INSTITUTE OF HYDROLOGY

REPORT No.5.

PROCESSING ERRORS IN THE ANALYSIS OF CATCHMENT DATA

A.N. Mandeville, D.T. Plinston, and Miss A. Hill (1969)

ABSTRACT

A revised system of processing catchment data recently introduced at the Institute of Hydrology is described. The features of particular interest are the form of the raw and final data, and the use of a quality control program. The results of applying the system to a record of previously hand-checked data are tabulated. Possible sources of processing errors are examined and the success of the new system in minimizing them is discussed.

INTRODUCTION

Since its foundation in 1962 the Institute has initiated experiments to investigate the hydrology of six natural catchments in different parts of Britain. Some of these are used purely to study the rainfall-runoff relationship, while others are designed to examine the hydrological consequences of changes in land use, i.e. of vegetation or drainage.

All catchments possess a dense network of instruments; streamflow, rainfall and climatological measurements are recorded at each by the local river authority, by voluntary observers, or by members of the Institute. The gauging structures are either Crump weirs or trapezoidal flumes, and river stages are recorded automatically on charts or punched paper tape. Daily and storage rain gauges are read manually, while the hourly distributions are obtained from the charts of a limited number of Dines autographic tilting-siphon gauges. Daily climatological observations are taken manually for calculation of the Penman estimate of evaporation. In addition, on most catchments soil moisture changes are measured by neutron scattering or gravimetric sampling techniques, and in a chalk catchment the water-levels of numerous wells are recorded. All these data are sent back to the Institute where they are punched on cards and then processed by digital computer to give the hydrographs and areal estimates of the variables for each catchment.

Handling such large quantities of diverse data can result in two major problems. The first of these is the proliferation of errors, which may occur either in the preparation of the basic data or during the primary stages of processing. For example, the rainfall readings may be ascribed to the wrong time of day, or the wrong calibration may be used for converting the water levels to discharges for a particular flume.

The accuracy required of the final data is governed to a large extent by the uses to which it may be put. Various rainfall-runoff models may be tested on a particular catchment, or its waterbalance may be compared with that from a neighbouring catchment. A combination of the investigations would require not only total, but also continuous accuracy, i.e. the shape of an individual recession curve would be of equal importance to the cumulative totals of the variables over a given period. However, no consideration is made in this paper of either instrumental error or theoretical error in the standard hydrological methods used, as these can only be examined properly when the final computed results are not obscured by unnecessary data processing errors.

The second problem, that of complexity, is produced by many factors, of which the following sample illustrates their influence on the processing of the data. Several different variables are measured on each catchment; others, such as an ecological measurement, may be added at a later date. Although the data are in sequential order they are not collected continuously, but sent back to be processed in batches. In addition to the drawn out changeover from British to Metric units, the measuring instruments are occasionally recalibrated entirely. The quantity of data itself is an important consideration; an input of 4450 readings and an output of 2250 readings is a months quota for a typical catchment! Obviously if separate programs were to be used for each combination of conditions, the number of programs would soon reach astronomical proportions.

In addition to accuracy and simplicity, the ideal system of data processing should possess the virtues of efficiency and versatility. The storage of input and output data should be permanent and compact and should allow for rapid retrieval of information. It is naturally impossible to maximize any one of these requirements except at the expense of the others, but if a choice is to be made, greatest weight should be attached to accuracy.

METHODS

Although it is many years since digital computers were first used for data processing, not much information is available about their use in hydrology for checking large quantities of information. It is suspected that ideas in this field have remained mainly theoretical, and it has been left to other sciences, notably meteorology, to apply them in practice. Hand-checking large quantities of data was found to be both laborious and unreliable, and the computer's ability to make high speed comparisons of two sets of data was the essential ingredient for the success of the following methods.

