# Data Assimilation Research Centre (DARC)
# Data Management Plan

## Executive Summary

The Data Assimilation Research Centre (DARC) has numerous data interactions from accessing satellite observations to producing 4-dimensional NWP model outputs. This document outlines the key datasets in terms of inputs and outputs to DARC research. Archival of key data products is discussed and key emphasis is placed on Envisat data that is of interest to the wider academic community. BADC plans and funding implications are included.

## 1. Introduction and Scope

It is NERC policy to ensure as wide as practical archiving of data for future use. In fact, the NERC Data Policy Handbook states: "NERC grant-holders in academia are required to offer to lodge with NERC a copy of the data resulting from the supported research when it is completed, together with documentation/metadata describing these data" – see http://www.nerc.ac.uk/data/policy.shtml for a copy of the handbook.

Successful archiving is difficult unless the task and the significant effort involved are recognised at the outset of activity, and are fully built into plans. Typically this discipline will be undertaken for all continuous measurement data and for field studies where generation of data for shared use or long term reference is an aim. A sensible judgement of the level of data to be archived is required.

Produced by the BADC, this Data Management Plan is the result of discussions with:

- The DARC Director
- DARC PIs and their research teams
- Other ENVISAT users and Earth Observation scientists

Ongoing issues of data management will be agreed by the above parties.

The stated aim of the DARC is the assessment, combination and synthesis of Earth Observation data with numerical models in order to reproduce the evolution of the earth system and to forecast its behaviour. To achieve this aim the DARC must locate, extract, ingest, process, archive, and create data.

The purpose of the DARC Data Management Plan is to set up a coherent approach to data issues relating to data used by, shared amongst and produced by DARC research groups. Of additional interest to the wider research community is the archival and dissemination of Earth Observation datasets, in particular those from the Envisat satellite. Another issue that is mentioned relates to datasets outside the remit of the DARC that may be of particular use for portable data assimilation (one of the aims in the original DARC proposal).

The objective is to ensure that:

- Appropriate data support is provided to the scientists involved with the DARC.
- DARC datasets are archived and distributed in a suitable manner.
- Distribution conditions and data usage do not infringe on the individuals' rights to publish their own work.
- Potentially scientifically valuable data are kept for the long-term.
- A high quality documented DARC data archive is created.
- Data and documents can be distributed more widely to the scientific community.
- Envisat data is being archived and distributed around the scientific community.
- Funding issues are addressed in relation to all DARC data management.

Appendix 2 shows the DARC Data Protocol that all participants should abide by.

## 2. Key issues in the DARC Data Management Plan

### 2.1 The Structure of the DARC – A virtual centre

The Data Assimilation Research Centre (DARC) is a NERC Earth Observation (EO) Centre of Excellence. The DARC has funding for 5 years and is a virtual centre, being distributed across the University of Reading (Depts of Meteorology and Mathematics), University of Cambridge (Dept of Chemistry), University of Oxford (Dept of Atmospheric, Oceanic and Planetary Physics), University of Edinburgh (Dept of Meteorology) and the Rutherford

Appleton Laboratory (Atmospheric Science Division). The Met Office is also a key collaborator on many of the DARC projects.

Further information about the DARC is available from http://darc.nerc.ac.uk

### 2.2 Limitations on DARC data support and archival

This document has the broad remit of bringing together all the data-related issues of numerous research groups. EO datasets are typically very high volume and therefore entail significant resources to archive and document. At present there is no funding to provide full data archival for DARC datasets. The BADC recently gained funding to archive data from the Envisat instruments MIPAS, SCIAMACHY, GOMOS and MERIS. Since no funding is provided for direct data support and archival of non-Envisat data, the BADC will only provide these where there is not a significant impact on other resources.

## 3. Plans for archival and long-term data integrity

### 3.1 Data flows and the DARC

Appendix 1 presents a breakdown of the major data flows affecting the DARC.

### 3.2 Current BADC Envisat holdings

The BADC currently holds around 250 GB of SCIAMACHY L1 data and 800 GB of MIPAS L1 data spanning the time period December 2002 – present. This data is being collected by the RAL receiver and stored in an archive only made available to select DARC and Envisat users.

This data is being supplied by the usual BADC mechanisms (web registration and application for dataset) to a select group of RAL, Oxford and Leicester scientists working on Envisat and DARC projects. The BADC has now gained funding to consolidate its Envisat datasets and become the ESA-designated UK archive for academic use.

### 3.3 BADC as the UK academic archive of Envisat data

In recognition that multiple research communities require Level 1 and Level 2 data from Envisat the BADC has joined forces with three key sponsors of this work: the DARC, and the UTLS-Ozone and CWVC thematic programmes. Through the EEDAC (Expediting Use of Envisat Data in Data Assimilation and Cloud Climatology Development) proposal the BADC has permission to archive level 1B products from MERIS, MIPAS, SCIAMACHY and GOMOS aboard Envisat. The data volumes amount to a maximum of about 6 TB per year and the data of interest is not available on the network so delivery will take place via CD-Rom or DVD (approximately 9500 CDs or 1200 DVDs). The BADC will begin housing the data on-line (and near-line) at RAL in mid-2004. The data will also be available via the NERC Earth Observation Data Centre.

