SR No.4

DATA SYSTEMS AND LIBRARY

STATUS REPORT

GAUTAM ROY

NATIONAL INSTITUTE OF HYDROLOGY

JAL VIGYAN BHAVAN

ROORKEE-246 667 (UP)

INDIA

1983-84

PREFACE

The National Institute of Hydrology is a society under the Ministry of Irrigation, Govt. of India. Its main objective is research in basic and applied hydrology as well as water resources engineering. Data systems and library forms one of the important areas of activity at the Institute. As a requirement, before proceeding for training abroad under UNDP project, the sub-committee of Technical Advisory Committee recommended that the concerned scientist should prepare a status report in his area to enable him to obtain sufficient background of the status as well as need for further research/training in the area. This status report has, therefore, been prepared for the area of Data Systems and Library for Hydrologic Analysis.

This report aims at a comprehensive survey of hydrology-related data banks, - their component functions and adopted methodologies. One primary concern is with the systematic storage of data whereby they can be manipulated interactively and retrieved easily. This calls for the use of data base management systems whose application in hydrological data banks is recent, but holds tremendous promise for the future.

The report is presented in six chapters. Chapter I introduces the tasks of a data bank and the significant properties of data. ChapterII deals with the types of data used in hydrological studies, and the standards and methods of their collection. In Chapter III the various data processing activities-quality control,

editing, analysis and synthesis - have been reviewed. The techniques of data storage and retrieval systems have been discussed in Chapter IV. Some prominent hydrological data banks of the world have been reviewed, and their salient features discussed in Chapter V. The concluding chapter highlights need for data bank of NIH using available computer facilities of VAX-11/780 system.

CONTENTS

LIST OF FIGURES

ABSTRACT

The advent of sophisticated water resources and hydrological studies has necessitated the development of hydrologic data libraries in recent times. This report discusses the operating methodologies of such libraries. The report is introduced with a comprehensive description of the types and aspects of information usually encountered in hydrological investigations. The operational procedures of data banks are reviewed systematically. They are:

i) Collection of raw data which involves the standardization of measurement/observation procedures and the establishment of data collection network.

ii) Processing of raw data, comprising quality control, editing, analysis and synthesis, to prepare them for ready use in hydrological studies.

iii) Storage and retrieval of data using data base management principles. The storage and retrieval techniques are discussed separately, followed by a survey of the integrated data base management process. Illustrative cases of their use in hydrology-related areas are discussed.

The current developmental status of some major hydrologic data banks are reviewed with emphasis on the scientific and organizational aspects. The data banks reviewed cover a wide spectrum of hydrologic interests, and indicate the areas of current research. The study is concluded with a synoptic discussion of the need for a

hydrologic data library in this Institute, the operational strategies that can be envisaged for it, and the scientific background necessary for the same.

1.0 INTRODUCTION

The recognition of the vital role of hydrology in human life has led to tremendous growth and proliferation of hydrological studies in recent times. The successful and efficient execution of these studies depends on a vast and diverse amount of information. Where sufficient information is not available quantification of hydrological processes is possible only with limited accuracy, and plans for hydrologic control and development need to make compensatory provisions. However, such provisions cannot be perfect surrogates of the relevant information themselves. Hence it is essential to organise a systematic body of information in the form of a data library to fulfil the data needs of quantitative hydrological investigations.

1.1 Functions of Hydrologic Data Library

A hydrologic data library forms the vital link between raw data collected in the field, and their ultimate use in hydrological studies. Figure 1 shows the organisation of a typical hydrological data bank. There are several discernible operational procedures of a data library. These may be summarily described as follows:

a) Data collection: Raw data are acquired from various measurement and collection agencies, and preprocessed before entering into the computer. Preprocessing depends on the medium used for data

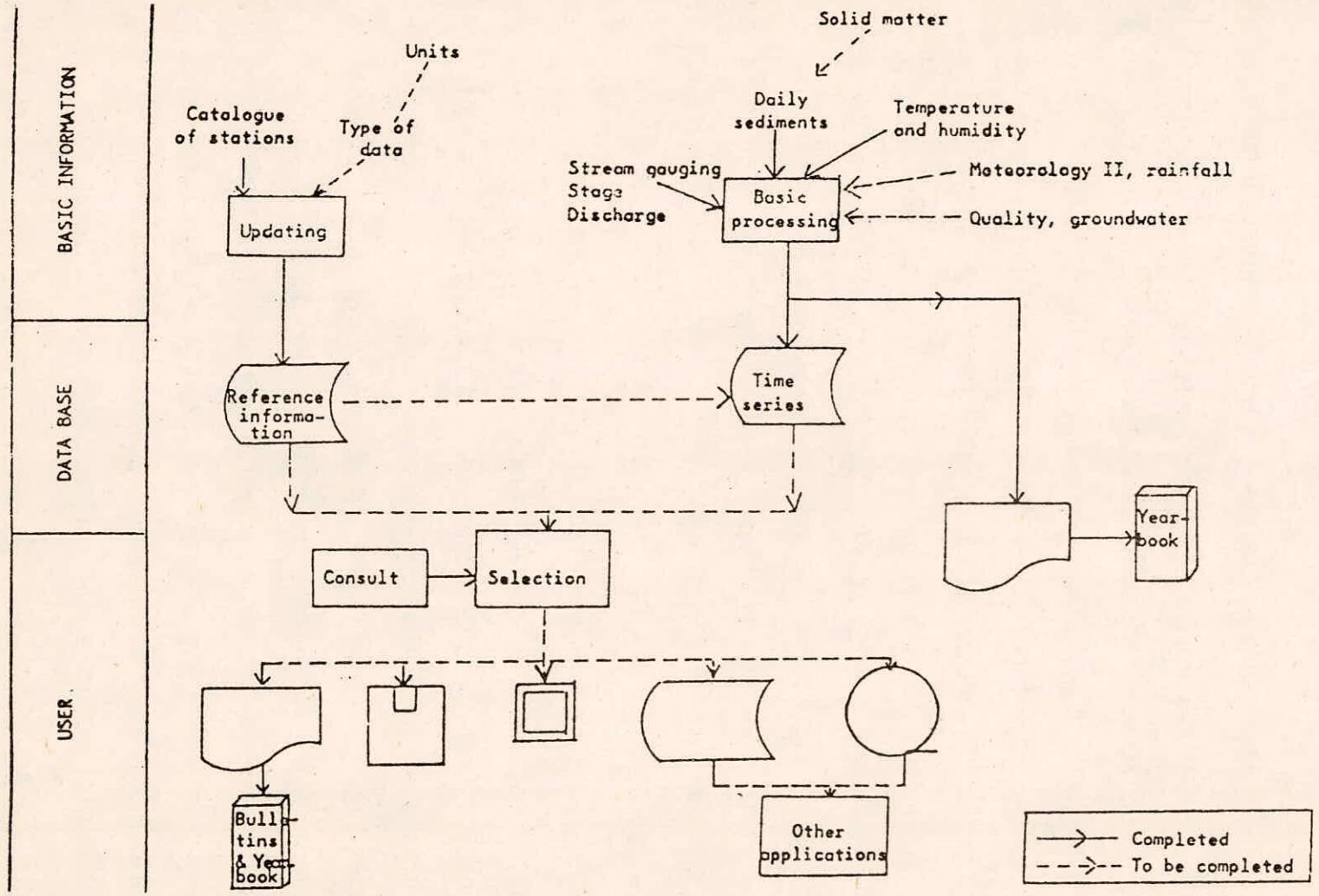

FIGURE 1 - HYDROMETEOROLOGICAL INFORMATION SYSTEM (XMO, 1981)

recording, and the input peripherals of the computer.

b) Data processing: Processing and dissemination of data involves quality control measures, classification of data, their conversion into consistent units and formats, and analysis and synthesis of data. The latter two procedures involve computation of physically meaningful or operationally useful parameters, filling-in of missing data, etc.

c) Storage and retrieval: Data base management concepts are used in storing information with relevant site/instrument specifications and quality indices. Data are stored in groups in direct, sequential or random access files depending on the data base structure.