Firstly, all the catchment parameters needed during the computations, such as the Thiessen areas, the height of the anemometer, or the flume correction factor are added to the respective batch of data when it is punched up on cards. This ensures that the programs are entirely general and can handle data from many different catchments, without further information being added. Secondly the complete batch of data is subjected to a quality control program. This consists of a series of simple tests, which ensure that the data are complete, in the correct sequence and free from gross inconsistencies.

Thirdly, an indexing system is used to add a unique identity to each batch of data. This ensures that the correct calibration tables and processing programs are selected, and serves as the basis of the data retrieval system. Fourthly, the basic raw data as well as the final processed data are mounted on magnetic tape. The former allows any future modifications of the standard hydrological techniques to be rapidly introduced, while the latter ensures that the final data are free from rounding errors and in the most suitable form for further use.

INPUT DATA

One calendar month was selected as the most convenient length of record for handling the incoming field data. A separate batch

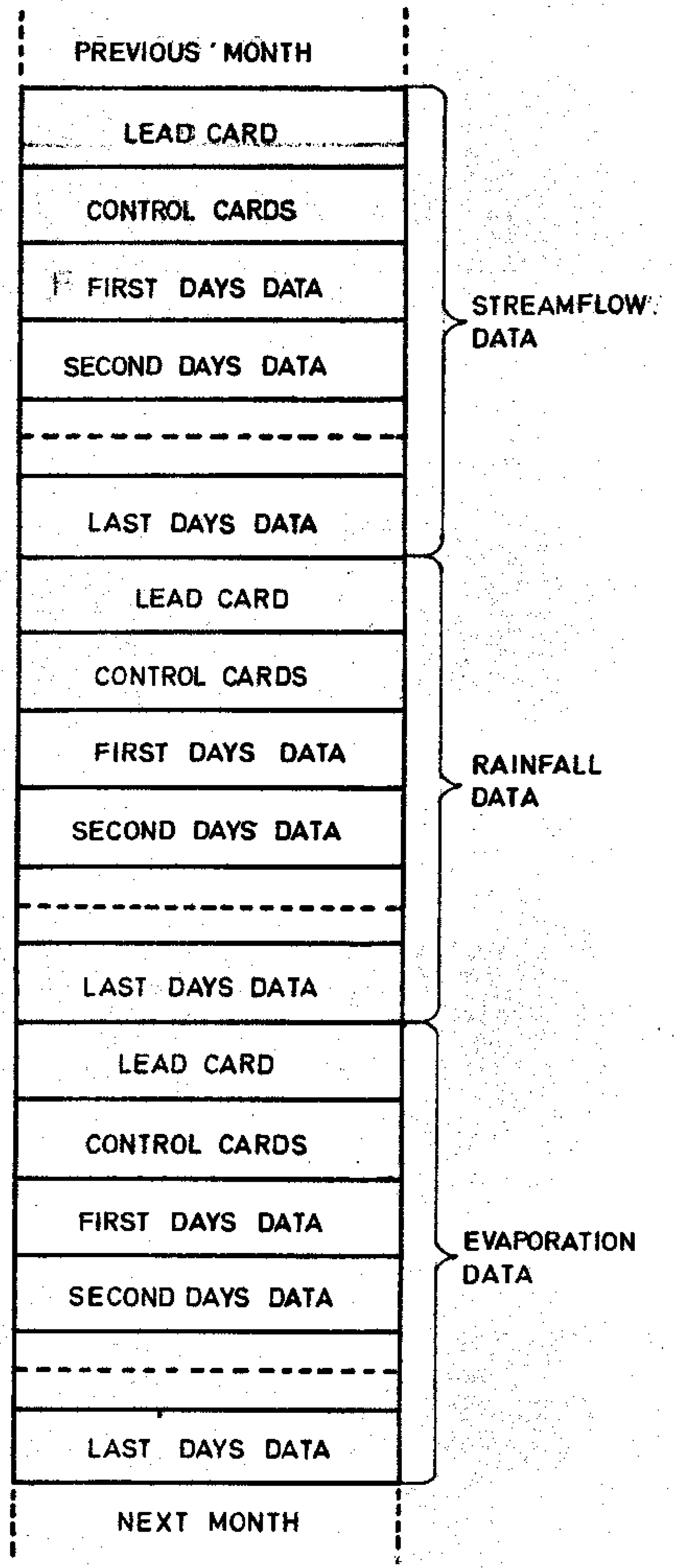

FIG. 1. THE ORDER OF A MONTHS DATA ON THE TAPES.

