2002

Centre for
Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

Report of the Working Group on software for processing, quality control, archiving and dissemination of hydrological data from research catchments

J W Finch
A Bayliss
R V Moore
C Watts

15 July 2002

EXECUTIVE SUMMARY

Terms of reference:

I. To produce software requirements specification for LOCAR, CHASM and AWQARD for:-
   - processing and quality control of field data
   - receiving, archiving and dissemination of field data, including receiving data from third parties, e.g. EA.

II. Review functionality and documentation of existing relevant in-house software.

III. Review functionality of commercial software.

IV. Identify financial resources available for software procurement.

V. Identify options and costs for meeting CEH Wallingford requirements.

VI. To prioritise software procurement.

VII. Consult with CEH, BGS and CCS staff as appropriate.

VIII. Recommend management structure and staff for delivering required software functionality.

IX. To recommend work programmes for software procurement from December 01 to March 03.

Inefficiencies and risks in the current procedures for handling field data at CEH Wallingford are:

- loss of data due to lack of backup procedures;
- loss of data on media which are either obsolete or decaying;
- loss of data knowledge due to staff leaving;
- wasted effort due to duplication of programs/macros used to quality control data;
- inefficiencies due to staff having to learn by their own mistakes rather than being able to follow established procedures.

Discussions with staff at CEH Wallingford found a general consensus on a number of issues on the requirements for data input, processing and QC, including a clear communication line from the data users, the data archive, to those who collected the data, to deal with errors detected when using the data.

No in-house software provides the facility of graphical data editing on a PC that is an essential feature for field data acquisition. With this exception HYDATA provides a significant number of the features, although there is some risk involved in its without a commitment to a longer-term programme of upgrading and development. Similar criticisms can be made against SWIPS plus it uses an internal DBMS, but it provides a functionality not available from commercial systems in terms of handling manual soil water measurements. WIS was designed to provide a generic database with flexible and efficient querying of disparate data types. With the exception of having a graphical data editor, it is not optimised for input and quality control of field data. Rather, its strength lies in supporting a Data Centre. However, in order to continue to fulfil this role for existing Thematic programmes, let alone forthcoming ones, it is urgently in need of upgrading.

There are three sets of commercial software, HYDSYS, WISKI and Hydrolog, that fit most of the main criteria and where there is a clear commitment to the long-term support and development of the software by its supplier. A limited technical evaluation has been carried out which concluded that WISKI was best suited to the requirements of this site.

We have dismissed the possibility of either upgrading or developing in-house solutions as impractical, due to the time scale involved and the lack of resources. We have also concluded that it is not necessary for any single software package to deliver the complete solution. Rather we have identified the software that meets most of the requirements and then identified ways of meeting the rest of the requirement

Data input, processing and QC

There is total agreement that raw data must be retained but there is disagreement about how this should be achieved. It is recommended that raw data should be stored in the archive, unless the data volumes are significantly greater than the calibrated values, e.g. in the case of data from Hydras, when CDs should be used.

The analysis of the software to perform the function of data input, processing and quality control indicates that the most cost effective solution is provided by the WISKI, Hydrolog or HYDSYS software systems. A limited technical evaluation established that WISKI was better suited to the requirements of this site than the other two.

WISKI, HYDSYS, Hydrolog and HYDATA cannot handle efficiently manual measurements of soil water because they do not consider the possibility of data being a depth (or height) series as well as a time series. SWIPS is developed to handle these data and so could fulfil this role for LOCAR and CHASM. However, there are no plans for any further development of SWIPS and so this can not be considered a solution beyond the life time of these programmes.

The anticipated direct cost is about £45k to purchase the commercial software WISKI, which will be from the AWQARD computing budget. There will be some hidden costs in terms of staff training.

Archiving

There is a clear first choice for the use of the Oracle database management system by staff at Wallingford, and other CEH sites, to archive data. However, the Oracle DBMS may not be suitable for all data sets, e.g. work based overseas, but this could be catered for by using Microsoft Access as the DBMS but using, as far as possible, the same data model as used on the Oracle DBMS.

There are advantages in using the same software system for archiving data as that used for input, processing and quality control of the field data. WISKI, Hydrolog and HYDSYS can use the Oracle DBMS and therefore are capable of acting as archives for much of the data. HYDATA would need to be tested before it was confirmed that it could use the Oracle DMS and there is no long-term commitment to its development. The most cost effective strategy is to use the commercial software, i.e. WISKI.

SWIPS does not use the Oracle DBMS and so some other method of archiving data must be sought for manual soil water data. The pragmatic solution is to copy the data into WISKI as time series for each measurement depth.

It would simplify matters considerably if there was an agreed data model, which is likely to be that of the software that is used by the field teams. The use of commercial software means that the data model may not be known to us, however there will be routines to access the database. It will also be necessary to provide training in the form of a basic introduction to database concepts and specific skills necessary to work with the Oracle database on this site.

Data Centre

The function of the Data Centre has been separated from the consideration of archives in general because the Data Centre's terms of reference include data for the whole of the LOCAR and CHASM programmes and thus will include data types not regularly dealt with at Wallingford, e.g. bank erosion measurements, subsurface geological data, remotely sensed measurements and data derived from these etc.

Given that WISKI is designed for hydrological time series, it would not fulfil the task as well as WIS. In addition, the Data Centre staff are already experienced in using WIS. However, the WIS software must be updated but not all the functionality of WIS is required for the Data Centre role by further developing an existing prototype software into an operational system. Relatively few staff at CEH Wallingford have experience of using WIS or of accessing the data from the WIS data model on the Oracle DBMS. Therefore the code for routines to retrieve and access data from the WIS data model will have to be written. In addition, there will also be a need for staff to be trained in the use of WIS. However, there is a fallback position of using WISKI to manage time series data, ARC/GIS for spatial data, and some bespoke developments of software to manage any data sets not catered for otherwise.

The only person at CEH Wallingford with the experience required to lead this activity is Mr Roger Moore. It is proposed to develop the functional specification and high level design using existing CEH staff but it will be necessary to subcontract out the programming work to a software house.

Funding for part of this can be found from a number of sources:

LOCAR	£20k
CHASM	£10k
AWQARD	£37k
Right (BNSC funded)	£10k

In addition, there are some funding proposals which, if successful, would also contribute:

HarmonIT	£10k
HarmoniRiB	£10k

However, this leaves a shortfall of about £100k which would need to be made up from central funds over a period of 4 years.

Dissemination

Significant uncertainty remains until the LOCAR and CHASM Steering Committees make a decision on the mode of dissemination they want. It is possible to carry on with the present system, i.e. using ftp, diskette or CD in response to requests. However, it is clear that the Web is seen by most people as their first choice of searching for and acquiring data. The inability to provide this does not give a positive impression of the capabilities of this site. Ideally, there should be an initiative at a CEH level to share expertise and resources and to make sure that specific data sets had a common appearance and functionality. Enquiries to CCS established that this is not currently the position

It is not possible to identify any resources available at this time. Currently, CEH Wallingford is very poorly placed to meet these expectations as no staff currently have the required skills let alone any experience.

Spatial data input, processing QC and archiving

Although the amount of spatial field data is minimal for LOCAR, CHASM and AWQARD, we have briefly considered spatial data as it is relevant to other projects. Given that the handling of spatial data at the Wallingford site is dominantly carried out using Arc-Info and/or Arc-View, it is sensible to continue with their use. The one proviso is that the supplier, ESRI, has launched a major upgrade in the form of ArcGIS 8 in 2001 which can use the Oracle DBMS. Therefore it would be sensible if any new projects were to use ArcGIS 8 and there should be a migration of spatial data from ArcInfo and ArcView to this new system.

CONTENTS

Page

1 INTRODUCTION

This report considers the issue of software for the processing, quality control, archiving and dissemination of hydrological data from research catchments. Particular emphasis has been placed on the forthcoming LOCAR, CHASM and AWQARD programmes of NERC but it also considers other projects at CEH Wallingford where there is a requirement for such software. The objective has been to recommend timely and cost effective solutions.

1.1 TERMS OF REFERENCE

I. To produce software requirements specification for LOCAR, CHASM and AWQARD for:-

II. processing and quality control of field data

III. receiving, archiving and dissemination of field data, including receiving data from third parties, e.g. EA.

IV. Review functionality and documentation of existing relevant in-house software.

V. Review functionality of commercial software.

VI. Identify financial resources available for software procurement.

VII. Identify options and costs for meeting CEH Wallingford requirements.

VIII. To prioritise software procurement.

IX. Consult with CEH, BGS and CCS staff as appropriate.

X. Recommend management structure and staff for delivering required software functionality.

XI. To recommend work programmes for software procurement from December 01 to March 03.

1.2 THE PROCESS OF TIME SERIES DATA MANAGEMENT

The process of data management has a cyclical nature, as illustrated by Figure 1. Quality control is required in each of the cycles but addresses different issues. Thus, the quality control in the data acquisition cycle is concerned primarily with visual inspections to detect major problems, e.g. sensor failure. At the monthly cycle a more comprehensive level of checking is possible which could include automated checking of bounds, e.g. $0<pH<14$ and ion balances. It is also possible to check for consistency over short periods of time, e.g. drift in the calibration of a sensor. It is at this stage that the data are processed to produce 'secondary' values, e.g. flow from stage or evaporation from climate data. The long-term cycle is concerned with inter-station comparisons and modelling to test for consistency

Figure 1 includes the division of responsibilities between the field teams and the data centre for the LOCAR programme. A similar division applies to the NRFA with the data acquisition and monthly cycle being the responsibility of the data suppliers, i.e. the EA. However, many groups at Wallingford would be responsible for the whole process since the data acquired is primarily required for their own use. In addition, they would not necessarily identify archiving as an activity or, if they did, they would place it at the end of the process because archiving is defined as applying to data that is not in regular use. In general terms the process is the same for all measurements, it is the allocation of responsibility that changes according to the circumstances.

