Weltdatenbank Abfluß
Bundesanstalt für Gewässerkunde
Koblenz, Deutschland

Global Runoff Data Centre
Federal Institute of Hydrology
Koblenz, Germany

Report No. 16

The GRDC Database
- Concept & Implementation -

Johannes Pauler, Thomas Decouet

July 1997

56068 Koblenz, Kaiserin-Augusta-Anlagen 15-17
Phone +49-261-1306-224, Fax +49-261-1306-280
e-mail (RFC 822): grdc@koblenz.bfg.bund400.de
e-mail (X.400): c=de;a=bund400;p=bfg;o=koblenz;s=grdc

DATABASE TABLES

Typically for the relational database system is the grouping of information in a logical way and store these groups in different tables. This prevents data inconsistency and redundancy because data is changed at only one point and is then available all through the system. These tables can be seen as logical units.

GRDC-NO	RIVERNAME	STAT-NAME	LATITUDE	LONGITUDE	AREA
6229500	Vaenern	Vaenersborg	58.38 N	12.32 E	46830
6337100	Weser	Vlotho	52.17 N	8.85 E	17618
2907500	Pur	Samburg	67.08 N	78.15 E	95100

Table 1: Example of a database table

Each table is divided in rows (records) and columns (fields). Each column in a table is of constant length. For example the column GRDC-NO is 7 characters of length, the Column RIVERNAME 40 characters and so. This again leads to a constant length of every row in that table. Now the following problems can arise:

- ☞ The information for one field is smaller than the column width: In this case disk space will be wasted. The information will be stored correctly, the missing characters will be filled with blanks.
- ☞ The information for one field is longer than the column width: Here the information will be cut.

Therefore it is necessary to fix a proper column length when designing a table so that neither information will not be cut nor too much disk space will be wasted. The design of the most important tables used in the GRDC database are shown in APPENDIX A

INDICES

Each row in a table is represented by it's primary key. This is a special field or a group of fields. It is not allowed to use one primary key more than once in a table. The primary key field(s) must normally be the first column(s) in the table. The database system software uses these keys to build relations between different tables and for fast querying a table. The primary key field(s) always needs a valid content (NULL values are not allowed). In the GRDC database the GRDC-No. is the content of nearly all primary keys. In addition, all other fields can be declared as secondary keys. These keys don't need to be unique and may be of NULL value.

The system stores the primary and secondary keys and the physical storage addresses of the relating rows in a special table, the so called index. These tables are very useful (and sometimes the only way) for data manipulation. For example, if the above table is queried for a specific river and the column RIVERNAME is not indexed the system will query the table sequentially (this means column by column) until the end of the table. If there exists an index the system will query the index table by using the optimal search algorithm and then directly accesses the fitting rows by using their addresses. The system itself decides which algorithms it will use (self-optimising query language). Care must be taken that not too many secondary key fields are declared for one table and therefore the rows grow too long.

From time to time (after updates or value changes) the database must be re-indexed. By doing this the database tables will be written to the harddisk in a new order and the index tables will be newly defined and generated. The system allocates a certain diskspace for each table. When this diskspace is filled the additional records will be written all around the harddisk. If now a new record is added to the table the index table will be newly sorted and the relative position of the new record in the diskspace is computed from the index and the storage adress. The indices of the records outside the allocated diskspace are only appended at the index table. When a query is started the system first looks in the diskspace by using special algorithms. If it doesn't find the record the remaining indices are looked up sequentielly. By re-indexing the table a new diskspace is computed and allocated and the records outside the former diskspace will be integrated in the system. Another point is to prevent or mend address conflicts in the index tables.

STEPS TO BUILD UP A RELATIONAL DATABASE

The following list shows the different steps for planning a relational database. It is recommended to work out these steps in the order below:

1. defining the attributes (fields) needed (all information that shall be stored)
2. gathering the attributes belonging together in tables (objects)
3. examine the attributes for further subdivision (e.g. name -> prename and name)
4. examine the tables for further subdivision
5. defining field types and length (=> record length)
6. defining primary keys (in special circumstances these keys need to be constructed)
7. defining secondary keys
8. ordering the columns of a table in a logical sense

OUTLOOK: OBJECT ORIENTED DATABASE MANAGEMENT SYSTEMS (ODBMS)

In future the object oriented database models will gain importance. But this change will not be drawn as sharp as from the hierarchical to the relational model in the early 90's, because modern relational systems adopt items from the object oriented concept (e.g. Data Warehouse, object oriented query languages, distributed databases...). At the present there exists no standardization of ODBMS and compatibility to other standards is not established. The table below shows the principle characteristic differences between relational and object-oriented database systems.

