ISSN 0260-6925

Bangor Occasional Paper No. 9

A DATA BANK FOR A GEOCHEMICAL CYCLING STUDY

B K Wyatt
Institute of Terrestrial Ecology

Institute of Terrestrial Ecology
Bangor Research Station
Penrhos Road
Bangor
Gwynedd

July 1981

Contents

A Data Bank for the Geochemical Cycling Project

1 Scope of this Document

This paper is intended to describe the suite of programs written for the storage and retrieval of data collected in the course of the ITE Geochemical Cycling Project (ITE 594), to function as a comprehensive manual for users of the programs, and to provide complete systems documentation to facilitate subsequent program modification and development.

2 System Overview (See Figure 1)

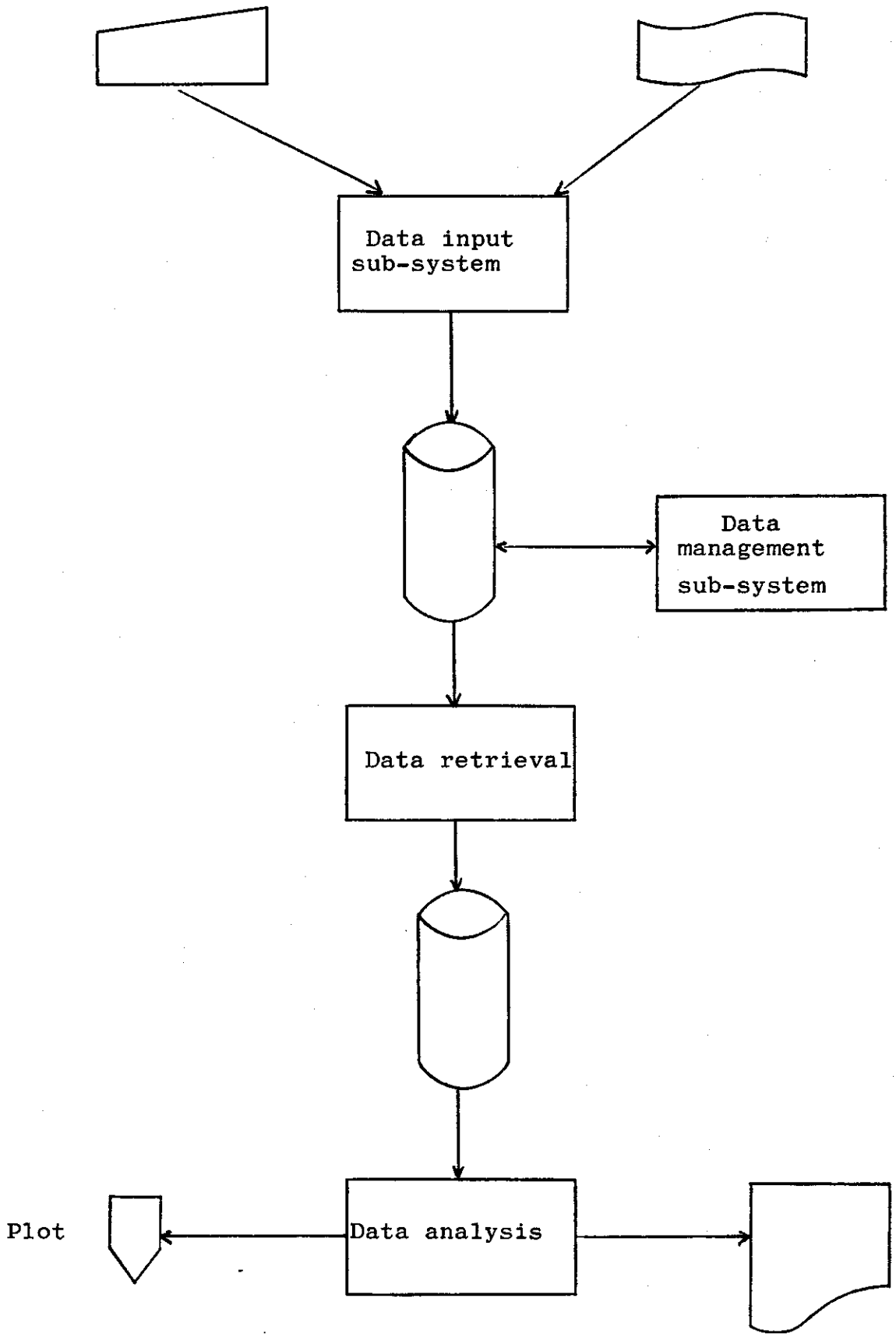
The ITE Geochemical Data Bank and its associated programs is intended for the storage and retrieval of results of chemical analysis of water samples from the ITE Geochemical Cycling Project. It is envisaged that the system will be extended in due course to accommodate other data (e.g. meteorological data, hydrologic data) acquired in the course of this Project.

