station code and station data comprises:

- ◊ station code
- ◊ station name
- ◊ district
- ◊ country
- ◊ latitude
- ◊ longitude
- ◊ altitude
- ◊ catchment area and
- ◊ agency

Station logbook

To maintain control on the status of processed and executed series, the actions can be stored in a station logbook. The logbook can only be opened and updated for the station or series location with the following features:

period : validity period of the remark in year, month and day.
date : the date of the log-book opened or updated
remark : remark on series or its processing
status : status of action
action : required action
user : name of user making the remark

Station history

Station history with specified information about the station is kept properly for future references. There is no constraint on the contents of the history file. Any information of the station, state of processing, maintenance should be entered under station history. New information of a station is to be added to the previous record.

Series characteristics

Series characteristics, which are indispensable for data processing, are measured properly for data processing. The following are included in series characteristics:

- ◊ station code
- ◊ data type, and
- ◊ time interval

The following characteristics are additionally needed:

- ◊ data unit
- ◊ time interval
- ◊ basic time interval
- ◊ time shift
- ◊ missing value
- ◊ minimum or maximum, and
- ◊ rate of rise or fall

However, non-equidistant series does not comprise time interval, basic time interval, or time shift.

2.9.2 Initial quality check

Regular checks should be done to make sure that standards are being followed. This may include regular testing of data added to the data set or may involve spot checks. This would allow users to pinpoint difficulties at an early stage to assist the correction of errors. A sample of collection stations can be resurveyed to check their accuracy and precision. If too many errors crop up, or if the surveyed area has changed greatly, the work is updated and corrected. Preliminary data checking can be done by different ways such as:

- ◊ cross checking
- ◊ general observation
- ◊ graphical presentation, and
- ◊ simple statistical analysis

2.10 Quality at different levels

Different levels are checked for the quality of data. The errors or mistakes can easily be detected during the use of data for different purposes. Data can be disseminated to different internal and external users to detect any accidental or gross error. Detected error can be returned to relevant organizations. The following directions are proposed to the different user levels:

- ◊ Cross-examination should be done before use or any kind of operation on data by users;
- ◊ Overall study on data is to be done regarding capturing and quality processing by professional users; and
- ◊ If any kind of gross mistake is found by users the data sets should be sent back to the database administrator to make necessary corrections.

Chapter 3

Data Processing

3.1 What is data processing

Data processing is aimed at ensuring logical and scientific integrity of databases. Processing contributes to data quality. The interface places many demands on the robustness and completeness of the descriptions of data structure. It requires standardization in terms of the data model and the storage formats.

3.2 Why processing

Data collection at field level has many influences over quality control. Different types of data are collected by different methods. The human factor is an important consideration for any field measurement. Properly collected data gives an absolute measure of the quality of data, which is the proper root for analysis of hydrometric and hydro-meteorological data. Different types of errors such as random, systematic or non-homogenous errors may be introduced in the observation/sensing, transmitting or recording during measurements. Although it is very difficult to maintain certain standards of reliability and accuracy of time series data, the collection and processing of data are crucial for many areas, especially for national databases.

3.3 Basic needs of data processing

The first step of data processing is to understand the following characteristics of the data set in order to develop a QC method, an analysis technique, and a presentation tool. The data character items are:

- ◊ basic statistics (averages, max. min, range, percentiles),
- ◊ data intervals / inconsistencies,
- ◊ missing data / incomplete data,
- ◊ frequency distribution,
- ◊ identification of problem area,
- ◊ setting up a strategy for QC and analysis, including guidelines on the use of blanks and flagging.

3.3.1 International standards for data processing

The processing steps for the data set are:

a) Reformatting of the acquired station-related data;
b) Quality control of the metadata (station co-ordinates);
c) Preliminary filling the missing Point Data Bank (PDB), i.e. merging the station-related data from different sources (national data sets, other collections) to one world-wide data set;
d) First objective analysis using uncontrolled records;
e) Automatic quality control of the data record;
f) Visual expert quality control and interactive correction with graphical workstation assistance (comparison with climate maps, orographical data, extreme-event-catalogs, etc.);
g) Second objective analysis using controlled data;
h) Filling the results of the second analysis run (step 7) into recorded data Grid Data Bank (GDB);
i) Correction of the results from step 8 with regard to systematic measuring errors;
j) Calculation of grid-related stochastic errors of precipitation results (step 8) from station density, value variance fields, individual data uncertainties, and uncertainty of the systematic measuring error corrections.

3.4 Present processing procedures

Data is collected and processed by agencies for their own purposes, such as for publishing annual reports or for specific research work and project development or monitoring. Agencies follow their own procedures to process data with whatever facilities they have available, some of which are described below:

3.4.1 Processing of BWDB

The BWDB Hydrology Division is one of the earliest users of computers for data storage and processing. The Processing and Flood Forecasting Circle of the BWDB processes water level, discharge, sediment discharge, rainfall and evaporation data. Water level data received from the field are adjusted for vertical shifts following check leveling and then keyed into the computerized system to display a graphical presentation. In cases of apparently incompatible data, reference is made to the field office for validation. After this initial check, the data is transferred to the database, and a second check is made by drawing a superimposed water level hydrograph of the upstream and downstream stations.

A spreadsheet developed for entering **discharge data** allows several trial curves such as WL vs. discharge, cross section area, etc, to be plotted to identify errors or incompatible data, and then transferred into the database. At the end of the year or when appropriate, data is retrieved for development of a rating curve. From the database, observed discharge data can be downloaded in a spreadsheet file and displayed as graphs of level vs. area, width, velocity and discharge. When calculating the mean daily discharge from observed water levels, the BWDB regularly corrects for shifts, applying corrections when the rating curve changes with time due to changes in the cross-sectional characteristics along the cross-section.

The BWDB maintains a network of stations collecting **rainfall data** over the country. Field sheets of daily rainfall and rainfall recorder charts from the Automatic Rainfall Recorder (ARFR) are received monthly, and hourly rainfall is entered from the charts. In examining the recorder chart and comparing the 24-hour rainfall total with daily rainfall from the manual gauge to check test the consistency of the data, the following points should be considered:

- ◊ Consistency of daily records with daily data of the adjacent stations.
- ◊ Consistency of monthly rainfall totals with monthly and annual rainfall totals in a sheet of 13 maps and with areal rainfall distribution by drawing isohytes.
- ◊ Depth duration data on an annual basis for each station. Plotting of this map also provides an additional check in the areal distribution of rainfall.
- ◊ Long-term trend in the average and standard deviation of rainfall by double mass analysis (and correction applied).

On receipt of the **monthly evaporation** statement, daily evaporation is calculated by balancing the rainfall and amount of water added/removed. High values of vaporisation are caused by rain associated with strong winds when the splashing of raindrops against the water in the pan, or leakage in the pan or unrecorded removal of water from the pan (drinking by birds and animals) cause extra loss of water. If the evaporation on a rainy day exceeds the highest daily evaporation recorded on a non-rainy day of the month, the evaporation value is bracketed as doubtful and replaced by estimated value. When daily evaporation data of a station is missing for one to three consecutive days, the missing data is estimated, but if the gap is longer, it is left unfilled.

The BWDB Groundwater Circle (GWC) collects data on observation wells, aquifer tests, lithology and water quality in a number of stations. Observation well data are collected every Monday and processed in hardcopy within 45 days and sent to the processing division. Consistency and error checking procedures are available, mainly using hydrographs superimposed on records from the same

station for different years (long term hydrograph correlation). Data digitization normally lags by three years due to a shortage of operators.

Aquifer tests are generally performed to estimate the transmissivity and storage co-efficient. The BWDB conducted over 200 tests at production wells installed by the BADC. Deep tubewells are used to pump from the wells at a constant rate. Sometimes shallow tubewells are used for pumping tests, but they cannot be run continuously for 72 hrs, so there is replenishment of the water surface during pumping. Data is stored in hardcopy in the processing circle and processed data is available in both hardcopy graphical and tabular formats. Test holes, drilled throughout Bangladesh, are concentrated in areas on which sub-surface data are the least available. The program consists of drilling deep (300-460m) and shallow (90-100m) holes, sampled at 3m intervals. Analyses in the field are done to produce borelogs and sent to the RRI for laboratory test (sieve analysis) and further confirmation of materials. The GWC has the facilities to enter the data in a database file as developed under the UNDP project and a small part of it has been digitized. The GWC also collects samples of **water quality data** in 117 stations during March-April every year. These samples are sent to the RRI for chemical analysis of 19 parameters. Data are recorded, processed and filed in hardcopy and used to prepare water quality maps by district.

The Flood Forecasting and Warning Center (FFWC) of the BWDB produces daily flood bulletins using data from 44 water level and 46 rainfall stations all over the country. The bulletins are sent in by wireless. The 3-hourly data are sent at 9am to 10am every morning including holidays and entered in computers within an hour for analysis. The telemetering transmission system cannot transfer the data acquired directly into the flood-forecasting model and the data has to be manually re-entered.

Data validation software is used to check for errors and consistency of the data, which are then processed in a forecasting model to generate flood warning information for each of the stations within two to four hours. The daily bulletin lists water levels above danger level and makes 24-hour and 48-hour forecasts.

3.4.2 Bangladesh Meteorological Department (BMD)

The BMD processes most of the meteorological parameters at the Processing Division both for surface and upper air all over Bangladesh round the clock. After receiving data from the Forecasting Section as teleprinter output, the data is entered into a digital format for further checking and processing. Hardcopy printouts are taken for correcting errors in punching. If errors are found, they send back the data to the field office using the same device. A technical committee consisting of SPARRSO, the DoE, and the Flood Monitoring Cell of the BWDB provides suggestions on the processing procedure and long term forecasting techniques. The Processing Cell produces tables of maximum, minimum and average weekly, fortnightly, monthly and annual values in report format. Digital data are sent to the World Meteorological Organization.

3.5 Need for improved processing facilities

Gauge readers living in the remotest areas of the country and facing some of the most adverse physical situations are responsible for the collection of data. These data have to be correct, mutually compatible and consistent with the laws of hydraulics and river mechanics. To be useful, the collected data require supervision and quality control at the field level and validation and quality control at the data processing office. A number of persons have been trained for data processing in different agencies, but it was observed that during the implementation of the FAP 24 training program, trained people were not available at the BWDB to operate the computer system for hydrology. Most professional engineers are reluctant to work in the processing circle for financial reasons, as no planning allowances are provided. The most important outcome is that the officers trained under the project have either been transferred or retired. The officers trained in data processing should not be transferred frequently. Arrangements are needed for the transfer of technology from the trained personnel to those who come to replace them.

In the past the Design Directorate had arranged a presentation on the completion of designs which was compulsory for all officers to attend. The design was improved with discussion and queries and junior officers were able to learn about new technology. The BWDB should adopt this principle to keep their technical manpower up-to-date about modern technology. Computerized departments need equipment and accessories (paper, disks/CDs, ink, backup facilities, etc) and software for optimum use. Sufficient funds have to be assigned to departments for these needs and future system upgrades. However, funds are rarely available on a sustained basis. The following are required to improve processing facilities:

- *Data collection* - methods and processes for the collection of new data.
- *Data analysis procedures* - the methods for computing, assembling a data set for a specified product.
- *Data integration* - Data integration procedures are the methods for combining various data sets into a unified, geographically harmonious data set
- *Quality control and quality assurance* - Quality control and quality assurance processes are respectively the methods followed to achieve a specified quality and the methods to check the quality of an existing data set.

3.6 Data processing at the NWRD

WARPO processes data for its own needs but it does not have sufficient manpower to handle the processing of different types of data for other users. It has modern equipment and processing facilities with a central database system. It acts as a data-collating agency, processing data and correcting it when errors are found and filling in gaps. Discrepancies are reported to agencies for comment and improvement of mutually approved data layers. WARPO shares information with other collecting agencies and aims to create a sustainable database with facilities for easy update in the future. However, none of this is sustainable without a legal framework, or, as a minimum interim measure, strong administrative support of the NWRD.

3.6.1 Data processing tools and applications

NWRD data are usually processed by the Data Collecting Agencies (DCA) prior to being sent to WARPO. However, there is considerable scope for adding value to it within WARPO so that it can better serve user needs. A number of tools have been developed which can be applied to transform, combine and present data in different ways. ArcView, web-based and Oracle-based tools and applications have been developed for front-end users for searching, downloading, exporting, viewing, analyzing and data sharing. A number of system management tools have also been developed for updating the database. A brief description of some of the tools developed under the NWRD is given in the following and are more fully described in a user manual.

The first way of adding value is to standardize formats so that the tools can be applied to all data of a particular type, irrespective of the format used by the original DCA. An Import Tool may therefore be develop to allow easy transfer of data from the formats used by particular agencies into the NWRD format. It is not considered necessary or desirable to insist that all agencies use a common format. Since potential users use different formats, an Export Tool has been developed to allow users to take copies of the data formatted to their own requirements. These tools also allow DCAs to take back copies of their own databases if needed, either as the original set or with further processing added. To export data to different external formats from the NWRD, a Generic Export Tool has been created as a stand-alone application developed in Visual Basic,

which can export the data to Excel, Access, Comma and Tab Delimited Text format. The subset of tabular data maintained in the Oracle database can be selected and exported to any location. This tool can also export shape files with associated information to any desired location.

A Time Series Viewer can be used to display, edit, export and import any time series data. Data from the SQL server can be retrieved for display as chart for a single station or more than one station for any specified time interval. This list will be extended in the future in response to user demand. Modern computing facilities make it relatively easy to develop tools and it is tempting for WARPO to develop further tools simply because it is possible. This is time-consuming and with scarce human resources, it is important to concentrate on tools for which there is an established demand from users. In addition, powerful tools make it easy to forget the limitations of the original data set. A smooth surface can be created and presented as a coloured contour map from very few point observations. The above tools were developed after discussion with NWMP consultants and WARPO planners in order to assist in different types of analysis. The potential users of these tools are still not well defined. Future users of these tools may be different projects and organizations dealing with spatial and time series data (such as CEGIS or the IWM).

3.7 Types of Error

In the historical data series for a hydrological event (e.g. rainfall) if any abnormal value is found which cannot be obtained from the deterministic or stochastic processes of hydrology, it can be concluded that the data for a particular year, month or day used in analyses are faulty. Several types of errors may exist, including random, systematic or input data errors.

Systematic errors

a. Systematic error occurs when the sign of the error tends to persist over a number of time intervals.

b. Inadequate representation or misrepresentation by a hydrological process model is one of the causes of systematic error.

c. A further cause of systematic error is the use of a period of record in the calibration of parameters which is not representative of the long term and which, for example, does not contain enough events of critical magnitude to calibrate key parameters. Consequently, hydrological responses of critical magnitudes (e.g. floods or low flow sequences) may be simulated incorrectly.

d. In the solution of storage problems for synthesizing flows, the existence of systematic errors may be very serious.

Random errors

a. Random errors can be defined as errors, which occur when a model shows no tendency to over estimate or underestimate for a number of successive time intervals.

b. Random errors can be identified by checking for the conservation of means and variances.

Attribute data errors

a. The existence of data errors must also be recognized. These input errors include errors in rainfall, evaporation, temperature and stream flow.

b. The following types of errors are associated with flow and rainfall data:

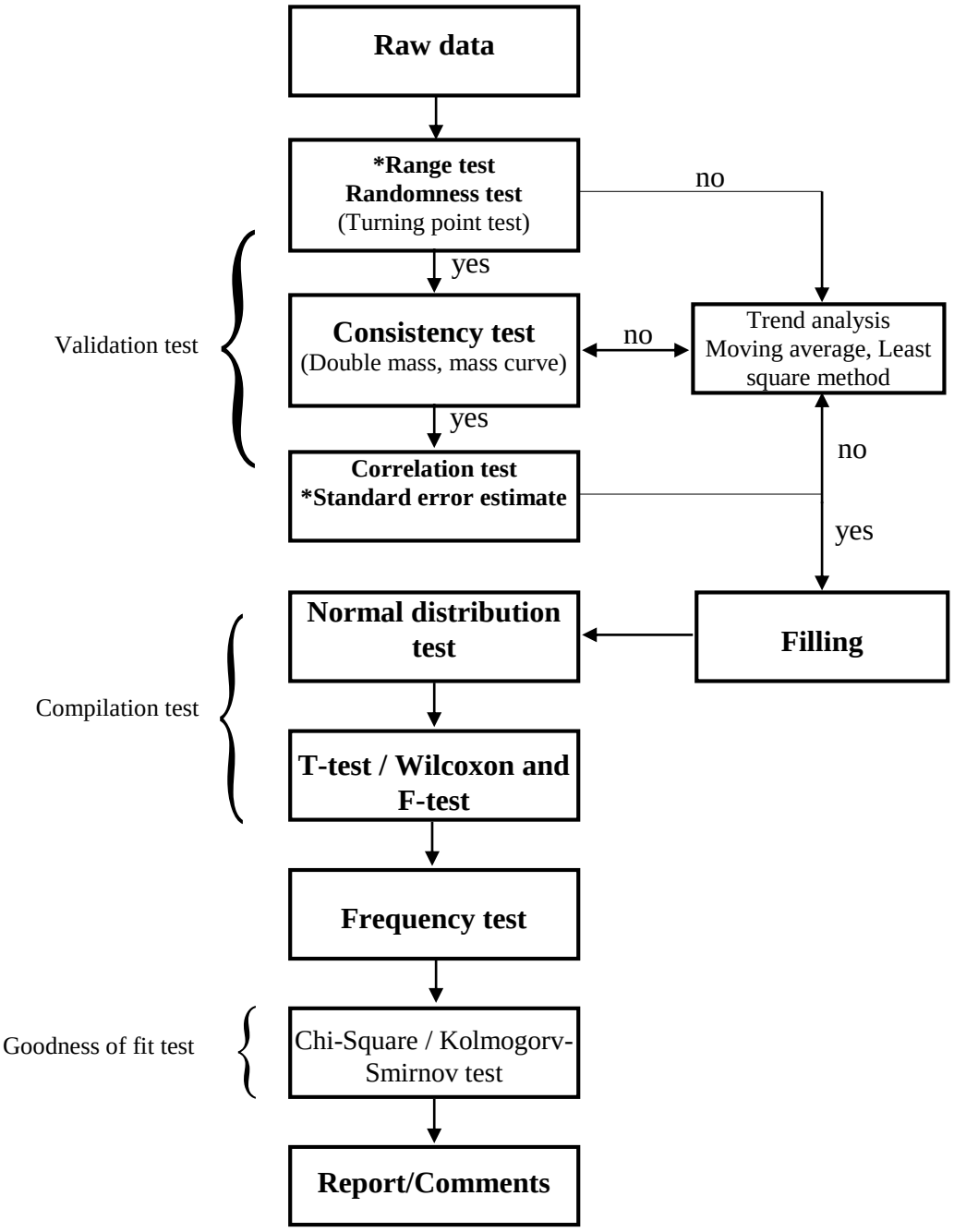
    ◊ Method of measurement errors;

- ◊ Flow data errors due to inaccurate rating curves;
- ◊ Mechanical and digitizing errors associated with flow data;
- ◊ Flow data errors due to poor or inaccurate manual observations, and
- ◊ Flow data errors due to the influence of catchment development or rainfall data errors.

It could thus be concluded that most errors can be detected by normal checking procedures and common sense and that basic checks should be carried out on all hydrological data, since some less obvious errors can influence model result significantly.

3.8 Generalized data processing steps

Flow chart



Note: '' means sometimes essential.*

The definition of the tests can be expressed by the following tests:

Raw data: Data collected from different organizations.

Randomness test: Turning point test is done for checking the randomness of data.

Consistency test: Double mass analysis shows the trend or inconsistency of a series.

Trend analysis: Moving average method is used for trend correction.

Regression and Correlation test: To find the best association of data in a series.

Filling: Suitable method is used for filling gaps.

Normal distribution: Normal distribution is used for the selection of suitable significant test (T-test/Wilcoxon test).

Frequency test: Mass curve is used for the test to see any shifting on the tested and base series.

Significance test: T-test or Wilcoxon test is used for finding the significance value.

Goodness to fit test: Chi square test is used for checking the good fit of a series.

3.9 Recommendations

Data storage and aggregation

Data can be stored in paper base files, computer files or in databases. It is advantageous to store data in computer files or in databases. However, the original paper files must be well guarded and analyzed carefully for what information to be stored in a database and in which time step. Data with significant change in values are picked up by data aggregation to avoid unnecessary readings.

Data acquisition and transfer

Planning, designing and management of water resources or water usage systems require information on past and present behaviors in time and space of all process and boundary conditions affecting relevant water quantity and quality parameters. This information is provided by an Information System, which covers data operations of the following kinds:

Data collection comprising

- ◊ Sensing
- ◊ Recording and
- ◊ Transmission

Data processing comprising

- ◊ Preparation
- ◊ Entry and transfer to the database
- ◊ Validation
- ◊ Correction
- ◊ Filling in missing data
- ◊ Compilation and analysis
- ◊ Retrieval and
- ◊ Dissemination or publication of annual report.

3.10 Quality gradation

The following quality control levels are applied to NWRD data layers:

QC 1: Identification of obvious errors, including human error and instrumental failure.

QC 2: Identification of errors that becomes apparent when data from different sources are compared.

QC 3: Adjustment by processing designed for the intended use of data, e.g. adjustment for biases, change of algorithms, etc.

3.11 Approach/ Techniques of hydrologic processes

3.11.1 Hydrologic process

Any hydrologic environment consists of water inputs, environmental responses and the output. For example, with rainfall as the input, infiltration and run-off are the environmental response and output respectively. Usually the output from one environment becomes input for another. This union of three-in-one input, response and output may be described as the basic parts of a hydrologic system, while each of these parts represents a hydrologic process. Natural hydrologic processes are never deterministic but are a combination of various deterministic and stochastic processes.

3.11.2 Deterministic processes

These are the processes of hydrology that are the results of physical, chemical and biological deterministic laws. For example, a rating curve of the stage-discharge relationship of a river cross-section with a fixed bed is a unique function and is thus a deterministic relation giving the same discharge for the same discharge. The case is opposite when the bed is changeable where many random factors influence the stage-discharge relationship.

3.11.3 Stochastic processes

These are the processes of hydrology, which are governed by the laws of chance, e.g. phenomena such as precipitation, evaporation, runoff, etc. Strictly speaking, there are no pure deterministic hydrologic processes. It is , in fact, a combination of deterministic and stochastic processes or predominantly stochastic processes. Therefore, hydrology cannot be fully understood, described and applied to water resources developmental projects without the extensive use of methods of probability theory. It is important to stress that the terms chance, random, probabilistic and stochastic are considered synonyms in this text; they all refer to phenomena subject to the laws of chance.

Chapter 4

Time Series Analysis

4.1 Concept

A sequence of values collected over time on a particular variable is a time series. Time series may be conveniently divided into two types: continuous and discontinuous. A time series which is given at a series of discrete time points, t_i where, i = 1,2...n is defined as a discontinuous time series, even though the variable itself may be continuous. A continuous time series is one in which the time variable is continuous, even though the phenomenon being described may not necessarily be continuous for a given 't'. Continuous time series must be converted into a discrete form prior to analysis.

A time series is a sequence of values arranged in order of their occurrence and characterized by statistical properties. The time interval for a discrete series might vary from a day or less to a month or a year. The selection of the time interval for representing a continuous series by a discrete series has two aspects:

- ◊ the extraction of optimal information in observing a variable, or from hydrologic data.
- ◊ the optimization between the accuracy of results and the computation cost in solving various problems.

Four types of time series are common in hydrology. These are:

- ◊ Full series
- ◊ Annual series
- ◊ Partial-duration series, and
- ◊ Extreme value series

4.2 Analysis of time series

There are two main goals of time series analysis: (a) identifying the nature of the phenomenon represented by the sequence of observations, and (b) forecasting (predicting future values of the time series variable). On the other hand, it may be referred to as the detection and quantitative description of the generating process that is characteristic of a given set of observations. Time series analysis includes the following types of analysis:

- Correlation analysis
- Spectral analysis
- Range analysis
- Run analysis
- Storage analysis

4.2.1 Correlation analysis

Correlations are done to identify the independence of time series. Correlation tests with a certain bend of confidence limit albeit double mass analysis are generally the most appropriate means of checking data consistency. Time constraints mean that it was impracticable to produce them for all stations. A prior quality check by cross-correlation can be carried out to identify stations where double mass analysis would be particularly beneficial. A correlation matrix can be produced for a group of ten neighboring stations, and any station, which showed poor correlation with more than three of the other nine stations, can be assumed to require further checking. The correlation can be carried out on monthly data for the wet season only, since differences in dry season rainfall is less significant than overall mean

rainfall. In general, the correlation coefficients would be higher if dry season months are included in the analysis. The definition of a "poor" correlation is necessary. There is no particular high value for correlation between monthly records at nearby stations, but a higher threshold would indicate that many stations have doubtful data. If more time is available on this or a subsequent project is undertaken, it would allow further examination of stations with borderline correlation. Correlation analysis covers the computation of:

- ◊ Auto-covariance function
- ◊ Auto-correlation function
- ◊ Cross-covariance function
- ◊ Cross-correlation function.

4.2.2 Spectral analysis

Spectral analysis provides an alternative approach to identifying the generating process that underlies an observed time series. On the other hand, auto correlation analysis operates in the time domain. The relation between spectra density (y-axis) and frequency in cycles per unit of time (x-axis) are plotted for spectral analysis. The smoothed auto-spectral estimate C_{xx} (f), for f=0, ½ is calculated from: (H.M.V.3.00):

$$C_{xx}(f) = 2\{c_{xx}(0) + 2\sum_{k=1}^{M-1} c_{xx}(k)w(k)\cos(2\pi fk)\} \qquad (4.1)$$

where :

f = frequency in cycles per time interval, computed at spacing $1/(2N_f)$, where, N_f is 2 to 3 times M

N_f = number of frequency points

C_{xx} (k) = autocovariance function at lag k

M = truncation point or maximum lag of the autocovariance function used to estimate the autospectrum; clearly M is conditioned by:

$M \leq L_{max}$

w(k) = window function.

Following window w(k) for k = 1, M-1 according to Tukey is used to smooth the spectral estimate:

$$w(k) = \frac{1}{2}(1 + \cos(\frac{\pi k}{M})) \qquad (4.2)$$

4.2.3 Range analysis

The quantities are computed from the accumulative departures from the mean S_i for i = 0, N and with S_0:

$$S_i = \sum_{j=0}^{i} (X_j - m_x)c_f \qquad (4.3)$$

where:

m_x = average of x(i), i = 1, N

c_f = conversion factor to transfer intensities into volumes.

4.2.4 Run analysis

A run is and excursion above or below the crossing level i.e. bounded by an upcrossing and a downcrossing or a downcrossing and an upcrossing. The positive and negative runsums RS^+ and RS^- respectively, are computed from:

$$RS^+ = \sum_{i=j}^{K} (x_i - x_c) c_f \qquad (4.4)$$

where:

j = location of an upcrossing
k = location of the next downcrossing
c_f = conversion factor to transfer intensities into volumes.
x_c = crossing level

$$RS^- = \sum_{i=k}^{m} (x_c - x_i) c_f \qquad (4.5)$$

where:

k = location of the next downcrossing
m = location of the next upcrossing

4.2.5 Storage analysis

The sequent peak algorithm water shortage or equivalently storage requirements without running dry are computed for various draft levels from the reservoir. The algorithm considers the following sequence of storage:

$$S_i = S_{i-1} + (x_i - D_x) c_f$$

$$i = 1, 2N \; ; \; S_0 = 0 \qquad (4.6)$$

where:

x_i = inflow
D_x = $D_L . m_x$
m_x = average of x_i, i = 1, N
D_L = draft level as a fraction of m_x
c_f = multiplier to convert intensities into volumes (time units per time Interval)

The local maximum of S_i large tan the preceding maximum is sought. Let the locations be k2 and k1 respectively with k1. The largest non-negative difference between S_{k1} and S_i, $_i$ =k1.., k2.., is then determined, which is the local range. This procedure is executed twice the actual series x_i; hence the series x_i is used twice in sequence: $x_i = x_{N+i}$. In this way initial effects are eliminated. Besides the above-mentioned range analysis, run analysis can be done for time series analysis.

4.3 Interpolation

The following interpolation methods can be considered to fill missing data:

- ◊ linear interpolation,
- ◊ block-type filling: block type filling data is the replacement of missing data by the last non-missing values before the gap.
- ◊ use of series relation: relation equations are used for filling the missing data.

◊ spatial interpolation.

4.3.1 Linear interpolation

In a number of cases gaps in a series can be filled by linear interpolation between the last value before the gap and the first one after, provided that the distance over which interpolation takes place is not too large. The use of linear interpolation requires the following options:

◊ selection of types of linear interpolations

◊ selection of series interval

◊ period to be considered for filling.

◊ maximum interpolation distance (expressed as a number of time intervals). This means that gaps larger than this maximum will not be filled-in.

4.3.2 Block-type filling-in

Filling–in data according to block-type filling-in comprises the replacement of missing data by the last non-missing value before any gap.

4.3.3 Series relation

Series relation is used to fill-in missing data, provided that the standard error in the fit is small. Different types of equation (polynomial, simple linear, exponential, power, logarithmic, hyperbolic, multiple linear etc.) are used for fill–in missing data.

4.3.4 Spatial Interpolation

Complex data sets are simplified into a low dimension to filter errors. The spatial interpolation technique is applicable to quality and quantity parameters with spatial characters, such as rainfall, temperature, evaporation, etc., but sampled at a number of stations. Missing data at a test station are estimated by weighted averages of observations at neighbour stations. The weights are inversely proportional with some power of the distance between the test station and the neighbour stations.

Chapter 5
Basic Statistics

5.1 General

Statistics is a set of methods that are used to collect, analyse, present, and interpret data. Statistical methods are used in a wide variety of occupations and help people identify, study, and solve many complex problems. The statistical methods enable decision-makers and users to make informed and better decisions about uncertain situations.

To compete statistical analysis successfully users and decision-makers must be able to understand the information and use it effectively. Statistical data analysis provides hands on experience to promote the use of statistical thinking and techniques to apply in order to make educated decisions according to need.

Studying a problem through the use of statistical data analysis usually involves four basic steps.

1. Defining the problem
2. Collecting the data
3. Analyzing the data
4. Reporting the results

5.2 Objectives of statistics

The objectives of statistics in hydrology may be listed as follows:

- ◊ Interpretation of observations;
- ◊ Search for hydrology probabilistic regularities;
- ◊ Extraction of maximum information from hydrologic data; and
- ◊ Presentation of hydrologic information in condensed form as graphs, tables of numbers, and mathematical equations, basically for decision-making in water resources planning.

5.3 Scope of analysis

The main subjects covered in statistical analysis is:

Plausibility of hydrological data: All measured hydrological data or the value derived therefrom contains errors. Quality control test for the examination of possible errors in hydrological data is recommended.

Theoretical probability distributions: It covers the concept of distribution functions, moment and quintiles, empirical distributions, plotting positions, partial series, and distribution functions, including the Gaussian or normal distribution, log normal, Pearson type III, log Pearson III and Gumbel.

Correlation and regression analysis: Theoretical considerations are useful to indicate the existence of correlation between hydrological variables. The problem is then to determine the type and degree of correlation.

Time series analysis: The main objective of time series analysis in hydrology is to disaggregate the trend, the periodic and the stochastic elements from a measured hydrograph. The methods of time

series analysis are also used to evaluate the accuracy estimates of statistical parameters and to determine whether a time series is sufficiently long.

5.4 Some basic statistics

The following basic statistics are needed in statistical analysis:

Minimum: $x_{min} = \min (x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n)$ (5.1)

Maximum: $x_{max} = \max (x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n)$ (5.2)

Median: Middle value of the ranked series X_i

Mode: Value of X, which occurs with the greatest frequency; i.e. the middle value of the class with the greatest frequency. If classes have equal greatest frequency then the middle value of the class with the lowest class levels will be indicated as mode

Skewness: The lack of symmetry of a distribution is called skewness or asymmetry. The population skewness is given by

Skewness = 3(Mean - Median)/Standard deviation

5.4.1 Product Moment

Several summary statistics can describe the character of the probability distribution of a random variable. Moments and quantiles are used to describe the location or central tendency of a random variable. Mean of a random sample is the first moment, the second moment about the mean is the variance. The standard deviation is the square root of the variance and describes the width or scale of a distribution. These are examples of product moments because they depend upon powers of *X*.

Arithmetic mean:

$$\bar{x} = m_1 = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5.3}$$

where, $\bar{x} = m_1$ = Arithmetic mean and n = No. of data x_i = Individual sample.

Standard deviation (SD):

$$\text{SD, } m_2 = \left[\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right]^{1/2} \tag{5.4}$$

Coefficient of Variance (CV):

$$CV = m_2 / \bar{x} \tag{5.5}$$

Coefficient of Skewness (CS):

$$\text{CS, } m_3 = \frac{\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^3}{m_2^3} \tag{5.6}$$

Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have

heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

$$\text{Kurtosis, } m_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{m_2^4} \tag{5.7}$$

5.4.2 Probability Weighted Moment

A new class of moments, called Probability Weighted Moments (PWM) was introduced by Greenwood et al., (1979). The PWM is particularly suitable for distributions whose CDF can be inverted.

The PWM estimator b_0 is the sample mean. To estimate PWMs of a sample, the data are arranged in ascending order of magnitude so that $x_1 \le x_2 \le \ldots\ldots \le x_i \le \ldots\ldots \le x_n$ where n is the sample size.

