

# **Aerosol Properties, Processes And Influences on the Earth's climate (APPRAISE) Data Management Plan**

## **1. Introduction**

NERC's designated data centre for APPRAISE is the British Atmospheric Data Centre (BADC). The purpose of the APPRAISE data management plan is to set up a coherent approach to data issues during the programme. Its objective is to ensure that,

- Appropriate data support is provided to the scientists within the programme.
- Data are made available to collaborators within the programme in a timely fashion.
- Distribution conditions protect the individuals' rights to publish their own work.
- Potentially scientifically valuable data are kept for the long-term.
- A high quality documented APPRAISE archive is created.
- Data and documents are eventually distributed to the broader scientific community.

A summary of the suggested conditions for data access and publication are given in the data protocol at Annex 1. All applicants to the data will be asked to abide by this protocol.

## **2. Types of data generated by APPRAISE**

Annex 2 shows the information collected on data deliverables from the project plans. The next stage will be to visit the PIs of each project to get a clearer picture of the data expected. APPRAISE projects will be involved in new measurements, the gathering of existing data, data synthesis and model output generation.

For new measurements, raw data in the form that they have on acquisition will in general not be archived at the data centre, but it is each PI's responsibility to ensure that they are stored safely with the relevant processing software or, alternatively, with documentation on retrieval algorithms. Details of the raw data in existence should be documented at the BADC. Processed data, i.e. observation data that have been subject to some treatment or formatting, or data derived from these, will be archived at the data centre and made available to the APPRAISE community. Existing datasets are likely to be already subject to a data protocol, but if not, these datasets can be included in the APPRAISE dataset.

Production of synthesised data consists of the compilation and harmonisation of existing observation data from a variety of origins. This usually involves some kind of modelling and the resulting datasets are more likely to be superseded by new versions than are datasets from one single source. For these two reasons, archival and curation of synthesised and model data present similarities, and these data will be subject to the BADC policy and guidelines for archiving simulation data which can be found in Annex 3. Synthesised data and selected model output, together with the required metadata, will be archived at the BADC. Selection of model output worth archiving will be done jointly by the project PIs and the BADC, based on the selection criteria listed in the policy.

## **3. Data archive at the BADC**

A website will be set up for APPRAISE at the BADC, which will be the gateway to all APPRAISE data: <http://badc.nerc.ac.uk/data/appraise/>. This page will include links to all relevant documentation and external sites and will be updated as the programme unfolds.

For the duration of the Programme, the BADC will

- liaise with the APPRAISE researchers to get updates on their data deliverables and needs, and develop the present data management plan accordingly;
- provide support to APPRAISE investigators on issues related to format, metadata and submission;
- answer data related queries;
- maintain and update the APPRAISE archive, data portal and uploader;

- integrate APPRAISE data into the NERC Data Grid (NDG);
- monitor access applications;
- release the data to the public in due course.

In addition to the observation and model data, the APPRAISE archive will hold all documentation that would be too extensive to be archived as metadata within the data files themselves (in particular, source codes of models used to generate some model output).

The data held at the BADC will be preserved for the long-term unless the dataset review (as discussed in annex 3) concludes that the data have been superseded by new or better versions. If new versions are submitted, the option of keeping the old ones will be envisaged. In this case, it will be made very clear which one is the most recent version.

Data will be backed up at regular intervals and duplicates will be saved on tape.

#### **4. Dataset catalogue and metadata search**

An entry will be created for APPRAISE in the BADC Data Catalogue. As data populate the archive, the catalogue entry will be completed with input about the instruments, the models, etc. The underlying metadata scheme involves the production of MOLES (Metadata Objects for Links in Environmental Science) records. The records will be integrated into the NERC Datagrid, which allows the search of metadata pertaining to datasets throughout the network of all NERC Data centres.

At a later stage, part of the MOLES records will be completed by information retrieved automatically from the files, which assumes (and underlines the importance of) properly formatted data.

#### **5. Formats**

APPRAISE data will be formatted in NetCDF or NASA-Ames. NetCDF is a widely used in the geophysical community and a range of software exists to produce, read and handle NetCDF files. If preferred, NASA Ames may be used for the storage of less voluminous sets of measurements. These two formats will underlie the BADC catalogue search engine and the NERC Data Grid, so that APPRAISE datasets will be integrated swiftly into these two instruments, provided that files do include CF-compliant metadata in the formatted fields (see the next section).

Documentation, tools, templates, examples and help on these formats are provided at the following web pages.

- NetCDF format (binary) — <http://badc.nerc.ac.uk/help/formats/netcdf/>
- NASA Ames format (ASCII) — <http://badc.nerc.ac.uk/help/formats/NASA-Ames/>

The BADC will provide support to data suppliers using NetCDF or NASA Ames.