Relational DBMS	Object-Oriented DBMS
Table oriented	Object oriented
Information splitting (entity principle); sharp seperation between data and functions	Unity of information and functions (class principle)
	Class heritage
Manual generated primary keys	System generated object-identifiers
Descriptive query language; embedded SQL	Object-oriented languages (Smalltalk, C++)
No user-defined data types; Complex data types (maps, pictures, sounds...) only as Binary Large Objects (BLOBs)	User-defined data types and functions (user needs to define the query in Smalltalk or C++ by using class heritage)
Access modes depend on the used operating system	Access modes depend on type declaration as public, published or private
Query result as a temporary table in RAM or on harddisk	Query result as an object in RAM
Short response time	Long response time; resource consuming

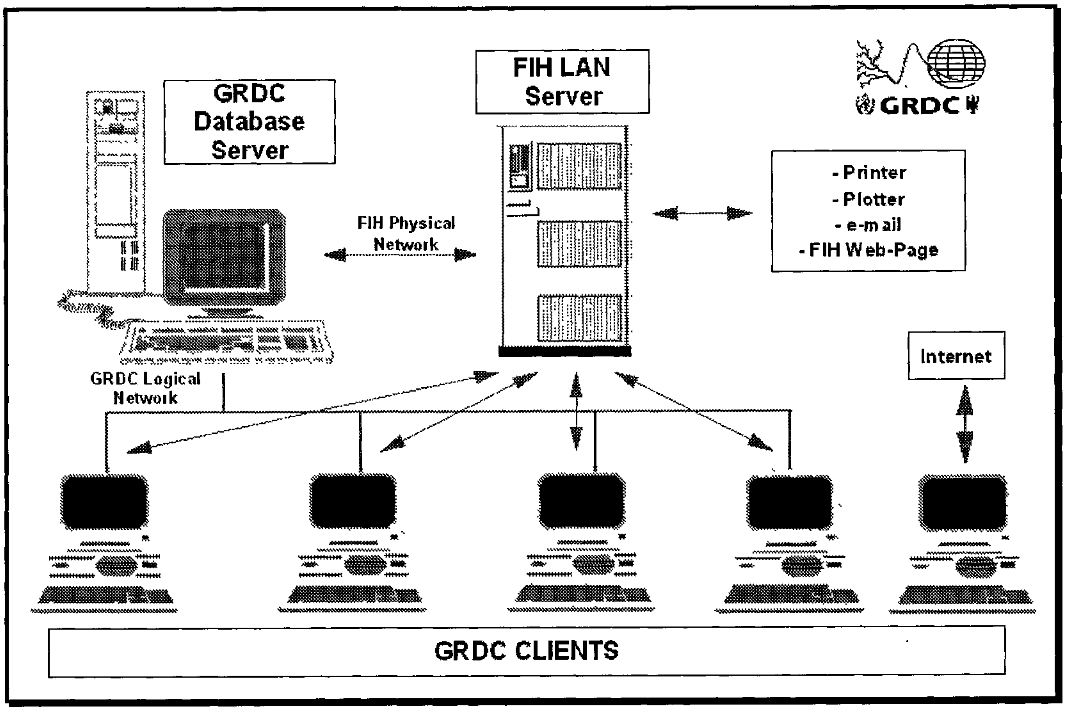


Figure 2: Network configuration of the GRDC

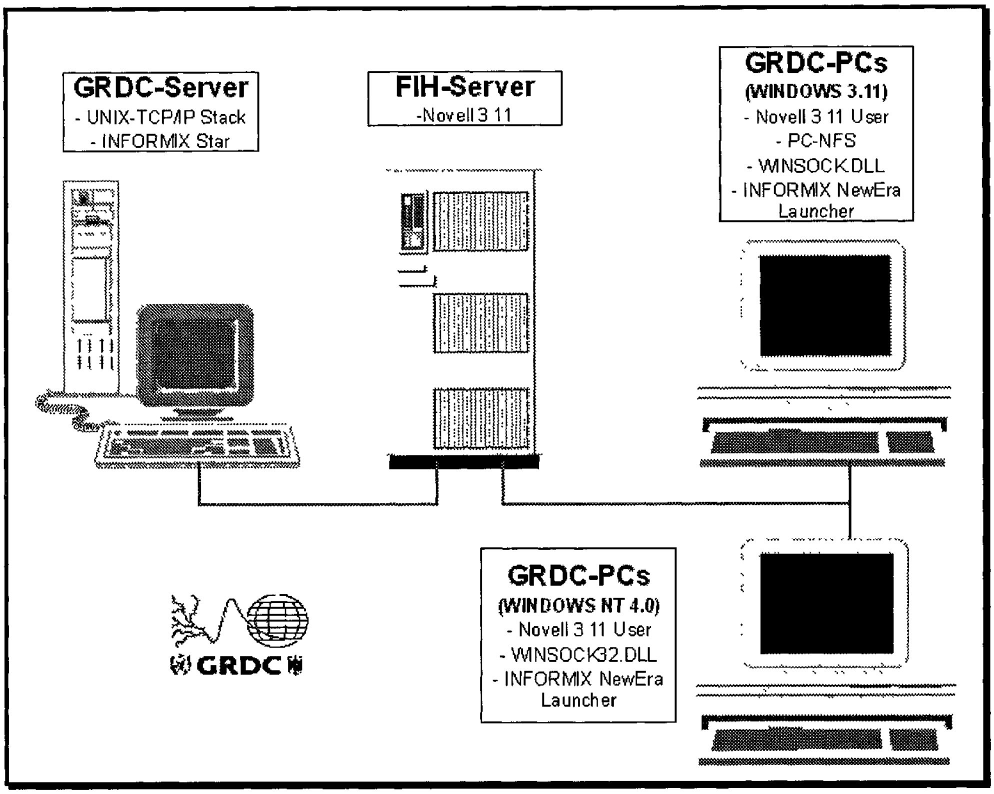


Figure 3: Network protocols and drivers (using WINDOWS 3.11 clients)

GENERATING CATALOGUES

Generating the catalogues of available and missing data is done after each major data import. These interim catalogues are for the internal use of the GRDC staff. Public new catalogues are published twice a year. These catalogues are available by e-mail, diskette or can be downloaded from the GRDC website on the WMO or FIH webserver. In addition, a *Catalogue Query Tool for WINDOWS* completes this package. This tool allows the user to query the catalogues in an easy way. It will be explained on page 36 later in this report. Because the catalogues are too voluminous they are not sent as a printout.
Figure 6 shows the dataflow and the affected tables when a catalogue is generated.