The first four unbiased PWM estimators are

$$b_0 = x_{mean} \tag{5.12}$$

$$b_1 = \frac{1}{n}\frac{\sum_{i=1}^{n}(i-1)x_i}{(n-1)} \tag{5.13}$$

$$b_2 = \frac{1}{n}\frac{\sum_{i=1}^{n}(i-1)(i-2)x_i}{(n-1)(n-2)} \tag{5.14}$$

$$b_3 = \frac{1}{n}\frac{\sum_{i=1}^{n}(i-1)(i-2)(i-3)x_i}{(n-1)(n-2)(n-3)} \tag{5.15}$$

It is obtained from these equations that PWM estimators are linear functions of data.

5.4.3 L-Moment

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

$$l_1 = b_0 \tag{5.16}$$

$$l_2 = 2b_1 - b_0 \tag{5.17}$$

$$l_3 = 6b_2 - 6b_1 + 6b_0 \tag{5.18}$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{5.19}$$

The analogous of conventional product moment ratio estimators, such as CV, CS and kurtosis are L-moment estimators L-cv, L-cs and L-kurtosis.

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

$$t_r = l_r/l_2 \tag{5.20}$$

These are dimensionless quantiles, 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 co-efficient of variation (standard deviation divided by mean), is the L-cv, defined by

$$t_2 = l_2/l_1 \tag{5.21}$$

It takes values between 0 and 1.

L-moment ratio diagram can be used to compare the L-skewness and L-kurtosis relations of different distributions and data samples. This gives a visual indication of which distribution may be expected to give a good fit to a data sample or samples.

The goodness of fit for candidate distributions are compared by using the probability plot correlation co-efficient and the root mean square deviation. Then the capability of a distribution to simulate the site to site variations in the statistical behavior of flood samples are assessed with the help of L-moment ratio scatter diagram.

5.4.4 Commonly used Distributions

Binomial

Gives probability of exactly successes in n independent trials, when probability of success p on single trial is a constant. Used frequently in quality control, reliability, survey sampling, and other industrial problems.

Negative Binomial

Gives probability similar to Poisson distribution when events do not occur at a constant rate and occurrence rate is a random variable that follows a gamma distribution.

Poisson

Gives probability of exactly x independent occurrences during a given period of time if events take place independently and at a constant rate. May also represent number of occurrences over constant areas or volumes. Used frequently in quality control, reliability, queuing theory, and so on.

Normal

The normal distribution is an extremely important probability distribution in many fields. It is also called the Gaussian distribution. It is actually a family of distributions of the same general form, differing only in their *location* and *scale* parameters: the mean and standard deviation. The standard normal distribution is the normal distribution with a mean of zero and a standard deviation of one. Because the graph of its probability density resembles a bell, it is often called the bell curve.

Gamma

A basic distribution of statistics for variables bounded at one side - for example x greater than or equal to zero. Gives distribution of time required for exactly k independent events to occur, assuming events take place at a constant rate. Used frequently in queuing theory, reliability, and other industrial applications.

Exponential

Gives distribution of time between independent events occurring at a constant rate. Equivalently, probability distribution of life, presuming constant conditional failure (or hazard) rate. Consequently, applicable in many, but not all reliability situations.

Log-normal

Permits representation of random variable whose logarithm follows normal distribution. Model for a process arising from many small multiplicative errors. Appropriate when the value of an observed variable is a random proportion of the previously observed value. In the case where the data are lognormally distributed, the geometric mean acts as a better data descriptor than the mean. The more closely the data follow a lognormal distribution, the closer the geometric mean is to the median, since the log re-expression produces a symmetrical distribution.

Weibull

General time-to-failure distribution due to wide diversity of hazard-rate curves and extreme-value distribution for minimum of N values from distribution bounded at left. The Weibull distribution is often used to model "time until failure" In this manner, it is applied in actuarial science and in engineering work. It is also an appropriate distribution for describing data corresponding to resonance behavior, such as the variation with energy of the cross section of a nuclear reaction or the variation with velocity of the absorption of radiation in the Mossbauer effect.

Extreme value

Limiting model for the distribution of the maximum or minimum of N values selected from an "exponential-type" distribution, such as the normal, gamma, or exponential.

5.5 Types of statistical test

5.5.1 Parametric statistics

- ◊ Appropriate for interval/ratio data
- ◊ Follow normal distribution
- ◊ For large data parametric tests are so robust
- ◊ For small data parametric tests are not so robust.

5.5.2 Non-parametric test

- ◊ Used with normal/ordinal data
- ◊ Does not follow the normal distribution
- ◊ For large data nonparametric tests are so powerful.
- ◊ For small data nonparametric tests are not powerful.

5.5.3 Recommends for choosing a statistical test

This guideline has discussed different statistical tests. To select the right test we need to know the followings:

- ◊ Collected data type and
- ◊ Goal that should be resolved.

5.5.4 Null hypothesis test

A statistical procedure designed to assess the degree of plausibility of a given statement or null hypothesis, concerning for example the value of the population mean, on the basis of the evidence provided by a sample. Thus a test conducted at the 5% significance level means that the null hypothesis will be accepted if the given sample value is one of those which would be expected from 95% of all random samples. It is important to realize that a hypothesis test cannot lead to a definite conclusion. Generally, a statement theorizing that there is no difference between (the null) groups of subjects, or factors to be studied.

5.6 Statistical features:

Descriptive statistics

- Arithmetic mean
- Standard deviation
- Standard error
- Skewness
- Kurtosis
- 95% confidence limit for mean
- Min value
- Max value
- Sample range
- Number of samples
- Median
- Geometric mean
- Variance
- Average deviation

Parametric tests

- Frequency analysis
- T-test
- Regression
- Correlation
- F-test
- Z-test of a correlation coefficient
- Fisher cumulant test for normality of a distribution.
- Chi-square test
- Kolmogorov-Smirnoff
- Sign test

Nonparametric tests

- F-test
- Chi-squared
- Wilcoxon's test
- Wilcoxon-Mann-Whitney U-Test
- Spearman rank correlation test
- Kolmogorov-Smirnov test
- Median test
- Serial correlation test
- Turning point test for randomness of fluctuations
- The difference sign test for randomness in a sample
- Rank correlation test for agreement in multiple judgments.

Chapter 6
Data Validation

6.1 General

The following options are typically used for data validation purposes. Time series data are applied for equidistant time series. But some of them are applied for non-equidistant or combined equidistant and non-equidistant time series. Data validation should go through the following stages:

*Screening***:** Data screening is performed for proper listing of series for easy references and first check on the range of data.

Time series plots: Time series plots show the sum or accumulated differences from the mean.

Relation curves: A relation curve gives a functional relationship between two series which is used for detection or random errors, systematic errors, missing data and forecasting purposes.

Double mass analysis: Double mass analysis is a technique to detect possible inhomogeneities in series, like jumps, trends etc.

Series homogeneity tests: It is performed by different statistics, which have been discussed later.

Spatial homogeneity test: The test is applicable to quality and quantity parameters with spatial characters like rainfall, temperature, evaporation, etc. to the neighbor stations.

6.2 Test of trend

Purpose and objectives

Trend analysis is used for finding the inhomogenity of a series. Before estimating trend the first thing to find is whether or not any trend is present at all. And to check this, tests for randomness following test are performed on the time series. If any trend is found then trend analysis is necessary for removing the trend before analysing the data.

6.2.1 Turning point test

The turning points test is used check whether the time series verify the following hypothesis:

- ◊ H_0: The series is a random no trend series.
- ◊ H_a: The series has trend and/or has autocorrelation errors.

Description

Trend of a series is estimated in this step. For an independent stationary series, N is number of data, T is approximately normally distributed with,

$$\text{mean} = 2(N - 2)/3 \text{ and}$$

$$\text{standard deviation} = [(16N - 29)/90]^{0.50}$$

Standard value of T,

$$t = \text{T-mean} / \text{Standard deviation}$$

If *t* value ranges within ±1.96 for 5% level of significance, it is decided that the series is of random sequence. If series is not random, the trend value should be estimated by Least Square Method or Moving Average Method.

Input

Range of data set is divided by a certain block and arranged in ascending/descending order to determine the number of turning.

Output

Each limit identifies one turning point. Cumulative number of turning is determined whether data is random.

Operational requirement and restrictions

Equal limit is used for finding turning point and record of data set should be long.

6.2.2 t-Test for detecting linear trend

The t-test is used to check whether the time series data verify the following hypothesis:

- ◊ *H_0: The series is a random no trend series.*
- ◊ *H_a: The series has linear trend.*

Description

Assume that y_t, $t = 1, \ldots\ldots\ldots, N$ is an annual time series and n = sample size. A simple linear trend can be written as

$$y_t = a + bt$$

where a and b are the parameters of the regression model. Rejection of the hypothesis b = 0 can be considered as a detection of a linear trend. The hypothesis that b = 0 is rejected if

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right| > T_{1-\frac{\alpha}{2},\nu}$$

in which r is the cross-correlation coefficient between the sequences $y_1,\ldots\ldots,y_N$ and $1,\ldots\ldots\ldots\ldots,N$, and $T_{1-\frac{\alpha}{2},\nu}$ is the $1-\alpha/2$ quantile of the student t distribution .

6.2.3 Mann-Kendall Test

The Mann-Kendall test is used to check whether a time series follows a trend or not. The hypotheses are as follows:

- ◊ *H_0: The series is random no trend series*
- ◊ *H_a: The series has trend either upward or downward*

Description

This is a non-parametric test which tests for a trend in a time series without specifying whether the trend is linear or nonlinear. Consider the annual time series y_t, $t = 1, \ldots\ldots\ldots, N$. Each value $y_{t'}, t = 1, \ldots\ldots\ldots, N-1$ is compared with all subsequent values $y_t, t = t'+1, t'+2, \ldots\ldots\ldots, N$, and a new series z_k is generated by

$$z_k = 1 \qquad \text{if } y_t > y_{t'}$$

$$z_k = 0 \qquad \text{if } y_t = y_{t'}$$

$$z_k = -1 \qquad \text{if } y_t < y_{t'}$$

in which k = $(t'-1)(2N-t')/2+(t-t')$. The Mann-Kendall statistics is given by the sum of the z_k series

$$S = \sum_{t'=1}^{N-1} \sum_{t=t'+1}^{N} z_k$$

This statistic represents the number of positive differences minus the number of negative differences for all the differences considered.

The test statistic for N>40 may be written as

$$u_c = \frac{S+m}{\sqrt{V(S)}}$$

where, S = Mann-Kendall statistic, m = 1 when S < 0 and m = -1 when S > 0 and

$$V(S) = \frac{1}{18}\left[N(N-1)(2N+5) - \sum_{i=1}^{n} e_i (e_i - 1)(2e_i + 5) \right]$$

in which e_i is number of data in ith group (tied group). The statistic u_c is assumed to be zero if S = 0. then the hypothesis of an upward or downward trend cannot be rejected at the α significance level if $|u_c| > u_{1-\alpha/2}$.

For small sample size,

$$u_c = \frac{\tau - \mu_\tau}{\sigma_\tau}$$

Where, $\mu_\tau = 0$, $\sigma_\tau = \sqrt{\frac{2(2N+5)}{9N(N-1)}}$ and $\tau = \frac{N_u - N_d}{N(N-1)/2}$ in which N_u = the number of upward pairs, N_d = the number of downward pairs and N = sample size.

6.3 Double mass analysis

Purpose

Double mass analysis is a technique commonly employed to determine corrections to hydrological data to account for changes in data collection procedures or other local conditions. The changes may result from a variety of things including changes in instrumentation, changes in observation procedures, or changes in gauge location or surrounding conditions. Double mass analysis detect the possible inhomogeneities in series, like jumps, trends, etc. by investigating the ratio of accumulated values of two series:

- ◊ the series to be tested, and
- ◊ the base series.

Computational procedure

If Yi, (i =1,N) be the test series and Xi, (i =1,N) the base series. The double mass analysis then considers the following ratio of two series:

$$rc_i = \frac{\sum_{j=1}^{i} Y_j}{\sum_{j=1}^{i} X_j} \tag{6.1}$$

$$pc_i = \frac{\sum_{j=1}^{i} Y_j}{\sum_{j=1}^{N} Y_j} \cdot \frac{\sum_{j=1}^{N} X_j}{\sum_{j=1}^{i} X_j} \tag{6.2}$$

Or expressed as a ratio of the percentages of the totals for N elements:

In the tabular results 9 columns can be presented as follows:

1) Time
2) Value of series X (base series)
3) Accumulated value of series X
4) Accumulated value as a percentage of the total of X
5) Value of series Y (test series)
6) Accumulated value of series Y
7) Accumulated value as a percentage of the total of Y
8) Ratio (item 6)/(item 3), equation (1)
9) Ratio (item 7)/(item 4), equation (2)

The last column is also presented in the double mass plot.

Now from any set of data we can calculate and plot the percent of total series (base over X-axis-column no.4 and test over Y-axis -column no.7).

Input

Average cumulative value of stations in X direction and cumulative value of tested station in Y direction.

Output

Particular point from where trend starts. A ratio is determined for multiplying the deviated record(s) to get homogeneous series.

Operational requirements and restriction

Stations of cumulative average values must be similar topographical features and climates with the station of tested series.

6.4 Mass curve

Purpose

To find the variation of a series with respect to time is shown graphically by a mass curve. Variation of tested series is compared with the average of a series of surrounding stations.

Description

A mass curve is a plot of cumulative value of hydrological data (rainfall, discharge etc.) against time. From the mass curve total depth of rainfall and intensity of rainfall at any instant of time can be found. The amount of rainfall for any increment of time is the difference between the ordinates at the beginning and end of the time increments, and the intensity of rainfall at any time is the slope of the mass curve (i.e. $\Delta P/\Delta t$) at that time. A mass curve of rainfall is always a rising curve and may have some horizontal sections, which indicates periods of no rainfall. The mass curve for the design storm is generally obtained by maximizing the mass curves of the severe storms in the basin.

Input

Data of base series and tested series is plotted with respect to time.

Output

Dissimilar of the point of tested and base series is found.

Operational requirement and restrictions

Data must be collected in same time and similar topographical features. Short-term interval (day/week) for rainfall data is not used for mass curve analysis.

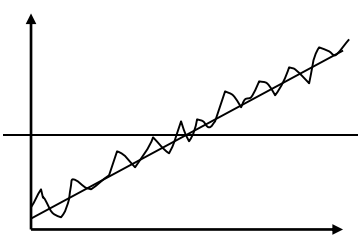
6.5 Residual mass curve

A residual mass curve represents accumulative departures from the mean. It is an efficient tool e.g. to detect climatic variabilities or other inhomogenates. The departure of the mass curve from the normal may be plotted against time. In other word the mass curve is plotted about a horizontal axis obtained by rotating the average slope line of the mass curve, to the horizontal. Such a plot is called a "residual mass curve".

The difference between the maximum and minimum values of a residual mass curve for a given period n is known as the 'range' for the period n. If R is the range of a period n years of annual runoff record whose sample standard deviation is σ.

6.6 Removing Trend

Hydrologic time series exhibit, in various degrees, trends, shifts or jumps, seasonality, autocorrelation, and nonnormality. The common ones are trends in the mean and in the variance. A linear trend in the mean is shown in Figure 6.1a. The trend $\overline{y_t}$ can be removed by the difference yt - $\overline{y_t}$ as shown in fig. 6.1b. The variance of such difference series, expressed by st2, may be either a function of time or may be a constant. The trend in variance can be removed by $\frac{y_t - \overline{y_t}}{s_t}$. The residual series may still have other properties such as correlation structure which can be decomposed and removed.



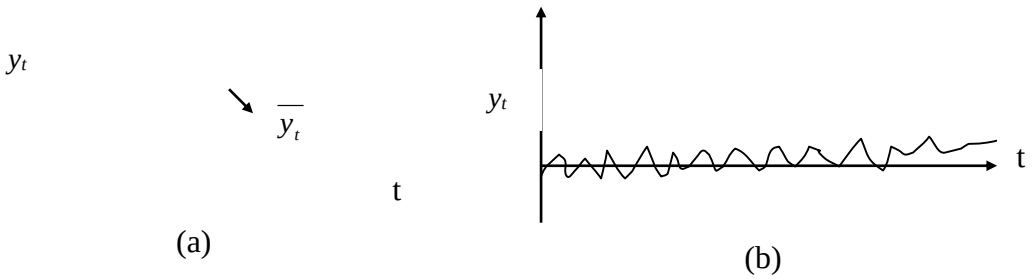


Figure 6.1 Removing trends in the series

6.7 Moving averages

Purpose and objective

To investigate the long term variability or trends in series moving average curves are useful. The procedure of smoothening the series consists in averaging preceding and succeeding values to a given values to a given value x_i or a total of *2m* or *2m + 1* successive members of series. The new smoothened value x_3 is at the i^{th} position of the series. The smoothened procedure may be repeated n times. Generally, a polynomial is fitted to *2m + 1* points.

Description:

Consider, for example, fitting a cubic equation to seven points. Let *t = -3, -2, -1, 0, 1, 2, 3* denote the time indices of the seven points. Let $a_0 + a_1t + a_2t^2 + a_3t^3$ represents the cubic equation to be fitted to seven points by the least square method. Then

$$S = \sum_{t=-3}^{+3}\left(x_t - a_0 - a_1t - a_2t^2 - a_3t^3\right)^2 \qquad \text{(a)}$$

To minimize S, we have

$$\frac{\partial S}{\partial a_j} = -2\sum_{t=-3}^{t=+3}\left(x_t - a_0 - a_1t - a_2t^2 - a_3t^3\right)t^j = 0 \qquad \text{(b)}$$

Where j = 0, 1,2,3. It may be noted that $\sum t^i = 0$ when *i* is an odd power, e.g., when i =1,

$$\sum t = -3-2-1+0+1+2+3 = 0.$$

$$\sum t^2 = (-3)^2 + (-2)^2 + (-1)^2 + 0 + (1)^2 + (2)^2 + (3)^2 = 28$$

From equation (b), we get

$$\frac{\partial S}{\partial a_0} = \sum\left(x_t - a_0 - a_2t^2\right) \quad \text{for } j = 0$$

Therefore,

$$\sum x_t = 7a_0 + a_2\sum t^2$$
$$= 7a_0 + 28a_2 \qquad \text{(c)}$$

Also,

$$\frac{\partial S}{\partial a_2} = \left(x_t - a_0 - a_2t^2\right)t^2 = 0 \text{ for } j = 2$$

$$\sum t^2 x_t = a_0 \sum t^2 + a_2 \sum t^4$$
$$= 28a_0 + 196a_2 \qquad \text{(d)}$$

Therefore, by solving equation (c) and (d), we get

$$a_0 = \frac{1}{21}\left(7\sum x_t - \sum t^2 x_t\right)$$

$$= \frac{1}{21}\left(7x_{-3} + 7x_{-2} + 7x_{-1} + 7x_0 + 7x_1 + 7x_2 + 7x_3 - 9x_{-3} - 4x_{-2}\right.$$
$$\left. - x_{-1} - 0 - x_1 - 4x_2 - 9x_3\right)$$
$$= \frac{1}{21}\left(-2x_{-3} + 3x_{-2} + 6x_{-1} + 7x_0 + 6x_1 + 3x_2 - 2x_3\right)$$
$$= \frac{1}{21}\left(-2, +3, +6, +7, +6, +3, -2\right)$$
$$= \frac{1}{21}\left(-2, +3, +6, +7\right)$$

Thus the smooth value at the middle position of seven points can be obtained by coefficients derived above by the least square method and given in equation above.

Input

Average of five or more (odd number) weights is moved successively.

Output

Weights or coefficients sum to unity and they are symmetric about the middle points. The middle value of the average is the treated value.

Operational requirement and restrictions

Series should be long. Short series does not show better filling. Initial and final values can not be filled in moving average method.

Advantage of moving average

- ◊ We get the same trend fit whether we fit forward or backward.
- ◊ It is easy to up-date.
- ◊ The moving average coefficients of a fit to (2m+1) points are the same for polynomial of degree 2p and 2p+1.

Disadvantage of moving average

- ◊ Data at the beginning and end of a series are lost.
- ◊ This may generate cycles or other movements that were not present in the original data.

6.8 Least square method

Purpose and Objective

The method of finding the line that best fits the collected data is done by Least Square Method. This is used to find the equation of an appropriate trend line, which can later be used to compute the trend values T.

Description

The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (*least square error*) from a given set of data. Suppose that the data points are (x1, y1), (x2, y2)…….(xn, yn) where x is the independent variable and y is the dependent variable. The fitting curve f(x) has the deviation (error) d from each point i.e. d1 = y1-f(x1), d2 = y2-f(x2)……. dn = yn-f(xn). According to the method of least square, the best fitting curve has the property that:

$$MSE = d_1^{\,2} + d_2^{\,2} + ... + d_n^{\,2} = \sum_1^n d_i^{\,2} = \sum_1^n \left[y_i - f(x_i)\right]^2 \qquad (6.3)$$

MSE = Maximum square error

Recommendation

Statistical inference of trend existence is the best way to detect trends in hydrologic series. If a jump is in the mean, variance first serial correlation coefficient, or any other parameter, then with their known sampling distributions and the tolerance level fixed, it is possible to determine whether a jump is significantly different from zero or not. Similarly a straight line trend x = a + bi fitted as a regression line by the least squares through the x-time series which is tested for b being significantly different from zero, or as a polynomial of the second or higher order having coefficients that are or are not significantly different from zeros. This approach in trend detection and description is more reliable than the moving average.

6.9 Normal distribution function

Purpose and Objective

Normal distribution is used to find out the distribution shape of data set. It measures the bell shaped for qualitative data and skewed for irregular data.

One reason the normal distribution is important is that many psychological and educational variables are distributed approximately normally. Measures of reading ability, introversion, job satisfaction, and memory are among the many psychological variables approximately normally distributed. Although the distributions are only approximately normal, they are usually quite close. A second reason the normal distribution is so important is that it is easy for mathematical statisticians to work with. This means that many kinds of statistical tests can be derived for normal distributions.

Description

The normal distribution can be fitted to data either analytically or graphically. The analytical test to check the goodness of set of a distribution function to an empirical distribution will be dealt with in the ensuing text. Goodness of fit, however, is generally checked graphically. In order to facilitate this fit, a

special scale for probabilities is transformed in such a way that the normal distribution function becomes a straight line. The probability density function (PDF) of a random continuous variable and is given as:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-(x-\mu)^2/2\sigma^2)$$

The two parameters of the distribution are the mean, μ and the standard deviation, σ for which X and s, derived from sample data, are substituted in above equation. By a simple transformation the distribution can be written as a single parameter function only. Thus when $t = (x-\mu)/\sigma$, $dx = \sigma\, dt$, the PDF becomes

$$f(t)=\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$$

And the continuous distribution function (CDF),

$$F(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t}\exp(-u^2/2)du$$

The variable t is called the standard unit, it is normally distributed with zero mean and unit standard deviation. This transformation is called the standardization of normal distribution variable.

Input

Number of data in a certain range is a plotted in y-direction and the range is plotted in x-direction.

Output

Average line of plotted data shows whether data set is normally distributed or bell shaped in condition. It helps to make further statistical decision.

Operational requirement and condition

Large number of data set gives more accurate curve.

6.9.1 Decision of normality test

The result of this test is expressed as ‘accept normality or ‘reject normality’ with P value. If P value is higher than 0.05, it may be assumed that the data have a Normal distribution and the conclusion accepts Normality is displayed.

In normality test, if the P value is less than 0.05, then the hypothesis that the distribution of the observations in the sample is normal should be rejected. In the later case the sample cannot accurately be described by arithmetic mean and standard deviation, and such samples should not be submitted to any parametrical statistical test or procedure, such as t-test. To test the possible difference between not normally distributed samples, the Wilcoxon test should be used, and correlation can be estimated by means of rank correlation.

When the sample size is small, it may not be possible to perform the selected test and an appropriate message will appear. In this case visually evaluate the symmetry and peakdness of the distribution using the histogram or cumulative frequency distribution.

6.10 Pearson's Correlation test (Parametric)

Purpose and objective

Pearson's "r" is a measure of the strength of the linear relationship between two variables: the extent to which increases in one variable are associated with increases (or decreases) in another variable. Pearson's "r" is a *parametric* test.

Procedures for using Pearson's correlation test (parametric)

It is used when both of the variables satisfies certain requirements and assumptions. These assumptions are that both variables are

(a) continuous;
(b) measured on an interval or ratio scale;
(c) normally distributed.

Pearson's "r" is measured on a scale which goes from +1 (perfect positive correlation, where increases in one variable are perfectly matched by increases in another variable) to -1 (perfect negative correlation, where increases in one variable are perfectly matched by decreases in another variable).

Formula for Pearson's "r"

$$r = \frac{\sum XY - \frac{\left(\sum X\right)*\left(\sum Y\right)}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)} * \sqrt{\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}}$$

To get a clear idea about the correlation one has to determine the square of the correlation coefficient "r". r^2 is called the "coefficient of determination" and it shows how much the variance in one of the variables can be accounted for by a knowledge of the other.

Note that calculating the coefficient of determination emphasizes the fact that the scale on which Pearson's r is ranged (i.e., -1 to +1) is not a linear scale: a correlation of 0.8 is not twice as good as one of 0.4, it is much better. A 0.8 correlation between two variables means that knowledge of one variable accounts for 64% of the variance in the other (0.8 squared = 0.64 = 64%). In contrast, a 0.4 correlation accounts for only 16% of the variance, and a 0.2 correlation accounts for only 4% of the variance. In short, the closer a correlation is to -1 to +1, then the stronger it is, and this process accelerates as one nears -1 or +1.

Input

Two sets of data to find the difference.

Operational requirement and restriction

The method is applicable for the data which are normally distributed and continuous.

6.11 Spearman's rank correlation test (non-parametric)

Purpose and Objective

Spearman's rank correlation test is a technique used to test the direction and strength of the relationship between two variables. Square value of difference is determined and helps to determine R_s for taking null hypothesis decision.

Procedure for using Spearman's rank correlation

This is used in much the same circumstances as Pearson's r. The chief difference is that it makes fewer assumptions about the nature of the data on which the correlation is to be performed: the data need only be measurements on an ordinal scale of measurement (i.e., as long as they can be meaningfully placed in rank order, this correlation can be used).

Another important difference is that whereas Pearson's "r" is a measure of the strength of the linear relationship between two variables, Spearman's rho is a measure of the strength of the monotonic relationship between them. If a monotonic relationship exists, it simply means that one of the variables increases (or decreases) by some amount when the value of the other variable changes. A linear relationship is thus a special case of a monotonic relationship. Thus, if there is a monotonic but non-linear relationship between two variables, it's better to use Spearman's rho because Pearson's "r" will tend to underestimate the strength of the relationship. Spearman's rho won't be "fooled" by the fact that the relationship isn't a linear one.

Method to calculate

Rank both sets of data from the highest to the lowest. Make sure to check for tied ranks.

Subtract the two sets of ranks to get the difference *d*.

Square the values of *d*.

Add the squared values of *d* to get Sigma d_2.

Use the formula $R_s = 1 - (6 \Sigma d^2/(n^3-n))$ where *n* is the number of rank and d is difference of similar position.

As with Pearson's "r", Spearman's rho is on a scale which goes from -1 (perfect negative correlation) to +1 (perfect positive correlation).

If the R_s value…

…is –1, there is a perfect negative correlation.

…falls between –1 and 0.5, there is a strong negative correlation.

…falls between –0.5 and 0, there is a weak negative correlation.

…is 0, there is no correlation

…falls between 0 and 0.5, there is a weak positive correlation.

…falls between 0.5 and 1, there is a strong positive correlation.

…is 1, there is a perfect positive correlation between the 2 sets of data.

If the Rs value is 0, state that null hypothesis is accepted. Otherwise, it is rejected.

Input

Two sets of data are ranked in similar order to find the difference of ranks.

Output

Correlation of two values. Square value of difference is determined and helps to determine Rs for taking null hypothesis decision.

Operational requirement and restriction

It is applicable if the data sets are normally distributed.

6.12 Median run test

Purpose

The runs test can be used to decide if a data set is from a random process.

Description

In the median run test the number of runs Nr of series A_i, (i=1, N) above and below the median is counted. Run is defined as an excursion above or below the median, denoted by A_m:

- ◊ A positive run is bounded by an up-crossing and down-crossing,
- ◊ A negative run is bounded by a down crossing and an up-crossing.

If the number values above and below the median are denoted by m and n respectively, the quantity N_r for a random series is asymptotically normally distributed with N (μ_r,σ_r):

$$\mu_r = 2mn/(m+n) + 1$$
$$\sigma_r^2 = 2mn\{2mn-(m+n)/\{(m+n)^2 (m+n-1)\}$$

where, μ is mean and σ is variance.

Operational requirement

The normal approximation holds for m and n >20. For smaller values of m and n may be used to obtain critical values of Nr at a 5% significance level.

6.13 Difference sign test

The difference sign N_n between successive values of series A_i, (i =1,N): $A_{(i+1)}$ - $A_{(i)}$. Let the maximum of the two be given by N_{ds}:

$$N_{ds} = \max (N_p, N_n)$$

For an independent stationary series of length Neff (Neff +N – zero differences) the number of negative or positive differences is asymptotically normally distributed with N (μds,σds):

$$\mu_{ds} = \frac{(N_{eff} -1)}{2}$$

$$\sigma_{ds}^2 = \frac{(N_{eff} +1)}{12}$$

The absolute value of the following standardized test statistic is computed:

$$|u| = \frac{|Nd_s - \mu d_s|}{\sigma_{ds}}$$

6.14 Series homogeneity test

Depending on the type of analysis series must fulfill one or more of the following requirements:

- *Stationary*: i.e. the properties or characteristics of the series do not vary with time;
- *Homogeneity*: i.e. all elements of a series belong to the same population;
- *Randomness*: i.e. series elements are independent.

Following statistical tests investigate stationary, homogeneity or randomness of a series:

- *Median run test*: a test for randomness by calculating the number of runs above and below the median:
- *Turning point test*: a test for randomness by calculating the number of turning points:
- *Difference sign test*: a test for randomness by calculating the number of positive and negative differences;
- *Spearman rank correlation test*: the Spearman rank correlation coefficient is computed to test:
    - ◊ the existence of correlation between two series,
    - ◊ the significance of serial rank correlation, and
    - ◊ the significance of a trend;
- *Arithmetic serial correlation coefficient test* : a test for serial correlation:
- *Wilcoxon-Mann-Whitney U-test*: a test to investigate whether two series are from the same population;
- *Student t-test*: a test on difference in the mean between two series;
- *Wilcoxon W-test*: a test on difference in the mean between two series;
- *F*-test for variance of two series;
- *Linear trend test*: a test on significance of linear trend by statistical inference on slope of trend lie;
- *Range test*: a test for series homogeneity by the rescaled adjusted range.
- *Chi-Square test:* It is used for 'Goodness of Fit'

Notes:

- The Spearman rank correlation test may be used as a single or two series test; in the single series mode it tests the significance of correlation with time.
- Wilcoxon-Mann-Whitney U-test, Student t-test and Wilcoxon W-test are basically two-series tests; however, the test can also be used for a single series by means of the split-sample approach, where a series is divided into two parts, which are mutually compared.

6.15 Spatial homogeneity test

Like rainfall, temperature, evaporation, etc., but sampled at a number of stations (point measurements) to investigate the reliability of point observation at a station called the test station, the observations are compared with weighted averages of the rainfall at neighbor stations. The weights are inversely proportional to some power of the distance between the test station and neighboring stations. Now the test is considering the difference between the observed and estimated value at the test station. If the absolute difference between observation and estimation exceeds specific limits (absolute and relative), the observation will be flagged (but not deleted!) to stress the need for further investigation. The test generally performed for rainfall, like estimation of point rainfall.

Chapter 7
Data Compilation

7.1 General

Aggregation and disaggregation of series are executed in data compilation. Aggregation of a series implies the evaporation of a series with a large time interval, by adding or averaging data of the series with the smaller interval: for example aggregation of series with an interval equal to one day to a series with an interval of one month, or from month to year, etc. Disaggregation is the opposite process: by disaggregation series with a smaller time interval is created. In aggregation series two cases are distinguished:

- ◊ For instantaneous observations the aggregated data are averages of the originals, e.g. discharges in m3/s.
- ◊ For accumulative observations the aggregated data are the sums of the originals e.g. rainfall or runoff in mm.

In disaggregation series following options exist:

For instantaneous observations

- ◊ Basic and disaggregated series data are equal,
- ◊ Disaggregated series data are interpolated linearly between the mid-point values of the basic series data (the original series, the series to be aggregated, is called the basic series.)

For accumulative observations

Disaggregated series data are fractions of the basic series data; if there are n disaggregated series elements in a basic series interval, then the fraction is $\frac{1}{n}$: disaggregated series data = $\frac{1}{n} \times$ basic series data.

7.2 Minimum, mean and maximum series

For a given series the computation of minimum, maximum and mean values for specific time periods are needed. Minimum, maximum and mean values for following time periods can be obtained:

- ◊ day
- ◊ month
- ◊ year
- ◊ period within the year.

7.3 Fitting distributions

For fitting distributions normally following theoretical frequency distributions are used:

- ◊ Normal Distribution
- ◊ Two-Parameter Log-Normal Distribution,
- ◊ Three-Parameter Log-Normal Distribution, Exponential distribution,
- ◊ Extreme Value Type I or Gumbel Distribution

- ◊ Extreme Type II or Frechet Distribution,
- ◊ Extreme Type III Distribution,
- ◊ Generalized Extreme Value Distribution,
- ◊ Pearson Type III or Gamma distribution
- ◊ Log-Pearson Type III distribution
- ◊ Box-Cox transformation to normality,

For each distribution one can obtain:

- ◊ Estimation of parameters,
- ◊ Summary of observed and theoretical probabilities,
- ◊ Goodness of fit-test:
  - – Binomial,
  - – Kolmogorov-Smirnov,
  - – Chi-squares
- ◊ Computation of extreme values for specific return periods, either related to probability of non-exceedance or exceedance, and
- ◊ Plot of distribution function.

