AN INTEGRATED SOFTWARE PACKAGE FOR THE ANALYSIS AND DISPLAY OF HYDROLOGICAL OR WATER RESOURCES TIME SERIES DATA

DA Hughes • D Forsyth • DA Watkins

WRC Report No. 867/2/00

Water
Research
Commission

AN INTEGRATED SOFTWARE PACKAGE FOR THE ANALYSIS AND DISPLAY OF HYDROLOGICAL OR WATER RESOURCES TIME SERIES DATA

Final Report to the Water Research Commission
on the Project :

"Integration and Application of Daily Flow Analysis and Simulation Approaches within Southern Africa"

by the

Institute for Water Research
Rhodes University
Grahamstown
South Africa

DA Hughes, D Forsyth and DA Watkins

IWR Report No. 2/2000
WRC Report No. 867/2/00
ISBN No. 1 86845 656 0
ISBN Set No. 1 86845 657 9

Table of Contents **Page**

LIST of FIGURES | Page

LIST OF TABLES **Page**

ACKNOWLEDGEMENTS

The project was funded by the Water Research Commission of South Africa under the direction of Research Manager, Mr Hugo Maaren to whom the project team are grateful for the support and guidance given during the project and for his very able chairmanship of the Steering Committee. The other members of the Steering Committee are also thanked for their contributions and constructive comments :

Dr GC Green, Water Research Commission
Prof. B Kelbe, University of Zululand
Mr BD Wolf-Piggott, WSR, University of the Witwatersrand & DWAF (Water Resources Planning)
Mr E Nel, DWAF (Hydrology)

Thanks are also due to the following individuals and/or organisations :

All those organisations who took part in the workshop held at the beginning of the project and listed in section 2 of this report.

Dr Mark Dent and Mr Shiresh Nundlall of the Computing Centre for Water Research for their advice and assistance in developing links between the IWR software and ArcView.

Mr Andre Greyling of Ninham Shand Inc. for help and advice on the use of database tables and storing time series data in BLOBs.

Dr Graham Jewitt, Mr Andy Pike and others of the School of Bioresources Engineering and Environmental Hydrology at the University of Natal, Pietermaritzburg for useful comments on the developed software.

Dr Alan Gustard, Mr Roger Moore, Dr Kevin Sene and others at the Institute of Hydrology, UK for sharing their ideas and concepts about the efficient storage of spatial and time series data and for demonstrating some of their developing software.

EXECUTIVE SUMMARY

The original proposal for this project referred to promoting the more widespread use of observed or simulated daily streamflow data in various fields of water resource decision making and management. One of the areas that was to be addressed was the determination and implementation of instream flow requirements (IFRs), which the hydrology section of the Institute for Water Research had started to get involved with during 1995. The perception was that there was room for improvement in the way in which the hydrological information, used within IFR studies and elsewhere, could be generated and made use of. Two main improvements were considered :

- The rapid preparation of daily time series of natural and present day streamflows.
- The presentation and use of this information by the non-hydrological specialists at the IFR workshops.

The first point has been addressed mainly by Dr Vladimir Smakhtin (who has subsequently left the Institute) who is the author of a separate report (WRC Report No. 867/1/2000) associated with this project and which deals with pragmatic approaches to estimating daily flow time series in data sparse areas. The second point is addressed through this report which outlines the development of software to display and analyse time series data. The intention of this component of the project was to rapidly develop tools for presenting the hydrological time series data to non-hydrological specialists in an informative, flexible and interactive manner. The software is still under development, but a prototype version has been available for some time and has been used in several IFR workshops, as well as being transferred to other University Departments where there are research groups working on related problems. It is believed that, although many improvements can be made to the functionality of the software, it has proved to be useful and has facilitated the transfer of hydrological understanding to other specialists involved in IFR workshops.

The need for a more generic approach to displaying and analysing time series data was identified by several groups, some that are involved in the development of related software (such as hydrological and water resource models) and some that are simply involved in the use of time series data (whether observed or simulated). Many of those involved in the use of such data are non-hydrological specialists who can be frequently confused by using different software packages all the time and who have voiced the opinion that they would prefer a single package that performs all the functions required. One of the major issues that arose during the initial design period was the format of storage for the time series data. There was a great deal of discussion and disagreement amongst the various groups consulted and it was apparent that no one format would be acceptable to all parties. The concept of a common format was therefore dropped in favour of ensuring that (eventually) the software would work seamlessly with a wide range of formats and that others could be supported by simple importing routines.

The beginning of the report discusses some of the design considerations that formed the basis of the software development (section 2), while the bulk of the report contains descriptions of the various components of the software (section 3). Section 3 is therefore a cross between a manual and an explanation of the reasons for including certain components. It should be recognised that the software is far from complete, in terms of what it has the potential to do. Section 3 therefore concludes with a 'to do' list, some of which have been identified by the development team and others by individuals or groups who have had an opportunity to evaluate the software. It is the intention of the project team to continue with the development of the software and to add the new facilities that have been identified as being required.

One of the initial objectives during the design phase was to be able to link the time-series display software with spatial data. A relatively simple (and somewhat clumsy) procedure has already been established using Avenue script within ArcView to define a search rectangle. The coordinates of the search are then used to select time series files or database records on the basis of their spatial position attributes and build a 'project', which is simply a reference to the sources of the data. The 'project' is the basis of the main display and analysis program and is used during applications to select which time series to work with. The fact that the interface with ArcView is somewhat restrictive caused the development team to look for alternatives; specifically the use of Map Objects - a package that allows ArcInfo coverages to be accessed and manipulated within Delphi program code. However, this search only took place toward the end of the project and there was only time to develop a revised design approach and not time to implement it. Section 4 describes the approaches that the development team will be attempting to implement in the near future. Some progress has already been made in this direction and at least a prototype version should be available during 2000.

The conclusions refer to the fact that the new software has not yet been integrated with the IWRs existing modelling software (developed some time ago using C language in a DOS environment). It is now considered that further work from 2000 onwards using some of the revised concepts will allow this integration to proceed rapidly and in a format that will be a great deal easier for users to understand and apply than was previously the case. The vision is the ability to set up and run hydrological models, analyse and display time series data, display other kinds of data (text memos, bitmaps, etc.) or generate new spatial coverage information all from a single spatial interface.

1. INTRODUCTION

This report forms the second of three parts of the final report on the project 'Integration and application of daily flow analysis and simulation approaches within southern Africa'. The other two reports address different components of the project but are all linked.

The first report (Smakhtin, 2000) describes some of the research that has concentrated on pragmatic methods of generating daily time series of flows for South Africa catchments using readily available data. The third report represents a component of the project that was expanded in importance during 1998 and 1999 due to the requirements of the Department of Water Affairs and Forestry in terms of quantifying the water quantity component of the ecological reserve (part of the new Water Act). This report (Hughes and Münster, 2000) describes the research that the IWR has conducted over the last three years into methods of providing the hydrological information and methodology requirements for determining instream flow requirements (with varying degrees of confidence) as a contribution to the techniques required to quantify and implement the Reserve component of the new South African Water Act.

While this report largely describes the development of some computer software, it has not been written in the normal style of a software manual. An attempt has been made to explain the motivation behind the development and the principles that guided the design process. The report does, however, include sections that explain what the software is capable of, as well as those that provide some guidance to its use. The final part of the report, which discusses future developments, is a reflection of the rapidly changing environment in which software development takes place these days. This section indicates that some radical changes are proposed to the approaches that were used during the project. However, that does not mean that the products of this project will not be of any future value. In fact, the software that has been generated by this project will be able to fit into the newly proposed structure quite well without a great deal of modification.

1.1 The need for integrated time series analysis and display software

Any hydrologist who has dealt with the vast amount of information associated with time series data is readily aware of the need for efficient tools for visualising and analysing such data. This is particularly true for those who grew up in the science before the advent of fast personal computers, spreadsheets and other proprietary software. Even as recently as the 1980s, the analysis of moderately long time series of hydrological data was never a simple task and there were very few facilities available for the generation of a wide range of graphical displays with which to visualise the information. Those facilities that were available were frequently associated with a specific software package and restricted to particular input data formats, a knowledge of which was frequently not available to anybody but the developers. Fortunately, this situation has improved and there is now a great deal more transparency about data formats, as well as the fact that there is probably more standardisation. Part of the reason for this is that, in the past, the computer storage space and the speed at which data could be read and processed were considered to be critical issues and every software developer had their own ideas about which approach was best. Some designers adopted a simple approach to data storage, using text data files that could be easily viewed and edited, but at the expense of reading and storage efficiency. Others adopted approaches based on reading and storage efficiency at the expense of being able to view the data using simple text viewers. Many of these issues have become somewhat redundant as storage space and reading speed are less critical today then before. There are also far more standard data management tools (database software, for example) available now than before and there is less need to write new dedicated software. However, there are still a great many arguments surrounding the issue of standard data formats for time series data and these will probably continue for some time to come.

Amongst hydrologists, those that have dealt with the design and application of hydrological time series models (whether deterministic or stochastic) should be the most aware of the need for a flexible and integrated approach to analysing and displaying time series data. There can be nothing more frustrating than having to spend a great deal of time manipulating the results of a simulation, simply to discover how well the model is performing. The more time that is taken to view and analyse the results, the less time is available to ensure that the results are the best that can be obtained - this applies equally to applying models as well as to developing them. Having to export data from a modelling system to an external graphics software package takes time and effort. However, the alternative has often been the duplicate development of post-processing software for results analysis every time a new model is developed.

There are also many non-hydrological specialists who are demonstrating an increasing awareness and interest in the value and use of hydrological time series. These potential users place different demands on software as they are frequently not as familiar with some of the standard approaches to displaying and analysing time series data as hydrologists are. They may also have particular requirements that are not standard as far as hydrologists are concerned.

1.2 Previous developments within the IWR at Rhodes University

The Institute for Water Research (and the former Hydrological Research Unit) have had a long history of model development and application, but it was only during the early 1990s that an integrated approach was adopted. Once again this was related to developments in computer hardware and software as well as a realisation that a careful investment in an integrated model system would be more than repaid in efficiency of model application at a later stage. This prompted the development of the HYMAS (HYdrological Model Application System) package which was designed to include a number of pre-processing (data input, re-formatting and editing procedures) facilities, model running facilities, as well as post-processing visualisation and analysis procedures within a single modelling shell that could be used to apply a range of different models (Hughes, et al., 1994). The package was written in 'C' for DOS-based PC systems and has been extensively added to and used by the IWR and others (for research and practical applications) over nearly a decade. While the original design of the system can be considered to be reasonably successful, it has been overtaken in many respects by recent developments in computer hardware and software. The most obvious development is the advent of the Windows operating system and the appearance of object oriented visual languages. Other developments, in terms of the speed of hardware, the amount of memory typically included on a standard PC, the removal of DOS restrictions in the handling of large data arrays and the large reduction in cost of data storage have all made some of the original design specifications somewhat redundant. HYMAS development started on relatively fast (at the time) 386 PCs with math-coprocessors , but could operate quite successfully on 286 PCs, as long as the user was prepared to be patient. A number of the data handling procedures were designed around the restrictions of the existing hardware that are no longer applicable.

While many of the original components of HYMAS were carefully designed, some of the later ones were forced into the original design and are less than satisfactory. In addition, extensive use of the software over several years has revealed shortcomings that are not very easy to correct without extensive re-design of the system. A decision was therefore made not to try and fix most of the problems (except where they were identified as critical errors), but to begin a new system that would take advantage of the more recent developments in computer technology. However, it was also recognised that to replace the whole of HYMAS would be a mammoth task for which the IWR did not have the available resources in the short term. As the immediate requirement seemed to be more flexible tools for analysing and displaying time series data from a variety of observed or simulated sources, it was decided to initiate the new developments with post-processing graphical display software and develop some interim tools for converting data used within HYMAS to a format suitable for use with the new approach. In adopting this approach it was also considered that issues would emerge that would be of value when time and resources became available to re-design the remainder of the HYMAS system. As parts of the rest of this report illustrate, this has indeed been the case and the Institute believe that they now have a realistic vision of the way forward, in terms of software development, for the next few years.

The IWR made a decision during 1996 to move away from 'C' as their standard language for software development towards DELPHI (object orientated, visual PASCAL), largely in response to a survey of opinion of other groups developing computer code under similar circumstances. Initial tests of the language and program development environment indicated that the choice was justified as the amount of time required to structure quite complex programs was found to be relatively short and a substantial improvement on non-visual languages. It is quite possible that other visual languages would have served the Institutes requirements just as well.

2. BASIC DESIGN CONSIDERATIONS

As the final product was to be designed to satisfy at least some of the requirements of a range of users, it was considered essential to discuss some of the design issues with a representative group prior to beginning the development of code. A further reason for such a discussion was to explore the possibilities of carrying out cooperate developments and the sharing of code. A workshop was therefore held during March 1997 to explore various aspects of the development of an integrated time series analysis and display package. The following groups participated in the workshop :

Computing Centre for Water Research, University of Natal, Pietermaritzburg
Department of Agricultural Engineering, University of Natal, Pietermaritzburg
Department of Civil Engineering, University of Stellenbosch
Department of Hydrology, University of Zululand
Directorate of Hydrology, Department of Water Affairs and Forestry
Institute for Water Research, Rhodes University
Ninham Shand Inc., Consulting Engineers, Cape Town
Water Research Commission (WRC)

The WRC, in the role of workshop convener, identified the following three points as important issues to consider during the workshop :

- The objective of the workshop should be to determine how best to maximise cooperation in the development of TS analysis and display software development.

- There are many users, including hydrologists and non-hydrologists, who make use of TS data and analysis procedures and therefore who could make use of any software developed. There is also a perceived need for more standardisation in the analysis procedures and the display techniques (especially for non-hydrologists).