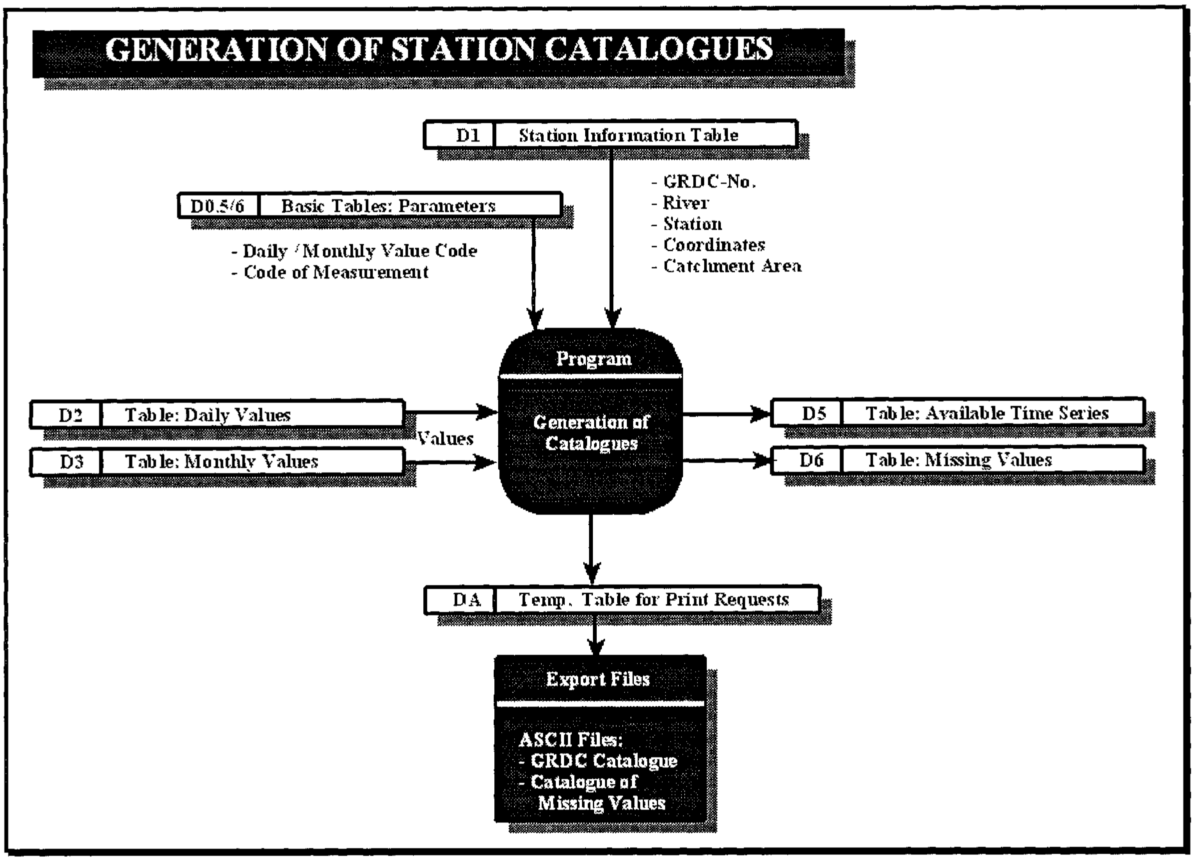


Figure 6: Dataflow Generating Catalogues

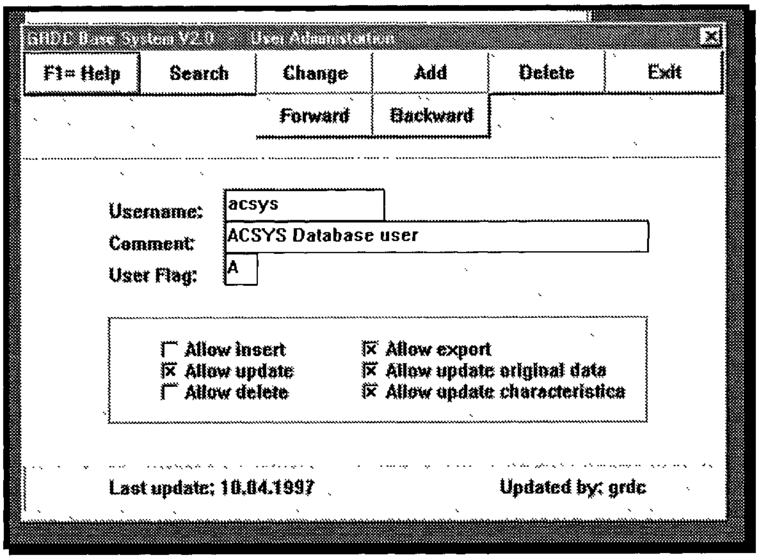
The GRDC Base System: Generating Catalogues of Available or Missing Time Series

The generation of the catalogue of available data works the same way as with the catalogue of missing values. The only difference is that the catalogue of available data shows the complete time series (date of the first available record to the date of the last available record) including the missing values and is more thought as a station catalogue, while the catalogue of missing values lists all missing time intervals of each station.

To generate a catalogue the function "Generate Catalogue" in the dialogue 'Catalogue of Stations' or 'Catalogue of Missing Values' is used. Here the following options are available:

- Generate Catalogue for new Daily Data: When new daily values are imported, the corresponding record in the table 'grdc' will be marked. Now only the information for these marked stations will be updated (or added) in the catalogue tables.
- Generate Catalogue for all Daily Data: Here the catalogue tables are cleared and newly filled by scanning all stations in the table 'grdc'.

User Administration



Username: acsys
Comment: ACSYS Database user
User Flag: A

☐ Allow insert
☒ Allow update
☐ Allow delete
☒ Allow export
☒ Allow update original data
☒ Allow update characteristics

Last update: 10.04.1997
Updated by: grdc

Figure 7: Dialogue User Administration

This task increases with the growth of GRDC's cooperation in international projects. The administrator is able to define a special flag for each project. So the GRDC does not need to generate special databases for each of these projects (and consequently, the need to administer these databases!).

The administrator defines a database user 'acsys' and identifies him/her with the flag 'A'. Now he/she marks all stations in the Register of Stations that belong to the ACSYS project with this 'A'. When the user 'acsys' connects to the database, the system recognizes this and will show only the information and data of the ACSYS project.

