UNITED STATES

DEPARTMENT OF THE INTERIOR

GEOLOGICAL SURVEY

VERIFICATION PLAN FOR HYDROLOGIC DATA

By Clyde W. Alexander

U.S. Geological Survey

Open-File Report 82-374

Portland, Oregon

1982

UNITED STATES DEPARTMENT OF THE INTERIOR

JAMES G. WATT, Secretary

GEOLOGICAL SURVEY

Dallas L. Peck, Director

For additional information write to:

U.S. GEOLOGICAL SURVEY
847 NE. 19th Ave., Suite 300
Portland, Oregon 97232

CONTENTS

ILLUSTRATIONS

VERIFICATION PLAN FOR HYDROLOGIC DATA

By Clyde W. Alexander

ABSTRACT

Water-data users now have access to computerized data files containing unpublished, unverified hydrologic data. Quality control of hydrologic data can be performed by computerized data-verification routines before data are stored in user-accessible files.

A single, unified concept describing a master data-verification program with multiple special-purpose subroutines, and a screen file containing verification criteria, can probably be adapted to any type and size of computer-processing system. Some traditional manual-verification procedures can be adapted for computerized verification, but new procedures can also be developed that would take advantage of the powerful statistical tools and data-handling procedures available to the computer. Prototype data-verification systems should be developed for all data-processing environments as soon as possible. The WATSTORE system probably affords the greatest opportunity for long-range research and testing of new verification subroutines.

MANAGEMENT OVERVIEW

Data verification refers to the performance of quality control on hydrologic data that have been retrieved from the field and are being prepared for dissemination to water-data users. The traditional high standards of accuracy maintained for published data have been accomplished using manual methods of verification. Recent developments in the field of computerized data processing and automated data-retrieval systems have accelerated rapidly. Water-data users now have access to computerized data files containing unpublished unverified hydrologic data. Therefore it is necessary to develop techniques and systems whereby the computer can perform some data-verification functions before the data are stored in user-accessible files.

The U.S. Geological Survey uses three types of computer systems to process and disseminate hydrologic data: (1) WATSTORE (National Water Data Storage and Retrieval System), that contains the historical data and is used for processing data in non-real time; (2) HYDRECS (Hydrologic Data Real-Time Computer-Processing System), that is the central real-time system designed to accept data transmitted automatically from field sites via Earth-orbiting communications satellites; and (3) local processing, or distributive processing, that is done on small computers at the District level and can be designed to operate in

either real time or non-real time.

Computer programs can be developed that will perform data verification in all three of the above systems. A single unified concept describing a master data-verification program, using multiple special-purpose subroutines, and a screen file containing verification criteria can probably be adapted to any type and size of computer-processing system.

Some of the traditional manual-verification procedures can be adapted for computerized verification, but new procedures can also be developed that would take advantage of the powerful statistical tools and handling procedures available through the computer.

Prototype data-verification programs should be developed for all three data-processing systems as soon as possible. The WATSTORE system probably affords the greatest opportunity for long-range research and testing of new verification subroutines.

The concepts and suggestions presented are derived, in part, from an intensive interagency study of data-verification requirements intended for application to CROHMS (Columbia River Operational Hydromet Management System). Some of the suggested data-verification techniques have been tentatively explored by the Northwest Water Resources Data Center, Portland, Oreg., but none of them has been implemented or tested in an on-line real-time environment.

INTRODUCTION

Hydrologic-data collection, compilation, and dissemination activities of the U.S. Geological Survey include various quality-control procedures designed to ensure the scientific validity of the data. Quality control begins with the operation and maintenance of the data-collection facility in the field. It includes periodic inspection and preventive maintenance of sensors, recorders, and related equipment and observation of possible changes in hydrologic and hydraulic conditions at the field site. Data verification refers to quality-control procedures designed to detect errors in data that have been retrieved from the field and are being processed and prepared for dissemination to water-data users. Manual processing and computation of retrieved data includes data verification at each step. Many of the manual computation procedures have been replaced by automatic data-processing procedures. Automatic data processing is faster and allows the hydrologist more time for data analysis and interpretation, but it also eliminates some of the manually performed data-verification procedures.

Automated data-relay systems result in further isolation of the hydrologist from the data-processing procedure. Data from remote field sites can now be automatically retrieved, processed,

and disseminated to water-data users in near-real time. This eliminates the manual verification of hydrologic data prior to dissemination. Quality control must be done by the computer as an integral part of the data-processing procedure.

The purpose of this paper is to (1) describe some of the essential elements of a computerized data-verification system and their application to real-time and non-real-time data-processing systems and (2) formulate a preliminary plan for the development and implementation of data verification within the Geological Survey.

OVERVIEW OF DATA-PROCESSING SYSTEMS

Most of the hydrologic data collected by the Geological Survey are recorded by analog-to-digital recorders on 16-channel paper tape. The tapes are physically retrieved by hydrographers and returned to the field office. The standard data-processing procedure in use at the present time (1981) is to insert the paper tapes into a tape reader in the field ofice and transmit the data by telephone lines to the Automatic Data Processing Section of the Geological Survey in Reston, Va., where the data are recorded on magnetic tapes and processed into WATSTORE. Many different data-processing programs resident in WATSTORE are used by field personnel to analyze and prepare the data for publication in various reports.

An alternative to manual retrieval of data from the field site is to use an automated data-relay system. This involves electronic transmission of data from field sites to a central receive site and data-processing facility. The communications network is typically either land based, using VHF (very high frequency) or microwave frequencies, or dependent on the use of Earth-orbiting satellites for data relay. The Geological Survey has implemented a satellite data-relay system using the NOAA/NESS /GOES (National Oceanic and Atmospheric Administration/National Environmental Sciences Service/Geostationary Earth Orbiting Satellite). Data are transmitted from the field at intervals during the day and relayed through a receive site into HYDRECS. The data-communication equipment in the field is generally driven by standard digital recorder equipment which also records the data on paper tape.

The HYDRECS system resides in the Geological Survey Honeywell—/ computer in Reston. HYDRECS receives data from two

---

/ The use of brand names in this report is for identification purposes only and does not imply endorsement by the U.S. Geological Survey.

minicomputers. One is operated by the Satellite Data Relay Group in Reston and relays data from the NOAA/NESS receive site. The other one is operated by the COMSAT General Corporation and relays data from the COMSAT receive site. After the data have been processed into the HYDRECS data base, they can be accessed by registered users of HYDRECS. Data can also be automatically transferred from HYDRECS into WATSTORE for final processing and storage.