The BADC will provide data usage information to both NERC and ESA. The level 1 MIPAS data is required by the DARC, to allow the production of new datasets based on new and improved retrieval algorithms (current DARC funding has allowed the production of algorithms, but does not allow the production of a complete dataset). Such new datasets will be of direct use in quantifying the global budgets of key atmospheric constituents. The level 1 MIPAS data will also be of use in the development of new assimilation schemes to make the best use of this data.  The level 1 MERIS data will be used to construct a TOA radiance gridded dataset, and it and the level 2 MERIS data will be provided to the CWVC programme as correlative datasets.  Meteorological data from Envisat will be also used by the DARC.

## 3.4 Access restrictions to DARC datasets

### 3.4.1 Datasets covered by this Data Management Plan

The DARC datasets covered by the plan are defined below in table 1.

**Table 1.** Definition of 'DARC datasets'.

| | | |
|---|---|---|
| **(i)** | **Supporting Datasets** | Data from sources outside the DARC (such as Envisat and other EO data) that is analysed, assimilated and processed by DARC scientists. Some of this data might be held by DARC participants or at the BADC. |
| **(ii)** | **Established Datasets** | These are datasets that have been used previous to the DARC and are still in use. They are already located at an available archive and do not need to be considered in great detail by this plan. |
| **(iii)** | **DARC-derived Datasets** | These datasets are of most interest as they are the products of DARC research. These datasets may vary greatly in size and nature. Some will be considered useful beyond the DARC and others will not. Decisions on where to archive the data will depend on the perceived future usage. |

### 3.4.2 Restrictions on Data Access

The access restrictions on DARC-related data are dependent on a number of issues:

- Current availability from known sources.
- Additional agreements for limited access.
- Academically restricted datasets.
- Data available only from links at contacts in the Numerical Weather Predicting Agencies (Met Office and ECMWF).

The access restrictions on data will therefore depend on the nature of the dataset. Where possible, data will be made publicly available but many sensitive datasets will be restricted to academic users only and in some cases data will be restricted to DARC participants.

### 3.4.3 Data restrictions in the case of NERC Thematic Programmes

If any DARC-related data is also considered a part of a NERC Thematic Programme which will have its own data management rules. In such cases the data restrictions for the Thematic Programme will apply to the data unless specific exception is made by the PIs of the project.

## 3.5 The DARC data archive

### 3.5.1 Archive location and long-term integrity

Since the DARC will use and create a range of datasets they will be dealt with in different ways. Where it is envisaged that the community will have an interest and a long-term archive is appropriate the data should be located at the BADC (resources willing) which will act as the archive of last resort. In the case of datasets being made available to the BADC the data provider is also responsible for providing documentation, metadata and software to decode and visualise the data. The data provider will also be expected to field a reasonable number of user queries that are received by the BADC.

### 3.5.2 Archiving policy on raw datasets and theoretical studies

In recognition that validated raw data (i.e. quality controlled data prior to additional processing) potentially represent an invaluable source of information for the future, DARC participants should archive them in a way that guarantees longevity and accessibility. Although not necessarily located at the BADC, validated raw datasets and their access should be fully documented at the BADC. In addition, investigators are encouraged to submit model results which would be the basis of theoretical studies or would illustrate the use of their models.

### 3.5.3 File Formats

All data produced by the DARC should be stored in standard file formats such as NetCDF (following the Climate and Forecasts (CF) Metadata Convention[1]), HDF, NASA Ames and GRIB. When deciding on an output format DARC participants should consider accessibility and future use. Compliance with the CF convention enables the inclusion of valuable metadata within the data files. Documentation on formats is available from the BADC (http://badc.nerc.ac.uk/help/formats/), as well as links to downloadable free software packages to read and write various file types.

[1] The CF convention is fully documented at: http://www.cgd.ucar.edu/cms/eaton/cf-metadata/index.html

Data received directly from EO sources such as Envisat instrument data will be stored in the format it is supplied in with tools made available to read the data.

### 3.5.4 Archive directory structures at the BADC

Datasets held at the BADC will follow the filenaming convention set out below. Groups holding data elsewhere should consider this convention when naming their files. The directory structure for the data should follow the convention:

```
/badc/PROJECT/data/YYYY/[MM]/[DD]/[hh]/[mm]/[ss]/filename
```

Where:

`PROJECT` is the code for the DARC project or EO instrument (e.g. `envisat-mip1b`, `envisat-mip1b`).

`YYYY` is the year (4 digits).

`MM` is the month (2 digits). Note that the `[MM]` directory is optional.