In addition, it should be recognised that the procedures required are generic. It is extremely rare for any measurement system not to require calibration. Therefore, somewhere in the processing procedure the 'raw' measurements will have calibration coefficients applied to generate a hydrological parameter rather than the sensor output. In some cases this may occur more than once before the data is in a form that is useful, e.g. in the case of converting from the mV output of a pressure transducer to river stage, and then the application of a rating curve to produce flow. Similarly identifying errors is also a generic procedure in that practically all measurements systems are prone to sensor failure, calibration drift, the occurrence of spikes etc.

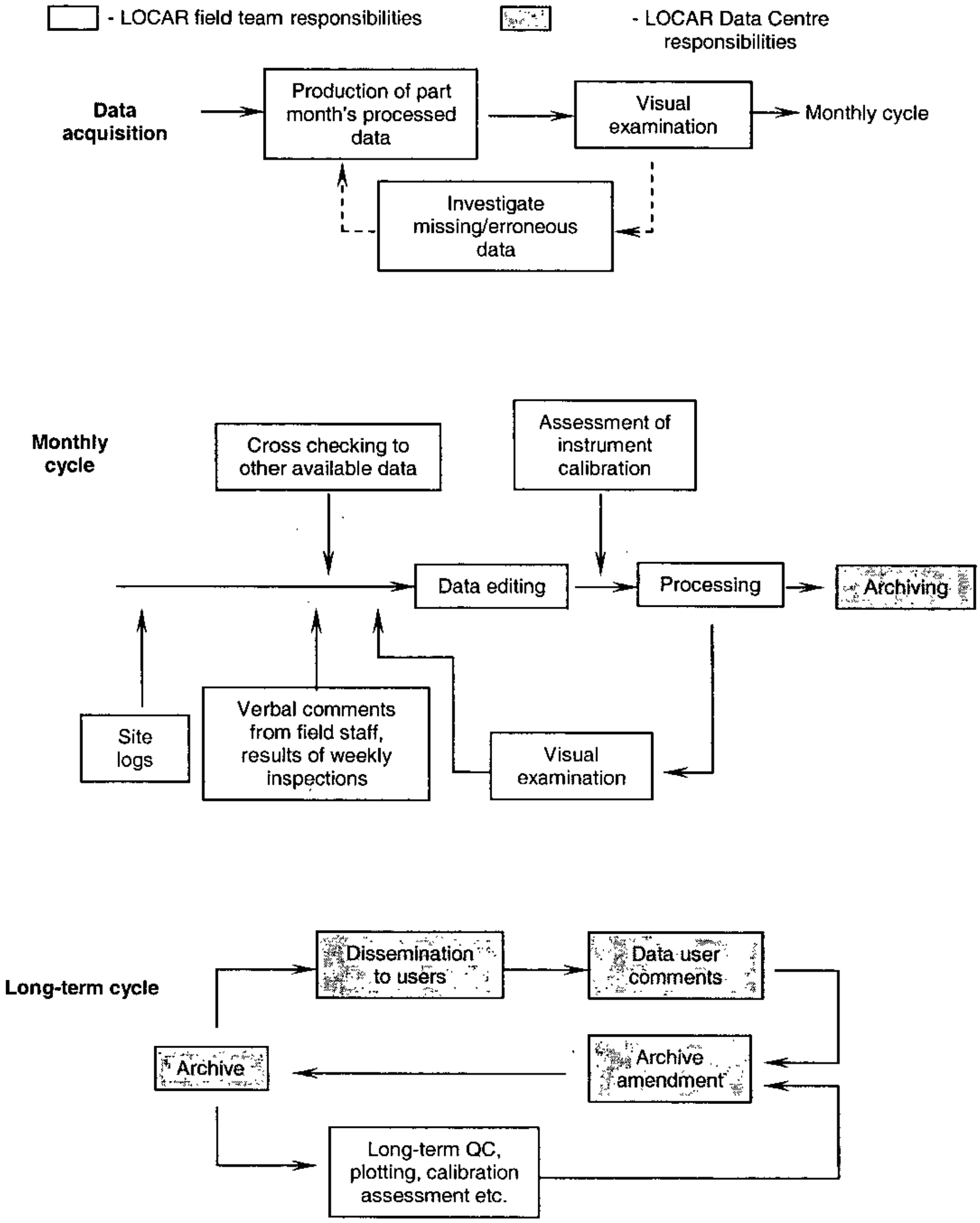
- LOCAR field team responsibilities
- LOCAR Data Centre responsibilities
Data acquisition
Production of part month's processed data
Visual examination
Monthly cycle
Investigate missing/erroneous data
Monthly cycle
Cross checking to other available data
Assessment of instrument calibration
Data editing
Processing
Archiving
Site logs
Verbal comments from field staff, results of weekly inspections
Visual examination
Long-term cycle
Dissemination to users
Data user comments
Archive
Archive amendment
Long-term QC, plotting, calibration assessment etc.

Figure 1 The cycles of processing field data

1.3 THE CURRENT SITUATION

At the time of writing this report, the procedures used at CEH Wallingford for processing, quality control, archiving and dissemination of hydrological data are very much *ad hoc*. The National River Flow Archive (NRFA) probably has the most established quality assurance (QA) system, using a variety of programs to carry out a rigorous set of quality control algorithms which only exclude modelling. In comparison, data from Plynlimon are not subjected to any automated checks. The data downloaded from loggers are graphed and an initial visual inspection of the data made to check for major errors. The raw data files are ftped to Wallingford, where they are stored in original form. The data are then bulked into monthly groups, graphed and visually checked again. For flow data, if errors are detected in the front-line logger it is substituted with data from the backup logger before being loaded into the Oracle database. A variety of procedures are used on other data sets, depending on the experience of the staff involved and the objectives of the project, but these procedures are generally carried out using Microsoft Excel and, in a few cases, Microsoft Access. A wide variety of methods and amount of detail are also used to deal with the 'site diaries' to record the measurement procedures used, although many are based on hard copy. No editing of data makes use of graphical editing facility, with the result that much of the editing is done by time consuming manual changes to the data values.

There are also no consistent established procedures for archiving data. Although some data sets are held using the Oracle database management system (DBMS), notably the NRFA, Plynlimon, LOIS, the Acid Waters Monitoring Network and flood data, many data sets are held on individual scientist's PCs, usually as Excel spreadsheets. Some data have been put into international archives, e.g. HAPEX-Sahel, LAPP etc. In the past, there has been some loss of data during the process of migrating from one computer system to another, e.g. from the Honeywell to the IBM mainframe.

The handling of spatial data is dominated by the use of Arc-Info and Arc-View. These are available under the CHEST agreement and are the most widely used GIS globally. These products include extensive facilities for data entry and quality control. An important exception to the use of Arc Info/View is the use of Fortran and UNIRAS routines on workstations that are used to handle some of the gridded datasets, e.g. land cover, rainfall, DTM etc.

Dissemination of data is not an issue for most staff at Wallingford. The exceptions are the NRFA, Plynlimon and LOIS. For these data, in response to a request (generally in the form of an email), selected data are manually retrieved into a comma-separated-value (CSV) file and either emailed or placed in the anonymous ftp area (depending on the quantity of data involved).

There are clearly inefficiencies and risks involved in the current procedures:

- loss of data due to lack of backup procedures;
- loss of data on media which are either obsolete or decaying;
- loss of knowledge about data due to staff leaving;
- wasted effort due to duplication of programs/macros used to quality control data;
- inefficiencies due to staff having to learn by their own mistakes rather than being able to follow established procedures.

2 THE CHARACTER OF CEH WALLINGFORD PROJECTS

This section describes the projects which have a significant requirement for the input, quality control and archiving of hydrological data.

2.1 LOCAR/CHASM

It is assumed that the reader is familiar with the NERC LOCAR and CHASM programmes. If not, information can be found on the NERC website, http://www.nerc.ac.uk.

LOCAR has three catchments, the Tern, Frome/Piddle and the Pang/Lambourne, of which Wallingford staff will be responsible for field data collection in the latter. Staff from Wallingford will be responsible for field data collection in the Upper Severn catchment for CHASM. The remaining three catchments are the Oona, Feshie and Eden.