- The WRC are under pressure to avoid financing duplication within existing and future projects supported by them. This particularly applies to the field of software development.

While there was a great deal of general agreement about the need for some measure of standardisation, both in terms of the language to use for code development, as well as the format of data storage, there was also a great deal of disagreement on the best approaches to be used. These issues could not be satisfactorily resolved at the workshop and many of them are still unresolved nearly three years later. However, on a more positive note, it was realised that standardisation of data storage methods is not a very critical issue as long as interfaces to several formats are available within the software and that these are transparent to the user. The workshop participants also highlighted the range of analysis and display options that were considered necessary.

Hughes (1997) attempted to summarise some of the outcomes of the workshop, recent developments that have contributed to a need for an integrated software package, some of the potential applications and some of the technical considerations related to developing the software. The majority of the points raised within this paper and further developed during 1998 and 1999 are discussed below.

2.1 Interaction with other software

While one of the principles of the basic design was that the software should be able to be used independently of any other package, it was also considered to be essential that it could interface with related software as seamlessly as possible. Specifically, the ever increasing trend towards accessing hydrological and other geo-referenced type information through a spatial interface suggests that the software should be able to be linked to GIS coverages in some way. There are clearly a number of different ways in which this type of link can be achieved and they will be somewhat dependent on the GIS package being used. Ultimately, the link and the way in which it operates will depend upon the efficiency of the design of the spatial and time series databases and how they are related to each other.

The initial design considerations also attempted to account for the fact that the time series analysis and display software would be required to link to time series simulation and modelling packages, that are in current use as well as those that might emerge in the near future as revisions of existing approaches. Some models (those that operate at daily time steps and lower, that operate as spatially distributed or semi-distributed models and those that simulate a number of internal process variables), can generate extremely high volumes of data. Some of these data are important during the model setting up and calibrating stages, but are not necessarily important at later stages. While such issues may not be very important under some circumstances, it is possible that the format of data storage considered to be most suitable for the analysis and display software may not be suitable in terms of the efficiency of the operation of the modelling software. A decision was taken, quite early in the project, that the method of data storage would be designed for the analysis and display software, but that it would be a flexible and straightforward method that would make it suitable for use with a range of other software. It was also accepted that as they are no real standards of data storage within South Africa it is inevitable that interface routines would have to be written to link with other software packages.

2.2 Data storage considerations

Many of the design considerations that fall within this category have arisen from the points already raised in the previous section. The Institute team working on the project could not identify any consensus amongst hydrologists working within South Africa on the most efficient and suitable methods to store time series data and therefore those adopted were the ones that best matched the project teams perceptions. It was also recognised that every effort would be made to assist other groups who wished to use the developing software but with an alternative type of data format.

During the development and application of HYMAS, the IWR had a great deal of success with the use of relatively simple binary file structures. They are efficient in terms of memory usage and, unlike text type data storage, the amount of memory required is not dependent upon numeric precision (within the range used for hydrometorological data). Searching and moving around in binary files is also more efficient and much easier to program than with text files, important considerations when dealing with the high volumes of data that can be associated with time series. The frequently made observation that binary files cannot be easily read and edited is considered irrelevant to the project team as it is a trivial matter to write software to translate the binary data to text (or vice versa), as long as the binary format is simple and well understood.

A further consideration, revealed during the workshop, was that many groups favour the use of database tables for storing information and that adopting this type of approach would tend to make the software more compatible with recent trends in commercial software developments.

The final choice made by the development team was to adopt two parallel data storage methods, both of which would be supported directly by the software. It was anticipated that support for other methods could be added at a later stage in the development without having to carry out a great deal of restructuring of the program.

- ❑ Binary File Format : This format is based on individual binary files being used to store the time series data and that each individual file would have a fixed size 'header' that would contain all the information, referred to as the time series attributes (see next section), required to be able to interpret the remaining contents of the file. The time series component of a file can consist of four different types of data :

- Single variables (stored as 4 byte real numbers), designed for situations where the time interval between the data is fixed (1 hour, 1 day, 10 days, 1 month, etc.).

- Time (stored as either intervals or cumulative values from a defined base and as 4 byte integer or 4 byte real numbers) and single variables (stored as 4 byte real numbers), designed for situations where the time interval between the data is variable (as with digitised breakpoint rainfall or streamflow data, for example).

- Multiple variables (all stored as 4 byte real numbers), designed for situations where it is convenient or efficient to write data from several sources to the same file. This option was included to account for storing output data from semi-distributed simulation models. For example, the VTI model which is the main daily time step, rainfall-runoff model within HYMAS, can generate output consisting of up to 22 variables (rainfall, soil moisture, evaporation, interception, various runoff components, final sub-area runoff, etc.) for each time step and sub-area in the spatial distribution system. If all of these data were to be written to separate files it would involve keeping open, or opening and closing 22 files for every sub-area. It was therefore concluded that a realistic compromise between the requirements of a future version of the model and the time series analysis and display software would be to write all the input, state and output variables for each sub-area to separate files.

- Time (stored as either intervals or cumulative values from a defined base and as 4 byte integer or 4 byte real numbers) and multiple variables (stored as 4 byte real numbers). This approach is designed for situations where the model being run may operate with variable time intervals.

❑ Database Table Format : In this format the time series attribute information is stored in the first set of fields (see next section) of the database table, while the last field of the table is a Binary Large Object (or BLOB) that can contain information of variable length. The BLOB is identical in most respects to the time series component of the binary files referred to above and can in fact be accessed within a computer program in more or less the same way. A decision was taken to assume that the database table format would be the preferred option by the IWR development team and therefore would represent a more permanent form of storage than the binary file option. As the internal details of model simulations are not normally considered for permanent storage, it was decided not to allow for the 'multiple variable' option within the database table format. The concept, in terms of linking the analysis and display software with a model, would therefore be that the model would generate multiple variable output files (one for each sub-area), which could be accessed by the display software. An option in the modelling software would be to write some of the variable/sub-area combinations to separate records within a database table for more permanent storage.

Towards the end of the project a great deal more thought was given to interfacing time series and analysis and display approaches to spatial data. While these considerations have given rise to some different ideas about how to structure time series data within a more complete database incorporating spatial feature coverages, non-time series tabulated data, photograph and diagram bitmaps as well as text type memo data, the basic concepts behind the structures discussed above are unlikely to change. More details on likely future developments are included at the end of this report.

2.3 Data attribute information

Regardless of the format of the files or database fields that are read, certain attribute information is required so that the software can interpret the data correctly and enable or disable certain options. Some of this information may be purely descriptive to identify the data source, some may be required to provide a spatial context, some to specify the format of the stored data, some to identify the data type and units and some to specify codes used for missing data.

In designing a standard header of attribute information to be used with all time series data generated at the IWR, the main consideration was flexibility and the requirement to be able to use these attributes to define almost any kind of time series data related to hydrology and water resources. The following list provides the details of the information contained within the attribute header of the binary files and notes whether or not that information is included as a field in the time series database tables.

A.1 Main description of the data source (stored as a 42 byte string, including the null terminator, giving a maximum length of text of 41 characters).

A.2 Geographical or site description of the data source (stored as a 42 byte string, including the null terminator, giving a maximum length of text of 41 characters).

In the database tables these two text descriptors are replaced with three text fields, each containing a maximum of 30 characters. The fields are referred to as 'Geog. ID', 'Data ID' and 'Data Type', which illustrates the reasoning behind having three fields. However, any descriptive information can be entered.

A.3 Latitude (stored as decimal degrees as a 4 byte real number)

A.4 Longitude (stored as decimal degrees as a 4 byte real number)

A.5 Hemisphere (N or S - upper or lower case, stored as a 1 byte character)

A.6 Hemisphere (E or W - upper or lower case, stored as a 1 byte character)

A.7 Area (stored as a 4 byte real number and assumed to be in km^2 - this will not be relevant to all types of data and could be left as zero).

A.8 Data object type describing the main structure of the data in the time series part of the file (or the BLOB in the case of a database table). This is stored as a 2 byte integer in the range 0 to 3 where :

0 represents data stored are T/S of a single variable.
1 represents data stored are T/S of data pairs of time plus single variable.
2 represents data stored are T/S of multiple variables.
3 represents data stored are T/S of data groups of time plus multiple variables.

Later attributes in the list qualify some of the details of the format of storage.

A.9 Units of the time data. This is stored as a 2 byte integer as a code value for minutes, hours, days or months. The code values are taken from the standard system units file which is reproduced in Table 2.2. This attribute applies to all data object types.

A.10 The time interval, in units given by A.9, if the data are stored with a fixed interval (i.e. type 0 or 2 data objects). This is stored as a 2 byte integer.

A.11 Data type of time value stored as a 2 byte integer. This code is used to identify how the time data appear in files which contain type 1 or 3 data objects (see A.8). Note that the non-time variables are assumed to be always stored with floating point numbers.

Where : 0 represents an unsigned integer (2 bytes)
1 represents a signed integer (2 bytes)
2 represents a long integer (4 bytes)
3 represents a floating point number (4 bytes)
4 represents a double precision floating point number (8 bytes)

In the database tables, this attribute is not used and the time data are assumed to be stored in the same way as the non-time data, i.e. as floating point numbers.

A.12 The type of the time interval data if the data are stored with a variable interval as data object type (A.8) 1 or 3 in units given by A.9. Stored as a 2 byte integer.

Where : 0 represents time steps/intervals.
1 represents cumulative time from base1800.
2 represents cumulative time from base 1900.

A13 Data type of the non-time variables stored as a 2 byte integer. This attribute is no longer used and the software assumes that the data are stored as 4 byte floating point numbers.

A.14 Variable type class (i.e. length, area, etc.) stored as a 2 byte integer and refers to the class type code in the standard systems units file (Table 2.2). This attribute applies to all database table BLOBs, but only data object types 0 or 1, as other data objects have more than one variable and their class and units are provided in a linked text file (see A.18 and A.19).

A.15 Variable type units (i.e. mm, cm, etc.) stored as a 2 byte integer and refers to the unit type code in the standard systems units file (Table 2.2). This attribute applies to all database table BLOBs, but only data object types 0 or 1, as other data objects have more than one variable and their class and units are provided in a linked text file (see A.18 and A.19).

A.16 Scaling factor (0.001 - 0.1 - 1000) stored as a 4 byte floating point number. This attribute applies to all database table BLOBs, but only data object types 0 or 1. For example, streamflow data may be stored in the file in $ft^3 s^{-1}$, but the attribute code under A.15 may refer to $m^3 s^{-1}$. The scaling factor would then have the value of 0.0283.

The multiple variable file types (attribute A.8 with codes 2 or 3) require information about the list of variables associated with each time step. This is achieved by having an attribute that identifies how many variables are present (A.17), referencing a text file (A.18) that contains the name of the variable and its variable type and class and identifying which of the variables referred to within the text file have been included in the data file (A.19). *None of these three attributes are included in the database tables.*

A.17 Number of variables contained within in a multiple variable file. This code is stored as a 2 byte integer and applies to type 2 or 3 data object types only.

A.18 Full file path and name of a text file (see Table 2.3 for an example) containing the variable names, variable type and unit class codes (refer to Table 2.2) that could appear as part of a multiple variable file. The file specifications are stored as a 64 byte string (including the null terminator), allowing a full path of 63 characters. This attribute only applies to data object types 2 and 3.

A.19 This attribute identifies which of the variables included in the filename given by A.18 are actually included in the time series data file. The information is stored as a 44 byte array to represent the presence (byte value = 1) or absence (byte value = 0) of up to 44 variables. Note that the number with a byte value of 1 should not exceed the number given by attribute A.17 and that if the file given in A.18 refers to less than 44 variables (as most will) some of the elements at the end of the array will be ignored. This attribute only applies to data object types 2 and 3.

A.20 The value within the non-time data that will be recognised as missing data. Stored as a 4 byte floating number.

A.21 The starting year of the stored data. Stored as a 2 byte integer which must include all 4 digits of the year.

A.22 The starting month of the stored data. Stored as a 2 byte integer.

A.23 The starting day of the stored data, Stored as a 2 byte integer.

A.24 The starting hour of the stored data, Stored as a 2 byte integer.

A.25 The ending year of the stored data. Stored as a 2 byte integer which must include all 4 digits of the year.

A.26 The ending month of the stored data. Stored as a 2 byte integer.

A.27 The ending day of the stored data, Stored as a 2 byte integer.

A.28 The ending hour of the stored data, Stored as a 2 byte integer.

Within the database tables there are four fields to represent the start and end dates (date fields) and the start and end times (time fields).

A.29 The number of records (individual time steps) within the data file or database table BLOB. Stored as a 4 byte long integer.

There is an additional attribute field in the database tables to represent a unique lookup value for locating specific records.

It should be reasonably apparent that the data storage system being used within the IWR is relatively straightforward and that to write an interface program to convert other data files to, either a binary type file, or a record in a database table is a relatively trivial task for a competent computer programmer. Similarly, should there be other important database structures already in existence, it should not be too difficult or time consuming to incorporate a module into the IWR software that allows the user the option to select that data format, rather than one of the IWR formats.

Table 2.1 Size composition of attribute record and the order in which they appear in the DELPHI code 'attribrec' packed record (similar to a 'structure' in other languages).