Effective retrieval of data requires not merely their storage in an organised structure (data base) but also familiarity of the user with the data base components and their interrelationships. It is necessary to provide instructions and software devices to search, display and retrieve data in the desired forms. The instructions for locating data are usually aided by a query facility to serve the common user.

The basic operational links of a computerized data bank are illustrated in figure 2. As evident from the figure the operations of a data library are closely and sequentially connected. As the magnitude and diversity of the information that need to be stored increase the complexities of the processing, storage and retrieval procedures increase proportionately. This calls for great skill in the functioning of vast and growing data libraries.

Depending on the way in which data are collected and processed data banks may be centralized, or distributed (coordinated). A fully centralized hydrological data bank is physically located at one site

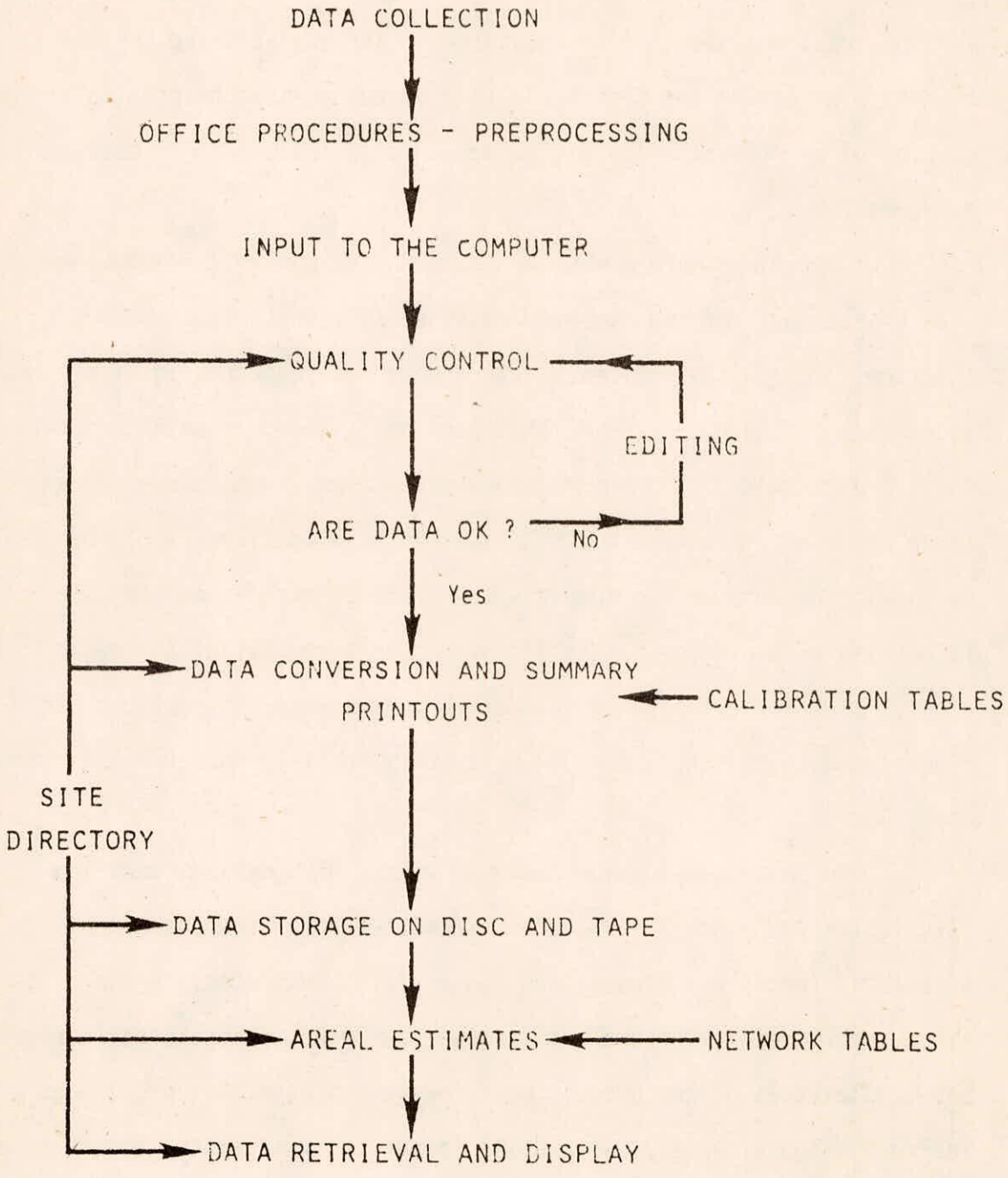

FIGURE 2 - HYDROLOGICAL DATA LIBRARY FRAMEWORK (I.H., 1981)

and comprises data covering all hydrological variables covering the entire country. Such organisations permit substantial economy of processing and storage, software and management. However due to the complexity and largeness of structure they may be impractical in big countries with a variety of source agencies and user requirements.

A distributed (coordinated) data bank is a coordinated or cooperative arrangement between different data banks, or their operating agencies. Such banks are more reliable, manageable, and adaptable to changing technologies or users'needs. They are typically used in real-time applications where the end user must have easy access to interactive use and control of data in the data bank.

Each data bank needs to be organised according to its specific needs and limitations. Individual data libraries incapable of meeting all user-needs directly often maintain two parallel modes of functioning, viz.,(a) storage of data as a data base, wherefrom data can be retrieved per se , and (b) indexing of data to provide requisite information on the types and brief description of data available from alternative sources.

1.2 Characteristics of Hydrology-Related Data

There are several aspects of hydrology-related information whose understanding is required in order to organise a data bank. These characteristics may be broadly described as follows:

1.2.1 Experimental and historic data

Information derived from controled experiments are invariant subject to the experimental constraints. Provided these constraints

are known such data are readily applicable or adaptable for specific purposes. On the other hand data pertaining to natural conditions (such as natural resources and socio-economic information) are historic in nature as these conditions are continually changing. Hence such information must be compiled and updated periodically.

1.2.2 Time and space coordinates

A hydrological event is the consequence of manifold processes whose domain in space and time are indefinite. However, the relative importance of each factor depends on its spatial and temporal configurations. Hence time and space coordinates must be directly linked with all historic data.

1.2.3 Quality

The data as acquired by human efforts depend on human limitations, environmental conditions and the means used to acquire them. Thus an estimation of the reliability and accuracy of data is necessary.

1.2.4 Quantitative and qualitative data

Quantitative data are relatively easy to deal with as they are amenable to mathematical representation and analyses. Qualitative data cannot be subjected to mathematical or statistical interpretation without at least some loss of information. In connection with automated data storage and retrieval systems distinction is made between 'hard' and 'soft' data. 'Hard' data is the term applied to numerical and narrative data that can be processed by a digital computer. 'Soft' data are those which cannot be processed by digital computers such as

information contained in reports, photographs and file cabinets.

1.2.5 Unit and format

Quantitative information of the same type may be expressed in various units. For ready use they must be converted to same units, consistent with international conventions, and stored in specific formats.

1.2.6 Use-orientation of data

Hydrological model studies often require that raw data in altered forms. Models may be deterministic, parametric or stochastic. For the latter two types of models parameters in frequent use constitute part of basic data of a data bank such as geometrical parameters and statistical indices. The handling of data also depends on their use-schedule viz., real time and non-real time studies.

The full import of the above characteristics is realised in the efficient design of data libraries.

2.0 DATA COLLECTION

Scientific data consisting of invariant physico-chemical properties are empirically determined, and are usually well-documented in scientific literature. However, historical hydrology-related information are constantly growing in quantity and diversity. Hence a systematic procedure must be adopted for the collection and compilation of hydrology-related data. This necessitates the understanding of the specific data needs of hydrological studies.