The programs operate on the PDP-11/34 computer at ITE, Bangor, where all data input and initial data analysis has been performed. However, it is intended that long-term data storage will be within ITE's Terrestrial Environment Information System (TEIS), using the G-EXEC data management system on a NERC main frame computer, which will offer scope for more powerful statistical analysis and more sophisticated graphics output than is available locally on the PDP-11 minicomputer.

The system comprises data files held in matrix form and programs which perform the following functions on the data files;

(i) add new data from various sources

Figure 1 - System Overview



(ii) tabulate and list the raw data

(iii) sort and edit the data files

(iv) retrieve selected sub-sets

(v) perform simple statistics

(vi) produce limited graphic outputs

3 Program Functions

(i) Data Input (See Figure 2)

Analytical data are received from four independent sources in four different formats. These data sources are:

(a) laboratory notebook records of flame photometer readings with associated calibration data. These records require conversion to concentrations before they are incorporated in the data bank.

(b) Chemical Data Sheets from ITE's central analytical laboratory at Merlewood

(c) analytical data punched directly to paper tape from the data processing system in the analytical laboratory at Merlewood.

(d) data sheets recording pH values, particulate concentrations, bicarbonate levels and results from a direct-reading atomic absorption spectrophotometer.

Six programs (AACAL, PLYNØ2, PLYNØ9, PLYNØ6, PLYN12 and PLYNØ5) have been written to accept data from these sources, carry out rudimentary data validity checks and insert valid records into the data bank.

(ii) Data Management (See Figure 3)

Program PLYNØ1 is used to initialise the data files. A simple editor (PLYNØ8) permits