Attribute		Char. length	2 byte Integer	4 byte Integer	4 byte Real
A.1	Main description	42			
A.2	Site description	42			
A.3	Latitude (Dec. Deg.)				1
A.4	Longitude (Dec. Deg.)				1
A.16	Scaling factor				1
A.7	Area				1
A.20	Missing data value				1
A.5	Hemisphere (N or S)	1	1		
A.6	Hemisphere (E or W)	1	1		
A.8	Data Object Type (0, 1, 2, 3)		1		
A.9	Time data units (see table 2.2)		1		
A.10	Time interval (1 - 59)		1		
A.11	Data type of time (0 - 4)		1		
A.12	Time interval type (0 - 2)		1		
A.13	Data type of variable (not used)		1		
A.14	Variable Class type (see Table 2.2)		1		
A.15	Variable Units type (see Table 2.2)		1		
A.17	No. of variables		1		
A.18	File reference for variable names	64			
A.19	Variables included bitmap	44			
A.21	Start Year (1800 - ????)		1		
A.22	Start Month (1 - 12)		1		
A.23	Start Day (1 - 31)		1		
A.24	Start Hour (0 - 23)		1		
A.25	End Year (1800 - ????)		1		
A.26	End Month (1- 12)		1		
A.27	End Day (1 - 31)		1		
A.28	End hour (0 - 23)		1		
A.29	Number of records			1	
Total size		194	19*2=38	1*4=4	5*4=20

The total size of the header is therefore 194 + 38 + 4 + 20 = 254 bytes = 256.

Table 2.2 Standard system file of variable classes and unit types.

Variable Class type 1 : LENGTH			(7 Unit types)			
1 : mm	2 : cm	3 : m	4 : km	5 : ins	6 : ft	7 : miles
Variable Class type 2 : AREA			(5 Unit types)			
1 : m^2	2 : hectares	3 : km^2	4 : ft^2	5 : $miles^2$		
Variable Class type 3 : VOLUME			(7 Unit types)			
1 : litres	2 : m^3	3 : Ml	4 : $m^3 * 10^6$	5 : ft^3	6 : gallons	7 : gall $*10^6$
Variable Class type 4 : FLUXES			(5 Unit types)			
1 : mm h^{-1}	2 : m s^{-1}	3 : m h^{-1}	4 : m d^{-1}	5 : ft s^{-1}		
Variable Class type 5 : RATES			(7 Unit types)			
1 : litres s^{-1}	2 : m^3 s^{-1}	3 : m^3 d^{-1}	4 : Ml d^{-1}	5 : $m^3 * 10^6$ d^{-1}	6 : ft^3 s^{-1}	7 : gall. d^{-1}
Variable Class type 6 : TIME			(5 Unit types)			
1 : mins	2 : hours	3 : days	4 : months	5 : years		
Variable Class type 7 : WEIGHT			(6 Unit types)			
1 : mg	2 : gm	3 : kgm	4 : tonnes	5 : cwt	6 : tons	
Variable Class type 8 : CONCENTRATION			(8 Unit types)			
1 : mg l^{-1}	2 : gm l^{-1}	3 : kgm m^{-3}	4 : tonnes m^{-3}	5 : cwt ft^{-3}	6 : cwt gal^{-1}	7 : tons ft^{-3}
8 : tons gal^{-1}						
Variable Class type 9 : OTHER			(3 Unit types)			
1 : %	2 : Fraction	3 : mS m^{-1}				

Table 2.3 Example of a text file listing the variable class, unit and names references. This example relates to the full range of output from the semi-distributed VTI model that forms part of HYMAS

Number of variables	22	
Variable Class	Variable Unit	Description
1	1	Rainfall
1	1	Potential Evaporation
1	1	Canopy Evaporation
1	1	Potential Dam Storage
1	1	Upper Zone Moist. Storage
1	1	Lower Zone Moist. Storage
1	1	Soil Evaporation
9	1	Saturated Area
1	1	Saturated Area Runoff
1	1	Intensity Runoff
1	1	Soil Baseflow
1	1	Spring Baseflow
1	1	Ground Water Baseflow
1	1	Ground Water recharge
3	2	Ground Water Outflow
1	1	Perched Storage
1	3	Ground Water Depth
5	2	Upstream Inflow
5	2	Total Sub-area Flow
1	1	Available Loss Storage
3	3	Transmission Losses
5	2	Final Routed Runoff

2.4 Accessing time series within the software

During the initial design process, it was considered that users would frequently return to a set of time series files or database records for display and analysis and should ideally not have to reselect them from a directory of files or a larger database table every time. It was therefore decided to adopt the principle of a 'profile' of time series data that would store references to and summary details of several time series data sets. Thus, at the start of the analysis and display program a user could select an existing profile and use it, add time series to it, or remove time series from it. It should also be possible to build a completely new profile for later use.

A further consideration is that once a profile has been made there will still be user selections to be made in terms of which data to display and how to display these data. It was therefore recognised that a facility to save and load profile and display options would be extremely useful. This is particularly true for repeated assessment of the results of a simulation model. After the first run of the model, a profile could be established with some of the simulated variables included as well as some observed data for comparative purposes. The format of the graphical displays could be selected and the whole application saved. The next time the model is run, the display software is loaded and the application restored so that the same graphical image is generated, saving time in terms of selecting a profile and setting up the graphs. This may seem like a trivial issue, but may be quite significant in terms of efficiency when calibrating a model and requiring many repeated displays of the same information.

2.5 Analysis and display options

During the initial phase of the design, it was decided to use simple graphical plots of the data (time on the X-axis, data units on the Y-axis) as the basis for several other types of analysis and it was decided to allow up to two sets of axes to be available on the screen. Other analyses or display options would then be added as overlays onto the simple plots. However, there are several potential options that could be included to allow a great deal of flexibility in the manner in which the data are displayed. It was also recognised that the more options which are included, the more time and effort it will take to generate simple graphs. Some of the considerations that contributed to the design were :

- Some users will wish to plot data from different sources on the same set of axes. These differences may mean that the time series have different time steps (overlaying daily and monthly data, for example), or that they are stored using different units (displaying streamflow data stored in $m^3 s^{-1}$, togther with TDS concentration data in $mg l^{-1}$, for example).

- The need to allow for more than one data unit type on the same graph suggests that both the left hand and right hand Y-axes must be available.

- While there may be situations where two or more data sets stored with different time intervals should be displayed togther using their original intervals, there will also be situations where it will be required that they are all reduced to a common interval. An example could be superimposing simulated daily streamflow data with observed (or simulated using a different model) monthly streamflow data. The software therefore needs to allow for the data series to be integrated over a range of time periods that are longer than the original.

- It was recognised that it is frequently useful to display data in a standardised manner, dividing the actual values by catchment area, or the mean value of the data series, for example.

- It should be possible to select a period from the full data series for plotting, despite the fact that on-screen panning and zooming within the graphs will be possible.

- It should also be possible to select seasonal periods (a group of months) from several years and plot these on a single graph. This type of facility has frequently been identified as being useful to investigate seasonal similarities (or differences) within a group of generally wet or dry years in terms of annual totals.

The above all relate to the selection of the manner in which the data will be plotted. The design also recognised the need for some further facilities for modifying the display once the graphs are plotted.

- ❑ It should also be possible to manually scale both axes and be able to plot the data on the Y-axis using a logarithmic scale.

- ❑ It should be possible to remove series (as well as add them) from the plotted graphs.

- ❑ It should be possible to display one or both of the graphs on the full screen.

- ❑ It should be possible to print out the graphs.

- ❑ A method of reviewing the source of a series is required (some sort of key).

Apart from the standard time series display facilities there should also be a number of optional analysis facilities that may or may not include additional graphical displays. The workshop held at the beginning of the project identified a number of potential options :

X-Y Scatter plots (including observed v simulated analyses) and comparative statistics
Flow duration curves
Curve fitting
Spell/Run analysis
Frequency (high and low flow) analysis
Seasonal distributions
Baseflow analysis and separation
Recession curve analysis
Spectral analysis
Hydrograph separation
Water Quality piper diagrams
Box and Whisker plots
Groundwater analysis routines (not specified in detail)
Residual flow diagrams

In terms of the immediate use to which the IWR intended to put the software (comparison of simulation results with observed data, providing hydrological data display facilities for instream flow requirement workshops and other situations involving non-hydrological specialists) the development team identified some of these as priorities and decided to leave others until the basic software design was further developed. The following were therefore identified as being either relatively simple to implement, or of being of sufficiently high priority to require implementation immediately.

- ❑ Simple statistics of a selected time series (i.e. means and standard deviations of both un-transformed as well as log transformed values).

- ❑ Seasonal distribution plots of one or more selected time series.

- ❑ Duration curve plots of selected time series data with the additional facility to generate curves based on data from individual calendar months. It should be possible to point to anywhere on a duration curve and obtain the data value and the percentage time equalled or exceeded.

- X-Y scattergrams, including two variable analyses (observed versus simulated data, for example), single mass curve plots (cumulative time versus cumulative data) and double mass curve plots (two cumulative data series plotted against each other). For the two variable analyses the software should also be able to generate some comparative statistics.

- Baseflow analysis and separation. The separation algorithm(s) should be flexible and a there should be a facility for saving the separated baseflow time series to a file or database record.

- High and low flow frequency analysis. The difficulties with this option are related to the choice of which theoretical frequency distributions to include.

- Spell and run analysis. Facilities of this type are included in HYMAS and while they have been found to be very useful, more thought is required about the methods used to set up the analyses and display the results.

3. DESCRIPTION OF THE COMPONENTS

It should be noted that the development of the software is expected to continue beyond the end of the current project and the compilation of this report. The description that is included here therefore represents the only the first phase of the development. A later chapter of this report provides further information about future developments that have either already started, or are being considered.

The current name for the software is TSOFT, although this is likely to change at a later date.

3.1 Data preparation

Sections 2.2 and 2.3 described the structures of the two existing options (binary files and database records) for the storage of time series data to be accessed by the software. At present, only two data preparation utility programs have been written and these are used to convert data stored in HYMAS type files to either the new binary format (including the standard attribute header) or to records within a database table. The reasoning behind this approach was that HYMAS already contains a wide range of data conversion routines (from standard data file formats used in South Africa, as well as several other common formats) and the development team did not consider it necessary to reproduce these at this stage. A further factor is that quite a substantial amount of the raw data that the IWR have handled in the last few years has not been compatible with changes in centuries (Y2K problem) because the year is stored using two digits. Many of these formats will therefore have to be modified which will effect the data conversion programs that will be required in the future.

3.1.1 Program to convert HYMAS type files to TSOFT binary type files

This program allows the user to select and open one of several types of HYMAS files, enter or edit the attribute information and save the attribute and time series data to a standard TSOFT type binary format file. Figure 3.1 is a reproduction of the initial window and illustrates that there is a facility to list the records of the HYMAS file to ensure that the data are being read correctly. In the case of multiple variable HYMAS files (such as the *.mos files used for storing model results), it is necessary to know the number of sub-areas and number of variables that are included in the file. These values are interpreted automatically from the file structure.

Figure 3.2 shows the window that is used to enter or edit the attribute information in several of the TSOFT programs. Many of the entries are reasonably self explanatory, while others require some guidance in their use and may need to be supported by a help facility. Similarly, many are set automatically depending upon the type of HYMAS file that has been opened. The data object type in the example has been set to multiple variable as a model output file has been selected. The variable class and units entries are turned off because there are several variables in the file and it is necessary to select a variable names file (see table 2.3) and highlight the variables that are present to populate the array that forms attribute A.19 (see the bottom right of figure 3.2). In the case of single variable files it will be necessary to set the class and units. As the model output files of HYMAS do not contain information about the start and end dates, these details also have to be entered, although the difference between the start and end dates will be correct if the time data units and interval is 1 day. For other HYMAS files the start date can be read from the file and the end date is calculated from the number of records and the time step. Clearly, this is not possible if the data are stored with variable intervals.

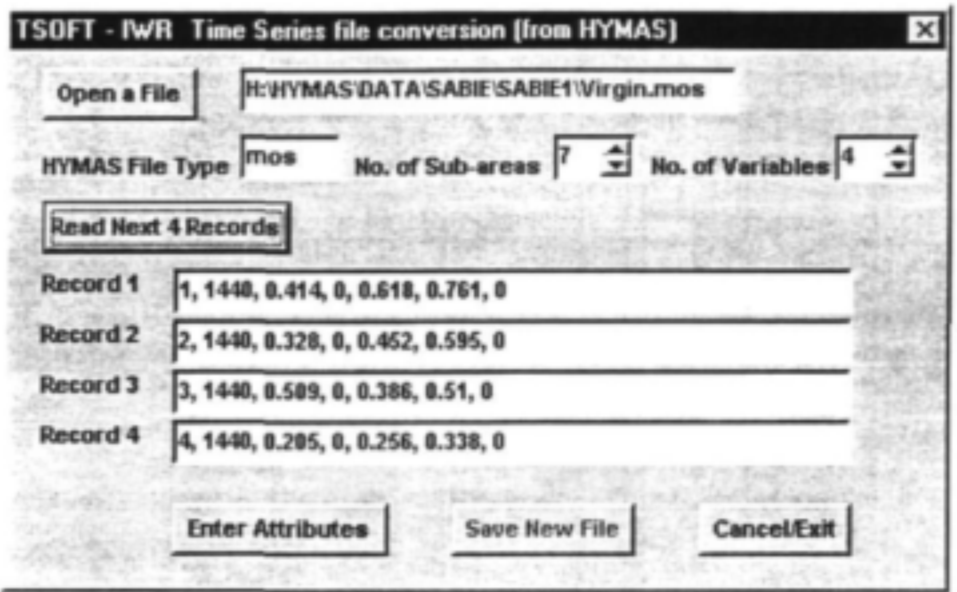


Figure 3.1 Initial screen of the program used to convert HYMAS files to TSOFT files.