`[DD]/[hh]/[mm]/[ss]/` are optional directories for day, hour, minute and second. The depth of directories included to describe the time field will depend on the amount of files present. A rule of thumb is that a new level should be included if more than 100 files exist in the directory.

`filename` is the name of the data file itself (see section below on file-naming).

### 3.5.5 File-naming

Considering the diverse range and locations of DARC datasets it does not seem practicable to enforce a file-naming convention throughout. Each group should document its file-naming convention so that others accessing the data can locate the information they need to use the data.

### 3.5.6 Data submission

The BADC provides an automatic web-based file uploader accessible via the *Data* and *My BADC* tabs on the web site (or directly from: http://badc.nerc.ac.uk/data/submit.html). Online assistance is provided for this service. Alternatively, files can be submitted by FTP. Both methods are fully documented on the BADC web site.

### 3.5.7 Updates to datasets

On occasion a new version of a dataset or subset will be processed by a data provider. Files should contain the version number or history in the metadata so that newer versions can be distinguished from previous versions. When a new version of a dataset is received the BADC will send a message to all known users of the data to inform them of the change.

### 3.5.8 Documentation

Metadata are a crucial part of any archive since they ensure the readability of the data. It is therefore essential that metadata are submitted at the same time as the datasets to which they pertain. Metadata pertaining to all DARC-related data archived elsewhere should also be supplied to the BADC.

To guarantee data archive quality, full documentation on all validated raw and processed data, as well as on models and model results, should be provided to the BADC. Standard metadata will be archived within data files. An example of the sort of metadata that should be provided is detailed at:

http://badc.nerc.ac.uk/help/metadata

In addition to the standard metadata, investigators are encouraged to archive all relevant information at the BADC, including references, papers, reports etc. Designated directories will be created for this purpose.

### 3.6 Supporting Envisat collaboration with Collaborative Workspaces

The BADC will set up a collaborative workspace dedicated to Envisat users. This will be a secure web space available to registered users only where scientists can share results, documents and preliminary data files. Users will have to apply for access to this space which will be accessible from the BADC workspaces page at:

http://badc.nerc.ac.uk/community/workspaces.html

## 4. Funding issues and the DARC data archive

Producing a DARC archive at the BADC will involve significant resources in staff time and storage costs. The funding proposal for Envisat products has already been discussed but no funding is currently allocated for the archival of output data products from the DARC. Where there is agreement that a dataset should be held at the BADC the DARC may be asked to either provide staff support or additional funding to enable an appropriate archive to be established. User queries for datasets may also be fielded by DARC personnel with the BADC acting as a referrer only. Funding of datasets will be dealt with on a case-by-case basis.

## 5. Portable Data Assimilation – possible data implications

The original DARC proposal introduced the intention of liberating data assimilation from meteorological forecasting institutions and allowing research scientists to run assimilations outside of the Met Office and ECMWF. Appendix 3 provides an outline of the portable data assimilation project.

The data implications of porting the Met Office data assimilation suite would involve the requirement of:

i. An observations database accessible outside the Met Office.

ii. A new community-wide database containing auxiliary information such as obstore files and assimilation statistics. Note that the BADC has extracted the 44-years of ECMWF ERA-40 observations and holds the data which is not yet visible to users.

iii. A new archive of output from data assimilation experiments and their associated model runs.

During the lifetime of the DARC it is still unclear as to how the porting process will be undertaken. If significant funding is provided then portable data assimilation would be supported by additional staff at the Met Office and in the academic community, providing both infrastructure and expertise vital to expanding assimilation research.

## 6. Near Real Time Data Assimilation – data implications

The DARC hopes to eventually perform Near Real Time (NRT) data assimilation combining research satellite EO data with conventional observations currently used in operational NWP. Such NRT access to data would involve significant resource implications and would necessitate new data access methods to daily observations located at the Met Office or ECMWF.

This may be delivered as a wider outcome of the portable data assimilation project. Otherwise, further funding would be required in order for the BADC to create one of the following:

- A copy of the Met Office MetDB (meteorological database) which is updated daily in NRT.
- A copy of the ECMWF ODB (Observations database) which is updated daily in NRT.
- New data streams and (GRID-enabled) technology to extract observations from either the MetDB or ODB.

NRT access to the required EO datasets, such as HiRLDS, would also need to be implemented. These will be dealt with as and when NRT data is required by the community for assimilation experiments.

## 7. Sharing and re-distribution of DARC software

The DARC already provides a number of simple model codes to explain and allow simple Data Assimilation experiments. These include a 4D-Var pendulum written in Fortran, some IDL/PV-Wave code for Optimal Interpolation that shows the effect of statistical analysis parameters and an illustration of 4D-Var applied to the Lorenz system written in Matlab. These are all freely available via the DARC website at: http://darc.nerc.ac.uk/models/index.html

Where software has an application and potential user group wider than the DARC participants and is open-source it should be offered to the BADC by DARC participants. The BADC can then locate the data in a DARC or other appropriate location in its archive for future use. All software should come with full documentation by the author(s) and ongoing support will be provided by those and not by the BADC unless otherwise agreed.