7.3.1 Log-normal distribution function

Many hydrologic variables show a marked right skewness, partly due to the influence of natural phenomena having values greater than zero or some other lower limit, and being unconstrained, theoretically in ht upper range. In such cases frequencies will not follow the normal distribution, and instead, variables fortunately are often functionally normal and their logarithms follow a normal distribution.

$$p(x) = \frac{1}{x\,\sigma_X\sqrt{2\pi}} \exp\left[-\frac{(y-\mu_y)^2}{2\,S_y^{\,2}}\right] \qquad \text{for } X > 0. \tag{7.1}$$

where, μ_y = location parameter
S_y = Shape parameter

7.3.2 Standard incomplete gamma function

This function forms the basis of a number of distribution functions. It has very wide applications in hydrology studies. But Pearson Type III, a special case of Gamma distribution is all the more useful. This distribution has been widely adopted as the standard method for flood frequency analysis in a form known as the log-Pearson type III in which the transform Y = log X is used to reduce skewness. Although all the three moments are required to fit the distribution, yet it is extremely flexible in the sense that a zero skew will reduce the Log-Pearson type III distribution to lognormal. One of the greatest advantages in practical application of gamma and normal varieties lies in the fact that the sum of two such variables retains the same distribution. This function forms the basis of a number of distribution functions. It has the following form:

$$f(x) = x^{k-1}\exp(-x) / \Gamma(k) \qquad \text{for } x > 0. \tag{7.2}$$

A random variable x with this density is said to have the gamma distribution with shape parameter k. The following exercise shows that the family of densities has a rich variety of shapes, and shows why k is called the shape parameter.

7.3.3 Pearson's type III or gamma distribution function

This is a skew distribution with limited range in the left direction, usually bell shaped. The Pearson curve, is truncated on one side of the axis of the variate, i.e. below a certain value of the variate the probability is zero, but it is infinite converging asymptotically to the axis of the variate. This means even values infinitely large (or small) have a certain probability of occurrence.

$$p(x) = \frac{\lambda^{\beta}(x-\varepsilon)^{\beta-1} e^{-\lambda(x-\varepsilon)}}{\Gamma(\beta)} \tag{7.3}$$

where, $\lambda = \frac{s_x}{\sqrt{\beta}}, \beta = \left(\frac{2}{C_s}\right)^2, \varepsilon = \bar{x} - s_x\sqrt{\beta}$

The parameters can be estimated by:

- ◊ method of moments, and
- ◊ modified maximum likelihood method

7.3.4 Raleigh distribution function

$$P_R(X) = F_G(Z) \tag{7.4}$$

where:

$Z = (X - X_0)/\beta)^2$ and $\gamma = 1$
X_0 = location parameter, $(X_0 < X)$
β = scale parameter

The parameters X_0 and β are estimated from the moments.

7.3.5 Exponential distribution function

$$P_E(X) = F_G(Z) \tag{7.5}$$

where:

$Z = (X - X_0)/\beta$ and $\gamma = 1$
X_0 = location parameter, $(X_0 < X)$
β = scale paramerer

The parameter X_0 and β are estimated from moments

7.3.6 General Pearson distribution function

$$P_{GP}(X) = 0.5 + (K/|K|)(F_G(Z) - 0.5) \tag{7.6}$$

where:

Z = $((X-X_0)/\beta)^k$,
β = scale parameter,
k = type paramete (integer)
1: Pearson Type III (Exponential distribution for γ =1)
2: Raleigh (γ = 1) and Maxwell (γ = 1.5) distributions.
-1: Pearson Type V distribution.

The type parameter k is provided by the user. The parameters X_0, (X_0<X) or may be provided by the user (2-parameter distribution) or can be estimated (3-parameter distribution). The parameters X_0, β and γ are estimated by a mixed moment maximum likelihood method.

7.3.7 Log Pearson type III distribution function

$$p(x) = \frac{\lambda^{\beta}(y-\varepsilon)^{\beta-1} e^{-\lambda(y-\varepsilon)}}{x\Gamma(\beta)} \tag{7.7}$$

where, $\lambda = \frac{s_y}{\sqrt{\beta}}, \beta = \left(\frac{2}{C_s(y)}\right)^2, \varepsilon = \bar{y} - s_y\sqrt{\beta}$

Parameters are estimated by mixed moment-maximum likelihood method, and described in terms of sample moment.

ε = location parameters,
β = shape parameters,
γ = scale parameters.

7.3.8 Extreme type I or Gumbel distribution

$$p(x) = \frac{1}{\alpha}\exp\left[-\frac{x-\mu}{\alpha} - \exp\left(-\frac{x-\mu}{\alpha}\right)\right] \text{ for } -\infty \leq x \leq \infty \tag{7.8}$$

where: μ = location parameter,
α = scale parameter.

The parameters μ and α can be estimated by :

1) method of moments,
2) modified maximum lidelihood method, (with or without censoring).

7.3.9 Extreme type II or Frechet distribution

$$P_{EV-II}(X) = \exp\left(\frac{X - X_0}{\beta}\right)^{\frac{1}{k}} \tag{7.9}$$

where:

X_0 = location parameter, (X_0<X)
β = scale parameter,
k = shape parameter, (k<0)

The location parameter may be given (2-parameter distribution) or estimated (3-parameter distribution). The parameters can be estimated by the modified maximum likelihood method.

7.3.10 Extreme type III distribution

The Extreme Value Type III (EV3) distribution arises when the extreme is from a parent distribution that is limited in the direction of interest. This distribution has found its greatest use in hydrology as the distribution of low stream flows. Naturally low flows are bounded by zero on the left. The EV3 distribution is also known as Weibull distribution, (Haan, 1979). Distribution function can be described as follows:

$$p(x) = \alpha x^{(\alpha-1)} \beta^{(-\alpha)} \exp\left[-(x/\beta)^{\alpha}\right] \tag{7.10}$$

where: β = scale parameter,
α = shape parameter,

The parameters of the Weibull distribution can be estimated by the method of moments

7.3.11 Goodrich/Weibull distribution

$$P_{GO}(X) = 1 - \exp\left(-\frac{(X - X_0)}{\beta}\right)^{\frac{1}{k}} \tag{7.11}$$

where:

X_0 = location parameter, ($X_0 < X$)
β = scale paramerer
k = shape parameter ($k > 0$)

For the estimation of parameters the same applies as for the Extreme Type II distribution.

7.3.12 Pareto distribution

$$P_{PA}(X) = 1\text{-}e^{-z} \quad ; 0 < Z < \infty \; ; \theta = 0 \; (GP - I) \tag{7.12}$$

$$= 1 - (1 - \theta Z)^{\frac{1}{\theta}} \quad ; 0 < Z < \infty \quad ; \theta < 0 (GP - II)$$

$$= 1 - (1 - \theta Z)^{\frac{1}{\theta}} \quad ; 0 < Z < \frac{1}{\theta} \quad ; \theta > 0 (GP - III)$$

where:

$Z = (X\text{-}X_0)/\sigma$

X_0 = threshold (to be specified by the user),

σ = scale parameter

θ = shape parameter:

0 : general Pareto type –I distribution,

< 0 : general Pareto type-II distribution, and

> 0 :general Pareto type-III distribution.

The domain of X is,

For $\theta \leq 0$: $X > X_0$

For $\theta > 0$: $<X_0 < X < X_0 + \sigma/\theta$

The parameters are estimated either by the maximum likelihood method ($\theta \leq 0$) or by the method of moments ($\theta > 0$).

7.3.13 Peaks over threshold (POT) method

The POT-method uses the Pareto distribution, which is a skewed, heavy-tailed distribution that is sometimes used to model the distribution. An additional parameter λ is introduced, which is the average number of exceedance per year, A return period of T years corresponds in the POT method to one occurrence of the extreme in a series of λT exceedances above a fixed threshold X_0. Hence the related probability of non-exceedance in a series of all exceedances above a threshold is $1-1/(\lambda T)$.

7.4 Selection of Probability Distribution for Frequency Analysis

Among all the above-described probability distributions all the methods are not applicable for all types of data. Distributions are dependent on statistical properties of the data set. Depending on statistical properties of the data set, different data sets follow different types of distribution. Selection of appropriate probability distribution can be done by probability plotting on a probability paper and goodness of fir test. In Flood Hydrology Study (1992) under the Flood Action Plan (FAP 25) has included the results of an investigation on suitability of six probability distribution functions for fitting annual maximum water level and annual maximum discharge data for 5 key stations (Bahadurabad, Serajganj, Baruria, Mawa, Bhairab Bazar). They have followed following methods:

- ◊ Visual inspection of plotting on probability paper
- ◊ Goodness-of-fit tests using Kolmogorov-Smirnov (KS) and Chi-Square (CQ) tests
- ◊ Comparison of the extrapolation properties of the distribution.

It is reported that all distributions fit fairly well on probability papers and visual inspection was not conclusive. The criteria for judging an extrapolated value have not been mentioned there.

Chowdhury and Karim (1993) have followed different approaches for selecting appropriate probability distribution. These approaches are described in Article 7.4.1 of this chapter. They have compared five widely used method of frequency distribution for water level and discharge data for 48 and 31 gauging stations. The methods are Log-normal, Gumbel, generalized extreme value (GEV), Pearson Type-3 and Log-Pearson Type-3 (LP3). From their study report GEV distribution has been found to be the most appropriate followed by LN2, Gumbel and LN3 distributions for frequency analysis of annual maximum discharge data in Bangladesh. The GEV distribution has been found to be most appropriate followed by LN2 and P3 for annual maximum water level data.

In FAP 4, (Southwest Area Water Resources Management Project) final report (1993), frequency analysis has been carried out for the annual peaks at non-tidal gauging stations. The Gumbel and Extreme Value Type-3 distributions were fitted to the data and distribution which gave the best for events at higher return periods was chosen for that station. Other distributions have not been used in that study.

7.4.1 Methodology for selection of probability distribution

Normalization of the sample

Normalization of a sample of data means dividing every data of sample by the sample mean ($\bar{x}$). This can be expressed mathematically as: $h_i = x_i/\bar{x}$, where h_i is the normalized data.

Generation of random number

A random number is defined as a number selected at random from a population of numbers in such a technique that every number in the population has an equal chance of being selected. Random sample can be generated from a probability distribution by making use of the fact that the CDF for the continuous variate is uniformly distributed over the interval 0 to 1. Thus for any random variable X with PDF p(x), the variate $P(x) = \int_{-\infty}^{x} p(x)(t)dt$ is uniformly distributed over (0,1), (Chowdhury and Karim, 1993). Therefore, random numbers from a distribution can be drawn by generating random numbers from the uniform distribution and using the inverse of the CDF of that distribution. One example of generating data for Gumbel Distribution have been using the following equation:

$$x_p = \varepsilon - \alpha \ln[-\ln[p]]$$

where ε is a location parameter and Euler's constant, which is equal to 0.57721, α is a scale parameter.

Normal random numbers can be generated from the equation

$$Zp = \sqrt{\frac{q^2((4q+100)q+205)}{((2q+56)q+192)q+131}}$$

Where q = ln [2p]; p is the probability for different return periods.

The log-normal random numbers can be obtained by exponentiating the generated normal random numbers. Equations for obtaining the random numbers for GEV and P3 are

$$x_p = \varepsilon + (\alpha/k)[1-\{-\ln(1-1/T)\}^{k}]$$

$$x_p = \mu_x + K_p\sigma_x$$

For P3 distribution Z_p obtained from normal random number is inserted in place of K_p. The LP3 random numbers can be obtained by exponentiating the generated P3 random numbers.

Moment ratio diagram

Two approaches are available, one is to construct product moment ratio diagram and the other is to construct L-moment ratio diagram. For annual maximum data, the moment ratio diagram should be constructed by using unbiased estimators of moment ratio. Two moment ratio diagram can be constructed. One is a plot of L-CS versus L-CV and the other is L-kurtosis versus L-CS. Generally the first one is used for the two parameter distributions. Equation 5.16 to 5.21 is used to plot this diagram. L-moment ratio diagram would be plotted for the observed data and by using different probability distributions. Selection of distribution is made by examining which candidate distribution produces a L-kurtosis versus L-CS plot that is close to that obtained from observed data samples in the region. It is better to apply this method to the regionally homogeneous data sample.

Regional distribution of L-CS

The skewness is an important parameter since the shape of the tail end of a distribution is dependent on this. L-CS of the observed data and L-CS obtained for different distributions will be plotted against EV1 standardized variate. Plotting position can be determined by using the following formula:

$$F_i = i/(n+1)$$

By comparing the plot of observed data sample with the plot obtained from different distributions selection will be made.

Regional distribution of the largest data

For this approach sample size and period should be same for the selected station. A new data set can be obtained by picking the largest value from each year for each station. Average of theses values for each station will be plotted against EV1 standardized variate. For every parent distribution, similarly a new data set can be obtained by picking the largest value from every generated sample. Similar plot will be made for different distributions. Selection of distribution will be made by comparing the plot of observed sample data with generated data.

Probability plot correlation

Probability plots are extremely useful for visual revealing the character of a data set. Plots are effective ways to determine if fitted distributions appear consistent with the data. The PPC coefficient can be used as a criterion for comparing the goodness-of-fit of alternative distributions. It gives a measure of the correlation between the ordered observations h_i and the corresponding fitted quantiles $w_i = G^{-1}(F_i)$, Determined by plotting positions $F_i = i/(n+1)$

For each h. Values of correlation coefficient ρ near 1.0 suggest that the observations could have been drawn from the fitted distribution. Essentially, ρ measures linearity of the probability plot providing a quantitative assessment of fit. ρ can be obtained from following equation:

$$\rho = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(w_i - \overline{w})}{\left\{\sum_{i=1}^{n}(h_i - \overline{h})^2 \sum_{i=1}^{n}(w_i - \overline{w})^2\right\}^{0.5}}$$

where: $\overline{h} = \sum_{i=1}^{n} h_i/n$ and $\overline{w} = \sum_{i=1}^{n} w_i/n$

The quantiles w_i can be obtained for distributions is obtained from Eqs. 7.2, 7.4, 7.6, 7.8, 7.9, 7.11, 7.13 with the help of plotting position formula.

Root mean square error in fit

When a distribution is fitted to a sample of observed data at a station, the deviation of the fitted distribution from observed data can be evaluated by root mean square error (RMSE).

The average RMSE can be used as a criterion for comparing the goodness-of-fit of alternative distributions. In the case of regional analysis, it is advantageous to use the relative average RMSE. (Chowdhury, 1993)

RMSE can be calculated by:

$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_i - \hat{h}_i}{h_i}\right)^2\right]^{0.5}$$

where h_i be the observed value and $\hat{h}_i$ be the computed value from the fitted distribution. The values of $\hat{h}_i$ can be obtained from Eqs. 7.2, 7.4, 7.6, 7.8, 7.9, 7.11, 7.13.

Results

A summary of ranking of the distributions can be made from the results of above described five approaches

7.5 Test for stability of variance and mean

Test for stability, of variance and mean (F–test and t–test) (for stationary, consistency and homogeneity). To test the stability of variance and mean, which is an indication of the homogeneity, stationary and consistency of the time series, the F-test for stability of variance and t-test for stability of mean is used [Dhamen and Hall].

7.5.1 T-test

Purpose

Testing for equality of the variances and means of two samples.

Description

Now we want to know whether could seeding significantly affects rainfall, i.e. whether one set of numbers is larger or smaller than the other. The t-test is used to test for the mean of two sets of numbers being equal. You compute the t-statistic, which is given by this formula:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{SE_{12}} \tag{7.13}$$

The denominator is computed with:

$$SE_{12} = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where:

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where: $\bar{x}_1$ and $\bar{x}_2$ are the means of the two sets of numbers, S_1 and S_2 are the square roots of the variances of the two sets, and n_1 and n_2 are the number of points in each data set.

Operational requirements

Normal or almost normal distribution is necessary.

T-test Condition

It would seem logical that, because the t-test assumes Normality, one should test for Normality first. The problem is that the test for Normality is dependent the sample size. With a small sample a non-significant result does not mean that the data come from a Normal distribution. On the other hand, with a large sample, a significant result does not mean that we could not use the t-test, because the t-test is robust to moderate departures from Normality- that is, the P value obtained can be validly interpreted. There is something illogical about using one significance test conditional on the results of another significance test. In general it is a matter of knowing and looking at the data. One can ‘eyeball’ the data and if the distributions are not extremely skewed, and particularly if (for the two-sample t-test) the numbers of observations are similar in the two groups, then the t-test will be valid. The main problem

is often that outliers will inflate the standard deviations and render the test less sensitive. Also, it is not generally appreciated that if the data originate from a randomised-controlled trial, then the process of randomisation will ensure the validity of the t test, irrespective of the original distribution of the data.

The t-test is used to determine whether sample (s) have different means. Essentially, the t-test is the ratio between the sample mean difference and the standard error of that difference. The t-test makes some important assumptions:

- ◊ interval/ratio level data.
- ◊ one or two levels of one or two variables
- ◊ normal distribution
- ◊ equal variances (relatively)

Types of t-test

There are three types of t-tests:

The one sample t-test-

- ◊ This statistic tests a sample mean against a known population mean
- ◊ The null hypothesis tests if the sample mean is equal to the population mean.

The independent samples t-test-

- ◊ This statistic tests whether the mean of one sample is different from the mean of another sample.
- ◊ The null hypothesis tests if the mean of sample 1 is equal to the mean of sample2.

The paired group t-test

- ◊ This statistic tests if two groups within the overall sample are different on the same dependent variable.
- ◊ The null hypothesis tests if the mean of (var1- var2) is equal to 0.

7.5.2 Wilcoxon-Mann-Whitney U-test

Purpose

The two-tailed (Wilcoxon) Mann-Whitney U test can be used to test whether the medians of two independent distributions are different. So the test is used for comparing the medians of two samples.

Description

The Wilcoxon-Mann-Whitney test uses the ranks of data to test the hypothesis that two samples of sizes m and n might come from the same population. The procedure is as follows:

- ◊ Combine the data from both samples
- ◊ Rank each value
- ◊ Take the ranks for the first sample and sum them
- ◊ Compare this sum of ranks to all the possible rank sums that could result from random rearrangements of the data into two samples.

If step 4 reveals that the rank sum for the observed first sample is larger (or smaller) than nearly all the random orderings, this indicates that the first sample is significantly different from the second sample.

The sum of the raw scores for the control group are computed from

$$S = \{N\,(N+1)/2\} - W_{group}$$

Where: S = sum of the control group
N = total no. of rank
W_{group} = sum of working group

$$r_1 = \frac{\dfrac{\sum_{i=1}^{N-1}(A_i - m_A)(A_{i+1} - m_A)}{N-1}}{\dfrac{\sum_{i=1}^{N}(A_i - m_A)^2}{N}} \qquad (7.14)$$

with: m_A = mean of A_i
r_1 = Correlation co – effient
N = No. of data

The statistic used to measure the significan ce of r_1 is :

$$|t| = |r_1|\sqrt{\{(N-3)/(1-r_1^{\,2})\}} \qquad (7.15)$$

The formula for this z score computation is…

$$z = \frac{U + 0.5 - n_1(N+1)/2}{\sqrt{\dfrac{n_1 n_2 (N+1)}{12}}}$$

Where the value of U is the sum of the ranks in the important group, n_1 and n_2 are the no. of ranks of two groups and $N = (n_1 + n_2)$ = is the sum of total no of ranks.

Input

Two columns of measured or counted data.

Output

Mean value of the sample data sets.

Operational requirements

This test is non-parametric, which means that the distributions can be of any shape. This test is valid for only if both samples have N>7.

7.5.3 F-test

Purpose

The F test employs the statistic (F) to test various statistical hypotheses about the mean (or means) of the distributions from which a sample or a set of samples have been drawn. It is noteworthy that as demonstrated in this tutorial, the t test is a special form of the F test.

Description

The F or Fisher distribution is the distribution of the variance-ratio of samples from a normal distribution. Also, the F-test or Fisher-distribution gives an acceptable indication of stability of variance of the sub-sets of the time series even if the samples are not from a normal distribution. The test statistic is the ratio of the variance of two split, non-overlapping, sub-sets of the time series. F-test is the best choice as it always gives the exact P value. Whether the chi-square test is simpler to calculate but yields only an approximate P value. Chi-square test should be definitely avoided when the numbers are larger.

$$f(x)=\frac{\Gamma\left((m+n)/2\right)}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\frac{x^{(m-2)/2}}{\left(1+(m/n)x\right)^{(m+n)/2}}, \qquad x>0$$

where,

m, n	= degree of freedom
m/(n +1)	= mean
$n\sqrt{((m+2)/(m(n-2)(n-4))}$	= standard deviation

if n > 4, $\quad var(x)=2\left(\frac{n}{n-2}\right)^{2}\frac{m+n-2}{m(n-4)}$

Input

Two columns of measured or counted data.

Output

Variance of medians

Operational requirements

Random data set is required for f-test.

7.6 Goodness to fit test

7.6.1 Chi-square test

Purpose

Testing for equal distribution of compartmentalized, counted data.

Description

The chi-square test is used to test if a sample of data comes from a specific distribution. An attractive feature of the chi-square goodness of fit test is that it can be applied to any univariate distribution for which one can calculate the cumulative distribution function. The chi-square goodness of fit test is applied to binned data. This is actually not a restriction since for non-binned data one can simply calculate a histogram or frequency table before generating the chi-square test. However, the value of chi-square test statistic is dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires sufficient sample size in order for the chi-square test is that it requires sufficient sample size in order for the chi-square approximations to be valid.

The chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness of fit tests. The chi-square goodness of fit test can be applied to discrete distributions such as the binomial and the Poision. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

Often, the interest in practice is to know whether or not the observed frequencies differ significantly from the expected frequencies. A measure of the discrepancy existing between observed and expected frequencies is supplied by the statistic chi-square,

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - e_i)^2}{e_i} \right] \qquad (7.16)$$

where:

O_i = observed frequency
e_i = expected frequency = N*1/k, where, N = no. of observed data.
k = no. of class (interval) frequency.

If $\chi 2 = 0$, it indicates that the observed and theoretical frequencies agree exactly, while if $\chi 2>0$, they do not agree exactly. The sampling distribution of χ is approximated very closely to the chi-square distribution with (k-1) degrees of freedom provided the expected frequencies can be computed without having to estimate population parameters from the sample statistics. If we know (k-1) of the expected frequencies the remaining frequency can be determined. The degrees of freedom would be (k - h - 1) if the expected frequencies can be computed only by estimating h population parameters.

From the percentile of chi-square distribution table the critical value of $\chi 2$ is measured with 95% confidence level.

The expected frequencies are computed on the basis of a hypothesis H_0.

Input

Two columns of counted data in different compartments (rows).

Output

Numerical value for goodness of fit. The larger Chi-square value is the more probable the null hypothesis is false.

Operational requirements

- ◊ all the individual items in the samples should be independent; and
- ◊ the differences between small observed and expected frequencies at the ends of a distribution have a great effect upon Chi-square. Hence it is necessary to consider together two or more classes at each end. It has been suggested that no group should contain fewer than five expected frequencies.
- ◊ Chi-square test should be avoided if the number of data less than six.

7.6.2 Kolmogorv - Smirnov test

The 'two-tailed' Kolmogorov-Smirnov test determines whether two independent samples have been drawn from the same population. If the two samples have in fact been drawn from the same population, then the cumulative distributions of both samples may be expected to be fairly close to each other, i.e. they should show only random deviation from the population distributions. If the two sample cumulative distributions are too far apart at any point this suggests that they come from different populations. Thus a large enough deviation between the two sample cumulative distributions is evidence for rejecting the null hypothesis.

The Kolmogorov-Smirnov two-tail test focuses on the maximum deviation, D.

$$D = \text{maximum} \mid S_{Na}(X) - S_{Nb}(X) \mid \quad (7.17)$$

$S_{Na} = \frac{K}{N_a}$ and $S_{Nb} = \frac{K}{N_b}$ where K = No. of required observation. N_a and N_b = Total no. of observation.

For large samples (N>40) Kolmogorov-Smirnov tables show that the value of D must equal or exceed the value of:

$$1.36\sqrt{\frac{N_a + N_b}{N_a N_b}} \quad (7.18)$$

To reject the null hypothesis at the 5 per cent level, that is, that they are not from the same population.

The 'one-tailed' Kolmogorov-Smirnov test determines whether the two samples have been drawn from the same population or whether the values of one sample are stochastically larger than the values of the population from which the other sample was drawn. The maximum deviation is again calculated using equation (7.13) and the significance of the observed value of D can be computed by reference to the chi-squared distribution. It has been shown that for large samples:

$$\chi^2 = 4D^2 \frac{N_a N_b}{N_a + N_b} \quad (7.19)$$

It has a sampling distribution which is approximated to the chi-square distribution with two degrees of freedom. A chi-squared table for reference is given below.

If Kolmogorov-Smirnov value is less than D value then null hypothesis will be rejected. That is, in this case there is a significant difference between the two samples.

Which is less than the maximum deviation, and thus we can reject the null hypothesis at the 5% level. That is, in this case there is a significant difference between the two speed samples.

7.7 Test for relative consistency and homogeneity

Test for relative consistency and homogeneity with double–mass analysis will be carried out for all data. Double mass analysis can be carried out using the records for wet season months only. Inclusion of dry season months would make very little difference to the resulting plots. The cumulative rainfall at the station of interest could be compared to the cumulative average rainfall at a group of between the neighboring stations. It should be noted that missing data values have not been replaced by long-term mean as is often done in double mass analysis; instead, the months of missing data have been omitted from the period of analysis.

The variations in slope of the rainfall pattern may also consider which can be attributed to natural variation in rainfall patterns. The checking of these records might identify particular errors in the data and thereby improve the overall reliability. The double mass analysis can clearly identify some periods of questionable data, and examination of the data led to some adjustments. Where a period or periods of data showed consistently high or low rainfall the records have been adjusted by a factor chosen to bring the slope of the double mass plot into line with the long-term situation represented by the average of the check station. It has not been possible to identify the reason for such consistent errors in the data, but it is considered that the adjustments result in inaccurate data being replaced by a more realistic record. Possible reasons for changes in slope of the double mass plot include the re-location of a rain gauge, construction of nearby buildings, removal of trees, the use of an incorrect measuring cylinder, and incorrect conversion from inches to millimeters. After the double mass analysis further examination of the records would identify additional corrections.

7.8 Test for regional homogeneity

Stations can be grouped subjectively by site characteristics (i.e., latitude, longitude, elevation and mean annual precipitation). Although that purpose of the homogeneity test is to identify the stations with frequency distributions that are identical apart from a station-specific scale factor. For testing the homogeneity between different stations discordancy measure test is used. Stations can also be grouped objectively using cluster analysis procedures. Cluster analysis is a multivariate statistical analysis procedure for partitioning a data set into groups. The procedure involves assigning a data vector to each site. Sites are then divided into groups based on the similarity of their data vectors. The data vectors can consist of site statistics, site characteristics, or some combination of both.

For regional frequency analysis it is required that data series come from a homogeneous region. A homogeneous region is group of sites having data series which are assumed to have been drawn from the same frequency distribution. A nonparametric test based on a discordancy measure suggested by Hosking and Wallis (1993) was used to determine whether a given data series was discordant with a group of data series as a whole. The theme of the method is that if the L-moments (L-cv, L-skewness and L-kurtosis) of a site can be considered as a point in the three dimensional space, a group of sites will yield a cluster of such points. Any point which is far from the center of the cluster can be marked as discordant. Let $u_i = [t_1^{(i)}\ t_3^{(i)}\ t_4^{(i)}]^T$ be a vector containing at-site estimators of L-moment ratios for a data series of site *i.* Define the average over the N sites in an hypothesized region

$$u_{mean} = \frac{1}{N}\sum_{i=1}^{N} u_i$$

and the sample covariance matrix as

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(u_i - u_{mean})(u_i - u_{mean})^T$$

The discordancy measure for site I can be expressed as

$$Di = \frac{N}{3(N-1)}(u_i - u_{mean})^T S^{-1}(u_i - u_{mean})$$

Hosking and Wallis (1993) tabulated critical values of the discordancy statistic Di for various numbers of sites in a region at a significance level of 10%. Threshold Di ranges from 1.333 for 5 sites to 3.00 for 15 or more sites in a given region.

Chapter 8
Data Infilling

8.1 Necessity of data infilling

The occurrence of missing data varies considerably depending upon the different factors, e.g. instrumental problem, human error, recording error, systems involved in delivering the data. Missing or incomplete data are a serious problem in many fields of research. An added complication is that the more data that are missing in a database, the more likely that it is needed to address the problem of incomplete cases, yet those are precisely the situations where imputing or filling in values for the missing data points is most questionable due to the small proportion of valid data points relative to the size of the data set.

8.2 Methods used in filling data

8.2.1 Preliminary infilling

Different methods and techniques are used in filling missing values depending on the nature of data. Sometimes the following methods are only used for this purpose. No further approaches are followed. In this guideline we have considered the methods in filling as a preliminary filling and furthermore have considered for training the filling data series to get more nearest output to real data. The most widely used methods are-

If the large number of missing values-

- ◊ correlation weightage method,
- ◊ linear interpolation method,
- ◊ double mass analysis

If the small number of missing values-

- ◊ moving average method,
- ◊ normal ratio method,
- ◊ arithmetical average method and
- ◊ national weather services method

Double mass analysis and moving average methods have been discussed at the statistical and exercise sections.

Correlation weighted method

Correlation is used in a series of data to denote some form of association. It is used between two quantitative variables. It is assumed that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. Pearson method is used if the nature of data series is parametric and Spearman's method is in nonparametric series.

Prior filling by correlation weightage, nearest stations are selected and correlation among the nearest stations is determined. On the preference of correlation values some sort of series are selected and normalized their correlation values. After getting the correlation factors other series are arranged according to the similar date. The normalized correlation factor is used for determining the missing value of a similar period. The calculated value may be shifted from the natural trend of its own series.

The shifting is happened due to having the variation of tested data series from the other selected data series. After fitting the shifting data series is fed into the model for training.

Linear interpolation method

Linear interpolation method is one of the widely used methods in filling missing of hydrological data. Statistical tools, SPSS, gives the best technique to do so in modern data analysis. Based upon study it is recommended that it gives better output if the fewer number of missing values exist in the series. However, it could be used for preliminary filling, later on which will be fed into secondary infilling method, Artificial Neural Network intelligence.

Normal ratio method

If the normal annual precipitation at any of the index stations differ from that of the station in question by more than 10%, the arithmetical average method does not work. Under such circumstances, the normal ratio method is used in which the amounts at index stations are weighted by the ratios of the normal annual precipitation values. Thus, precipitation at station A, say P_A, is given by

$$P_A = \frac{1}{N}\left(\frac{N_A}{N_1}.P_1 + \frac{N_A}{N_2}.P_2 + \frac{N_A}{N_3}.P_3.....\right) \quad (8.1)$$

Where, N stands for normal annual precipitation.

N_A = Average annual rainfall

N = Average annual rainfall of surrounding station

P = Precipitation record.

Arithmetic average method

Many precipitation stations have breaks in their records due to some reason or the other. To fill this missing data, three stations as close to and as evenly spaced as possible around the particular station with the missing record are selected such that the normal annual precipitation at each of the index stations is within 10% of that for the station. Then a simple arithmetic average of the precipitation of the index stations provides the estimated amount. Thus, if P is the amount of rainfall and suffix A is for a particular station, and 1,2,3 are for index stations, then

$$P_A = \frac{P_1 + P_2 + P_3}{3} \quad (8.2)$$

National weather services method

This is a more suitable method in practice and has been verified on both theoretical and empirical bases. The rainfall at point *A* is computed by establishing a set of axes running through *A*. The following equation is used for filling missing.

$$P_A = \frac{\sum PW}{\sum W}$$

where: $W = 1/D^2$

D = distance between station *A* and the index station.

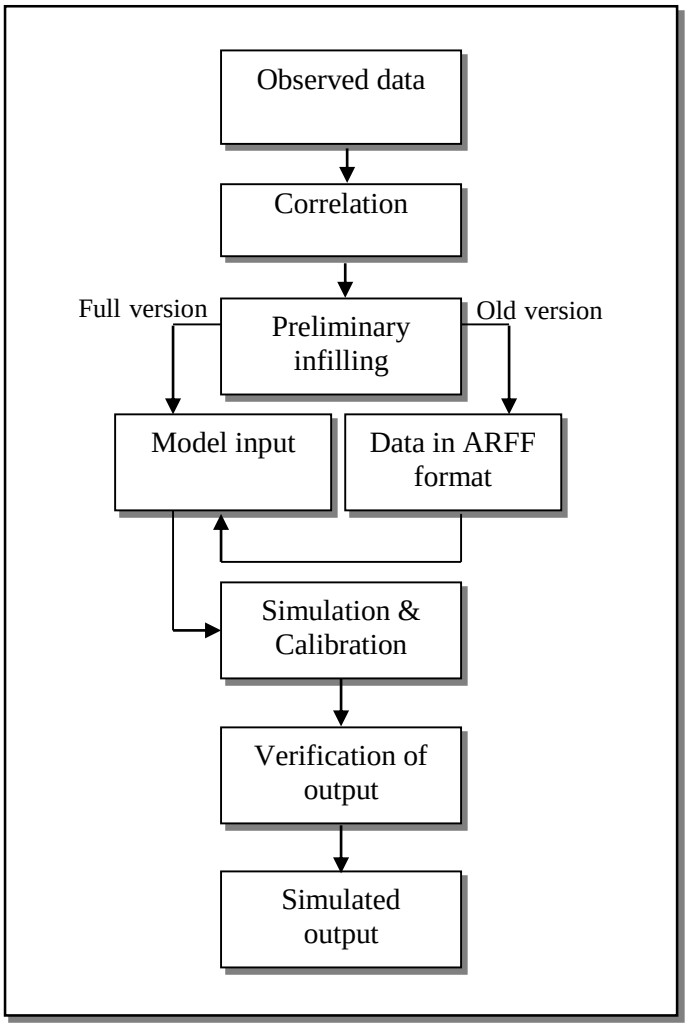
8.2.2 Secondary infilling

Artificial Neural Network (ANN) is one of the most up to date techniques in data infilling so far. The main advantage of it that generates the whole series regarding the input values as a training data set. It is widely used in developed countries. NWRD is studying on it for the application of the hydrological database. The preliminary fitted data has been set as an input data set in the model. Simulation shows better performance over preliminary treated data. If no preliminary filling, model assumes an average value for each missing values. Indeed, it checks out the variation of measured and targeted values and minimizes the error to get the nearest values of real data. However, it can train over without preliminary filling data for which more iteration is needed and Root Mean Square Error (RMSE) shows higher error or poor simulation. The output of ANN's simulation could be improved by clustering the data in seasons or similar time period.