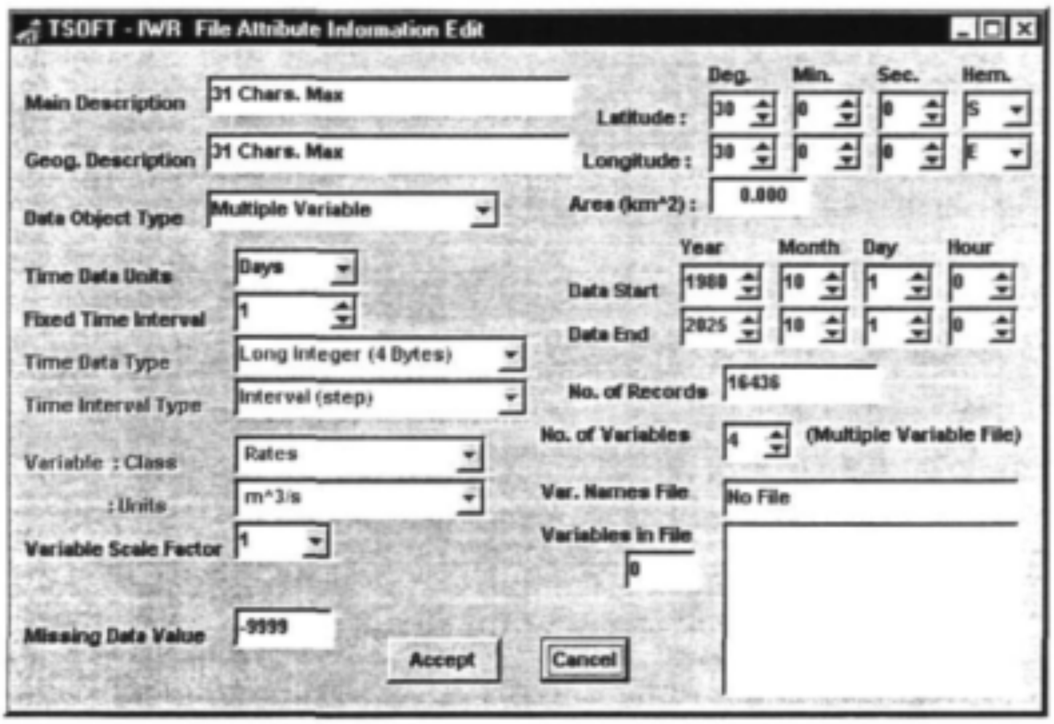


Figure 3.2 Attribute entry or editing window (used within several TSOFT programs)

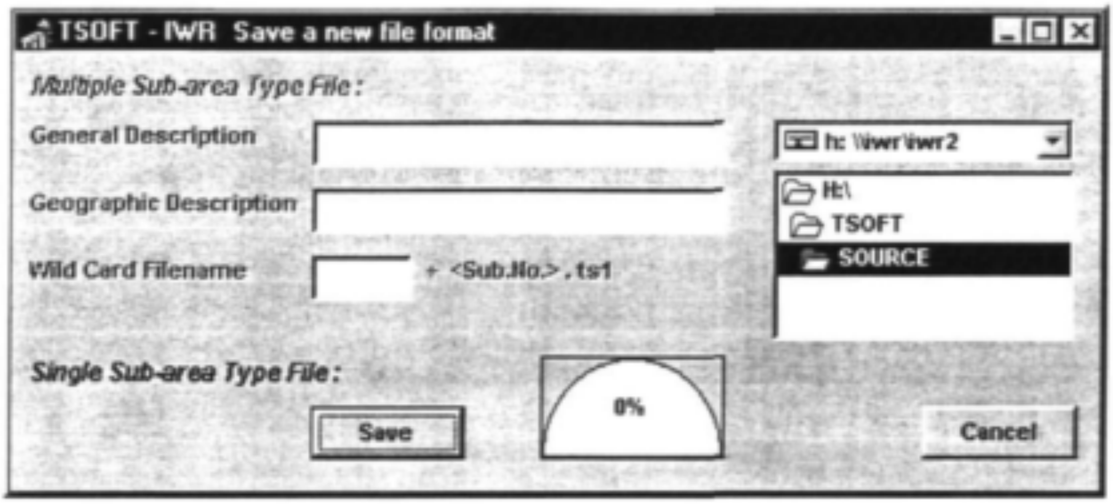


Figure 3.3 File saving window associated with the data conversion program

Figure 3.3 illustrates the file saving window that represents the final stage of the data conversion program. There are two possible saving situations; the first if data for a single sub-area are stored within the HYMAS file and the second if data for more than one sub-area are stored. In the former case (as illustrated) clicking on the 'Save' button brings up a standard Windows file selection window and the user simply selects the directory and types in the filename required. In the latter case, the user is expected to select the directory using the facilities on the right of the screen and enter a wild card filename, to which will be added the sub-area number for each data series in the HYMAS file. Thus, if a HYMAS data file with time series for seven sub-areas is selected and the wild card filename is specified as 'RIVER' then seven TSOFT data files will be created (RIVER1.TS1 to RIVER7.TS1).

3.1.2 Program to convert HYMAS type or TSOFT files to records in a database

This program performs a similar function to the previous one, except that the data are written to a record in a database table. As this form of data storage has become the IWR preferred method, the option to convert previously created TSOFT binary files has been included. Figure 3.4 indicates that the program can also be used for managing a database table of time series data, either deleting records or editing the attribute information, as well as viewing the time series data. The 'Import TS1 type file' simply calls up a file access window that allows the user to select the required directory and choose a TSOFT type binary file (with extension TS1) to import. All the attribute and time series data are then added to a record in the database which is allocated a unique record number.

The 'Import HYMAS data' option calls the same module as represented by figure 3.1 so that the user can select the required HYMAS type file and enter or edit the attribute information where necessary. The only difference between this program and the previous one is that before the data are saved the user has the opportunity to select for saving one or more of the available variables in a multi-variable file. For example, assume a model results file with 5 sub-areas and 4 variables (Rainfall, Soil Evaporation , Upstream Inflow and Final Routed Runoff) has been selected for conversion and the user has selected Rainfall and Final Routed Runoff to be saved. A total of 10 new database records will be added; a Rainfall and Final Routed Runoff time series for each of the 5 sub-areas. Further editing of some of the attributes is then possible by entering the table display.

3.1.3 Interfacing alternative data sources with TSOFT

The detailed information contained within sections 2.2 and 2.3 has been provided so that users with alternative data sources are able to write software that will carry out the necessary data conversions. A further possibility is that the project team be contacted to consider adding into TSOFT itself components to access time series data stored in different ways. The only critical considerations are that the required attribute information can be accessed from the new type of data.

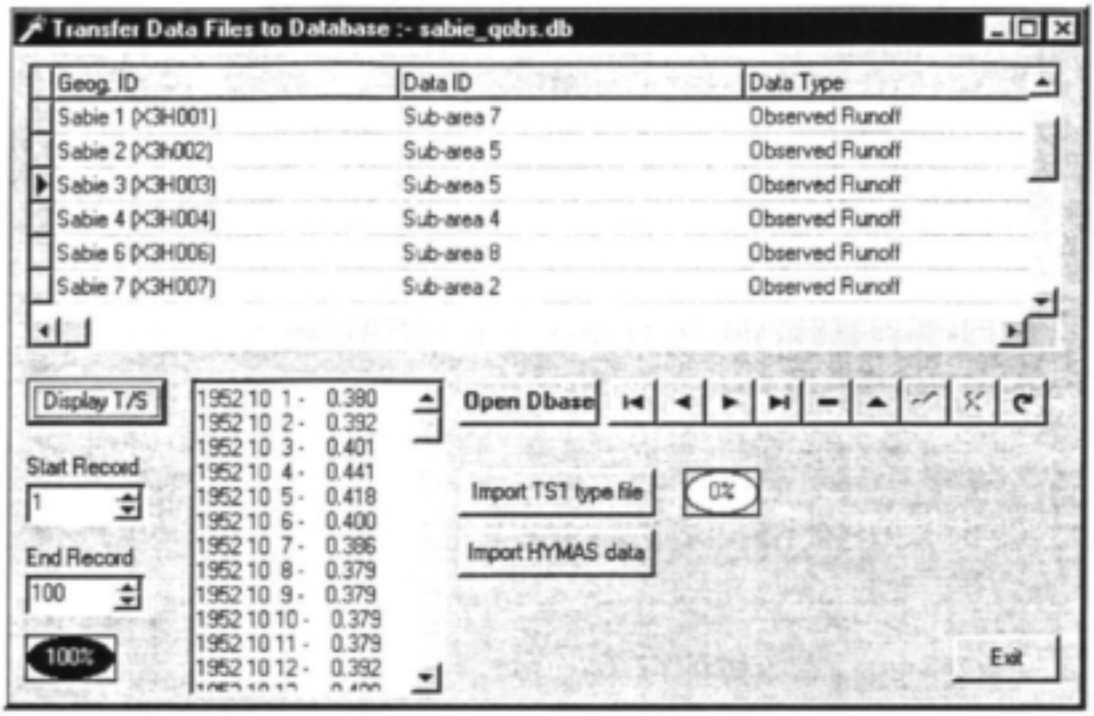


Figure 3.4 Initial window of the program designed to allow HYMAS data to be transferred to a database record.

3.2 Time series selection and data profiles

There are a number of menu options available when running the main graph display program, however, most of them are not accessible until a 'Profile' of time series data has been loaded. This profile is simply a list of references to either binary time series files or records in a database table. Figure 3.5 illustrates the profile selection process and demonstrates that there are several profile management options included.

The simplest use of the screen given in figure 3.5 is to select a profile (*.prf type files) from existing ones and click on the 'Use Profile' button. This loads the profile and allows the other menu options in the main program to be activated. There are also two options to allow new time series to be added to a profile (either an existing or a new one) from database tables or stored in a file. A third option is available to remove a time series from the profile , as well as an option to save the contents of an edited profile to disk. At a later stage in the development of the software it is anticipated that options to load time series from other sources will be included. The main consideration is that the programs attribute information memory record can be populated from the data source and that an internal routine for reading the time series data from the new source can be programmed.

The 'View/Edit Data File' option has been included to allow the attribute information stored in a file or database record to be viewed and edited. If the data are stored in a file (see the first three entries in the example profile loaded in figure 3.5), then this button activates an attribute header editing window (see figure 3.2), while if the data are in a database table (fourth entry in the figure 3.5 example) a window similar to the top part of figure 3.4 is activated.

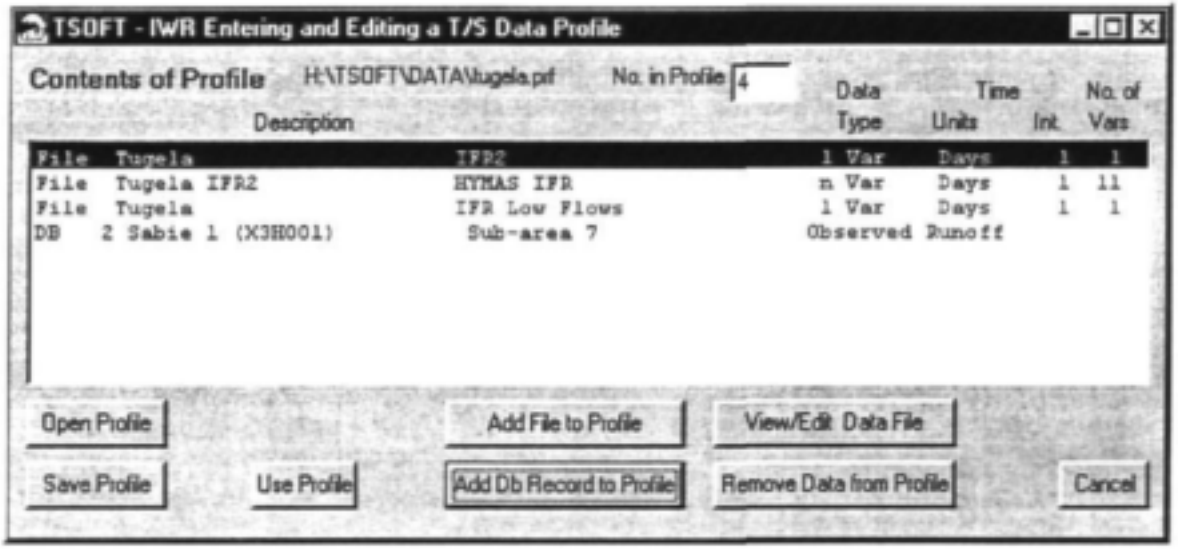


Figure 3.5 Time series data 'Profile' opening and editing window

3.3 Adding time series graphs to the display

Figure 3.6 shows the full screen of the main graph display program togther with the window that allows the user to select a specific time series for display. To switch between adding graphs to the top or bottom set of axes is simply a matter of clicking on that part of the screen and ensuring that the required graph heading (-Graph 1- or -Graph 2-) is highlighted. If the selected profile reference is associated with a multiple data series (the second line in the example of figure 3.6) then a list of the stored variables is displayed and the user may select one or more for display.

3.3.1 Time series conversion options

One of the criticisms of the HYMAS options for displaying model results is that data stored with different units or time steps could not be displayed together on the same graph and that there are no facilities for summarising data stored with a fine time resolution and displaying them at a coarser resolution.

Figure 3.7 illustrates that there are four possible conversion options; Integration Period, Integration Type, Standardisation Type and Scaling Factor. These options are not available if more than one variable has been selected from a multiple variable time series type file, for the obvious reason that each one selected could be stored with different units and have different restrictions. If several variables from a multiple variable file are required to be displayed with conversions then they must be selected one at a time.