2.1 Classification of Hydrology-Related Information

Hydrological phenomena being part of the complex of incessant global processes, several of the latter play important roles in hydrological investigations.

2.1.1 Hydrologic data

Surface hydrology: These include information on streamflow, springs, reservoirs, lakes and oceans. The pertinent information needed are (i) storage, or discharge rate; (ii) water level or stage; (iii) concentration, or transport rate of sediment; (iv) water quality data; (v) water temperature and conductance; (vi) heights and velocities of waves and surges; (vii) topographical features of the water holding or conveyance system.

Subsurface hydrology: It comprises of water level and storage water quality, temperature, piezometric head, infiltration and physical properties of the conveying medium of ground water.

2.1.2 Meteorological data

Relevant meteorological information consists of precipitation (form, amount and intensity), snow depth, water equivalent of snow cover, air, temperature, wind speed and direction, himidity, solar radiation and pan evaporation.

2.1.3 Ecological data

Various constituents of the ecological system are of concern in hydrological investigations. Of particular interest are the vegetal cover over a region, microbial organisms in water (such as planktons) and in the biosphere, and small plants and animals that have a direct relation with water quality such as algae, weeds, shrimps, and fishes.

2.1.4 Geologic data

Quantitative assessment in these areas is often difficult, and much information exists in 'soft' form. The pertinent information consists of topographic maps, sedimentation, lithology, bathymetry, hotogeology and environmental geology.

2.1.5 Land use data

Human activities such as industrialization and agriculture, with their accompanying paraphrenalia, affect and alter the hydrological regime in countless ways. The major land use forms are alteration of vegetal cover (such as forests to grasslands or farmlands), occupancy and use of land for residential and commercial puirposes, and changes introduced in water conveyance and storage systems. Urbanization has also led to rapid changes in water needs and usage for domestic

and industrial purposes.

2.1.6 Socio-economic data

Hydrological planning and management are largely dependent on socio-economic considerations. These include demography, business and occupational economy, educational, welfare and recreational activities, socio-economic developmental trends and programmes, and government and local rules and regulations.

2.2 Integrated Data Collection Procedure

The data enlisted in the previous section are diverse in nature, and are measured/observed by sundry organisations, each with its specific modes and standards of measurement. It is, therefore, necessary to systematize the procedure of data collection. In several developed countries data collection has been organised into a streamlined process by designing data networks. A network connects all the data collecting agencies to an apex body which prescribes quantitative methodologies for the observation and measurement of data. Such specifications refer to:

i. site locations for data measurement ;

ii. types of data to be collected/measured ;

iii. frequency of data collection/measurement;

iv. duration of data collecton activities; and

v. instrumentation and techniques for data colection/measurement

Standard procedures for the above are framed in order to reduce errors, and to ensure a certain level of uniformity in the quality of data. Cognizance must be taken of the practical limitations as well as the specific uses of data in hydrological studies. Appropriate method

for data acquisition have been recommended by the WMO, US Department of Interior, and other national and international organisations.

As hydrological stdies are varied in nature, there are two possible ways of organising data networks.

A) A single multipurpose network of all data collecting agencies. The disadvantage in this method is that the resulting aggregate of collected information is huge and difficult to handle. Attempts have been made to minimize this problem by recommending different intensities of data collection in different regions. But such measures may not be effective as data needs of hydrological studies cannot be decided in advance.

B) Moss (1979)[9] envisaged the setting up of separate use-specific data networks. The collecting agencies may be common to the networks but since each network serves only one purpose, it needs to deal with considerably lesser volume of data. At the same time interactive functioning between the networks is also possible. However, the proliferation of networks leads to increased expenditure and manpower in setting up and maintaining the infrastructure.

Keeping in mind the continually changing data needs of hydrological investigations no prior decision can be taken as to which of the above alternatives will be optimally cost-efficient in a given situation. A consideration of paramount importance is the adaptability of the data collection network, wherein the scope of development is seeded in its design.

3.0 DATA PROCESSING

3.1 General

Prior to the storage of inflowing data it is imperative that the information be tested, suitable rectifications be made, and some measure of its quality established. The first concern arises with respect to the time base. The data collected from various sources is not likely to have a common time base. A computer-based control and data-logging system is therefore necessary, not only for precise time correlations, but also for real time data used in water resources systems operation and management. Apart from real time considerations, data processing is concerned with quality control, editing, analysis and synthesis of the collected information. The former two processes are sometimes collectively called primary processing, and the latter secondary processing.

3.2 Quality Control

Users of data must be aware of the potential problems arising in hydrological studies as a result of the quality of data. The measurement of data is affected not merely by the method employed but also the field conditions at the time of measurement. With reference to the former a list of standard measurement/observation procedures are recommended for use. For the field situation specific factors need to be recorded which represent the environmental condition.

Errors and inconsistencies in raw data exist due to a variety of causes. The main types of errors are ;

i) Procedural errors: arising due to improper techniques, measuring devices, base information or calibration charts.

ii) Operational errors: caused by the faulty operation of the adopted mechanism or procedure, or the result of human limitations.

iii) Peripheral errors: are independent of the procedure adopted, or their execution, such as unknown or unpredictable environmental factors.

iv) Non-chargeable errors: occur after measurement, i.e. while recording or transferring information. They my also arise due to misinterpretation of procedural instructions.

Due to the varied and uncertain nature of errors many of the errors may not be identifiable or rectifiable. Thus quantitative checking cannot completely root out the errors, though values can be labelled suspect where appropriate. Routine error corrections are useful where a definite trend or correlation is expected, and are usually made by computerized methods. However, automated quality control programmes are limited in scope, and may not be completely reliable. Hence they must be supplemented by manual intervention.

The commonly employed quality and consistency checks are:

i) Testing for absolute or physical limits of data (such as negative inflows to reservoirs, measurements outside range of instrument capacity), i.e. physical consistency.

ii) Checking for mutually conflicting sets of data values within a set of data i.e. internal consistency.

iii) Double mass curve analysis (for rainfall, runoff etc). of contemporary sets of data of neighbouring stations.

iv)Time series analysis to detect changes in the homogeneity in time series. This is a valuable supplement to double mass analysis.

v) Cross correlation and regression analyses to compare between related types of data.

From the above tests conflicts between data values are resolved, and missing data filled up where possible. Correlation coefficients and regression equations are also computed where useful and physically meaningful.

3.3 Editing

Editing consists of conversion of the semi- processed data into those with consistent units and formats. Usually the following international conventions are used:

i) Universal Decimal Classification (UDC): Sponsored by the Federation Internationale de Documentation (FID) for subject and miscellaneous classification.

ii) I.S.O. recommendation for common elements and their units.

The above recommendations do not encompass the whole spectrum of hydrology related data. Additional nomenclature, notations and units are sometimes required. Such specifications have been compiled by WMO, NAWDEX and other large-scale organisations. In addition to the symbols and units, formats are standardized according to the type of data and economy of storage.

3.4 Analysis and Forecasting

Many of the hydrological variables recorded are not directly used for modelling purposes. Often a set of data is reduced to.

represented by or augmented with additional parameters (statistical or otherwise) in actual computational studies. Since these parameters are frequently used it is convenient to develop routine programmes for computing and storing them in the data banks. Such computations include areal estimates of data (precipitation, runoff, evaporation, etc.) statistical moments, and correlation and regression analyses. Typical simulation and optimization models are also run on the data depending on their expected needs.

Information available in analogue form often need to be digitized for purposes of processing storage and ready reproducibility. Spatial information may be stored in two or three dimensional grid systems and the necessary quality control tests performed on them. The process of digitization may be achieved manually by digitizing tables or by automated or semi-automated digitizing instruments. The conversion process can be checked by computer plotting of the data into overlays and comparing with the original traces.

The acquiring of fresh data necessitates the updating of the existing data file by comparison with the new data values, and recomputing parameters if necessary. Thus time series and auto-correlation analyses are re-invoked to ensure consistency of data in time, and to establish long-term trends and statistical moments. Moreover if the files in which data are to be stored are of pre-defined lengths, it is necessary to check for storage problems, and reallocation of storage space is necessary.