of cards is punched for each variable measured on a catchment, and consists of a lead card, control cards and the daily data in sequence. These batches are then assembled in a standard order; one month's data for a typical catchment are shown in Fig.1.

Lead Card

A simple 8 digit code is punched on the lead card to identify the batch of data uniquely. The variable is represented by the first two digits, the catchment by the next two and the month and year by the remaining four. For example,:- 03021065 indicates that the batch is the third variable - the evaporation, the second catchment - the river Ray, for the month of October, 1965.

Control Cards

In addition to the instrumental readings there is certain information about the catchments that is required for the quality control and processing programs. These parameters fall under one of the following four headings: catchment characteristics, instrumental network, calibration of instruments or administrative details. Some parameters are constant for a given catchment, while others vary from month to month, or change completely when new instruments are introduced. On account of their importance, the parameters are punched on the control cards and thus become an integral part of the data, passing unchanged through the programs to the final display. Since they form the basic reference for testing the data during the quality control these cards are checked extremely carefully by hand. The code used on a typical control card is shown in Fig.2 with an explanation of the corresponding parameters.

CODE	PARAMETER	TYPE
1856.1	Area of catchment in hectares	Catchment characteristic
03	Number of recording gauges	Instrumental network
1	First recording gauge symbol	Calibration
23	No. of standard gauge corresponding to first recording gauge	Instrumental network
2	Second recording gauge symbol	Calibration
14	No. of standard gauge corresponding to second recording gauge	Instrumental network
4	Third recording gauge symbol	Calibration
02	No. of standard gauge corresponding to third recording gauge	Instrumental network
23	Total number of standard gauges in catchment	Instrumental network
15	Number of standard gauges in operation	Calibration
1	Index to denote measurements taken in inches	Calibration
02	Catchment No.	Administrative
31	Number of days in month	Administrative
12	Month	Administrative
67	Year	Administrative

FIG. 2 RAINFALL CONTROL CARD

Daily Data

For each catchment these data comprise daily meteorological data, daily and hourly raingauge readings, and river stages recorded, depending on their rate of change, at various intervals between 15 minutes and 3 hours. To assist identification the catchment number and date are punched on each card, and where the order of two consecutive cards is not clear, an additional numbering system is used.

Another problem to be faced is the continual change in the recording gauge network due to instrument failure. This is overcome by assigning one symbol from the geometric scale 1, 2, 4, 8 etc to each of the recording gauges. A symbol on the standard gauge card representing the sum of these symbols for example 5, indicates uniquely the combination of recording gauges in action.

Catchment Tables

Some catchment information, such as the flume calibration contains too much data to be mounted on the control cards and is stored instead on a separate magnetic tape entitled the TABLES TAPE. A symbol on the control card indicates the position on the tape of the tables required by a particular catchment during the processing program.

ANALYSING THE DATA

After the incoming field data have been punched on cards, the analysis proceeds in three distinct steps (Fig.3). Firstly the cards are run through a quality control program to eliminate errors, and then copied onto a magnetic tape called the COPY TAPE. Secondly the processing program, with this tape as input, uses standard hydrological techniques to derive the required final values, which are output onto another magnetic tape called the PROCESSED TAPE. Finally another program displays the data on this tape in various ways such as printing out, plotting on graphs and punching daily totals.