The second aspect of the user administration is the management of user accounts like changing, adding or deleting values in the database. So for example the administrator may give the right to change values for the user ACSYS but not the right for adding or deleting, this user can now use the Plausibility Tool and overwrite the original data with corrections but he/she is not able to delete suspect values or add manually new data sets. By doing this the original data are still available in special fields of the table. The GRDC uses this possibility to prevent accidental manipulation of its data by users who are not confident in the use of the GRDC tools e.g. guest scientists.

Generation of new Station Catalogues

The generation of a new set of catalogues (Catalogue of Available Time Series and the Catalogue of Missing Values) is time consuming and only needs to be done after updating a large amount of stations. Therefore it is useful to do this on a special machine with an operating system that allows real multitasking. For example the generation of a catalogue of missing daily values will take more than 9 hours (end 1996). The resulting ASCII files are transferred to the GRDC's staff computers by FTP or telnet.

Export of Data to ASCII Files

This function is used whenever a user needs the data of special stations. Here the user marks the data by using the Database User Tool and if he is not allowed to export the data by himself addresses the administrator to extract them for him. The purposes may be a data request from customers or the generation of special data products with tools which cannot directly access to

ADDITIONAL GRDC TOOLS

The tools briefly described in this chapter improve the utilities of the GRDC Database System with regard to data queries by users, quality check and the visualization of the database for monitoring purposes. The description of the tools is presented in an 'overlook' way and is not meant to describe the tools and their functionality in detail.

GRDC PLAUSIBILITY TOOL

This chapter describes in an informative way the main functions of this tool as an extension of the GRDC database system.

The Plausibility Tool allows the graphical check of mean daily and mean monthly discharge hydrographs of one or several stations. The tool is meant to be operated by an experienced hydrologist who can make sound plausibility judgements on the basis of graphical inspections and filling of data gaps using the methods described below. Due to the large variability of hydrological regimes an automated process is not desireable.

