DATABASES AT THE UK MET. OFFICE

C.T.Little
U.K. Meteorological Office
Bracknell, U.K.

1. INTRODUCTION

The UK Met Office has a large complex system of communications and computers to meet a wide variety of requirements and commitments. This paper describes those collections of data which have standardised software to store, manage and access the data. This is a pragmatic rather than a theoretical definition of a database. Data on the telecommunication computers, which have only a store and forward message switching role, are not discussed. There is a variety of computers used in real time systems to deliver satellite imagery, radar imagery and locally processed satellite soundings, and these are not considered within the definition either. Similarly, at present, the supercomputer is used for number crunching only, and it is not considered appropriate to provide databases on it for reasonably widespread access. All of the functionally important databases are on the IBM mainframe, and fall into two categories: observational and field orientated.

2. FIELD DATABASES

2.1 Output from the Numerical Weather Prediction (NWP) suite is interpolated to standard levels, locations and variables suitable for use by forecasters and the individual files of variables are stored in a database called the PRINTFILE. Its main role is to isolate the forecasters and their support programs from changes in the NWP system, such as differing levels and coordinate systems. The Printfile has existed in various versions for nearly 20 years.

2.2 Recently, an IMAGERY database has been set up to simplify access to the wide range of satellite and radar imagery used by forecasters. It is relatively simple, consisting essentially of an index of currently available images.

2.3 Output from both the imagery database and the Printfile is combined into complete pictures, for dissemination to forecasters and customers. This PRODUCT database reduces the work required to supply commonly required products, such as North Atlantic Surface Pressure charts.

2.4 All of the above databases will be mentioned in subsequent talks by the staff who support the Central Forecast Office.

3. OBSERVATIONAL DATABASES

3.1 Synoptic Data Bank

The central observational database is the Synoptic Data Bank (SDB) which collects GTS, UK and other data from the telecommunications systems and makes them available in near real time.

Data can be retrieved via a standard subroutine call, or via an interactive, menu-driven, interface. The interactive interface can also be used to set up jobs to retrieve old data banks. Data is stored in 12-hourly files, with satellite data in separate files. The most recent five days' data is kept online (about 100 MBytes) and older data is copied to tape and kept for five years.

Data is received in bulletin format and then split into individual reports and quality controlled. Quality control flags are then stored with each report. No observations are overwritten but the best, usually latest, version of a report is labelled as 'preferred'.

Retrieval of reports can be in characters, as stored, or converted to integers and consistent units. Retrieval can be by individual elements or station number, WMO block, satellite pass etc.

3.2 Metar/Tafs Database

Metars and Tafs received by the SDB are directed into a separate database, controlled by the commercially supplied CA-Cullinet IDMS database. When the Heathrow forecast office moved from London Airport to Bracknell, there was a requirement to give the aviation forecasters rapid access to the most recent (24 hours) Metars and Tafs, which they had previously accessed on their obsolete minicomputer system. Instead of adding the extra data to the SDB, it was decided to have a separate database implementation. This

was because:

a) The SDB, originally designed about 20 years ago, consisted of about 1 million lines of Assembler, with a small, but increasing proportion of Fortran 77, and required a team of 6 to maintain. A commercial database without the overheads of in-house software maintenance was worth considering against any possible drop in performance.

b) The potential volume of Tafs and Metars may have swamped the rest of the SDB with deleterious effects on SDB response times.

The result of the pilot project was that the service was delivered working, on-time, with reasonable response time (<1 second). However, some programming effort was needed to convert data (current databases seem very character orientated) and performance is disappointing. After 6 months of tuning, the database is still 30 times bigger and slower than the equivalent SDB software, in terms of machine resources. Twenty four hours' data takes 27 MBytes.

3.3 IDMS

There are a number of other databases using IDMS. The administration and accounting database uses 225 MBytes of online storage and the library bibliographic database uses 160 MBytes. The biggest, using 265 MBytes, contains a mixture of climatological, agrometeorological and forecast data to support agricultural customers. A similar database of 50 MBytes supports the ship-routeing service. These databases are distributed between two simultaneously running copies of the IDMS software. One copy is considered meteorologically operational, the other more developmental.

