HYDROLOGY PROJECT

Technical Assistance

World Bank & Government of The Netherlands funded

Training module # SWDP - 02

Understanding Surface Water data processing activities under HIS

New Delhi, November 1999

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: dhvdelft@del2.vsnl.net.in

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

1. Module context

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	Understanding Surface Water data processing activities under HIS
Target group	:	Assistant Hydrologists, Hydrologists, Data Processing Centre Managers
Duration	:	One session of 60 min
Objectives	:	After the training the participants will be able to*:* • Know various hydrological data processing activities under HIS
Key concepts	:	• Frequency of receipt of data • Data validation • Data correction, completion and compilation • Data analysis • Data reporting • Data transfer
Training methods	:	Lecture
Training tools required	:	OHS
Handouts	:	As provided in this module
Further reading and references	:	

3. Session plan

No	Activities	Time	Tools
1	***Important aspects data processing under HIS:*** • Data processing under HIS	2 min	OHS 1
2	***Stages in Surface Water data processing:*** • List of various stages	2 min	OHS 2
3	***Receipt of data:*** • Record of receipt of data • Illustration - structured storage of manuscripts and analogue records	2 min	OHS 3 OHS 4
4	***Data Entry:*** • SW data entry	2 min	OHS 5
5	***Data validation:*** • Data validation and its objectives • Figure - Classification of measurement errors • Typical errors at observation stations • Validation - a sequential and complementary process	10 min	OHS 6 OHS 7 OHS 8 OHS 9
6	***6.1 Level of validations*** • Levels of data validation	2 min	OHS 10
7	***6.2 Primary validation*** • Primary validation process • Illustration 1 • Illustration 2	4 min	OHS 11 OHS 12, OHS 13
8	***6.3 Secondary validation*** • Secondary validation process • Illustration 1a • Illustration 1b • Illustration 1c	4 min	OHS 14 OHS 15 OHS 16 OHS 17
9	***6.4 Hydrological validation*** • Hydrological validation process • Illustration • Data validation - guiding factors	4 min	OHS 18 OHS 19 OHS 20
10	***Data in-filling (completion) and correction*** • Data in-filling and correction process • Illustration	4 min	OHS 21 OHS 22

11	***Data compilation*** • Data compilation • Illustration 1 • Illustration 2	2 min	 OHS 23 OHS 24 OHS 25
12	***Data Analysis*** • Hydrological data analysis • Illustration 1 • Illustration 2	2 min	 OHS 26 OHS 27 OHS 28
13	***Data Reporting*** • Aspects of data reporting	3 min	 OHS 29
14	***Data transfer*** • Data transfer requirements	2 min	OHS 30
15	***Wrap up***	15 min	

4. Overhead/flipchart master

5. Handout

Add copy of Main text in chapter 8, for all participants.

6. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

7. Main text

Contents

Understanding Surface Water data processing activities under HIS

1. Important aspects of data processing under HIS

- **The HIS processes, stores and disseminates groundwater as well as surface water data** but, as data collection and processing of surface and groundwater is done separately, this module and course are concerned with surface water data only. There is however, considerable overlap in the principles of data management.

- **Surface water data entry and processing are carried out almost exclusively by computer. Processing of data is accomplished using dedicated hydrological data processing software HYMOS.**

- **Processing of hydrological data is not a single step process.** It is carried out in a series of stages, starting with preliminary checking in the field, through receipt of raw data at Sub-divisional offices and successively higher levels of validation, before it is accepted as fully validated data in the State or Regional data storage centre.

- **The progress of data from field to data storage is not a one-way process.** It includes loops and feedbacks. The most important link is between the field station and the lowest processing level at Sub-divisional offices with frequent feedback from both ends but there will also be feedback from State and Divisional offices downward on the identification of faulty or suspect data through validation. Facilities for feedback from data users must also be maintained.

- **Processing and validation of hydrological data require an understanding of field practices.** This includes the principles and methods of observation in the field and the hydrological variable being measured. It must never be considered as a purely statistical exercise. With knowledge of measurement techniques, typical errors can be identified. Similarly knowledge of the regime of a river will facilitate the identification of spurious data. For example for river level or flow, a long period at a constant level followed by an abrupt change to another period of static level would be identified as suspect data in a natural catchment but possibly due to dam operation in a regulated river.