The tool was developed by the GRDC in close collaboration with TRITON GmbH.

Purpose

The main idea for developing the GRDC Plausibility Tool was:

- ☞ Checking the time series of the GRDC Database for inconsistency and suspicious entries
- ☞ Filling the missing values or correcting suspicious data by using different algorithms or manually

Suspicious data normally can be recognized by a visual inspection of the time series curve as:

- ☞ a divergence between chronological 'neighbourhooded' data of the same station
- ☞ a divergence between the long year means of the same station
- ☞ a divergence between neighbour stations in the same time series.

Implementation

The tool consists of two seperate programs:

1. the database access and the computing routines (written in INFORMIX NewEra)
2. the graphical visualization (written in C++)

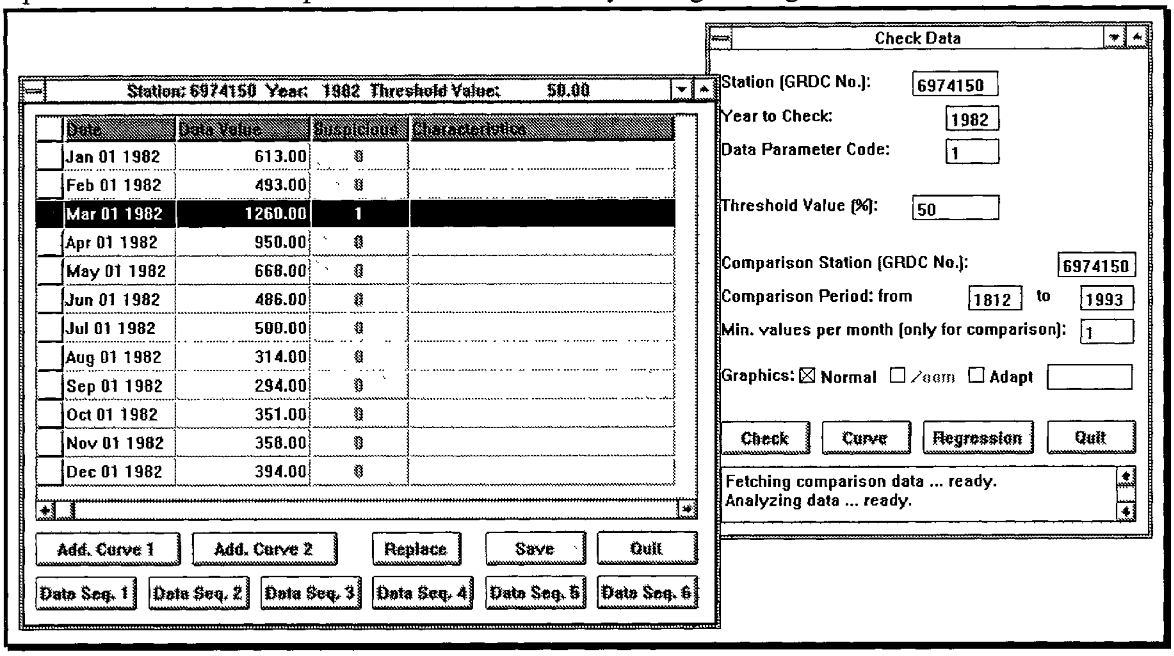
This is because INFORMIX NewEra works as an interpreter and is for this reason slower than a compiled program. The tool itself directly accesses the database and is implemented on Windows 3.11 and Windows NT 4.0.

Software handling

As mentioned above this application is directly accessing the database. This means that the data is directly read from the database and the changes are directly written back to the proper table if the user owes the rights to do so. To prevent loosing the original values there exist two columns for each discharge value in the tables 'mome' and 'tame' (see Appendix A). When importing discharge data each value is written in both columns. So when changes are made by using the Plausibility Tool the corrected value is only transferred to one column while the original value still remains in the second column. Additionally a flag is set for the changed value to show the method of correction. If for example a typing error is detected the original value can be corrected by the database administrator, too.

Presently customers receive the original datasets if they do not ask for the corrected data, while the GRDC uses the corrected data when generating data products. One handicap at present is that the customer is not able to get to know about the correcting method, but it is planned to implement new export file formats where the flags for each value are exported, too.

The checking of daily and monthly data work the same way. Before starting a session the username and password need to be entered and will be checked by the database system. Then the user is connected to the database. Now the 'Check Data' window (figure below) appears and the user needs to enter the information for the session. Now the requested data are drawn from the databank and the data of the year to check is shown in the data selection table (figure below). In the 'Suspicious' column a flag is set by the program if a value is detected that is out of the threshold limitation ('Threshold Value' in the figure below). This threshold value can be pre-determined to adapt the tool to different hydrological regimes.