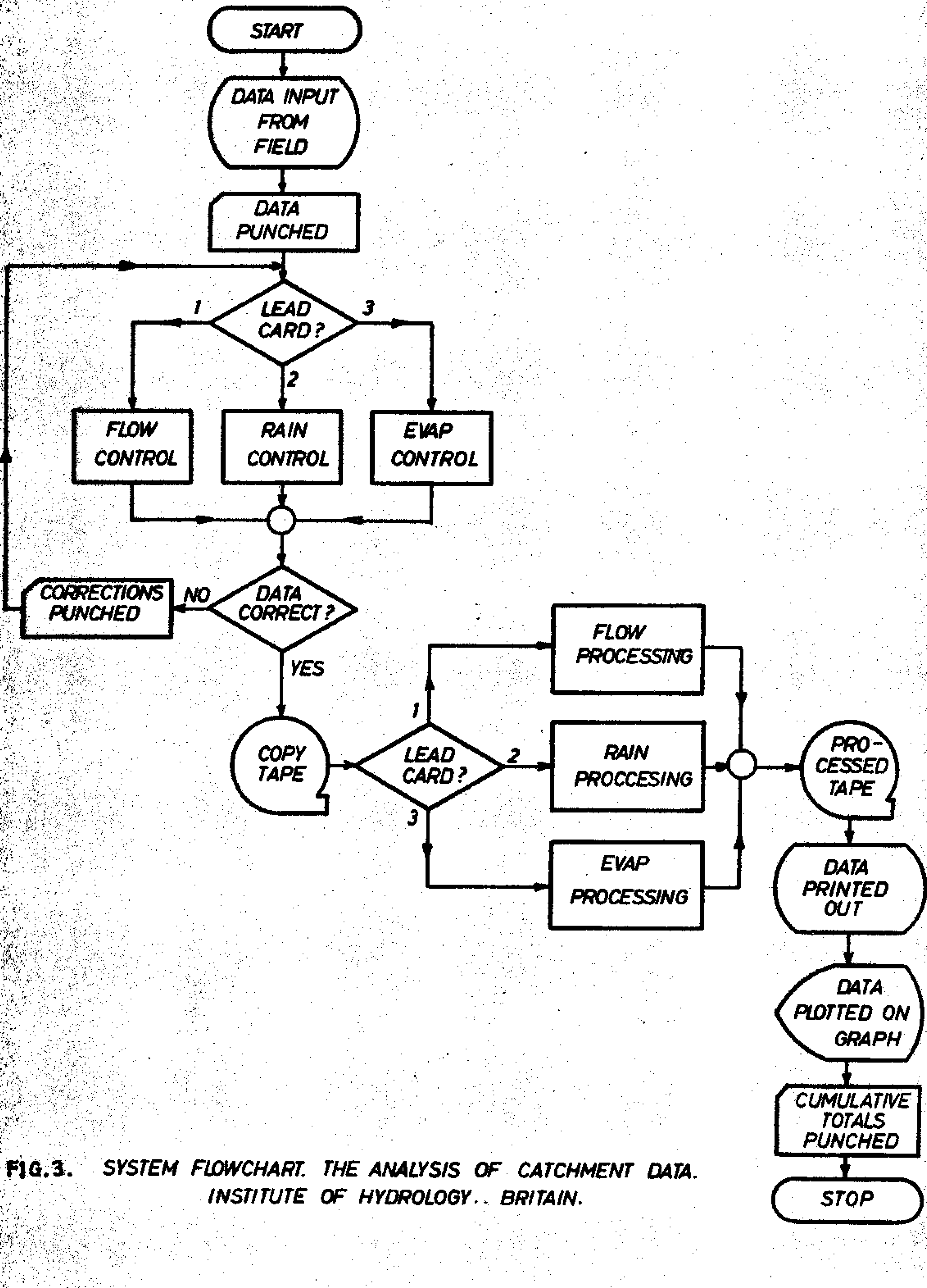

FIG. 3. SYSTEM FLOWCHART. THE ANALYSIS OF CATCHMENT DATA. INSTITUTE OF HYDROLOGY.. BRITAIN.