Raw data (if any), software and images will be archived in their original formats. Text documents should be archived in PDF.

#### **6. Metadata**

Metadata (“data about the data”) contain the necessary information to find (“discovery” metadata), read, understand, interpret and use the data.

The Climate and Forecast (CF) Metadata Convention — developed for NetCDF but applicable to any geophysical dataset — will be used to select and format the metadata related to data recorded in the above types of formatted files. For example, as far as possible, CF standard names should be used to name recorded variables, even in NASA Ames files. The set of CF standard names is an evolving nomenclature, open to new soundly founded proposals. Proposals can be submitted to the CF Convention through the CF mailing

list — and soon via the web. If requested, the BADC will provide advice to researchers in forming new name candidates.

The metadata should be integrated into NetCDF files through the use of local and global attributes. The NASA Ames (ASCII) data files include a header which contains, in a predefined display, basic information on the data recorded in the file, so that these metadata are inseparable from the data to which they pertain. Both file types provide room for comments, which may include any information that cannot be provided in the formatted fields, although this information, very valuable for the user, will be less useful in terms of automated search.

The CF Metadata Convention is available from <http://www.cf-conventions.org/> and the BADC provides additional guidelines on CF at [http://badc.nerc.ac.uk/help/formats/netcdf/index\\_cf.html](http://badc.nerc.ac.uk/help/formats/netcdf/index_cf.html)

Metadata should be as specific, explicit, accurate and complete as possible. They should be formulated in a transparent way, avoiding unexplained assumptions and implicit references to unavailable or undocumented conventions.

The checklist provided by BADC at <http://badc.nerc.ac.uk/help/metadata/#Elem> includes the following main metadata elements.

- Information on the (physical or theoretical) experiment.  
*Date when experiment or model simulation started. Site or trajectory bounding box or domain limits. Platform, instrumentation. Model name.*
- Information on the data originator(s).  
*Names, affiliation, contact address including e-mail, telephone number. Research programme name, research project code.*
- Information on the independent variables (in geophysical datasets, usually a spatio-temporal grid).  
*Names, units, domain of definition of independent variables. Interval values when appropriate.*
- Information on the data.  
*Version number. Date of last revision. Processing level (nature of raw data, derivation method). Nature, name, units, scaling factors, accuracy of dependent variables.*
- Information on the format.  
*Type of format + reference of format documentation. File structure. Number of lines in file header if any. Record structure.*
- Any additional relevant information, such as instrument description and specifications, essential model features and configuration, conditions in which the data were collected, algorithms used to derive the recorded data from the raw data, reference publications, etc.

CF standards exist for part of the above and should be applied. NASA Ames provides formatting rules for some of the elements above. Any information that cannot be integrated into the data files as local attributes (NetCDF) or in the file header top section (NASA Ames) should either be inserted in the files as global attributes (NetCDF) or comments (NASA Ames), or (if substantially large) be provided in separate supporting documentation. This additional documentation may include collection methods, algorithms, model parameterisations, references, advice to the users, plots, pictures, etc. and will be archived alongside the data. Texts should be submitted as PDF files. Source codes can be archived if they help the understanding of the archived model output; in this case, it is preferable to archive also a set of standard model input.

## 7. Data file names

Data file names will follow the BADC file name convention documented at [http://badc.nerc.ac.uk/help/file\\_naming.html](http://badc.nerc.ac.uk/help/file_naming.html)

File names are composed of three (optionally four) fields separated by underscore signs, and an extension separated from the rest by a dot. Each field may only include lower case letters, digits and the hyphen sign. The file name template is

*instr\_loc\_YYYYMMDD[hh[mm[ss]]][\_extra].ext*

where fields inside square brackets are optional and where

*instr* is a standard name for one of the following

- an instrument — sometimes denoted by the quantity it measures (e.g. “uea-peroxides”);
- a set of instruments (e.g. “core-cloud-phy” for the set of FAAM core instruments measuring cloud physics data);
- a production method or its result (e.g. “syn-o3” for ozone data synthesis);
- a model (e.g. “mcm-short” for the Master Chemical Mechanism of short lived species).

When the instruments or model are operated by one group, it should be composed of two parts separated by a hyphen, with the first part being the name of the university or institution owning the instrument or model (e.g. “uea-doas” for the Differential Optical Absorption Spectrometer operated at the University of East Anglia).

*loc* is a standard name for one of the following

- a location or a region (where the data was collected, where the modelled phenomena take place, where the computed trajectory starts or ends, etc.); if the model does not reproduce the conditions at a given particular place, it can be the type of landscape which is simulated; if the data cover the globe, *loc* can be set to “globe”; if none of this applies, it can be the location where the model was run;
- an itinerant platform (such as a van, a balloon, a ship, an aircraft, a satellite, etc.).