Station: 6974150 Year: 1982 Threshold Value: 50.00

Date	Data Value	Suspicious	Characteristics
Jan 01 1982	613.00	0	
Feb 01 1982	493.00	0	
Mar 01 1982	1260.00	1	
Apr 01 1982	950.00	0	
May 01 1982	668.00	0	
Jun 01 1982	486.00	0	
Jul 01 1982	500.00	0	
Aug 01 1982	314.00	0	
Sep 01 1982	294.00	0	
Oct 01 1982	351.00	0	
Nov 01 1982	358.00	0	
Dec 01 1982	394.00	0	

Station (GRDC No.): 6974150
Year to Check: 1982
Data Parameter Code: 1
Threshold Value (%): 50
Comparison Station (GRDC No.): 6974150
Comparison Period: from 1812 to 1993
Min. values per month (only for comparison): 1
Graphics: ☒ Normal ☐ Zoom ☐ Adapt

Fetching comparison data ... ready.
Analyzing data ... ready.

Figure 13: Session selection and data selection table

By using the buttons 'Add. Curve' two additional time series can be added for graphical inspection, while additional time series for computation can be added with the buttons 'Data Seq. 1' to „Data Seq. 6".

Session Features

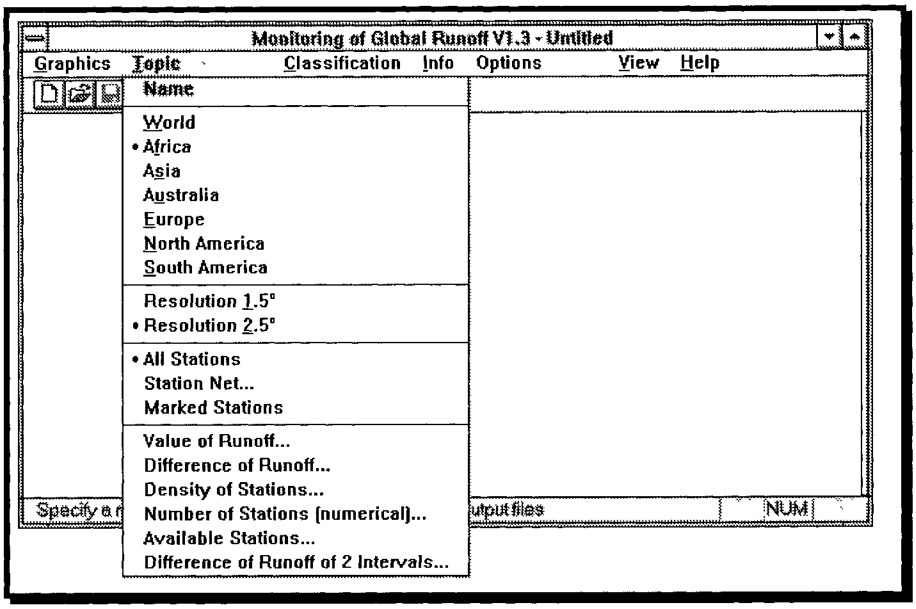


Figure 16: Session features

For each session the features must be selected in the following way:

1. the map theme (world or one of the WMO regions)
2. the grid resolution (1.5° or 2.5° gridding)
3. the kind of stations that shall be affected
4. the kind of information that shall be visualized

The user now starts the query and the necessary information is drawn directly from the GRDC Database. After the query has ended a map or a series of maps will be drawn. The selected information will be visualized in coloured grids depending on different categories. The classification is computed automatically, but the user owes the possibility to change the category limits and colours manually (figure 17).