The programs are run on an ICT Atlas computer; this is a fast, time-sharing machine with 48K ferrite core and 96K magnetic drum store. One inch wide 12 track magnetic tapes are used to store information, and the Ampex TM2 tape mechanism gives a transfer rate for reading or writing of 64,000 6-bit characters per second. Input is on an ICT card reader at a speed of 600/min., and output on an Anetex printer at 1,000 lines/min., an ICT card punch at 100/min and a Benson-Lehner Model J Graph Plotter.

Quality Control Program

Amongst the many types of error that this program eliminates, some are obvious such as data being in the wrong order, or missing entirely. Many are handling errors, such as mis-copying from the charts or punching the cards incorrectly. Others are internal inconsistencies such as a false peak in the river stage data or a lack of equality between recording and standard raingauge totals at the same site. All these errors will have a significant effect on isolated parts of the processed data. However, there are other types of error, for example instrumental drift, which may affect continuous quantities of processed data and will require more sophisticated elimination tests than those now being used.

Only a brief description follows of the three methods used to detect errors, but lists of the main tests and references to the files containing full details of the programs may be found in the appendices. Firstly a direct comparison can be made with the relevant parameter on the control card; for example, the sum of the Thiessen areas should equal the catchment area. Secondly groups of readings are continuously compared and if their differences exceed a pre-determined limit, the position of a possible error is printed out. Lastly the total of the variables for each day, and the total of each variable for the month are printed out; these may be compared with the hand-worked totals off the original data. For each possible error indicated, the appropriate cards are compared with the original data and corrections made when necessary. A batch of data is repeatedly amended and run through the quality control program until an error free run is recorded. Then the batches are added in monthly order to the COPY TAPE in Binary Coded Decimal, and an off-line tape listing obtained which is a permanent record of the input data.

Processing Program

The variety of catchments and variables requires that several different methods of analysis be used, and these are certain to be replaced by more sophisticated methods in future. Thus each method is written as a separate subroutine of the processing program, and the code on the lead card ensures that the correct subroutine is applied to each batch of data as it is read off the COPY TAPE.

The same method of analysis may often be used on more than one catchment, the only difference being a change in instrumentation. Thus the calibration tables are preceded by the catchment numbers which are checked against the control card before they are read off the TABLES TAPE into store. If the calibration is dependent upon two quantities, for example a symbol and a control card parameter, it is written in the form of a matrix. Using the parameter to indicate the row, and the symbol to indicate the column, the correct corresponding dependent value may be extracted from the table using a simple search technique.

The output from this program is stored in binary form on the PROCESSED TAPE, in the same order as the COPY TAPE (Fig. 1). The information on the lead and control cards is transferred verbatim in order both that a permanent record is kept and that the tape may be used effectively for further work. It is logical that the form of the output data should reflect the frequency of observation of the original data. It is also wasteful to include either dates, since the data run consecutively, or daily totals, since these may be computed when required. Following these principles the PROCESSED TAPE contains, for each day, a number of instantaneous discharges measured in cubic metres per second with their corresponding times, hourly values of runoff, hourly areal rainfall and various daily estimates of potential evaporation, all measured in mm. over the catchment.

Display Program

This program reads a month's consecutive data off the PROCESSED TAPE into store, calculates daily and monthly totals, and prints out the data in a new arrangement as follows. The first page contains all the information on the control cards, the second page is a monthly summary of the daily totals of the variables, and then full details of one day's results are set out on each of the remaining pages. The instantaneous discharges, hourly rainfall values and one daily potential evaporation estimate are plotted against the same time axis on graph paper. Finally the daily cumulative totals of runoff, rainfall and evaporation are punched on cards with the catchment number and data for further analysis of the catchment water balance.

RESULTS

The effect of each run of the quality control program on 4 years' data, which had been hand-checked previously, is shown in Fig.4. No errors, however, were discovered on either the lead or control cards.