Figure 2 - Data Input

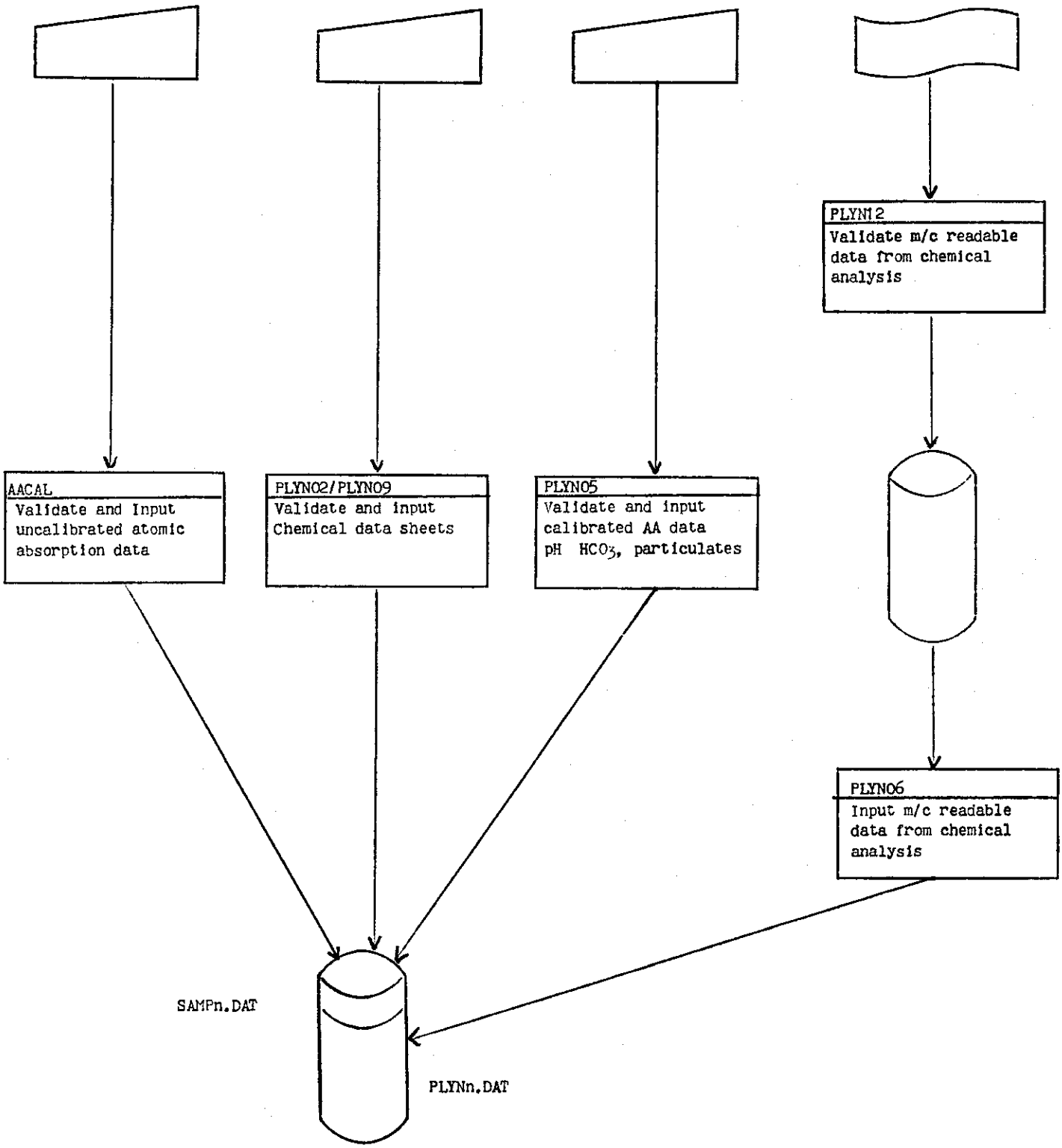
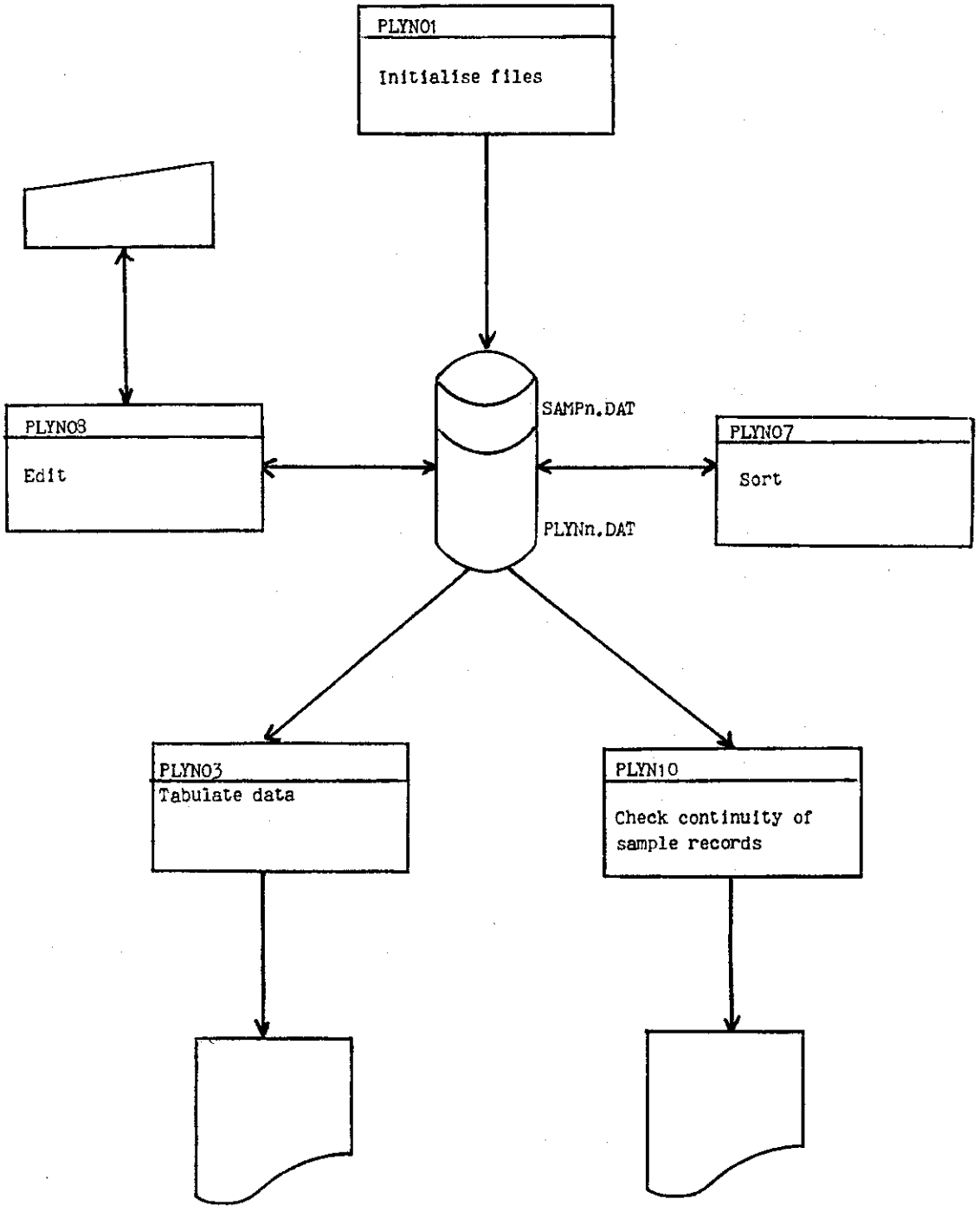


