WORKSHOP

ON

MODELLING OF HYDROLOGIC SYSTEMS

4-8 September, 2000

Data Processing using HYMOS

by

B. Venkatesh

Organised by

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

DATA PROCESSING USING HYMOS

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

1.0 Introduction

The concept of the hydrologic cycle forms the basis for the hydrologist's understanding of the source of water at or under the Earth's surface and its consequent movement by various pathways back to the principal storage in the ocean. Two of the greatest problems for the hydrologist are quantifying the amount of water in the different phases in the cycle and evaluating the rate of transfer of water from one phase to another within the cycle. Thus measurement within the cycle is a major function.

Hydrological data hold the key to the orderly and efficient development and control of water resources. Advances in scientific Hydrology and Engineering Hydrology are dependent on good, reliable and continuous recording, processing and analysis of hydrologic variables. The present hydrologic components are dealt with separately by each individual organisation. There is no single body within India that is responsible for all hydrological service. Though data collected by any one Organisation can be utilised by others. Unfortunately in the present system the same data is collected by the different agencies independently. Even the collected data by these agencies suffer from the following two major drawbacks. 1). The data collected by any single Organisation independently is observed to be inadequate to define the entire water system, 2). The data, which is not really required or data pertaining to unrelated elements are frequently observed and collected.

Once the data collect, the next important and necessary step is their proper storage. In the conventional ways, the data are mostly kept in manuscript form in registers or files. Usually no processing is done before the data is stored in the registers/files. With the advancement of technology, computer memories, magnetic tapes, microfilms etc., these conventional storage media are being replaced. In the modern procedures, the data are stored only after a particular level of processing.

The term 'Data Processing' covers data preparation, data entry and transfer to the database, data validation, data correction, filling-in-of missing data, data compilation and analysis, data retrieval and data dissemination/publication of yearbooks. Basically, the aim of data processing is to manipulate the raw data and to put it in a proper form and extract the required information from it. The processing activities are divided into primary and secondary processes. The primary processing consists of the data validation i.e., removing the random and systematic errors if any from the raw data and putting it in a proper format. The secondary processing includes deriving some information from the data such as fixing of rating curves using gauge-discharge data, Flood forecasting using precipitation data etc.

2.0 Concept of Data Processing

The past few decades have witnessed tremendous growth of utilisation and management of water resources. The successful and efficient execution of these studies/decisions requires a vast amount of data of interrelated phenomenon. The accuracy of results obtained depends upon the quality and quantity of information used. This information is provided by an information system, which cover, the data operation of the following kind: 1). Data Collection and 2) Data processing. Generally a large amount of data is involved, making a proper tuning of all phases in the collection and processing of data a necessity.

Data processing may be defined as any systematic procedure through which basic information is transposed into more accessible or more directly usable forms. It covers all activities after transmission upto and including a data dissemination. It comprises editing and organising functions (detecting and correcting errors, classifying and indexing records, sorting and collating data prior to use, etc.), carrying out necessary calculations and transcribing both data and results in prescribed format to a suitable storage and processing medium.

The data processing consists of three basic steps: input, processing, and output.

Input or data entry :

In this step, the initial data are prepared in some convenient form for processing. The form will depend on the processing machine.

Processing

In this step, the input data are changed and usually combined with other information to produce data in amore useful form. This step usually involves a sequence of certain basic processing operations.

Output

The results of the processing steps are collected. The particular form of the output depends on the use of the data.

Frequently, three more steps are added to the basic data processing cycle to obtain the expanded data processing cycle. They are Organisation, Distribution and storage.

Organisation

It is the process of collecting the original data. and original recording of data is called a source document

Distribution

This refers to the distribution of the output data. Recordings of the output data are often called report documents. Sometimes, the report documents may become the source document for future data processing

Storage

This step is crucial in many data processing procedures. Data processing results are frequently placed in storage to be used as input data for further processing at a later data. A unified set of data in storage is called a file. A collection related files are called a database.

3.0 Data Processing Operations

A general data processing procedure normally consists of a number of basic processing operations performed in some order, they are

- Receipt of data
- Data entry to computer/ register
- Data validation - primary and secondary
- Data completion and correction
- Data compilation
- Data analysis
- Data reporting
- Data transfer

Receipt of Data

A record of receipt and data of receipt for each station record is maintained for each month of the year in suitably formatted registers. Receipt will be recorded on the day of delivery. Such registers will be maintained in each office through the system. These registers have two purposes:

- To provide a means of tracking misplaced data
- To identify the cause of delay beyond a target data whether late from the field or delay at a processing office and hence to follow up with corrective measures.