The roles of the Catchment Technical and Support Services Teams (CST) and the Data Centre are defined for the LOCAR programme. Both functions have been laid out in tender documents in some detail. In effect, the onus for quality control has been put on the CST with the Data Centre responsible for checking consistency in the data. However, the Data Centre may have to be responsible for some quality control for data from Special Topic projects as it is not certain to what level these data will have been checked. The situation is not as clear yet for CHASM but, since staff at Wallingford will be responsible for the Data Centre and making measurements in the upper Severn Catchment, there will be a requirement for data processing and quality control at Wallingford for some of the CHASM data.

2.1.1 Catchment technical and support services teams

The types of data that the CST will be responsible for are essentially the same in LOCAR and CHASM. For LOCAR they are defined in 'The Duties and Responsibilities of the Lowland Catchment Research (LOCAR) Catchment and Technical and Support Service Teams (May 2001)' and so only a summary will be given here. The measurements include:

- precipitation
- climate (using AWS)
- surface fluxes (using Hydras)
- soil water content
- soil water potentials
- river flow
- water quality (continuous monitoring and bulk samples)
- river sediment (turbidity)
- groundwater levels
- groundwater chemistry (continuous monitoring and bulk samples)

There is also a requirement to collect bulk water samples, generally at monthly intervals, which will be sent to laboratories and the CSTs will quality control the results of the analyses. The time interval of the data vary between 15 minute and monthly. The only spatial data that will be generated by the core programme are yearly land cover maps.

There are currently plans for 25 sites in the Pang/Lambourne, seven of which will be visited weekly, seven fortnightly and eleven monthly. The exact facilities at each site varies, e.g. ten of the sites only have boreholes.

There is a requirement for the CST to secure the raw data. Quality control is specified to occur "as close in time and space to the moment and point of observation as practicable" and includes:

- clock errors such as flat battery, out of sync with GMT, running fast/slow;
- datum errors such as local bench mark against OS bench mark, instrument reading against local bench mark, sensor drift, spike in record, unusual rates of rise or fall;
- range checks such as site specific range checks (e.g. level), variable specific range checks (e.g. $0<pH<14$);
- inter site checks within the catchment and within-site checks against similar instruments, continuous records against period totals, e.g. rainfall;
- data are internally logical, e.g. dry bulb> wet bulb temperature, solar > net radiation, solar < theoretical maximum solar;
- all checks currently regarded as forming part of best practice.

The CSTs will need the skills, hardware and software to:

- make field measurements;
- interrogate and download loggers;
- receive telemetered data;
- receive third party data, e.g. flow data, groundwater levels and fish counts from the EA;

- maintain a catchment diary;
- store the data securely for the duration of the project;
- browse and edit the data as text and graphically;
- quality control the data;
- process the data;
- track data that are being processed by other organizations (e.g. water quality samples, samples for radio carbon dating, etc);
- create and maintain meta-data;
- create and maintain appropriate QA documents;
- transmit the data to the Data Centre;
- receive data returned for error correction.

The current estimate of the date for the start of data acquisition is in March 2002, at the earliest.

2.1.2 The Data Centre

The definition of the duties of the Data Centre are given in the LOCAR Data Management Plan. Implicit in these duties is a requirement for skills, hardware and software that will enable the Data Centre to:

- receive data from the CSTs;
- receive data from the Special Topic PIs and third parties;
- reformat data including changing format, unit and code conversion, data model conversion;
- carry out limited quality control of data;
- perform higher level quality control, e.g. water balance checks;
- edit data;
- maintain an audit trail of change;
- load data into an archive, including database integrity protection;
- maintain meta-data;
- selectively retrieve data;
- export data;
- dispatch data to users;
- provide facilities for direct access to the archive by programs;
- create and maintain web sites;
- publish data;
- exploit data;
- maintain OPI's.

In addition to the data collected by the CSTs, the Data Centre will be responsible for archiving the data from the science projects. At this time an initial survey of the data requirements of these projects is underway but it is too early to be certain what the data types will be. Nevertheless, it is likely that much will be similar to that being a collected by the CST. However, there will be some types that are not and some of these will be spatial datasets. Therefore there is a requirement for the Data Centre to be able to handle spatial data.

2.2 AWQARD

This infrastructure award is primarily concerned with providing improved field instrumentation across CEH. The full description can be found in the original proposal so only a summary will be given here.

Wallingford – 9 Hydras (capable of measuring H_2O and CO_2 fluxes) with telemetry

Edinburgh – 1 Hydra and a truck based laboratory for measuring gas emissions
Windermere – Automatic Water Quality Monitoring Stations (AWQMS), 4 bio and 11 thermo
Bangor – 6 Automatic River Monitoring Systems (ARMS), 4 Wallingford Integrated System for Environmental monitoring in Rivers (WISER), 1 LiCOR closed system CO_2 flux station.

Although there is no overt provision for a data centre, NERC scientific projects are now required by the NERC Data Policy to offer their data to the appropriate Designated Data Centre. This is to ensure their long-term security and availability to future projects. Therefore, it should be anticipated that data from these instruments will be available for the Data Centre at Wallingford. It is also clear that these instruments will generate significant amounts of data which will need quality assurance to be defined and quality control to be carried out. At the very least, staff at Wallingford will be downloading data from nine Hydras and carrying out quality control on these data. It would seem sensible that the procedures used at Wallingford should be available to staff at other CEH sites to handle similar data.

2.3 OTHER PROJECTS

Field measurements are predominantly carried out in the Process Hydrology and Water Quality Divisions. Nevertheless, there is a requirement in the Water Resources and Environment and Hydrological Risks Divisions to quality control field data and data acquired from other sources. Thus all the Divisions at Wallingford have a requirement to quality control, edit and archive data. The hydrological data types in these other projects are mainly covered by those for LOCAR and CHASM. However, there are some additional data types, notably economic and social measures, plant physiology data, and spatial data of all kinds. Spatial data can be a significant part of some projects and, in particular, a number of digital spatial data sets are available for Plynlimon.

In addition, a number of projects are not based in the UK so there are occasions when it will not be possible to use the computing facilities at Wallingford. This may include occasions where one of the outputs of the project will be to leave a data archive and data processing facilities in-country. There are often occasions when data acquisition is not in the UK with the result that there will be a need for data processing and quality control in-country.

3 THE SOFTWARE REQUIREMENT

3.1 DATA INPUT, PROCESSING AND QC

Discussions with staff at CEH Wallingford produced a general consensus on a number of issues:

- Quality control should be carried out by the staff responsible for the data collection whenever possible.
- The best platform for data input and quality control software is a PC, because of the ease of connection to loggers etc. and also because the staff who will do the processing are familiar with PCs but not workstations.
- Automatic quality control is of limited value and tends to either detect only major errors or reports so many false warnings as to lose credibility.
- Graphical display and manual inspection is a very effective means of checking data.
- It is vital that the full history of the measurement system is available if errors are to be confirmed and corrected, implying that these data must be recorded electronically and kept with the raw data.
- Efficient procedures for data editing are required, e.g. a graphical data editor.
- The system must be easy to use.
- The system should be upgradable as new operating systems, database systems etc. become available.

It is also recognised that use of the data often shows up errors that had not been noted during the field quality control. Therefore it is very important that there is a clear communication line from those using the data, through those responsible for maintaining the data archive, to those who collected the data.

Although the preference is to use PCs for the data processing and quality control, this does not imply that the database being used will also be on the PC. If the processing is being done at Wallingford, then the PC will be able to use the Oracle DBMS through the LAN and thus the database can be on the database server.

3.1.1 Data loggers

Campbell Scientific data loggers are the most commonly used at CEH Wallingford so it would be desirable for any software to be able to link to these. However, it would appear that loggers being purchased for CHASM and LOCAR will be from a number of different manufacturers. It s unlikely that any software system is going to be able to download directly from all these. However, most logger manufacturers supply standalone software for PCs that enables the data to be downloaded and stored in simple files. In addition, data from the Hydras involves a significant amount of pre-processing and so will not be directly downloaded. Realistically, this implies that some or all the field data will have to be downloaded from the loggers into CSV files, or similar, and may need some reformatting before it can be loaded into the quality control software. A possible solution is the use of Excel with macros developed for the different instrument types.

3.1.2 Data management

For LOCAR, CHASM and other such projects, it is likely that the field teams will effectively be setting up their own archives, in parallel with that of the Data Centre, in order to be carry out the data quality control. At regular intervals they will retrieve data from their archive and send this to the Data Centre. If errors are shown up in the data after it has been sent to the Data Centre, care will be essential in ensuring that the corrected data is maintained in both the field teams archive and the Data Centre, and any users of the data are notified of the change to the data.

3.2 ARCHIVING

There is a clear first choice for the use of the Oracle database management system by staff at Wallingford, and other CEH sites. This has the advantages of:

- a rigorous backup procedure in place to protect the data;
- procedures available to limit or exclude retrieving, editing and/or input of data;
- long-term security of data independent of media/hardware involving evolution rather than radical change;
- a guaranteed upgrade path.

However, there are a number of issues which will need to be addressed if there is to be increased use of the Oracle database as the main archive.

Firstly, it would simplify matters considerably if there was an agreed data model (i.e. the nature of the tables and the relationships between tables). If we opt for commercial software for data processing and quality control then the data model for this is likely to become the de-facto standard but we are unlikely to know what that data model is. However, the commercial software provides routines to access the database, but there is an element of risk if we do not purchase updates to the software, or the software is discontinued.