*YYYYMMDD* is one of the following

- the date when the first data record was collected (for observation);
- if the data are made of monthly or yearly averages, *DD* [resp. *MM*] can be replaced by the first or last day in the month [resp. in the year];
- the start date of the recorded scenario (for a realistic time-dependent simulation);
- the date when the first recorded model result was computed;
- if none of the above applies, the date when the dataset was produced.

*hh*, *mm* and *ss* (hours, minutes, seconds) can be added if necessary, so that this field can have one of the forms *YYYYMMDD*, *YYYYMMDDhh*, *YYYYMMDDhhmm* or *YYYYMMDDhhmmss*.

*extra* is an optional field with a free content.

*ext* is the format extension; for example,

- *ext* = na for NASA Ames,
- *ext* = nc for NetCDF.

Accepted values for the first two fields and for the extension are listed at [http://badc.nerc.ac.uk/cgi-bin/filespec\\_doc?id=INSTRUMENTAL&hfile=1](http://badc.nerc.ac.uk/cgi-bin/filespec_doc?id=INSTRUMENTAL&hfile=1)

Data providers should submit new values of *instr* and *loc* to the BADC before uploading their data, in order to ensure that their files are not rejected by the web uploader described in the next section.

## 8. Data submission and ingestion

If needed by other APPRAISE groups, preliminary data should be made available to them as soon as possible, if possible via the BADC. Processed data and model results should be supplied to the BADC as soon as they are ready, and no later than the project end date. Individual project archives should be complete by the end date of the project. The current submission schedule is given by the information recorded in the Expected Delivery Date column in Annex 2.

Instructions on data submission will be made available from the BADC at <http://badc.nerc.ac.uk/data/appraise/>

Before submitting their data to the BADC, investigators should,

1. Form file names as described in Section 7; if adequate field names cannot be found in the BADC lists of standard instrument, location and extension names; contact the BADC with suggestions or requests for new names.
2. Choose one of the accepted formats described in Section 5 and format their data accordingly.
3. Ensure that all required metadata (see Section 6) are included in the data files and are formulated in a clear and unambiguous way.
4. Check sample files against the online tools provided by the BADC at
  - <http://titania.badc.rl.ac.uk/cgi-bin/cf-checker.pl> for NetCDF files;
  - [http://badc.nerc.ac.uk/cgi-bin/dataex\\_file.cgi.pl](http://badc.nerc.ac.uk/cgi-bin/dataex_file.cgi.pl) for NASA Ames files.

Data are uploaded to the BADC APPRAISE incoming area, where appropriate sub-folders will also be made. To access this area, data suppliers have to register for APPRAISE, and this will also give access to all APPRAISE data.

Data should be uploaded through the APPRAISE web uploader which will be set up for that purpose. In three steps (selection of a folder, selection of a format, selection of a file on the home computer), this online tool will enable participants to upload any data file or supporting documentation. Directions will be provided online from the BADC APPRAISE web page. Alternatively, if the number of files to be submitted is very large, data and metadata can be uploaded via FTP by connecting to the BADC FTP server at <ftp.badc.rl.ac.uk>

Format and filename checks will be performed on files submitted to the BADC via the web file uploader, to spot errors of file name or format. Files with wrong names or format errors will be rejected. Such checks will be performed manually for files submitted by FTP.

An e-mail will be sent automatically to a member of the BADC staff every time new files appear in the incoming area. The files will then be transferred from the BADC incoming area to the APPRAISE archive. Once the structure of the archive is fixed, this ingestion may be automated based on the file names and the incoming folders.

## **9. Data access and conditions of use**

Access to data will be restricted to APPRAISE participants until one year after the end of each project, after which they will be made public. Whilst the data are restricted from the public domain, a password protected system will be used whereby participants will be prompted to agree with the APPRAISE Data Protocol (see Annex 1) in order to access the archive. Applications for access to restricted data will be forwarded to the relevant principal investigator, who will decide whether access can be granted. To speed up this process, the BADC will maintain lists of participants in each project who can be automatically accepted.

After release of the data to the public domain, anonymous users will be requested to contact the relevant data providers before using the data and to acknowledge the APPRAISE programme and the data suppliers in any publication using the data. These rules should be in the comments of each data file, will be on the BADC information pages and will be in readme files in the directories.

## **10. Other services**

An APPRAISE online workspace could be made available if this would be useful to the programme. The workspace would be visible and accessible only by APPRAISE participants and would be intended to ease the exchange of ideas, documents and preliminary data between the members of the programme. It is not an alternative to data submission to the BADC but must rather be considered as a discussion forum or an area for data, reports and papers in the validation and preparation phase.