ANN simulates better output with the large data series, as it is required. Some outputs of ANN's have been annexed.

ANN methodology

Following flow-chart shows the ANN's methodology. During training the expertness of the modeller plays an important role over good fitting of data.



8.3 Filling of missing rainfall data

8.3.1 General method

Normal ratio method is widely used to carry out the infilling. This method is applicable for a unit missing of data. Other methods are also used depending on the users needs. The procedure is summarized as follows:

Step 1

For each station a group of nearby stations are chosen from which the missing values will be estimated. Among large number of stations correlation method is preferred for the selection. The size of the group varies depending on the location of potential infilling stations. Stations of doubtful data quality are not included as filling stations.

Step 2

When a whole month data is missing, the monthly rainfall is estimated first by the normal ratio method from those stations the specified group which had data for the whole month and then this monthly rainfall is distributed in proportion to the rainfall pattern at one of the stations (the closest one for which data is available). This method provides the best estimate of the monthly rainfall and gives it a daily distribution which is realistic for the time of year and the location; the alternative procedure of using all the stations (effectively a weighted average of daily rainfalls) will tend to produce a smoothed series with many modest daily rainfalls but few extreme values. If data is missing at all the check stations then the record remained as missing data.

For an example group of three neighboring stations (a, b and c), all of which have data for the required month, the rainfall in month j at station i (the one being infilled) was found as follows:

$$Rm_i = (M_{i,j}/3)*\{Rm_a/M_{a,j} + Rm_b/M_{b,j} + Rm_c/M_{c,j}\}$$
$$Rd_i = Rd_a * Rm_i/Rm_a$$

where, $M_{x,j}$ = Mean rainfall at station x for month j
Rm_x = Monthly rainfall at station x
Rdx = Daily rainfall at station x

Step 3

Where data for part of month is missing, the missing days are infilled from the closest of the neighboring gauges, factoring the daily value at that gauge by the ratio of mean rainfalls for the relevant month at the two stations:

$$Rd_i = Rd_a * M_{i,j}/M_{a,j}$$

If data at the nearest gauge is also missing, the required value is filled from the next closest gauge, which has data available. In the event that data is missing at all the specified gauges the value at the station being infilled remained as missing.

This infilling provided estimates for almost all the periods of missing data.

8.3.2 Proposed method

Following methods are proposed for filling the missing rainfall:

Step1

Correlations of surrounding rainfall stations are determined.

Step2

Data series are arranged according to the correlations. Missing values are filled by ordinary method, e.g. correlation method, linear interpolation method etc.

Step3

Cluster the data series as Training and Verification sets. Verification set comprises one-fourth to one-third of the data series and the rest is Training set.

Step4

Data sets are fed into the model and different parameters are set for better simulations.

Step5

From the simulated output results are picked out for verifications.

8.4 Methods of handling in missing data

Some of the most popular methods for handling missing data described below. This list is not exhaustive, but it covers some of the more widely recognized approaches to handling databases with incomplete cases.

- *Listwise or casewise data deletion:* If a record has missing data for any one variable used in a particular analysis, omit that entire record from the analysis. This approach is implemented as the default method of handling incomplete data by many statistical procedures. Pairwise data can be deleted also.
- *Mean substitution:* Substitute a variable's mean value computed from available cases to fill in missing data values on the remaining cases.
- *Regression methods:* Develop a regression equation based on complete case data for a given variable, treating it as the outcome and using all other relevant variables as predictors. Then, for cases where Y is missing, plug the available data into the regression equation as predictors and substitute the equation's predicted Y value into the database for use in other analyses. An improvement to this method involves adding uncertainty to the imputation of Y so that the mean response value is not always imputed.
- *Expectation Maximization (EM) approach:* An iterative procedure that proceeds in two discrete steps. First, in the expectation (E) step the expected value of the complete data is computed. In the maximization (M) step the expected value is substituted for the missing data obtained from the E step and then maximize the likelihood fuction as if no data were missing to obtain new parameter estimates. The procedure iterates through these two steps until convergence is obtained.

Chapter 9
Processing of Hydrological Data

9.1 Concepts

Processing of data is essential as prime factor for various applications in connection with water resources planning, drainage design, reservoir design and operation, irrigation as well as hydrological forecasting and flood control. After processing a period of data the nature of the data set and its quality is determined.

The standard gauge is used to record daily readings. The size of the aperture and the height varies among various countries but it is usually standardized for each country. Therefore, the two points, types of collected data and goal are supposed to be decided before processing. Following concepts should go through data processing:

- ◊ enter data in the database and retrieve data
- ◊ plot and edit data
- ◊ carry out hydrological observations and measurements
- ◊ assist in the compiling and dissemination of hydrological data and information
- ◊ take part in field work

9.2 Steps for data processing

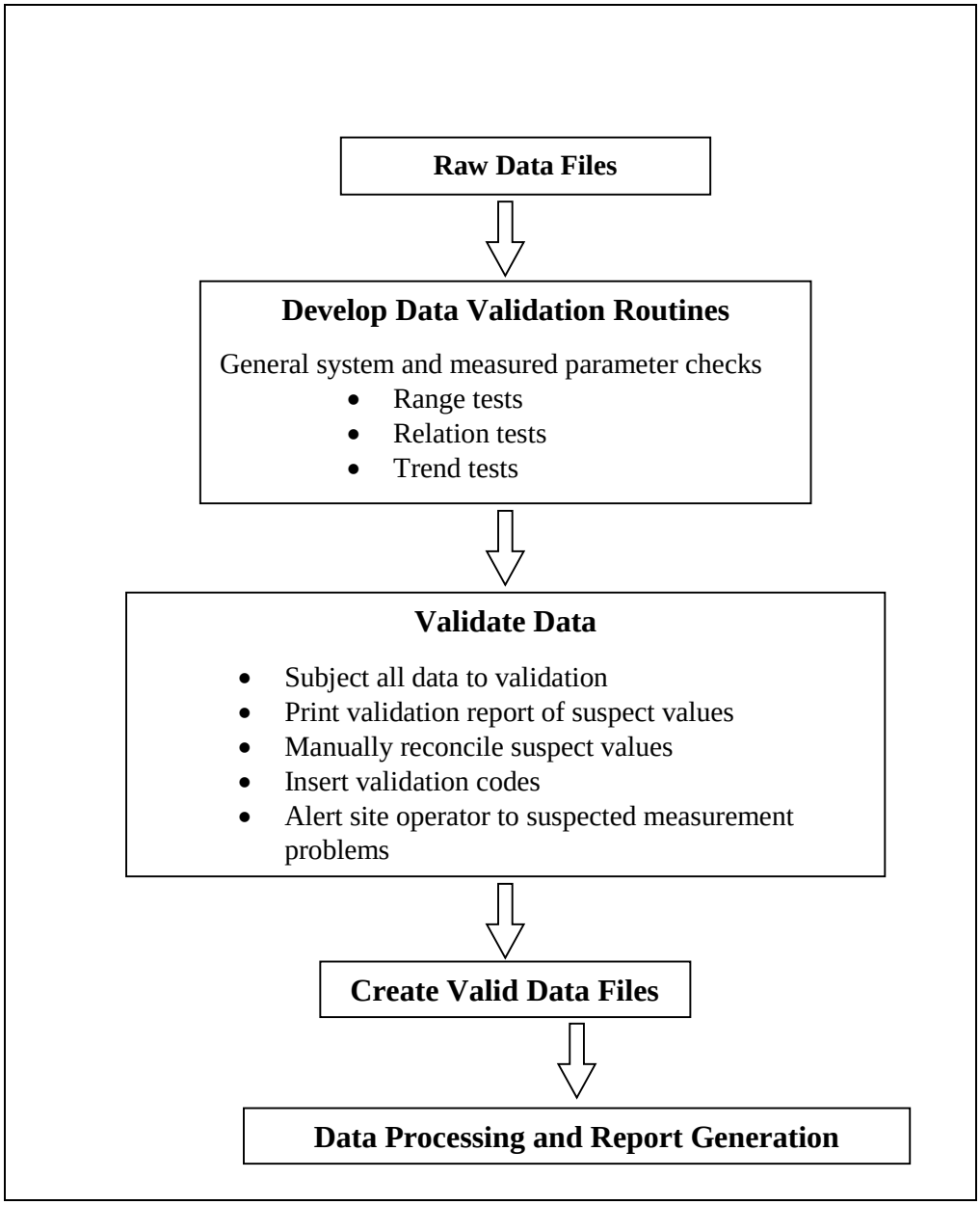
Handling of observational data until they are in a form ready to be used for a specific purpose. To make the proper use of data, data should be stored in such a way that all-possible errors can be avoided and that the data can be made easily accessible. Following steps are required to get a proper storage of data:

- ◊ validation;
- ◊ screening
- ◊ verification
- ◊ filling of missing value;
- ◊ compilation.

9.2.1 Validation

After the field data are collected and transferred to office computing environment, the next steps are to validate and process data, and generate reports. The flow chart presented below illustrates the sequence and roles of these steps:

Data Validation Flowchart



Data validation is defined as the inspection of all the collected data for completeness and reasonableness, and the elimination of erroneous values. This step transforms raw data into validated data. The validated data are then processed to produce the summary reports you require for analysis. This step is also crucial to maintaining high rates of data completeness during the course of the monitoring program. Therefore data must be validated as soon as possible, within one-two days, after they are transferred. The sooner the site operator is notified of a potential measurement problem, the lower the risk of data loss.

Data validation methods

Data can be validated either manually or automatically (computer-based). The latter is preferred to take advantage of the power and speed of computers, although some manual review will always be required. Validation software may be purchased from some data logger vendors, created in-house using popular

spreadsheet programs (e.g., Microsoft Excel, Lotus 123), or adapted from other utility environmental monitoring projects. An advantage of using spreadsheet programs is that they can also be used to process data and generate reports. These programs require an ASCII file format for imported data; the data logger's data management software will make this conversion if binary data transfer is used. There are essentially two parts to data validation; data screening and data verification.

9.2.2 Data screening

Erroneous data should be corrected before storing and should go through screening. As the first step of processing, screening of data is performed to obtain proper listing of series for easy reference and first checks on the range of data. Data screening procedure consists of five principal steps. These are:

- ◊ rough screening of the data and compute or verify the totals for the hydrological year or season.
- ◊ plot these totals according to the chosen time-step (e.g. month, year, season and note any trends or discontinuities).
- ◊ test the time series for absence of trend by suitable statistical method (e.g. Spearman's rank-correlation method, moving average method etc).
- ◊ apply the F-test for stability of variance and the t-test for satiability of mean to split, non-overlapping, sub-sets of the time series.
- ◊ apply Chi-square test for 'Goodness to Fit'

The above procedures form what we call the basic procedure. If necessary, one can expand the basic procedure to include two additional steps. These are:

- ◊ test the time series for absence of persistence by computing the first serial-correlation coefficient;
- ◊ test the time series for relative consistency and homogeneity with double-mass analysis.

9.2.3 Data verification

Validation routines are designed to screen each measured parameter for suspect values before they are incorporated into the archived database and used for site analysis. They can be grouped into two main categories, general system checks and measured parameter checks.

1. General system checks: Two simple tests evaluate the completeness of the collected data:

Data records: The number of data fields must equal the expected number of measured parameters for each record.

Time sequence: If there are any missing sequential data values this test should focus on the time and date stamp of each data record.

2. Measured parameter checks: These tests represent the heart of the data validation process and normally consist of range tests, relational tests, and trend tests.

Range test: These are the simplest and most commonly used validation tests. The measured data are compared to allowable upper and lower limiting values. The limits of each range test must be set so they include nearly (but not absolutely) all of the expected values for the site. Technicians can fine-tune these limits as they gain experience.

Relation test: This comparison is based upon expected physical relationships between various parameters. Relational checks should ensure that physically improbable situations are not reported in the data without verification.

Trend test: These checks are based on the rate of change in a value over time. Different statistics have been discussed in statistical section for removing trend of a time series data.

Data validation code

Data validation code is assigned for suspect values which provide particular information about data.

Table 9.2: Data validation code

Code	Rejection Criteria
-900	Rejection
-990	Unknown event
-991	Operating error
-992	Equipment malfunction
-993	Missing data (no value possible)
-994	Missing data (value possible)

9.2.4 Filling missing value

Filling missing value has been discussed in infilling chapter.

Treatment of suspect and missing data

After the raw data are subjected to all the validation checks, what should be done with suspect data? Some suspect values may be real, unusual occurrences while others may be truly bad. Here are some guidelines for handling suspect data:

1. Generate a validation report (printout or computer-based visual display) that lists all suspect data. For each data value, the report should give the reported value, the date and time of occurrence, and the validation criteria that it failed.
2. It should be examined to determine their acceptability. Invalid data should be assigned and replaced with a validation code. A common designation for data rejection is assigning a –900 series validation code, with numbers that represent various rejection explanations.
3. If redundant sensors are used, replace a rejected value from the primary sensor with a substitute one from the redundant sensor as long as the redundant sensor's data passed all the validation criteria.
4. Maintain a complete record of all data validation actions for each monitoring site in a Site Data Validation Log. This document should contain the following information for each rejected and substituted value:
    - ◊ File name
    - ◊ Parameter type and monitoring height
    - ◊ Date and time of flagged data
    - ◊ Validation code assigned and explanation.
    - ◊ The source of the substituted values.

9.2.5 Data compilation

Data compilation encompasses with some statistical techniques to analyze or summarize the characteristics of the data set, which is included with data aggregation, frequency analysis, fitting of

distribution functions, etc. After validation and filling of missing value data compilation ensure the data quality.

Frequency analysis is measured to find the time impact over the data. Different graphical plots ensure the quality. Procedure involved in interpreting a past record of hydrological events in terms of future probabilities of occurrence, e.g. estimates of frequencies of floods, droughts, storages, rainfall, water quality, waves. Fitting distributions are also used to examine the fitness of data set such as chi-square test.

9.3 Discharge data generation

Daily or continuous discharge data cannot practically be obtained directly. But it is possible to obtain daily or continuous stage data and from that a continuous discharge record can be estimated based on the relationship of water level and flow. The result is a correlation called the stage-discharge relationship. The Stage-Discharge relationship is known as rating curve. The rating curve is constructed by plotting successive measurements of discharge and gauge height on a graph. Uses of the rating curve are mainly:

- ◊ To convert water level record into flow rates.
- ◊ To estimate the design flood discharge by extrapolating rating curve.

The rating curve must be checked periodically to ensure that the relationship between the discharge and the gauge height has remained constant; scouring of the stream bed or deposition of sediment in the stream bed or deposition of sediment in the stream can cause the rating curve to change so that the same recorded gage height produces a different discharge.

Construction of rating curve

To develop this relationship, discharge measurements are obtained at the gauging station over the maximum range of gauge heights possible. A history of the relationship evolves over time, as each discharge measurement and corresponding stage is plotted, and a smooth curve is drawn that best represents these points. This curve is converted to a table of discharge values for incremental gauge heights, which in turn is referred to as a stage-discharge table. Daily or continuous discharge can be derived from this table using a daily or continuous stage record. In other case equation for the plotted curve will be derived and daily or continuous discharge can be obtained from that equation. The constant values of the equation need not be the same for full range of stage.

Methodology for developing a stage discharge curve:

- ◊ The stage discharge relationship curve can be plotted in a arithmetic or logarithmic paper. One example is shown in *Figure 9.1*.
- ◊ The relationship between the stage and discharge is a single- valued relation which is expressed as

  $$Q = C_r(G-a)^{\beta} \quad (1)$$

  Where, Q = stream discharge, G = Gauge height (stage), a = a constant which represents the gauge reading corresponding to zero discharge, C_r and β are rating curve constants. Values of C_r and β need not be the same for the full range of stages.

For arithmetic plot stage versus discharge can be plotted and for logarithmic plot (G-a) versus Q can be plotted. Logarithmic plot is advantageous as in logarithmic paper equation plots as a straight line.

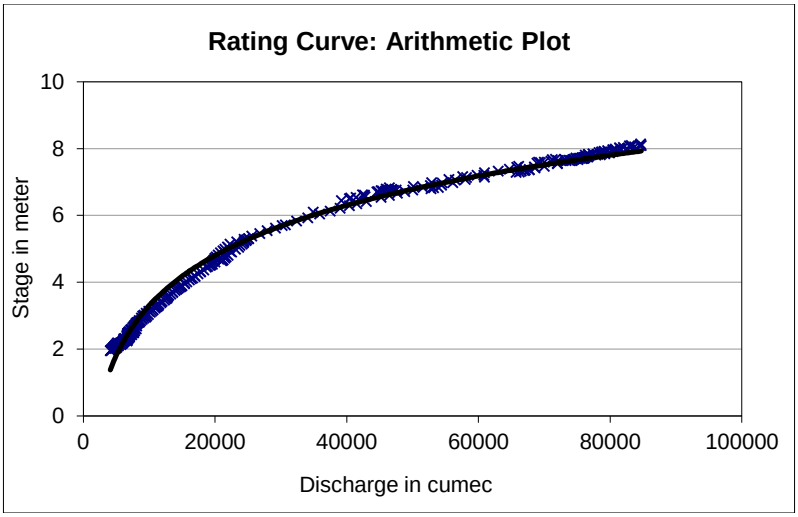


Figure 9.1 Rating Curve: Arithmetic Plot

Determination of Rating Curve constants:

Determination of C_r and β

The best values of C_r and β in equation for a given range of stage are obtained by the least-square-error method. Thus by taking logarithms:

$$\log Q = \beta \log (G-a) + \log C_r \qquad (2)$$

it can be compare with the equation $Y = \beta X + b$

where, $Y = \log Q$, $X = \log (G-a)$ and $b = \log C_r$

For the best-fit straight line of N observations of X and Y

$$\beta = \frac{N\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{N\left(\sum X^2\right) - \left(\sum X^2\right)}$$

$$b = \frac{\left(\sum Y\right) - \beta\left(\sum X\right)}{N}$$

Determination of a

There are four approaches to determine 'a'

1. Plot Q vs G on an arithmetic paper and draw a best fit curve. By extrapolating the curve by eye judgment find a as the value of G corresponding to Q = 0. Using this value of a, plot log Q vs log (G-a) and verify whether the data plots as a straight line. If not, select another value, which is nearest to the previous value and by trial and error get an acceptable value that gives a straight line in logarithmic plot.
2. A graphical method due to running as follows: The Q vs G data are plotted to an arithmetic scale and a smooth curve through the plotted points are drawn. Choose three points in the curve like A, B and C such as their discharges are in geometric progression, i.e.

   $$Q_A/Q_B = Q_B/Q_C$$

   At A and B vertical lines are drawn and then horizontal lines are drawn at B and C. The intersecting points of these vertical and horizontal lines are D and E. Two straight lines Ed and

BA are drawn and extending to join at F. The Y ordinate of F point is the value of a. This method assumes the lower part of the stage discharge curve to be a parabola.

3. Plot Q vs G to an arithmetic scale and draw a smooth good-fitting curve by eye-judgment. Select three discharges Q_1, $Q_2$and Q_3 such that Q1/Q2 = Q_2/ Q_3 and note from the curve the corresponding values of gauge readings G_1, G_2 and G_3. From Equation (1)

   $(G_1\text{-}a)/(G_2\text{-}a) = (G_2\text{-}a)/(G_3\text{-}a)$

   $$a = \frac{G_1G_3 - G_2^2}{(G_1 + G_3 - 2G_2)}$$

4. A number of optimization procedures that are based on the use of computers are available to estimate the value of a. A trial and error search that gives the best value of the correlation coefficient is one of them.

9.4 Procedure of statistical analysis

Following steps have been considered for statistical analysis of hydrological data:

Step1: Raw data

Arranging raw data.

Step 2: Validation test

This step is typically used for data validation purposes. Range test is done before trend of a series is estimated in this step. Data range helps to arrange data series in some ranges for a particular series. Before estimating trend, the first thing to find is whether or not any trend is present at all. And to check this, tests for randomness, turning point test are performed on the time series. If t value ranges within ±1.96 for 5% level of significance, it is decided that the series is of random sequence. If series is not random, the trend value should be estimated by Least Square Method or Moving Average Method. On the other hand it goes through the consistency test.

Step3: Consistency checking

For checking the consistency of a set of data Double Mass analysis has been done to detect possible inhomogeneities in a series, like jumps, trends, etc. by investigating the ratio of accumulated values of two series. It should not be done with any missing value.

Step4: Correlation test

Correlation test is performed for determining the correlation of neighboring stations with the tested station. The missing value of the tested station should be filled with comparing the most correlated station. If the correlation value does not satisfied it should be gone through the trend analysis.

Step 5: Filling missing value

After getting the correlation value the best correlated series with tested series is taken as base series. Normal ratio method is done for filling missing value.

Step 6: Normality test

The primary reason to test whether data follow a normal distribution is to determine if parametric tests procedures can be applied or not. Use of a larger α-level will increase the power to detect non-normality, especially for small sample size.

Step 7: Test for shift in the mean

After filling missing value normal distribution is done before significant test. If data set is normally distributed T-test is performed for further analysis otherwise Wilcoxon test is performed. It can be also defined that if the sample cannot accurately be described by arithmetic mean and standard deviation, and such samples should not be submitted to any parametrical statistical test or procedure, such as T-test. The Wilcoxon test should be used for alternative analysis.

Step 8: Trend test

Trend test is necessary to determine presence of trend in the data series. If there is any trend in the data series it should be removed before further analysis.

Step 9: Trend analysis

If data series contains trend then it can be removed by trend analysis (by using moving average method and/or slope)

Step 10: Selection of frequency distribution

Before starting with distribution function for a set of data it has to be ensured that the selected distribution is applicable for that data set that is the inherent assumption is the sample is drawn from the selected distribution. For selecting the true distribution five methods can be used which is described in the Article 7.4. Goodness of fit test is also used for this purpose.

Step 11: Frequency analysis

After selecting the type of distribution frequency analysis can be done. The objective of frequency analysis of hydrological data is to relate the magnitude of the extreme events to their frequency of occurrences through the use of probability distribution.

Step 12

After analysis of data quality report or comments is prepared.

Bibliography

BWDB, 1972. The Hydrometric Apparatus and Techniques used in the Hydrology Directorate (BWDB water supply paper-361)

Chow V. T., Maidment David R., Mays Larry W., 1988. Applied Hydrology

Chowdhury, J.U. & Karim, M.A. (1993), Selection of Probability Distribution for Flood Frequency Analysis in Bangladesh, Final Report.

Chowdhury, J.U. (1995), Flood Frequency Analysis**,** Supporting Study on Selection of Probability Distribution for Flood Frequency Analysis in Bangladesh, Volume 7.

Hosking, J.R.M. and Wallis, J. (1993) Some statistics useful in frequency analysis, Wat.Resour. Res. 24,588-600

Hossain, M.M, (1995), BUET, Lecture notes on Sediment Transportation.

EGIS, 1999, Application of AEZ database in drought management and water availability assessment.

FAP 2, 1992 North West Regional Study. Draft final report, summary.

FAP 2, 1991, North West Regional Study. Interim report- volume 3, annex 2: hydraulic studies, annex 3: hydrology, annex 4: groundwater.

FAP 2, 1991, North West Regional Study. Interim report- volume 2, annex 1: engineering.

FAP 4, 1993, Southwest Area Water Resources Management Project, Final Report, Volume 5.

FAP 6, 1993, North East Regional Water Management Project. Interim report.

FAP 6, 1993, North East Regional Water Management Project. Specialist study on river sedimentation and morphology: draft final.

FAP 23, 1992, Flood Proofing. Interim report.

FAP 24, 1994, River Survey Project. Report on mission project adviser, annexure I: sediment gauging strategies, annexture III: remote sensing and bathymetry in morphological assessment, annexture IV: comments on FAP 24 - a flow and sediment gauging.

FAP 24, 1993, River Survey Project. Interim report, volume I: main report.

FAP 24, 1996, River Survey Project. Overland flow and floodplain sedimentation measurements.

FAP 24, 1996, River Survey Project. Optimization of sediment measurements.

FAP 24, 1996, River Survey Project. Bathymetric survey

FAP 24, 1996, River Survey Project. Water level and gauging stations (Special report 2)

FAP 24, 1996, River Survey Project. Final report: main volume

FAP 24, 1996, River Survey Project. Optimization of hydraulic measurements. (Special report no. 11)

FAP 24, 1996, River Survey Project. Final report- annex 3: hydrology.

FAP 25, 1992, Flood Modelling and Management, Flood Hydrology Study, Annex 1**.**

Haan Charles T., 1979. Statistical Methods in Hydrology

Hymos Manual, 1992, Version 3.0, delft hydraulics

Jaisawal R. K., Goal N.K., Singh P., Thomas T., (2003), L-moment based flood frequency modeling.

Klinting A., 2000, National Water Sector Database, Database and Application Design.

Kumar R., Chatterjee C., Panigrahy N., Patwary B. C., Singh R.D. (2003), Development of regional flood formulae using L-moments for gauged and ungauged catchments of North Brahmaputra River system.

Mutreja K N, 1990. Applied Hydrology

NWRD, WARPO, 2001 Technical Notes on Surface Water Data Group

Ogink H.J.M., 1993, Lecture notes on hydrology 2

Raghunath H M, 1990. Hydrology (Principles, Analysis, Design)

Scarborough James B., 1996, Numerical Mathematical Analysis,

Shahin Mamdouh, Oorschot H.J.L Van, De Lange S.J., 1993. Statistical Analysis in Water

Resources Engineering by Hydrosystems Engineering

SSWRDSP, LGED, 1998, Standard Design Catalogue, volume 1: Hydrology.

Approximations for use in constructing L-moment ratio diagrams, *Research Report RC 16635*, IBM Research Division, Yorktown Heights, N.Y., 1991]. The approximations are also given in Appendix A.12 of the book *Regional frequency analysis: an approach based on L-moments*, by J. R. M. Hosking and J. R. Wallis.

SWMC Handbook, 1996 Surface Water Modelling Centre

WEB Sites:

http://www.crwr.utexas.edu/gis/gisenv98/class/risk/ex1/ex1soln.html#hist

http://www.dianthus.co.uk/statistics/fisher.htm

http://www.deming.eng.clemson.edu/pub/tutorials/qctools/flowm.htm

http://www.epa.gov/region1/oeme/toc.html

http://www.filebox.vt.edu/artsci/stats/waterman/7.1A.htm

http://www.fao.org/docrep/X5560E/x5560e04.htm

http://www.fmi.uni-sofia.bg/vesta/Virtual_Labs/special0/special5.htm

http://geography.uoregon.edu/mcdowell/geog427w02/Exercises/Ex%201/x1.html

http://math.uc.edu/~brycw/classes/147/blue/tools.htm

http://mathworld.wolfram.com/topics/StatisticalDistributions.html

http://www.munro-group.co.uk/met.htm, http://www.web.utk.edu/~toddc/xsqtan.html,

http://www.physics.csbsju.edu/stats/t-test.html,

http://www.roads.dft.gov.uk/roadsafety/goodpractice/39.htm#01

http://www-stat.stanford.edu/~nars/jsm/FindProbability.html,

http://www.stats.gla.ac.uk/steps/glossary/nonparametric.html#kst

http://traffic.ce.gatech.edu/nchrp2045/glossary.html#217

http://www2.truman.edu/shaffer/266ch13_2001.htm

http://www.up.ac.za/academic/geol/meteo/WKD162c03.doc

http://www.warpo.org/NWRD/HydStn.html,

http://www.wmo.ch/web/homs/homs/h76205.html,

http://www.xhuoffice.psyc.memphis.edu/statistics/xhu/all/lecture18/node18.html,

Appendix A
Examples of steps for data processing

Checklist for quality monitoring of time series data

Raw Data	Format: Data Processing Software:
Timeline (Mention Missing Data Information)	Period:
Metadata	Yes/No (If yes, informative or not)
Bundle information (Field description and units)	Yes/No (If yes, informative or not)
Technical Report	
Validation Test	Method and result
Consistency Checking	Method and result
Correlation test	Method and result
Filling missing value	Method and result
Test for Normality	Method and result
Test for shift in the mean	Method and result
Trend test	Method and result
Trend Analysis	Method and result
Selection of distribution	Method and result
Frequency Analysis	Method and result
Comments	Very good / Good/Moderate/Poor

Some Exercises has been practised following the above checklist is presented below:

Exercise A1: Rainfall data

The following table shows the monthly average rainfall data (in mm) from 1962-1995 for the base station Atghoria (Station ID 1) and the 13 neighbouring stations. The stations are selected on the basis of distance from the base station and there correlation has been investigated. The correlation among the stations with base station is given in the Table below. From the observed correlation 3 stations are selected which shows better correlation with the base station. Further analysis of the base stations is done considering the 3 surrounding stations for statistical tests.