## 8. Data distribution

As stated above, we cannot make a general rule for data access restrictions for all DARC and EO datasets. A password-protected access system can be set up at the BADC to reflect the defined permissions. Distribution of DARC data held at the BADC will take place via the Internet and FTP. During any restricted period, entitled DARC participants who have applied for access to the data will be allocated an account at the BADC allowing them to directly download the data from the archive. This facility will be extended to external collaborators who will have been personally authorised to access the data by DARC PIs.

The DARC website can act as a link to the data holdings at the BADC and other locations. DARC data held at the BADC would benefit from future development of data access technologies. Facilities currently under development include the e-Science NERC DataGrid, and a *Live Access Server*[2] allowing data subsetting, visualisation and format conversion via a web-interface.

## 9. Publication

Results coming out of the DARC-related research will be published in the usual way. During the data validation period of a particular campaign or project, each investigator will have the right to refuse the use of his/her results in a publication or a presentation prior to the investigator's own publication of that work. If measurements or model results from other DARC-related research are used in a DARC participant's publication, joint authorship should be offered. This will not necessarily have to be accepted, particularly in cases where due credit and acknowledgement can be given in other, possibly more appropriate, ways. References of publications should be communicated to the BADC where a list of published works will be held.

## 10. The Future of the DARC

It is clear from the information presented in this plan that the issues of DARC data management would have been more effectively dealt with if greater resources were made available. These lessons should be learned for future proposals for continual funding for the DARC.

In the planning stage there should be significant scoping of the data requirements for such instruments as GOME-2 (aboard ESA's METOP) and ADM on ESA's Explorer Mission. BADC's costings and expertise should be used to calculate the necessary resources to deliver a more complete data management programme.

[2] Live Access Server (LAS) provides a data manipulation interface to data via the web, see http://ferret.wrc.noaa.gov/Ferret/LAS/ferret_LAS.html for more details.

## Appendix 1 – Summary of DARC data interactions

Since the DARC is mainly involved with Envisat and other satellites other non-DARC research groups have been contacted regarding their use of the satellite data. The aim was to understand where overlap was taking place and to avoid multiple groups having to obtain the same datasets.

The key datasets in term of volume are:

- Envisat datasets from MIPAS, SCIAMACHY and AATSR instruments.
- Other Earth Observation datasets such as HiRDLS.

The next two sections explain the main flows of the data into the DARC/Envisat community and those that are being generated by new DARC research.

**Data flows into the DARC**
The following list includes the main data flows into the DARC. Other less significant data flows and minor Envisat users (in terms of volume) have been omitted.

### 1.1 EARTH OBSERVATION DATA FROM SATELLITE SOURCES

### 1.1.1 Envisat

1.1.1.1 MIPAS
The Reading group will be assimilating MIPAS data starting with L2 but then moving onto L1 data (with code developed by Edinburgh and Oxford). They receive the data on CD from ESA but can also access the BADC archive.

The RAL group require L1 data so that they can devise and apply improved algorithms for higher level products to the DARC (i.e. they will also want to compare the processed L2 data to their own algorithm data). Only limb data is required from MIPAS. They are looking at retrieving water vapour and ozone fields to the lowest possible altitudes. Data volumes are around 4.2GB/day, some of which is received via the RAL receiver, made available by the BADC.

The Oxford group are developing code to assimilate MIPAS data. This have begun with retrieved profiles which they will characterise for other DARC groups (Reading in particular). They will then move onto assimilation of raw radiances. They are receiving data by CD-Rom, FTP and from the RAL DDS receiver (via the BADC).

The Leicester group will require all of the entire mission of all MIPAS L1B spectra and MIPAS L2 trace gases products. The data volumes are as above and they are receiving the data from a variety of sources including CD-Rom, FTP, Exabyte Eliant tapes and the RAL DDS receiver (via the BADC).

The Edinburgh group may be interested in using some MIPAS data.

1.1.1.2 SCIAMACHY
The Reading group will be assimilating SCIAMACHY data, received on CD but can also available via the BADC archive.

The RAL group require L1 nadir data to retrieve ozone, water vapour jointly with aerosol from nadir observations. They will be modifying the L2 algorithm of ozone height resolves and also looking at water vapour, and chemical species. Data volumes are around 2.5GB/day and some of this is received via the RAL receiver, made available by the BADC.

The Leicester group require L2 products and maybe L1 data. The data volumes will be as above and their data sources will be as for MIPAS.

The Edinburgh group may be interested in using some SCIAMACHY data.

1.1.1.3 AATSR
The RAL group require L1 radiances (brightness temperature in visible and reflectance in the IR) at the 1 km x 1 km resolution. This are used to characterise cloud properties with SCIAMACHY chemical constituent data. Data

volumes are around 9.2GB/day for the entire dataset. The RAL group receives some of the data on CD direct from ESA as part of the UK PAC (Cal/Val).