Third-party data required by projects and held at the BADC, such as ECMWF and Met Office data sets, will be made available to the participants, subject to current access conditions. If required, BADC will endeavour to retrieve data sets from other sources at no cost or will negotiate their acquisition at the best possible cost.

## Annex 1

### APPRAISE Data Protocol

The aims of the Data Protocol are,

- to encourage rapid dissemination of scientific results from the APPRAISE programme;
- to protect the rights of the individual scientists funded by APPRAISE;
- to have all the involved researchers treated equitably;
- to ensure the quality of the data in the APPRAISE data archive.

These aims conflict at times, and it is hoped that the provisions of the protocol resolve these conflicts fairly. It is recognised that this cannot always be achieved to everyone's complete satisfaction; there are bound to be cases where individual interests clash with those of the APPRAISE programme. Therefore, to try to meet these aims, all PIs involved in APPRAISE, in accordance with and on behalf of their co-investigators, must agree to abide by the following conditions:

1. Data and model results of interest to APPRAISE groups that will be produced during the programme will be made available to all APPRAISE participants, and APPRAISE 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 a principal investigator's request, access may be extended to personally authorised collaborators.
2. The designated APPRAISE data centre is the BADC.
3. When relevant, preliminary data must be made available to APPRAISE collaborators as soon as possible. Any corrections or amendments to the preliminary data should be announced as soon as possible.
4. All validated processed data (i.e. data sets in their final form), as well as quality-checked raw data that will have been recognised of general interest to the community, will be archived at the BADC. Archival must take place no later than the end of the concerned project.
5. If an error in the data is signalled to the BADC after archival, the updated version provided by the originator of the data will be archived alongside the old version or will replace the old version, depending on the nature of the error and in agreement with the data provider. If the error is detected by a user, the originator of the data will be consulted and only changes provided by him/her will be archived, following the same procedure as above. Both errors and updates will be reported and documented in the metadata attached to the respective data files.
6. Data submitted to the BADC must be in the data format agreed between APPRAISE principal investigators and the BADC (namely NetCDF and NASA Ames).
7. All agreed metadata describing data, models and model results, regardless of their archival location, must be supplied to BADC.
8. It is each principal investigator's responsibility to ensure that the data used in publications are the best available at that time.
9. If measurements or model results from other APPRAISE research groups are used in a publication by an APPRAISE participant, during or after the programme, 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, more appropriate ways.
10. Whilst the data are restricted from the public domain (see Clause 1), each principal 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 PI's own publication of that work.
11. Whilst the data are restricted from the public domain, no data should be transferred to a third party without the originator's consent.
12. In the event of dispute, the final decision rests with the APPRAISE Programme Advisory Group.
13. At any time, APPRAISE data distributed via the BADC may not be used for commercial purposes. Requests for commercial use must be addressed to the owner of the IPRs, whether NERC or the data originator(s).

## Annex 2 Details of Data within Projects

### 1. ACES

WP		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
1.1	Emissions studies in smog chambers	Prof C N Hewitt, University of Lancaster	Normalised BVOC emissions per unit leaf area by GC-MS and PTR-MS		April 2008	1.0			
1.2	Single precursors for Secondary Organic Aerosols in Smog chambers	Dr G B McFiggans, University of Manchester	Aerosol size distributions from DMOPS and optical particle spectrometers Water uptake from HTDMA Component classifications from Q-AMS CCN measurements from DMT		May 2008	1.0			
		Dr P S Monks, University of Leicester	VOC measurements from CIR-TOF-MS and HT-TOF-MS		May 2008	1.0			
		Prof A C Lewis, University of York	Aerosol composition from GCxGC-MS and LC-MS/MS	APPRAISE core post at York	Dec 2008	1.0			
1.3	Organic aerosols from compound ensembles in smog chambers	Dr G B McFiggans, University of Manchester Prof A C Lewis, University of York Dr P S Monks, University of Leicester	Normalised BVOC emissions per unit leaf area by GC-FID, GC-MS and PTR-MS		Jan 2009	2.0			
2.1	In-canopy flux characterisation add-on to Danum program	Prof A C Lewis, University of York Prof C N Hewitt, University of Lancaster	Ambient organic aerosol identification		Nov 2008	2.0	OP3 at Danum Valley in 2008 BVOC emission data	University of Manchester	OP3-Danum-08 to receive all ACES data
		Dr G B McFiggans, University of Manchester	Bioaerosol detection from a WIBS-2 UV fluorescence		Nov 2008	2.0	C4 and C15 compound emissions from Dr Geron of the US EPA		
		Dr E Nemitz, CEH Oxford	Bio-aerosol identification and diversity of microbial communities using DNA and microscopic analysis Abundance and distribution of IN		Nov 2008	2.0			