For economy of storage as well as for ease of data handling it is necessary to reduce data to a minimum. While some redundant data can be eliminated by processing, quality control and corrective measures may actually increase the volume of information. Further, due to uncertainty in the nature of data itself it is desirable to

preserve the original information in addition to the corrected values. One way of reducing the amount is by wiping out data values which, within a certain tolerance interval, can be interpolated between neighbouring values. This procedure can significantly compress the body of information for continuously recorded data.

Water resources and water quality regulating services like reservoir operation, toxicity warning and flood forecasting, may necessitate real-time data handling. In such cases measurement stations must be equipped with suitable devices (such as telemetering equipment, on-line connectiuons to computer, etc.) for immediate transmission of data to the data bank. Real time data processing involving quality control is difficult since manual checks may not be possible and automated quality control can misinterpret and distort data resulting in crucial errors.Real time data handling is highly complex and depends largely on the specific use of data and the availble computer software.

4.0 INFORMATION STORAGE AND RETRIEVAL

4.1 General

Data can be stored in computer-processable forms in various ways but an organizational structure is essential. The collective storage of groups of data constitutes a data base. The data base organisation depends on two aspects:

a) The types of data (entities) and their mutual relations

b) The number and types of intended use of the data.

Several problems arise in the common storage of data in data library. Since the same data records & files are accessed by several users it is necessary to distinguish between different users and their specific needs. Further, while users are free to manipulate data for their intended purposes, the stored data must remain unaltered in the data bse. The efficient tackling of these problems comprise the tasks of data base management systems (DBMS). While DBMS is essentially an integrated process, conceptually the storage and retrieval processes need to be split in order to study their salient features.

4.2 Information Storage

Computers serve as useful media for DBMS since vast and varied amounts of data can be processed, stored and retrieved with ease. The storing data in a data base it is necessary to compile an inventory of the stored data. The organisational modelling of data in a data base constitutes a scheme or data model, which provides

the complete list of name code, position, size and brief descriptions of each component of the data base.

The use of a schema renders the data base organisation independent of the amount of data. The record segments enquired for a particular application (set of operations) constitute the subschema for that application. The order of data fields may be rearranged in a subschema record.

The management of individual data files depends on the mode of storage of data. Data are usually stored in arrays in direct, sequential or random access files. Where data are collected by fully instrumented and regulated procedures the quality and frequency of data are relatively constant and they are stored in files of constant lengths, chronologically compiled. For data recorded manually at variable time intervals they must be stored in variable length files. The time and conditions of measurement and quality indices must be specified in conjunction with the data.

The medium of data storage is selected according to the type of data and their expected future needs. Frequently used media are casette tapes, magnetic tapes and disks, hollorith cards, paper tapes and microforms. Real-time data used in water resources/quality management or operation are stored on magnetic disks, while magnetic tapes are employed for storing historical information. Magnetic tapes have the advantage of rapid access by a computer and indefinite reuse. However, magnetic codes are vulnerable to gradual fading, and may be destroyed by accidents or unintentional over-writing. Hence it is desirable to maintain such data in duplicate as a precautionary measure.

Microfilms and other microform devices hold information in very compact forms, and with near-permanance under a wide range of

environmental conditions. However, they can not be reused which deters their frequent use in data storage. As a rule, microfilms are used to store data that cannot be precisely digitized such as maps, charts, file cabinets, reports, and photographs. Whenever, possible, however, data in analogue forms are digitized and stored on tapes or disks in addition to microform storage.

4.3 Data Retrieval

In digital computers data are stored in alphanumeric form in simultanity with specifications and descriptive statements. The size, format and quality are also recorded in respective data files. Some of the data specifications are represented by codes or abbreviations to save storage space.

Depending on the internal structure of the information base the data files are accessed by suitable software devices. Computer programmes allow for searching, editing, display and output of data in suitable forms. The data retrieval system is facilitated by appropriate user instructions. These pertain to:-

i) Description of the data base structure and size, indexing and components of the data base.

ii) The type and range of data values for each component

iii) Explanation of abbreviations, codes, quality indices and other specifications.

iv) Information on the updating of data.

All the components of a data base may not be indexed in order to save storage space, as for instance when some components are frequently used, or when the number of components is significant in comparison to the number of data held therein. In such cases indexed

and non-indexed components in the same schema are designated by respective keys. Appendix 1 illustrates a typical display from a data library (NAWDEX).

Most data accessing jobs are achieved through on-line terminals though batch mode is usable when the data base structure and software procedures are familiar. In general, instructions for software procedures and commands must be available to the common user. These instructions are on:

i) locating and display of data ;

ii) prevention of display of unwanted data;

iii) specification of selected data records;

iv)controlling the sequence and format of data output;

v) analysing the data to obtain statistical or other parametric information; and

vi) printing of data from different data base levels, and from disjoint schema records.

Several means of data output may be required depending on user needs, such as computer printouts, microfiche, paper tapes, magnetic tapes, etc. For converting digital data into analogue form, as in the case of maps and charts use of computer graphics and digital plotters is necessary

4.4 Integrated DBMS

The fundamental problems in the design of a DBMS have been outlined at the outset of this chapter. Three major steps (see figure 3) can be discerned in setting up a data base:

a) Data analysis and modelling consisting of (1)

recognizing relations between data; and (ii) functional analysis which is a modelling activity resulting in a data model or functional ens model

b) Design of the logical data structure, i.e. the logical format of the schema, from the users'perspective of the data organisation.

c) Design of the physical data structure, or the physical schema format, which is the software organisation of the data model.

The physical format conveys the logical relations between data which are:

i)Reflexive, symmetric, and transitive.

ii)One:one; One:many; and many:one

Depending on the type(s) of relations used, three types of data base structures exist, viz., the tree or hierarchical structure the plex or network, and the relational data base.

In the hierarchical data base each data type (entity) is related to a single parent data type. This interactive operations of a number of entities may require linearization of large trees traversing a large number of thereby increasing costs. Such structures are easy to vizualize but interactive data processing between any two or more entities is difficult. Nevertheless, due to the fact that common hydrologic data processing does not often involve unconnected data types, the hierarichical data base has been frequently used hydrological data libraries. Typical hierarichal DBMS are IMS (Information Management System marketted by IBM Co., and used in the Hydrological Data Bank of Bavaria, FRG), CDMS 500 (marketted by Digital Equipment Corp.,USA for PDP 11 computers), MARK IV (marketted by Informatics,USA for IBM and Univac computers; and used

at NAQUADAT, Canada) and SYSTEM 2000 (marketted by MRI Systems, USA; implemented IBM , Univae and CDC computers, and used by NAWDEX, USA).

The network structure is a powerful representation of many to many relationship. It relies heavily upon loops and pointers. The pointers data segments in one segment, file to those in the same or some other file. Junction segment, secondary segments , are needed to convey many to many relations. Typical examples of network DBMS are FORDATA (developed and used by CSIR, Canberra one CDC Cyber computers, and using FORTRAN user language) and IDMS(developed by Culliane Corp.,USA and implemented on several major computer systems).

In the relational data base each relation describes a type of entity. Thus it depends heavily upon tables and their mathematical equivalent matrices. Tables also convey the relations that prevail between and among other tables and are hence the most important construct of the relational data base. The relational data base is more recent than the previous two and took firm shape only after the CODASYL conference in 1971. A typical example of a relational DBMS is INGRES (developed at the University of California for DEC PDP 11 computers).

The running of a DBMS consist of the following activities:-

i) Routine activities involving reference, posting and maintenance.

ii) The writing of a new application programme or extending the scope of the data base, necessitating software development.

iii) Query facility operation to help users in searching, manipulating and retrieving data, which involves the use of an easily communicable query (programming) language.

The query processing facility is an extra load on the DBMS software, and hence dispensed with when it is possible to instruct or educate users in data handling procedures.