The Leicester group will receive the AATSR L2 (spatially averaged (AST) and gridded (GST)) datasets for the entire mission (as part of the validation programme). Some specific L1 scenes are also required. Data is received on CD in standard Envisat format for the entire mission.

The Edinburgh group may be interested in using some AATSR data.

1.1.1.4 GOMOS

The Reading group will use temperature, ozone and water vapour from GOMOS in assimilation experiments. These will eventually be combined with the MIPAS and SCIAMACHY assimilation schemes.

The Leicester group will receive some GOMOS L2 data. Volumes should be around 200MB/day for L2 so this should be transferable via FTP or CD without a problem.

The Edinburgh group may be interested in using some GOMOS data.

1.1.1.5 MERIS

The level 1 MERIS data will be used to construct a TOA radiance gridded dataset, and it and the level 2 MERIS data will be provided to the CWVC programme as correlative datasets. The BADC will hold this archive of MERIS data.

### 1.1.2 Eos-AURA

1.1.2.1 MLS

The Edinburgh group will receive the Eos-AURA UARS data as part of their involvement in the Cal/Val. They have their own source for the data although the BADC might want to get a copy or mirror of the data. Raw radiances of the MLS data will be 4GB/day and will be received on DVD.

1.1.2.2 HiRDLS

The BADC will be the UK designated archive for academic use of HiRDLS data. We will usually receive the data via the electronic network (either via FTP or the Unidata Local Data Manager) and will create an archive that is accessible to academic scientists (and possibly other users). The volumes of the different levels of HIRDLS data (taken from the Science Data Management Plan) are:

| Level type | Data volume |
|---|---|
| Level 0 | 540 Mbyte/day |
| Level 1 | 864 Mbyte/day |
| Level 2 | 87 Mbyte/day |
| Level 3 | 60 Mbyte/day |
| Total | 1551 Mbyte/day |

The current expected launch date for HiRDLS is mid-2004.

The Oxford group will aim to get at least 2 years of HiRDLS data to produce work on assimilation of profiles, radiances and then singular vectors (following on from their ISAMS and MIPAS work). This data will be available from the BADC.

The Edinburgh group will want access to HiRDLS data for assimilation and possibly other data streams from Eos-AURA.

### 1.1.3 ERS-2

1.1.3.1 ATSR-2

ATSR-2 data is available from the RAL ATSR-2 group. This is available on request and is currently extracted on tape for individual requests.

1.1.3.2 GOME

The RAL Earth Observation group receive GOME data on CD directly from ESA directly. This is relatively low in volume compared to Envisat instruments. The BADC also holds a GOME dataset (~5GB in total) from 1995-1999, documented at:

http://badc.nerc.ac.uk/data/gome

The RAL group has plans to re-process the GOME archive in 2004.

### 1.1.4 UARS

1.1.4.1 MLS

The BADC holds the MLS data from the UARS satellite, it can also be accessed from Goddard Space Flight Center (GSFC) Distributed Active Archive Center (DAAC) in the US. This data is made available to Reading and other DARC groups via the usual BADC archive, documented at:

http://badc.nerc.ac.uk/data/uars

The Reading group have used L3 MLS data and software provided by the BADC for some initial assimilation experiments.

The Cambridge group have been using MLS data from GSFC DAAC for testing their assimilation scheme. At the time of writing there had not been a final decision on which other datasets to include although GOME and MIPAS (L2) are possibilities.

1.1.4.2 HALOE

The BADC holds HALOE data at level 2 (uninterpolated profiles at measurement locations), version 19 and also HALOE data at level 3A and version 19 (HALOE L3). The data volumes are small compared to other EO datasets.

The Reading group are looking at HALOE data and comparing water vapour in the stratosphere with that in the ECMWF ERA-40 dataset.

1.1.4.3 ISAMS

The BADC holds ISAMS data at level 2 (uninterpolated profiles at measurement locations) and version 8 and also ISAMS data at level 3A and version 10 (ISAMS L3), the latter is publicly available.

The Oxford group are characterising ISAMS radiometer data received on CD or obtained from the BADC archive. The volumes are relatively small compared to other EO datasets, spanning from 1991 to 1992.

### 1.1.5 MSG

The BADC has an agreement to act as an archive for Meteosat Second Generation (MSG) imagery data. At the time of writing the BADC is investigating an FTP link to the data source. Discussions with the Met Office have also been held regarding format conversion to HDF.

## 1.2 MET OFFICE DATASETS

### 1.2.1 Unified model output (used to drive assimilation).

Met Office Unified Model (UM) datasets will be primarily used by the Reading DARC group for assimilation experiments. Output from previous model runs is used to initialise a new assimilation run. These data will mainly be used and stored on the Met Office computing architecture where the assimilation experiments are run.

Other groups including those at Cambridge and RAL will also use Met Office Unified Model output that is archived routinely at the BADC.