2.2	In-canopy VOC precursor and BOA flux characterisation	Prof A C Lewis, University of York Dr E Nemitz, CEH Oxford	Concentration profiles of primary BVOCs and oxidation products from automated profiler system PTRMS system In-canopy size distribution profiles from GRIMM optical particle counters and an aerosol spectrometer Chemical characterisation by AMS Aerosol flux dynamics from aerosol flux systems Derivation of vertical distribution of sources and sinks		Jan 2009	10.0	Turbulence measurements from OP3		
2.3	VOC precursor and BOA flux characterisation above an oil palm plantation	Dr E Nemitz, CEH Oxford	VOC emissions, aerosol deposition rates, aerosol production fluxes from PTR-MS, CPC, eddy-covariance AMS system.		Sep 2008	10.0			
3.4	Master Chemical Mechanism (MCM) modelling - Mechanism reduction	Dr M E Jenkin, Imperial College London	Possible archival of model fields used in publications on the major precursors to SOA formation and major SOA components			100.0			
4.1	Development of MEGAN BVOC model	Dr P I Palmer, University of Edinburgh Prof C N Hewitt, University of Lancaster	Estimates of SOA precursor emissions from GLOMAP and MEGAN BVOC flux model			30.0	OP3 BVOC fluxes from satellites for Borneo MODIS and MISR over Borneo for AOD, plume height, size.		
4.2	Impact on AOD and PAR	Dr P I Palmer, University of Edinburgh	2 years of data at GAW tower in Danum valley, Malaysia, giving aerosol properties and leaf area index,		Continuous for 2 years from Jan 2008.	50.0	Satellite observations of HCHO		Sunphotometer will be used in AERONET
4.3	Climate feedback between BVOC and cloud radiative properties	Dr P I Palmer, University of Edinburgh					BVOC flux and aerosol properties for Abisco, Sweden Radiosonde data from Tawau in Sabah	Dr Arneeth, Lund University	

212.0

## 2. ICE

WP		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
1.3	Data from aircraft observations	Prof T Choularton, Manchester Prof A Blyth, Leeds Prof H Coe, Manchester	a) Core instruments b) Turbulence probe c) Aerosol Mass Spectrometer and TOF-AMS d) Soot photometer (SP2) and UHSAS, 2D-S e) Cloud Particle Imager f) Cloud Particle Counters and Differential Mobility Analyser g) Filter analysis by IC, SEM and EDAX h) GRIMM Optical particle counter i) Small Ice Detector and SID2	FAAM Cambridge/UFAM Dr P Williams, Manchester Hertfordshire Dr M J Flynn, Manchester  Hertfordshire	Chilbolton flights Feb 2008 and May 2009 Scotland, Pennine and Munich flights July and Nov 2008	200.0	EUCARRI data sharing		
2.1	Chilbolton ground-based observations during FAAM flights	Prof A J Illingworth, Reading Dr R J Hogan, Reading	CAMRa radar 94 GHz cloud radar	Dr C Walden (RCRU) Dr E O'Connor (Reading)	Feb 2008 and May 2009	200.0			
2.2	Long-term Chilbolton measurements	Prof A J Illingworth, Reading Dr R J Hogan, Reading	CT75K lidar ceilometer 1.5 um Doppler lidar 355 nm UV lidar Raman UV lidar CIMEL sun photometer Vertically pointing 3 GHz radar Vertically pointing 35 GHz cloud radar	Dr E O'Connor (Reading) " " Dr J Agnew (RAL) Dr S Ventouras (RAL) Mr D Ladd (RAL) Dr J Nicol (Reading/UFAM)	Nov 2007 - Nov 2009 continuous data	500.0			