The *Integration Period* options are dependent upon the time interval of the raw data and clearly only a longer time step can be selected. The full range of options are daily, 7 days, 10 days, 30 days, month, 3 months, 6 months and 1 year.

The *Integration Type* options are only activated if an integration period is selected and include ordinary and running means, as well as sums.

The *Standardisation Type* options include standardisation by catchment area (in km^2, hectares or m^2), by mean daily flows or volumes, by mean monthly flows or volumes or by mean annual flows.

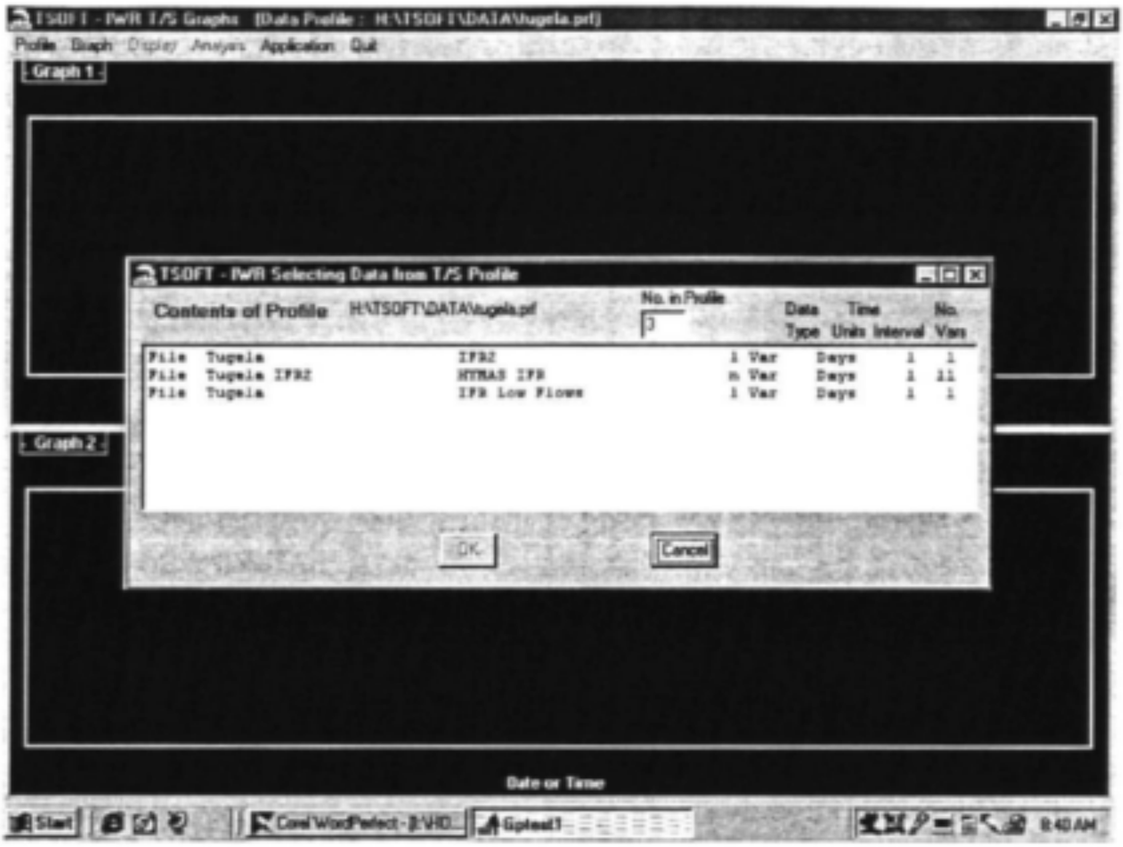


Figure 3.6 Full screen of the main graph display program together with the time series selection option.

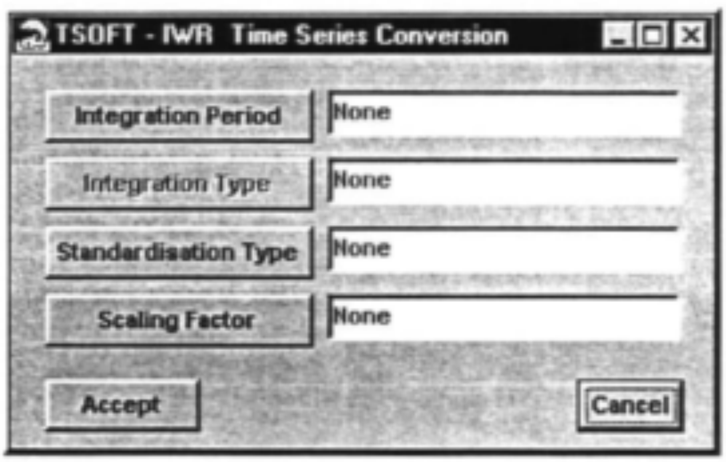


Figure 3.7 Setting the time series conversion options

The available *Scaling Factor* options are dependent upon the units in which the original data are stored. For example, data stored in $m^3 s^{-1}$ can be converted to $m^3 day^{-1}$, $Ml\ day^{-1}$ or million $m^3 day^{-1}$. Although the units conversion menu refers to volumes per day, if the integration period has been set to convert daily data to a lower time resolution and the integration type to be a sum, then the volumes are absolute values and not per day.

3.3.2 Time period selection options

Figure 3.8 illustrates the available options for selecting sub-sets of the available data series for display and further analysis. There are essentially two main options. The first is a simple matter of specifying the start and end date of the data to display (within the constraints of the start and end dates of the original data) and becomes largely redundant due to the panning and zooming options that are available with the main display window. However, if duration curve analyses, for example, are required on sub-sets of the total data series, then they must be selected at this point as all further analysis methods apply to the full data set loaded onto the main screen (and not just those data seen in the current display after zooming).

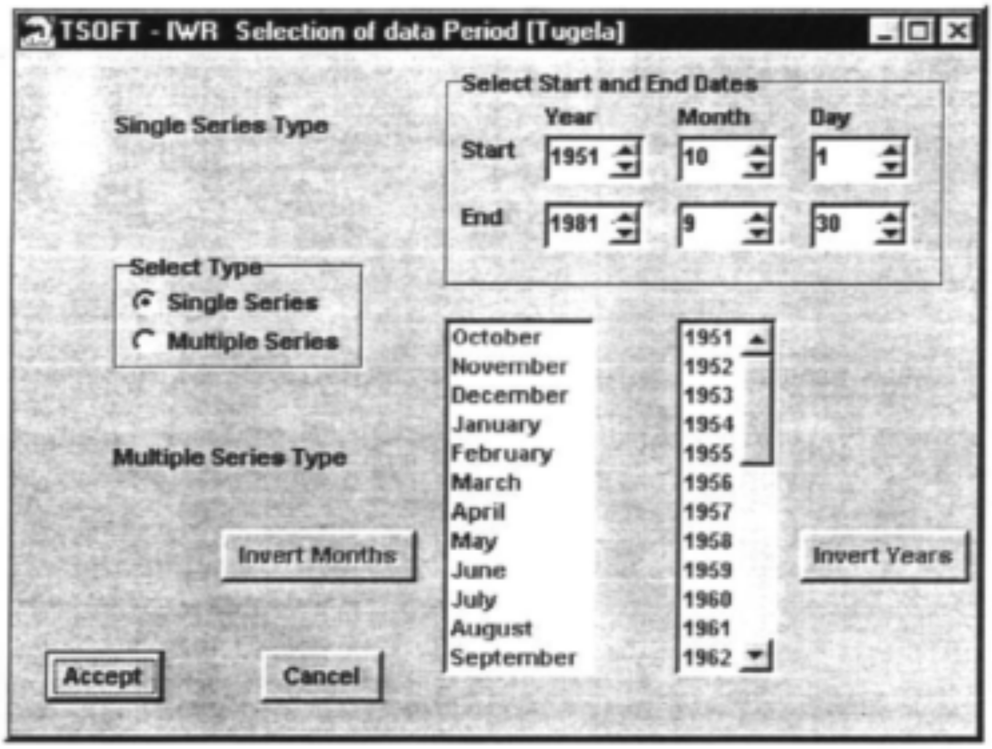


Figure 3.8 Time period selection options

The second option (Multiple Series) allows the user to select a group of months (in sequence) and a group of years. A graph line is then drawn for each of the selected years using data spanning the selected months. This facility has been included to allow a range of seasonal hydrographs taken from several years (all wet or all dry, for example) to be superimposed on the same set of graph axes.

3.4 Manipulating displayed graphs

Before explaining the various additional analysis options, there are several main menu options and other facilities that can be used to modify the form of the displayed graphs. Zooming can be achieved using the left mouse button (to zoom in, hold the left button down and move the cursor down and to the right, to zoom out hold the left button down and move the cursor up and to the left), or by selecting the 'Scale' sub-option of the 'Display' main option. If data series with two different units have been selected for display on either the top or bottom graph then both the right and left vertical axes are used and the scale options allow both axes to be scaled independently.

A 'Show' sub-option of 'Display' allows the user to toggle between showing both graphs or a single graph (top or bottom) on the full screen, while a 'Colours' sub-option can change the background colour from black to white (in which case the first selected series is shown in black rather than white).

The 'Graph' option has sub-options of 'Print' (which is self explanatory) and 'Key' which allows the user to display a summary key of the time series displayed. Clicking on one of the items in the list will change the background colour of the key display window to the same colour as the line used to display the selected series. Some brief summary statistics of the data displayed are also provided.

The 'Application' menu item allows the current setup (selected profile, selected time series and conversion options) to be saved to a file (with extension *.app). This option can also be used to rapidly retrieve a saved application and redraw all the selections without going through all the steps outlined in sections 3.2 and 3.3.

3.5 Time series analysis options

At the present level of development the main graph display program includes four additional analysis options; monthly distributions, duration curves, X-Y scatterplots and baseflow analysis. Additional components will be added as time allows and as the need arises.

3.5.1 Monthly distributions

Plotting monthly distribution curves is a simple matter of selecting from a list of all the data sets that are plotted on one of the main screen graphs. The data are sorted into months and the means of all the non-missing data values in each calendar month used to draw the distribution curves. The vertical axis used for any data set remains the same as on the main graph, as do the units of measurement (see example in figure 3.9). If the original data are streamflows given in $m^3 s^{-1}$ and the monthly distributions are required to be plotted in $m^3 * 10^6$, then it will be necessary to use the data conversion routines to plot the series as monthly values in $m^3 * 10^6$ on the main graph before drawing the monthly distributions.

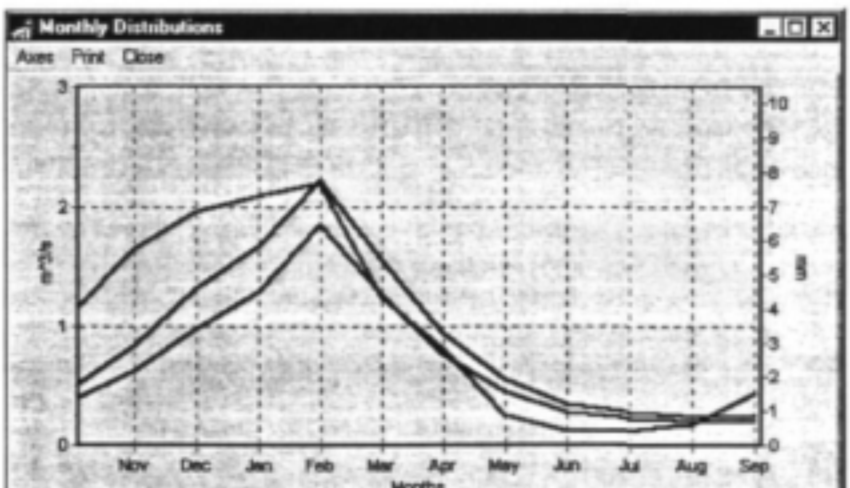


Figure 3.9 Monthly distribution plot window

No key information is supplied on the monthly distribution graph at this stage and the same colours (except for the first graph, which plots as a black line instead of white) are used as for the lines on the main graph.

3.5.2 Duration curves

The duration curve plotting function operates in a very similar manner to the monthly distribution plots. All data sets drawn on the main graph are listed and the user can select which to use in the duration curve analysis. As with the monthly duration plots, the line colours, units and vertical axis used are the same as on the main graph. A further option is available to plot calendar month duration curves for a single data set (figure 3.10).

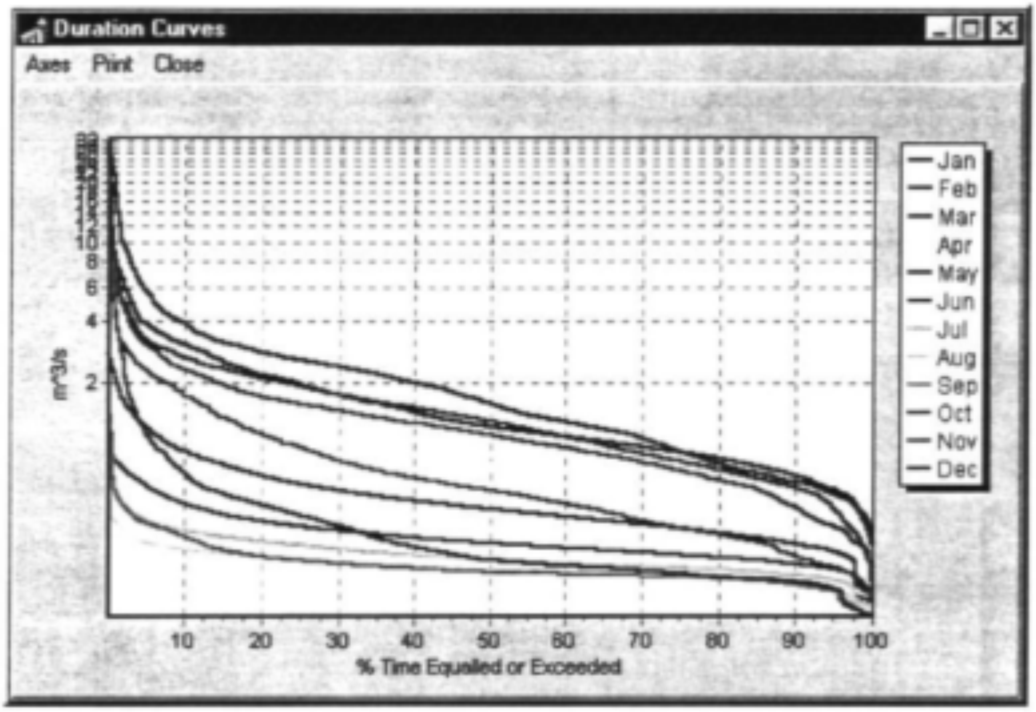


Figure 3.10 Example of duration curve plotting. The 'Monthly curves for a single time series' option has been selected and the vertical axis set to logarithmic.