Secondly, there is a general lack of awareness of, and the necessary skills to make use of, the Oracle database at CEH Wallingford (although there are some groups with considerable expertise). Therefore, it will be necessary to provide training in the form of a basic introduction to database concepts and specific skills necessary to work with the Oracle database on this site.

In addition, it is recognised that the Oracle DBMS may not be suitable for all data sets, e.g. work based overseas, but this could be catered for by using Microsoft Access as the DBMS but using, as far as possible, the same data model as used on the Oracle DBMS.

There is some disagreement as to whether raw data should be archived or not (there is total agreement that it must be retained). With the exception of a Hydra (24 Mb per day), the data volumes involved are not sufficiently large as to preclude the inclusion of raw data in a database. It is recommended that raw data (with the exception of Hydras) should be stored in the archive because all other data are derived from the raw data and experience has shown that the opportunity for errors in calibrating data, correcting for drift, infilling missing data etc. are major and happen all too often. Thus there is often a need to re-calculate data sets as errors become apparent with subsequent use of the data. Also, provided the amount of raw data is comparable to the amount of derived data that is archived then it is simpler to store and manage the data sets together, with a consequent reduction in the risk of the two data sets becoming separated.

Where raw data is not held in an archive, provision must be made for its long-term preservation. This implies stability of the physical medium, the hardware needed to access that medium, and the format the data are stored in. Currently, CDs fulfil these criteria best. However. even these have a finite life span so consideration must be given to copying the data at intervals of about ten years.

It would be preferable if no calibration was carried out in the loggers themselves but it seems unlikely that this will happen. It is very important that great care is taken to ensure the appropriate calibration values are maintained in the data loggers and that these are documented in electronic form and clearly linked to the relevant data.

3.3 DATA CENTRE

The function of the Data Centre has been separated from the consideration of archives in general because, although the Data Centre is by definition an archive, the converse is not true, i.e. there is a requirement at Wallingford to archive data which is not part of the Data Centre.

The Data Centre's terms of reference include data for the whole of the LOCAR and CHASM programmes and thus will include data types not regularly dealt with at Wallingford, e.g. bank erosion measurements. Based on the RACS(R) component of LOIS, this diversity of data types will include spatial data, e.g. remotely sensed measurements and data derived from these, subsurface geological data etc.

3.4 DISSEMINATION

The Data Centre also has a requirement to disseminate the data, in response to requests from the UK scientific community and so there is a need for the user community to browse meta-data in some form. Whilst LOIS was active, the Data Centre was receiving 50 to 100 requests for data per year (this does not include direct access to the data base by models).

The Wallingford site already makes Plynlimon data available to CEH users and the wider research community. There are about 30 requests for Plynlimon data (excluding chemistry) per year.

Currently, the LOCAR and CHASM Steering Committees have not made a decision about the method of data dissemination. For LOIS, data is generally disseminated by ftp-ing a file in response to an email request. In addition, a series of CDs were produced (print runs of 500). It is planned that data for URGENT will be available as flat files, i.e. not in a database, via the Web. It would seem likely that the future lies with some form of Web-based dissemination and so the presumption must be that this is what the LOCAR and CHASM steering committees will request.

In a Web based system, users would be given an id and password in order to control the data they have access to and to monitor usage, e.g. for OPIs. Users will expect to be able to browse the data, selecting on site location(s), variables and time period. They will also expect to see 'thumbnail' graphs of the data before selecting the data they require.

4 FUNCTIONALITY OF IN-HOUSE SOFTWARE

This does not consider *ad hoc* systems, such as macros for use with Excel. An in depth analysis has not been carried out but a set of criteria of an ideal system for data input, processing and quality control has been defined, in consultation with staff at Wallingford, and used as a basis in the evaluation, see Annexe 3.

4.1 NRFA SYSTEMS

The NRFA has a range of software used to quality control the flow data received from the Environment Agency. These carry out a thorough quality control and are dedicated to the format the data is received in and the particular Oracle tables used. They have been developed using a variety of programming languages. It would take a considerable effort to convert these into a system for wider use.

4.2 HYDATA

HYDATA was designed for hydrometric data and does fulfil many of the required criteria, although it does not have graphical data editing facilities. It runs under Windows 4.2 but it is in need of investment to bring it up to a fully functional form appropriate for modern computer systems. It is ODBC compliant and has been tested with the Access DBMS but not fully tested with the Oracle DBMS. There is a good user manual and internal training but the programs are not well documented. There are commercial systems that are capable of carrying out many of the same functions so the case for making the investment needed to upgrade HYDATA is not obvious.

4.3 SWIPS

Designed to handle manual measurements of soil water content and tension data obtained from neutron probes and tensiometers, this software fulfils most of the required criteria for these data types. However, it does not use either the Access or Oracle DBMS, although it is likely that it could be modified to do this without major effort. This limits the users ability to develop additional functions through software designed to link directly into the database. There is a user manual and the programs are documented adequately. There are no comparable commercial systems.

4.4 WIS

A very extensive software system which uses the Oracle DBMS in a client-server form. It includes a graphical data editor. However, the software is UNIX based (i.e. incompatible with PCs) and makes use of an obsolete graphics library which requires it to run under an operating system (SUNOS) that is no longer supported by the supplier. The result is that the software has to be run on hardware that already exceeds its design life. WIS would need major investment to enable it to work on workstations with the latest operating systems and even more to enable it to run on PCs. Nevertheless, it does have much of the functionality required for the role of a Data Centre, and might even be considered over specified. It currently handles data from NERC Thematic Programmes such as LOIS and URGENT.

4.5 LOWFLOWS2000

This is included because it incorporates a data loader that was developed out of WIS but has been updated and modified to work under Windows on PCs. Therefore the software may be a basis for something to be used by the LOIS/CHASM Data Centre, although the data model only operates with Access.

4.6 CONCLUSIONS

None of the in-house software provide the facility of graphical data editing on a PC that is seen by staff as being an essential feature for field data acquisition. With this exception HYDATA provides a significant number of the features, although there is some risk involved in its use due to the lack of a commitment to a longer term programme of upgrading and development. Although similar criticisms can be made against SWIPS, it provides a functionality not provided by commercial systems in terms of handling manual soil water measurements.

WIS was designed from the point of view of providing a generic database with flexible and efficient querying of disparate data types. As such, with the exception of having a graphical data editor, it is not optimised for input and quality control of field data. Rather, its strength lies in supporting a Data Centre. However, in order to continue to fulfil this role for existing Thematic programmes, let alone forthcoming ones, it is urgently in need of upgrading.

5 FUNCTIONALITY OF COMMERCIAL SOFTWARE

An exhaustive search for appropriate software has not been attempted. Rather this review has focussed on systems brought to the attention of the working group by staff at CEH Wallingford and by searching the web.

5.1 HYDROLOG

http://www.informetric.co.uk/Systems/systems.html

This is currently used by the EA for handling their hydrometric data (flow and water quality and claims to handle AWS) and fulfils many of the criteria required. However, the software, as used by the EA, is now elderly (it is essentially DOS based) and the EA is currently procuring a replacement for it. The company, HydroLogic, has put considerable effort into converting it to a single, Windows based, package and have made good progress in this direction and the product looks promising. However, they still have a lot to do to complete the task. The new product uses the Oracle DBMS.

5.2 HYDSYS

http://www.hydsys.com/

This software claims to have a wide range of functions and certainly does appear to be able to handle flow, water quality and groundwater meteorological data. It is a Windows based system with an intuitive arrangement of menus and was designed with water resources management in mind. It is recommended to use its own database management system but it is also capable of using other DBMS, such as Oracle. The company was taken over by Time Studio who are committed to maintaining it in the near future but plan to merge both products into a new system, Hydstra, over a period of a few years.

5.3 MAGPIE

http://www.mea.com.au/ms.html

This software appears to be dedicated to the Unidata Starlogger and MEA micrologger but is compatible with the SDI-12 sensor communication standard. It uses Dbase IV as its database. The promotional literature focuses on its use with AWS data and so it is unclear whether it is able to deal with other hydrological data types and what level of functionality it gives.

5.4 TIMESTUDIO

http://www.timestudio.com/su

This is one of the two systems that the EA considered as a replacement for Hydrolog to manage its hydrometric data. It appears to be a generic system, able to handle flow, water quality and meteorological data. It is Windows compatible and uses ODBC to communicate with the selected DBMS (which include Oracle and Access) and so is capable of acting as an archive. There is a dynamic link library that allows custom applications to connect directly. It has a direct link into Excel. There is an optional modelling module that includes flood and flow modelling. It was designed with real time data acquisition and modelling in mind. The company that has developed the software is Hydro Tasmania and there is a UK distributor (Ewan Associates Ltd., Stirling)

5.5 WISKI

http://www.jbsenergy.com/Instruments/Products/Software/software.html

This is the second of the systems evaluated by the EA and is the one that has been selected for their use. It is capable of being based on a LAN (i.e. there are client and server versions) and is Windows compatible. It is able to link to several types of DBMS (including Oracle and Access) and so is capable of acting as an archive. It is capable of handling a range of hydrological data types, including flow, rainfall, groundwater levels, water quality and snow depth. There are additional modules for discharge measurement quality control and rating curve creation. Frank Farquharson is exploring the potential for collaboration with the company and there may be scope for us to influence future developments of the software. The German company that has developed the software, Kisters AG, has about 160 employees and WISKI is one of three products aimed at different markets (although the software do share many basic functions).