3.1	Ground-based field experiments	Prof H Coe, Manchester Prof A Lewis, York University	Dec 2007 and Jan/Feb 2009 IOP data from UFAM mobile aerosol lab at Chilbolton which has a comparable instrument set to the FAAM core and non-core listed at 1.3 plus: a) Sample aerosol characterisation by LC-tandem MS b) Sample analysis by HPLC/H-NMR c) Ice nucleation by INC d) Sonic anemometers at 5 and 10 m e) Radiosonde data	Dr. P Williams, Manchester  APPRAISE core post, York  ISAC, Bologna	Mar 08 and May 09	100.0			
3.2	Closure tests of the Aerosol Diameter Dependent Equilibrium Model (ADDEM)	Dr D Topping, Manchester	Predictions of the subsaturated properties of the aerosol from ADDEM	PDRA2, Manchester	June 07 to Nov 09 continuous	100.0			
3.3	Long-term measurements of aerosols	Prof H Coe, Manchester	Subset of IOP data, including most of the instruments from the UFAM mobile aerosol lab	Dr P Williams, Manchester	Nov 07 to Feb 09 continuous	50.0			
4.1	Laboratory studies of cloud and ice nucleation	Prof P Kaye, Hertfordshire Dr M Gallagher, Manchester	Full AIDA instrument suite for aerosol characterisation during experiments	PDRA2, Manchester New PDRA, Hertfordshire	4 experiments Dec 07, May 08, Feb 09, July 09.	50.0			
4.2	Analysis of Aerosol Interactions and Dynamics in the Atmosphere (AIDA) databases	Prof P Kaye, Hertfordshire Prof T Choularton Dr M Gallagher, Manchester					TDLAS, FTIR, CPI, SID and AMS data from AIDA databases	Institute for Meteorology and Climate Research, at the Forschungszentrum, Karlsruhe, in Germany	
4.3	Ice nucleation using realistic aerosol	Prof P Kaye, Hertfordshire Dr M Gallagher, Manchester	Full AIDA instrument suite for aerosol characterisation during experiments plus TOF-AMS and SP-2 Chemical characterisation by ESEM and EDAX	New PDRA, Hertfordshire	4 experiments Dec 07, May 08, Feb 09, July 09.	100.0			

4.4	New inversion algorithms for small ice discrimination	Dr Z J Ulanowski, Hertfordshire Prof P Kaye, Hertfordshire	a) SID cloud probe data for well-defined ice in AIDA b) A comprehensive database of measurements of the ice nucleation properties of a library of characteristic aerosol-dust types	New PDRA, Hertfordshire	Sept 09	200.0			
5.1	Mountain wave cloud modelling	Prof T Chouarton, Manchester Prof K Carsaw, Leeds	Parcel model runs that support publications		June 10	100.0			
5.2	Chilbolton cloud modelling	Prof T Chouarton, Manchester Prof K Carsaw, Leeds	Large Eddy Cloud Resolving Model, and parcel model simulations of the observed vertical profile of aerosol extinction from the lidar that support publications	J Marsham, Leeds	June 10	100.0			
6.1	Global simulations of ice nuclei	Prof K Carsaw, Leeds	Global Aerosol model output	Leeds APPRAISE core post	June 10	200.0			
6.2	Global obs of mixed-phase clouds from the A-train	Dr R Hogan, Reading Prof K Carsaw, Leeds	LEM simulations of clouds	Dr R Hogan, Reading	June 10	100.0	a) Ice and liquid cloud retrievals from MODIS and the CloudSat 94-GHz radar, the Calipso depolarization lidar, and radiometers. b) Aircraft campaign data from ACE1 in Tasmania c) CERES data	University of Reading NERC project	

2000.0

Other requirements: BADC workspace

### 3. ADIENT

WP		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
1.2	FAAM flying campaign - ADIENT-PLUME	Dr E Highwood, Reading Prof H Coe, Manchester	145 hours of core and non-core FAAM data for flights in Nov 07, May-June 08 and Nov 08 and joint EUCAARI flights around Munich during 2008. Of note are ToF-AMS, DMA, SP2, UHSAS, VACC, ORAN, PAN and large radiometers.	Reading PDRA Manchester PDRA J. McQuaid and B Brooks for VACC, ORAN and PAN. J. Haywood for large radiometers	Nov 08	200.0	AERONET ground-based data for comparisons.		Pass flying data to Met Office
1.3	Satellite data in support of flights	Dr R Grainger, Oxford	Images of MODIS, MISR, AATSR, MSG-SEVIRI and CALIPSO radiances and aerosol products in NRT on ADIENT website Software extraction tool	R Grainger, Oxford	Nov 07 - Nov 08	500.0			
1.4	Ground-based data in support of flights	Dr E Highwood, Reading	Sunphotometer data Aerosol profiles from Lidar technology portable lidar	PML	Dec 08	100.0			
2.1	Properties of anthropogenic aerosol	Prof H Coe, Manchester	None		Nov 08		NOAA/Manchester Holme Moss experiment	Manchester	
2.2	Evolution of properties of anthropogenic aerosol	Dr E Highwood, Reading Prof H Coe, Manchester	None						
2.3	Simulations of aerosol evolution in pollution plumes	Dr G McFiggans, Manchester Prof K Carslaw, Leeds	CMAQ vanilla and CMAQ MADRID model simulations of aerosol properties during ADIENT-PLUME GLOMAP/UKCA model simulations of aerosol properties in nrt during campaigns		Jan 09 and Sept 09	500.0			
3.2	Determination of radiative impact of anthropogenic aerosol from GERB/SEVIRI observations	Dr H Brindley, Imperial	15 minute / 10 km resolution TOA SW fluxes in ADIENT-PLUME and ADIENT-BUDGET regions		Nov 09	100.0	MSG instruments GERB and SEVIRI	Imperial	