A. Step 1- Raw Data

Table A1.1: Average rainfall depth (mm) at 13 stations - Base Stations

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1962	18.8	10.9	8.5	107.8	234.6	254.5	340.0	232.4	127.7	108.7	0.0	0.0
1963	0.0	0.0	6.6	71.3	144.7	360.5	276.5	188.1	155.7	115.9	24.0	0.0
1964	1.5	10.9	5.2	110.1	186.6	429.4	384.4	205.9	119.3	352.7	1.1	0.0
1965	0.8	5.9	23.4	43.0	75.0	281.5	285.7	296.0	230.1	40.9	16.9	0.5
1966	16.7	0.0	3.1	49.4	104.7	270.9	259.7	272.0	145.9	107.7	14.4	8.0
1967	15.4	3.6	83.6	87.9	87.9	242.6	256.9	347.5	316.9	89.1	0.1	1.6
1968	0.0	2.3	24.9	47.8	117.3	368.1	287.9	267.0	131.4	101.4	32.2	0.0
1969	0.5	1.3	65.4	77.2	122.4	228.4	194.4	574.1	329.9	93.5	32.6	0.0
1970	39.4	15.1	19.6	34.9	102.0	287.7	391.4	231.5	353.0	328.6	4.3	0.0
1971	7.5	5.2	3.2	162.1	122.5	267.7	313.1	452.7	239.0	102.1	31.0	0.0
1972	1.4	14.4	0.6	49.0	201.0	242.9	143.5	265.9	145.4	29.6	0.0	0.0
1973	10.5	29.3	24.7	90.1	377.8	435.3	180.0	238.2	351.1	122.5	30.9	48.6
1974	0.6	0.0	75.0	124.0	122.8	196.0	504.4	226.8	200.0	75.3	0.5	0.0
1975	1.0	9.6	10.4	61.5	187.0	130.0	619.6	211.1	200.2	141.8	25.0	0.0
1976	0.0	14.8	10.5	49.9	352.0	367.9	378.6	378.6	200.8	91.2	7.1	0.0
1977	7.2	42.6	10.2	176.0	191.5	605.1	357.1	181.7	137.4	147.5	11.9	43.8
1978	1.1	18.7	53.6	95.5	227.3	229.8	175.5	164.3	226.6	89.5	0.4	0.0
1979	10.8	24.2	10.0	50.8	15.9	224.4	310.4	427.8	189.3	34.2	25.1	19.5
1980	25.4	13.6	26.7	27.4	341.0	221.6	211.6	208.6	201.3	175.7	0.0	0.0
1981	14.1	68.0	47.5	276.9	224.1	146.6	391.8	175.3	161.8	4.7	0.0	49.8
1982	0.6	5.3	80.7	131.1	90.1	198.0	166.7	190.3	100.4	47.6	51.3	0.0
1983	6.4	30.4	58.7	135.8	217.5	147.4	256.1	367.3	212.6	272.9	0.0	15.5
1984	3.9	0.0	0.4	61.6	195.8	590.0	386.4	244.1	368.4	149.9	0.0	0.0
1985	0.2	1.3	36.4	33.7	211.0	248.5	272.5	181.0	209.7	157.1	1.0	0.1
1986	1.8	6.1	2.6	89.7	139.1	226.3	274.5	215.2	626.8	236.4	62.9	1.9
1987	0.6	2.5	23.4	90.9	107.0	199.9	478.2	572.6	218.3	42.3	15.7	22.2
1988	0.0	51.5	47.5	95.8	325.1	447.7	305.0	220.2	200.4	118.1	130.1	1.5
1989	0.1	23.6	0.1	14.4	317.9	203.5	431.0	83.2	195.8	142.0	0.0	12.0
1990	0.0	72.1	97.7	137.7	168.4	259.3	380.8	221.2	349.3	128.4	26.7	0.7
1991	11.0	23.3	27.7	52.1	323.9	403.0	376.4	202.8	533.0	179.0	0.6	102.8
1992	1.0	55.5	6.3	15.5	129.6	144.8	373.9	219.6	281.4	75.5	7.3	1.2
1993	1.2	27.7	62.5	125.1	224.4	372.8	290.1	271.3	326.0	121.0	6.5	0.0
1994	30.0	44.4	18.9	68.3	226.8	291.3	153.0	187.1	115.5	124.1	4.8	0.0
1995	7.6	27.1	4.8	8.7	133.4	377.9	316.5	337.2	326.0	53.0	75.8	0.9

Table A1.2: Rainfall depth (mm) at Atghoria station – Test station

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1962				81.3	197.8	383.0	293.8	140.8	62.2	92.5	0.0	0.0	104
1963	0.0	0.0	8.3	25.1	109.7	306.6	175.8	245.8	164.8	97.5	27.0	0.0	97
1964	0.3	21.1	5.8	105.0	215.8	474.5	367.5	257.1	116.9	387.4	0.0	0.0	163
1965	0.0	2.0	36.3	75.3	85.9	423.7	312.9	385.6	216.5	18.3	12.7	2.3	131
1966	12.1	0.0	1.5	44.3	139.9	395.1	204.1	286.2	143.7	134.0	18.0	12.5	116

1967	20.3	2.0	94.7	155.0	117.6	375.3	370.8	499.4	486.7	226.1	0.0	1.8	196
1968	0.0	0.0	23.6	20.6	127.8	268.5	334.3	317.8	82.0	103.4	23.1	0.0	108
1969	1.3	0.0	83.0	129.4	113.9	222.0	183.5	782.6	535.3	141.4	36.4	0.0	186
1970	39.9	18.8	31.5	25.1	130.5	338.1	515.9	143.6	420.6	356.3	5.6	0.0	169
1971	10.1	7.4	0.0	197.9	181.2	462.8	507.0	559.1	319.6	76.6	42.9	0.0	197
1972	3.0	21.8	0.0	78.7	187.2	295.8	205.3	302.2	160.0	39.3	0.0	0.0	108
1973	19.0	6.4	15.0	116.5	348.0	437.4	162.8	298.0	376.8	108.1	30.4	63.5	165
1974	0.0	0.0	112.7	89.5	95.6	217.0	578.7	239.4	265.2	119.3	5.1	0.0	144
1975	3.3	1.3	18.7	76.7	184.8	77.0	599.2	164.8	125.8	158.3	42.7	0.0	121
1976	0.0	31.0	3.3	15.8	281.8	398.6	496.4	345.1	383.4	32.3	1.8	0.0	166
1977	12.7	39.1	0.0	176.5	204.3	751.4	383.7	142.5	151.4	154.2	10.4	44.5	173
1978	0.8	22.3	48.7	85.1	250.6	199.1	335.5	129.7	200.4	52.3	0.0	0.0	110
1979	6.9	21.8	15.5	61.8	10.4	248.7	255.7	509.9	316.8	30.1	41.6	32.8	129
1980	30.0	7.8	26.0	2.5	272.5	302.7	279.2	296.1	101.1	163.8	0.0	0.0	123
1981	45.5	120.6	89.4	426.1	211.8	208.3	450.0	227.2	162.4	6.3	0.0	69.5	168
1982	0.0	5.3	134.9	193.0	71.0	244.8	204.9	233.2	74.5	38.6	40.3	0.0	103
1983	13.0	38.1	52.0	215.6	220.7	132.2	264.0	427.6	387.9	467.1	0.0	20.6	187
1984	7.2	0.0	0.0	37.7	180.7	739.4	304.0	231.8	282.9	182.2	0.0	0.0	164
1985	0.0	0.0	85.1	16.8	148.1	284.7	369.7	130.8	186.8	118.0	0.0	0.0	112
1986	6.4	16.5	0.0	111.0	75.4	317.0	319.7	146.9	621.7	277.1	13.5	6.9	159
1987	0.0	0.0	43.6	104.6	85.8	260.5	500.9	658.7	231.6	32.0	17.8	16.0	163
1988	0.0	47.2	26.4	62.2	259.8	394.4	258.2	218.5	137.0	43.8	109.5	0.0	130
1989	0.0	62.0	0.0	25.6	261.5	240.4	549.9	110.6	170.2	209.0	0.0	30.4	138
1990	0.0	97.5	113.7	122.7	128.1	343.3	367.2	289.0	366.5	124.7	34.5	0.0	166
1991	23.2	17.8	9.8	51.7	351.7	320.0	281.1	180.2	552.6	115.0	1.8	76.3	165
1992	0.0	45.4	0.0	0.0	119.6	191.1	327.3	217.3	222.6	73.5	1.5	6.4	100
1993	1.2	29.2	80.0	111.2	161.5	324.9	268.9	344.5	221.7	74.3	19.2	0.0	136
1994	22.8	65.2	2.9	47.4	258.8	311.5	208.2	167.7	215.8	98.1	0.0	0.0	117
1995	10.5	18.1	12.9	12.2	90.7	201.9	343.5	282.6	437.4	83.9	99.4	0.0	133

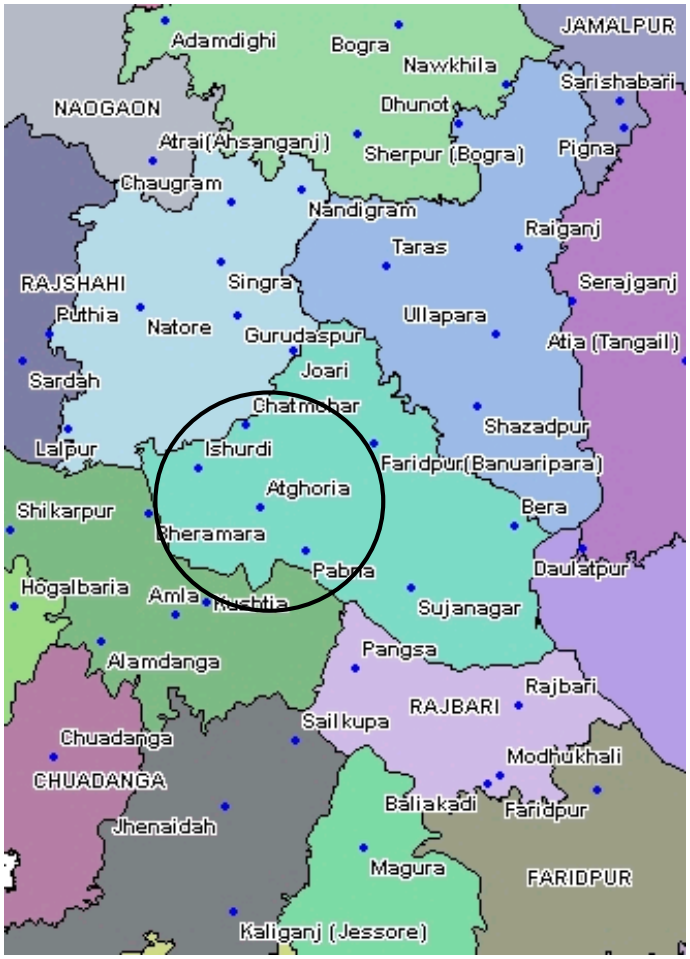


Figure A1.1: Neighbor stations of Atghoria

Table A1.3: Rainfall at Pabna (Station ID-25)

Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1962	17.8	16.8	26.2	128.2	237.6	305.2	271.8	221.4	140.2		0	0
1963	0	0	0	55.2	152	276.2	163.6	187.6	191.2	106.6	21.1	0
1964	1.5	14	0	82.3	123.8	371.4	268.1	207.6	173	489.2	0	0
1965	0	5.6	34.8	59.7	74.9	262.7	344.3	302.7	220	32.3	39.6	0
1966	0	0	6.8	24.4	87.7	201.7	226.9	224.3	126.6	97.5	20.6	20.2
1967	19.8	0	64.8	108.6	59.7	254.9	233.6	291.3	274.9	120.3	0	3.8
1968	0	3.3	26.2	86.6	131	356.5	219.2	213.3	97.6	54.6	25.9	0
1969	1	0.7	51.3	73.1	127.7	173	273.2	532.8	368	115.8	37.4	0
1970	49.5	8.1	33.6	64.3	60.6	318.2	387.9	184.2	541.7	385.5	0	0
1971	1.5	18.3	3.8	0	0	278.7	354.8	639	333.1	51.3	36.8	0
1972	0	22.9	0	37.4	312.4	199.4	183.3	262.1	135.8	24.9	0	0
1973	0	115	25.1	95.1	382.2	345.9	164.4	184.9	266	112.7	42.6	69.1
1974	0	0	105.3	108.4	98.5	241.6	448.7	271.9	184.2	118.8	0	0
1975	2	3.3	22.4	124.4	400	52.4	647.2	157	123.2	135.2	31	0
1976	0	15.5	11.4	12.5	286.7	335.2	268.1	365.7	244	90.1	0	0
1977	13.9	21	5.1	169.6	236.5	584.7	390.3	161.8	64.5	117.7	17	54.1
1978	2	18.8	37.9	120.1	222.2	212.1	177	136.5	253.3	51.5	2.5	0

Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1979	13.7	32	1.3	29.9	5.8	209.7	369.4	481.9	157.8	31.2	22.9	33.5
1980	45.7	8.4	26.6	25.4	316.9	227.1	201.7	188.2	169.8	220.5	0	0
1981	16.5	81.2	31.4	250.7	247.9	194.1	482.8	144.1	152.1	5.8	0	69.1
1982	0	8.1	98.6	118.7	81.5	172.2	148.7	218	65.3	74.1	33	0.5
1983	1.5	77.9	43.2	47.9	250	140.5	339.6	391.8	341.3	336.8	0	16
1984	2.5	0	0	81.2	155.9	701.5	249.5	264	339.9	145.3	0	0
1985	0	0	37.3	51.5	107.8	301.3	301.6	172.1	127.2	79.5	0	0
1986	0	0	0	38.7	118.8	247.1	234.8	210.2	655.1	196.8	48.8	2.3
1987	2	0.3	36.8	29.4	74.5	237	332.4	534.4	168.9	5.6	29.2	17.8
1988	0	43.2	36.8	36	392	384.1	320	180.5	181	80.2	126	0
1989	0	29.5	0	11.5	345	165.9	308.6	67.5	196	121.8	0	11.5
1990	0	100.5	109	101.5	150.8	248.7	367.8	264.5	223	90	14.5	0.5
1991	10.5	26	10	10	302	385.7	418	305	463.4	215	1	115.5
1992	2.5	45.5	0	0	158.5	132.8	419.9	196.3	244.4	167.7	5	0
1993	0.4	14.3	32	116.2	142.9	367.6	271.5	314.3	322.9	98	1.6	0
1994	37.4	56.3	43.1	45	349	267.5	123	193.5	132	112.5	0	0
1995	10	27.5	0.5	3.5	81	270.5	196.1	404	331	79.5	70.5	0
Sum	**251.7**	**814**	**961.3**	**2347**	**6273.8**	**9423.1**	**10107.8**	**9074.4**	**8008.4**	**4164.3**	**627**	**413.9**
Avg	**7.40**	**23.94**	**28.27**	**69.03**	**184.52**	**277.15**	**297.29**	**266.89**	**235.54**	**122.48**	**18.44**	**12.17**

Table A1.4: Rainfall at Ishwardi (Station ID-15)

Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1962	17.8	5.1	5.8	63.5	238.9	284.8	433.8	277.5		126	0	0
1963	0	0		82.3	177.9	616.6	216.3	275.1	252.1	172.8	33	0
1964	0	19.9	2.3	63.5	184	547.3	428.1	106	112.2	619.4	0	0
1965	0	6.3	36.9	48.5	85.6	210.5	425	236.5	215.4	90	11.7	0
1966	0.3	0	5.1	105.1	71.5	208.1	271.9	194.3	200.4	209.1	16.5	5.6
1967	14.5	0	94.5	83.3	70.3	172.6	289	533.6	213.2	96.5	0	0
1968	0	0	24.6	10.9	71.9	397.3	249.7	236	125.5	140.4	29.2	0
1969	0	0	65.7	70.6	62	189.5	127.6	590	396.2	107.2	45.9	0
1970	46.3	23.4	21.3	20	68	305.4	267.7	103.4	309.8	320.3	2.8	0
1971	16	4.1	0				71.1	452.5	472.8	104.2		
1972				17.8	67.6	147.9	130.4	103.8	279.4	12.7	0	0
1973	20.3	6.3	2.5	43.2	426.3	525.7	72.4	171.4	203.5	101.6	8.7	25.4
1974	1.3	0	0	161.3	21.6	116.9	146.5	45.6	37.6	49.1	0	0
1975	0	0	23.3	130.8	60.4	127	436.2	172.1	105.2	43.5	14.5	0
1976	0	34.6	0	40.7	108.5	331.6	141.3	252.9	101.1	124.7	10.2	0
1977	11.5	35.5	2.5	181.6	172.8	823.1	298.3	219.6	82.1	106.7	0	50.6
1978	0	36.1	71.6	88.8	223.5	167.6	125.8	218.9	184.9	69.4	0	0
1979	16	18.3	0	70.6	13.9	323.4	377	438.4	163.8	27.9	33	10.2
1980	30	0	21.6	0	276.1	245.2	135.8	208.1	289	164.5	0	0
1981	18.9	59.6	39.3	216.6	203.2	128.1	393.7	189	279.2	2.5	0	82.5
1982	0	6.3	78.8	129.7	70.4	263.4	87.4	168.8	143.4	21.6	35.6	0
1983	11.4	10.1	31.5	75	218.6	139.8	315.1	342.9	177.3	332.7	0	0

Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1984	2.5	0	0	64.8	62.8	571.4	334.4	243.4	390.9	130.8	0	0
1985	0	0	29.7	55.3	161.8	157.2	186.9	161.7	250.4	133.6	0	0
1986	0	5.3	5.1	65	119.1	163.6	338.4	218.5	480.3	173.5	42.5	4.8
1987	0	0	16.3	94.5	96.2	139.5	520.4	918	161.1	19	10.2	29.5
1988	0	28.2	98.8	28.8	354.6	341.2	241.1	238.6	218	79	97	0
1989	0	58	0	6	406.4	103.1	494.2	68.2	208.2	103.9	0	11.5
1990	0	62.7	132.7	116.1	252.4	435.6	510.4	143.2	376.6	113.2	38.5	0
1991	0	8	18	135.5	188.2	471.5	222.2	179.8	425.1	110.3	0	88.6
1992	0	28.1	0	21.5	180.3	105.3	305.6	319.3	345.4	51.5	3.5	1
1993	2	3.1	49.1	121.9	243.8	309.2	232.2	194	322	89.5	8	0
1994	24.5	41.5	0.2	35	209	326	151.1	154.5	142	68.5	0	0
1995	14	28	31	0	151	224	243	300	337	24	94	5
Sum	**247.3**	**528.5**	**908.2**	**2448.2**	**5318.6**	**9619.4**	**9220**	**8675.6**	**8001.1**	**4139.6**	**534.8**	**314.7**
Avg	**7.274**	**15.54**	**26.71**	**72.006**	**156.43**	**282.92**	**271.2**	**255.16**	**235.33**	**121.75**	**15.73**	**9.256**

Table A1.5: Rainfall at Chatmohar (Station ID-7)

Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1964				55.3	231.1	808.2	576.9	134.2	56.8	185.7	0	0
1965	8.1	5.8	0	0	42.2	219.4	50	254.3	51.6	20.6	24.1	0
1966	31.8	0	3.3	11.7	60.6	223.6	133.4	229.2	97	10.4	1.3	0.3
1967	0.5	0.5	39	62.2	58.8	153.4	163.7	224.4	226.3	28.4	0	0
1968	0	0	1.8	6.4	12.3	224.6	143.3	360.3	158.1	49.7	13.5	0
1969	0	0	76.9	65.5	102.4	199.7	268.1	739.1	162.7	54.4	6.6	0
1970	11.7	7.1	0	26.2	99	314.7	424	167.7	276.9	398.9	9.4	0
1971	0.8	0	0	163.6	42.8	45.8	159.2	128.6	473.8	103.9	9.1	0
1972	0	0	0	0	130.8	452.6	67.8	238.4	150.6	9.6	0	0
1973	13.2	14.8	20.3	122.2	415.4	845.2	293.9	282.1	591.6	85.6	21.4	6.9
1974	0	0	95	196.6	166.2	142.5	619.8	178.1	102.4	46.8	0	0
1975	0	0	0	47.1	94.4	163.1	816.2	178.3	106.9	45.2	30.5	0
1976	0	3	4.8	26.4	711.8	467.8	343.6	474.2	48	42.9	3	0
1977	3.6	58.9	7.9	96.4	108	983.5	141.5	135.6	57.4	87.8	0	15.2
1978	0	4.3	53.9	105.1	127.6	144.9	38.3	69.9	70	16	0	0
1979	8.9	15.5	0	44.7	8.1	206.5	233.7	292.2	75.8	25	8.4	3.8
1980	5.1	5.9	20.8	14.3	239.4	214.6	115.5	271	29.3	58.7	0	0
1981	12.6	74	52.8	144.3	101	123.4	390.7	161.4	86.2	0	0	18.8
1982	0	0	33.2	71.4	22.4	142.8	73.6	140.1	44.9	24.9	34.5	0
1983	4.8	20.8	39.6	125.7	218.2	112.8	118.4	0	0	0	0	0
1984	0	0	0									
1985				42.2	195	298	271.2	197.8	183.1	365.2	0	1.5
1986	2.3	5.8	8.9	102	82.5	238.5	250.7	204.7	760.3	152.9	62.7	0
1987	0	0	7.1	63.8	117.1	155.3	538.7	536.2	225	80	3.8	16.3
1988	0	79.3	55.2	87.4	216	369.1	269.4	197.5	290.2	132.1	133	0
1989	0	45	0	0.1	223.7	182.9	515.4	38.7	334.3	101.5	0	0
1990	0	63	67.5	112.8	145.4	211.5	289.7	173.4	157.8	81.3	11	0

Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
1991	3	0	33	15	386.6	209	421.4	139.7	458.1	154.7	0	88.4
1992	0	23	0	69.5	100.1	125	376.2	225.2	258.4	44.8	4	0
1993	2	5	74.5	140.5	306	401	222	301	284	261	0	0
1994	28	50	17	44	168	249	125	172	168	96	7	0
1995	0	35	0	11	114	295	366	290	272.8	18.5	72.6	3
Sum	**136.4**	**516.7**	**712.5**	**2073.4**	**5046.9**	**8923.4**	**8817.3**	**7135.3**	**6258.3**	**2782.5**	**455.9**	**154.2**
Avg	**4.547**	**17.22**	**23.75**	**66.884**	**162.8**	**287.85**	**284.43**	**230.17**	**201.88**	**89.758**	**14.71**	**4.974**

B. Step 2- Validation Test

Test for Randomness of the data: Turning point test

The Turning point test is used to check whether the time series verify the following hypothesis:

H_0: The series is a random no trend series

H_a: The series has trend and/or auto correlated errors.

With this test one can check yearly or monthly randomness of data set. The example describes yearly randomness checking of the data set:

Mean = 2(N – 2)/3 and

Standard deviation = $[(16N - 29)/ 90]^{0.50}$

Standard value of t,

$$t = \frac{(T- \text{Mean})}{\text{Standard deviation}}$$

Table A1.6: Turning-point test of station Atghoria

Annual average rainfall	Range	Turning points	Cumulative Turning points	Annual average rainfall	Range	Turning points	Cumulative Turning points
104				129	110<129>123	yes	14
97	104>97<163	yes	1	123	129>123<168	yes	15
163	97<163<131	no	1	168	123<168>103	yes	16
131	163<131>116	yes	2	103	168>103<187	yes	17
116	131>116<196	yes	3	187	103<187>164	yes	18
196	116<196>108	yes	4	164	187>164>112	no	18
108	196>108<186	yes	5	112	164>112<159	yes	19
186	108<186>169	yes	6	159	112<159<163	no	19
169	186>169<197	yes	7	163	159<163>130	yes	20
197	169<197>108	yes	8	130	163>130<138	yes	21
108	197>108<165	yes	9	138	130<138<166	no	21
165	108<165>144	yes	10	166	138<166>165	yes	22
144	165>144>121	no	10	165	166>165>100	no	22
121	144>121<166	yes	11	100	165>100<136	yes	23
166	121<166<173	no	11	136	100<136>117	yes	24
173	166<173>110	yes	12	117	136>117<133	no	24
110	173>110<129	yes	13	133			
							T = 24

Therefore, T = 24 and N=34

T follows normal distribution with

Mean = $\frac{2}{3}$ (N-2) = $\frac{2}{3}$ (34-2) = 21.33 mm

Standard deviation = $((16N - 29)/90)^{1/2} = ((16*34-29)/90)^{1/2} = 2.392$

Therefore,

$$t = \left|\frac{(24 - 21.333)}{2.392}\right| = 1.115$$

$t_{1-\alpha/2}$ = 1.96 (at 5% significance level) [Area under the normal curve]

Since $t < t_{1-\alpha/2}$, Null hypothesis accepted that the data set is random and No trend in series. No trend analysis is necessary at this stage.

C. Step 3- Consistency Checking

i) Double mass analysis

Table A1.7: Cumulative average of test and base station

Year	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
Base Station	120	232	383	491	596	723	838	982	1132	1274	1366	1527
Test Station	104	201	364	495	611	807	915	1101	1270	1467	1575	1740
Year	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	
Base Station	1654	1787	1942	2101	2208	2320	2441	2571	2659	2803	2969	
Test Station	1884	2005	2171	2344	2454	2583	2706	2874	2977	3164	3328	
Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	
Base Station	3082	3239	3387	3549	3667	3821	4007	4117	4269	4374	4513	
Test Station	3440	3599	3762	3892	4030	4196	4361	4461	4597	4714	4847	

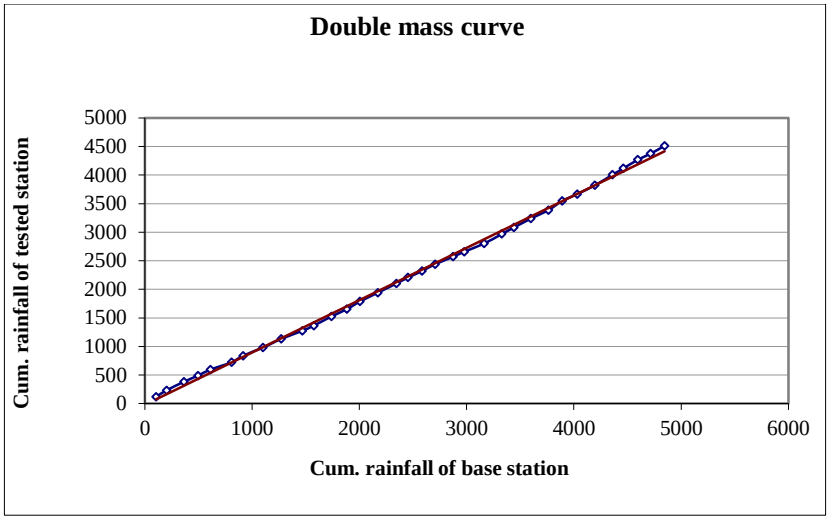


Figure A1.2: Double mass plot

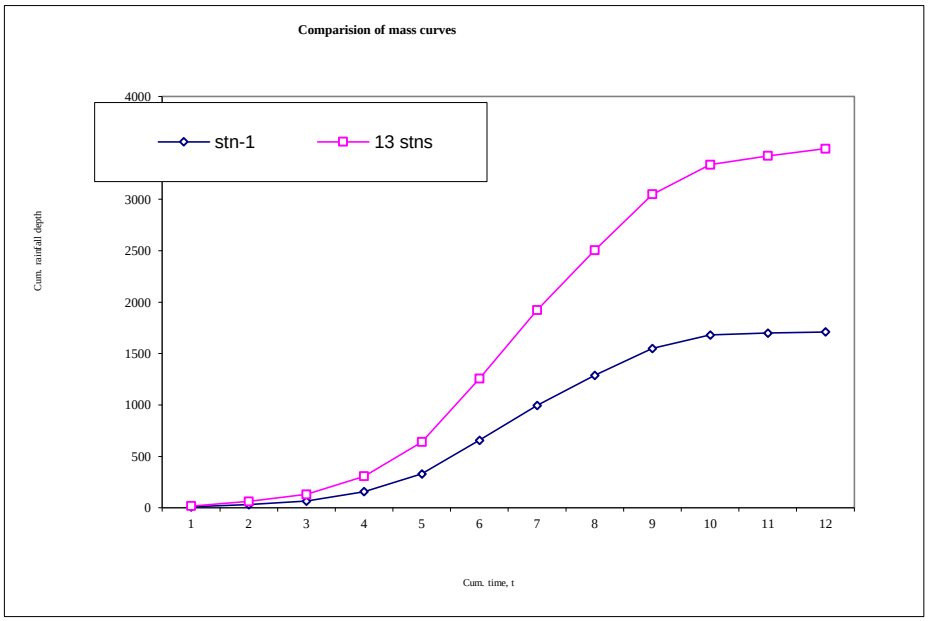
Since, there is no significant change in trend in the double mass plot, the data is considered as consistent.

ii) Mass curve

The variation of rainfall with respect to time may be shown graphically by a mass curve. A mass curve of rainfall is a plot of cumulative depth of rainfall against time. From the mass curve total depth of rainfall and intensity of rainfall at any instant of time can be found.

Figure A1.3: Comparison of mass curve of cumulative average rainfall depth at station– 1 and the 13 stations vs. time.

iii) Relation curve between tested station and base station:



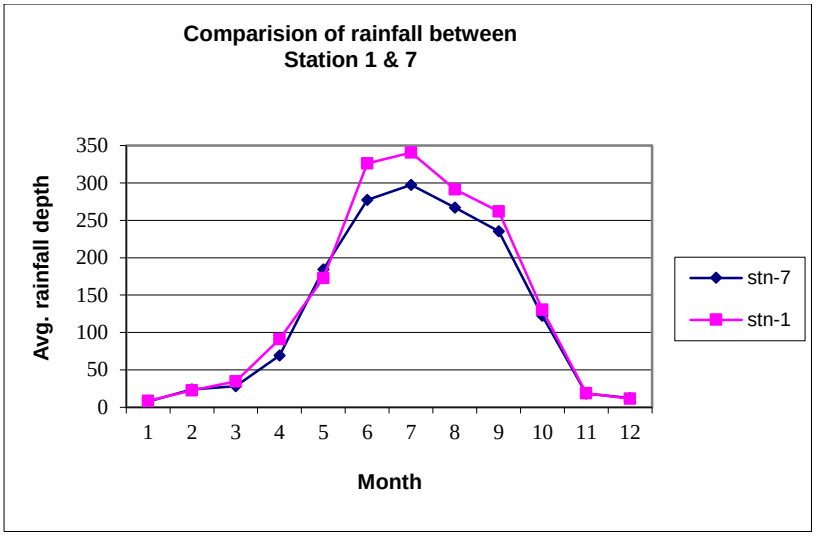


Figure A1.4: Comparison of rainfall depth between rainfall station-1 and station-7

It seems that the rainfall depth of the station-1 and station-7 has relatively change.

D. Step 4- Correlation test

If the data set comes from normal distribution then parametric test is applicable for that data set otherwise nonparametric test is applicable. For correlation test Pearson's correlation test is parametric test and Spearman's rank correlation test is nonparametric test. Example of two types of correlation test is given here.

i) Spearman's rank correlation test

Table A1.8: Spearman's rank correlation test

Avg. rainfall of Test station	Rank	Avg. rainfall of Base station	Rank	d	d^2
120	22	104	31	-9	81
112	26.5	97	34	-8	56
151	10.5	163	13.5	-3	9
108	29	131	20	9	81
104	32	116	26	6	36
128	19	196	2	17	289
115	24	108	29.5	-6	30
143	13.5	186	4	10	90
151	10.5	169	6	5	20
142	15	197	1	14	196
91	33	108	29.5	4	12
162	3.5	165	10.5	-7	49
127	20	144	16	4	16
133	17	121	24	-7	49
154	7.5	166	8.5	-1	1
159	5	173	5	0	0
107	30	110	28	2	4
112	26.5	129	22	5	20
121	21	123	23	-2	4
130	18	168	7	11	121
89	34	103	32	2	4
143	13.5	187	3	11	110
167	2	164	12	-10	100
113	25	112	27	-2	4
157	6	159	15	-9	81
148	12	163	13.5	-2	2
162	3.5	130	21	-18	306
119	23	138	17	6	36
154	7.5	166	8.5	-1	1
186	1	165	10.5	-10	90
109	28	100	33	-5	25
152	9	136	18	-9	81
105	31	117	25	6	36
139	16	133	19	-3	9

$$R_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

$$R_s = 1 - \frac{6\sum 2052}{34^3 - 34}$$

$$R_s = 0.6865$$

It shows that the series are strongly correlated. [R_s fall between 0.5 to 1 - strong positive correlation]

ii) Pearson's correlation test

Table A1.9: Pearson's correlation test

Avg. rainfall of Test station, X_i	Avg. rainfall of Base station, Y_i	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean})^2$	$(Y_i - Y_{mean})^2$	$(X_i - X_{mean})*(Y_i - Y_{mean})$
120	104	-12	-39	154.28	1486.78	478.94
112	97	-21	-46	432.82	2075.61	947.82
151	163	18	20	318.47	417.84	364.78
108	131	-24	-12	597.20	133.61	282.47
104	116	-28	-27	804.92	705.37	753.50
128	196	-5	53	24.88	2855.96	-266.55
115	108	-18	-35	314.04	1194.31	612.42
143	186	11	43	111.56	1887.14	458.84
151	169	18	26	319.66	699.14	472.74
142	197	9	54	88.90	2963.84	513.32
91	108	-42	-35	1730.93	1194.31	1437.80
162	165	29	22	831.59	503.61	647.14
127	144	-6	1	31.69	2.08	-8.11
133	121	0	-22	0.13	464.78	-7.63
154	166	22	23	463.85	549.49	504.86
159	173	27	30	706.88	926.67	809.35
107	110	-26	-33	670.18	1060.08	842.87
112	129	-21	-14	435.95	183.84	283.10
121	123	-12	-20	136.21	382.55	228.27
130	168	-3	25	7.27	647.25	-68.59
89	103	-44	-40	1956.98	1564.90	1749.99
143	187	11	44	113.15	1975.02	472.73
167	164	34	21	1153.43	459.72	728.19
113	112	-20	-31	401.51	933.84	612.33
157	159	24	16	585.43	270.31	397.80
148	163	15	20	226.62	417.84	307.72
162	130	29	-13	850.44	157.72	-366.24
119	138	-14	-5	199.17	20.78	64.34
154	166	21	23	431.76	549.49	487.08
186	165	54	22	2868.02	503.61	1201.81
109	100	-23	-43	549.72	1811.25	997.84
152	136	20	-7	385.62	43.02	-128.80
105	117	-27	-26	750.55	653.25	700.21
139	133	6	-10	40.06	91.37	-60.50
133	143			550	876	484

$$r = \frac{\sum (X_i - X_{mean})(Y_i - Y_{mean})}{\left(\sum (X_i - X_{mean})^2 \sum (Y_i - Y_{mean})^2\right)^{0.5}}$$

r = 0.6972

It shows that the series are strongly correlated.

iii) Correlation among surrounding stations

Table A1.10: Correlation with Pabna Station

Average rainfall of Test station, X_i	Average rainfall of Base station, Y_i	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean})^2$	$(Y_i - Y_{mean})^2$	$(X_i - X_{mean})*(Y_i - Y_{mean})$
124	104	-5	-39	22.94	1486.78	184.69
96	97	-33	-46	1074.14	2075.61	1493.15
144	163	15	20	235.40	417.84	313.62
115	131	-14	-12	201.14	133.61	163.93
86	116	-43	-27	1806.88	705.37	1128.95
119	196	-10	53	91.98	2855.96	-512.54
101	108	-28	-35	768.16	1194.31	957.82
146	186	17	43	298.17	1887.14	750.13
169	169	41	26	1645.73	699.14	1072.66
143	197	14	54	201.90	2963.84	773.57
98	108	-31	-35	943.45	1194.31	1061.50
150	165	21	22	455.86	503.61	479.14
131	144	3	1	6.51	2.08	3.68
142	121	13	-22	158.99	464.78	-271.84
136	166	7	23	47.16	549.49	160.99
153	173	24	30	581.66	926.67	734.17
103	110	-26	-33	679.86	1060.08	848.94
116	129	-13	-14	172.68	183.84	178.17
119	123	-10	-20	94.23	382.55	189.86
140	168	11	25	115.40	647.25	273.31
85	103	-44	-40	1936.65	1564.90	1740.88
166	187	37	44	1342.68	1975.02	1628.44
162	164	33	21	1072.63	459.72	702.22
98	112	-31	-31	942.94	933.84	938.38
146	159	17	16	294.16	270.31	281.98
122	163	-7	20	42.78	417.84	-133.70
148	130	19	-13	377.04	157.72	-243.86
105	138	-24	-5	581.97	20.78	109.98
139	166	10	23	106.80	549.49	242.25
189	165	60	22	3553.27	503.61	1337.70
114	100	-15	-43	210.71	1811.25	617.77
140	136	11	-7	126.40	43.02	-73.74
113	117	-16	-26	244.11	653.25	399.33
123	133	-6	-10	36.69	91.37	57.90
129	143			602	876	517

Correlation coefficient between Base station with Pabna Station, r = 0.7123

It shows strong correlation between two stations.