5.6 OTHERS

There are others available, notably for automatic weather station data, but these tend to be provided by the manufacturer for their specific product and so do not fulfil the basic criteria of being generic.

5.7 CONCLUSIONS

There are three serious contenders, HYDSYS, WISKI and Hydrolog, that meet a reasonable number of the criteria and where there is a clear commitment to the long-term support and development of the software by the supplier. A more detailed evaluation is necessary to determine which of these is technically better in terms of meeting our requirements.

6 SOLUTIONS

We have considered the potential solutions under the three different functions of data acquisition, archiving and the Data Centre. However, in practice, these cannot be considered in isolation because the functions are inter-linked. For example, the output formats of the data processing and quality control stage should, ideally, be the same as the input formats to the archive and/or Data Centre. Therefore, we have taken these linkages into account.

We have dismissed the possibility of either upgrading in-house solutions or developing new ones as impractical, due to the time scale involved and the lack of resources. We have also taken the approach that it is not necessary for any single software package to deliver the complete solution. Rather we have identified the software that meets most of the requirements and then identified ways of meeting the rest of the requirement

6.1 DATA INPUT, PROCESSING AND QC

In terms of in-house software, the discussion in Section 4 shows that there are only two in-house software systems, HYDATA and WIS, that could form the basis of a solution in that they are capable of handling most of the types of measurements that will be made. HYDATA is PC based but it does not include a graphical data editor. WIS was not designed with field data acquisition in mind and cannot run on a PC. Nevertheless, it does use the Oracle DBMS and it has a graphical data editor. WIS is looking very dated as it has not been upgraded to the modern computing environment. This could be accomplished but it would take some time, ca one year, and would involve significant expenditure, ca £200K. HYDATA could be used but there would need to be some investment to enable it to run under the latest operating systems and a figure of about £100K would be needed. However, there is the potential to recover a proportion of these costs back through sales. In the short-term, HYDATA could be used in its current form to provide some of the functionality required.

There are three potential commercial software candidates, HYDSYS, WISKI and Hydrolog, see Section 5. All have good provision for input of data and routines for data quality control, including a graphical data editor. All can use Access or Oracle DBMS. On the information currently available, there is little to chose between them. External factors that might influence a preference are that the EA has selected WISKI (and thus data transfers between them and us might be simplified) and whether an agreement for additional collaboration can be reached with any of the suppliers

In order to help clarify the cost effectiveness of the potential solutions we have chosen a set of performance criteria and awarded a ranking to each of the software, i.e. HYDATA, WIS, HYDSYS, WISKI and Hydrolog. The ranking is very simple 1 = poor, 2 = satisfactory, 3 = good. The criteria are

1. how well does it meet the functions for input, processing and QC of field data?
2. software environment e.g. ease of use, uses current computing environment etc.;
3. initial cost, i.e. purchase price;
4. hidden costs, e.g. staff training, software support etc.;
5. long-term development and upgrade commitment of the supplier.

We have included two different scenarios for the in-house solutions: without any further development and with further development. This analysis is necessarily crude and subjective and so the results should be use with considerable caution.

Quotes from the three commercial suppliers were obtained against the same specification of a license for five years. However, the pricing structure for the products is different so a direct comparison is not simple. The results are summarised in Table 1.

The results, shown in Table 2, suggest that the most cost effective solution is either the HYDSYS or WISKI software. Hydrolog comes out third and HYDATA fourth, but only if investment is made to develop and upgrade the latter. The in-house software without further development does not score well, confirming that this is not a viable solution. Thus, the most cost effective solution is to use either HYDSYS or WISKI.

Following this, a limited technical evaluation was carried out by Matt Fry, the results of which are given in Annexe 4. A small group of CEH Wallingford users met to discuss the information and to achieve a consensus on which product would best serve our requirements.

Table 2 A comparison of the costs of three commercial systems

	Hydrolog	HYDSYS	WISKI
number of users at CEH Wallingford	20	unlimited	3 "named users"
purchase cost	35000	12000	22500
installation & training costs	unknown	3500 + 350 per day	installation included
5 years maintenance costs	42000	11400	13500
Total	**77000**	**26900 +**	**36000**
CEH Dorset and B'ham University		10320 (5 seats)	15000 (2 extra named users)

NOTE - these costs exclude VAT

Table 2 Cost effectiveness analysis of software for field data input , processing and QC

Software	Performance criteria					Average score
	1	2	3	4	5	
HYDATA – without upgrade	2	2	3	2	1	2.00
HYDATA – with upgrade	2	3	2	2	2	2.20
WIS – without upgrade	2	1	3	1	1	1.60
WIS – with upgrade	2	3	2	1	2	2.00
HYDSYS	3	3	1	3	3	2.60
WISKI	3	3	1	3	3	2.60
Hydrolog	3	3	1	2	3	2.40

Hydrolog gave the impression that, although it has been significantly improved over the last twelve months, it still requires significant investment to be technically on a par with the other two. The quote we have received is the highest but we would anticipate that they would be prepared to significantly reduce this if we were to approach them (the first quotes from the other suppliers were also high). HydroLogic are the smallest of the three companies and so it is reasonable to assume that there is increased risk in terms of long term commitment.

Although HYDSYS performed well, there was the impression that the software is still in the process of development. The company that produced HYDSYS has been taken over by a company that produced a rival product (TimeStudio). The merged company has plans to migrate the two products to a single product that will utilise the best points of both. This process will take several years so there is a risk of significant change occurring over the life time of LOCAR and CHASM. The company producing the software is owned by a large company, Hydro Tasmania, who use the software as part of their business. Therefore this option probably involves the least risk in terms of continuing development and long term commitment.

WISKI was assessed as technically the better product, basically because it gave the impression of its components all being at the same level of development and of a good, consistent design. The company is of a reasonable size (ca 160 employees) and has similar products aimed at two other markets. This would suggest relatively low risk in terms of long term commitment. This product has been adopted by the EA to handle its hydrometric data which might prove an advantage in terms of access to the EA's data. Our understanding of the definition of a "named user" is that this refers to projects/data sets and not individual user-ids. If this is correct then it would be best to get five named users (LOCAR, CHASM, AWQARD and two others for other activities). A new quote was obtained for five which gave a price of £45 k.

HydroLog and WISKI can work with Oracle 8.1.7 which is available at this site. (WISKI has been tested in this form but we have made no attempt to test HydroLog in this mode). HYDSYS using Oracle is at the stage of beta testing. The suppliers have indicated that the products can be installed within 4 weeks of receiving the order.

HYDSYS, WISKI, Hydrolog and HYDATA cannot handle efficiently manual measurements of soil water. This is because they do not consider the possibility of data being a depth (or height) series as well as a time series, with the result that they do not have the concept of a profile. This concept of a profile is important in the way the data is collected, quality controlled and processed. SWIPS is developed to handle these data and so could fulfil this role for LOCAR and CHASM. However, there are no plans for any further development of SWIPS and so this can not be considered a solution beyond the life time of these thematic programmes. As there are currently no comparable commercial software solutions, there is no obvious way out of this other than CEH Wallingford investing in a development programme for SWIPS. This does not need to start now and so the most pragmatic solution is to continue using the current version of SWIPS and to review the situation in about three years time when the situation may be different.

Although the amount of spatial field data is minimal for LOCAR, CHASM and AWQARD, we have briefly considered spatial data as it is relevant to other projects. Given that the handling of spatial data at the Wallingford site is dominantly carried out using Arc-Info and/or Arc-View (available through CHEST), and that these products have good facilities for the input, processing and quality control, it is sensible to continue with their use. The one proviso is that the supplier, ESRI, has launched a major upgrade in the form of ArcGIS 8 in 2001. This has some improvement in the algorithms available but is considerably easier to use than its predecessors. Although ESRI has not stated that they plan to phase out the previous products, it is fair to assume that this is their long-term objective. A further advantage of ArcGIS 8 is that it can use the Oracle DBMS, whilst the previous products used ESRI's own DBMS. Therefore it would be sensible if any new projects were to use ArcGIS 8 and there should be a migration of spatial data from ArcInfo and ArcView to this new system.

6.2 ARCHIVING

Data will be held by the Data Centre for LOCAR, CHASM and, presumably, AWQARD and so archiving is not really an issue. However, for other projects the data will not necessarily be held by the Data Centre and so archiving must be considered.

The commercial software systems, HYDSYS, WISKI and Hydrolog can use the Oracle DBMS and therefore are capable of acting as archives for much of the data. For the same reason, WIS is also capable of acting as an archive, but HYDATA would need to be tested before it was confirmed that this was possible.

There are advantages in using the same software system for archiving data as that used for input, processing and quality control of the field data:

- reduction in the staff effort required;
- minimising the risk of errors occurring in the process of transferring the data;
- the data in the archive is always up-to-date.

Therefore, the most cost effective long-term strategy is to use the commercial software, i.e. either WISKI or HYDSYS.