4.1.1	Synthesis of optical properties of key aerosol types	Dr E Highwood, Reading	Enhanced dataset of aerosol optical properties	Reading PDRA	Dec 09	10.0			
4.2.1	Regional aircraft in situ analysis	Dr E Highwood, Reading	1995-2005 regional aerosol description	Reading PDRA	Sept 09	50.0	SAFARI, DABEX, DODO, SHADE, TARFOX aircraft data	BADC, USA.	
4.2.2	Regional analysis of satellite and ground-based data	Dr R Grainger, Oxford	Derived dataset giving a statistical analysis of GlobAerosol and AERONET measurements		Sept 09	1.0	1995-2005 GlobAerosol and 1995-present sunphotometer AERONET data		
4.2.3	Regional UKCA-TOMCAT data	Prof K Carslaw, Leeds	Statistical analysis of the 1995-2005 aerosol fields at various scales from UKCA-TOMCAT model		Mar 09	1.0			
4.2.4	Synthesis of regional fields	Dr R Grainger, Oxford	Dataset providing a regional analysis of aerosol loading and properties		Sept 09, Dec 09 and Mar 10	10.0			
4.3.1	Comparison of model radiative effects with global measurements	Dr B Kerridge, RAL	Dataset of simulated satellite radiances		Mar 10	200.0	1995-present ATSR-2 and AATSR	RAL	

**1672.0**

Other requirements: project website

#### 4. Core 1 – York

Milestones		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
2	Analysis of historical SOA samples from EUPHORE smog chamber	Prof A Lewis Dr J Hamilton	Composition data form alpha-pinene and toluene precursor species	Dr J Hamilton, York	May 07 and May 09	1.0			
3	Analysis of aerosol samples from the EUPHORE smog chamber during 2006 EUROCHAMP campaign	Prof A Lewis Dr J Hamilton	Gas and aerosol phase composition and precursor data for reactions of p-xylene, isoprene, hexenol, and hexenol acetate with OH + O <sub>3</sub> Instruments used will be a) GCxGC-TOFMS instrument b) LC- microTOFMS instrumentation c) Bruker HCT Plus HPLC tandem ion trap MS	Dr J Hamilton, York	Nov 07	1.0	2006 EUROCHAMP smog chamber project data	EUROCHAMP consortium members will provide	
9	Analytical measurements on samples collected in APPRAISE projects	Prof A Lewis Dr J Hamilton	Low molecular weight species results	Dr J Hamilton, York	Nov 09	1.0			

3.0

## 5. Core 1 – Bristol

Milestones		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
1	The Hygroscopicity and Aging of Organic Aerosol	Dr J Reid Dr G McFiggans	Chamber studies of SOA aging. Sizing by mobility (DMPS), hygroscopicity from HTDMA, activation determination from CCND counter, composition from AMS	Bristol PhD student	Oct 08	5.0			
	Single particle studies of vapour/ particle partitioning for semi-volatile organics	Dr P Griffiths Dr Christine Braban Dr Tony Cox	Electrodynamic balance data on aerosol particle mass, size, refractive index and evaporation rate	Cambridge PDRA	Oct 08	5.0			
	Studies of effect of aging, association and accretion reactions	Dr J Reid	Hygroscopicity of organic aerosol Particle size and composition measurements	Bristol PhD student	Jul 08	5.0			
2	The Freezing Temperatures of Mixed Component Organic/Inorganic/Aqueous Aerosol	Dr Christine Braban Dr P Griffiths Dr Tony Cox	Flow-tube FTIR measurements of freezing temperatures and nucleation rates of low solubility organic salts	Cambridge PDRA	Mar 09	5.0			
	Corresponding single particle work	Dr Christine Braban Dr P Griffiths Dr Tony Cox	Phase transitions and freezing behaviour measured by the electrodynamic balance	Cambridge PDRA	Jun 09	5.0			
3	The Kinetics of Mixed Component Particle Growth	Dr J Reid	Equilibration time measurements for mixed component aerosol	Bristol PhD student	Jun 09	5.0			
	Studies on organic films	Dr J Reid	Permeability, evaporation and growth measurements	Bristol PDRA	Jul 09	5.0			
	Corresponding ensemble work in chamber	Dr Christine Braban Dr P Griffiths Dr Tony Cox	Growth factors from a tandem DMA system	Cambridge PDRA	Sep 09	5.0			
4	Single particle measurements of equilibrium partitioning of organic aerosol components	Dr J Reid	Particle sizes as a function of RH and T using optical tweezers and brightfield spectroscopy	Bristol PhD student	Sep 10	5.0			

	Dependence of mixing state, phase partitioning and salting	Dr J Reid		Cambridge PDRA	Sep 10	5.0		
--	--	-----------	--	----------------	--------	-----	--	--