### 1.2.2 Observational data from the MetDB.

A full assimilation and NWP experiment or simulation requires the input of observations into the Observation Processing System (OPS). DARC scientists running assimilations on the Met Office computing system should have access to observations held in the MetDB.

## 1.3 ECMWF DATASETS

### 1.3.1 Model output data from the ECMWF Integrated Forecasting System (IFS)

Model output will be required to run the PrepIFS assimilation scheme at the ECMWF. This is to be used by the collaborating Edinburgh and Reading groups to run a re-analysis of years during the 1990s. An improved set of UARS MLS data will be assimilated along as a parallel version of the ERA-40 data streams. This data will sit on the ECMWF computing systems where the re-analysis will be run.

Other groups including those at Cambridge and RAL will also use ECMWF Operational (and Re-analysis) Model output that is archived routinely at the BADC.

### 1.3.2 ECMWF Observations

Along with the model output from the ECMWF IFS the re-analysis simulations will require observations for the assimilation process. These have been retrieved by the BADC from the ECMWF ODB (observations database) that provides the input observations for ERA-40. This data is currently held in the original BUFR format at the BADC.

## 1.4 ATMOSPHERIC CHEMISTRY DATASETS

### 1.4.1 Cambridge

The Cambridge group has provided a 200GB dataset of constituent observations transformed to flow tracking co-ordinates for use in DARC assimilations. This data was provided in a proprietary format.

## 1.5 OCEAN DATASETS

### 1.5.1 ESSC Ocean Data

ESSC has a number of datasets that it has used in assimilation-related research. Most of these are supplied be established international links, such as the altimetry data which is a combined product from a number of sources including TOPEX, ERS-1 and ERS-2. They also use profile data from US databases. A high volume dataset is the OCCAM data which is estimated at 200GB in total. This will be served by a DODS server to convert from HDF format. The ESSC has access to some ECMWF and Met Office ocean data via established links. They may be interested in Envisat data later in the programme.

**Data flows generated by the DARC**

The main data flows created by the DARC are the following:

## 1.1 PROCESSED DATASETS FROM EARTH OBSERVATION DATA

### 1.1.1 Envisat

1.1.1.1 MIPAS

The Oxford group will be characterising MIPAS retrievals and making them (or the processing code) available to Reading for assimilation experiments. The DARC and BADC should consider whether this processed MIPAS data is of use to the wider community and whether a long-term archive should be established.

The RAL group will also be working on improving algorithms for MIPAS L2 data. When a final product is being produced this may be of interest to the research community in which case the data should be offered to the BADC for archival. Documentation and software should also be provided.

1.1.1.2 SCIAMACHY

The RAL group plans to modify the L2 algorithm of ozone height resolves for SCIAMACHY data. The same archival considerations should be made as for their MIPAS L2 data.

1.1.1.3 AATSR

The RAL group will characterise cloud properties using SCIAMACHY chemical constituent data. The same archival considerations should be made as for their MIPAS L2 data.

### 1.1.2 Eos-AURA

1.1.2.1 HiRDLS

The BADC intends to run processing code (provided by Oxford University) to process L0 to L1 data which will then be used for assimilation by the DARC. This will be available via the usual BADC archive.

## 1.2 ASSIMILATION EXPERIMENTS USING MET OFFICE SYSTEMS

1.2.1 Assimilation experiment outputs using the Met Office Unified Model.

Experiments run by the Reading group on the Met Office computing systems will produce output of potential interest to the community. The volumes are likely to be in order of <100GB and the data should be provided in one of the preferred data formats for model data (i.e. NetCDF or GRIB).

## 1.3 ASSIMILATION EXPERIMENTS USING ECMWF SYSTEMS

### 1.3.1 Assimilation experiment outputs using the ECMWF IFS.

Collaborations between the Reading and Edinburgh groups will create a Re-analysis dataset for at least a year during the 1990's. This data could potentially be of interest to the ECMWF as an improvement to the currently available stream from ERA-40. However, the data could also be useful to the wider community and may be lodged at the BADC.

Once the dataset has been validated the DARC and BADC will decide whether to archive a copy. The decision will depend on the expected use of the dataset by the atmospheric science community and the resource implications in creating and maintaining the archive. The volume of this dataset is likely to be in the order of a few hundred Gigabytes and should be provided in NetCDF or GRIB format.

## 1.4 ATMOSPHERIC CHEMISTRY DATASETS

### 1.4.1 Cambridge

As mentioned in section 4, the Cambridge group has provided a 200GB dataset of constituent observations transformed to flow tracking co-ordinates for use in DARC. This dataset is archived and documented at Cambridge with David Lary as the main contact for those interested in using it.

## Appendix 2 – DARC Data Protocol

The scope of this Data Protocol is intended to be datasets generated by DARC research. This may include research using both instruments and model datasets.

The aims of the Data Protocol are:

- to encourage rapid dissemination of scientific results from DARC;
- to protect the rights of the individual scientists producing data in the DARC;
- to ensure that all involved researchers are treated equitably;
- to ensure the quality of the data in the DARC data archive.