2. Stages in surface water data processing

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

In this introductory lecture an attempt will be made to generalise so that principles apply across all variables but with examples from particular variables.

3. Receipt of data

Data progress by stages through the processing system, from field to Sub-divisional office to Division and hence to State or Regional Data Processing Centres. **At each stage in the process target dates for receipt and for onward transmission are prescribed.**

A record of receipt and date 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 date 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 a Sub-divisional office or District office in a variety of media, as hand-written forms and notebooks, charts or digital data files on magnetic media.

Arrangements must be made for the storage of raw paper records after entry to computer files. In the case of Sub-divisional offices this will be a temporary storage but permanent storage will be in Divisional offices. Manuscripts must be maintained temporarily but in a well organised manner at the Sub-Divisional Data Processing Centres for a period of three years. Storage of hard-copy data (forms and charts) must also be logical and structured. This is to ensure that the original raw data remain accessible for further validation and checking. After a lapse of three years the manuscripts of all the data must be transferred to the respective Divisional Data Processing Centres for the purpose of archival.

Raw and processed data on the magnetic media will be sent from Sub-Divisional to Divisional Data Processing Centre and thereafter from Divisional to State/Regional Data Processing Centre.

4. Data Entry

The bulk of raw surface water data is 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, i.e. in the Sub-divisional data processing centres. This has the advantage that 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.

One exception to this practice is that data resulting from the analysis of samples in various water quality laboratories is entered at the laboratories themselves.

The Primary Module of dedicated hydrological data processing software HYMOS is available for accomplishing data entry from hand-written forms and tabulated autographic chart records. The software incorporates user-friendly data-entry screens which mimic the layout of all standard field forms to simplify direct entry from keyboard. Screens are available to select a station and data type from a list. Once selected, a screen with appropriate date and time labels is displayed against which data are entered.

The software provides automatic checking to reject unacceptable characters, e.g. alpha characters in a numeric field or duplicate decimal point. The validity of entered data is checked against pre-set limits and the program thus detects and rejects gross errors directly,

e.g. resulting from misplaced decimal point. A facility is available for assessing the hydrological integrity of the data through the simultaneous display of a graphical plot of the entered data.

Data entry software also allows entry of static and semi-static data.

5. Data validation

Data validation is the means by which data are checked to ensure that the final figure stored in the HIS 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 subject to errors and that undetected errors may also arise in data entry, in computation and, (hopefully infrequently) 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.

Measurement errors may be classified as random or systematic or spurious in nature (Fig. 5.1)**.**

- **Random errors** are sometimes referred to as experimental errors and are equally distributed about the mean or 'true' value. The errors of individual readings may be large or small, e.g. the error in a staff gauge reading where the water surface is subject to wave action, but they tend to compensate with time or by taking a sufficient number of measurements.

- **Systematic errors** or bias is where there is a systematic difference, either positive or negative, between the measured value and the true value and the situation is not improved by increasing the number of observations. Examples are the use of the wrong raingauge measure or the effect on a water level reading of undetected slippage of a staff gauge. Hydrometric field measurements are often subject to a combination of random and systematic errors. Systematic errors are generally the more serious and are what the validation process is designed to detect and if possible to correct.

- **Spurious errors** are sometimes distinguished from random and systematic errors as due to some abnormal external factor. An example might be an evaporation pan record where animals have been drinking from the pan, or a current meter gauging result using a very bent spindle. Such errors may be readily recognised but cannot so easily be statistically analysed and the measurements must often be discarded.