Local processing is a relatively new concept coming into use by the Geological Survey. It involves the use of minicomputers located in the separate District offices to perform the data-processing functions for each District. After the data have been processed and edited, the data are transmitted to the central computer in Reston for storage in the WATSTORE data base. Retrieval of field data for local processing can be done either manually or by use of automated retrieval procedures. Manual retrieval involves the removal of punched paper-tape records from the field station by hydrographers and reading the tapes into the local minicomputer for processing instead of transmitting them directly to Reston. Automated retrieval involves the use of local antennas designed to receive data from the GOES satellite system. This kind of installation will be a full-fledged satellite data-relay receive site and will allow the Districts to collect and disseminate data in near-real time.

A computerized data-collection network (CROHMS) has been established in the Pacific Northwest to serve the needs of the Columbia River Water Management Group agencies. CROHMS was designed and implemented under the auspices of an interagency committee (Hydromet Data Committee) that functions as a standing technical committee of the Columbia River Water Management Group. The U.S. Army Corps of Engineers owns and operates the CROHMS central computer system. The central computer receives a continuous stream of data from remote minicomputers and teletype terminals owned and operated by many different agencies in the Columbia River basin. Hydroclimatic data, including streamflow, rainfall, air temperature, snow depth, and water content of the snowpack are retrieved from several hundred remote field sites. Retrieval methods range from manual observations to automated retrieval under minicomputer control. Data communications used for automated retrieval include radio frequency, microwave, and Geostationary Earth Orbiting Satellite (GOES). Varying degrees of processing are performed on the data both before and after they are transmitted to the CROHMS central computer. Most of the streamflow data in CROHMS are obtained from Geological Survey gaging stations. Data communication and telemetry equipment installed in the gage houses is generally owned and maintained by the CROHMS agency which retrieves the data for operational purposes.

NEED FOR DATA VERIFICATION

The foregoing discussion of present and planned data-processing systems indicates two general data-processing environments, real time and non-real time, within which automated data verification should be performed. Data resident in the WATSTORE system are directly accessible by many Geological Survey cooperators. Data entered into WATSTORE by the standard non-real-time, or batch, process may contain errors until the responsible field office has reviewed the data and submitted corrections to the file. An automated data-verification system would serve the twofold purpose of (1) alerting the cooperator of the occurrence of erroneous or suspect data and (2) improving the efficiency of the review and correction procedures used by the responsible field office. The same rationale applies to local processing; efficient error detection at the local level will ensure that accurate data are available for storage in WATSTORE or for direct release to cooperators with less time lag between data collection and data dissemination.

Automated data-collection networks may not always be combined with automated real-time data dissemination, especially at the local level. In that event, data processing and data verification can be performed in batch mode. Generally, however, the justification for installing a real-time data-collection system is to provide the capability for dissemination of real-time data to the water-data-user community. This requires that data verification be automated and done in real time or near-real time.

It would be difficult and beyond the scope of this paper to assess the impact of faulty data on all the various data users. The development of automated data-collection dissemination procedures has been in direct response to the expressed needs of water-resource managers. Increasing demand for available water supplies has led to more intensive long- and short-term planning studies. Water-resource managers are using current real-time or near-real-time data as a basis for frequent management decisions. Such decisions based on faulty data can, in extreme cases, result in inefficient use of available water resources, economic losses, and other severe consequences.

Lystrom (1972) made a study of the potential magnitude and frequency of errors in real-time data, using historical daily mean discharge data and field information as a basis for the analysis. In a detailed study of 14 selected stations in the Pacific Northwest, he concluded that errors of 10 percent or greater could be expected for approximately 25 days per station year in a real-time environment. Lystrom's study did not include errors associated with data-transmission and data-communication interfacing devices.

Potential errors in hydrologic data can be categorized

according to their source as follows:

1. Sensing and recording equipment malfunction. Sensor errors relate to gaging-station equipment, such as intake pipes at stilling wells, gas-pressure lines and orifices at servo-manometer installations, and temperature and other water-quality-monitoring probes. All these devices are susceptible to being plugged, displaced, broken, buried, or otherwise rendered inoperable. Recorder errors relate to malfunctions of equipment installed in the gage house. Stilling-well floats can sink or hang up, and float tapes and drive chains can break or become dislodged from proper positioning on drive wheels. Loss of electrical power resulting from dead batteries or failure of other power sources disables the entire system. Clock stoppage, resulting from power loss or mechanical failure, is a common cause of lost data. Water-quality monitors are subject to a variety of problems caused by mechanical and electrical malfunctions unique to that kind of equipment.

2. Data-communications errors. Data communication begins at the gaging station and includes (1) the interface between the recorder and the DCP (data-collection platform); (2) the transmission path through space; (3) repeaters or communications satellites; (4) receive-site equipment; and (5) the interface between the computer system and the receive site. Failure or sporadic malfunction of any part of the data-communication system can introduce errors or cause loss of data.

3. Changes in the stage-discharge relation. Changes in the channel geometry at a gaging station generally change the stage-discharge relation. The stage-discharge relation may also be affected by variable factors other than channel geometry. The formation of ice in the stream channel is a common problem in many area. Floating trees and other debris may become lodged in the channel in such a way as to affect the rating temporarily or even permanently. The growth of moss and other plants on the streambed and along the banks is also a factor. These conditions do not affect the validity of the stage data recorded or transmitted from the site; however, until the stage-discharge rating change is detected and implemented, the computed discharges will be in error.

Errors in the data, from whatever source, result in one of two basic situations. Either the integrity of the raw data is destroyed, or the integrity of the processing and interpretation of the data is jeopardized or destroyed. Garbled data caused by communication errors results in lost or unusable data unless the error is detected and the data can be retransmitted or recovered from backup records. Sensor and recorder errors are responsible for most instances of complete loss of valid data.