These aims conflict at times, and it is hoped that the provisions of this protocol will resolve these conflicts fairly. It is recognised that this cannot always be achieved to everyone's complete satisfaction. There are likely to be cases where individual interests clash with those of DARC. Therefore to try to meet these aims, all PIs, Co-Is and Instrument Scientists involved in DARC must agree to abide by the following conditions:

1. DARC data produced under the auspices of a NERC Thematic Programme will be subject to the Data Protocol of that programme.
2. DARC datasets that fall under the auspices of NERC will be made available to all relevant participants, and to those participants only, during a *restricted access period* ending one year after the concerned project end date, after which data and model results will be released to the public domain. At an investigator’s request, access may be extended to personally authorised collaborators.
3. The designated data centre for DARC-derived data is the BADC although some data will be archived within other DARC institutes.
4. The longevity of validated raw data must be ensured in a secure archive, possibly, but not necessarily at the BADC. Details pertaining to the validated raw data (i.e. metadata), whether or not archived at BADC, must be sent to the BADC, as well as information on how to access the data.
5. Preliminary datasets will normally be made available to other DARC collaborators involved in a particular project as soon as possible. Any corrections or amendments to the preliminary data should be announced as soon as possible.
6. Validated processed data (i.e. datasets in their final form) must be archived at the designated DARC data centre with the required metadata. The data providers and the BADC should arrange a data submission date. Archival should take place no later than this agreed date.
7. Data produced outside of the auspices of NERC programmes and projects should be provided to the BADC where possible. The data owner may stipulate data access restrictions on such occasions.
8. Data submitted to the BADC must be in the data formats agreed between DARC investigators and the BADC (*NetCDF* (following the CF convention) preferred, *NASA Ames*, *HDF* and *GRIB* also accepted). All agreed metadata describing data, models and model results, regardless of their archival location, must be supplied to the BADC. Formats and metadata are documented at BADC.
9. It is the responsibility of each Investigator to ensure that the data used in publications are the best available at that time.
10. If measurements or model results from other DARC-related research is used in a publication by a DARC project participant, joint authorship must be offered. This does not necessarily have to be accepted, particularly in cases where due credit and acknowledgement can be given in other, possibly more appropriate, ways.
11. Datasets used by the DARC that are covered by agreements with Weather Forecasting Agencies (Met Office and ECMWF) that are archived will be subject to the access restrictions of those agreements.
12. Whilst the data are restricted from the public domain (see Clause 2), each investigator has the right to refuse to allow his/her work, whether measurement or calculation, to be used in a publication or presentation prior to the Principal Investigator’s own publication of that work.
13. Whilst the data are restricted from the public domain, no data should be transferred to a third party without the originator’s consent.
14. In the event of dispute, a scientific steering committee will be set up consisting of the DARC Principal Investigators of the specific campaign or project, who will make a final decision.

## Appendix 3 - Outline of portable data assimilation

Figure 1 presents an outline of the key components and data transfers between the data assimilation and NWP systems at the Met Office.

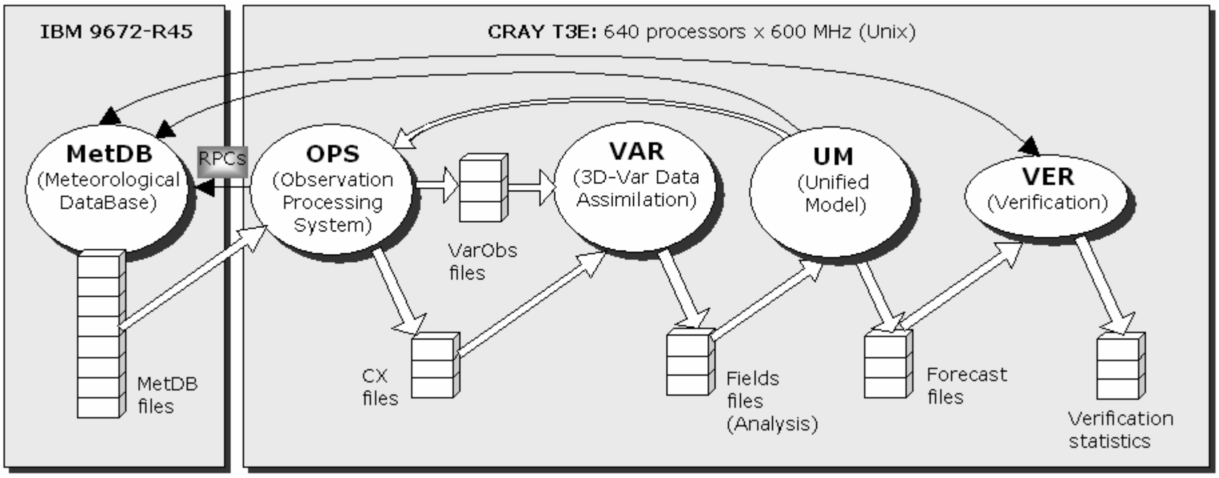

**Figure 1.** The components of the current Met Office data assimilation system. The four main components are the MetDB, OPS, VAR and the UM. The additional component, VER, provides useful information on the accuracy, skill and quality of a forecast.