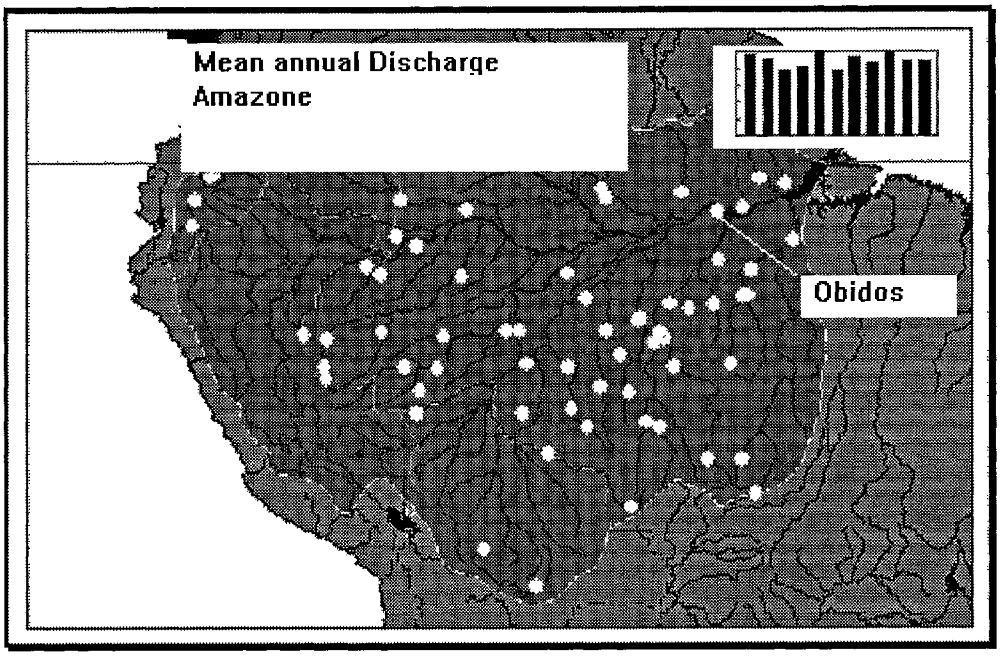


Figure 20: Snapshot with statistical graphs

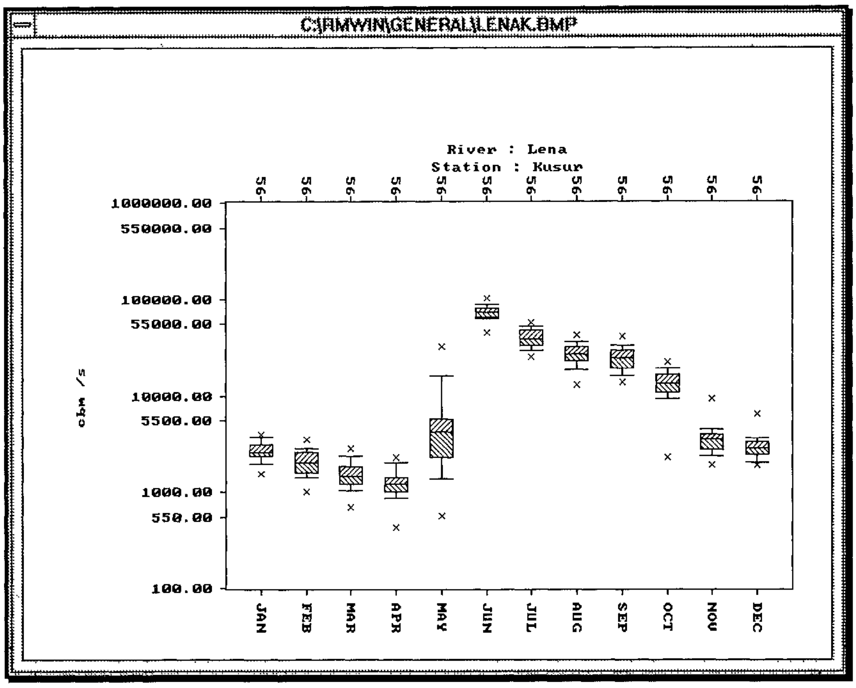
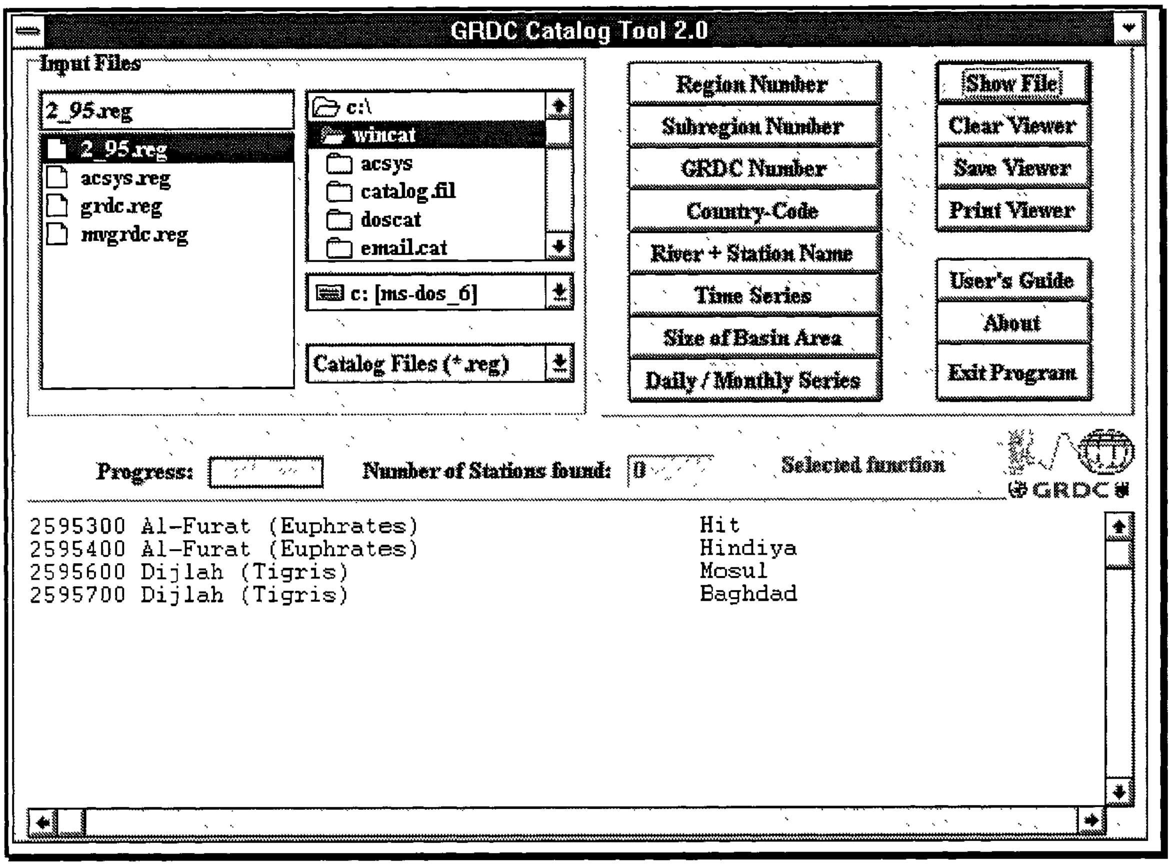