OVERVIEW OF PRESENT DATA-VERIFICATION PRACTICES

Prior to the age of computer processing, hydrologic data were subjected to a manual and subjective scrutiny at all phases of computation. Analog recorder charts were used as a basis for computing daily mean discharge. The continuous pen tracing provided a vivid, highly informative visual record of the performance of the recorder and the hydraulic conditions in the stream. Errors due to sensor and recorder malfunctions were generally identifiable. Temporary backwater conditions due to ice or debris in the channel were recorded and identifiable by examination of the characteristics of the record. To the extent that analog strip-recorder charts have been replaced by digital recorders, subjective verification of a continuous trace of river stage is no longer feasible. The punched paper tapes do not lend themselves to visual interpretation of the record. The tabulation of hourly gage-height data that results from processing digital tapes through WATSTORE can be examined visually for obvious gross errors but does not yield the same amount of information available from a strip chart. This has resulted in an increased emphasis on verifying processed daily discharges during preparation of data for publication. This problem is being corrected in many districts with local processing capability; programs have been developed that prepare plots of unit values read from digital tapes. Plots of this kind are probably the most important non-real-time verification tools available and should be incorporated into all such systems.

Manual data verification is highly subjective. Personal experience and knowledge of the local hydroclimatic environment are important. There are, however, a number of generally recognized graphical and statistical tools that can be used. The discharge hydrograph is probably one of the more useful and widely used graphical devices. A comparison of discharge hydrographs for two or more hydrologically similar streams provides the analyst with much useful information. Supplementary plots of rainfall, air temperature, and other hydroclimatic parameters are often used in the analysis. Some of the more objective tools commonly used for verification include statistical analysis of historical records. Regression analysis and other statistical techniques can be developed ranging from comparisons among watersheds to rather complex basin models. If there are several streamflow stations on a given stream, then flow routing or summation procedures, or both, can be used. A detailed survey of how extensively these procedures are used in various regions of the country is beyond the scope of this report. Personal experience and informal contacts indicate that complex, statistically oriented techniques are used for special studies but are not often applied to routine data-verification procedures.

Computerized verification of non-real-time data has not been well developed at the national level. Program E659 in the

WATSTORE system performs the computations on all types of digital-recorder data and is the primary input source to the WATSTORE Daily Values File. The data-verification capabilities of E659 are limited to detection of missing data points and detection of sequential pairs of punch readings that differ by more than a specified allowable test difference (rate-of-change check). Error flags are assigned to faulty data at processing time and printed on the primary computation-sheet output from the program. Because the error flags are not stored in the data base, they are not available to water-data users.

WATSTORE programs that process data in the Water-Quality File include more data-verification capability than is available in E659. For example, Program K441 includes three categories of data checking: (1) An alert system designed to identify water-quality data that exceed acceptable or legal limits; (2) chemical-logic tests designed to ensure mathematical consistency in the analysis of specific chemical constituents; and (3) consistency tests that check for unreasonable data, erroneous parameter codes, and so forth. In addition to these three checks that are applied automatically, an optional routine is available for user-supplied criteria that will perform statistical tests, regression-analysis checks, and data-range checks.

Computerized verification of real-time data is largely nonexistent in the HYDRECS system. HYDRECS does contain a set of screen codes that can be attached to erroneous data. However, at the present time these screen codes are used only to flag data-communication errors, such as parity errors and character errors that have been detected at the receive site by the signal-monitoring equipment. Data received by HYDRECS from NOAA/NESS (National Oceanic and Atmospheric Administration/National Environmental Sciences Service) is in ASCII (American National Standard Code for Information Interchange) and is converted to engineering units by HYDRECS. The NOAA/NESS system also sends to HYDRECS various messages in text format concerning each DCP. Text messages include information regarding poor transmission-signal characteristics and DCP's that are transmitting outside their assigned time slots.

A second satellite data-relay system operating at the national level is presently being tested under contract with COMSAT General Corp. COMSAT utilizes the GOES satellite as the communication relay, but operates its own receive-site and data-processing facility. Data received by HYDRECS from COMSAT have been converted from ASCII code to engineering units (gage heights) and subjected to some error analysis. The error analysis done by COMSAT consists of upper and lower exceedance limits, instrument dead band, and rate-of-change checks. Error codes are sent to HYDRECS along with the data and assigned specific screen code values.

Several districts are operating in a local processing mode

using minicomputers or microcomputers to translate paper tapes and perform various amounts of processing. Some districts (for example, California and New Jersey) are doing the data processing locally and transmitting verified daily values to WATSTORE. Other districts are translating and editing the tapes locally but use the WATSTORE system for processing. Data verification is being done manually in most places, aided by interactive programs for applying time and datum corrections to the unit values. Computers are being used in some places to generate plots of unit values, in a format similar to the analog recorder strip chart, for visual examination. The New Jersey District is incorporating a rate-of-change check and a no-change check into its processing system, which is probably the most highly developed non-real-time local processing system operating at the present time.

The CROHMS system does not yet include an operational data-verification capability. The Geological Survey agreed to participate in the development of the CROHMS data-verification system, and the Northwest Water Resources Data Center has been the participating Geological Survey project office. An interagency work group developed a conceptual design for a real-time data-verification system for CROHMS, but it has not been implemented because of a lack of computer resources in the CROHMS system. Data-verification development will proceed when the system is upgraded, probably in the 1982 fiscal year.

OBJECTIVE DATA-VERIFICATION TECHNIQUES

A computerized data-verification system cannot be expected to detect all possible errors and should not be viewed as a replacement for subjective error analysis. Computerized verification techniques can be used to alert the data user and the analyst to areas of questionable data that need to be used with caution or subjected to closer analysis.

The development of computerized data-verification techniques involves identification of those well-known subjective procedures that can be described in an objective manner and converted into algorithms the computer can use. The following paragraphs describe some of the more obvious and, perhaps, most readily developed techniques and the types of data errors to which they might be applied.

1. Maximum rate of change. An MRC (maximum rate of change) test has primary application to the detection of errors arising from sensing and recording equipment malfunctions and from some kinds of data-communications errors. It is probably most useful when applied to unit values, but it can also be applied to successive daily values. However, a singe-valued MRC applied to all seasons of the year is not adequate as an error-detection device. Maximum rate-of-change limits must be defined on at least a seasonal basis. During expected periods of low flow, the MRC value would be less than that applied to periods of high flow. Ideally, the MRC value should be

definable by a time-dependent or a parameter-dependent regression equation.