Correlation with Ishwardi station

Table A1.11: Correlation with Ishwardi station

Average rainfall of base station, X_i	Average rainfall of test station, Y_i	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean})^2$	$(Y_i - Y_{mean})^2$	$(X_i - X_{mean})*(Y_i - Y_{mean})$
132	104	6	-39	40.38	1486.78	-245.03
166	97	40	-46	1620.44	2075.61	-1833.95
174	163	48	20	2285.21	417.84	977.17
114	131	-12	-12	141.32	133.61	137.41
107	116	-18	-27	339.64	705.37	489.46
131	196	5	53	23.72	2855.96	260.29
107	108	-19	-35	347.06	1194.31	643.81
138	186	12	43	147.31	1887.14	527.25
124	169	-2	26	2.96	699.14	-45.51
160	197	34	54	1179.62	2963.84	1869.81
84	108	-41	-35	1710.19	1194.31	1429.16
134	165	8	22	67.03	503.61	183.73
48	144	-77	1	5995.32	2.08	-111.59
93	121	-33	-22	1089.29	464.78	711.54
95	166	-30	23	917.35	549.49	-709.98
165	173	40	30	1568.47	926.67	1205.59
99	110	-27	-33	722.06	1060.08	874.89
124	129	-1	-14	1.90	183.84	18.70
114	123	-12	-20	133.70	382.55	226.15
134	168	9	25	74.46	647.25	219.53
84	103	-42	-40	1761.58	1564.90	1660.33
138	187	12	44	146.71	1975.02	538.28
150	164	24	21	591.89	459.72	521.64
95	112	-31	-31	963.34	933.84	948.48
135	159	9	16	79.58	270.31	146.66
167	163	41	20	1706.01	417.84	844.30
144	130	18	-13	324.74	157.72	-226.32
122	138	-4	-5	17.05	20.78	18.83
182	166	56	23	3139.24	549.49	1313.38
154	165	28	22	794.05	503.61	632.37
113	100	-12	-43	151.19	1811.25	523.31
131	136	5	-7	30.02	43.02	-35.93
96	117	-30	-26	883.84	653.25	759.85
121	133	-5	-10	23.40	91.37	46.24
126	143			854	876	427

Correlation between test station and Ishwardi, r = 0.494

The value shows weak correlation between two stations.

Correlation between Test station and Chatmohor station

Table A1.12: Correlation with Chatmohor station

Average rainfall of base station, X_i	Average rainfall of test station, Y_i	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean})^2$	$(Y_i - Y_{mean})^2$	$(X_i - X_{mean})*(Y_i - Y_{mean})$
228	163	112	18	12642.39	317.29	2002.81
56	131	-59	-14	3457.17	201.29	834.19
67	116	-48	-29	2328.65	851.91	1408.47
80	196	-35	51	1251.23	2581.91	-1797.38
81	108	-34	-37	1176.91	1382.91	1275.76
140	186	24	41	599.14	1665.66	998.98
145	169	29	24	869.89	567.04	702.32
94	197	-21	52	448.29	2684.54	-1097.01
87	108	-28	-37	764.86	1382.91	1028.46
226	165	111	20	12301.16	392.54	2197.42
129	144	14	-1	190.73	1.41	-16.40
123	121	8	-24	69.48	585.04	-201.62
177	166	62	21	3842.21	433.16	1290.08
141	173	26	28	685.25	773.54	728.05
53	110	-63	-35	3923.70	1238.16	2204.12
77	129	-38	-16	1463.53	262.04	619.27
81	123	-34	-22	1150.75	492.29	752.66
97	168	-18	23	325.42	520.41	-411.52
49	103	-66	-42	4376.63	1779.79	2790.96
53	187	-62	42	3816.90	1748.29	-2583.22
0	164	-115	19	13257.08	353.91	-2166.06
173	112	58	-33	3309.39	1101.41	-1909.19
156	159	41	14	1664.82	190.79	563.58
145	163	30	18	908.15	317.29	536.79
152	130	37	-15	1390.84	230.66	-566.40
120	138	5	-7	24.94	51.66	-35.89
109	166	-6	21	32.37	433.16	-118.41
159	165	44	20	1930.34	392.54	870.47
102	100	-13	-45	167.86	2041.91	585.45
166	136	51	-9	2629.36	84.41	-471.11
94	117	-21	-28	461.08	794.54	605.26
123	133	8	-12	64.30	148.54	-97.73
115	145			2548	813	329

Correlation coefficient, r = 0.2286, weak correlation.

Table A1.13: Correlation between Tested station and Base stations

Station ID	Station name	Atghoria Station
25	Pabna	0.7123
15	Ishurdi	0.494
7	Chatmohar	0.2285

Hence, it seems that the tested station is most closely related with the Pabna rainfall station.

Step 5: Filling missing value

Data of January, February and March of 1962 at Atghoria station is missing. Atghoria station is highly correlated with Pabna station. So for estimating missing value Pabna station gets first priority.

Table A1.14: Estimation of missing value

Station name	January	February	March
Average rainfall at test station	8.77	23.20	35.62
Average rainfall at Pabna station	7.4	23.94	28.27
In 1962 rainfall at Pabna station	17.8	16.8	26.2
Filling missing value for Atghoria station 1962	21.1	16.3	33

Missing value of January at Atghoria, P_A =

$$\left(\frac{\text{Long time average for January of Atghoria}}{\text{No. of neighboring station}}\right) * \left(\frac{\text{Average discharge of neighboring station for that particular year}}{\text{Long time average for JAnuary of neighboring station}}\right)$$

= (8.77/1) * (17.4/7.4)

= 20.62 mm

Step 6: Test for Normality

H_0: The data are normal

H_1: The data are not normal

If $r < r_{critical}$ – reject null hypothesis

Determination of X_T:

Observed data are arranged in descending order

Calculate, p = (i – a)/(n+1-2a), a = plotting position for normal distribution a = 0.375, I = rank,

n = sample size = 27

Then, return period, T = 1/p, p and T determined for each sample and then determine Z_T for each sample by using following formula

Standard Normal Variate, $Z_T = \dfrac{\left(1-\frac{1}{T}\right)^{0.135} - \left(\frac{1}{T}\right)^{0.135}}{0.1975}$

$$= \frac{\left(1-\frac{1}{43.6}\right)^{0.135} - \left(\frac{1}{43.6}\right)^{0.135}}{0.1975}$$

= 2.0058

$$X_T = X_{mean} + Z_T \sigma_x$$

Where, X_{mean} = 143, σ_x = 30.04

So, $X_T = 143 + 30.04 * 2.102$

$X_T = 206$

Table A1.15: Determination of probability plot correlation coefficient

Ranked data	X_T	$X_i - X_{mean}$	$X_{Ti} - X_{Tmean}$	$(X_i - X_{mean})^2$	$(X_{Ti} - X_{Tmean})^2$	$(X_i - X_{mean}) * (X_{Ti} - X_{Tmean})$
197	206	54	63	2964	3986	3437
196	193	53	50	2856	2532	2689
187	186	44	43	1975	1844	1909
186	180	43	38	1887	1407	1629
173	176	30	33	927	1094	1007
169	172	26	29	699	857	774
168	168	25	26	647	671	659
166	165	23	23	549	522	535
166	163	23	20	549	400	469
165	160	22	17	504	300	389
165	157	22	15	504	219	332
164	155	21	12	460	153	265
163	153	20	10	418	100	205
163	150	20	8	418	60	158
159	148	16	5	270	30	90
144	146	1	3	2	11	5
138	144	-5	1	21	1	-5
136	141	-7	-1	43	1	7
133	139	-10	-3	91	11	31
131	137	-12	-5	134	30	63
130	135	-13	-8	158	60	97
129	133	-14	-10	184	100	136
123	130	-20	-12	383	153	242
121	128	-22	-15	465	219	319
117	125	-26	-17	653	300	443
116	123	-27	-20	705	400	531
112	120	-31	-23	934	522	698
110	117	-33	-26	1060	671	844
108	113	-35	-29	1194	857	1012
108	109	-35	-33	1194	1094	1143
104	105	-39	-38	1487	1407	1446
103	100	-40	-43	1565	1844	1699
100	92	-43	-50	1811	2532	2142
97	79	-46	-63	2076	3986	2877
143	143			29786	28377	28277

$$r = \frac{\sum (X_i - \overline{X}) * (X_{Ti} - \overline{X_T})}{\left(\sum (X_i - \overline{X})^2 * \sum (X_{Ti} - \overline{X_T})^2 \right)^{0.5}}$$

r = 0.973

$r_{critical}$ = 0.968

r > $r_{critical}$, null hypothesis accepted that data are from normal distribution. So, parametric tests can be applied for this data set.

Step 7: Test for shift in the mean

Parametric Test: t-Test

H_0: Equal mean, no change in the mean

H_a: Shift in the mean

After infilling the missing value one can check the data set if the mean of the new data set is deviated from the mean of the previous data set.

Formula: $T_c = \dfrac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}}$

where, $S = \sqrt{\dfrac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}}$, where N = no. of sample, $\overline{y}$ = mean of the sample and s denotes standard deviation, subscript 1 and 2 describes sample data before and after infilling.

Here, N_1 = 33, $\overline{y_1}$ = 8.77 and s_1 = 12.1

N_2 = 34, $\overline{y_2}$ = 9.14 and s_2 = 12.1

$$S = \sqrt{\frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}} = \sqrt{\frac{(33-1)*12.1 + (34-1)*12.1}{67 - 2}}$$

S = 3.5

$$T_c = \frac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}} = \frac{|8.77 - 9.14|}{3.5\sqrt{\dfrac{1}{33} + \dfrac{1}{33}}} = 0.4$$

$T_{1-\alpha/2}$ = 1.96

$T_c < T_{1-\alpha/2}$, so hypothesis is accepted and acceptation of hypothesis is considered as no detection of shift.

Non-parametric test: Mann-Whitney Test

Step 8: Trend test

T-test: test for linear trend with other station

H_0: no trend

H_a: Linear trend

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right|$$

where, N = number of sample = 34

r = Cross correlation co-efficient = 0.7123

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right| = \left| \frac{\sqrt{34-2}}{0.7123\sqrt{1-0.7123^2}} \right| = 11.315$$

$T_{1-\alpha/2,\nu} = 1.96$

$T_c > T_{1-\alpha/2,\nu}$, So null hypothesis is rejected. Rejection of hypothesis considered as a detection of linear trend.

This test also can check trend between two months of one station by determining correlation between those months.

Test for trend: Mann-Kendall test

H_0: The series is random no trend series

H_a: The series has trend either upward or downward

$$u_c = \frac{S+m}{\sqrt{V(S)}}$$

where, S = Mann-Kendall statistic, m = 1 when S < 0 and m = -1 when S > 0

Yearly average data has been taken for Mann-Kendall test.

Calculation for Mann-Kendall statistic is shown in table (A1-16)

Here, S < 0, m = 1

$$V(S) = \frac{1}{18}\left[N(N-1)(2N+5) - \sum_{i=1}^{n} e_i(e_i-1)(2e_i+5) \right]$$

Where, e_i is the number of data in tied group and n is the number of tied group. In this data set there is five tied group, n = 5 and e_i = 10.

N = 34

V(S) = 4545.333

$u_c = (34+1)/\sqrt{4545.333} = 0.1335$

$u_{1-\alpha/2} = 1.96$

$u_c < u_{1-\alpha/2}$, so the null hypothesis is accepted and acceptance of hypothesis indicates no trend random series. So, trend analysis is not necessary.

Table A1.16: Determination of Mann-Kendall Statistics

110	97	163	131	116	196	108	186	169	197	108	165	144	121	166	173	110	129	123	168	103	187	164	112	159	163	130	138	166	165	100	136	117	133	Sum
	-1	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	0	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	1	1	1	22
		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	32
			-1	-1	1	-1	1	1	1	-1	1	-1	-1	1	1	-1	-1	-1	1	-1	1	1	-1	-1	0	-1	-1	1	1	-1	-1	-1	-1	-6
				-1	1	-1	1	1	1	-1	1	1	-1	1	1	-1	-1	-1	1	-1	1	1	-1	1	1	-1	1	1	1	1	-1	1	-1	1	6
					1	-1	1	1	1	-1	1	1	1	1	1	-1	1	1	1	-1	1	1	-1	1	1	1	1	1	1	-1	1	1	1	17
						-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-26
							1	1	1	0	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	1	1	1	22
								-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-22
									1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-19
										-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-24
											1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	1	1	1	19
												-1	-1	1	1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	1	0	-1	-1	-1	-1	-11
													-1	1	1	-1	-1	-1	1	-1	1	1	-1	1	1	-1	-1	1	1	-1	-1	-1	-1	-3
														1	1	-1	1	1	1	-1	1	1	-1	1	1	1	1	1	1	-1	1	-1	1	10
															1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-12
																-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-16
																	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	1	1	1	13
																		-1	1	-1	1	1	-1	1	1	1	1	1	1	-1	1	-1	1	6
																			1	-1	1	1	-1	1	1	1	1	1	1	-1	1	-1	1	7
																				-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-12
																					1	1	1	1	1	1	1	1	1	-1	1	1	1	11
																						-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-12
																							-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-7
																								1	1	1	1	1	1	-1	1	1	1	8
																									1	-1	-1	1	1	-1	-1	-1	-1	-3
																										-1	-1	1	1	-1	-1	-1	-1	-4
																											1	1	1	-1	1	-1	1	3
																												1	1	-1	-1	-1	-1	-2
																													-1	-1	-1	-1	-1	-5
																														-1	-1	-1	-1	-4
																															1	1	1	3
																																-1	-1	-2
																																	1	1
																																	S	-10

Step 9: Trend Analysis

Not needed.

Step 10: Selection of distribution

Goodness of fit (Chi-square) test

Comparing the theoretical and sample values of the relative frequency or the cumulative frequency function can test the goodness of fit of a probability distribution. In the case of the relative frequency function, the Chi-square test is used.

Table A1.17: Goodness of fit test

Column	1	2	3	4	5	6	7	8
Interval, I	**Range(in)**	**N_I**	**$Fs(x_i)$ (n_i/n)**	**$Fs(x_i)$**	**z_i**	**$F(x_i)$**	**$P(x_i)$**	**χ^2**
1	0-55	5	0.417	0.417	-0.67	0.251	0.251	1.317
2	55-110	1	0.083	0.500	-0.25	0.440	0.189	0.713
3	110-165	1	0.083	0.583	0.17	0.567	0.127	0.183
4	165-220	1	0.083	0.666	0.59	0.722	0.155	0.401
5	220-275	1	0.083	0.749	1.01	0.844	0.122	0.150
6	275-330	2	0.167	0.916	1.43	0.922	0.078	1.219
7	330-385	1	0.083	0.999	1.85	0.968	0.046	0.357
	Total=12		**1.000**					**4.340**

Mean $y_m = 142.5$

Standard deviation $\sigma = 131.31$

Column 1: equally ranges data

Column 2: no. of data within range

Column 3: ration of the no. of certain ranges of data and the total no. of data

Column 4: Cumulative value of column 3.

Column 5: $z = (x- y_m)/\sigma = (41.5-143.5)/51 = -2$

Column 6: From cumulative probability of standard normal distribution table

Column 7: Intervals of the values of column 6.

Column 8: $n[fs(x_i) - P(x_i)]^2/ P(x_i)$

From the cumulative probability of the standard normal distribution table find the value corresponding to z values. To employ the table for z<0, use Fz (z) = 1 - Fz (|z|)

Where Fz(|z|) is the tabulated value.

To check the goodness of fit, the

$$\frac{n[f_s(x_1) - p(x_1)]^2}{p(x_1)} = \frac{12(0.417-0.251)^2}{0.251} = 1.317$$

$$p(x_2) = .440-.251 = 0.189$$

χ^2 test statistic is calculated by

The total of the values in column 8 is 4.340. The value of $\chi^2_{v,\ 1-\alpha}$ for a cumulative probability of $1 - \alpha = 0.95$ and degree of freedom v = m-p-1 = 7 – 2 – 1= 4 is $\chi^2_{7,\ 0.95} = 9.49$. Since this value is greater than $\chi^2_{c,}$ the null hypothesis (the distribution fits the data) cannot be rejected at the 95 percent confidence level; the fit of the normal distribution to the Simulbari station rainfall data is accepted. If the distribution had fitted poorly, the values of $f_s(x_i)$ and $P(x_i)$ would have been quite different from one another, resulting in a value of χ^2_c larger than 9.49, in which case the null hypothesis would have been rejected.

Hence the test of goodness to fit shows that the data are well fitted in the series. Other methods for selecting the fitted distribution are described in 'Technical notes on statistical analysis'.

Step 11: Frequency Analysis

Described in 'Technical notes on statistical analysis'

Step 12: Comments

It is analyzed that the set of data at the station at Atghoria (station-ID-1) is good quality. Different analysis shows that the data set is within the accepted limit. The data set is suitable for planning purposes.

Exercise A2: Discharge data

Following tables show the monthly average discharge data of Baruria Transit and Hardinge Bridge has been recorded from 1966 to 1994. Steps for quality checking of time series discharge data are described below. In this example Baruria transit is taken as test station.

Step 1: Raw Data

Table A2.1: Monthly average discharge data of Baruria Transit

Year	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1966 – 67	7814	11860	31953	48419	65219	61050	24771	12157	8915	6805	5934	6041	290939	24245
1967 – 68	6930	11957	25463	51594	49581	53517	30742	12770	8744	6895	6160	5977	270330	22527
1968 – 69	8202	13404	31800	60361	58990	42010	37345	13653	8899	6628	5735	5628	292655	24388
1969 – 70	6957	11376	28800	51939	61142	58957	31897	14040	9503	7238	6481	6944	295273	24606
1970 – 71	9627	20661	34520	57832	71116	55380	39797	16597	9712	7503	6580	6328	335654	27971
1971 – 72														
1972 – 73	9597	17826	30470	49390	57948	53780	29194	15393	10204	7450	6363	6675	294290	24524
1973 – 74	7914	14752	38187	47871	79181	66030	49571	22547	12500	8555	6655	6351	360113	30009
1974 – 75	9024	19829	34113	69690	96539	78923	41529	18677	10937	7583	6573	6205	399623	33302
1975 – 76	9911	15335	28133	65126	78819	75647	42968	18403	9971	6750	5530	5722	362315	30193
1976 – 77	7429	12318	28210	53561	66526	66213	28926	14693	10057	7315	6416	7054	308719	25727
1977 – 78	11186	19345	38490	53587	75003	65357	42423	18990	11067	7788	6261	6215	355712	29643
1978 – 79	8467	16039	37723	60755	74352	62793	39068	16757	10792	7073	5983	5817	345617	28801
1979 – 80														
1980 – 81	9188	18287	32980	60845	93532	76313	37965	17377	9036	6028	5365	6056	372973	31081
1981 – 82	8922	12757	23147	66145	76465	60803	27416	12150	7420	5853	4571	5309	310957	25913
1982 – 83	9152	17371	28983	54410	64032	72373	29629	14380	11229	9951	9380	6260	327150	27263
1983 – 84	11736	20935	28167	50235	57803	82770	52716	19280	9347	6911	6016	5393	351310	29276
1984 – 85	9684	19184	39530	70071	68487	88217	36758	16367	9707	6930	5804	6981	377720	31477
1985 – 86	10992	15241	37233	67129	66081	69983	56674	20990	10415	7138	6255	6447	374580	31215
1986 – 87	9343	13687	19890	54155	63406	58103	40690	15870	9014	6559	4530	5229	300477	25040
1987 – 88	9260	11703	22987	55090	91319	81113	44223	19710	10309	6678	5606	7105	365104	30425
1988 – 89	10288	19619	37953	66800	92477	86250	40810	16813	10577	7314	6046	5928	400877	33406
1989 – 90	8119	17458	40373	66174	59303	62907	47816	19043	10531	7310	6321	6856	352212	29351
1990 – 91	10829	19745	40576	66603	77003	65447	56561	18300	10155	6812	4975	5551	382557	31880
1991 – 92	9970	20381	39270	69823	75926	81833	38442	16483	10027	7120	5462	5576	380313	31693
1992 – 93	10725	15048	19983	47810	55535	57497	32739	17500	9225	6492	5756	5403	283713	23643
1993 – 94	7299	19077	37433	63839	76087	77183	49955	18853	10446	6776	5603	5572	378124	31510
Sum	238563	425198	836370	1529255	1851874	1760450	1030623	437793	258742	185455	156363	158622		
Average	9176	16354	32168	58817	71226	67710	39639	16838	9952	7133	6014	6101		

Table A2.2: Monthly average data of Hardinge Bridge

Year	April	May	June	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1966 – 67	1656	1724	3208	13011	31874	25670	7630	4138	3128	2309	1951	1551	97850	8154
1967 – 68	1528	1615	2832	15247	30019	43390	11985	5339	3783	3064	2710	2212	123726	10310
1968 – 69	1776	1745	4259	22018	36984	20543	17667	5999	3736	2729	2204	1863	121521	10127
1969 – 70	1741	1904	4532	16547	43790	34940	18526	7234	4532	3065	2660	2507	141979	11832
1970 - 71	2207	2305	4763	19452	33310	34197	15202	6263	3843	2638	2294	2197	128669	10722
1971 - 72														
1972 - 73	2337	2562	3969	11866	22129	28343	12196	6250	4292	3170	2838	2389	102341	8528
1973 - 74	2013	2746	7382	17506	41029	40767	26494	9545	5476	3886	2936	2393	162173	13514
1974 - 75	2394	2487	4281	18274	45432	34220	15368	6775	4004	2722	2345	1946	140249	11687
1975 - 76	1880	1713	3466	29313	44039	39563	20042	6978	3405	1882	1470	830	154581	12882
1976 - 77	728	1323	4413	14911	34558	45097	13067	4565	2435	1424	1100	912	124533	10378
1977 - 78	967	1167	2484	21282	43035	33693	17308	5828	3379	2281	1762	1550	134738	11228
1978 - 79	1729	2415	5921	26381	52355	44007	20526	6649	3467	1884	1731	1413	168476	14040
1979 - 80	1269	1501	1482	14259	29448	15119	7807	2754	1955	1335	1029	911	78869	6572
1980 - 81	943	1216	3294	25274	48013	44517	14487	5106	2727	1639	1334	991	149542	12462
1981 - 82	1107	1578	2401	22768	41406	29023	13429	4099	2494	1441	1383	1281	122409	10201
1982 - 83	1577	1840	4526	10155	33319	47750	9781	4482	2779	1382	1136	869	119597	9966
1983 - 84	820	1534	2235	12468	25668	44767	24939	7198	3323	2105	1551	1240	127846	10654
1984 - 85	981	1393	7831	22794	33803	46173	11273	4455	2381	1532	1147	873	134635	11220
1985 - 86	771	877	1675	16523	38803	35830	37129	12197	4441	2729	1822	1705	154503	12875
1986 - 87	1284	1652	2455	25574	40745	30090	17093	6050	3193	2266	1432	993	132827	11069
1987 - 88	1066	1299	2019	14810	44629	56450	14348	6200	2524	1375	1085	994	146800	12233
1988 - 89	1030	1475	2957	20689	51865	40703	10610	3592	2032	1280	856	512	137600	11467
1989 - 90	607	998	5139	17890	26339	26900	15992	4299	2153	1304	672	788	103080	8590
1990 - 91	870	1547	4497	30583	44119	32023	19024	5077	2255	1459	775	577	142806	11901
1991 - 92	730	1151	5299	11692	27855	40297	9910	3786	2228	1475	810	499	105732	8811
1992 - 93	430	555	902	5848	21939	27637	9524	4438	2228	1199	544	316	75558	6296
1993 - 94	402	1081	2052	10670	23861	36330	17958	5162	2679	1286	1111	636	103228	8602
Sum	34841	43403	100271	487805	990368	978039	429315	154458	84872	54862	42686	34947		
Average	1290	1608	3714	18067	36680	36224	15901	5721	3143	2032	1581	1294		

Step 2: Checking for validation of data

Test for randomness of the data: Turning point test

The Turning point test is used to check whether the time series verify the following hypothesis:

H_0: The series is a random no trend series

H_a: The series has trend and/or autocorrelated errors.

With this test one can check yearly or monthly randomness of data set. The example describes yearly randomness checking of the data set:

Mean = 2(N-2)/3 and Standard Deviation = $[(16N-29)/90]^{0.50}$

Standard Value of T, t = (T-mean)/standard Deviation

Table A2.3: Turning point determination at Baruria Transit within the data range years

Yearly Avg. Discharge in cumec	Range	No of turning	Cumulative no. of turning
24245			
22527	24245>22527<24388	yes	1
24388	22527<24388<24606	no	1
24606	24388<24606<27971	no	1
27971	24606<27971>24524	yes	2
24524	27971>24524<30009	yes	3
30009	24524<30009<33302	no	3
33302	30009<33302>30193	yes	4
30193	33302>30193>25727	no	4
25727	30193>25727<29643	yes	5
29643	25727<29643>28801	yes	6
28801	29643>28801<31081	yes	7
31081	28801<31081>25913	yes	8
25913	31081>25913<27263	yes	9
27263	25913<27263<29276	no	9
29276	27263<29276<31477	no	9
31477	29276<31477>31215	yes	10
31215	31477>31215>25040	no	10
25040	31215>25040<30425	yes	11
30425	25040<30425<33406	no	11
33406	30425<33406>29351	yes	12
29351	33406>29351<31880	yes	13
31880	29351<31880>31693	yes	14
31693	31880>31693>23643	no	14
23643	31693>23643<31510	yes	15
31510			
			T = 15

Therefore, T = 15 and N = 26

T follows normal distribution with, Mean = 16 and

$$\text{Standard Deviation} = [(16N-29)/90]^{0.50} = 2.074$$

Therefore, t = 0.482

$$t_{1-\alpha/2} = 1.96$$

Since $t < t_{1-\alpha/2}$, the null hypothesis says that the series is a random no trend series.

Step 3: Consistency Checking

Double mass analysis

Table A2.4: Cumulative average of test and base station

Year	1966 - 67	1967 - 68	1968 - 69	1969 - 70	1970 - 71	1971 - 72	1972 - 73	1973 - 74	1974 - 75
Test Station	24245	46772	71160	95766	123738	148262	178271	211573	241766
Base Station	8154	18465	28591	40423	51145	59674	73188	84876	97757
Year	1975 - 76	1976 - 77	1977 - 78	1978 - 79	1979 - 80	1980 - 81	1981 - 82	1982 - 83	1983 - 84
Test Station	267493	297135	325937	357018	382931	410193	439469	470946	502161
Base Station	108135	119363	133403	139975	150176	160143	170796	182016	194891
Year	1984 - 85	1985 - 86	1986 - 87	1987 - 88	1988 - 89	1989 - 90	1990 - 91	1991 - 92	
Test Station	527201	557626	591032	620383	652263	683956	707599	739109	
Base Station	205960	218193	229660	238250	250151	258962	265258	273860	

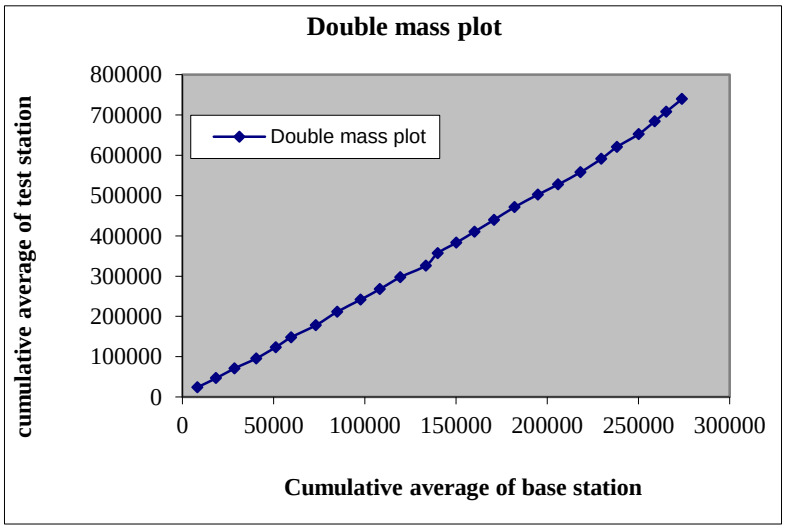


Figure A2. 1: Double mass plot

Since there is no significant change in the trend in double mass plot, data can be considered as consistent data.

Step 4: Correlation Test

Correlation between Baruria Transit and Hardinge Bridge

Table A2.5: Determination of correlation

X_i	Y_i	X_i-X_{mean}	Y_i-Y_{mean}	(X_i-X_{mean})*(Y_i-Y_{mean})
24245	8154	-347832	-136518	47485231719
46772	18465	-325305	-126207	41055785855
71160	28591	-300917	-116080	34930545329
95766	40423	-276311	-104249	28805060543
123738	51145	-248340	-93526	23226297311

X_i	Y_i	X_i-X_{mean}	Y_i-Y_{mean}	(X_i-X_{mean})*(Y_i-Y_{mean})
148262	59674	-223815	-84998	19023855309
178271	73188	-193806	-71484	13853933760
211573	84876	-160504	-59796	9597524790
241766	97757	-130311	-46914	6113467902
267493	108135	-104585	-36537	3821168104
297135	119363	-74942	-25308	1896665966
325937	133403	-46140	-11269	519948053.1
357018	139975	-15059	-4696	70725441.07
382931	150176	10854	5504	59742675.91
410193	160143	38116	15471	589686411.5
439469	170796	67392	26125	1760587990
470946	182016	98869	37344	3692166106
502161	194891	130084	50219	6532721455
527201	205960	155124	61288	9507257884
557626	218193	185549	73522	13641847184
591032	229660	218955	84988	18608622703
620383	238250	248306	93578	23236063095
652263	250151	280186	105479	29553672712
683956	258962	311879	114290	35644550837
707599	265258	335522	120586	40459282940
739109	273860	367032	129189	47416339694
Mean	Mean			Sum
372077	144672			4.61103E+11
Stdev	Stdev			
221689	83263			

X_{mean} = 28427

Y_{mean} = 10605

S_x = 25017

S_y = 13336

Co-efficient of correlation, r = 0.0.961

In Ganges there is only two discharge station so correlation with other station cannot be judged for infilling the missing data. If there remains more than two station correlation between test station with other station should be checked for selecting the station which would be used for infilling the missing data.

Step 5: Missing data

Missing data years at Baruria Transit are as follows:

- ◊ 1971 – 72
- ◊ 1979 – 80

Since, data of 1971 – 72 at Hardinge Bridge is also missing, infilling of data 1979 – 80 is only possible with the help of Hardinge Bridge data

Infilling of Missing data

Table A2.6: Infilling of missing data with the help of neighboring station

Station name	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Annual Discharge Baruria Transit (test)	9176	16354	32168	58817	71226	67710	39639	16838	9952	7133	6014	6101
Annual Discharge Hardinge Bridge (base)	1290	1608	3714	18067	36680	36224	15901	5721	3143	2032	1581	1294
In 1979 - 80 data at Hardinge Bridge	1269	1501	1482	14259	29448	15119	7807	2754	1955	1335	1029	911
Filling missing value for B. Transit for 1979 - 80	9021	15270	12840	46420	57183	28260	19464	8105	6189	4688	3913	4292

Missing value of April at Baruria Transit, D_A =

$$\left(\frac{\text{Long time average for April of B. Transit}}{\text{No. of neighboring station}}\right) * \left(\frac{\text{Average discharge of neighboring station for that particular year}}{\text{Long time average for April of neighboring station}}\right)$$

=(9176/1) * (1269/1290)

= 9021 m^3/sec

Step 6: Normality Test

H_0: The data are normal

H_1: The data are not normal

If $r < r_{critical}$ – reject null hypothesis

Determination of X_T:

Observed data are arranged in descending order

Calculate, p = (i – a)/(n+1-2a), a = plotting position for normal distribution a = 0.375, I = rank, n = sample size = 27

Then, return period, T = 1/p, p and T determined for each sample and then determine Z_T for each sample by using following formula

Standard Normal Variate, $Z_T = \dfrac{\left(1-\frac{1}{T}\right)^{0.135} - \left(\frac{1}{T}\right)^{0.135}}{0.1975}$

$$= \frac{\left(1-\frac{1}{43.6}\right)^{0.135} - \left(\frac{1}{43.6}\right)^{0.135}}{0.1975}$$

= 2.0058

$$X_T = X_{mean} + Z_T \sigma_x$$

Where, X_{mean} = 28023, σ_x = 2792

So, $X_T = 28023 + 2792 * 2.0058$

$X_T = 33623$

Table A2.7: Normality test

Ranked data	X_T	$X_i - X_{mean}$	$X_{Ti} - X_{Tmean}$	$(X_i - X_{mean})^2$	$(X_{Ti} - X_{Tmean})^2$	$(X_i - X_{mean}) * (X_{Ti} - X_{Tmean})$
32674	33623	4651	5600	21636076	31364577	26050074
31693	32382	3670	4359	13466970	18996730	15994637
31510	31660	3487	3637	12161713	13229113	12684190
30373	31123	2350	3100	5521126	9611720	7284746
30355	30683	2332	2660	5436305	7073216	6200980
30319	30301	2296	2278	5271296	5188026	5229496
30040	29958	2017	1935	4070158	3745160	3904279
29993	29643	1970	1620	3880783	2625738	3192165
29974	29349	1951	1326	3805769	1757142	2585977
29882	29068	1859	1045	3454142	1092693	1942760
29644	28799	1621	776	2626427	601567	1256970
29459	28536	1436	513	2062210	263403	737016
29358	28278	1335	255	1781076	65274	340968
29303	28023	1280	0	1639221	0	0
29108	27767	1085	-255	1177856	65274	-277279
27815	27510	-208	-513	43258	263403	106744
26316	27247	-1707	-776	2913833	601567	1323959
26209	26978	-1814	-1045	3290200	1092693	1896095
25989	26697	-2034	-1326	4135567	1757142	2695696
25976	26403	-2047	-1620	4189498	2625738	3316704
25815	26088	-2208	-1935	4875573	3745160	4273149
25296	25745	-2727	-2278	7434569	5188026	6210535
24716	25363	-3307	-2660	10933714	7073216	8794118
24388	24923	-3635	-3100	13213393	9611720	11269580
24245	24386	-3778	-3637	14273597	13229113	13741435
23643	23664	-4380	-4359	19186444	18996730	19091352
22527	22423	-5496	-5600	30200567	31364577	30777070
Mean	Mean			Sum	Sum	Sum
28023	28023			202681343	191228719	190623413

$$r = \frac{\sum (X_i - \overline{X}) * (X_{Ti} - \overline{X_T})}{\left(\sum (X_i - \overline{X})^2 * \sum (X_{Ti} - \overline{X_T})^2 \right)^{0.5}}$$

r = 0.9682

$r_{critical}$ = 0.961

$r > r_{critical}$, null hypothesis accepted that data are from normal distribution. So, parametric tests can be applied for this data set.