To date, no simple method of implementing a probability scale on the horizontal axis and retaining the zoom and pan facilities has been found, although the development team are still working on this and expect to be able to implement an improved horizontal scaling facility in the near future. Using the mouse to click on any of the lines will show the variable and % time equalled or exceeded values so that users can obtain more precise information than can be obtained through visual interpretation of the graph.

3.5.3 X-Y Scatterplots

There are three possible ways in which data can be displayed using the 'XY Plots' option. The first is a straightforward bi-variate analysis of two time series (selected from the list of all those plotted on the relevant main graph). The user highlights the X data and clicks on the 'Select X data' button, then highlights the Y data and clicks on the 'Select Y data' button, then clicks the 'Plot' button. Figure 3.11 provides an example of the resulting plot. A further option is then available (apart from the standard options to change the axes to logarithmic scales and to print the graph) to display some statistics of the relationship between the two data sets (see figure 3.11).

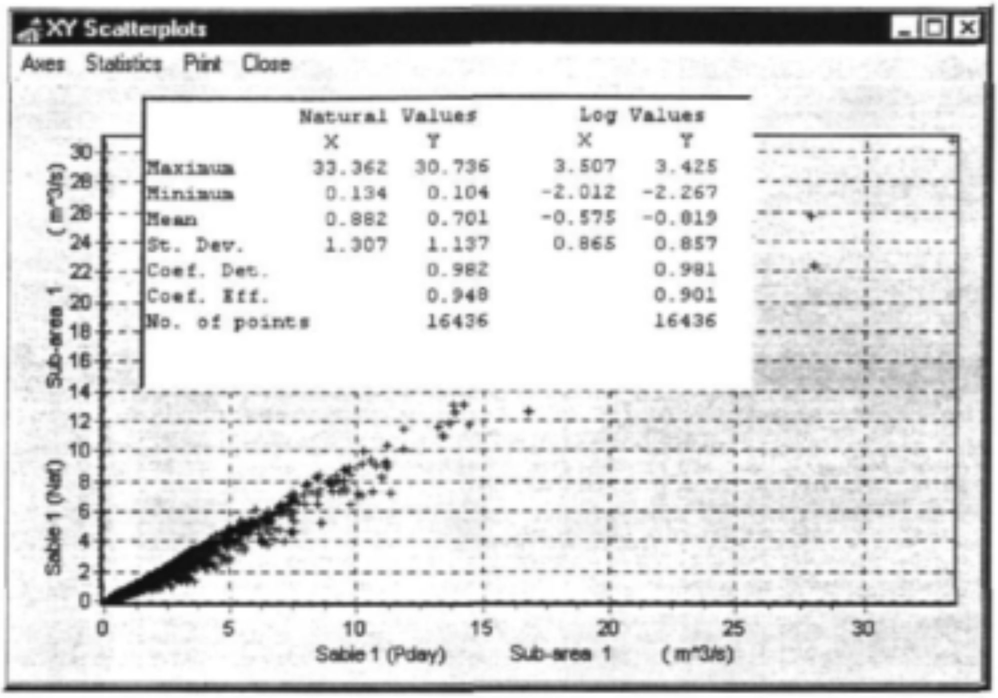


Figure 3.11 Example of an X-Y scatterplot where the 'Statistics' option has been selected.

The two comparison statistics are the coefficient of determination (R^2) and the coefficient of efficiency (a measure of the one-to-one fit, which is sensitive to systematic differences in the two data sets and therefore very valuable in comparing observed versus simulated values).

The second option allows the user to perform a single mass curve analysis of a data series. The 'Cumulative Data' option is selected and no X-data series are highlighted before clicking the 'Select X data' button, while the Y data are selected in the normal way. The third option is a double mass curve plot, in which case the 'Cumulative Data' option is selected and then the X and Y data selected as in the normal generation of an bi-variate scattergram.

3.5.4 Baseflow separation

The baseflow analysis option includes two different methods of separating baseflows from total flows. The first is a well known standard method of digital filtering that was used by Smakhtin and Watkins (1997) in their report on low flow analysis of the flow regimes of South African rivers and originated from Nathan and McMahon (1990). Either of the sources can be consulted for more details on the methodology. The method makes use of a single parameter (less than 1.0 and usually in the range 0.05 to 0.99) that controls the separation process. The second method is currently being tested as an alternative and makes use of rate-of-rise (increases and decreases in the value of the variable over time steps) criteria to determine the separation. For this method the user decides over what time step to calculate rates-of-rise and the program generates a histogram of these rates for the whole total flow time series. The modal value then becomes the default value for the maximum rate of rise at which separation starts, although this parameter can be modified as well. There is also a scaling parameter that controls the shape of the separated line after initial separation. As the method is still being tested there are no real guidelines for the values of the parameters and potential users are advised to experiment with various combinations before selecting appropriate values for their specific data sets.

Regardless of the option selected, the resulting time series of baseflows can be saved to an existing database table (the same one that contains the total flow source) and added to the data profile that is currently active for the main program. This means that it is a simple matter to leave the baseflow analysis options, return to the main graph and add the time series of separated baseflow as a line on one of the sets of axes. These data can then be subjected to further display or analysis options.

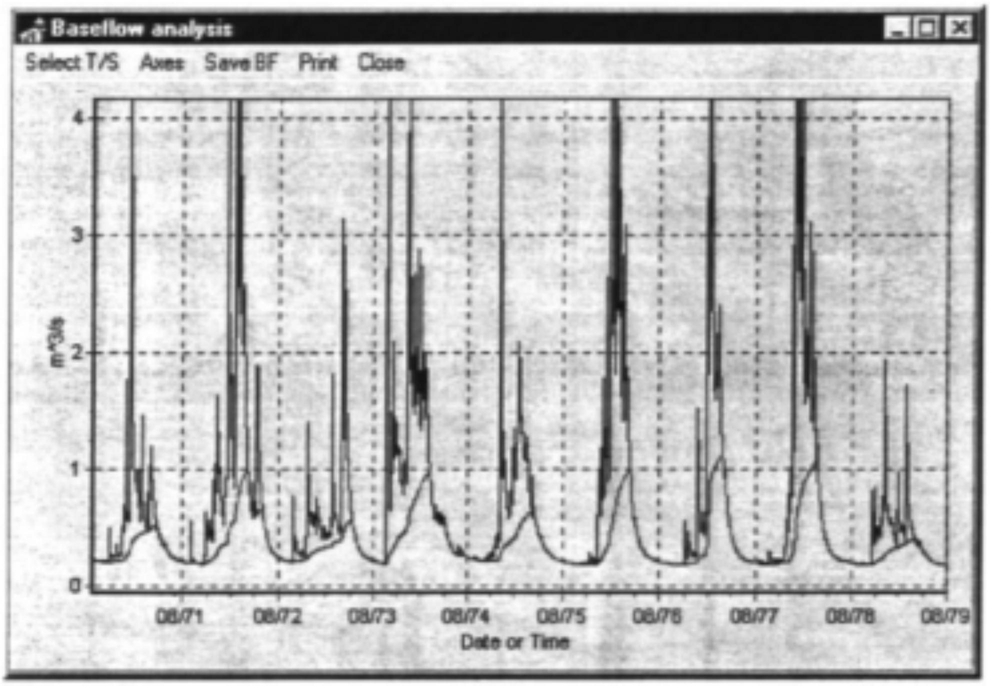


Figure 3.12 Example of a baseflow separation

3.5.5 Planned analysis options

The analysis methods referred to in 3.5.1 to 3.5.5 are the only ones that are currently implemented in the main display programme, but several others have been identified for inclusion within the near future.

High flow (or other variables) frequency analysis has been identified as a important analysis requirement for instream flow assessments to guide the geomorphological specialists in the establishment of suitable channel forming discharge recommendations. One problem is that many of the streamflow time series available for South African gauges do not have accurate flood peak data due to frequent exceedence of their stage-discharge rating curves. However, there still appears to be some value in adding a high flow, or flood peak, frequency analysis method. The other issue to resolve is then what type of flood frequency analysis is to be included, a question that mainly revolves around the theoretical frequency distributions that should be included. There are, however, guidelines from other South African publications that might be of value (see Alexander, 1991 for example).

Low flow frequency analysis methods were included in the HYMAS package during the project on low flow estimation in South African rivers (Smakhtin and Watkins, 1997) and have proved to be quite useful for planning run-of-river abstraction schemes. The HYMAS facilities allow the user to select a range of durations (such as 1, 5, 7, 10 and 30 days - assuming on the original data uses a daily time step) and annual low-flow extremes can then be estimated for a range of return periods using the sample values or a fitted Weibull distribution. It is intended to implement a similar method within the TSOFT software.

A third analysis method that is also included in HYMAS that has proved to be valuable is spell, or run analysis, where the amount of time and the frequency that flows (or other variables) occur above, or below, a set of threshold values are analysed. An extension to this analysis is the calculation of deficit volumes; i.e. the volume of water between a threshold and the flow values for the periods of time that the flow is below the threshold. This type of analysis has also proved to be useful in some water resource studies where abstractions are planned without significant storage. When these routines were added to HYMAS it was noted that there are a wide range of options available for displaying the results and that it is difficult to display them in a manner that is clear to a non-specialist. Before implementing them inside TSOFT it will therefore be necessary to plan the displays carefully.

In addition to the additional options referred to above, which have been identified by the development team as being required to meet their own requirements, several others were identified by the staff of the School of Bioresources Engineering and Environmental Hydrology at the University of Natal, Pietermaritzburg. As these were passed to the developers near the end of the project they had not been included at the time of writing this report. However, many of them are quite straightforward to implement and will be added as time allows. It is expected that most of them will be added during 2000.

- Additional data integration methods, including minimum, maximum and last day values over the integration period (a trivial addition that can be implemented almost immediately).
- The ability to set user definable colours, headings and axes text (should be relatively easy to implement).
- The ability to modify line types and user lines, markers or bars (not a very simple modification due to the possible options and limitations that need to be placed on the users choice, especially with bars - will be investigated).
- Printing quality and ability to print to graphics files of different types.
- Synchronization of the zoom period for both graphs (should be relatively easy, but there may be some complications).
- Ability to superimpose time series graphs on mass curve plots (relatively straightforward).
- Ability to plot cumulative differences as part of a mass curve plot (can be established as an additional analysis option as the existing options are incompatible with this approach).
- Ability to plot differences between two time series as time series plots and as one axis of a scatterplot (also needs to be a new analysis option).
- Display of regression lines and 1:1 fit lines on bi-variate scatterplots and to include some additional statistics such as the y-intercept and slope of the regression line (this is a relatively straightforward modification).
- Seasonal colour variations in the points plotted on a bi-variate scatterplot.
- Export of time series summary statistics to a comma delimited file (straightforward).
- Links to spatial plots (see sections 4 and 5 of this report).

4. INTERFACE WITH ARCVIEW

The following section outlines some of the future proposed developments that are associated with the use of MAP OBJECTS as the spatial data visualisation method within Delphi programs. However, the IWR has not reached that stage in the development of the software and yet still required the software to have some interface with spatial data. The initial approach that has been used is to write an AVENUE script component inside ArcView (with assistance from Rajesh Nundlall of the Computing Centre for Water Research) that allows the user to define a rectangular area within a particular coverage and then pass the coordinates of the corners to an external Delphi program (currently called **arctest.exe**). The contents of the script are given below and it can be seen that at this stage of the development the script is very simple.

```
'Script to select box
'Return coords and display selection box (highlighted)

 TheView = av.GetActiveDoc
 aDisplay = TheView.GetDisplay

 SelRect = TheView.ReturnUserRect
 aFill = Symbol.Make(#SYMBOL_PEN)
 aFill.SetColor(Color.GetGreen)

'--------------------------------------------------------
'Send co-ordinates of rectangle to file

system.execute("h:\tsoft\source\arctest.exe"++SelRect.GetLeft.asString++Sel
Rect.GetTop.asString++SelRect.GetRight.asString++SelRect.GetBottom.asString)

'---------------------------------------------------------
'Zoom to selection

 ZoomRect = SelRect.ExpandBy(0.001)
 aDisplay.ZoomToRect(ZoomRect)
```

Figure 4.1 shows that the coordinates of the top left and bottom right hand corners of the search rectangle are passed to and displayed by the Delphi program. These then form the basis of the following search routines that are carried out by the Delphi program.

The '*TS1 files search*' button is used to allow the user the option to select one or more directories in which TSOFT binary files are located and search through the headers of these files for longitude and lattitude attributes that fall within the search rectangle. Those found will be displayed in the '*Files Found*' box.