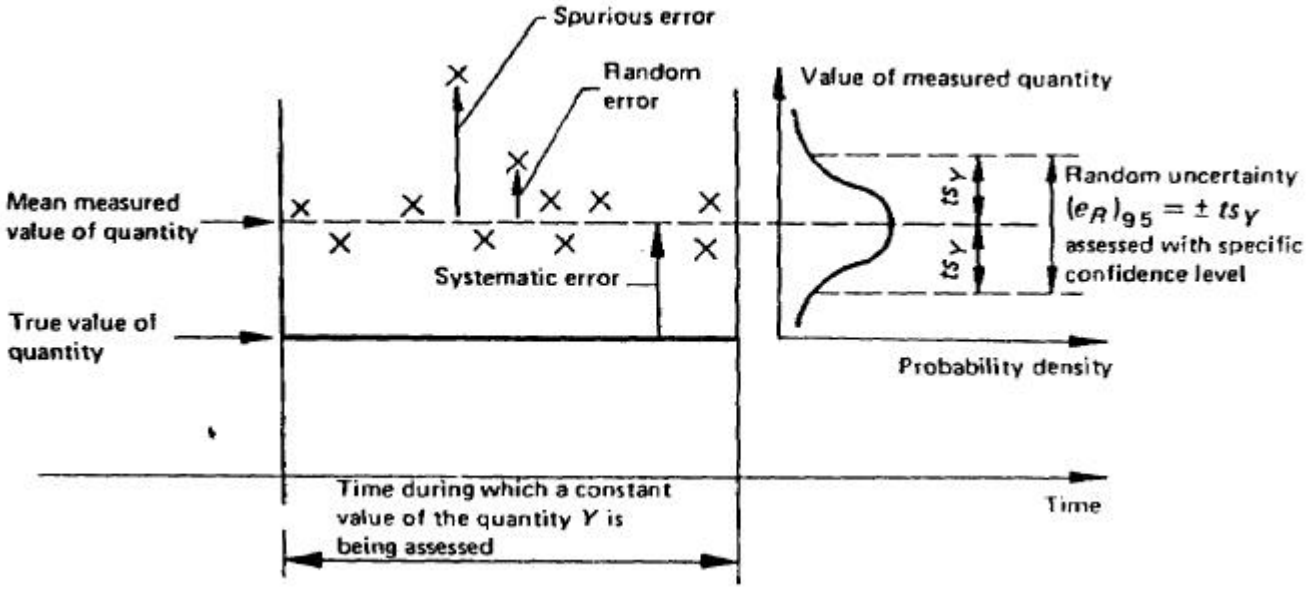

Fig. 5.1: Classification of measurement errors

Typical errors at observation stations are as follows:

- from faulty equipment, e.g. thermometer with air bubble in column
- from malfunction of instrument, e.g. slippage of float tape in level recorder (systematic but not constant)
- from improper instrument setting by observer, e.g. level recorder compared with staff gauge (systematic)
- from exposure conditions, e.g. stilling well blocked so that measured level in well differs from the river (systematic)
- personal observation errors, e.g. gauge misread or value interpolated away from site (random or spurious)
- transcription error in writing the observed reading
- error in field computation, e.g. current meter measurements.

Variables ultimately stored in the HIS may be directly measured (e.g. rainfall) or they may be derived using a relationship with one or more other variables (e.g. discharge). In the latter case the error in the derived value depends both on the field measurements and on the error in the relationship which is both random and systematic, the latter being particularly important if based on a relationship extrapolated beyond the limits of observation.

As a consequence of such errors it is important not only to ensure the use of good equipment and observational procedures but also to monitor the quality of all the data received. Validation procedures must be applied in a rigorous and standardised manner.

Validation involves a process of sequential and complementary comparisons of data and includes:

- for a single data series, between individual observations and pre-set physical limits
- for a single series between sequential observations to detect unacceptable rates of change and deviations from acceptable hydrological behaviour most readily identified graphically

- between two measurements of a variable at a single station, e.g. daily rainfall from a daily gauge and an accumulated total from a recording gauge
- between two or more measurements at neighbouring stations, e.g. flow at two points along a river
- between measurements between different but related variables, e.g. rainfall and river flow

Improvement in computing facilities now enables such validation to be carried out whereas in the past the volume of data the time required to carry out manual validation was prohibitive.

5.1 Levels 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 both 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 been divided so that those steps which primarily require interaction with the observation station are carried out in close proximity (i.e. at Sub-divisional office) whereas the more complex comparisons are carried out at higher levels.

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

(a) Primary data validation,
(b) Secondary data validation and
(c) Hydrological data validation.

Since, every hydrological variable has distinct characteristics it is necessary that data validation techniques be explained for different variables separately - Modules 8 and 9 for rainfall, 16 and 17 for climate, 22 and 23 for water level, 28 and 32 for stage-discharge data, and modules 35 to 38 for discharge. However in this section general principles are provided which apply to variables in general.

5.2 Primary data validation

Primary validation is carried out at Sub-divisional data processing centres immediately after keying-in or transferring the raw data.

Primary validation is primarily involved with comparisons within a single data series and is concerned with making comparisons between observations and pre-set limits and/or 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 out primary validation for example with respect to daily rainfall data. SWDES provides options to facilitate the primary validation with as little effort and ambiguity as possible.