An example of the type of error that might be detected with an MRC check is shown in figure 1. This is a typical ADR/DCP (Analog to Digital Recorder/Data Collection Platform) interface error, where the paper tape was punched correctly but the DCP data were stored and transmitted incorrectly.

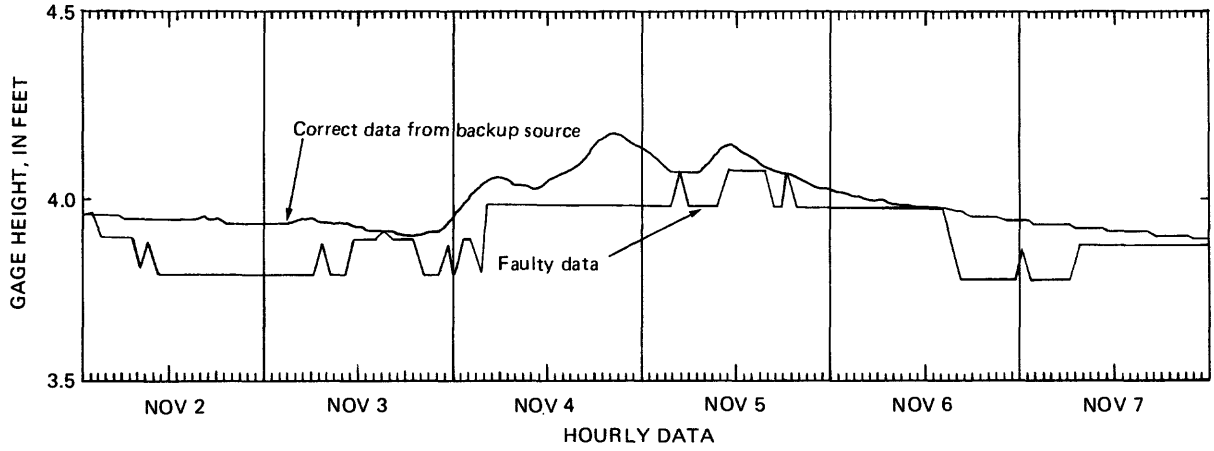

FIGURE 1. – Graph showing real data for North Fork Bull Run River near Multnomah Falls, Oreg. November 2-7, 1979.

If the MRC limit is exceeded, the error would be detected. It might happen, however, that the MRC limit is not exceeded. In this case, the error could possibly be detected by looking for oscillating rates of change among three or more successive values. Also worth noting in connection with the MRC check is the fact that maximum rate of change on a rising stage may be considerably different from the MRC on a falling stage at a specific station. Depending on the degree of refinement possible in defining the MRC limits at a specific site, the distinction between MRC limits for rising and falling stages may or may not be possible or practical.

2. An NC (no change) test has direct application to detection of sensor and recording-equipment malfunctions. Clock stoppage, broken or dislodged float tapes and drive chains, and ice conditions are some of the problems that can cause a station

to report unchanging values for a prolonged period of time (fig. 2). The NC limit defined for computer use would be expressed as the maximum allowable number of successive identical unit values. This limit must also be defined on at least a seasonal basis. It should be applied to the primary parameter, such as gage height, rather that to derived values. Stage recorders normally report values to the nearest 0.01 ft. It might be normal for some stations to report identical

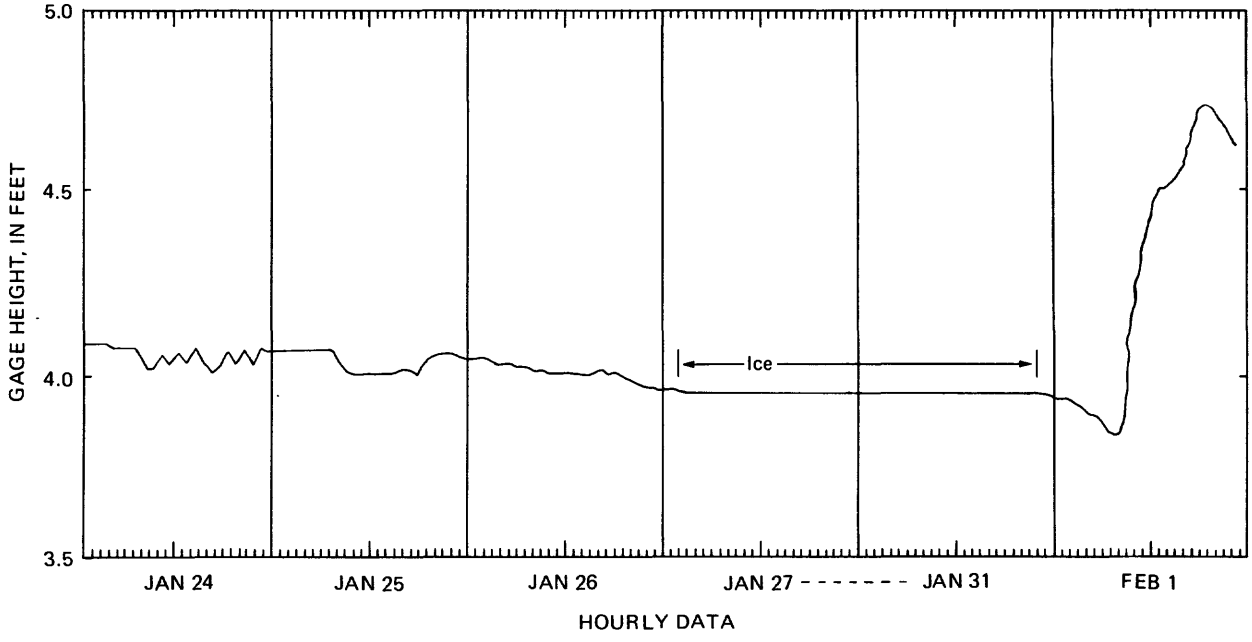

FIGURE 2. – Graph showing real data for Nork Fork Bull Run River near Multnomah Falls, Oreg., January 24-February 1, 1980

readings for prolonged periods during some seasons of the year, but for many stations, slight changes should be expected diurnally or from day to day even during periods of base flow.