The '*DB Tables Search*' botton is used to search through one or more database tables to find records that have spatial attributes falling within the search rectangle. In the example provided in figure 4.1, the spatial coverage selected is the sub-area and hydrometeorological coverage for the Sabie catchments and four database tables have been accessed, containing sub-area rainfall data, observed streamflow data, simulated present day streamflows and simulated natural streamflows. Once all the files and database tables have been selected, the '*Search & Extract Data*' button is used to carry out the search and identify which files or records match the spatial criteria. The '*Save to Profile file*' button is then used to create a TSOFT data profile containing references to all the data series found (within either files, table records or a combination of both). Finally, the '*Run Graph*' button is used to launch the main TSOFT graph generation program with the newly created profile already loaded, as illustrated in Figure 4.2. It is now possible to display any of the data series that were selected and to carry out the normal array of possible analysis options. It is, of course, also possible to load up a different profile or add and remove data series from the profile created by the program arctest.exe.

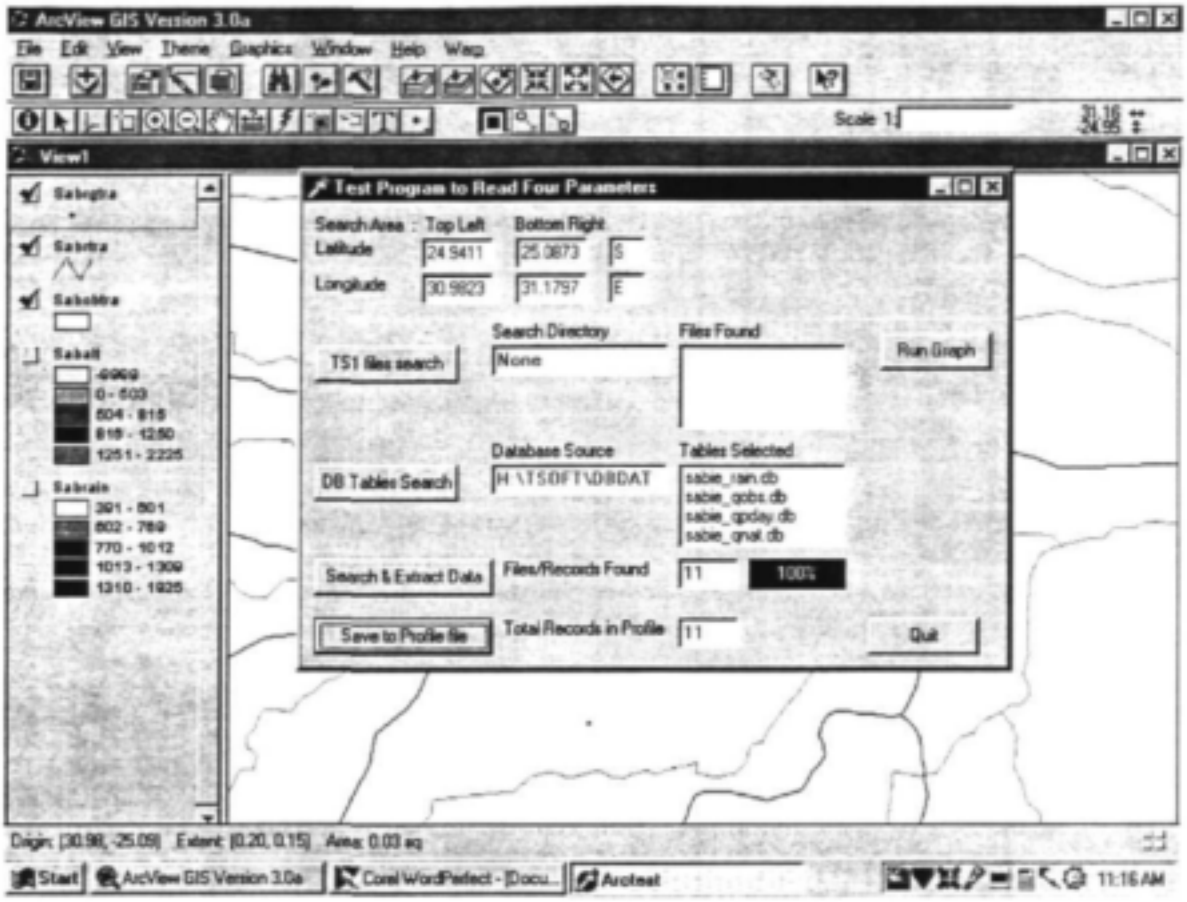


Figure 4.1 Result of calling the Delphi program arctest.exe from within ArcView (using the icon 3rd from the right on the lower menu line of ArcView).

While the facilities that have been included in this relatively simple example of an interface between the TSOFT facilities and ArcView are very limited, they do serve to illustrate what is possible and indicate that quite complex and detailed links could be established, largely dependent upon what information is shared between the spatial database (of ArcView) and the time series database (of TSOFT).

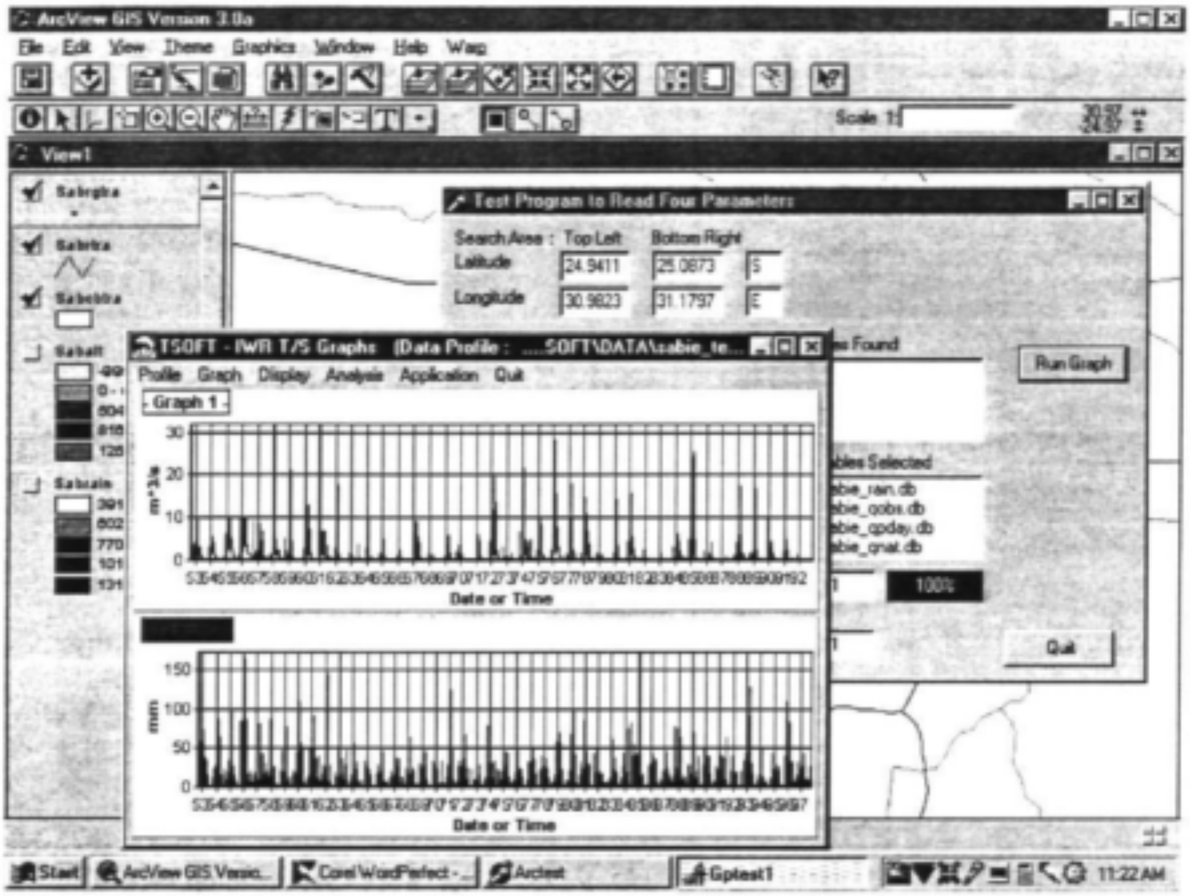


Figure 4.2 Screen image illustration with the main graph plotting program launched, one data set added to each graph panel, both panels colour reversed (turning off the black background) and the graph window decreased in size.

5. REVISED DATABASE AND SPATIAL INTERFACE STRUCTURE FOR TIME SERIES ANALYSIS AND DISPLAY SOFTWARE.

The earlier part of this report has explained the ways in which the IWR have been developing some tools for time series data visualisation and analysis, part of which has involved developing modifications to the way in which the Institute stores and accesses time series data. The eventual long term aim was to re-write all of the modelling, pre- and post-processing software that is contained within the DOS based HYMAS system. This was expected to be an ambitious project that would take quite a lot of time and resources. It is also a project that has been largely overtaken by other events.

During 1999, the IWR became intensively involved in the whole Ecological Reserve process, actually setting Reserves and developing software to facilitate the quantification process at various resolutions. These developments led to a proposal to the WRC for a project designed to develop a more complete decision support system for quantifying the Reserve. As part of the Reserve determination process involves the use of simulated hydrological data, it makes a great deal of sense to abandon any attempts to re-write the HYMAS system until such a time that the Reserve DSS specifications have been defined. The models and time series processing, display and analysis programs could then become an integral part of the Reserve DSS. Looked at in another way, the Reserve DSS could be considered as an integral part of a much more flexible software package that can deal with a variety of water resource and hydrology related problems.

During September 1999, Prof. Hughes visited the Institute of Hydrology in the UK and was shown the recent developments being carried out on their 'Micro-Lowflows' software. These developments were very impressive and the basics of the data access design appeared to be very useful, flexible and relatively straightforward to implement. Some of the design philosophy explained below has been largely based on what was noted during this visit.

The essence of the proposed design is to make use of a spatial interface to drive access to other data and to have a series of additional utilities that either use those data, or generate and store additional data. The utilities could be part of the main program, or they could be external programs that can be called from the main program. Example utilities could include data importing routines, database management routines, model parameter estimation and editing routines, time series simulation models and graphical display programs (such as those described elsewhere in this report). The eventual details of the design should be such that the number of possibilities are limitless and their implementation quick and efficient.

5.1 Proposed spatial interface

The proposed spatial interface will be through the use of MAP OBJECTS within a DELPHI program making use of ARCINFO coverages which will have been digitised externally. Each of the coverages will be identified as a ***feature*** which will have ***attributes*** associated with them. While it is possible to have some attributes defined as part of the ARCINFO coverages, it is proposed that these be stored separately from the ARCINFO tables within a database structure as defined below. The ARCINFO coverages will therefore be used to define the ***features*** and their spatial characteristics only. The main reason for adopting such an approach is that ARCINFO table files are not readily capable of storing (as attributes) some types of data. There is, however, no reason why attributes stored in existing ARCINFO coverages could not be saved to additional database tables as part of the facilities of the program.

5.2 Proposed main database structure

The ARCINFO tables have the following general structure, the specific fields after the ID field being dependent upon the coverage type (i.e. line, polygon or point) and whether or not additional attribute information has been included.

ARCINFO coverage tables

ID	Other information (including spatial) depending on coverage type			

The **DATA DICTIONARY** of the proposed database will be made up of four components, the first being the list of ***features***. The following table illustrates the structure of the first component, which provides the link to the spatial data.

DATA DICTIONARY component 1

Feature Code	Feature Name	Feature Source	ID Field	Desc. Field

Feature Code: A unique number that is associated with a single ARCINFO shape file.

Feature Name: The name of the ***feature*** for display and selection purposes.

Feature Source: The path and filename of the shape file.

ID Field: This is the field in the coverage table file that uniquely identifies a polygon, point or line.

Desc. Field: This is the field in the coverage table file that the user selects as the main coverage attribute (i.e. part of the shape file definition) that is to be used to describe the cover component (for example, the quaternary catchment name in the WR90 polygon coverage).

The second component of the data dictionary uniquely associates external (i.e. not part of the shape file) ***attributes*** with a feature defined in data dictionary 1.

DATA DICTIONARY component 2

Feature Code	Attribute Name	Attribute Code	Datatype

Feature Code: A unique number associated with a feature (shape file) contained within data dictionary 1.

Attribute Name: The name of the attribute.

Feature Code: A unique number associated with the attribute.

Datatype: The type of data associated with the attribute that also defines the type of table that the attribute data will be stored in :

0 = Text : Simple text string with a limit of (say) 80 characters.

1 = Integer : Single integer number value.
2 = Real : Single real number value.
3 = T/S : A set of values that represent a time series of data.
4 = Bitmap : A bitmap image.
5 = Array : A set of values representing a single attribute (not time series)
6 = Memo : A text document or similar set of strings.

The third component of the data dictionary provides the link between the list of attributes and the database tables in which the data are stored.

DATA DICTIONARY component 3

Attribute Code	Table Code	Database Alias	Table Name	Max. Rec

Attribute Code: The link to data dictionary 2.

Table Code: A unique code to identify the table in which the data are stored.

Database Alias: The alias name (which also specifies the path and type) of the database in which the data are stored.

Table Name: The name of the table in which the data resides which must have a structure (number and type definition of fields) that is compatible with the Datatype from component 2.

Max. Rec: The maximum number of records in the table defined by Table Name. This is stored here so that when additional data are added it is straightforward to increment the Table Record of data dictionary 4.

The fourth component of the data dictionary links the Table Codes from data dictionary 3 with the records in the spatial database (using the ID Field of data dictionary 1) and the records in the database table associated with the Table Code.