These databases use the strength of commercial databases:- presenting uniformly, a wide variety of disparate, but associated, data, and allowing new user dialogues to be constructed relatively quickly.

3.4 Observation Processing Database

The Observation Processing Database (OPD) also retrieves its data directly from the SDB. All the data used by the NWP system is assocciated with further quality control results, and statistical error values, produced by the NWP system. It is written in Fortran 77, and will be discussed in a subsequent talk.

It was decided to construct a separate database because it would have entailed a large scale change to the existing SDB, and hence would be a high technical risk. The modellers could also develop something more quickly when initially not constrained by operational change control procedures. Also the contents of the OPD are dependent on the formulation of the NWP suite and therefore this division maintains an SDB relatively independent of the suite.

3.5 Climate Data Bank

The Climate Data Bank retrieves the bulk of its data from the SDB after three days, when most GTS data has arrived, and subjects it to much more stringent quality control procedures. Climate data are also received on postal forms and on Automatic Weather Station data cassettes. After the exhaustive quality control, the data is kept indefinitely, in two forms: as time series for a given station, and in monthly records for all stations. The data, both online and offline, is stored and accessed by a common set of indexed sequential I/O routines called GPAccess, written nearly twenty years ago.

The series of data include:

Global Upper Air Data	since 1968, offline
Global Marine Surface Data	since 1854, offline
Global Monthly Climat Data	since 1738, 10 years online
UK Hourly Synoptic Surface Data	since 1948, 18 months online
UK Hourly Climat Surface Data	since 1948, 18 months online
UK Weekly Soil Moisture Data	since 1969, 1 week online
UK Hourly 5km Rainfall data	since 1981, offline

The total amount of data is currently 18 GBytes.

A subset of the SDB data is also stored offline, for at least 25 years, with the same GPAccess method, but without any further quality control procedures.

The GPAccess software has various problems: it is too low level, and the users have to know detailed record formats and blocking structures; existing datasets cannot be easily expanded to encompass new data types. Many of the datasets use different software for access, causing a serious maintenance problem (30 people are involved in CDB maintenance and quality

control). As the format is based on 16 bit words, and the date was stored as a Century Day, all file formats had to alter on 14 September 1989, the 32768th day of the century.

4. CURRENT DEVELOPMENTS

4.1 Climsyn

The CDB team had a requirement, to meet customers needs, to store, online, the latest 3 months' global surface data. A pilot project was set up with the SDB team to store this data, after extraction from the SDB, as BUFR messages. Two messages are stored for each station: one with 3 hourly synoptic data in sequence and one with the daily climatological data. No quality control information is stored at present, neither are 8 or 9 groups, though work has started on the latter. So far, one month's data has been accumulated. Storage is compact (an estimated 0.1 GByte for the full 3 months of 5000 stations with 8 and 9 groups, compared to 1 GByte for the current SDB). Of course the current SDB does contain a number of versions of any given report. It is estimated that about 500 CPU seconds of IBM 3084 time will be required daily to remove old data and add new data to the rolling 3 month database. This is based on the technique of expanding the messages, removing the old report, adding the new one at the end and recompressing the BUFR message. A less CPU intensive approach, but more demanding of I/O, would be to have separate messages for each day. Intermediate strategies, such as compressing data on a weekly basis are being explored. The software was originally written in Fortran 77. An optimising compiler gave a 30% speed up and coding the bit manipulation kernels in Assembler gave another 30%.

3.2 This year a comprehensive review was carried out and new user requirements were established for the SDB and the CDB, with view to establishing common interfaces across both databanks. Work will start soon on the design of the new SDB and CDB. Backward compatibility with existing series of data will have to be maintained. Almost certainly more data will be kept online as the cost of storage and computing decreases.