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 was any error while making computations in the field or during keying-in the data. If it is found that the entered value(s) are different than the recorded ones then such entries are immediately corrected. Where such data values are

found to have been correctly entered they are then flagged as doubtful with a remark against the value in the computer file indicating the reason of such a doubt.

Apart from data entry errors, suspect values are identified and flagged but not amended at the Sub-divisional level. However the flag and remarks provide a basis for further consideration of action at the time of secondary and final data validation.

5.3 Secondary data validation

Secondary data validation is done at Divisional offices after primary validation has been carried out.

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 statistics 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 (e.g. convectional rainfall), 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 of good quality. If all the observation stations are given the status of being equally reliable 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 do not enjoy the status of being absolutely perfect.

As for the primary data validation, for the secondary data validation the guiding factor is also that none of the test procedures must be considered as absolutely objective on their own. They must always be taken as tools to screen out certain data values which can be considered as suspect. The validity of each of these suspect values is then to be confirmed on the basis of other tests and corroborative facts perhaps based on information received from the station. It is only when it is clear that a certain value is incorrect and an alternative value provides a more reliable indication of the true value of the variable that suitable correction should be applied and the value be flagged as corrected.

If it is not possible to confidently conclude that the suspected value is incorrect then such values will be left as they have been recorded with proper flag indicating them as doubtful. All those data which have been identified as suspicious at the level of primary validation are to be validated again on the basis of additional information available from a larger surrounding area. All such data which are supported by the additional spatial information must be accepted as correct and accordingly the flags indicating them as doubtful must be removed at this stage.

5.4 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 (water level) but is more often applied to derived variables (flow, runoff). 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 (or few) such checks have been applied, hydrological validation may become the principal 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. In the end it may suggest that a particular record is unreliable for particular periods or ranges but it will not always provide the means of correcting a faulty record.

6. 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 or whether doubt remains as to reliability but a more reliable correction is not possible and the original value then 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 same 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 correction is to be done using similar procedures as used for completing the data series.

7. 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 a process of aggregation. Occasionally disaggregation, for example from daily to hourly is also required.

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

Derived series can also be created, for example, maximum, minimum and mean in a time interval or a listing of peaks over a threshold, to which a variety of hydrological analyses may be applied.

8. Data analysis

The HIS is not designed to provide a comprehensive range of hydrological analysis techniques. However, procedures used in data validation and reporting have a wider analytical use. The following are examples of available techniques:

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

9. Data reporting

Past practice has been to publish all available data for a state or river basin. With the larger amount of digital data to be stored in the HIS this is now, neither practical nor desirable as legitimate users can easily be provided with the precise data they need in the format they require.

What is now required is a readily available document indicating what information is available and held in the HIS. This should include the following:

- maps showing observation stations within their catchment and administrative contexts
- lists showing the stations and the period of record available
- summary description of salient facts associated with stations
- summary hydrological information for all stations, e.g. annual and monthly totals
- significant trends in the behavior of the hydrological variables or alarming situations which need immediate attention of planners and designers
- selective listings or graphs(e.g. daily values) to give examples of available formats.

Periodic publication of special reports showing long term statistics of stations or special reports on unusual events may also be prepared. In addition a catalogue of data held in various HIS databases is prepared periodically.

Otherwise specific data can be provided to users on the basis of need. This can be prepared in digital form on magnetic media thus avoiding the need to re-key the data to computer. A wide range of tabular and graphical formats is available, for example showing comparisons of current year values with long-term statistics, thematic maps of variables such as annual and seasonal rainfall, duration and frequency curves, etc. More detailed information such as stage discharge ratings can be provided to meet specific needs.

10. Data transfer

Since data are processed by stages at several locations, rapid and reliable transfer of data from one location to another is essential. The decision on the optimal methods of transport have to be taken on the basis of volume, frequency, speed of transmission and the cost.

Where analogue data must accompany digital data for example from the observation station through to Divisional Data Processing Centres, it is expected that equivalent digital data are prepared on diskettes and CD-ROMs to accompany these data using physical data carriers.

Where there is little accompanying analogue data for example from Division to State Data Processing Centre then electronic transmission may prove more effective. The final decision on the availability of communication links will be forthcoming only after various data processing centres are functional since the communication technology is evolving at a very fast rate and new services are introduced in rapid succession.