More important than query language are the data description and manipulation languages (DDL and DML). DML is used to request input/output and positioning for the application programmes reference to the data base. Application programmes are used in the processing and handling of data and include inbuilt software devices or additional programmes (such as statistical computations). The DDL performs a declarative function. It describes the entire data in the data base for all possible applications at three levels-physical level, schema level & schema levelThus it is used to relate between data at each level, i.e. between different data records or segments.

For this purpose it employs keys to identify and order the segments. Among the DDL's of major importance are DL/1 of IBM's Information Management System, and CODASYL's DDL. A typical programme for creating a new component within a data base is shown in Appendix II. It uses FORTRAN-based programming language to update the schema and the data entry.

4.5 DBMS Applications in Hydrology-Related Areas

Various types of DBMS have been used in Hydrology-related data banks. While some of these DBMS's are marketted by reputed firms others were indigenously developed. Figure 3 shows the data base structure of a national hydrological data bank developed indigeneously to meet its specific needs. The following cases, illustrates the application of DBMS in hydrologic and similar field:

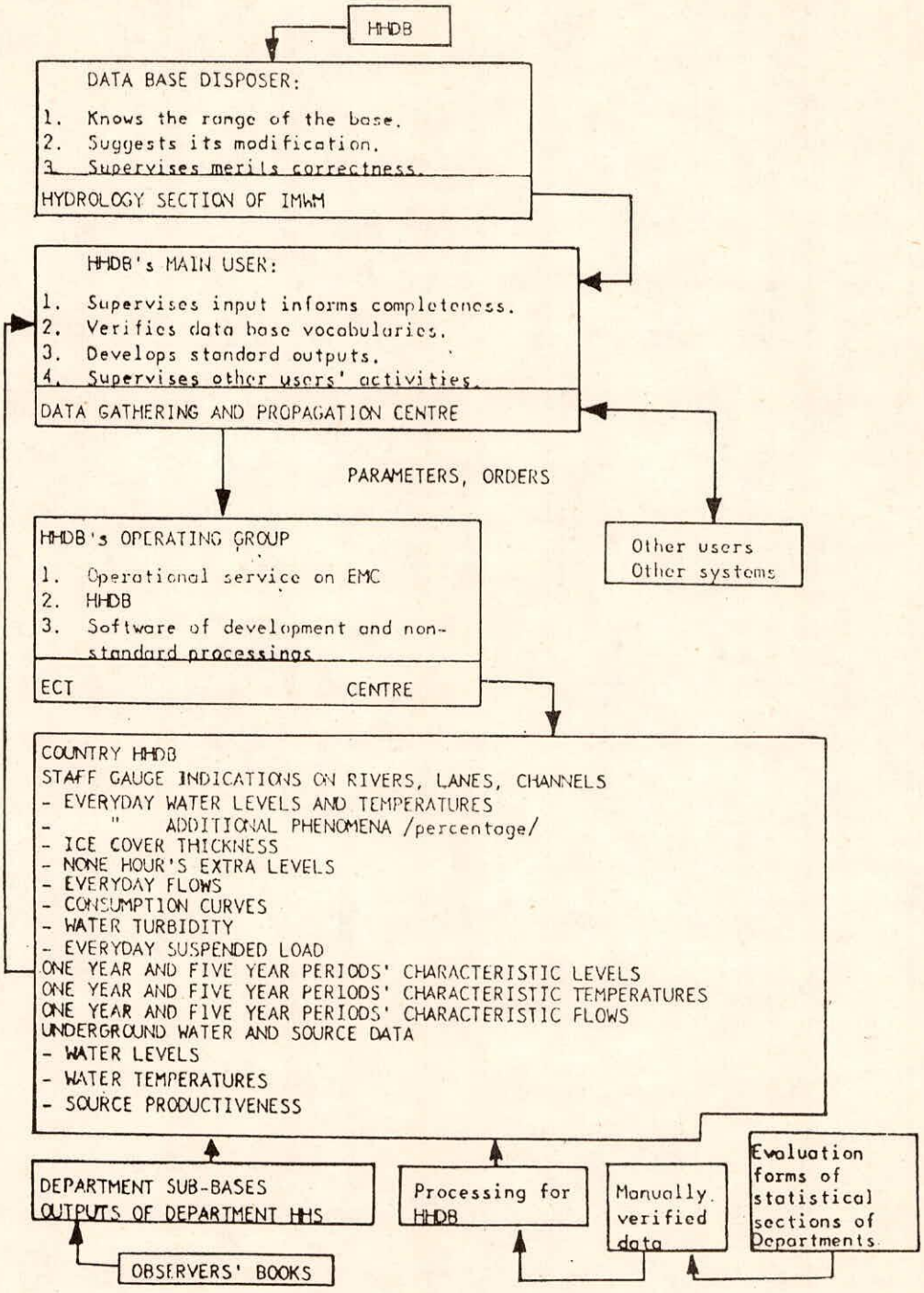

FIGURE 3 - THE HISTORICAL HYDROLOGY DATA-BASE STRUCTURE

4.5.1 ADAPT

(Areal Design and Planning Tool) is a methodology rather than a model to (i) develop a data base for spatial land/water system, and (ii) utilize the data base in conjunction with analytical mathematical techniques.Since spatial data do not have any logical sequential order their handling presents difficulties. ADAPT divides the region by a triangulated irregular grid network which allows variation of information concentration of different regions. Latitude, longitude and elevation of each triangle vertex gives the terrain or surface characteristics, while triangle sides represent surface boundaries. The spatial data base, opposed to the commonly used regular grid, enables integrated model studies of the region with great accuracy. The ADAPT system has been successfully used, such as in rainfall-runoff and stream profiles studies in Ohio, and to study the effects of alternative land development patterns on the storm and sanitary sewer system in Wyoming, USA.

4.5.2 FORDATA

The DBMS FORDATA has been developed by CSIRO,Canberra, modelled on the CODASYL 1971 proposals. FORDATA which has been implmented on CDC Cyber computers has network data base structure and uses Fortran language for data manipulation by users. The data base comprises a schema, subschema and records. Accessing of records is done by keys for indexed sequential or random access modes. Since 1974 FORDATA has been in increasing use in Australia for various types of land, hydrological and environmental data processing and storage.

4.5.3 IMS

The information Management System (IMS) of IBM is a multi-hierarchical data base organisation. It engages host-language system accessed by CALL, and has a fully developed query processor and recovery support. The data description language DL/1, especially developed for IMS is a significant software advantage of this DBMS. IMS is used at the Hydrological Data Bank of Bavaria, FRG, on an IBM/3033 computer. The IMS is supported by a schema, and uses PL/1 programming language. The IMS data base has been used exhaustively in this data bank to serve as a storage and retrieval system for surface water groundwater, water quality and precipitation data.

4.5.4 System 2000

This DBMS, commercially available from MRI System, Comp., USA uses a hierarchical data base structure. It has been used by the US Geol. Sur. for maintaining its Water Data Sources Directory and Master Water Data Index Data Bases. The data base schema defines relevant specifications about the data, which are represented by a system of alphanumeric codes, and components accessed by keys. System 2000, which has a developed query processor with easy query language for data retrieval, has found extensive use in NAWDEX data bases.