Data collected in the field are delivered first to the organisation that is maintaining the instrument in a variety of media, as hand-written forms and notebooks, charts or digital data files on magnetic media.

Data Entry

The bulk of raw data are in the form of time series of hydrological and hydro-meteorological, water quality and quantity data. All such data are entered to computer at the lowest level in the data processing system. This has the advantage that the supervisory field staff share neighbouring offices with data processing staff and can easily be made aware of observer's mistakes or instrumental errors, and feedback given to the field personnel

Data Validation

Data validation is the means by which data are checked to ensure that the final figure stored is the best possible representation of the true value of the variable at the measurement site at a given time or in a given interval of time. Validation recognises that values observed or measured in the field are subjected to error and that undetected errors may also arise in data entry, in computation and from the mistaken 'correction' of good data.

Validation is carried out for three reasons:

- to correct errors in the recorded data where this is possible
- to assess the reliability of a record even where it is not possible to correct errors
- to identify the source of errors and thus to ensure that such errors are not repeated in future

Level of validation

It is preferable to carry out the data validation as soon as the data is observed and as near to the observation station as possible. This ensures that information, which may be essential to support the inferences of data validation, is fresh in minds of the field staff and supervisors and that interaction between field and processing staff is possible. However, to provide full validation close to observation sites is impractical in terms of computing equipment and staffing and a compromise must be reached which recognises both the wide geographical spread of observation stations and the staff and equipment available. The sequence of validation steps has therefore divided so that those steps, which primarily require interaction with the observation station, are carried out in the close proximity, whereas the more complex comparisons are carried out at higher level of validation.

Thus, data validation to be carried out is grouped into three major categories:

- Primary data Validation
- Secondary data validation
- Hydrological data validation.

Primary Data Validation

Primary validation is primarily involved with comparisons within a single data series and is concerned with making comparisons between observations and statistical range of a variable or with the expected hydrological behaviour of a hydrological phenomenon. However, information from a few nearby stations within a limited area may also sometimes be available and these may be used while carrying our primary validation for example with respect to daily rainfall data.

Primary data validation highlights those data, which are not within the expected range or are not hydrologically consistent. These data are then revisited in the data sheets or analogue records to see if there way any error while making computations in the field or during keying-in the data.

Secondary Data Validation

Secondary validation consists of comparisons between the same variable at two or more stations and is essentially to test the data against the expected spatial behaviour of the system. Secondary validation is based on the spatial information available from a number of neighbouring observation stations within a comparatively large area. The assumption, while carrying out such comparison, is that the variable under consideration has adequate spatial correlation within the distances under consideration. Such correlation must be confirmed in advance on the basis of historical records and the experience thus gained in the form of various types of statistic is utilised while validating the data. Qualitative evaluation of this relationship is not very difficult to make. For certain hydrological variables like water levels and discharges, which bear a very high degree of dependence or correlation between adjoining stations, the interrelationship can be established with a comparatively higher level of confidence. However, for some variables which lack serial correlation and show great spatial variability, it is difficult to ascertain the behaviour with the desired level of confidence, In such circumstances, it becomes very difficult, if not impossible, to detect errors.

While validating the data on the basis of a group of surrounding stations, the strategy must always be to rely on certain key stations known to be good quality. If all the observation stations are given the status of being equally eligible then data validation will become comparatively more difficult. This is not done merely to make the data validation faster but on the understanding based on field experience that the quality of data received from certain stations will normally be expected to be better than others. This may be due to physical conditions at the station, quality of instruments or reliability of staff etc. It must always be remembered that these key or reliable stations also can report incorrect data and they donot enjoy the status of being absolutely perfect.

Hydrological Validation

Hydrological validation consists of comparing one record with one or more others, for interrelated variables at the same or adjacent stations and is designed to show up inconsistencies between the time series or their derived statistics. Hydrological validation may be applied to a measured variable but is more often applied to drive variables. This is usually done through regression analysis or simulation modelling.

If a record has been subjected to thorough field checking and primary and secondary validation, soon after the record has been obtained, then hydrological validation should reveal no more than is already known. However, for historical data to which no such checks have been applied, hydrological validation may become the principle check on the reliability of the record. Where data are to be used for design purposes, hydrological validation is essential. Otherwise hydrological validation may be selective both in terms of the stations and of the tests applied. Thorough hydrological validation requires a high level of professional expertise and can be very time consuming.

Data in-filling (completion) and correction

Raw observed data may have missing values or sequences of values due to equipment malfunction, observer absence etc. these gaps should, where possible, be filled to make the series complete. In addition, all values flagged as doubtful in validation must be reviewed to decide whether they should be replaced by a corrected value of whether doubt remains as to reliability but a more reliable correction is not possible and the original value them remains with a flag.