Figure 3 - Data Management



incorrect records to be deleted, new records to be inserted and individual fields in a particular record to be corrected. A sort program (PLYNØ7) enables records entered in random order to be sorted in ascending sequence of sample identification codes. Data may be tabulated and listed using program PLYNØ3. Finally, PLYN1Ø checks the database for continuity of sample records and lists any records which may be missing.

(iii) Retrieval (See Figure 4)

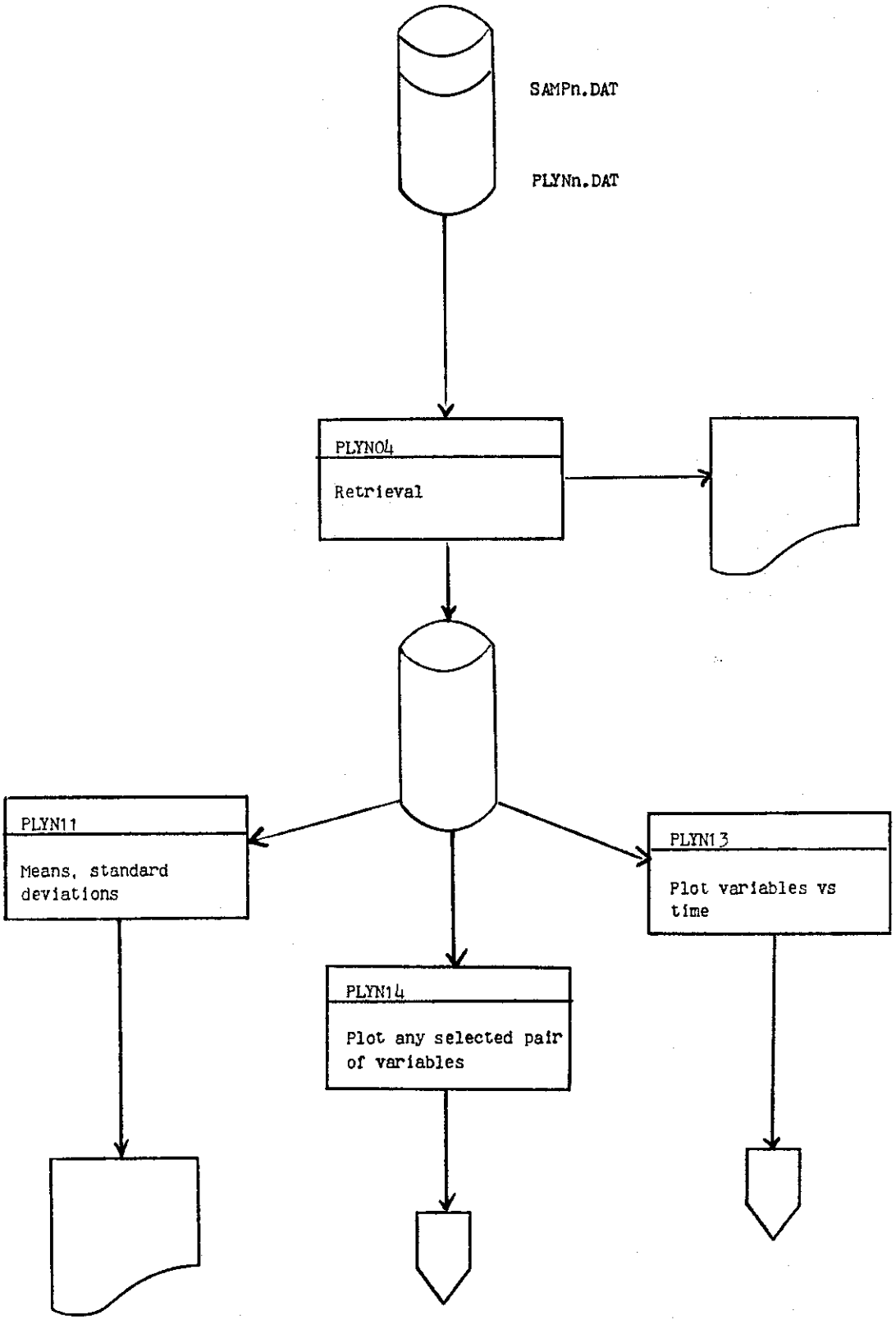
The retrieval program (PLYNØ4) permits subsets of the data bank to be extracted on the basis of selection criteria specified by the user when the program is run. These subsets are written to a secondary file which may be listed or may be passed to other programs for statistical analysis or for plotting. Retrieval may be for individual analytical results (e.g. pH, ammonium ion concentrations), for a given source or source type (e.g. sampler S1, rainwater, rivers), by date, soil horizon, etc.(See detailed program specification for a complete list of retrieval options).