DATA DICTIONARY component 4

Table Code	Spatial Record	Table Record

Table Code: The reference to the table defined in data dictionary 3.

Spatial Record: The number in the ID Field of the table associated with the shape file referred to in data dictionary 1.

Table Record: The number in the RECID field of the table associated with Table Code in data dictionary 3.

The remainder of the database structure is to be made up of as many tables as necessary containing the real data. Thus far in the design seven table types have been identified as required to store the different types of data commonly used in a variety of water resource related problems. These correspond to the definitions given by the ***Datatype*** field in component 1 of the data dictionary.

Text type tables : Made up of a simple structure of two fields.

RECID	Text (80 characters maximum)

Integer type tables : Made up a simple structure of two fields

RECID	Integer value

Real type tables : Made up a simple structure of two fields

RECID	Real value

T/S type tables : Made up of a RECID field, several time series attribute fields (start and end date, units, data formats, etc.) and a binary large object (BLOB) containing the actual time series data as just data values or as pairs of time and data values (defined within the attribute fields). These are very similar to the database tables that currently used to store time series data within the IWR and could be accessed by the existing TSOFT programs without a great deal of modification.

RECID	Time series attribute fields (similar to existing time series database tables)												T/S BLOB

Bitmap type tables: Made up of a RECID field, several bitmap attributes (such as title, x and y sizes, date and source type - jpg, bmp, etc.) and a binary large object containing the bitmap image.

RECID	Bitmap attribute fields (to be decided)								Bitmap BLOB

Array type tables: Made up of a RECID field, several array attributes (such as title, date, format, etc.) and a binary large object containing the array data. This table type could be used to store model parameter values or a table of IFR results for example. The assumption is that one of the array attributes will point to a source which defines what each element of the array actually represents (the format field).

RECID	Array attribute fields (to be decided)								Array BLOB

Memo type tables: Made up of a RECID field, memo attributes (title, date, etc.) and a formatted memo field containing the text memo. This type of table could be used for storing notes about a flow gauging station and related data quality for example.

RECID	Memo attribute fields (to be decided)								Memo

5.3 Interface Design

The spatial interface design is expected to be a set of menus that allow the user to select the features required and then the attributes. It will also be possible to select certain components of a feature (one or more polygons or points, for example). Further activity will depend upon the type of feature and attribute data selected and could involve some standard utilities within the main program, or could involve the running of external programs (such as a time series graphing utility similar to TSOFT referred to in this report, or a rainfall-runoff model). The utilities will also have to include facilities to load and edit attribute data, associate these with the spatial attributes of the features and to manage the data dictionary.

5.4 Perceived design difficulties

Some of the development difficulties are expected to be associated with linking the spatial attributes of features with the other data attributes. This point may be particularly relevant when importing data sets from external sources and associating them with spatial attributes of existing features.

One of the expected difficulties might be establishing a satisfactory approach to identifying the upstream-downstream connectivity of channels and/or sub-catchments. It would be very useful to have these approaches automated, but the required algorithms are never simple.

5.5 Perceived Advantages

From what is currently understood about the ease of use and flexibility of MAP OBJECTS, development time should be relatively fast, such that the developers can have prototype versions available for limited use after a relatively short time (several months). The system should be flexible enough to allow new facilities to be added very quickly and by several groups of developers at the same time (once the details of the design specifications have been agreed upon and finalised).

One of the main advantages will be to have a system that can be used by a wide variety of users at a range of different levels. A hydrological modeller should be able to use the system to set up and run models, as well as to evaluate the results. It should also be possible for a non-hydrological specialist to be able to use the system to view and compare data from different sources, or to integrate different types of data.

6. CONCLUSIONS and RECOMMENDATIONS

Although there is nothing very new about the research that has been carried out for the component of the project discussed in this report, the results have proved to be extremely useful even at this stage of the development of the software. The graph display and analysis program has been used in several Instream Flow Requirement workshops to illustrate various aspects of single flow time series and to make comparisons between flow time series. In previous workshops, the HYMAS software was used which was always less efficient and did not allow some types of analysis or graphing options to be used. The new TSOFT program allows the hydrological specialist to answer queries more successfully than before and the reactions from the non-hydrological specialists have all been very positive.

From a development point of view, the IWR staff always considered that HYMAS was relatively quick and easy to modify or add to. However, the TSOFT approach has proved to be even easier and quicker, largely as a result of greater standardisation and simplification of input data formats, but also a consequence of the use of a visual, object orientated language such as Delphi.

The degree of cooperation in software development that was originally planned for by this project was disappointing. One of the main purposes of the workshop held at the start of the project was to foster such cooperation, however, even at that stage it became apparent that the different groups involved had development concepts which differed to others, often in a fundamental manner. This made it difficult to find sufficient common ground with which to move forward in a productive cooperative manner. Other problems related to different groups having their own short term priorities and not being able to wait for others to complete their developments. Despite these rather negative comments, the IWR development team did learn quite a lot from the other groups who participated at the beginning, as well as from comments on the software once development had started. The software has now been passed on to the staff of the School of Bioresources Engineering and Environmental Hydrology at the University of Natal, Pietermaritzburg where they are developing interfaces to ACRU model output and the WDM database.

Even though the new software has not reached a level where it is fully integrated with existing hydrological modelling software, it has still proved to be useful for assessing model results generated within HYMAS. If the planned enhancements to the design of the database and the inclusion of the spatial component are carried out, the authors believe that an extremely powerful time series modelling, data display and analysis package will be the result. It is possible that the developments referred to in section 5 of this report that are being carried out by the Institute of Hydrology in the UK on their 'Micro-Low Flows' software will satisfy many of our requirements. If additional components can be added to that software by South African specialists for South (or southern) African purposes, then overlaps in development work and time may be avoided. It will be necessary to explore such avenues of future cooperation before expending too many resources on our own development work. However, even if that cannot be accomplished, specialists in this country can still learn from the concepts used by the UK developers (as the IWR already has), thus reducing our own development time. It should also be noted that similar developments related to linking spatial information to other data are being carried out, or are planned, by several other groups within South Africa. It is important that such developments be reasonable well coordinated so that unnecessary duplication does not take place and that groups working on similar developments are in contact and can learn from each other.

While many of the future developments in this type of hydrological tool carried out within the IWR will be driven by the requirements of the Institute staff for research and consultancy project purposes, comments and suggestions from other groups will continue to be heeded and incorporated as far as reasonably possible. The Institute will be embarking on a new research project in the year 2000, part of which is designed to develop a decision support system to provide many of the tools (hydrological and non hydrological) required to support the quantification of the Ecological Reserve. This will involve a cooperative effort where a wide range of specialists involved will have to make their requirements known so that they can be incorporated. From a hydrological perspective, many of the required tools are already developed in one form or another and 'simply' need to be put together into a more cohesive package.

Many of the concepts that have emerged and been clarified during this project will be used in the new project and should enable the development time to be reduced quite considerably compared to a situation in which development was starting from scratch. It has become apparent in the last few years that new approaches to the use of existing tools and ideas for the integration of a range of tools are constantly emerging. Any software that is designed to assist in the efficient application of such tools must be relatively flexible and be able to be updated without massive restructuring. It would of course be highly beneficial if the updating and revising could be carried out by several different groups so that the resources of a single group are not stretched too far, or that there is a long gap between suggestions for further development and their implementation. However, this involves a high level of inter group cooperation and collaboration that seems to be difficult to achieve with respect to software development. This point also suggests that a certain amount of duplication of effort will be almost inevitable. Fortunately, the software development tools that are now available ensure that the resources required to achieve results are much less than in the past and development times much shorter. Consequently, duplication becomes less of an issue than previously. Despite the difficulties, it is recommended that cooperation and collaboration be strongly encouraged, but not at the expense of crushing individual initiative which can occur if one approach (be it a model or a software package) becomes 'institutionalised' in favour of alternatives which may serve slightly different purposes and therefore still have value.

7. REFERENCES

Alexander, W J R (1991) **Flood Hydrology for southern Africa**. The South African National Committee on Large Dams (SANCOLD), Pretoria.

Hughes, D A (1997a) **Southern African "FRIEND" - The application of rainfall-runoff models in the SADC region**. Report to the Water Research Commission by the Institute for Water Research, Rhodes University. WRC Report No. 235/1/97, Pretoria.

Hughes, D A (1997b) The cooperative development of a hydrological time series analysis and display software package. **Proc. 8th South African National Hydrology Symposium, Pretoria**, Nov. 1997

Hughes, D A and Münster, F (2000) **Hydrological information and techniques to support the determination of the water quantity component of the ecological reserve for rivers.** Report to the Water Research Commission by the Institute for Water Research, Rhodes University. WRC Report No. 867/3/2000.

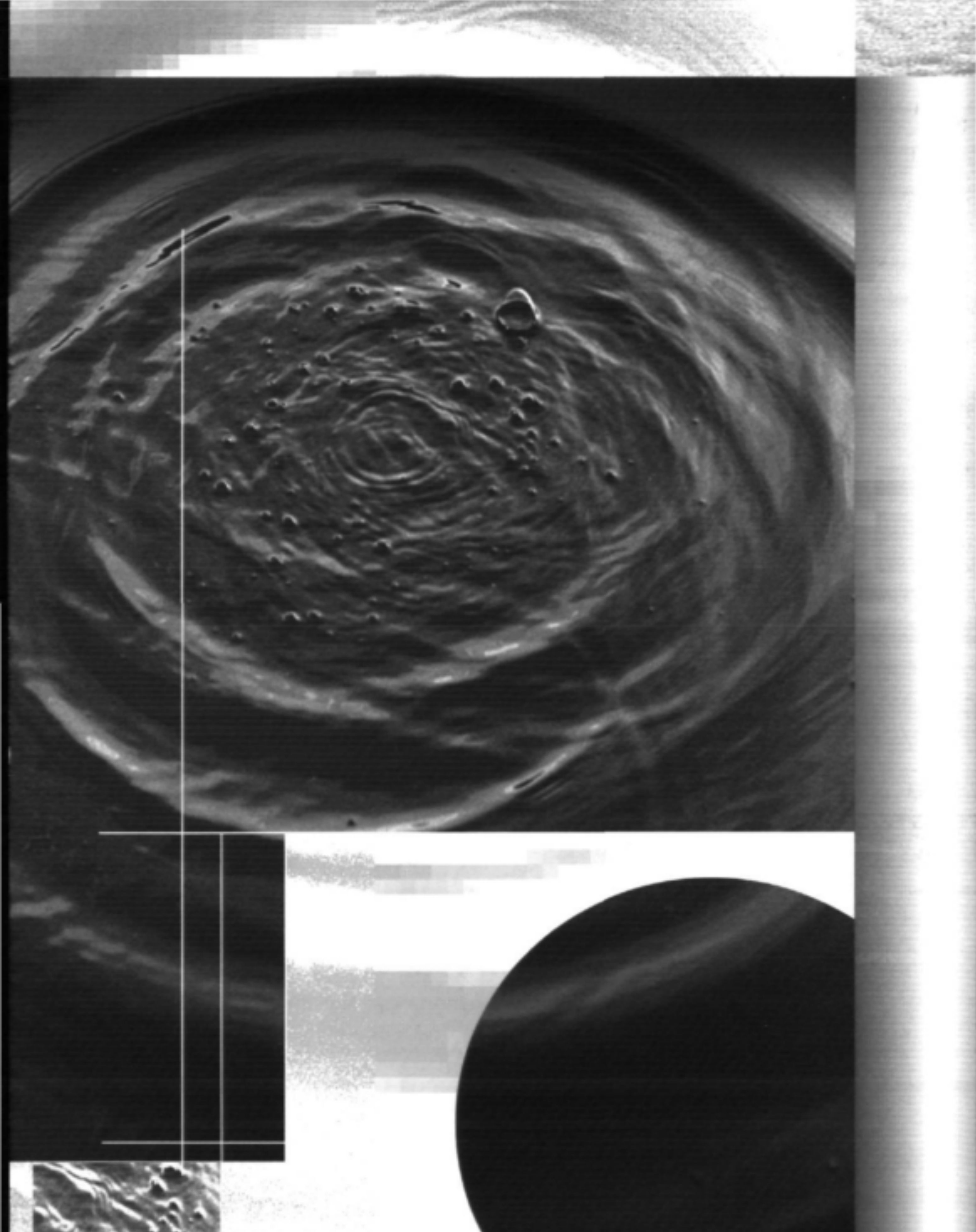
Hughes, D A, Murdoch, K A and Sami K (1994) A Hydrological Model Application System - a Tool for Integrated River Basin Management. In C Kirby and W R White (Eds) **Integrated River Basin Development**, John Wiley & Sons, Chichester, UK, 397-406.

Hughes D A, Sami, K and Murdoch, K A (1993) **Hydrological Models - Development and Application.** Report to the Water Research Commission by the Institute for Water Research, Rhodes University. WRC Report No. 138/2/93, Pretoria.

Nathan, R J and McMahon, T A (1990) Evaluation of automated techniques for base flow and recession analysis. **Water Res. Res.**, 26, 1465-1473.

Smakhtin, V Y (2000) **Simple methods of hydrological data provision.** Report to the Water Research Commission by the Institute for Water Research, Rhodes University. WRC Report No. 867/1/2000.

Smakhtin, V Y and Watkins, D A (1997) **Low Flow Estimation in South Africa**. Report to the Water Research Commission by the Institute for Water Research, Rhodes University. WRC Report No. 494/1/97, Pretoria.



Water Research Commission

PO Box 824, Pretoria, 0001, South Africa

Tel: +27 12 330 0340, Fax: +27 12 331 2565

Web: http://www:wrc.org.za