It is evident from the foregoing discussions that the design and/or adoption of a particular DBMS plays a significant role in the utility of a data library. Essential to the implementation of a DBMS is the schema. Figure 4 illustrates the tasks underlying the evolution of a suitable schema (moore et al.,1981). It involves two major step,-

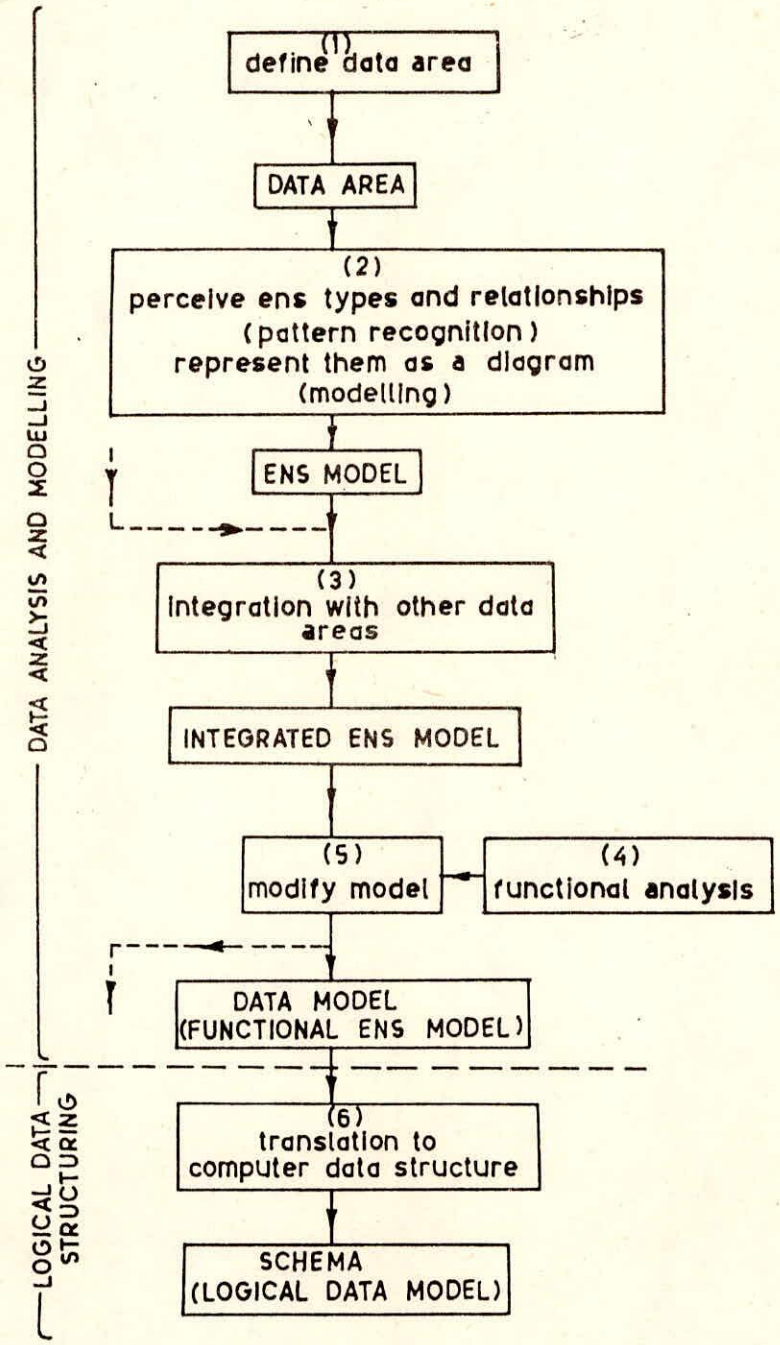

FIGURE 4 - DATA ANALYSIS AND MODELLING FOR A SCHEMA

the design of the information structure that models the generated data, and its transformation to a logical data structure (schema) that is compatible with a given DBMS. The former is a pattern recognising and mathematical analysing process, while the latter requires knowledge of the DBMS software very often, the evolution of the schema is a continuing and iterative process necessary for the development of the data base in response to the growing needs of the data library.

5.0 CASE STUDIES OF HYDROLOGIC DATA STORAGE AND RETRIEVAL SYSTEMS

5.1 National Water Data Storage and Retrieval System ,USA

The nationwide US Hydrological data base WATSTORE is operated by the Water Resources Divison (WRD) of the U.S. Geological Survey. Its data are available through OWDC and NAWDEX which also supply data of sister data banks such as STORET. The WATSTORE system is also available through 62 remote terminals in non-USGS offices. The computer centre division of USGS provides for the entire data processing needs of WATSTORE, which are conducted by data processing specialists of the WRD.

The WATSTORE system consists by one informational and four data files, viz.,

i) Station Header File containing pertinent information on the identification, location and physical description of over 200 000 sites for which data are stored in WATSTORE

ii) Daily Values File containing about 200 million water data parameter (daily values on streamflow, river stage, reservoir content, water temperature, specific conductance, sediment concentration and transport rate, and groundwater levels.

iii)Water Quality File consisting of about 1.4 million values on chemical, physical, biological and radio-chemical proportion of surface water and groundwater.

iv) Peak Flow File comprising of around 400 000 annual maximum flow and gauge height measurements that are used in flood frequency

analysis (with Log-Pearson Type III frequency distribution).

v) Groundwater Site/Inventory File containing geographical, geohydrological and other pertinent types of information of about 700 000 sites of wells, springs and other sources of groundwater.

The data measurement and collection procedure is highly systematized. Digital recorders are userd at many locations to record river stages, conductivity, water temperature, turbidity, wind direction and chemical contents. Field data are recorded paper tapes and transmitted to the data bank via telephone lines. Water quality data are sampled and analysed with uniform techniques by NASQUAN at 525 stations. Two central laboratories analysing about 150 000 water samples per year also contribute information to WATSTORE.

Quality control of data involves routine checking by computer programmes from an auxiliary file to ensure an acceptable level of consistency. Secondary processing involves computation of areal estimates, parametric models, statistical analysis and graphical plots. Real-time data processing has been in operation as part of a satellite data collection testing programmes since 1972.

In order to reduce high costs involving on-line disk files, only current data (about 20% of the whole) are held on disks while historical data are maintained an magnetic tapes. Among the five files, the ground water site inventory file uses the SYSTEM 2000 data base management system. Application programmes allow for retrieval of data from current and/or historical records. Data can be output as computer printouts, digital plots, or on punch cards or magnetic tapes. Current research in WATSTORE is predominantly towards the development of fully automated hydrological data base, standardization of data collection procedures, and real-time data processing.

5.2 National Water Quality Data System, Canada

The National Water Quality Data System (NAQUADAT), Canada is a water quality data base set up in 1969 by the Water Quality Branch, Inland Waters Directorate, Environment Canada. The data are collected by water quality networks through five regional offices of the Water Quality Branch in Canada. The data base system of NAQUADAT is quite efficient and has been implemented in other countries too.

The system comprises three interrelated files:

i) Station file contains description of sampling locations and includes geographical coordinates of about 6000 stations. Each station is assigned a numeric code which reflects the type of water and its location in the Canadian basin sub-basin scheme.

ii) Dictionary file contains information and description of about 750 analytical procedures. Each method is assigned a parameter code as well as other necessary coded information.

iii) Data file: The actual analytical results are stored in the data files with encoded parameters for pesticides, hydrocarbons, dissolved nutrients, bacteriologic, inorganic and orgnic classification, mean daily and monthly flows, etc.

NAQUADAT operates commercially with time-sharing facilities on IBM computers and with interactive software programmes in COBOL MARK IV and FORTRAN languages. The software devices are used to update, search and retrieve information. Considerable flexibility is allowed in specifying the information to be retrieved. Browsing and

interactive reporting are facilitated by two time-sharing utility programmes STS: WYLBUR and STS: INFO. Limited statistic are generated in the reporting programmes. Elaborate statistical analysis is achieved via integrated statistical packages such as SPSS or BMDP. The data can be retrieved as computer outputs, microfiche, digital plots and standard magnetic tape output. Current developmental efforts are devoted to interactive graphical presentation and overall efficiency of operation.

5.3 Hydrological Data Bank of Western Australia

The Australian Water Resources Council (AWRC) coordinates on a national scale different hydrological data banks in different regions of Australia. The data are primarily surface and groundwater measurements, rainfall, and water quality assessment. While the first three types of data are collected on the field by a limited number of government agencies, water quality information is collected via about 200 sundry organisations.