Step 7: Check for shift in the mean

Parametric Test: t-Test for shift in the mean

H_0: Equal mean, no change in the mean

H_a: Shift in the mean

After infilling the missing value one can check the data set if the mean of the new data set is deviated from the mean of the previous data set.

Formula: $T_c = \dfrac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}}$

where, $S = \sqrt{\dfrac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}}$, where N = no. of sample, $\overline{y}$ = mean of the sample and s denotes standard deviation, subscript 1 and 2 describes sample data before and after infilling.

Here, $N_1 = 26$, $\overline{y_1} = 28427$ and $s_1 = 24802$

$N_2 = 27$, $\overline{y_2} = 28040$ and $s_2 = 24629$

$$S = \sqrt{\frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}} = \sqrt{\frac{(26 - 1) * 24802 + (27 - 1) * 24629}{53 - 2}}$$

$S = 24714$

$$T_c = \frac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}} = \frac{|28040 - 28427|}{24714\sqrt{\dfrac{1}{26} + \dfrac{1}{27}}} = 0.057$$

$T_{1-\alpha/2} = 1.96$

$T_c < T_{1-\alpha/2}$, so hypothesis is accepted and acceptation of hypothesis is considered as no detection of shift.

Non Parametric test: Mann-Whitney test for shift in the mean

Step 8: Trend test

Test for linear trend with neighboring station

H_0: no trend

H_a: Linear trend

$$T_c = \left|\frac{\sqrt{N - 2}}{r\sqrt{1 - r^2}}\right|$$

where, N = number of sample = 27

r = Cross correlation co-efficient = 0.9422

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right| = \left| \frac{\sqrt{27-2}}{0.9422\sqrt{1-0.9422^2}} \right| = 15.84$$

$T_{1-\alpha/2,\nu} = 1.96$

$T_c > T_{1-\alpha/2,\nu}$, So null hypothesis is rejected. Rejection of hypothesis considered as a detection of linear trend.

This test also can check trend between two months of one station by determining correlation between those months.

Mann-Kendall Test for trend

H_0: The series is random no trend series

H_a: The series has trend either upward or downward

This can check trend of yearly average data. This is more useful to check the low flow or high flow trend of long time period. The trend is decreasing if Mann-Kendall statistics, S is a negative large number otherwise the trend is increasing. No trend hypothesis indicates a stable condition.

Formula: $z_T = \frac{\tau - \mu}{\sigma}$, where, $\tau = \frac{N_1 - N_2}{N(N-1)/2}$, $\mu = 0$, and $\sigma = \sqrt{\frac{2(2N+5)}{9N(N-1)}}$

N_1 (1st is smaller than the 2nd) and N_2 are the upward downward pairs respectively

Here in this example trend of yearly average data has been checked. Calculation of Mann-Kendall statistics is shown in table 2.8.

Here in this example from Table 2.8 values for N_1 and N_2 can be obtained as follows:

$N_1 = 243$, $N_2 = 108$

$$\tau = \frac{N_1 - N_2}{N(N-1)/2} = \frac{243-108}{27(27-1)/2} = 0.3846$$

$$\sigma = \sqrt{\frac{2(2N+5)}{9N(N-1)}} = \sqrt{\frac{2*(2*27+5)}{9*27*(27-1)}} = 0.1111$$

$$z_T = \frac{\tau - \mu}{\sigma} = \frac{0.3846-0}{0.1111} = 3.462$$

$z_{1-\alpha/2} = 1.96$,

$z_T > z_{1-\alpha/2}$, So null hypothesis rejected and it indicates presence of trend in the series.

From Table 7 Mann-Kendall statistics S = 135, the no is positive and it considered as upward trend of the series. For removing trend in the data trend analysis is necessary.

Table A2.8: Calculation of Mann-Kendall Statistics

24245	22527	24388	24606	27971	24524	30009	33302	30193	25727	29643	28801	17970	31081	25913	27263	29276	31477	31215	25040	30425	33406	29351	31880	31693	23643	31510	Sum
	-1	1	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	20
		1	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23
			1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	20
				1	-1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	17
					-1	1	1	1	-1	1	1	-1	1	-1	-1	1	1	1	-1	1	1	1	1	1	-1	1	8
						1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	17
							1	1	-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	0
								-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-17
									-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	-2
										1	1	-1	1	1	1	1	1	1	-1	1	1	1	1	1	-1	1	11
											-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	0
												-1	1	-1	-1	1	1	1	-1	1	1	1	1	1	-1	1	5
													1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
														-1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	1	-1
															1	1	1	1	-1	1	1	1	1	1	-1	1	8
																1	1	1	-1	1	1	1	1	1	-1	1	7
																	1	1	-1	1	1	1	1	1	-1	1	6
																		-1	-1	-1	1	-1	1	1	-1	1	-1
																			-1	-1	1	-1	1	1	-1	1	0
																				1	1	1	1	1	-1	1	5
																					1	-1	1	1	-1	1	2
																						-1	-1	-1	-1	-1	-5
																							1	1	-1	1	2
																								-1	-1	-1	-3
																									-1	-1	-2
																										1	1
																										S	135

Step 9: Trend analysis

Moving Average method

Let t = -3, -2, -1, 0, 1, 2, 3 denote the time indices of the seven points. Let $a_0 + a_1t + a_2t_2 + a_3 t_3$ represents the cubic equation to be fitted to seven points by the least square method. Then

$$S = \sum_{t=-3}^{+3} \left(x_t - a_0 - a_1 t^1 - a_2 t^2 - a_3 t^3\right)^2 \qquad \text{(a)}$$

To minimize S, we have,

$$\frac{\partial s}{\partial a_j} = -2\sum_{t=-3}^{t=+3} \left(x_t - a_0 - a_1 t - a_2 t^2 - a_3 t^3\right) t^j = 0 \qquad \text{(b)}$$

Where j = 0, 1,2,3. It may be noted that $\sum t^i = 0$ when i is an odd power, e.g., when i =1,

$\sum t = -3 - 2 - 1 + 0 + 1 + 2 + 3 = 0.$

$\sum t^2 = 28$

From equation (a), we get

$$\frac{\partial s}{\partial a_0} = \sum \left(x_t - a_0 - a_2 t^2\right) \quad \text{for } j = 0$$

Therefore,

$$\sum t^i = 7a_0 + a_2 \sum t^2$$

Also,
$$= 7a_0 + 28a_2 \qquad \text{(c)}$$

$$\frac{\partial s}{\partial a_2} = \left(x_t - a_0 - a_2 t^2\right) t^2 = 0 \text{ for } j = 2$$

Therefore,

$$\sum t^2 x_t = a_0 \sum t^2 + a_2 \sum t^4$$
$$= 28a_0 + 196a_2 \qquad \text{(d)}$$

Therefore, by solving equation (c) and (d), we get

$$a_0 = \frac{1}{21}\left(7\sum x_t - \sum t^2 x_t\right)$$

$$= \frac{1}{21}\left(7x_{-3} + 7x_{-2} + 7x_{-1} + 7x_0 + 7x_1 + 7x_2 + 7x_3 - 9x_{-3} - 4x_{-2}\right)$$
$$- x_{-1} - 0 - x_1 - 4x_2 - 9x_3$$

$$= \frac{1}{21}\left(-2x_{-3} + 3x_{-2} + 6x_{-1} + 7x_0 + 6x_1 + 3x_2 - 2x_3\right)$$

$$= \frac{1}{21}\left(-2, +3, +6, +7, +6, +3, -2\right)$$

$$= \frac{1}{21}\left(-2, +3, +6, +7\right)$$

Thus the smooth value at the middle position of seven points can be obtained by coefficients derived above by the least square method and given in equation above.

Now, Smoothened value = $\frac{1}{21}$ (-2 × 24245 + 3 × 22527 + 24388 × 6 + 24606× 7 + 27971 × 6 + 24524 × 3 + 30009 × - 2) = 24716 m

Table A2.9: Smoothened value of Tested station

Year	Average monthly discharge	Smoothened Values (m)
1966 - 67	24245	24245
1967 - 68	22527	22527
1968 - 69	24388	24388
1969 - 70	24606	24716
1970 - 71	27971	25815
1972 - 73	24524	27815
1973 - 74	30009	30040
1974 - 75	33302	29993
1975 - 76	30193	30373
1976 - 77	25727	29974
1977 - 78	29643	26209
1978 - 79	28801	25976
1979 - 80	17970	25989
1980 - 81	31081	25296
1981 - 82	25913	26316
1982 - 83	27263	29108
1983 - 84	29276	29358
1984 - 85	31477	29882
1985 - 86	31215	29303
1986 - 87	25040	29644
1987 - 88	30425	29459
1988 - 89	33406	30355
1989 - 90	29351	32674
1990 - 91	31880	30319
1991 - 92	31693	31693
1992 - 93	23643	23643
1993-94	31510	31510

Removing trend by using slope

A linear trend in the mean is shown in Figure A2.2. It can be removed by the difference $y_t - \overline{y_t}$ as shown in Table A2.10.

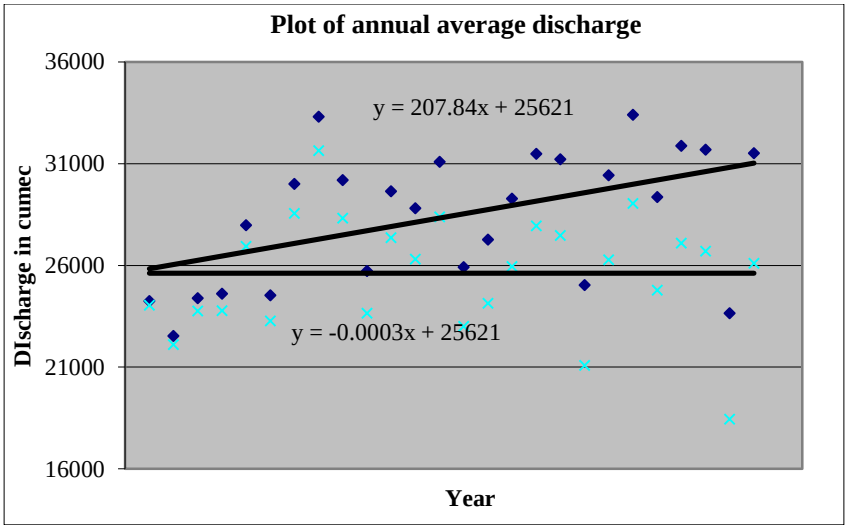


Figure A2.2 Linear trend in annual average discharge data

Table A2.10: Average discharge data with and without trend

Year	Yearly Average	Data without Trend
1966 - 67	24245	24037
1967 - 68	22527	22112
1968 - 69	24388	23764
1969 - 70	24606	23775
1970 - 71	27971	26932
1972 - 73	24524	23277
1973 - 74	30009	28555
1974 - 75	33302	31639
1975 - 76	30193	28322
1976 - 77	25727	23648
1977 - 78	29643	27356
1978 - 79	28801	26307
1980 - 81	31081	28379
1981 - 82	25913	23003
1982 - 83	27263	24145
1983 - 84	29276	25950
1984 - 85	31477	27943
1985 - 86	31215	27474
1986 - 87	25040	21091
1987 - 88	30425	26269
1988 - 89	33406	29042
1989 - 90	29351	24779
1990 - 91	31880	27099
1991 - 92	31693	26705
1992 - 93	23643	18447
1993 - 94	31510	26107

Step 10: Selection of distribution

Goodness of fit (Chi-square) test

Ho: The data are from specified distribution

Ha: The data are not from specified distribution and comes from other distributions

Formula: $\chi_c^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$

Null hypothesis is rejected if $\chi_c^2 > \chi_{1-\alpha,k-p-1}^2$

This test is a method to compare the observed and expected number of frequency (expected according to the distribution under test) that fall in the class interval.

Table A2.11: Goodness of fit test

Range of data	Observed Frequency, O_i	Z	Area under the normal curve	Area between range	Expected Frequency, E_i	O_i-E_i	$(O_i-E_i)^2/ E_i$
17500-20000	1	-2.79	0.4974	0.014	0.378	0.622	1.023
20000-22500	0	-2.13	0.4834	0.0542	1.4634	-1.4634	1.46
22500-25000	6	-1.47	0.4292	0.1411	3.8097	2.1903	1.26
25000-27500	4	-0.805	0.2881	0.2729	7.3683	-3.3683	1.54
27500-30000	5	-0.143	0.561	0.3625	9.7875	-4.7875	2.34
30000-32500	9	0.519	0.1985	0.1825	4.9275	4.0725	3.36
32500-35000	2	1.181	0.381	0.0861	2.3247	-0.3247	0.045
		1.84	0.4671				
Sum	27				30.0591	-3.0591	11.038

Mean = 28040

Standard deviation = 3776.7

Column1: Equal ranges of the data

Column2: Number of data falls in the range described in column1

Column3: Z = (X_i – mean)/Standard Deviation = (17500-28040)/3776.7 = -2.79

Column4: Area under the normal curve corresponding to Z value

Column5: Area under the range described in Column1= 0.4974-0.4834 = 0.014

Column6: E_i = Column5*Total observed frequency = 0.014*27 = 0.378

Column7: O_i-E_i = 1-0.378 = 0.622

Column8: $(O_i-E_i)^2/ E_i$ = $(0.622)^2/0.378$ = 1.023

$\chi_c^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$ = 11.038

$\chi_{1-\alpha,k-p-1}^2$ = 9.49 (for α = 0.05 and degrees of freedom, k-p-1 = 7-2-1 = 4)

Since $\chi_c^2 > \chi_{1-\alpha,k-p-1}^2$, null hypothesis rejected. That means the data set does not come from normal distribution.

This is one way of testing the goodness of fit or selection of probability distribution. Other methods are described in 'Technical Notes on Statistical Methods'.

Step 11: Frequency Analysis

After selection of type of probability distribution frequency analysis can be done according to the steps described in 'Technical Notes on Statistical Methods'.

Step 12: Reports and comments

Exercise A3: Water Level Data

The following table shows the monthly average water level data (in m) are recorded from 1983-2003 at Baruria Transit in Ganges. Steps for data quality checking for these data has been described here.

Step 1: Raw Data

Table A3.1: Monthly average water level data at Baruria Transit

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	2.7	4.18	5.3	7	7.29	7.94	6.64	4.5	2.94	2.33	1.9	1.88	54.64	4.55
1984-85	2.81	4.28	6.4	7.6	7.4	7.88	5.88	4.01	2.78	2.13	1.84	2.16	55.1	4.59
1985-86	2.9	3.59	5.9	7.2	7.24	7.3	6.71	4.77	3.48	2.69	2.2	2.23	56.17	4.68
1986-87	2.85	3.79	4.5	6.9	7.26	7.25	6.43	4.71	3.3	2.59	2.16	2.11	53.84	4.49
1987-88	3.14	3.74	5.3	7.3	8.46	8.2	6.49	4.63	3.35	2.61	2.28	2.58	58.1	4.84
1988-89	3.2	4.49	6.2	7.4	8.18	8.06	6.46	4.4	3.35	2.6	2.3	2.28	58.91	4.91
1989-90	2.81	4.25	6.3	7.3	7.33	7.49	6.64	4.54	3.18	2.44	2.15	2.26	56.69	4.72
1990-91	3.12	4.58	6.6	7.7	7.91	7.39	6.86	4.49	3.18	2.46	2.03	2.08	58.32	4.86
1991-92	3.01	4.69	6.4	7.4	7.68	7.9	6.12	4.24	3.07	2.49	2.08	2.08	57.13	4.76
1992-93	3.2	3.8	4.5	6.5	7.01	7.16	5.93	4.24	2.88	2.29	2.03	2.06	51.67	4.31
1993-94	2.49	4.46	6.1	7.3	7.75	7.79	6.69	4.51	3.16	2.43	2.17	2.21	57.09	4.76
1994-95	3.3	4.02	6.2	6.9	7.74	7.25	5.88	3.82	2.73	1.96	1.75	1.86	53.39	4.45
1995-96	2.39	4.66	6.5	8	7.9	7.83	6.57	4.59	3.22	2.42	2.05	2.22	58.39	4.87
1996-97	2.9	5	5.5	7.8	7.78	7.5	6.29	4.75	3.29	2.54	2.17	2.3	57.88	4.82
1997-98	3.07	3.59	5.5	7.2	7.42	7.3	5.66	3.92	3.39	2.72	2.07	1.96	53.85	4.49
1998-99	2.82	4.83	6.3	8.1	8.73	8.41	7.02	5.06	3.46	2.56	2.06	1.81	61.15	5.10
1999-00	2.19	3.91	5.9	7.7	7.82	8.05	6.96	4.94	3.29	2.37	1.92	1.91	56.96	4.75
2000-01	2.83	4.24	6.6	7.5	8.08	8.14	5.99	3.95	2.81	2.06	1.77	1.69	55.67	4.64
2001-02	2.23	3.46	5.6	6.8	7.57	7.68	6.64	4.2	2.77	1.98	1.74	1.71	52.36	4.36
2002-03	2.55	3.78	5.5	7.5	7.89	6.99	6.03	3.83	2.67	1.95	1.68	1.75	52.05	4.34
Sum	56.5	83.3	117	147	154	154	128	88.1	62.3	47.62	40.4	41.14		
Average	2.83	4.17	5.9	7.4	7.72	7.68	6.39	4.41	3.115	2.381	2.02	2.057		

Neighboring water level stations of Baruria Transit are as listed below:

1. Hardinge Bridge
2. Talbaria
3. Mohendrapur
4. Sengram
5. Sardah
6. Rampur Boalia
7. Goalunda Transit

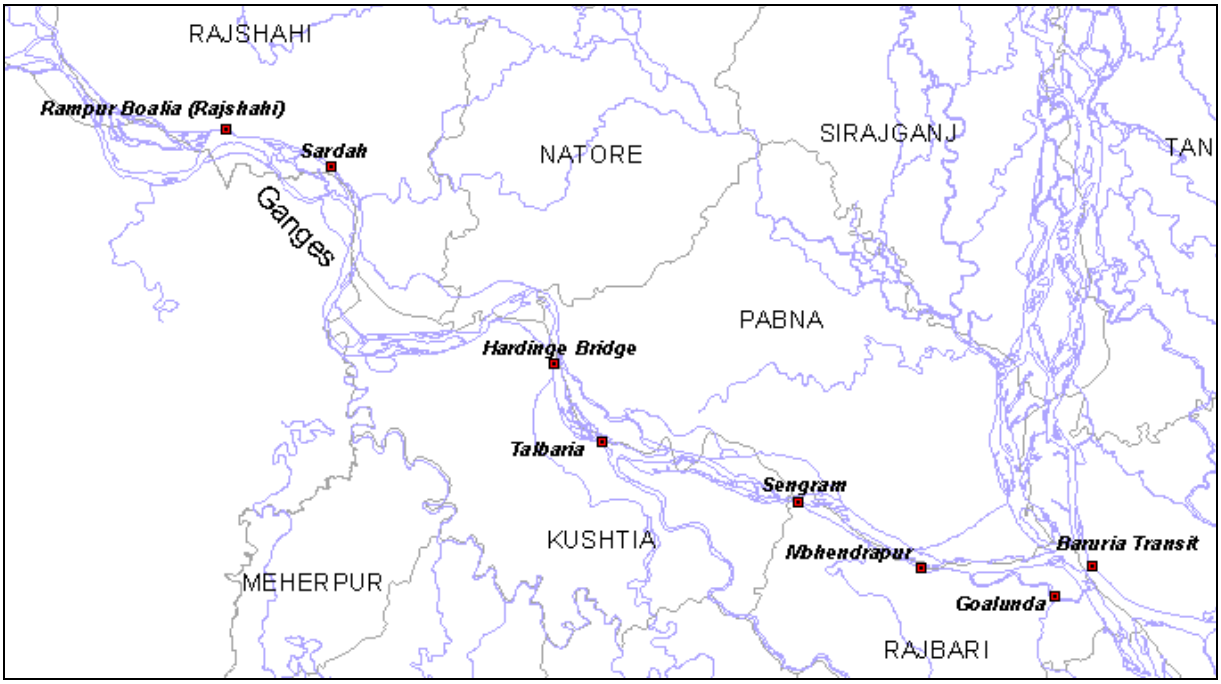


Figure A3.1: Location of upstream stations of Baruria Transit

Table A3.2: Average water level data of Goalunda Transit

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	2.95	4.46	5.72	7.31	7.72	8.44	7.13	4.87	3.29	2.67	2.26	2.27	59.09	4.92
1984-85	3.14	4.59	6.68	7.96	7.84	8.32	6.2	4.32	3.12	2.5	2.2	2.52	59.39	4.95
1985-86	3.31	4	6.22	7.6	7.77	7.78	7.13	5.15	3.82	3.1	2.58	2.57	61.03	5.09
1986-87	3.21	4.13	4.87	7.26	7.76	7.76	6.93	5.14	3.71	2.97	2.53	2.51	58.78	4.90
1987-88	3.54	4.16	5.76	7.71	8.9	8.61	6.87	5.02	3.68	2.92	2.6	2.96	62.73	5.23
1988-89	3.61	4.92	6.65	7.99	8.72	8.53	6.86	4.77	3.76	2.95	2.61	2.56	63.93	5.33
1989-90	3.08	4.48	6.65	7.78	7.76	7.94	7.11	5.07	3.62	2.83	2.54	2.67	61.53	5.13
1990-91	3.55	5.02	7.03	8.12	8.39	7.79	7.3	4.87	3.51	2.82	2.43	2.48	63.31	5.28
1991-92	3.46	5.14	6.82	7.89	8.1	8.34	6.55	4.62	3.48	2.91	2.49		59.8	5.44
1992-93	3.69	4.23	4.99	6.88	7.23	7.56	6.44	4.56	3.24	2.67	2.42	2.55	56.46	4.71
1993-94	2.99	4.93	6.67	7.88	8.31	8.36	7.23	5.06	3.71	3.03	2.77	2.84	63.78	5.32
1994-95	3.95	4.67	6.85	7.41	8.33	7.85	6.39	4.28	3.19	2.44	2.22	2.33	59.91	4.99
1995-96	2.86	5.16	7.1	8.57	8.44	8.36	7.05	5.06	3.73	2.97	2.61	2.73	64.64	5.39
1996-97	3.41	5.5	6.09	8.48	8.43	8.09	6.75	5.31	3.81	3.1	2.76	2.87	64.6	5.38
1997-98	3.72	4.31	6.25	7.87	7.96	7.92	6.26	4.52	4.04	3.44	2.88	2.88	62.05	5.17
1998-99	3.95	4.83	6.99	8.71	9.31	8.85	7.01	5.64	4.01	3.04	2.57	2.31	67.22	5.60
1999-00	2.71	4.37	6.39	8.2	8.36	8.74	7.58	5.53	3.87	2.95	2.49	2.43	63.62	5.30
2000-01	3.4	4.83	7.21	8.07	8.67	8.76	6.61	4.51	3.37				55.43	6.16
2001-02				7.79	8.34	8.31	7.29	4.91	3.48	2.63	2.41	2.33	47.49	5.28
2002-03	3.11	4.33	6	7.99	8.47	7.56	6.68	4.4	3.24	2.48	2.22	2.3	58.78	4.90
Sum	63.6	88.1	120.9	157.5	164.8	163.9	137.4	97.6	71.7	54.4	47.6	46.1		
Avg	3.35	4.63	6.37	7.87	8.24	8.19	6.869	4.88	3.58	2.86	2.50	2.56		

Table A3.3: Average water level data of Talbaria

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	4.48	5.71	6.53	9.8	11.8	13.13	11.66	8.79	7.14	6.43	5.92	5.54	96.93	8.08
1984-85	5.23	5.76	8.64	12.04	12.77	13.12	9.85	7.79	6.73	6.24	5.88	5.48	99.53	8.29
1985-86	5.23	5.39	6.71	10.3	12.49	12.51	12.25	9.61	7.7	6.75	6.08	5.9	100.9	8.41
1986-87	5.49	5.89	6.33	10.93	12.33	11.83	11.1	8.55	7.06	6.45	5.74	5.21	96.91	8.08
1987-88	5.3	5.55	6.59	10.12	12.91	13.11	10.39	8.57	6.86	6.02	5.74	5.56	96.72	8.06
1988-89	5.65	6.27	7.72	11.29	13.15	12.52	10.16	8.06	6.92	6.51	5.8	5.2	99.25	8.27
1989-90	5.13	5.66	8.29	10.91	11.97	12.04	10.6	8.01	6.64	5.67	4.7	4.81	94.43	7.87
1990-91	4.91	5.99	8.17	11.71	12.6	11.83	10.84	8.12	6.71	5.93	5.09	4.69	96.59	8.05
1991-92														
1992-93	4.32	4.74	5.42	8.66	11.32	11.51	9.41	7.66	6.25	5.19	4.08	3.68	82.24	6.85
1993-94	3.77	5.4	7.01	9.68	11.57	12.28	10.5	8.24	6.93	5.74	5.42	4.48	91.02	7.59
1994-95	4.41	4.97	7.54	10.79	12.73	12.04	9.71	7.68	6.62	5.47	4.98	4.67	91.61	7.63
1995-96	4.07	5.52	8.06	10.92	11.75	12.13	10.35	8.33	6.84	5.96	5.51	4.87	94.31	7.86
1996-97	4.48	6.18	7.22	11.45	12.77	12.72	10.26	8.08	6.52	5.41	4.73	4.17	93.99	7.83
1997-98	4.3	4.7	6.05	10.37	11.73	11.78	9.42	7.82	7.77	6.96	5.45	4.44	90.79	7.57
1998-99	4.3	4.7	7.93	11.1	12.31	13.41	11.95	10.48	8.59	6.34	5.6	4.82	101.5	8.46
1999-00	4.6	5.17	7.29	11.33	12.29	12.84	11.42	8.99	7.23	6.13	5.42	4.91	97.62	8.14
2000-01	4.8	5.76	8.82	11.28	11.84	12.2	9.47	7.51	6.23	5.31	5.42	4.58	93.22	7.77
2001-02	4.44	5.16	7.73	11.13	12.04	11.93	10.85	8.73	7.26	6.13	5.79	5.28	96.47	8.04
2002-03	5.15	6.04	7.51	10.28	12.09	11.41	9.91	7.74	6.58	5.64			82.35	7.94
Sum	90.06	104.56	139.56	204.09	232.46	234.34	200.10	158.76	132.58	114.28	97.35	88.29		
Average	4.74	5.50	7.35	10.74	12.23	12.33	10.53	8.36	6.98	6.01	5.41	4.91		

Table A3.4: Average water level data of Mohendrapur

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	3.8	5.33	6.67	8.8	9.81	10.71	9.19	6.4	4.53	3.8	3.3	3.14	75.5	6.29
1984-85	-	-	-	-	-	-	-	-	-	-	-	-		
1985-86	4.11	4.88	7.15	9.16	10.09	10.1	9.52	6.98	5.27	4.47	3.94	3.72	79.4	6.62
1986-87	3.72	4.85	5.61	8.75	9.49	9.28	8.13	5.94	4.59	3.95	3.34	3.1	70.8	5.90
1987-88	3.78	4.39	6.08	8.82	10.6	10.48	8.14	6.11	4.5	3.69	3.33	3.42	73.3	6.11
1988-89	3.9	5.24	7.05	9.3	10.56	10.12	8	5.86	4.39	3.68	3.15	2.94	74.2	6.18
1989-90	3.39	4.89	7.36	9.18	9.68	9.61	8.25	5.8	4.45	3.65	3.01	3.17	72.4	6.04
1990-91	3.8	5.26	7.47	9.53	10.05	9.32	8.52	5.75	4.27	3.49	2.86	2.78	73.1	6.09
1991-92	3.64	5.39	7.51	8.97	9.83	10.04	7.43	5.27	4.31	3.79	3.09	2.86	72.1	6.01
1992-93	3.95	4.65	5.42	7.81	8.89	9.02	7.45	5.48	4.31	3.42	2.78	2.59	65.8	5.48
1993-94	2.98	5.09	7.01	8.69	9.63	9.86	8.35	6.14	4.39	3.32	3.05	2.81	71.3	5.94
1994-95	3.79	4.54	6.89	8.37	9.76	9.07	7.13	4.91	3.83	2.98	2.61	2.51	66.4	5.53
1995-96	2.84	5.2	7.42	9.47	9.63	9.64	7.97	5.81	4.47	3.64	3.13	2.99	72.2	6.02
1996-97	3.35	5.46	6.19	9.21	9.51	9.16	7.28	5.56	3.95	2.99	2.5	2.4	67.6	5.63
1997-98	3.06	3.5	5.79	8.42	8.85	8.8	6.81	5.06	4.75	4.11	3.02	2.64	64.8	5.40
1998-99	3.5	4.73	7.28	9.88	10.83	10.48	-	6.25	4.5	3.51	2.92	2.35	66.2	6.02
1999-00	2.61	4.15	6.2	9.42	10.07	10.3	8.51	6.08	4.3	3.33	2.83	2.67	70.5	5.87
2000-01	3.4	4.78	7.56	9.33	9.92	10.23	7.71	5.54	4.15	-	-	-	62.6	

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
2001-02	-	-	-	8.89	10.09	10.09	8.87	6.1	4.63	3.67	3.42	3.24	59	
2002-03	3.91	5.12	6.84	9.41	10.23	8.93	7.72	5.37	4.15	3.18	2.91	2.9	70.7	5.89
Sum	63.53	87.5	122	171	187.5	185.2	145	110	83.7	64.7	55.2	52.2		
Average	3.53	4.86	6.75	9.02	9.87	9.75	8.05	5.81	4.41	3.59	3.07	2.90		

Table A3.5: Average water level data of Hardinge Bridge

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	5.54	6.6	7.41	11	13.12	14.18	12.73	9.72	8.04	7.31	6.8	6.41	108.9	9.07
1984-85	6.08	6.59	9.53	12.75	13.46	14.05	10.67	8.49	7.41	6.77	6.33	5.9		
1985-86	5.73	5.88	7.2	11.24	13.54	13.52	13.29	10.6	8.42	7.37	6.71	6.52	110	9.17
1986-87	6.1	6.46	6.94	12.33	13.36	12.84	11.51	8.94	7.75	7.15	6.43	5.95	105.8	8.81
1987-88	5.97	6.23	7.2	10.98	13.94	14.21	11.36	9.43	7.74	6.91	6.53	6.33	106.8	8.90
1988-89	6.32	6.83	8.2	12.14	14.16	13.31	10.86	8.58	7.38	6.95	6.21	5.63	106.6	8.88
1989-90	5.57	6.15	9.12	11.88	12.86	12.89	11.35	8.51	7.11	6.22	5.22	5.39	102.3	8.52
1990-91	5.45	6.52	8.76	12.63	13.66	12.9	11.79	8.89	7.35	6.57	5.64	5.23	105.4	8.78
1991-92	5.39	6.3	8.88	11.22	13.17	13.71	10.35	8.35	7.33	6.75	5.8	5.1	102.4	8.53
1992-93	4.9	5.25	5.9	9.4	12.33	12.7	10.46	8.52	6.99	5.92	4.82	4.37	91.56	7.63
1993-94	4.46	6	7.6	10.69	12.73	13.45	11.51	9.13	7.59	6.39	6.09	5.26	100.9	8.41
1994-95	5.01	5.27	7.88	11.55	13.51	12.83	10.4	8.37	7.32	6.24	5.73	5.4	99.51	8.29
1995-96	4.74	5.94	8.92	12.08	13.06	13.36	11.11	8.92	7.42	6.58	6.07	5.41	103.6	8.63
1996-97	4.83	6.19	7.29	11.93	13.43	13.53	11.24	9.21	7.75	6.74	6.13	5.55	103.8	8.65
1997-98	5.49	5.78	6.92	11.79	12.98	12.86	10.32	8.69	8.66	7.94	6.39	5.41	103.2	8.60
1998-99	5.68	6.76	8.37	12.53	14.01	13.84	11.81	10.3	8.32	7	6.27	5.41	110.3	9.19
1999-00	5.11	5.7	8.12	12.39	13.62	13.96	12.5	9.75	7.91	6.82	6.09	5.56	107.5	8.96
2000-01	5.44	6.47	9.97	12.84	13.44	13.83	10.71	8.55	7.47	6.42	5.74	5.31	106.2	8.85
2001-02	5.13	5.69	8.24	12.02	13.14	13.08	11.71	9.09	7.59	6.43	6.09	5.62	103.8	8.65
2002-03	5.49	6.21	7.87	11.42	13	12.23	11.61	8.41	7.11	6.12	-	-	89.47	
Sum	108	123	160.3	234.8	266.5	267.3	227.3	180	153	134.6	115	105.8		
Average	5.42	6.14	8.016	11.74	13.33	13.36	11.36	9.02	7.63	6.73	6.06	5.566		