3. Acceptable data window. An ADW (acceptable data window) check uses a predefined window or range of acceptable data values defined by an upper-bound value and a lower-bound value. Window width must be allowed to vary according to normal seasonal variations in the data. Multiple windows can also be used, defining zones of acceptable, marginal, and absolute error. Several different statistical techniques can be used to define the window, using historical daily values as a basis for the analysis. For example, the upper and lower bounds of a window might be based on the historical daily maximum and minimum values. Percentage exceedance values or average daily values could also be used as a basis for window definition. Summary hydrograph analyses of this kind are readily available

in WATSTORE, as described in volume 1, chapter 4, sections F, G, I, an K.

Regression analysis can probably be used in some situations. If there is a definable correlation between two or more stations, the acceptable-value window for station X might be expressed as a function of the concurrent value at station Y plus or minus some allowable error factor. The allowable window size could be expressed in terms of standard error or standard deviation resulting from the regression analysis itself. Regression analysis probably has greater potential at multiparameter stations such as water-quality-monitor sites.

Another approach to dynamic window definition is to use predictive basin-modeling techniques. If auxiliary data, such as rainfall, temperature, and snow depth, are available, the value to be checked can be compared with a computed value provided by the model. The allowable window size can be expressed as a confidence band, standard error, or other convenient statistic. Basin modeling is probably of limited value as a generalized data-verification tool because the necessary auxiliary data may not be available. It has greater potential in specialized applications, such as the CROHMS system, where models used for flood forecasting are well developed and large amounts of auxiliary data are routinely collected for that purpose.

Models based on time-series might be easier to develop and more useful than basin models. The Northwest Water Resources Data Center has experimented with a modeling procedure based on the kind of trend analysis used in the business and financial world. In this procedure, the annual hydrograph is separated into a trend line, a seasonal component and a residual, or an irregular component. Seasonal adjustment factors can be computed from the historical data on a daily or monthly basis. These factors are applied to recent antecedent data to produce a sequence of "deseasonalized" values. The deseasonalized sequence defines a trend line that can be projected; the appropriate seasonal factors are recombined with the projection to produce an expected next value. The result is a self-calibrating dynamic model of current conditions. Whether or not this technique is a practical and useful tool for data verification is not known, but it seems to have sufficient potential to warrant further study. Stochastic models in general should be studied for possible application to data verification. A stochastic model is one whose outputs are predictable only in a statistical sense (Haan, 1977).

ADW testing should probably be used as a secondary verification procedure in much the same way as the visual rcomparative hydrograph analysis is used. Studies done in the Northwest Water Resources Data Center indicate that the regression-analysis technique is capable of detecting discharge errors as small as 30 percent when applied to an ideal watershed

containing a large number of well-correlated stations

APPLICATION OF COMPUTERIZED DATA-VERIFICATION TECHNIQUES

Design Requirements

Computerized data verification will undoubtedly be done on many different kinds and sizes of computer equipment, and it will be utilized in several different ways. Two basis approaches can be described as on-line and off-line verification. Off-line verification involves the use of computer programs to access data that have already been processed into the user-accessible data base, and to generate statistical and graphical output suitable for manual error analysis. This method is, by definition, a non-real-time procedure. On-line verification, on the other hand, can be done in either real time or non-real time. On-line verification is done at processing time as an intermediate step between processing the raw data and entering the processed data into the user-accessible data base.

The data-verification techniques described in the preceding sections have general application to both off-line and on-line verification. The following discussion of design requirement and application techniques deals primarily with the special requirements and problems associated with on-line verification, whether it be real time or non-real time. The basic design requirements discussed below are derived from an intensive study done several years ago by the several agencies involved with the CROHMS system.

1. Incoming data should be subjected to data-verification routines before being stored in the user-accessible data base.

2. Error flags should be stored with the data in the user-accessible data base.

3. The basic file and software requirements include:
   a. A randon-access "screen file" to contain verification criteria and option codes for each station.
   b. A system of specialized data-verification subroutines, each designed to perform a specific type of test.
   c. A master data-verification control program. This program would receive incoming data from the main data-processing program; access the correct screen file; and, on the basis of option codes stored in the screen file, pass the data and screen-file information to the appropriate subroutine for verification. After verification within the subroutine, the master data-verification program returns both the data and any error flags to the main processor for storage in the data base (fig. 3).

Screen Files

The primary problem associated with computerized verification is how to store and use the data-verification criteria once they are developed. Maximum rate-of-change limits, length of allowable no change, and upper and lower bounds of predefined acceptable data range windows all must be allowed to vary with time or in response to some other index of current conditions.