A disadvantage of using HYDSYS, WISKI or Hydrolog is that they are not able to handle all the hydrological data types collected. WIS is capable of handling a much more diverse range of data types and so is better placed to provide a single, integrated archive. However relatively few staff have experience of using WIS or of accessing the data from the WIS data model on the Oracle DBMS. Therefore the code for routines to retrieve and access data from the WIS data model would have to be written. In addition, there will also be a need for staff to be trained in the use of WIS.

Given that much of the time series data handled by staff at Wallingford can be handled by WISKI or TimeStudio and much of the spatial data by ArcGIS 8, then the most cost effective solution is to use the same software for archiving as is used in the input and quality control of field data, provided that the software uses the Oracle DBMS at Wallingford. This would also apply to staff at other CEH sites.

SWIPS does not use the Oracle DBMS and so some other method of archiving data must be sought for manual soil water data. The pragmatic solution is to copy the data into WISKI or TimeStudio as time series for each measurement depth.

6.3 DATA CENTRE

The Data Centre will be required to handle not only data from the CSTs but also the science projects in LOCAR. Thus it is likely to handle a wide range of data types, which would be simplified if a single system were able to do this. Given that HYDESYS, WISKI and TimeStudio are designed for hydrological time series, they would not fulfil the task as well as WIS. In addition, the Data Centre staff are already experienced in using WIS. However, as described in Section 4, the WIS software must be updated but not all the functionality of WIS is required for the Data Centre role. It is the data model that is the essential element.

To achieve the update, existing prototype software for data loading, meta-data and dictionary preparation and maintenance will have to be developed into an operational system. The work will include the development of SQL macros and an object model to give modellers direct access to the database. The loader addresses many of the issues above and includes: modes of operation (listen, interactive, batch); quality control (database, range checks, historic record, laws of physics/chemistry); quality assurance (audit trail, meta-data); database update modes; dictionary driven; and unit and code conversion. The tasks to develop the proposed system are:

1. Document the input format.
2. Convert the data loader from Access to Oracle.
3. Convert the loader database from WIS version 3.0 back to WIS version 1.8.
4. Add any data types (real, integer, char, co-ordinate, paired values, etc) required by LOCAR and not in the WIS version 1.8.
5. Write a set of SQL routines and an object model to provide user access to the database.

6.4 DISSEMINATION

There is significant uncertainty until the LOCAR and CHASM Steering Committees make a decision on the mode of dissemination they want. It is possible to carry on with the present system, i.e. using ftp, diskette or CD in response to requests. However, it is clear that the Web is seen by most people as their first choice of searching for and acquiring data. This perception is likely to increase rather than diminish. The inability to provide this does not give a positive impression of the capabilities of this site.

Ideally, it would be best if there was an initiative at a CEH level to share expertise and resources and to make sure that specific data sets had a common appearance and functionality. Enquiries to CCS established that this is not currently the position

7 RESOURCES REQUIRED

7.1 DATA INPUT, PROCESSING, QC AND ARCHIVING

The anticipated direct cost is about £45k to purchase the commercial software, WISKI. from the AWQARD computing budget. There will be some hidden costs in terms of staff training .

7.2 DATA CENTRE

It is proposed to develop the functional specification and high level design using existing CEH staff. Ideally, the programming would also be carried out within CEH as this would utilise the experience from developing the prototype and knowledge of the database design to best effect. However, there are a number of projects in hand with the result that existing staff are unlikely to be available so it will be necessary to subcontract out the programming work to a software house.

Funding for part of this can be found from a number of sources:

LOCAR	£20k
CHASM	£10k
AWQARD	£37k
Right (BNSC funded)	£10k

In addition, there are some funding proposals which, if successful, would also contribute:

HarmonIT	£10k
HarmoniRiB	£10k

However, this leaves a shortfall of about £100k which would need to be made up from central funds over a period of 4 years.

The only person at CEH Wallingford with the experience required to lead this activity is Dr Roger Moore

7.3 DISSEMINATION

It is not possible to identify any resources available at this time. Currently, CEH Wallingford is very poorly placed to meet these expectations as no staff currently have the required skills let alone any experience. The Water Resources Section has used a Sandwich Course student to make some progress towards this. here is an initiative with ITSS, who have offered to contribute half the cost of creating a Web-enabled GIS system for the spatial data from Plynlimon. This could be used as the prototype for other systems. However, there is no source of funding for the remaining half.

8 ANNEXE 1 - LIST OF ABBREVIATIONS

ARMS	Automatic River Monitoring Systems
AWQMS	Automatic Water Quality Monitoring Stations
AWS	Automatic Weather Station
BGS	British Geological Survey
CCS	CEH Computer Services
CEH	Centre for Ecology and Hydrology
CHASM	Catchment Hydrology and Sustainable Management
CHEST	Combined Higher Education Software Team
CST	Catchment Technical and Support Services Team
DBMS	database management system
DTM	Digital Terrain Model
EA	Environment Agency (UK)
GIS	geographic information system
HAPEX-Sahel	Hydrology-Atmosphere Pilot Experiment in the Sahel
iTSS	Information Technology Solutions and Services
LAPP	Land Arctic Physical Processes
LOCAR	Lowland Catchment Research
LOIS	Land Ocean Interaction Study
NERC	Natural Environment Research Council
NRFA	National River Flow Archive
OPI	Output Performance Indicators
PI	principal investigator
QA	quality assurance
QC	quality control
URGENT	Urban Regeneration and the Environment
WIS	Water Information System
WISER	Wallingford Integrated System for Environmental monitoring in Rivers

9 ANNEXE 2 –GLOSSARY OF COMPUTING TERMS

Of necessity, this report has used a number of technical terms which may be unfamiliar to those not actively involved in the topic of databases and software. Definitions of the most common of these terms are given below. Definitions of other terms can be found at http://foldoc.doc.ic.ac.uk/foldoc/index.html

Archive

A long-term storage area for backup copies of data or for data that are not in active use.

Attribute

A named value or relationship that exists for some or all instances of some entity and is directly associated with that instance, e.g. pH would be an attribute of a water sample.

Client

A computer system or process that requests a service of another computer system or process (a "server") using some kind of protocol and accepts the server's responses. It is part of a client-server software architecture.

CSV

A widely used portable file format. Each line is one entry or record and the fields in a record are separated by commas. Commas may be followed by arbitrary space and/or tab characters which are ignored. If field includes a comma, the whole field must be surrounded with double quotes.

Database

An integrated collection of data that supplies information in a variety of forms and for a variety of applications.

Database administrator

An individual responsible for the design and management of the database and for the evaluation, selection and implementation of the database management system.

Database management system (DBMS)

A suite of programs which typically manage large structured sets of persistent data, offering ad hoc query facilities to many users. It can be an extremely complex set of software programs that controls the organisation, storage and retrieval of data (fields, records and files) in a database. It also controls the security and integrity of the database. The DBMS accepts requests for data from the application program and instructs the operating system to transfer the appropriate data. Data security prevents unauthorised users from viewing or updating the database. Using passwords, users are allowed access to the entire database or subsets of the database. The DBMS can maintain the integrity of the database by not allowing more than one user to update the same record at the same time. The DBMS can keep duplicate records out of the database; for example, no two customers with the same customer numbers (key fields) can be entered into the database. Examples are Oracle and Microsoft Access.

Database server

A stand-alone computer in a local area network that holds and manages the database. It implies that database management functions, such as locating the actual record being requested, are performed in the server computer.

Data centre

An organisation tasked with receiving, storing and disseminating data. It may also be responsible for some or all of the following activities: field work, quality control, analysis, quality assurance, publication in various forms, maintenance of meta-data, exploitation, software development

Data loader

Software that uses data input from flat files to build a multi-dimensional database.

Data model

The product of the database design process which aims to identify and organise the required data logically and physically. A data model says what information is to be contained in a database, how the information will be used, and how the items in the database will be related to each other. For example, a data model might specify that a customer is represented by a customer name and credit card number and a product as a product code and price, and that there is a one-to-many relation between a customer and a product.

Field

An area of a database record , or graphical user interface form, into which a particular item of data is entered.

Flat file

A single file containing flat ASCII representing or encoding some structure, e.g. of a database, tree, or network. Flat files can be processed with general purpose tools and text editors but are often less efficient than some kind of binary file. They are more portable between different operating system and application programs than binary files, and are more easily transmitted in electronic mail.

ftp

A client-server protocol which allows a user on one computer to transfer files to and from another computer over a TCP/IP network. Also the client program the user executes to transfer files.

GIS

A computer system for capturing, storing, checking, integrating, manipulating, analysing and displaying data related to positions on the Earth's surface. Typically, a GIS is used for handling maps of one kind or another. These might be represented as several different layers where each layer holds data about a particular kind of feature (e.g. roads). Each feature is linked to a position on the graphical image of a map. Layers of data are organised to be studied and to perform analysis.

LAN

A data communications network which is geographically limited allowing easy interconnection of terminals, servers, peripherals and computers within a single building and, if required, adjacent buildings.