Table A3.6: Average water level data of Sengram

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	3.98	4.89	6.22	8.85	10.3	11.41	10.03	7.27	5.82	4.85	3.96	3.61	81.19	6.77
1984-85	3.73	4.95	7.88	10.32	10.99	11.75	8.47	6.09	4.79	4.16	3.72	3.42	80.27	6.69
1985-86	3.85	4.9	7.2	9.58	10.9	10.37	10.34	7.53	5.96	5.28	4.81	5.01	85.73	7.14
1986-87	4.83	5.16	5.76	9.94	10.86	10.34	9.23	7.06	5.77	5.14	4.4	3.97	82.46	6.87
1987-88	4.45	5.17	6.48	9.86	11.86	11.66	9.16	7.38	5.71	4.76	4.32	4.28	85.09	7.09
1988-89	4.7	5.83	7.57	10.31	11.74	11.04	8.76	6.54	5.29	4.58	3.99	3.62	83.97	7.00
1989-90	3.25	4.38	6.9	10.02	10.33	10.46	9.43	6.99	5.71	4.89	4.03	4.14	80.53	6.71
1990-91	4.49	5.62	7.85	10.57	11.1	10.3	9.61	6.94	5.47	4.73	4.38	4.13	85.19	7.10
1991-92	-	-	-	-	-	-	-	-	-	-	-	-		
1992-93	4.27	5	5.81	8.2	10.22	10.63	8.6	6.66	5.33	4.45	3.61	3.36	76.14	6.35
1993-94	3.58	5.37	7.15	9.22	10.52	10.92	9.23	6.88	5.49	4.36	4.14	3.74	80.6	6.72
1994-95	4.4	5.1	7.51	9.59	11.34	10.73	8.81	6.87	5.86	4.95	4.57	4.33	84.06	7.01

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1995-96	4.04	6.07	8.31	10.96	11.5	11.52	10.19	7.91	6.51	5.27	4.62	4.21	91.11	7.59
1996-97	4.18	6.16	7.08	10.67	11.46	11.21	9.17	6.96	5.54	4.48	3.89	3.55	84.35	7.03
1997-98	3.9	4.33	6.23	9.39	10.07	10.22	7.96	6.3	6.07	5.49	4.19	3.53	77.68	6.47
1998-99	4.06	5.38	7.75	10.78	11.91	11.42	9.18	7.67	5.62	-	-	-	73.77	
1999-00	-	-	-	-	-	-	-	-	-	-	-	-		
2000-01	-	-	-	-	-	-	-	-	-	-	-	-		
2001-02	-	-	-	-	10.92	11.23	9.78	7	5.8	4.67	4.36	4.08	57.84	
2002-03	4.58	5.15	6.2	9.55	10.97	10.25	8.77	6.28	4.94	4.02	-	-	70.71	7.07
Sum	66.3	83.46	111.9	157.8	187	185.46	156.7	118.3	95.68	76.08	62.99	58.98		
Average	4.14	5.22	6.99	9.86	11.00	10.91	9.22	6.96	5.63	4.76	4.20	3.93		

Table A3.7: Average water level data of Sardah

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	7.48	8.5	9.11	13.49	16.09	17.52	15.72	11.95	9.91	8.4	8.14	7.6	133.91	11.16
1984-85	7.58	8.05	11.43	15.47	16.6	17.47	13.2	10.88	9.72	9.08	8.43	7.98	135.89	11.32
1985-86	7.93	8.04	8.95	13.89	16.84	16.59	16.45	13.3	11.14	9.03	8.54	8.13	138.83	11.57
1986-87	8.7	9.1	9.5	15.1	16.49	15.71	14.15	11.45	10.17	10.13	9.43	9.17	139.1	11.59
1987-88	8.18	8.42	9.2	13.28	16.81	17.56	13.97	12	10.21	9.49	8.66	8.09	135.87	11.32
1988-89	8.57	8.98	10.22	14.78	17.34	16.03	13.23	11.09	9.9	9.27	8.81	8.55	136.77	11.40
1989-90	8.05	8.55	11.55	14.4	15.68	15.82	14.12	10.87	9.59	9.42	8.6	8.06	134.71	11.23
1990-91	8.14	8.97	10.91	15.56	16.78	15.85	14.3	11.52	10.2	8.86	7.94	8.14	137.17	11.43
1991-92	8.36	8.82	11.45	13.83	15.99	17.03	13.06	11.19	10.34	9.57	8.66	8.27		
1992-93	8.01	8.1	8.54	12.01	15.18	15.66	13.29	11.64	10.24	9.81	8.8	8.21	129.49	10.79
1993-94	8.16	9.2	9.8	13.01	15.38	16.49	14.38	11.81	10.39	9.32	8.47	8.14	134.55	11.21
1994-95	7.84	7.92	10.22	14.41	16.94	16.22	13.41	11.49	10.51	9.25	8.98	8.12	135.31	11.28
1995-96	8.19	8.9	11.95	14.8	16.04	16.52	13.78	11.71	10.32	9.59	9.18	8.89	139.87	11.66
1996-97	7.83	8.56	10.06	14.59	16.7	16.88	14.01	11.92	10.56	9.5	9.18	8.59	138.38	11.53
1997-98	8.25	8.57	9.3	14.46	15.96	15.83	13.03	11.75	11.76	9.43	8.83	8.21	135.38	11.28
1998-99	8.95	9.8	10.9	15.37	17.33	17.25	14.81	13.07	11.46	11.02	9.61	8.67	148.24	12.35
1999-00	9.1	9.5	11.11	15.04	16.77	17.47	15.54	12.67	10.97	10.48	9.89	9.3		
2000-01	9	9.69	12.59	15.47	16.31	16.88	13.48	11.66	10.64	10.04	9.46	9.07		
2001-02	8.98	9.45	11.45	14.95	16.12	16.07	14.73	12.47	11.21	9.84	9.34	9.01	143.62	11.97
2002-03	9.49	9.96	11.05	14.14	15.81	15.34	13.55	11.46	10.31	10.19	9.96	9.6	140.86	11.74
Sum	166.8	177.1	209.3	288.1	327.2	330.2	282.2	235.9	209.6	191.7	178.9	169.8		
Average	8.34	8.854	10.46	14.4	16.36	16.51	14.11	11.8	10.48	9.586	8.946	8.49		

Table A3.8: Average water level data of Rampur Boalia

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1983-84	7.94	8.96	9.49	13.92	16.62	18.3	16.35	12.47	10.44	9.6	8.94	8.46	141.44	11.79
1984-85	8.18	8.67	12.23	16.37	17.66	18.5	13.94	11.51	10.25	9.61	9.09	8.65	144.63	12.05
1985-86	8.46	8.72	9.59	14.73	17.76	17.6	17.48	14	11.64	10.58	9.88	9.64	150.05	12.50
1986-87	9.18	9.49	9.97	16.04	17.46	16.7	-	-	-	-	9.28	9.11	97.24	
1987-88	9.2	9.43	10.22	14.14	17.83	18.6	14.74	12.66	11.03	10.24	9.85	9.66	147.58	12.30
1988-89	9.68	9.99	11.05	15.75	18.22	16.9	14.01	11.85	10.62	10.17	9.31	8.73	146.24	12.19

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Sum	Avg
1989-90	8.73	9.28	12.34	15.28	16.64	16.8	15.02	12.15	10.84	10.02	9.18	9.31	145.57	12.13
1990-91	9.3	10.09	11.86	16.36	17.55	16.7	15.15	12.32	10.82	10.06	9.34	9.12	148.69	12.39
1991-92	9.24	9.57	12.28	14.62	16.83	18	14.01	12.14	11.33	10.84	9.88	9.14	147.86	12.32
1992-93	8.85	9.11	9.57	12.91	16.15	16.6	13.93	12.4	11.02	10.14	9.25	8.97	138.88	11.57
1993-94	8.82	9.81	10.71	14.22	16.47	17.4	15.54	12.83	11.31	10.26	9.96	9.11	146.45	12.20
1994-95	8.83	8.86	11.09	15.22	17.81	17.1	14.19	12.2	11.25	10.27	9.8	9.51	146.08	12.17
1995-96	9.4	10.16	13.23	16.42	17.74	17.6	14.95	13.38	12.24	11.46	10.99	10.33	157.85	13.15
1996-97	9.58	10.24	11.47	15.61	17.51	17.9	15.41	13.46	12.27	11.31	10.73	10.16	155.61	12.97
1997-98	10.12	10.29	10.77	15.37	16.74	16.7	14.97	13.87	13.07	12.25	10.97	10.06	155.2	12.93
1998-99	10.32	11.19	12.21	16.56	18.37	18.4	16.04	14.36	12.39	11.42	10.79	10.16	162.16	13.51
1999-00	10.05	10.31	12.07	15.97	17.77	18.5	16.79	13.81	12.33	11.35	10.67	10.18	159.75	13.31
2000-01	10.04	10.82	13.78	16.27	17.38	18	14.84	12.89	11.77	10.91	10.35	9.92	157	13.08
2001-02	9.8	10.46	12.78	16.01	17.16	17.2	15.8	13.2	11.87	10.83	10.66	10.24	156.02	13.00
2002-03	10.14	10.7	11.93	15.15	16.87	16.3	14.31	11.92	10.69	9.72	9.5	9.36	146.63	12.22
Sum	185.9	196.2	228.6	306.9	346.5	349	287.5	243.4	217.2	201	198.4	189.8		
Average	9.29	9.81	11.43	15.35	17.33	17.47	15.13	12.81	11.43	10.58	9.92	9.49		

Step 2: Validation Test

Checking for randomness of data: Turning point test

Mean = 2(N-2)/3 and Standard Deviation = $[(16N-29)/90]^{0.50}$

Standard Value of T, t = (T-mean)/standard Deviation

Table A3.9: Determination of turning points

Average WL	Range	Remarks of Turning	Cumulative No. of Turning
4.55			
4.59	4.55<4.59<4.68	no	0
4.68	4.59<4.68>4.49	yes	1
4.49	4.68>4.49<4.84	yes	2
4.84	4.49<4.84<4.91	no	2
4.91	4.84<4.91>4.71	yes	3
4.72	4.91>4.72<4.86	yes	4
4.86	4.72<4.86>4.76	yes	5
4.76	4.86>4.76>4.31	no	5
4.31	4.76>4.31<4.76	yes	6
4.76	4.31<4.76>4.45	yes	7
4.45	4.76>4.45<4.87	yes	8
4.87	4.45<4.87>4.82	yes	9
4.82	4.87>4.82>4.49	no	9
4.49	4.82>4.49<5.10	yes	10
5.10	4.49<5.10>4.75	yes	11
4.75	5.10>4.75>4.64	no	11
4.64	4.75>4.64>4.36	no	11
4.36	4.64>4.36>4.34	no	11
4.34			T = 11

Therefore, T = 11 and N = 20

T follows normal distribution with, Mean = 12 and

Standard Deviation = $[(16N-29)/90]^{0.50}$ = 1.798

Therefore, t = 0.5561

For 5% level of significance $t_{1-\alpha/2}$ = 1.96

Therefore, $t < t_{1-\alpha/2}$. So, null hypothesis is accepted and the series considered as a random/no trend series.

Step 3: Consistency Checking

Double mass analysis

Table A3.10: Average and cumulative average of test station and neighboring station

Test Station		Base Station	
Average Cumulative WL of Tested Station	Cumulative avg. Tested Station	Average Cumulative WL of Surrounding Stations	Cumulative avg. Base Station
4.34	4.34	8.29	8.29
4.36	8.70	8.94	17.23
4.64	13.34	9.24	26.47
4.75	18.09	8.98	35.45
5.1	23.19	9.05	44.50
4.49	27.68	8.20	52.70
4.82	32.50	8.43	61.13
4.87	37.37	8.61	69.74
4.45	41.82	8.13	77.87
4.76	46.58	8.20	86.07
4.31	50.89	7.63	93.70
4.76	55.65	8.75	102.45
4.86	60.51	8.45	110.90
4.72	65.23	8.23	119.13
4.91	70.14	8.46	127.59
4.84	74.98	8.43	136.02
4.49	79.47	8.11	144.13
4.68	84.15	8.64	152.77
4.59	88.74	8.72	161.49
4.55	93.29	8.30	169.79

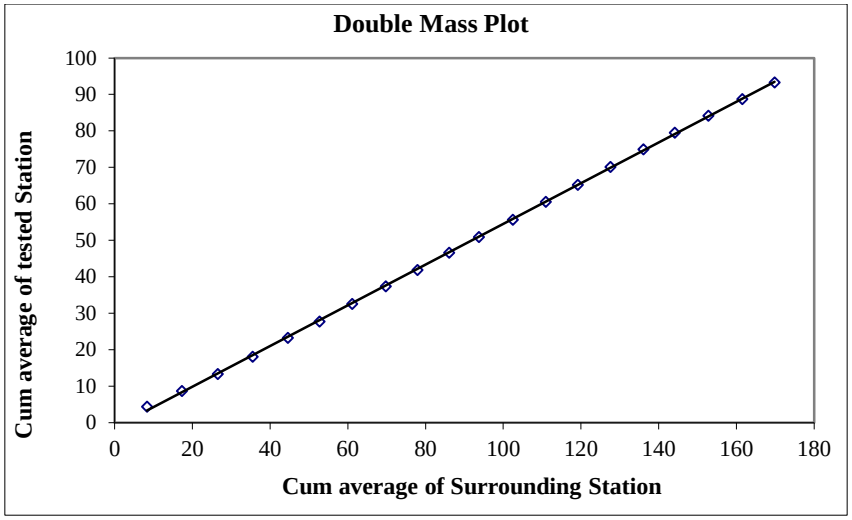


Figure 3.1 Double mass analysis of base station vs. tested station

From the double mass plot no trend change is observed here. Data series shows consistent pattern.

Step 4: Correlation test

Correlation between test station with neighboring station

Correlation between test station and Goalunda Transit

Table A3.11: Correlation with Goalunda Transit

X_i	Y_i	X_i-X_{mean}	Y_i-Y_{mean}	$(X_i-X_{mean})*(Y_i-Y_{mean})$
4.55	4.92	-0.11	-0.30	0.0330
4.59	4.95	-0.07	-0.27	0.0197
4.68	5.09	0.02	-0.14	-0.0023
4.49	4.90	-0.18	-0.32	0.0574
4.84	5.23	0.18	0.01	0.0010
4.91	5.33	0.25	0.11	0.0258
4.72	5.13	0.06	-0.09	-0.0057
4.86	5.28	0.20	0.05	0.0105
4.76	5.44	0.10	0.21	0.0207
4.31	4.71	-0.36	-0.52	0.1852
4.76	5.32	0.09	0.09	0.0087
4.45	4.99	-0.21	-0.23	0.0493
4.87	5.39	0.20	0.16	0.0332
4.82	5.38	0.16	0.16	0.0257
4.49	5.17	-0.18	-0.05	0.0091
5.10	5.60	0.43	0.38	0.1639
4.75	5.30	0.08	0.08	0.0066
4.64	6.16	-0.02	0.94	-0.0233
4.36	5.28	-0.30	0.05	-0.0164
4.34	4.90	-0.33	-0.32	0.1057
Mean	Mean			
4.66	5.22			Sum = 0.7079
Stdev	Stdev			
0.2132	0.3142			

Correlation between test station and Talbaria

Correlation between test station and Mohendrapur

Correlation between test station and Hardinge Bridge

Correlation between test station and Sengram

Correlation between test station and Sardah

Correlation between test station and Rampur Boalia

Correlation of Baruria Transit with surrounding Stations	
Station Name	**Correlation Co-efficient**
Goalunda Transit	
Mohendrapur	
Sengram	
Hardinge Bridge	
Talbaria	
Sardah	
Rampur Boalia	

Correlation between Baruria Transit and Goalunda Transit is best. So, infilling of missing data at Baruria Transit will be done with the help of Data of Goalunda Transit. But there is no missing data within the sample data used for this example. In all other stations some data is missing. List of missing data is given in Table A3.12:

Step 5: Filling missing data

Table A3.12: List of missing data

Name of Station	Period	
	Year	Month
Goalunda Transit	1991-92	March
	2000-01	Jan-Mar
	2001-02	April-June
Talbaria	1991-92	Whole Year
	2002-03	Feb-Mar
Mohendrapur	1984-85	Whole Year
	1998-99	Oct
	2000-01	Jan-Mar
	2001-02	April-June
Sengram	1991-92	Whole Year
	1998-99	Jan-Mar
	1999-00	Whole Year
	2000-01	Whole Year
	2001-01	Apr-June
Hardinge Bridge	2002-03	Feb-Mar
Rampur Boalia	1986-87	Oct-Jan

For infilling the missing data first step is to find out the correlation of that station with all other station. The station which gives best correlation with that station will be used for infilling the missing data. In other ways all neighboring stations also can be used for infilling. Missing data of Goalunda Transit is

filled here with both of the methods. Correlation between Goalunda Transit with other stations is not checked here but in case of Baruria Transit it has given the best correlation. So Baruria transit can be used for this station. But for infilling the other stations data correlation between target station with other stations should be checked.

In Table A3.13 shows the estimation of missing data by using the data of best correlated station Baruria Transit:

Table A3.13: Estimating missing data (Baruria)

Year	Jan	Feb	Mar	Apr	May	Jun
1991-92			2.59			
2000-01	2.48	2.20	2.10			
2001-02				2.64	3.85	6.03

For example in 1991-92:

Long term average for March at Goalunda Transit = 2.56 m

Long term average for March at Baruria Transit = 2.057 m

Data of March, 1991-92 at Baruria Transit = 2.08 m

Estimated missing data = 2.56*(2.08/2.057) = 2.59 m

Other year's data can also be estimated in this way.

Data also estimated by using all the neighboring station data and shown in Table A3.14:

Table A3.14: Estimating missing data (Neighboring)

Year	Jan	Feb	Mar	Apr	May	Jun
1991-92						
2000-01						
2001-02						

For example in January 2000-01:

Step 6: Normality Test

H_0: The data are normal

H_1: The data are not normal

If $r < r_{critical}$ – reject null hypothesis

Determination of X_T:

Observed data are arranged in descending order

Calculate, p = (i – a)/(n+1-2a), a = plotting position for normal distribution and n = sample size

Then, return period, T = 1/p, p and T determined for each sample and then determine Z_T for each sample by using following formula

Standard Normal Variate, $Z_T = \dfrac{\left(1-\frac{1}{T}\right)^{0.135} - \left(\frac{1}{T}\right)^{0.135}}{0.1975}$

$$= \frac{\left(1-\frac{1}{5}\right)^{0.135} - \left(\frac{1}{5}\right)^{0.135}}{0.1975}$$

$$= 0.8385$$

$$X_T = X_{mean} + Z_T \sigma_x$$

Where, X_{mean} = 4.66, σ_x = 0.2132

So, $X_T = 4.66 + 0.2132 * 1.876$

$X_T = 5.06$

Table A3.15: Estimating probability plot correlation coefficient

Ranked data	X_T	$X_i - X_{mean}$	$X_{Ti} - X_{Tmean}$	$(X_i - X_{mean})^2$	$(X_{Ti} - X_{Tmean})^2$	$(X_i - X_{mean}) * (X_{Ti} - X_{Tmean})$
5.10	5.0639579	0.432	0.400	0.18648	0.15997	0.17272
4.91	4.9634574	0.245	0.299	0.06011	0.08967	0.07342
4.87	4.9041712	0.202	0.240	0.04074	0.05768	0.04847
4.86	4.8593593	0.196	0.195	0.03842	0.03817	0.03829
4.84	4.8219785	0.178	0.158	0.03157	0.02496	0.02807
4.82	4.7890299	0.159	0.125	0.02539	0.01563	0.01992
4.76	4.7589156	0.097	0.095	0.00938	0.00901	0.00919
4.76	4.7306516	0.094	0.067	0.00874	0.00444	0.00623
4.75	4.7035586	0.083	0.040	0.00683	0.00156	0.00327
4.72	4.6771165	0.060	0.013	0.00362	0.00017	0.00079
4.68	4.6508835	0.017	-0.013	0.00028	0.00017	-0.00022
4.64	4.6244414	-0.025	-0.040	0.00062	0.00156	0.00098
4.59	4.5973484	-0.072	-0.067	0.00523	0.00444	0.00482
4.55	4.5690844	-0.111	-0.095	0.01225	0.00901	0.01050
4.49	4.5389701	-0.177	-0.125	0.03115	0.01563	0.02207
4.49	4.5060215	-0.177	-0.158	0.03145	0.02496	0.02801
4.45	4.4686407	-0.215	-0.195	0.04615	0.03817	0.04197
4.36	4.4238288	-0.301	-0.240	0.09040	0.05768	0.07221
4.34	4.3645426	-0.326	-0.299	0.10660	0.08967	0.09777
4.31	4.2640421	-0.358	-0.400	0.12828	0.15997	0.143251594
Mean	Mean			Sum	Sum	Sum
4.66	4.66			0.86368	0.80253	0.82174

$$r = \frac{\sum (X_i - \overline{X}) * (X_{Ti} - \overline{X_T})}{\left(\sum (X_i - \overline{X})^2 * \sum (X_{Ti} - \overline{X_T})^2\right)^{0.5}}$$

r = 0.9870

$r_{critical}$ = 0.951

$r > r_{critical}$, null hypothesis accepted that data are from normal distribution. So, parametric tests can be applied for this data set.

Step 7: Test for shift in the mean and variance

t-Test

t-Test for mean and shift in the mean. Table A3.16 shows the value of average water level before infilling and after infilling with two method, their average and standard deviation.

$$T_c = \frac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

where, $S = \sqrt{\frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}}$

Table A3.16: Average record before and after infilling

Month	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Avg	Stdev
Avg before infilling	3.35	4.63	6.37	7.87	8.24	8.19	6.87	4.88	3.58	2.86	2.50	2.56	5.16	2.249
Avg after infilling with 1st method	3.31	4.60	6.35	7.87	8.24	8.19	6.87	4.88	3.58	2.84	2.49	2.54	5.1478	2.257
Avg after infilling with 2nd method	3.34	4.63	6.38	7.87	8.24	8.19	6.87	4.88	3.58	2.86	2.50	2.55	5.1587	2.252

t-Test between original sample data with data set infilled with first method:

$$S = \sqrt{\frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}} = \sqrt{\frac{(12 - 1) * 2.249 + (12 - 1) * 2.257}{24 - 2}} = 2.253$$

$$T_c = \frac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{\left|\overline{5.1478} - \overline{5.16_1}\right|}{2.253\sqrt{\frac{1}{12} + \frac{1}{12}}} = 0.0134$$

$T_{1-\alpha/2} = 1.96$

$T_c < T_{1-\alpha/2}$, so hypothesis is accepted and acceptation of hypothesis is considered as no detection of shift

t-Test between original sample data with data set infilled with second method:

$$S = \sqrt{\frac{(N_1 - 1)s_1 + (N_2 - 1)s_2}{N - 2}} = \sqrt{\frac{(12 - 1) * 2.249 + (12 - 1) * 2.252}{24 - 2}} = 2.25$$

$$T_c = \frac{\left|\overline{y_2} - \overline{y_1}\right|}{S\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{\left|\overline{5.1587} - \overline{5.16}\right|}{2.253\sqrt{\frac{1}{12} + \frac{1}{12}}} = 0.0013$$

$T_{1-\alpha/2}$ = 1.96

T_c < $T_{1-\alpha/2}$, so hypothesis is accepted and acceptation of hypothesis is considered as no detection of shift

Value of T_c is found lower in case of second method.

Mann-Whitney Test

$$u_c = \frac{\sum_{t=1}^{N_1} R(y_t) - N_1(N_1 + N_2 + 1)/2}{\left[N_1 N_2 (N_1 + N_2 + 1)/12\right]^{1/2}}$$

Step 8: Determination and Testing of Trend

Parametric Test: Estimation of linear trend:

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right|$$

where, N = number of sample

r = Cross correlation co-efficient

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right| = \left| \frac{\sqrt{20-2}}{0.99996\sqrt{1-0.99996^2}} \right| = 474$$

$T_{1-\alpha/2,\nu}$ = 1.96

T_c > $T_{1-\alpha/2,\nu}$, So hypothesis is rejected. Rejection of hypothesis considered as a detection of linear trend between tested station Baruria Transit with neighboring station Goalunda Transit. In this way presence or absence of trend with other stations can be determined.

Nonparametric Test: Mann-Kendall test

$$u_c = \frac{S + m}{\sqrt{V(S)}}$$

where, S = Mann-Kendall statistic, m = 1 when S < 0 and m = -1 when S > 0

Yearly average data has been taken for Mann-Kendall test checking trend in the data series.

Calculation for Mann-Kendall statistic is shown in table A3.17

Table A3.17: Mann-Kendall Statistics

Early average Water Level data																			
4.92	4.95	5.09	4.90	5.23	5.33	5.13	5.28	5.19	4.71	5.32	4.99	5.39	5.38	5.17	5.60	5.30	5.26	5.15	4.90	Sum
	1	1	-1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	13
		1	-1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	12
			-1	1	1	1	1	1	-1	1	-1	1	1	1	1	1	1	1	-1	9
				1	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	0	13
					1	-1	1	-1	-1	1	-1	1	1	-1	1	1	1	-1	-1	1

Early average Water Level data																				
4.92	4.95	5.09	4.90	5.23	5.33	5.13	5.28	5.19	4.71	5.32	4.99	5.39	5.38	5.17	5.60	5.30	5.26	5.15	4.90	Sum
						-1	-1	-1	-1	-1	-1	1	1	-1	1	-1	-1	-1	-1	-8
							1	1	-1	1	-1	1	1	1	1	1	1	1	-1	7
								-1	-1	1	-1	1	1	-1	1	1	-1	-1	-1	-2
									-1	1	-1	1	1	-1	1	1	1	-1	-1	1
										1	1	1	1	1	1	1	1	1	1	10
											-1	1	1	-1	1	-1	-1	-1	-1	-3
												1	1	1	1	1	1	1	-1	6
													-1	-1	1	-1	-1	-1	-1	-5
														-1	1	-1	-1	-1	-1	-4
															1	1	1	-1	-1	1
																-1	-1	-1	-1	-4
																	-1	-1	-1	-3
																		-1	-1	-2
																			-1	-1
																				41

Here, S > 0, m = -1

$$V(S) = \frac{1}{18}\left[N(N-1)(2N+5) - \sum_{i=1}^{n} e_i (e_i - 1)(2e_i + 5) \right]$$

Where, e_i is the number of data in tied group and n is the number of tied group. In this data set there is one tied group and $e_i = 2$.

N = 20

$V(S)$ = 950

$u_c = (41-1)/\sqrt{950} = 1.298$

$u_{1-\alpha/2} = 1.96$

$u_c < u_{1-\alpha/2}$, so the hypothesis is rejected and rejection of hypothesis means no trend found. So, trend analysis is not necessary.

Step 9: Trend Analysis

Not necessary.

Step 10: Selection of distribution

Step 11: Frequency analysis

Step 12: Reports and comments

Appendix B
Tables

Table B1.1: Critical r* values for the probability plot correlation coefficient test of normality (Helsel and Hirch, 2002)

[reject H0: data are normal when PPCC r < r*]

Sample Size	α-level					
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.1**	**0.25**
3	0.867	0.869	0.872	0.879	0.891	0.924
4	0.813	0.824	0.846	0.868	0.894	0.931
5	0.807	0.826	0.856	0.88	0.903	0.934
6	0.82	0.838	0.866	0.888	0.91	0.939
7	0.828	0.85	0.877	0.898	0.918	0.944
8	0.84	0.861	0.887	0.906	0.924	0.948
9	0.854	0.871	0.894	0.912	0.93	0.952
10	0.862	0.879	0.901	0.918	0.934	0.954
11	0.87	0.886	0.907	0.923	0.938	0.957
12	0.876	0.892	0.912	0.928	0.942	0.96
13	0.885	0.899	0.918	0.932	0.945	0.962
14	0.89	0.905	0.923	0.935	0.948	0.964
15	0.896	0.91	0.927	0.939	0.951	0.965
16	0.899	0.913	0.929	0.941	0.953	0.967
17	0.905	0.917	0.932	0.944	0.954	0.968
18	0.908	0.92	0.935	0.946	0.957	0.97
19	0.914	0.924	0.938	0.949	0.958	0.971
20	0.916	0.926	0.94	0.951	0.96	0.972
21	0.918	0.93	0.943	0.952	0.961	0.973
22	0.932	0.933	0.945	0.954	0.963	0.974
23	0.925	0.935	0.947	0.956	0.964	0.975
24	0.927	0.937	0.949	0.957	0.965	0.976
25	0.929	0.939	0.951	0.959	0.966	0.976
26	0.932	0.941	0.952	0.96	0.967	0.977
27	0.934	0.943	0.953	0.961	0.968	0.978
28	0.936	0.944	0.955	0.962	0.969	0.978
29	0.939	0.946	0.956	0.963	0.97	0.979
30	0.939	0.947	0.957	0.964	0.971	0.979
31	0.942	0.95	0.958	0.965	0.972	0.98
32	0.943	0.95	0.959	0.966	0.972	0.98
33	0.944	0.951	0.961	0.967	0.973	0.981
34	0.946	0.953	0.962	0.968	0.974	0.981
35	0.947	0.954	0.962	0.969	0.974	0.982
36	0.948	0.955	0.963	0.969	0.975	0.982
37	0.95	0.956	0.964	0.97	0.976	0.983
38	0.951	0.957	0.965	0.971	0.976	0.983
39	0.951	0.958	0.966	0.971	0.977	0.983
40	0.953	0.959	0.966	0.972	0.977	0.984
41	0.953	0.96	0.967	0.973	0.977	0.984
42	0.954	0.961	0.968	0.973	0.978	0.984
43	0.956	0.961	0.968	0.974	0.978	0.984

Sample Size	**α-level**					
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.1**	**0.25**
44	0.957	0.962	0.969	0.974	0.979	0.985
45	0.957	0.963	0.969	0.974	0.979	0.985
46	0.958	0.963	0.97	0.975	0.98	0.985
47	0.959	0.965	0.971	0.976	0.98	0.986
48	0.959	0.965	0.971	0.976	0.98	0.986
49	0.961	0.966	0.972	0.976	0.981	0.986
50	0.961	0.966	0.972	0.977	0.981	0.986
55	0.965	0.969	0.974	0.979	0.982	0.987
60	0.967	0.971	0.976	0.98	0.984	0.988
65	0.969	0.973	0.978	0.981	0.985	0.989
70	0.971	0.975	0.979	0.983	0.986	0.99
75	0.973	0.976	0.981	0.984	0.987	0.99
80	0.975	0.978	0.982	0.985	0.987	0.991
85	0.976	0.979	0.983	0.985	0.988	0.991
90	0.977	0.98	0.984	0.986	0.988	0.992
95	0.979	0.981	0.984	0.987	0.989	0.992
100	0.979	0.982	0.985	0.987	0.989	0.992

Table B1.2: Areas under the Standard Normal Density from 0 to z (Haan, 1979)

z	0	1	2	3	4	5	6	7	8	9
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2703	0.2734	0.2764	0.2793	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3906	0.3925	0.3943	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4263	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4648	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4606
1.9	0.4713	0.4719	0.4736	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4915	0.492	0.4922	0.4924	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4986	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499
3.1	0.499	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Appendix C
Data Quality Report

Data quality implementation for Surface Water Level (Non Tidal) of BWDB

C.1 Product Specification

Dataset name	Water Level (Non Tidal) of BWDB
Product description	This data layer contains daily non-tidal water level data collected at 281 non-tidal water level stations of Bangladesh Water Development Board (BWDB).
Types of features (name, definition)	Water Level (Non Tidal) of BWDB
International boundary	Definition: International boundary of Bangladesh Source: SoB topographic map
Coastline	Definition: Coast line of Bangladesh Source: LANDSAT image
Other administrative boundary	Definition: Administrative boundary of districts, thana, union, pourshava Source: DLRS thana map

Data dictionary for features	
Name	Data
Water Level (Non Tidal) of BWDB	Time series surface water level data of BWDB

C.2 Data Quality Specification

Data quality sub elements	Product specification
Completeness	The data layer covers all non-tidal water level stations of BWDB.
Logical consistency	Requirement Unit: Units of measurement should be English measurement system Assurance: All attribute names and definition are verified
Quality	Data is available for all non-tidal water level stations (281) of BWDB (from 1-4-1910 to 31-3-2010). But extent of data availability varies from station to station. Some data are missing at some stations. The data those are to be obviously errors, have been excluded by NWRD.
Positional accuracy (Absolute)	The positional accuracy should be same of that source map.

C.3 Quality management in production

Assessment of source

Documentation

Data title	Water Level (Non Tidal) of BWDB
Organization	BWDB (Bangladesh Water Development Board)
Project under which the data was collected from source	National Water Resources Database (NWRD)

Name of individual who authorizes the data (if any)	System Analyst, Surface water processing division, BWDB
Unit of the data	Meter PWD
Year of publication	Regular update
Source of the content (in detail)	Bangladesh Water Development Board collected and stored this daily water level data from different non-tidal water level stations throughout the country for a long period of time.
Readability of features	Good
Readability of texts	Good
Final grading according to table 1	II

Grading

Grade	**Descriptions**	**Grade Name**
Grade I	Data information is quite good.	Very Good
Grade II	Everything is ok, but need some check.	Good
Grade III	Many errors but can be used giving little effort.	Fair
Grade IV	Quality is very bad which can't be usable as source data.	Poor

Data Capture	
Data digitisation	
Data preparation	Collecting this data layer from BWDB, NWRD has checked the obvious errors and stored in an ordered format.
Data editing	NWRD has checked the obvious errors and stored in an ordered format