In-filling or completion of a data series is done in a variety of ways depending on the length of the gap and the nature of the variable. The simplest case is where variables are observed with more than one instrument at the sae site (e.g. daily raingauge and recording gauge), the data from one can be used to complete the other. For single value or short gaps in a series with high serial correlation, simple linear interpolation between known values may be acceptable or values filled with reference to the graphical plot of the series. Gaps in series with a high random component and little serial correlation such as rainfall cannot be filled in this way and must be completed with reference to neighbouring stations through spatial interpolation. Longer gaps will be filled through regression analysis or ultimately through rainfall-runoff modelling. However, it must be emphasised here that various methods used for in-filling or correction will affect the statistics of the variable unless care is also taken with respect to its randomness. Nevertheless, it is not advisable to use completed or corrected data for the purpose of designing an observational network.

Data Compilation

Compilation refers primarily to the transformation of data observed at a certain time interval to a different interval. e.g. hourly to daily, daily to monthly, monthly to yearly. This is done by processes of aggregation. Occasionally desegregation, for example from daily to hourly is also required.

Compilation also refers to computation of aerial averages, for example catchment rainfall. Both aerial averaging and aggregation are required for validation, for example in rainfall-runoff comparisons, but also provide a convenient means of summarising large data volumes

Data Analysis

The procedures used in data validation and reporting have a wider analytical use. The following are examples of available techniques

- basic statistics (mean, standard deviations, etc.)
- statistical tests
- fitting of frequency distributions
- flow duration series
- regression analysis
- rainfall depth-area-duration
- rainfall intensity-frequency-duration

Data Reporting

It has been a practice to publish all available data for a state or river basin. The report usually contains all the informations regarding the state or river basin. This include the following:

- maps showing observation stations within their catchment and administrative boundary
- lists showing the stations and the period of record available
- summary description of salient facts associated with stations
- significant trends in the behaviour of the hydrological variables or alarming situations which need immediate attention of planners and designers
- selective listings or graphs to give examples of available formats

4.0 HYMOS as the Software for Data Processing

HYMOS is a database management and processing system for hydro-meteorological quality and quantity data, designed for use on personal computers. It arranges a convenient structuring of data in a database and provides an extensive set of tools for data entry, validation, completion, analysis, retrieval and reporting.

HYMOS is comprehensive, well tuned and easy to use. The graphical and map-based user interface offers a pleasant and efficient working environment. The information system facilitates processing of a wide range of water-related data. Field of application includes ;

- surface and ground water hydrology
- meteorology
- water quality and ecology
- water resources management

HYMOS data base is time series oriented with common facilities for spatial analysis. In combination with a GIS for comprehensive geographical data analysis, it will cover all data storage and processing requirements for planning, design and operation of water management systems.

HYMOS is developed to streamline the storage and processing of hydrological ad meteorological data. It is tailored for use by hydrological and meteorological data processing branches, water resources management authorities, water boards, and water engineering consultants and hydrological advisors.

4.1 Structure of HYMOS

HYMOS integrates the distinctive phase in the processing of hydrological data. The activities are carried out in specific processing modules: each module consists of a number of programs enclosing particular compilation and computations. The modules are structured according to a logical sequence of activities in data processing. All modules are linked to the HYMOS database, structured and controlled by a data management system. Different data base can be operational under HYMOS.

HYMOS comprises of following systems;

- a data base management system to create a database, to structure the database and to define user identification ;
- a data storage and retrieval system, covering data entry editing, reporting in tabular and graphical form as well as the transfer and retrieval of data ;
- a database processing system, including validation, series completion by interpolation, simulation and regression techniques, elaboration of flow measurement, data compilation, statistical analysis and time series analysis.

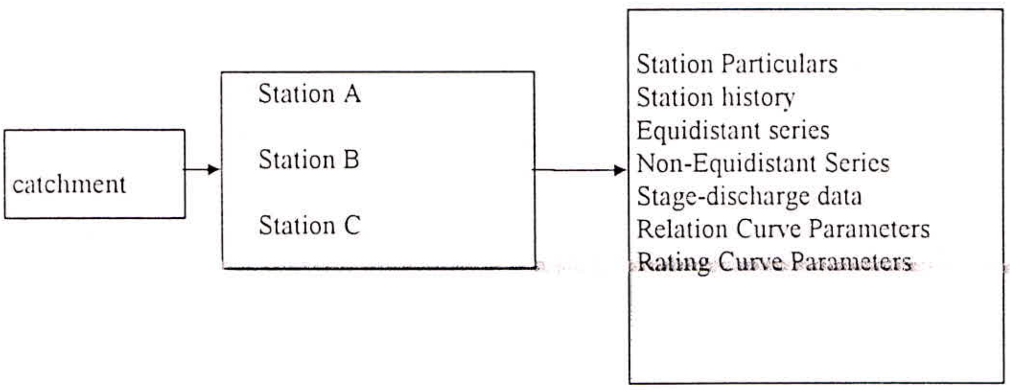
4.1.1 Database Management System

Database management deals with the creation and structuring of database and with the definition of user identifications.