VARIABLE	After one run	After two runs	After three runs
Streamflow	0	62	100
Rainfall	0	68	100
Evaporation	54	100	

Fig.4 Percentage of months free of error.

A total of 15,000 cards were added to the COPY TAPE in blocks of 500; the subsequent tape listing revealed only one additional error due to this process. An end of file card had been mistakenly put in back to front, but this fault was eliminated when a duplicate COPY TAPE was prepared.

Only one fault was discovered in the input that stalled the main program, yet escaped the quality control. A division overflow occured due to the sum of some recording gauges readings being zero, and as a result subsequently the quality control program was modified. Random hand-checks were made on the print-out once a month for each variable, and all agreed with the computed results.

At cost prices on the computer, the quality control amounted to £7/month, the processing program £1/month and the printing out £5/month with no charge for the graph plotting.

DISCUSSION

Accuracy

The primary aim of the methods presented in this paper is the avoidance of errors arising from either handling the input data or computing the final results. The former may be caused during the copying, punching and collating of the field data, or appear as internal inconsistencies between the readings themselves. The latter may be caused by using incorrect values of the catchment parameters, or by rounding off and interpolating within the analytical techniques used to obtain the areal estimates of the catchment variables. It should be remembered that no consideration is made here of the extent to which these techniques are theoretically valid. This question, as well as others such as instrumental drift, require further investigation.

To minimize copying errors the data should be transferred as few times as possible, and preferably punched directly from the original record which should be set out in the most suitable order. Provided there are sufficient staff available, punching errors may be eliminated using a verifier. In the absence of this equipment, a quality control program has the merit of eliminating both copying and punching errors at the same time, provided that possible errors are always checked against the original data. The results show that hand-checking large quantities of data is inadequate, and indeed the presence of a third run of the quality control program indicates that further errors are made when actually correcting the initial ones!

Collating errors, that is to say data missing or out of order, are avoided by rigorous checking against the control cards. For example, the data or catchment number on each card is printed out if it is wrong, while missing cards will merely stall the program. Checks must be made at every stage, as even trained computer operators can assemble data on the COPY TAPE imperfectly!

Internal inconsistences are eliminated by comparing readings which are adjacent in time or space. For example, standard gauge readings may be compared with adjacent ones while river stages must form a continuous distribution in time. Initially a limit for the difference between the readings must be chosen, but after a certain length of record has been obtained, statistical analyses will allow a more sophisticated approach. Although it may appear that the quality control program let through only one particular error, success can only be assessed relative to the tests employed, and no claim is made that the data will ever be free of errors entirely.

Errors in the catchment parameters differ in their effects on the processed results. For example, if the runoff is expressed as a depth over the catchment every one of its values will be affected by a wrong value of the catchment area, in contrast to which the areal estimate of rainfall will be affected by a small error in one of the Thiessen weighting areas only if there is a particularly uneven distribution of rain in the catchment. Since there are only a limited number of these parameters it should be possible to ensure that the correct values of each are punched on the control cards. Hand checking is considered adequate here, as the results suggest that the success of this method is inversely proportional to the amount of data checked. The argument against writing the parametric values permanently into the programs, instead of checking them each month, is that the changes from catchment to catchment and from year to year would demand a prohibitive number of programs. The control card system seems to provide a good solution, particularly as the opportunity of a second hand-check is provided by the quality control program print-out.

Rounding errors are minimized by using magnetic tape for the permanent storage of the calibration tables and processed results. These values, held to a large number of decimal places, should be used for further computation in preference to the rounded values on the print-out, which are purely for reference. This is clearly an advantage in mathematical model work where the numerous iterations involved would quickly generate large errors from rounded data.

As far as possible, standard, accepted techniques have been used to derive the areal estimates of the various parameters. When interpolations or approximations have been used, as, for example, in the derivation of the hourly areal run-off estimate from the discharge-time data, the methods were examined critically to ensure that the order of error introduced remained small compared with the probable error in the original data. Thus no unreasonable assumptions are made in computing the processed results: a factor of some importance to other users of the data!