(iv) Statistics and plotting (See Figure 4)

Program PLYN11 computes maximum and minimum values, means and standard deviations for up to 24 determinands. The program accepts as input the file generated by the retrieval program PLYNØ4.

Programs PLYN13 and PLYN14 also operate on the output file from PLYNØ4. PLYN13 plots on the line printer the values of up to 24 determinands

Figure 4 - Data Retrieval and Analysis



against time. Independently for each determinand, the user is able to change the scale of the Y axis and thus to 'stretch' or 'squeeze' the plot to suppress noise and clarify major discontinuities or long-term trends in the data. PLYN14 similarly plots any specified pair of variables. These are options to vary the axis scales independently, and to specify logarithmic transformation of both x and y data.

4 File Structure and Data Formats

It is practice in the Geochemical Cycling Project to assign each water sample, on collection, a unique code which records its origin (stream, rain, soil water etc.), the location of the sample point and, by means of a sequential number incremented from week 1 (23 June, 1979), the week of its collection. Table 1 provides interpretation of the complete notation. This notation is used in the data bank as the identification tag for each set of chemical analysis data stored there.

Associated with each sample are a number (currently 18) of results of chemical analysis (see Table 2). These data are stored in tabular or matrix form in 1250 x 24 element virtual array files. Each record, comprising analytical data from a single water sample, occupies one row of the array (24 elements) and each file has a capacity of 1250 records (about 6 months data). Each column of the array (or data field) thus corresponds to one of the chemical analyses listed in Table 2. 24 fields are available; the first 3 are used to record the absolute date of collection of the sample (field 1 = year, field 2 = month, field 3 = day): the next 18 record data from the 18 chemical analyses and 3 are spare.

Table 1 - Sample Coding System

Code format: nnn/A mm/H

nn: Trip number

1-999 (Trip 1 corresponds to 23 June 1979)

A Sample Type:

C - Stream Sample (Cyff)
G - Stream Sample (Gerig)
W - Stream Sample (Wye)
R - Rain Water
S - Soil Solution
L - Lysimeter
P - Snow

mm Site number
Numeric sequence assigned to individual samplers within each sample type.

H Horizon Code (S and L records only)

Site number	Horizon Code	Site number	Horizon Code
S 1	E	S18	C
S 2	C	S19	E
S 3	B	S20	B
S 4	E	S21	C
S 5	C	S22	E
S 6	B	S23	B
S 7	C	S24	C
S 8	E	L 2	P
S 9	B	L 5	P
S10	E	L 6	P
S11	C	L 7	O
S12	B	L 8	O
S13	E	L 9	O
S14	B	L10	O
S15	C	L11	P
S16	E	L12	O
S17	B	L13	O