Structure of Databases

A HYMOS created database comprises the hydrometeorological data of one or more catchment, or sub-catchment. As many databases are required can be used. Each database is stored in a separate directory.

In a particular database the data are, except for the catchment characteristic, structured station wise. Data type and time intervals are used for further identification of station data. The overall structure is shown in the following scheme;



Definition of Users

User ID's can be categorised into three parts;

a) group name
b) user name
c) Password

The access to a particular database is reserved to specified groups. Within a group user defined with different levels of authority, ranging for access to all facilities to data retrieval operations only. Each user must enter a unique password, which only is changed by the system manager.

4.1.2 Data Storage and Retrieval System

Under this, the main activity is that of creating the map layers, entry and editing of data observed. These options are described as follows

Maps and Monitoring Networks

The graphical interface provides you with interactive, graphical and map-based tools to define the monitoring network in the area of interest. You can add a new area of interest by setting up of map layers of the interested area. These maps can be produced with you own GIS systems. The following formats are supported:

- *.bna (Atlas GIS, but also PC-Arc-Info can export this format)
- *.mif (Map info Format)
- *.mos (MOSS format, a grid based GIS)
- *.shp (Universal shape format)

The monitoring network consists of stations for monitoring different types of parameters. In HYMOS, the following commonly used monitoring station types are pre-defined:

- water level
- discharge
- groundwater
- water quality
- meteorology
- rainfall
- snow height
- ecology structure
- spatial average

The only reason why these different types are defined was to easily trace back your data.

Series and Functions

HYMOS allows you to store general information, hydrological information (time series) and general characteristics, remarks on measured series etc. Hydrological information can be any information. Default parameters are flow data, rainfall, groundwater level etc. The hydrological information is identified by definition of :

- station code
- time (total interval and time step)
- data type or parameter

Station code: Each station contains a unique station ID, this is a unique set of characters with a maximum of 12 characters. The easiest way of coding a station is by using the station mane or abbreviation of it. When a station is added in the network you have to supply the station ID, never to be changed.

Time base : This is function is used to describe the time interval, time period to be used for entering the data and to analyse the data. There are two different types of time bases for any data:

- Equidistant time series ; a sequence of data with a fixed time intervals ; and
- Non-equidistant time series ; sequences of data unequally spaced in time. this implies that each data point has to be stored with the time label.

Data type or Parameters : Parameters are equivalent with the data type. The data type used in HYMOS is a three character specification of the type of data. Any combination of up to three characters is permitted and can be added to the available set. Examples,

- MPH : historical precipitation
- MPS : precipitation by standard rain gauge
- MPA : precipitation by autographic rain gauge
- HZS : water level by staff gauge
- HZA : water level by AWLR

Data Types used in HYMOS

The types of data handled by HYMOS can be categorised in the following groups;

- **Space Oriented data, Covering**

    Catchment characteristics
    Stations Particulars
    Station histories

- **Time oriented data, including**

    Equidistant time series
    Non-Equidistant time series

- **Relation Oriented data, divided in :**

    Stage-discharge data
    Relation curves

4.1.3 Data Processing System

The data processing system of HYMOS includes following modules:

- data validation
- data completion and regression
- flow measurement
- data compilation
- statistical analysis
- time series analysis

Data Validation

The data validation can be done using the following procedure

1. data screening by flagging, printing and tabular comparison of time series and computation of basic statistics.

2. graphical evaluation of time series (lines and bars), including

    - plotting of time series
    - residual time series graphs
    - residual mass curves
    - moving averages
    - water balances
    - graphical presentation of series availability in the database

3. relation curve analysis to establish relation equations and to investigate shifts in the relationships; it comprises of ;

    - graphical display of relations
    - fitting of relation curves
    - storage of relation curve parameters in the data base
    - comparison of relation curves of different time period

4. investigation of series homogeneity by means of double mass analysis, presented in graphs and label

5. statistical tests on data homogeneity and randomness

6. spatial homogeneity test (near neighbour-hood technique), where data at a base station are compared with weighted averages of neighbouring stations, selected on distance and orientation.