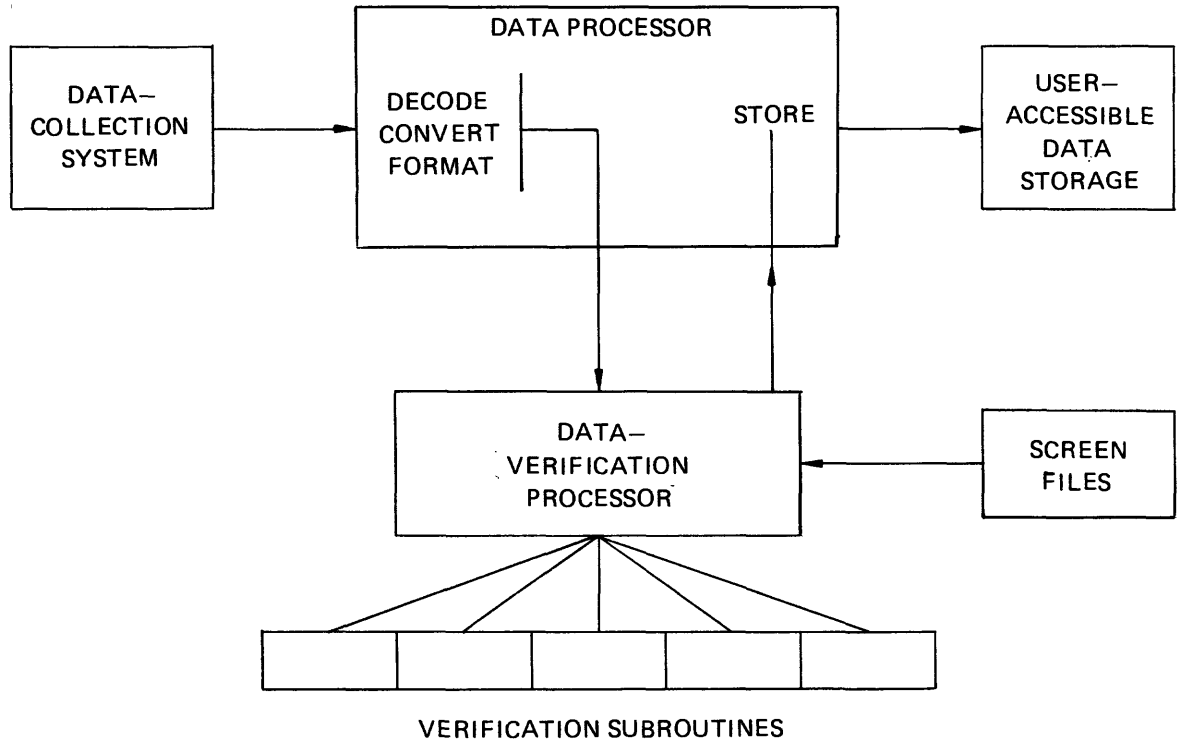

FIGURE 3. – Suggested configuration of a data-verification system.

One approach is to use a 365-day table of limits or a shorter table of limits defined on a seasonal basis. The table would be stored in the screen file and accessed by a table look-up procedure. Another approach is to subject the daily limits to a curve-fitting procedure so that they can be expressed as a time-dependent function. Lystrom (1972) explored this approach using polynomial equations to represent the annual hydrograph of the upper and lower bounds of a window.

A variation of the procedure proposed by Lystrom (1972) would be to use time-series analysis. The annual-discharge hydrograph

could be described as a time series composed of a deterministic (trend) component and a random component. The deterministic trend would probably be a sin-cosine function of the day of the year. The random component would be defined by analyzing the residuals from the deterministic trend. The acceptable data window for a given day could be computed as the deterministic trend plus or minus K times the standard deviation, where K is determined by the user. If this approach were applied to an average hydrograph defined on the basis of historical data, it would be sufficient to fit a single curve to the hydrogrpah rather than to fit two curves to the upper and lower bounds of a window.

Smoothing functions and data transformations are also potentially useful tools for developing more efficient curve-fitting techniques. They need to be explored in greater detail. The results of a preliminary study along these lines are illustrated in figure 4. In this example, an iterative smoothing procedure, termed 3RSSH by Tukey (1977), was applied to the daily 10-percent exceedance values generated by a summary hydrograph program. From this sequence of smoothed values, the maximum and minimum values are used to compute a midpoint halfway between the maximum and minimum. The smoothed value for each day is then expressed in radian measure as an angle between $+\pi$ and $-\pi$, computed from the maximum, minimum, and midpoint values (2π radians = 360°). A factor of $T*(2\pi)$ is added to each radian value, where T is the day number, ensuring that the resulting values are all positive and somewhat linearized. Standard curve-fitting procedures are applied to the transformed radian values, resulting in a function [f(T)] of T. This function is used in the general equation Q=R*Sin[f(T)]+m, where R=(max-min)/2 and m is the midpoint value. The usefulness of this approach depends on the degree to which the function f(T) is simplified by the transformation. If the curve-fitting procedure results in a

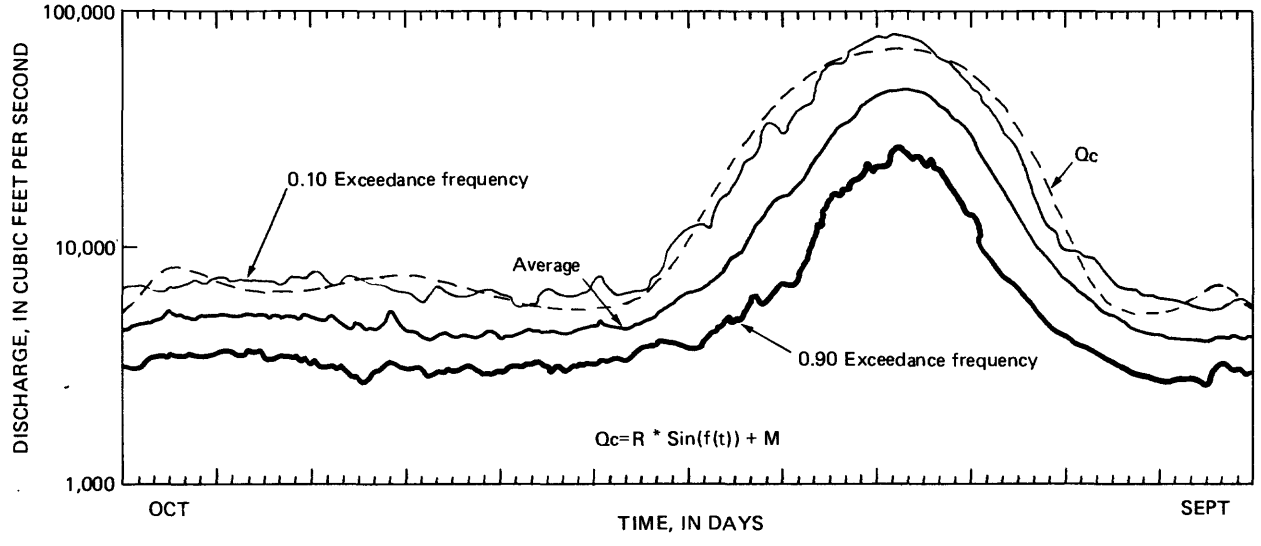

FIGURE 4. – Summary hydrograph showing exceedance-frequency curves for Salmon River at White Bird, Idaho, based on historical record (1928-72).

polynomial function of some complexity, then this method offers no advantage over a straight polynomial curve fitting of the smoothed data.

In any event, screen files should be designed to allow use of equation coefficients derived by these and other methods as they are explored and developed. In addition, the screen file should have the ability to store constants such as the maximum allowable number of sequential values reflecting no change. One essential element in the screen file is a verification option code, or codes, used by the master data-verification program to select the appropriate verification subroutines. Each subroutine would perform a specific kind of verification procedure and would use specific types of coefficients or constants previously stored in the screen file.