Meta-data

Data about data. In data-processing, meta-data is definitional data that provides information about or documentation of other data managed within an application or environment. For example, meta-data would document data about data elements or attributes, (name, size, data type, etc) and data about records or data structures (length, fields, columns, etc) and data about data (where it is located, how it is associated, ownership, etc.). Meta-data may include descriptive information about the context, quality and condition, or characteristics of the data.

ODBC

A standard for accessing different database management systems. An application can submit statements to ODBC using the ODBC flavour of SQL. ODBC then translates these to whatever flavour the database understands.

Quality assurance
A planned and systematic pattern of all actions necessary to provide adequate confidence that the output optimally fulfils expectations, i.e. that it is problem-free and well able to perform the task it was designed for.

Quality control
The assessment of output compliance with stated requirements. Quality control should be independent from processing.

Record
An ordered set of fields. The term is used in both files (where a record is also called a "line") and databases (where it is also called a "row"). In a spreadsheet it is always called a "row". In all these cases the records represent different entities with different values for the attributes represented by the fields.

Relational DBMS
A relational database management system allows the definition of data structures, storage and retrieval operations and integrity constraints. In such a database the data and relations between them are organised in tables. A table is a collection of records and each record in a table contains the same fields. Certain fields may be designated as keys, which means that searches for specific values of that field will use indexing to speed them up.

Server
A program which provides some service to other (client) programs. The connection between client and server is normally by means of message passing, often over a network, and uses some protocol to encode the client's requests and the server's responses.

SQL
An industry-standard language for creating, updating and, querying relational database management systems.

TCP/IP
The de facto standard European protocols incorporated into 4.2 BSD Unix. TCP/IP was developed for internetworking and encompasses both network layer and transport layer protocols. While TCP and IP specify two protocols at specific protocol layers, TCP/IP is often used to refer to the entire DoD protocol suite based upon these, including telnet and ftp.

10 ANNEXE 3 - CHARACTERISTICS OF FIELD DATA INPUT AND QC SOFTWARE

In order to evaluate both in-house and commercial systems we have, in consultation with other staff, defined a series of characteristics based on the assumptions that all data are time series and are 'point' measurements.

10.1 MANDATORY REQUIREMENTS

1. Software runs on a PC/laptop
2. It must be easy to use
3. Able to store/retrieve information to Access and Oracle DBMS
4. Store site location as: Site name, ID
5. Store instrument location as Site name/ID, X and Y grid co-ordinates, ground elevation, measurement height/depth.
6. Include instrument diary, i.e. when visited, changes made etc.
7. Input data from flat or CSV file or manually
8. QC procedures for simple error identification and reporting (e.g. outside user defined bounds)
9. Include flags to indicate data quality/history
10. Apply instrument and environmental calibrations
11. Calibrations are time variant, including corrections for long-term drift
12. Calibration procedures for linear, second order polynomial, user defined
13. Procedures available for developing rating curves
14. Handles measurements which are either event based or continuous
15. Store raw (protected against editing) and calibrated/edited data and include missing data.
16. Ability to graph raw and calibrated data as time plots (integrated with previous data)
17. Ability to edit data using a graphical editor
18. Ability for user to edit raw data values (but the edited values are not stored) prior to calibration.
19. No software limit to amount of data handled or stored.

10.2 DESIRABLE REQUIREMENTS

1. Runs under Windows 95,98, 2000 and XP
2. Links to any database by ODBC
3. Able to handle generic data (i.e. not specific to a particular measurement type)
4. Able to store instrument's geographic co-ordinates i.e. latitude and longitude
5. Indicators for whether measurements are time-interval or time-averaged, also for depth interval
6. Time variant description of instrument, e.g. manufacturer, model number, accuracy and precision, method of operation etc.
7. Dictionary of 'standard' instruments available to user
8. Facility to download data directly from loggers
9. Ability to 'remember' the input data format for particular sites, i.e. which columns are which variables.
10. Ability to store 'derived' data, e.g. calibrated data aggregated to a longer time period
11. Ability to optionally include lines showing min, max and average values on time plots
12. Ability to compare (graphically and statistically) and quality control using data from other sites
13. QC by checks with other variables (e.g. ion balance, solar radiation < net radiation etc.)

11 ANNEXE 4 – TECHNICAL EVALUATION OF COMMERCIAL SOFTWARE

Matt Fry evaluated three commercial systems over a short period (1 day each) and assessed their immediate usability in storing and analysing water level and flow data in comparison to HYDATA. The following gives his thoughts on each system from this brief evaluation as well as general comments that could be useful to other users.

The systems tested were:

- Hydrolog
- HYDSYS
- WISKI

Brief summary of results:

System	Environment	Major Advantages	Major Disadvantages
Hydrolog	Local drive	Very intuitive – easy to use and sensibly laid out Excellent flexible reporting	No user defined parameters or units Have to wait until next release (up to 3 months)
HYDSYS	Local drive	Lots of functionality including analysis routines Flexible parameters and units Good graphical editing facilities	Not sure about use on ORACLE Slightly confusing and non-intuitive
WISKI	Networked ORACLE database	Flexible parameters and units Sensible method for linking derived time series Good graphical editing Chosen by EA	Could be slow - this needs testing Cannot currently store rating equations – they claim to be changing this for the EA

11.1 BACKGROUND

Matt Fry's experience is mostly with HYDATA and so he evaluated these systems with reference to the functionality within HYDATA.

HYDATA is the institute's Hydrological Database and Analysis software used for national and project hydrometric databases around the world, particularly in Africa. It is now in version 4.2 – fully windows based with flexible graphing and good reporting options.

HYDATA stores time series data of any parameter – typically flow, stage, rainfall and storage but any new parameter can be defined. Multiple time series are stored against a station, usually representing a hydrometric or meteorological recording station. Data is stored at fixed intervals, either with a fixed gap between intervals (from 1 second to 1 day plus weekly, 10 daily, monthly and annual intervals) or at fixed times of the day (8:00, 18:00). Users tend to make full use of all of these options and all possible combinations of obscure interval data has been found and stored.

Users can enter data by hand or import data from files of a number of formats. Data can be edited within an editing window offering interpolation, datum adjustments, setting of quality flags, commenting of data but no graphical adjustment of data values.

There is a sophisticated gaugings and ratings editor where gaugings can be entered and rating curves (power law or polynomial) developed. Rating curves can be multi-part and there can be a number of versions each applying over a particular period.

There are a number of analysis options within HYDATA – FDC, double mass plots, low flow analysis, BFI plots and analysis, etc. of which FDC is the most common used.

A number of modules exist for more detailed importing and exporting and low flow analysis.

The 3 commercial systems were evaluated while performing the following tasks:

- Importing some sub-daily water level data exported by HYDATA
- Entering existing gaugings and rating curves, also trying to develop ratings
- Converting water level to flow using these ratings

The systems were not evaluated looking at water quality, meteorological or groundwater data and the issue of performance under heavy usage was not assessed.

11.2 CRITERIA

The systems evaluated on their performance in the following areas:

General

- What environment is the program running in?
- How intuitive is the system for the first time user?
- How fast are the standard operations?
- How are the help files?

Configuration

- How does the database store stations, time series, parameters, intervals, units
- How does the user set these up?
- What are the advantages / disadvantages of the system?

Data import

- How does the system import data?
- Is this easy and quick to use?

Data editing

- How does the standard data editing suite work?
- Can data be graphically edited?
- Can the data be edited in a table and the results be viewed on a graph?
- Are there options such as interpolation, missing data, data gaps?

Gaugings and ratings

- How does the database store gaugings and ratings?
- How does the software allow these to be entered and edited?
- Can rating equations and tables be stored?
- Can rating equations be developed easily?
- Can water level data be simply converted to flow?

Reports

- What are the available report options?
- How flexible are these reports in terms of graphing, tables, layout, text formats.

11.3 HYDROLOG

11.3.1 General

Hydrolog 4 (with some modules seemingly Hydrolog 3) running on the local drive of 1GHz, 512MB RAM PC, ACCESS database for (what must be very small volumes of) station information and flat binary files for the time series data. Apparently the system can run with ORACLE (v 8.1.7) – each PC needs the ORACLE client software installed.

The software proved to be very intuitive to the first time user. There are very few modules (though this may point to more limited functionality) – Time series Editor, Mapping, Station management, reports and importing. No analysis functionality.

The windows are laid out very well, consistently and easy to read. The user can see at all stages what stations, time series, available and archived data exists.

Most operations can be carried out very quickly, though it was run on a fast machine.

Help files seem quite comprehensive though consistently slowed down the computer.

The program crashed several times when editing time series and reading the Help files. A database manager user has sensibly been set up to allow other users to be logged off and locked tables freed up when the system crashes.

11.3.2 Configuration

Stations are set up in a sensible way.

Station names are limited length text fields of only 20 characters.

Time series are stored as parameters (e.g. stage, flow) for each station. The available parameters are fixed, new parameters can only be created in the next release of the software in next release (generally 3 months). **This will be a major disadvantage.**

Only one time series of each parameter is allowed at a station. This is not ideal (for instance for soil moisture / evaporation series at different depths / heights. They have parameters available for 'SM @ 10cm', 'SM @ 20cm', etc. but this does not seem ideal.