Series Completion and Regression

The completion and regression module comprises of

1. interpolation techniques for filling in missing data based on time and space interpolation
2. regression models to establish relationship, or to fill in missing data; the models may vary seasonally (maximum 12 periods), and can be of the following types;

   - polynomial equations
   - power equations
   - logarithmic equations
   - hyperbolic equations
   - exponential equations
   - simple and multiple linear regression equation
   - stepwise regressions

3. physically based lumped parameter rainfall-runoff model, SAMO, for filling in missing runoff data. SAMO is derived from the Sacramento Streamflow Simulation Model.

The model is system of parallel and serial reservoirs. In the simulation of runoff process a distinction is made between the land phase and channel phase. The land phase is approached by an explicit moisture accounting lumped parameter model. The catchment area is divided into one or more segments, discharging to a channel. Within every segment area homogeneity with respect to the rainfall and the basin characteristics is assumed. The propagation and attenuation of flood waves in the channel can be simulated by hydrological routing methods.

Flow Measurements

This module comprises a number of techniques for validation and elaboration of flow measurements and rating curves including;

1. Computation of stage discharge relationships given as parabolic and power type equations with
   - estimation of datum correction, and determination of coefficients of equation upto five water level ranges per relations for simple rating curves and compound curves
   - correction for backwater effects
   - corrections for unsteady flow
   - shift correction

2. Validation of rating equation for different periods and new measurements.

3. extrapolation of rating curves

   - computation of cross section capacities and parameters
   - graphs of cross sectional parameters versus stages
   - computation of synthetic stage-discharge data beyond the measured ranges

4. Stage-discharge transformation, using
   - rating curves stored in the database
   - rating equations of measuring and control structures for critical and sub-critical flow conditions and variable still level

Data Compilation

The data compilation module comprises of

1. aggregation and dis-aggregation of time series, where accumulative and instantaneous data are treated differently
2. series transformation with various arithmetic transformation options
3. minimum, mean and maximum series computation for selected time periods and transfer to the database
4. computation of aerial rainfall by
   - average of point rainfall data
   - thiessen polygon method
   - kriging method
5. interpolation and computation of best linear estimates and uncertainties in aerial quantities by point and block kriging method
6. computation of potential evapotranspiration, using
   - Penman method
   - Pan-evaporation method
   - Christian method
   - Radiation method
   - Makkink method
   - Jenson-Haise method
   - Blaney-Criddle method
   - Mass transfer method

Statistical Analysis

This module for statistical analysis includes ;

1. Computation of basic statistics and histograms
2. Fitting of probability distribution of the following types
   - Normal distribution
   - Log-normal distribution
   - Exponential distribution
   - Pearson-3 distribution
   - Log-Pearson distribution
   - Extreme type-I (Gumble) distribution
   - Goodrich distribution
   - Pareto Distribution for peaks over threshold

The parameters are estimated using the Method of Moments and a mixed moment-maximum likelihood method. Once the distribution is fitted for various exceedance probabilities can be computed. The Kolmogorov-Smirnov and Chi-square goodness of fit test are applied for hypothesis checking.

3. Computation of IDF-curves from monthly maximum precipitation.
4. Computation and plotting of

    - frequency curves
    - duration curves
    - average duration curves

Time Series Analysis

The time series analysis comprises of ;

1. auto and cross-correlation function computation
2. spectral analysis
3. run analysis ; computation of up- and down-crossings, run lengths and run sums
4. range analysis; computation of range of cumulative departures from the mean
5. analysis of storage requirements by sequence peak algorithm.

5.0 Remarks

Automated collection and processing techniques can lead to significant improvements in hydrological data collection, quality control, storage, accessibility, preparation of publications, and the use of data for analytical studies. However, the realisation of these improvements requires specialised training of the personnel involved in all phases of the work.

Reference

1. Chow.V.T., 1964. Hand book of Hydrology.

2. Chandramohan.T., 1992. Processing and Analysis of Hydrometeorological data. Status report no. IND/90/003, NIH, Roorkee

3. Delft Hydraulics, 1999. HYMOS users manual.

4. Hydrology Project, 1999. Surface Water Data Entry System (SWDES), users' manual.

5. Hydrology Project, 1999. Training Modules of Basic Surface Water Data Processing (HYMOS).

6. Jha, R., 1992. Hydrological Data Processing Technique. Status report no. IND/90/003, NIH, Roorkee

7. Rao, Y.R.S., 1993. Processing and Analysis of Hydrological Data. Status report no. IND/90/003, NIH, Roorkee