Application to Real-Time Verification

In a real-time processing environment, the primary constraint is that all processing must be done quickly and efficiently. The data-verification procedures must not cause a significant delay in processing time. The size of the computer and the details of the data-processing system will influence the ultimate design of the real-time data-verification software. Regardless of how such a system is designed, the data-verification software must receive at least four pieces of information as follows:

1. Station identifier (STID)
2. Parameter code (PC)
3. Value (V) of the data element to be verified
4. Time (T) at which the data element was sensed in the field. The verification program could be designed to receive from the data processor either of the following:
   1. A single data value (V) and its associated STID, PC, and T codes
   2. A block of data values, times, and parameter codes associated with an STID

To some extent, the design of screen files will depend on which verification option, single value or block of values, is chosen. If the single-value verification approach is used, the screen file should have the ability to store some antecedent values for use in the no-change check and other types of verification procedures. Alternatively, the data-verification program could have access to the data base containing data previously processed. Unless this access is done very efficiently, however, it may require more time than desirable in a real-time system.

A single-value data-verification system operating in real time would be limited to performing those tests that can be applied to unit values. A verification system designed to process a block of multi-parameter data could perform correlations among several parameters; this would be important

for water-quality monitor data.

Verification criteria stored in the screen files would need to be updated periodically. The updating procedure can be interactive or batch, depending on the configuration and size of the computer system.

Application to Non-Real-Time Verification

Data verification in non-real time can be done either on-line or off-line. If the raw data are being processed and stored in a user-accessible data base, on-line data verification should be utilized. Whether the processing and verification are done on a large centralized system such as WATSTORE or on smaller local processing systems, the basic concept of a screen file would apply. Design details will vary considerably, depending on the size and capability of the computer and the extent to which the data-verification software has access to data for more than one station. Verification can be done on blocks of data more readily in a non-real-time processing environment than in a real-time environment. Each block of data would consist of recorded instantaneous values for one or more parameters collected at a specific station, or at two or more stations, during a given time period. A combination of on-line and off-line verification may be most easily adapted to local processing on small computers. On-line verification would be used to detect large errors during the initial processing; off-line procedures would perform a second, more sophisticated verification of the processed data.

If data processing is done entirely on a local minicomputer and there is no outside-user access to the data, verification can be done entirely by off-line procedures. In this mode, all the data verification can be done locally, and final or corrected data transmitted to WATSTORE.

SOFTWARE SUPPORT REQUIREMENTS

On-line data-verification systems, operating in either real time or non-real time, will require predefined verification criteria for direct input or for storage in screen files. Some of the criteria required by techniques described in section entitled "Objective Data-Verification Techniques" can probably be developed manually (MRC, NC), but others will require statistical analysis of historical daily values. WATSTORE software described in volume 4, chapter 4, sections E, F, G, I, and K of the WATSTORE User's Guide provides a solid base of support for this requirement. Additional analysis procedures can probably be developed most easily by use of the SAS (Statistical Analysis System) presently available on the Geological Survey computer. Once developed, the criteria will have application to local processing as well as WATSTORE processing environments.

SAS is probably the most effective tool for off-line processing, due to the wide variety of graphic output products

that can be generated. Minicomputers performing local processing probably will not support SAS software, but they will probably be able to support specialized software based on SAS routines. SAS can be applied to the historical data files in WATSTORE to help identify those kinds of analyses that might be of value for off-line verification. Some of the more useful of these products should then be extracted for local in-house use.

SUGGESTED IMPLEMENTATION PLAN

Three different processing systems, WATSTORE, HYDRECS, and local processing are presently being used by the Geological Survey. On-line data verification can, and probably should, be incorporated into all three. New developments in electronic telecommunication and minicomputer capabilities are changing many traditional concepts regarding data processing. One of these changes is the shift in emphasis from centralized data-processing systems, such as WATSTORE and HYDRECS, to distributed processing represented by increased use of minicomputers and microcomputers at the local District level. It is unlikely, however, that all data processing will be done at the local level. As long as any significant amount of processing is done in the central systems and the results are made available to water-data users, there would seem to be a need for some data-verification capability within those systems. Local processing systems are just being developed and plans are being made to standardize some of the software these systems will use. Data-verification software for local processing should also be standardized to the extent possible.

The data-verification concepts and techniques described in this paper have not been tested in an on-line data-processing system. On-line tests should be done in both non-real-time and real-time environments. It would seen logical and cost effective to do this testing on the centralized processing systems presently operational, and to utilize the results for the design and development of standardized minicomputer software at a slightly later date.

The WATSTORE system offers the greatest opportunity for testing a wide variety of verification techniques with the least amount of constraint imposed by file size, data-base access, and other conditions. A prototype non-real-time verification system designed to interface with existing ADR programs would serve to test various kinds of data-verification techniques on a long-term basis. The software should be as modular as possible to allow flexibility of use and ease of modification. A conceptual and detailed documentation of the test system should be prepared and disseminated to field offices. Many field offices would welcome the opportunity to use a system and could provide information concerning performance, usefulness, and suggestions for improvements. This kind of input, along with more formal research by selected Districts and project offices would provide needed data-verification experience within the Water Resources

Division with minimum disruption of other Water Resources Division activities.

Off-line verification of data that have been processed and stored in WATSTORE can be done with SAS software. This kind of verification would produce numerical and graphical output for subjective analysis by field personnel as part of the normal record-preparation procedure. Figure 5 is an example of a SAS plot of specific conductance vs. discharge after applying a logarithmic transformation. The outlier in the lower left part of figure 5 is an erroneous year.

The SAS software contains all the necessary elements to provide the field hydrologist with extensive analytical tools. Some standardized data-retrieval and handling procedures (macros) are already available. Additional SAS macros should be developed specifically for data-verification activities and should be well documented for easy use by field personnel.