Porting is a non-trivial task and in order to benefit the academic community greater resources are required than initially anticipated. Since the DARC began there have been a number of developments to augment its commitment to delivering portable data assimilation. The main emphasis has been on the PARADISE (Portable Assimilation of Remotely Accessible Data In Support of E-science) proposal which was unsuccessful and the recent Data Assimilation Infrastructure Working Group.

This group has presented the issue as a Grand Challenge at the NCAS level with supportive papers put forward by the following groups:

- DARC (Reading).
- School of the Environment (Leeds).
- UWERN.
- JCMM (Reading).
- Met Office (Exeter).
- BADC (RAL).

The Met Office has expressed the will to provide its assimilation system to the community but is currently limited in resources. This effectively means that the human resources for undertaking this task must be found outside of the Met Office. The potential benefit to the Met Office is clearly in the development of assimilation techniques and addition of research assimilation streams into its operational system.

# Appendix 4 - Glossary of terms and acronyms used in this document

| | |
|---|---|
| 3DVar | 3D-Variational Data Assimilation |
| 4DVar | 4D-Variational Data Assimilation |
| ATSR | Along Track Scanning Radiometer (ATSR-1 was on ERS-1 and ATSR on ERS-2) |
| AATSR | Advanced Along Track Scanning Radiometer (on Envisat) |
| AOPP | Atmospheric, Oceanographic and Planetary Physics, Oxford University |
| ASAR | Advanced Synthetic Aperture Radar (on Envisat) |
| BADC | British Atmospheric Data Centre |
| Cal/Val | Calibration/Validation |
| DARC | Data Assimilation Research Centre |
| DORIS | Doppler Orbitography and Radiopositioning Integrated by Satellite (on Envisat) |
| ECMWF | European Centre for Medium-Range Weather Forecasts |
| Envisat | ESA satellite launched in 2002. Sun synchronous orbit with a mean altitude of 800 km and a descending node mean local solar time of 10:00 am. The orbit has a 35-day repeat cycle (as for the ERS-2 mission). One-day and three-day orbit subcycles provide a global sampling of the Earth matched to the requirement of assimilation into global meteorological and climate models. |
| EOS-Aura | Earth Observing System (EOS) Aura is a NASA mission to study the Earth's ozone, air quality and climate. |
| ERA-40 | ECMWF 40 year global re-analysis project (1957-2001). |
| ERS-1 | European Remote Sensing Satellite 1 |
| ERS-2 | European Remote Sensing Satellite 2 |
| ESSC | Earth System Science Centre, Reading University |
| Forward-model | An algorithm used to derive a retrieval from a radiance. |
| GMES | Global Monitoring of Environmental Systems |
| GODIVA | Grid for Ocean Diagnostics, Interactive Visualisation and Analysis |
| GOME | Global Ozone Monitoring Experiment |
| GOMOS | Global Ozone Monitoring by Occultation of Stars (on Envisat) |
| HALOE | HALogen Occultation Experiment |
| HiRDLS | High Resolution Dynamic Limb Sounder |
| ISAMS | Improved Stratospheric And Mesospheric Sounder |
| JCMM | Joint Centre for Mesoscale Meteorology |
| LAS | Live Access Server |
| LRR | Laser Retro Reflector (on Envisat) |
| MAPSCORE | Mapping of Polar Stratospheric Clouds and Ozone Levels Relevant to the Region of Europe |
| MERIS | Medium Resolution Imaging Spectrometer (on Envisat) |
| MIPAS | Michelson Interferometer Passive Atmospheric Sounder (on Envisat) |
| MLS | Microwave Limb Sounder |
| MSG | Meteosat Second Generation |
| MWR | Microwave Radiometer (on Envisat) |
| NWP | Numerical Weather Prediction |
| OCCAM | Ocean Circulation and Climate Advanced Modelling Project |
| OI | Optimal Interpolation |
| RA2 | Radar Altimeter 2 (on Envisat) |
| Radiance | Radiance (L) is the power emitted (dW) per unit of the solid angle (dW) and per unit of the projected surface (ds cosq) of an extended widespread source in a given direction (q).<br>L = d2W / (dW. ds. cosq) (in W.Sr-1. m-2) |
| RAL | Rutherford Appleton Laboratory |
| Retrieval | The instrument team’s calculated value of a geophysical quantity derived from the original radiance data from an instrument. |
| SCIAMACHY | Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (on Envisat) |
| SWIFT | Stratospheric Wind Interferometer for Transport Studies (ESA/Japan/Canada joint project). |
| UARS | Upper Atmosphere Research Satellite |
| Met Office | UK Meteorological Office |
| UM | Met Office Unified Model |