The hydrological data bank of the State of Western Australia is a prominent member of the AWRC data bank. The data bank is held under a CDC cyber 1972 computer with eleven disk drives and five tape drives. In addition a Xynetics flat-bed plotter, a CMC Key to disk system, a PDP 11/34 mini computer eight VDU's, and a Tektronix 4014 graphics display unit are used.

In its technical organisation and operations the data bank makes a fundamental distinction between continuously and non-continuously recorded information. The former category includes stream levels, groundwater rest levels (at observation bore holes), reservoir levels and water quality. Standardized measurement systems have been

adopted for such data. Graphical records are initially processed by field personnel, which is not required for automated digital records. Further checks, quality control and interpretation are provided by hydrologic experts and application programmes in the data bank. For each gauging station rating curves have been developed to make routine computations of daily flow volumes and peak flow values. Annual and average flow duration curves are also calculated. Analogue to digital conversions are checked from time to time by computer plotting of data into over lays and comparing with original charts.

Non-continuously recorded information consists primarily of water quality data and ground water rest levels. All water quality data are subjected to the same series of checking and updating programmes, results of field and laboratory analyses are recomputed, relevant correlations established and tests for range performed Further processing concerns research and investigation which are not part of routine work.

Comprehensive computer based system have been developed for the storage and retrievals of different types of data principally, rainfall and river level data are held, conisting of files containing time, value and quality information for each variable at each station. Most of the continuously recorded information are held in duplicate on magnetic disk and tape respectively. Old and little-accessed information are maintained solely in a magnetic tape-based archive. Non-continuous data are held entirely on disks.

Retrieval of data can be affected by interactive application programmes mostly written in FORTRAN IV language. For data analysis and processing a standarized working file structure has been developed for easy data manipulation. Data can be output in standard formats

on computer printouts, magnetic tapes and microfiche which fulfil most users'needs. Current developmental activities are in the field of hardware reconfiguration, the information system, and more efficient software for data-processing and retrieval.

5.4 Water Resource Data Bank of Bavaria, FRG.

The data bank of Bavaria was formed in 1976 by the Bavarian Ministry of Interior whose State Water Resource Office operates the hydrologic network through its local offices. Hydrometric data are obtained through the Federal Weather Service. The various data collected by this organisation pertain to quantitative and qualitative measurements-temperature and sediment load of surface water, ground-water, and precipitation.

The Information Mnagement Systems (IMS) serves as a data base management of the Bavarian Data Bank. Most of the data are available in analogue form (chronological graphs rating curves, etc.). These are digitized, recorded on magnetic tapes and transferred from mini-computers to the main IBM/3033 computer for processing and storage using PL-1 programming language.

Automatic quality control and manual checks for missing data, and regression analyses are employed. The parameters processed are water data levels, precipitation, humidity, temperature, radiation, pH, conductivity and turbidity.

The data base of the IMS storage and retrieval system is a hierarchical structure of data elements related by their logical dependencies. The hierarchical structure and data formats are described in the data base. The data base is supported by the data communications service which allows direct access to data files through

appropriate on-line terminals. Presently efforts are devoted to the development of real-time water quality collection of all inland waters.

5.5 Hydrological Data Bank of NVE, Norway

The hydrological data bank at the Norwegian Water Resources and Electricity Board (NVE) in Oslo is a centralized bank for surface water, and groundwater quality. The increasing hydropower developmental activities in Norway led to large-scale computerized data storage and retrieval facilities at NVE in 1970.

The data collected consist of groundwater stage in wells, stage and discharge in rivers and lakes, mass balance and meteorological measurements in selected glaciers, ice conditions, water temperature and sediment transport in glacier streams, urban runoff and precipitation, and water quality and tidal changes in fjords. The bulk of the incoming data is punched onto magnetic tapes, while analogue information is digitized manually or by curve tracers.

Quiality control and data processing involve manual correction for ice damming, spatio-temporal correlation and regression analyses, time-series analyses, precipitate-runoff models, stage-discharge rating curves, and simulation runs. The data are processed and handled through a CDC Cyber 171 computer with disk, magnetic tape. card reader and printer equipment.

The processed and regularly updated data are stored in several files-primarily direct or sequential access file on hard disk, or tapes. Additional information pertaining to the location, time, quality, environmental conditions and government regulations are also stored in the files.

The bulk of the data processing, is done in remote batch mode,

while retrieval and analysis are primarily by interactive programmes, usually written in FORTRAN language. The data can be reproduced on magnetic tapes and as computer printouts. Further development of the data bank is concerned with more elaborate and cost-efficient data collection network, and quality control measures.

6.0 CONCLUSION

Since inception, the National Institute of Hydrology has taken up the study and modelling in various areas of theoretical and applied hydrolody, as a consequence of which a vast amount of information has been assembled over the years. At present these are stored on magnetic tapes, computer printouts and manually compiled documents. It is necessary that these information, and more that is likely to be collected in future, be compiled and stored in a computer-retrievable library for ready use in forthcoming researches.

Of especial concern in setting up a data library are the types of data that need to be stored, the processing of quality control measures required, the intended/predictable uses of the data, and the software facilities available. As yet, data have been procured for the UGC command area Narmada, Hindon and Godavari basins, Damodar and Bhakra-Beas reservoir systems, etc. Some of these data are confidential and accessible to selected personnel only; some are restricted for use within the institute; the remaining information do not have any restriction on their use.

The data types of primary concern are hydrologic, geologic and meteorological. A typical example illustrates the types of data and data-processing necessary in hydrological studies. The Narmada basin data include rainfall (daily and hourly) flow (Hourly gauge, daily gauge and discharge), and river cross-section profiles. These data were procuredfrom the NBDDC (Baroda), IMD(Pune) and the M.P.Govt. Characteristic of the varied type of data measurement in vogue, wide

divergence was noticed in the quality of different sets of data. For instance, the NBDDC discharge data were measured with float type current meters. Some of their data were seen to be beyond the instrument ranges. Graphical plots of hourly stage data revealed gross abberational and necessitated corrective measures. Similar measures were needed for rainfall data also.

Certain data quality improvement measures are already in use in this Institute such as graphical curve-fitting. However, an exhaustive library of standardized programmes need to be developed for the regular processing of data. These include tests for:

i) Internal consistency of data and related parameters;

ii) Physical consistency of data and related parameters in space and time;

iii) Time series analysis;

iv) Correlation and regression analyses;

v) Data compression by deleting redundant values that can be approximately interpolated between neighbouring values.

The addition to routine computer processing, manual efforts are necessary to increase data fidelity and for corrections.

The same hydrologic information may be required in a number of studies. For example, the same precipitation data may be required for rainfall-runoff studies, groundwater analysis, and watershed modelling of a region. Hence these data must be stored in shareable files on computer-readable devices. Data that are infrequently used may be archieved on magnetic tapes, but data in regular and multi-use must be available on on-line disks. The latter task underscores the particular need of a DBMS in the institute. The institute also envisages the setting up of an experimental hydro-meteorological station for on-line hydrological studies. The need for multi-purpose

station for on-line hydrological studies. The need for multi-purpose on-line data processing also involves an efficient data base management system.

In using a computerized DBMS for hydrologic data library, the software facilities of the VAX-11/780 computer system need to be utilized. The record management services(RMS) of VAX-11/780 computer provides three file organisation-sequential, indexed and relative. The record access modes supported by the RMS are random, Keyed, sequential, dyamic and RFA(records and file address). Since file organisation on magnetic tapes can only be sequential, they are not suited for interactive DBMS. The indexed file organisation uses keys for programme-controlled location of records and their retrieval from disks. The sequential and relative file organisation can be processed using native and compatibility programming languages, but the indexed file organisation cannot be processed by the former language. Thus a careful choice of the file organisation in the DBMS is necessary.