Figure 21: Statistical Graphic as a bitmap

The statistic tool provides the tasks Basic Statistics, Hypothesis tests, Multiple linear regression and Frequency distribution plots:

- Basic Statistics (mean, mode, median, percentiles)
- Hypothesis tests (parametric / non parametric tests, ANOVA and time series)
- Frequency distribution (Normal, lognormal and cumulative distribution)

All statistical functions require at least one set of data, some require more data sets. Therefore the first task is to add one or more data sets into the statistic window (Figure 22).



```
2595300 Al-Furat (Euphrates)                    Hit
2595400 Al-Furat (Euphrates)                    Hindiya
2595600 Dijlah (Tigris)                         Mosul
2595700 Dijlah (Tigris)                         Baghdad
```

Figure 24: The GRDC Catalogue Tool

The following functions are available for querying the catalogue:

☞ Region Number (WMO Regions)
☞ Subregion Number (WMO Regions and Subregions)
☞ GRDC Number (the 7 digit GRDC Number)
☞ Country Code (the 2 character country code)
☞ River + Station Name (River and/or Station name)
☞ Time Series (stations within a certain time interval)
☞ Size of Basin Area (stations within an area size interval)
☞ Daily / Monthly Series (stations with daily / monthly data)

All functions work as well with the Catalogue of Stations as with the Catalogue of Missing Values, except the function "Time series". This is because of the different date formats in these files.

The results of all queries are added to the viewer and can be stored as a new catalogue file by using the option "Save viewer". After saving, the file-list on the left will be updated immediately and therefore an input file must be clicked before starting a new query. Besides this the viewer's content can be printed out on paper.

The file-list is masked by default for GRDC catalogues (*.stn). Using the file type selection window, any other ASCII file can be shown in the file list window and loaded to the viewer.

Table „*datv*“ (Available Time Series)

Col.-Name	Type (Width)	Key	Comment
da_grdc_nr	char (7)	A	GRDC station number
da_ta_mo	char (1)	A	Flag: Daily / Monthly data
da_me_code	smallint	A	Flag: Code of measuremnt
da_jahr_a	char (4)		Starting year
da_monat_a	char (2)		Starting month
da_tag_a	char (2)		Starting day
da_jahr_e	char (4)		Ending year
da_monat_e	char (2)		Ending month
da_tag_e	char (2)		Ending day
da_updater	char (12)		user who generated this record
da_flag	char (1)		Flag: Marked record
da_last_update	date		Date of record generation

Table „*datl*“ (Missing Time Series)

Col-Name	Type (Width)	Key	Comment
dl_grdc_nr	char (7)	A	GRDC station number
dl_ta_mo	char (1)	A	Flag: Daily / Monthly data
dl_me_code	smallint	A	Flag: Code of measurement
dl_jahr_a	char (4)		Starting year
dl_monat_a	char (2)		Starting month
dl_tag_a	char (2)		Starting day
dl_jahr_e	char (4)		Ending year
dl_monat_e	char (2)		Ending month
dl_tag_e	char (2)		Ending day
dl_updater	char (12)		user who generated this record
dl_last_update	date		Date of record generation

Table „*dbuser*“ (Database User Administration)

Col-Name	Type (Width)	Key	Comment
us_name	char (12)	A	User name
us_del	char (1)		Flag: allow delete
us_ins	char (1)		Flag: allow insert
us_upd	char (1)		Flag: allow corrected data update
us_exp	char (1)		Flag: allow data export
us_imp	char (1)		Flag: allow data import
us_upd_org	char (1)		Flag: allow original data update
us_upd_mwmerk	char (1)		Flag: allow update methode flag
us_bez	char (40)		comment
us_markz	char (1)		user flag (A=ACSYS...)
us_updater	char (12)		user who generated or changed this record
us_last_update	date		Date of record generation or change

APPENDIX B: EXAMPLE OF THE GRDC MONITORING TOOL