Table 2 - Data File Record Format

Field number	Record type	
1	Year	Sample Date
2	Month	
3	Day	
4	Sodium	mgl^{-1}
5	Potassium	
6	Calcium	
7	Magnesium	
8	Iron	
9	Manganese	
10	Aluminium	
11	Silicon	
12	Phosphorus	
13	Nitrate	
14	Ammonia	
15	Sulphate	
16	Chloride	
17	Bicarbonate	mg equ.
18	Total Organics	mgl^{-1}
19	Particulates	mgl^{-1}
20	pH	
21	Conductivity	μS
22	Not yet used	
23	Not yet used	
24	Not yet used	

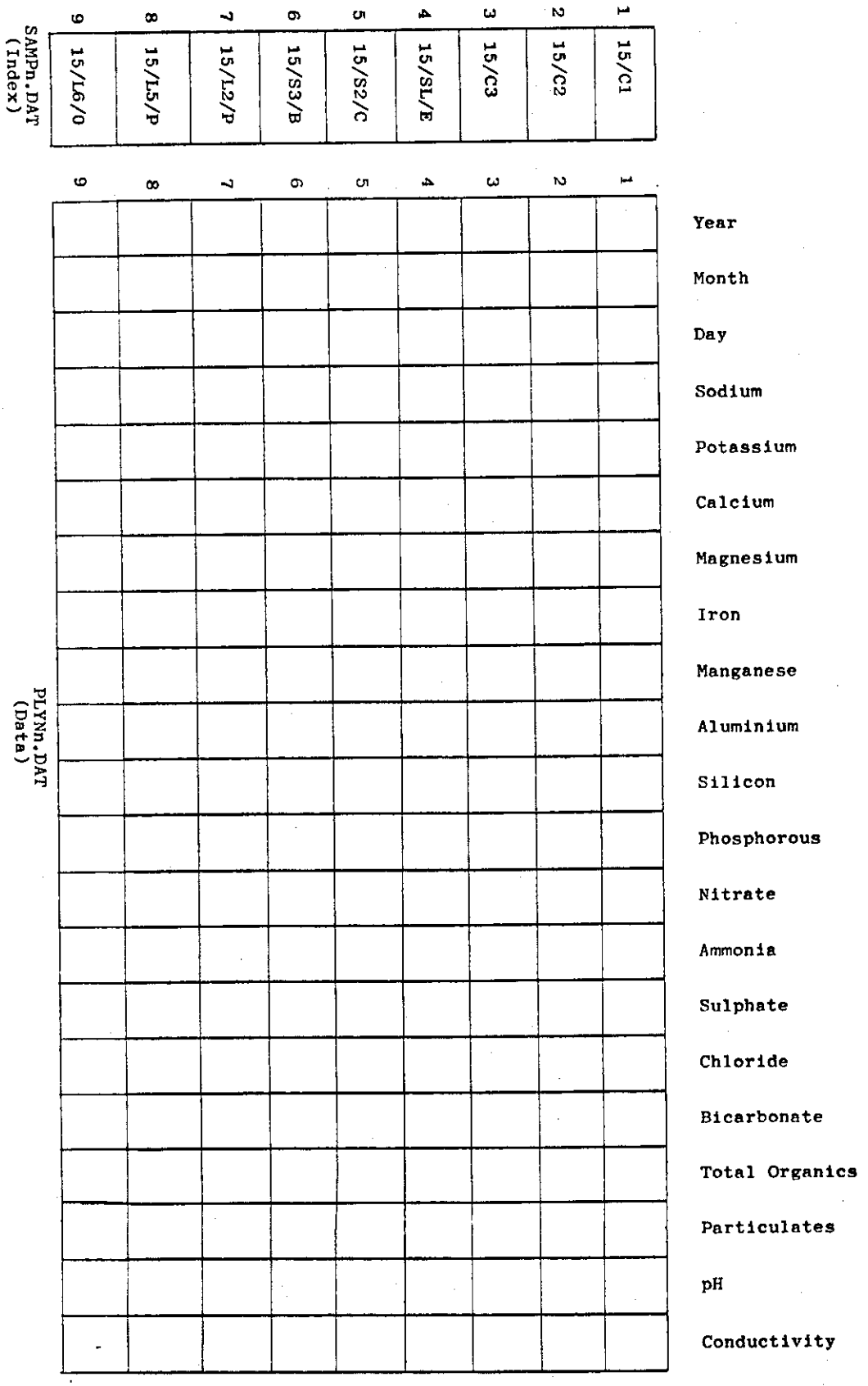
Since the main data array consists of real numbers, it is not possible to hold in the same array the sample labels, which consist of alphanumeric strings. Instead, a small file of sample labels is held separately and functions as an index to the main data file. To locate data for a particular sampler, (e.g. sample 16/81/E), the index file is searched serially until the required code is found: the address of the code in the index is then the same as the address (i.e. the record number) of the corresponding analytical data in the main file, which can now be accessed directly. (See Figure 5).