The various procedures discussed in this report need exhaustive study for designing a hydrologic data library. The comprehensive study of hydrological processes and data processing techniques is that first step in organising an information base. This includes the recent yet rapidly developing area of real time data processing which is increasingly used in the operation of hydrological systems . In addition, knowledge of data base management systems and their applications in hydrologic areas(such as the IMS in the Hydrologic Data Bank of Bavaria, FRG, and the SYSTEM 2000 in NAWDEX, USA) is necessary.

REFERENCES

1. Albertson, M.L. et al.(eds) (1971),'Treatise on Urban Water Resources Systems', Colorado State University, Fort Collins, Colorado.

2. Brown, R.H. et al.(eds) (1972), 'Ground Water Studies',Unesco Paris, Chaps.7 and 8.

3. C.W.R.D.M. (1983),'Hydrometeorological Data Bank', Surface Water Division, Centre for Water Resources Development and Management, Kozhikode, Kerala.

4. Digital Equipment Corporation, (1978), 'VAX-11 Software Handbook', Massachusetts.

5. Flores, Ivan (1980), 'Data Base Architecture', Van Nostrand Reinhold Co., New York.

6. I.H. (1981), 'The Processing of Hydrological Data', Institute of Hydrology, Wallingford, U.K.

7. Males, R.et al.(1980),'Application of the ADAPT Geo-Based Modelling System in Urban and Regional Water Pollution Management Proc.IFAC Symposium, Haimes, Y. and J. Kindler (eds) (1980), 'Water and Related Land Resources Systems', Pergaman Press.

8. Moore, A.W. et al (eds) (1981), 'Information Systems for Soil and Related Data', Centre for Agricultural Publishing and Documentation,Wageningen, Netherlands.

9. Moss, M.E.(1979),'Some Basic Considerations in the Design of Hydrologic Data Networks', Water Resources Research, Vol.15, No.6.

10. Poinke, M.B. and R.L.Kleckner (1979),'Hydrologic Data', ASA-CSSA CSSA-SSSA, Madoson(USA).

11. Sinha, BPC, et al.,(1983), ' Development of Ground Water Data Storage and Retrieval System', Proc. Vol.I, Seminar, CGWB, New Delhi.

12. US Army Corps of Engineers,' Resources Information and Analysis', Hydrologic Engineering Centre, California, Sept. 1979.

13. U.S. Department of Interior, Federal Advisory Committee on Water Data (1971), 'Design Characteristics for a National System to Stone,Retrieve & Disseminate Water Data,USGS, Washington, D.C.

14. US Department of Interior (1977),'National Handbook of Recommended Methods for Water Data Acquisition', Office of Water Data Coordination, Reston, Virginia.

15. William, O.O. and W.A. Knecht (1981), 'NAWDES System 2000 Data Retrieval Manual', USGS, Open File Report, 81-419,Reston.

16. World Met. Organisation (1969),' Automatic Data Handling and Processing for Climatological Purposes', WMO, Tech.Note No.100, Geneva.

17. WMO (1981), 'Case Studies of National Hydrological Data Banks', Operational Hydrology Report No.17, WMO-576, Geneva.

18. WMO (1981), 'Guide on the Global Data Processing Systems', WMO No.305, Geneva.

ANNEXURE I

Sample Output From a NAWDEX Directory File

NATIONAL WATER DATA EXCHANGE
WATER DATA SOURCES DIRECTORY

NAME OF ORGANIZATION	ORGANIZATION CODE	TYPE OF ORG.	TYPE OF ORIENTATION	NAWDEX MEMBER
NO NAME WATER AGENCY	VA999	O	E	YES

THIS ORGANIZATION COLLECTS THE FOLLOWING TYPES OF WATER DATA FOR THE SPECIFIED GEOGRAPHIC AREAS:

STATE, TTRY, CNTRY	SURFACE WATER QUANTITY	SURFACE WATER QUALITY	GROUND WATER QUALITY	GROUND WATER LEVELS	GROUND WATER PUMPAGE	GEOLOGIC DESCRIPTIONS
051	42	76*	21*	110	16	110

DATA AVAILABLE FROM THIS ORGANIZATION MAY BE OBTAINED FROM THE FOLLOWING LOCATIONS:

OFFICE CODE	OFFICE ADDRESS	RESPONDS TO PUBLIC REQUESTS	STORAGE MEDIA	AREAL COVERAGE
5101	CHIEF, DATA OPERATIONS NONAME WATER AGENCY 12201 SUNRISE VALLEY DRIVE RESTON VA 22092 TELEPHONE: 703-888-9999	YES	D	S

THE ABOVE OFFICE IS A NAWDEX ASSISTANCE CENTER:

TELEPHONE: 703-888-9999 TIME
OFFICE HOURS: M-F 8.00-4.30 EASTERN

COMMENTS:

ALSO HAS SPECIAL STUDIES ON SELECTED LAKES AVAILABLE. CHARGES ARE MADE FOR DATA. COST ESTIMATES PROVIDED ON REQUEST.

THE ABOVE OFFICE HAS DATA AVAILABLE FOR THE FOLLOWING GEOGRAPHIC AREAS:

STATE, TTRY, CNTRY	COUNTIES
051	003, 013, 059, 067, 121, 171, 199

DATA AVAILABLE FROM THIS ORGANIZATION ARE ALSO AVAILABLE FROM THE FOLLOWING OTHER SOURCES:

ORGANIZATION CODE	NAME OF ORGANIZATION	ALT/ PRF	DATA AVAILABLE	STORAGE MEDIA
USGS	CHIEF, USER SERVICES UNIT NATIONAL WATER DATA EXCHANGE USGS, 421 NATIONAL CENTER RESTON VA 22092	A	SQGLD	D

COMMENTS:

NAWDEX HAS LIMITED ACCESS TO THE DATA FILES OF THE AGENCY.

ANNEXURE II

Program to extend a FORDATA Data base

```
PROGRAM CREATE (INPUT,OUTPUT,TAPE5=INPUT)
        DIMENSION CARD(8)
        DATA(KERROR=Ø)
C NOMINATE THE SCHEMA LFN
        INVOKE SCHEMA 'SURVEY'.
        OPEN AREA 'SOIL', USAGE-MODE IS UPDATE.
C ZSTATUS IS SET BY FORDATA AND INDICATES THE RESULT OF EACH
        IF(ZSTATUS.NE.' )GOTO 41                                    COMMAND
1       READ (5,1ØØ)KEY,CARD
1ØØ      FORMAT(I1,7A1Ø,A9)
         IF(EØF(5)) GOTO 51
C KEY=1, SITE DATA; KEY=2, LAYER DATA
         IF(KEY.EQ.1)GOTO 11
         IF (KEY.EQ.2.AND.KERROR.EQ.Ø) 21,31
C STORE SITE DATA
11        KERROR=Ø
          CALL ERRSET (KOUNT,1ØØ)
          DECODE (36,1Ø1,CARD) TRAVERSE,SITENO,LANDFRM,
          TOP,DRAIN,LANDUSE,VEG
1Ø1       FORMAT(7(1X,I5)
C KOUNT ".Ø INDICATES ERROR ON DATA CARD
          IF(KOUNT.NE.Ø) GOTO 31
          STORE'SITE'
          IF (ZSTATUS.EQ.Ø) 1,31
C STORE LAYER DATA
21        CALL ERRSET (KOUNT,1ØØ)
          DECODE(37,1Ø2,CARD)TOP,BOTTOM,COLOUR
          TEXTURE,STONE,ALKALI,PH
1Ø2       FORMAT (2(1X,F4.2),1X,A11,2(1X,I2),1X,I4,1X,F3.2)
          IF(KOUNT.NE.0) GOTO  31
          STORE'LAYER'
          IF (ZSTATUS.EQ.0) GOTO  1
31         CALL ERROR(KEY,KOUNT,KERROR....)
C IF KERROR=1 SKIP TO NEXT SITE DATA
           KERROR=1
           GOTO 1
41          CALL ERROR(.....                    )
C AREA MUST BE CLOSED TO FLUSH BUFFERS
51         CLOSE AREA 'SOIL'
           END
```