Intervals seem strange. User can optionally define intervals for a parameter with good flexibility (down to 1 second) but can't convert between different intervals except at reporting stage. There is no 'data viewer' for graphs or numbers, except reports.
Units are also fixed, with few available options (mgl/day, etc.)

Sensible setup window and style making it very easy for first time user to create stations and time series and to start adding data. Very easy to find stations and parameters and periods of available data.

Gauging data cannot be stored within the software – only through an external (DOS) package.

11.3.3 Data editing

Data editing window works well, again very intuitive. Graphical data editing is minimal – user can edit comments, shift points up and down within same time interval.

The editor crashed or locked quite often.

User can't add missing flags – data must be deleted and a 'gap' denoted instead. User can't interpolate across periods of missing data. User can correct drift but not graphically.

No zoom out on X or Y-axis.

Manual graphical editing is quite poor. The editor is not in a standard windows interface and has a few quirks – similar to the HYDATA data editor and Gaugings and Ratings modules which used 3rd party components. These proved hard to maintain with time.

11.3.4 Ratings

No ratings creation within Hydrolog. This is performed from Gaugeman – DOS program, currently being developed for Windows.

Can add ratings equations to stations within station manager. The method for adding equations in parts and versioning is good. Only power law equation is available.

Conversion to flow is done real-time. Flow records can be viewed (as tables, not graphs) within the data editor and plotted as graphs and tables within a report. Flow data can also be summarised and summaries (of daily, monthly, etc. max, min, etc.) stored.

11.3.5 Reports

Good bespoke reporting. User can drag-and-drop labels, data tables and graphs to create report templates with multi-parameter graphs and single parameter tables – both fairly flexible. Templates can be saved and automatically applied to saved list of stations.

Report creation can be scheduled to a specific directory for viewing by non-hydrolog users. There is a Web version of this viewer currently under development.

11.4 HYDSYS

11.4.1 General

HYDSYS version 8 (.6.8) evaluation running on Windows NT, 1GHz, 512 RAM with databases (flat binary file) stored on local drive. Time series module only tested, not Mapping, Water quality, Groundwater, Modelling, Telemetry modules.

Module based program with very comprehensive list of modules, presumably with a lot of additional functionality. These run up from lists of menu items.
There are several aspects of the software that are quite confusing: the windows are different from standard Windows, many modules can be found in several different menus, there seems to be little consistency about forms and menu content.

Some operations (imports) are very fast. Some are slower.

Help files seem very comprehensive. Each module has a name starting with HY- (HYREPORT, etc.) and this can also be a little confusing.

There seem to be many analysis options.

11.4.2 Configuration

User must set up a site and each site has one single station, though it is not clear why. The station setup form is OK, though not the best way to show and edit this information – users can't simply list stations and time series without multiple clicks.

Each station has a number of data files, labelled from B to Z.
This is time series data of any parameter (letter's do not specify which parameter which can be confusing) that is being edited.
When editing is complete, data can be archived to a file labelled A (for archive).
Archived records can have any number of parameters in at once.

There is no need to create time series for stations as these are created on import as files with appendix (B-Z).

Parameters (rainfall, flow, stage) are denoted by numbers (stage is 100.00, flow is 140). This number doesn't really suggest which parameter it is and user must learn them or refer to lists provided within software.

This is not ideal for multi-parameter stations.

There is no way of summarising what data is stored at a station (parameters, start and end dates of each). The user may have multiple consecutive files (A, B, C, D) for stage, each imported from a raw data file.

Parameters are stored as base data and all conversions (e.g. level-flow) made in real-time. Data comes in files with date / time stamp to 1 second interval.

Files imported are stored as actual data at intervals read from file. This is therefore very flexible.

There are many parameters available as well as units and conversions but these were not tested. New parameters and units can be defined by the user. Not sure about whether series can be created for Max / Min / Mean, etc.

11.4.3 Data import

Many importing options and formats available. Only one tested. CSV files imported and by defining format of each line in file. i.e. "SSSS DDMMYYYY HHMMSS VVV.VVV QQ" meaning first four characters indicate station number, the next day, month, year, etc. This makes it very flexible, one limitation is that all values must be of same format (e.g. 4.56 must be written 004.560).

Easy to create files within Excel / Notepad. No header required. No strange formatting as it is user defined.

Worked fine – very fast at importing large amounts of data.

11.4.4 Data Editing

Data editing performed through the 'Data Manager's Workbench'. This allows several stations to be opened showing all data files in each. Each file can be opened either for editing or for brief viewing. Plots of different parameters can be drawn as a reference trace on both viewing and editing graphs.

The maximum volume of data is about 30000 values. Longer time series are divided into blocks.

The editor is good, fairly easy to use. Bulk operations can be done quickly. Data can be divided into monthly, annual or 'blocks'.

Graphs are OK, periods can be selected by zooming in and out. Points for editing can be easily selected and the data table is linked to the graph.

Some editing can be performed graphically – moving points up and down within a time interval, drawing a straight line, drawing a free line (questionable usefulness). Other operations can be performed on the data tables (adjust, set flags, n-point running filter, etc.) but not interpolation.

Graph properties are very messy and hard to understand. Graphs draw missing points as zero which is gives very poor graphs.

11.4.5 Gaugings and ratings

Gaugings are stored in a separate module – method of entering gaugings is very poor, very slow.

No time to look at fitting gaugings and ratings – looks detailed, but again – 5 or 6 gauging and rating modules make it very confusing.

Ratings are entered at station level. Not a particularly good method of entering ratings.

Flows only viewable as an output – not raw data.

On attempting to convert level to flow, received many errors if period had no rating, if data was above top of rating or below rating and flow series could not be viewed.

11.4.6 Reports

Seem to be many report options – predefined reports such as yearbooks, graphs, etc. Not that flexible in format though.

11.5 WISKI

11.5.1 General

WISKI v 5.5 running on the ORACLE database across the network on a P3 (?MHz) 64MB RAM PC with Windows 2000. Problems were encountered when installing ORACLE client software. Programs quite slow though this may be Win2000 and low memory or ORACLE connection across network. HYDATA runs quite fast on same machine – even across the network.

11.5.2 Configuration

The database stores stations in a standard way, time series are stored at each station, of many parameters.

Station names can only be 31 characters long. Stations and time series are easily viewed through the WISKI explorer. This also shows what period of data exists. Stations are created here also.

Some time series are automatically created for a station – Stage flow and Eta for a gauging station – and there are many default derived (ann max, month mean, etc.) time series.

Stations can apparently created from a user defined template which would be very useful. Stations can also be copied.

Parameters and units can be setup within the database –very flexible in the respect.

Time series have a base interval from 1 second upward. Time series are either manual – input – or derived from 1 or more other time series. The method for setting the 'origins' of a derived time series is good.

This method of storing the derived data within the database is good – and is the same as HYDATA – with real numbers stored for derived flows, means, annual max, etc. But they can automatically be recalculated e.g. if a rating is changed and the changes will pass on to the means, max, mins, etc.

These updates can also be scheduled for a low use period.

The large number of time series is originally quite confusing but well thought through.

Windows are generally quite intuitive. Graphs have lots of small buttons with indecipherable pictures on but this is probably because of extended functionality.

Gaugings are stored strangely, as a 'Q.Parameter' time series.

There are a few points where confusion is caused, maybe due to language differences.

No help files were available for the main system. The existence of these should be checked.

11.5.3 Data import

The system can import a number of different formats of data. I used the WISKI .zrx format – comma separated data with a header. This was created quite simply by formatting an existing CSV file in Excel and adding a header. The import was very quick.

There may be an issue of data being rounded to the nearest time interval on input (not good for 15-minute data stored at, i.e., 12:31:22, 12:45:22, etc.

Data can be very simply cut and paste from excel – gaugings were entered in this way. Again a small amount of formatting is necessary.

11.5.4 Data editing

The data editing facilities in WISKI are very good. Data is edited graphically with a linked table of the data displaying the actual values to the side. Many options exist for shifting points or periods of data graphically (drag-and-drop from graph, drift correction by fixing one end and dragging the other, inverting, etc.) and also by altering the numbers – interpolations, etc.

11.5.5 Gaugings and ratings

The gaugings and ratings are performed within a separate module called SKED.

This is quite slow to start up.

Gaugings are stored, as mentioned, in a time series. The flow series is set up to be derived from the level series – with a rating set as the method.

Within SKED the gaugings can be plotted and regressions applied. A number of ratings can be set with various preliminary / release, etc. versions each applying to a period of record. These ratings seem to be stored as a sort of table within WISKI – a number of stage / flow points between which flows are (presumably) interpolated.

There are a few places where it has been made to look as if the ratings can be stored as equations but these seem to be fixes. Apparently they are working on storing ratings as equations on behalf of EA but this needs looking into.

Currently a regression is drawn through a set of gaugings and then a number of points are selected (by the user's hand) along this regression line to effectively create a 'table' to simulate the equation. The problems are that this is not accurate or consistent and that it is very hard (almost impossible) to pick points from low to zero stage – so the equation is incomplete.

The method for calculating flows from stage (and other derived parameters) is very good as mentioned above.

11.5.6 Reports

There is an option for using Crystal Reports but it does not function. Graphs can be produced easily from the editing suite. Graph formats can be saved.