The structure of the sample labels can be exploited to identify groups of records with similar characteristics (e.g. all records from a specified trip, all records from the Cyff, lysimeter data from peaty layers, etc.), the data file once again being accessed via the index.

The maximum capacity of a PDP-11 BASIC virtual file is 1250 records of 24 elements. This results from the way in which such files are addressed, and the maximum integer size in PDP-11 BASIC. When this capacity is exceeded, additional files are opened on the same physical device to accommodate extra data. Successive generations of data are identified by a sequence number in the file identifier which is incremented as each new file is opened (e.g. PLYN1.DAT, PLYN2.DAT, etc.). Similarly, the identifiers of corresponding index files are also incremented (SAMP1.DAT, SAMP2.DAT, etc.).

The order in which records are held in these files has no significance as far as the programs which operate upon them are concerned. However, for purposes of data checking and, in particular, to identify missing records, it is useful to maintain the data in ascending

Figure 5 - File Structures



order of sample code, and a sort program (PLYNØ7) is available to do this.

5 Data Precision

With one exception, the system imposes no limits on the precision with which data are held (save only the capacity of a 16-bit computer word!). Data values are therefore stored in the same form and to the same level of precision as they were entered. The one exception is the program (AACAL) which computes concentrations from flame photometry data. These data are rounded to one decimal place before storage. The listing program (PLYNØ3) rounds data values before printing them, but this does not affect the actual value stored in the data bank.

6 Data Validation and Detection Limits

All data entry programs incorporate a number of data validation procedures. The format of the sample code is checked on entry as follows, and invalid codes are rejected.

Data checks on Sample Codes

(i) Check for presence of trip sequence number (week number), which must be all-numeric.

(ii) Check for presence of valid sampler letter, separated from trip number by a 'slash' ('/').

(iii) Check for presence of all-numeric sampler number.

(iv) Check for possible presence of suffix indicating a duplicate sample. If present, this must be 'A'.

(v) If sampler code letter is S (Soil Solution) or L (Lysimeter), add an appropriate soil horizon code letter (depending on the value of the sampler number), separated from the sampler code by a 'slash' ('/').

Data values are checked to ensure that they are numeric. In addition, most input programs ensure that values fall between specified upper and lower limits (See Table 3). The lower limits are set at the detection limits of the analytical methods used. Upper limits have been set arbitrarily to 'reasonable' values. Data falling outside these limits are rejected.

Results which are below the detection limits for a given analytical method are indistinguishable from zero, and should be punched as zero. Chemical data sheets frequently prefix values below the detection limit by '<' and input programs are designed to recognise '<' and interpret it as zero.

7 Absent Data

From the point of view of subsequent analysis, absent data falls into two categories:

(i) data absent because samplers were dry (e.g. no rainfall, streams dried up). In these cases, data values of -2 should be punched, and will appear in the data files. Data analysis programs will omit these samples when computing means, total chemical fluxes, etc.

(ii) data absent because no sample was taken (e.g. sampler buried in snow), because there was insufficient water to complete all analyses or because sample or data were inadvertently lost (e.g. cows devouring soil solution samplers!). In such cases, values of -1 should be punched. Subsequent data analysis will substitute the mean value of adjacent trips.

In the case of analysis for bicarbonate (HCO_3^-), it is not meaningful to attempt this analysis when pH is below 4.5. In these situations, record bicarbonate levels as -3.

Table 3 - Upper and Lower limits of validity

Observation	Upper limit	Lower limit
Conductivity	150	10
Iron	1	0.001
Manganese	1	0.01
Aluminium	1	0.1
Chloride	25	1
Silicon	10	0.05
Phosphate	1	0.02
Nitrate	1	0.005
Ammonia	10	0.5
Sulphate	10	0.01
Total Organic	20	0.1
Calcium	25	0.04
Magnesium	25	0.02

The main data files (PLYNn.DAT) are initialised with all elements set to 999. Results which are missing because they have not yet been entered therefore appear as '999' and can be readily distinguished from absent data or zero results.