Simplicity

In contrast to the problems of its development, the data processing system described possesses a number of features which ensure that it is easily applied. The choice of one month as the unit time-interval not only avoids the problem of unequal division of the year, but also permits a satisfactory turnover of work without delays to the incoming field data. The respective date and catchment number on each card, and different coloured cards for each variable, ease the handling of large quantities of data and aid quick identification of the position of errors. The lead and control cards permanently index each batch of data with all the information required for the successful execution of the quality control and processing programs. Finally the number of separate tapes and programs has been kept to a minimum to lessen the chance of data being misplaced or processed by the wrong method.

It is most important that one form of the final data should be suitable for further research without any part of the processing program having to be re-run. The essential lead and control card information transferred directly to the PROCESSED TAPE, combined with the non-format binary form of the processed data, allow a very high speed of transfer of information from tape to computer store. Attention should also be drawn to a similar feature of the display program; the output of cumulative daily totals of the variables on punched cards may be used in two ways. Either a simple balance between the values for a particular date gives the surfeit or deficit of the catchment storage from that at the beginning of the year, or alternatively the difference between the values of any variable for two different dates gives the total of the variable for the intervening period.

Efficiency

Due to the large quantities of data processed, the most economical format must be used at each stage. Cards were chosen for the quality control as they allow errors to be corrected cheaply and conveniently. When the data are satisfactory they are copied onto magnetic tape, as this allows much higher, and therefore cheaper, rates of transfer to and from the computer. A completely satisfactory method of displaying the data has yet to be devised, but it appears that the greatest use is made of the graphical output. With the increasing number of catchments, the storage and retrieval of displayed data may well become a problem, and consideration should be given to the microfilm output of the computer plotter which costs one twentieth of that from any other peripheral.

Although the overall computing costs were kept to a minimum by using a large fast machine, their breakdown reveals that the processing despite being much more extensive than either of the other two programs, accounts for less than one tenth of the total amount. The major cost was the quality control, and while it is true that this might be reduced by greater care in the error correction, it is certainly less than the cost of subsequent correction.

Versatility

Although there is a move in Britain towards standardisation of the techniques used for measuring catchment variables, at present the programs are designed to accept data in a variety of units. In particular, although the input may be in British or Metric units, the output of processed results is always expressed in mm. over the catchment to facilitate comparisons. Few other restrictions are applied to the output, and diverse combinations of variables at different intervals may be mounted on IBM tape, suitable for loan to outside organisations.

It is most important that the system used is flexible enough to allow for changes. For example, a program utilising separate subroutines for processing each variable allows extra subroutines to be added quite simply when either different catchments or new variables such as soil moisture are introduced. Slightly greater difficulties may be experienced with integrating the output from more advanced instrumentation in both the field and the office. It is hoped to replace the recording raingauges and climatological stations by automatic weather stations recording directly onto magnetic tape, and the problem of reading the stages off the charts may be alleviated by a digital pencil chart follower with a punched paper tape output.

Permanency

Besides being a very compact method of storing large quantities of data, magnetic tapes have the great merit of being both simple to use and quick to duplicate. As soon as a new block of data is added to either the COPY or PROCESSED TAPE, it is duplicated on a spare tape to insure against damage to the originals. To avoid deterioration, copies are made of the complete tapes every nine months which, since they are reproduced by the computer, are as clear as the originals when they were first compiled. When more advanced techniques of analysis are introduced, they may quickly be applied to the original unaltered data on the COPY TAPE and the cards, which are bulky and liable to deterioration, may be disregarded.

There are two important machine subroutines which form the basis for the rapid retrieval of information from the tapes. The first, called TPPOSN, indicates the current position of the tape, thus allowing a permanent address to be attached to the next block of data added. The other routine SEARCH will, if given an address, immediately find the required position on the tape. Although this method is adequate with the limited quantity of data existing at present, introduction of disks should greatly reduce the retrieval time.