A specific plan for developing and testing a computerized data-verification capability within the Geological Survey should contain several elements. The plan should provide for development of non-real-time and real-time systems operating at the central computer system and at the local processing level. The concept of a screen file and the general verification techniques described above can be applied to any of the various systems, with differences arising from the specific software interaction and data-flow requirements of the different systems and different computer sizes and types. Specific developmental activities are presented in what is believed to be a logical order of priority for a long-term program, as follows:

1. Modify the WATSTORE primary processing program (E659) to include additional data-verification routines for on-line processing.
   a. Add a no-change check and an acceptable data window check to the existing rate-of-change check; some developmental work along these lines has already been started. The program that deals with satellite data transferred from MULTICS now includes some of these checks. The WRD ADR Program (E659) was modified in August 1981 to edit, correct and flag potential errors in satellite input data. Current activities, however, do not make use of a screen file.
   b. Add a screen file and develop a master data-verification routine that would be used by program E659. The screen-file and verification subroutines should be capable of storing and using time-dependent equations representing various kinds of verification criteria.
   c. Modify the data base to include storage of error flags.

2. Modify the existing software in the HYDRECS system to develop a means of testing real-time verifications procedures.
   a. Design a screen file for HYDRECS. Perhaps one of the information files already available could be modified to

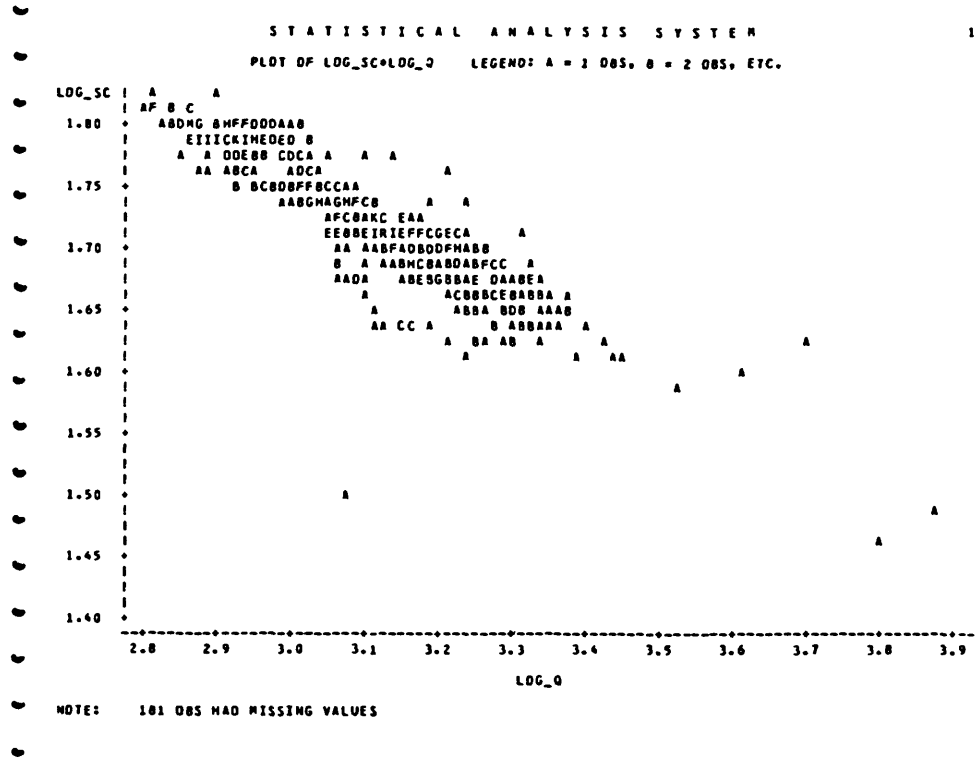

FIGURE 5. – Example of a logarithmic SAS plot of specific conductance versus discharge.

contain data-verification criteria. A relatively small file would suffice for testing purposes.

b. Utilize the existing system of screen codes to flag errors detected by the additional verification routines.

c. Develop a prototype data-verification program to run in real time for testing purposes.

d. Perform some operational testing of the prototype system and use the results as a basis for a more comprehensive real-time data-verification system designed for use either on HYDRECS or at the local processing level.

3. Development and testing of data-verification software for local miniprocessors or microprocessors should probably be done at headquarter's level. If a small computer is available for this purpose and can be programmed to simulate the function of such systems in a distributive processing network, the basic-data verification concepts and data-handling requirements can be tested for field use. Depending on perceived cost benefits and priorities, this work could be done in lieu of adding verification capability to HYDRECS. Because local processing will be performed in real time and in non-real time, both kinds of verification systems should be tested.

4. Develop non-real-time data-verification software designed to access WATSTORE daily-values files. This system would supplement the verification done by program E659 and would provide a test mechanism for research and development of more sophisticated kinds of error testing. It could be designed to use the same screen files designed for E659 and should have the capability of utilizing additional subroutines to test various approaches to verification.

5. Develop and document SAS procedures to access historical data and generate verification criteria for storage in screen files used by verification programs. These SAS procedures would supplement the information available from other WATSTORE statistical-analysis programs.

6. Develop and document additional SAS macros and procedures to retrieve and process streamflow data into a variety of graphical and statistical output formats for off-line verification by field hydrologists.

Development of a strong data-verification capability within the Water Resources Division should be viewed as a long-term project. Project leadership and coordination should be provided at headquarter's level. To obtain maximum utilization, the system must be relatively easy to use and provide demonstrable benefits to those involved with data collection and processing at the field level.

The Northwest Water Resources Data Center has been involved with the development of data-verification procedures applicable to CROHMS. Some of the results of that work can be documented in greater detail for use by the Water Resources Division. Requirements for data verification in the CROHMS system are somewhat different from those of the Water Resources Division, but there is a great deal of overlap. Data-verification studies done thus far by the Northwest Water Resources Data Center have centered on analysis and interpretation of long-term historical data. Some of this work has been done as a by-product of other studies done for and in conjunction with the Columbia River Water Management Group. As new ideas and techniques are developed, implementation of a Water Resources Division data-verification project will provide a focal point for contributions from project offices and Districts.

REFERENCES

HAAN, C. T., 1977, Statistical Methods in Hydrology: Ames, Iowa, The Iowa State University Press, 378 p.

Lystrom, D. J., 1972, Analysis of potential errors in real-time streamflow data and methods of data verification by digital computer: U.S. Geological Survey open-file report, 41 p.

Tukey, J. W., 1977, Exploratory data analysis: Reading, Mass. Addison-Wesley Publishing Co., 689 p.

☆U.S. GOVERNMENT PRINTING OFFICE: 1982—59