CONCLUSION

The results of applying the processing system described in this paper to the data from the River Ray catchment indicate its value in improving the quality and versatility of the basic and final data. A framework has been established for further developments in accuracy, effiøiency, retrieval and display.

ACKNOWLEDGEMENTS

The authors would like to thank members of the Catchment and Computer Sections, without whose considerable help the practical application of these ideas would not have been possible.

APPENDIX

(1) List of Quality Control Tests

(2) List of Program Files.

APPENDIX 1

DETAILS OF THE MONTHLY TESTS MADE ON EACH OF THE VARIABLES IN THE QUALITY CONTROL PROGRAM

STREAMFLOW

1) Read control card
2) Print control card
3) For each day of the month :
   a) Read the number of stages
   b) Test the date and catchment number of stage number card
   c) Read the times and stages
   d) Test the date and catchment number on time/stage cards
   e) Test that all times and stages have been read
   f) Test that first time is correct
   g) Test that time differences are correct
   h) Test that last time is correct
   i) Test for stage inconsistencies
   j) Test that first stage equals last stage of previous day
   k) Store last three times and stages
4) Test for the end of the month
5) Read control card and first day's data for following month
6) Test for stage inconsistencies in overlap between months
7) Test for an error free run

RAINFALL

1) Read first control card
2) Print first control card
3) Read remaining control cards with standard gauge numbers and Thiessen areas
4) Test order of area cards
5) Test date and catchment number of each card

6) Test that all area cards have been read

7) Test for sum of Thiessen areas equal to catchment area

8) Set cumulative stores to zero

9) For each day of the month :

   a) Read standard gauge amounts

   b) Test date and catchment number on standard gauge cards

   c) Test for nil rainfall

   d) Test consistency of standard gauge card symbol

   e) Test order of standard gauge cards

   f) Test that all standard gauge cards have been read

   g) Compute cumulative total of each gauge

   h) Compute total of gauge amounts for each day

   i) Test whether recording gauges operating

   j) Rearrange standard gauge rainfall in extended array

   k) For each recording gauge:

      i) Read recording gauge amounts

      ii) Test date and catchment number on recording gauge cards

      iii) Test for consistency of recording gauge card symbol

      iv) Compute cumulative hourly totals

      v) Test for correspondin g standard gauge

      vi) Test for total recording gauge amount equal to standard gauge amount

      vii) Test for total of recording gauges equal to zero

   l) Test that not too many recording gauges have been read

   m) Test that correct recording gauges have been read

10) Test for the end of the month

11) Print cumulative totals for each standard gauge

12) Print totals of standard gauge amounts for each day

13) Print cumulative hourly totals of recording gauges

14) Test for an error free run

EVAPORATION

1) Read the control card
2) Print the control card
3) Set cumulative totals to zero
4) For each day of the month:
    a) Read the elements of climatological data
    b) Test the date and catchment number on the data card
    c) Test the day number
    d) Compute cumulative total of each element
    e) Compute total of elements for the day
    f) Test for minimum temperature being less than or equal to maximum temperature
    g) Test for wet bulb temperature being less than or equal to dry bulb temperature
    h) Test for dry bulb temperature being less than or equal to maximum temperature
    i) Test for dry bulb temperature being greater than or equal to minimum temperature
5) Test for the end of the month
6) Print cumulative totals of each element
7) Print totals of elements for each day
8) Compute monthly means of each element
9) Print means
10) Test for an error free run.

APPENDIX 2

IDENTIFICATION SYSTEM FOR PROGRAM FILES

IH/Atlas/F4/A	Catchment Tables
IH/Atlas/F4/B	Quality Control
IH/Atlas/F4/C	Processing Program
IH/Atlas/F4/D	Display Program
IH/Atlas/F4/E	Tape Edit