**50.0**

## Core 2

Milestone		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
2.1.6	Laboratory measurements of mineral dust phase function	Dr Z Ulanowski, Hertfordshire	Dust phase data from electrodynamic levitation techniques	PhD student, Hertfordshire	Jun 09	10.0			
2.1.7	Benchmark high resolution radiative transfer modelling	Dr E J Highwood, Reading Dr A Dudhia, Oxford Dr K Shine, Reading	Model results for well-defined reference cases generated by the benchmark code.	Reading PDRA	Dec 09	100.0	a) DABEX, DODO, ADRIEX FAAM data b) Aerosol composition data c) SEVIRI and GERB data d) Measurements of single scattering albedo of organic aerosol	BADC  Manchester Imperial Bristol	
	Comparison to satellite aerosol retrievals	Dr B Kerridge, RAL				0.0	AATSR and SEVIRI	NEODC, RAL	
2.2.2	Lidar observations at Cardington	Prof R Jones, Cambridge	Lidar profiles of water vapour, ozone and aerosols		Throughout 2008	20.0			
2.2.3	Co-location of sunphotometers at Chilbolton for intercomparison with AERONET instrument	Dr E J Highwood, Reading Dr Tim Smyth, PML Dr Charles Wrench, RAL Dr John Foot, Met Office	Aerosol optical depth, size distribution and refractive index from Chilbolton Cimel sunphotometer Prede POM sunphotometer moved from Cardington UV lidar data from Chilbolton	Reading PDRA  Dr J Agnew, RAL	Jul 07 - Sep 08	10.0	AERONET sunphotometer network FAAM data e.g. AMMA, DODO	AERONET site  BADC	
2.2.4	Community available software for the derivation of aerosol radiative and microphysical properties from FAAM aircraft data	Dr E J Highwood,	Publication of software at BADC		Jun 09	0.0			
2.3.5	Formation of database of derived aerosol quantities	Dr E J Highwood, Reading	Database of global aerosol products from current and future satellite instrumentation	Reading PDRA	Dec 09	60.0	FAAM data	BADC	

200.0

### Core 3

Milestone		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
2	Model diagnostics for easy comparison against a wide range of instruments	Prof K Carslaw Dr G Mann Prof M Chipperfield	a) Hourly or 3-hourly output diagnostics from Global Model of Aerosol Processes (GLOMAP) for in situ instruments including TOF-MS, DMPS, CN counters, CCN spectrometers, HTDMA. b) Aerosol optical depth and standard optical properties for radiation diagnostics at AERONET sites	Leeds PDRA	Jan 09 and Jan 10	100	ECMWF cloud liquid and ice water contents		
3	Model evaluation	Prof K Carslaw Dr G Mann Prof M Chipperfield	Quality checked final model runs	Leeds PDRA	Jun 10	100	To be determined benchmark datasets to test model outputs		

200.0

## Core 4

Milestone		Investigators	Data deliverables	Data providers	Expected delivery date	Expected volume (GB)	Third-party data required	Third-party data location	Third-party data use
1.1	Development of a parcel model	Dr P Connolly Prof T Chouarton Dr M Gallagher Dr G McFiggans	None			0.0	a) AIDA Chamber simulations of ice nucleation and water activation on soot and dust b) CLACE data	Manchester	
1.3	Explicit Microphysics Cloud Resolving Model	Dr Z Cui Prof T Chouarton Dr P Connolly Prof A Blyth Prof K Carslaw	None			0.0	ICEPIC and ACTIVE data	Leeds	

0.0

## **Annex 3.**

---

### **Archiving of Simulations within the NERC Data Management Framework: BADC Policy and Guidelines**

#### **Introduction**

1. Issues associated with archiving information about the environment made by measurement are relatively well understood. This document outlines a general policy for archiving simulated and/or statistically predicted data<sup>1</sup> within NERC and provides specific policy and guidelines for the activities of the British Atmospheric Data Centre.
2. In the remainder of this document we use the term simulation to cover deterministic predictions (or hindcasts) based on algorithmic models as well as statistical analyses or composites of either or both of simulations and real data.
3. This policy has been developed in response to external legislative drivers (e.g. Freedom of Information Act and Environmental Information Regulations), external policy drivers (e.g. the RCUK promulgation on open access to the products of publicly funded research), as well as the existing NERC data management policy which is based around ensuring that NERC funded research is exploited in the most efficient manner possible.
4. The major question to be answered when considering simulated data is whether the data products are objects that should be preserved in the same way as measured products. In general the answer to this question is non-trivial, and it will be seen that guidelines are required to implement a practicable policy.

#### **Data Management and Simulated Data**

5. In general the information provided by models and the information provided by measurements are of a different nature. Simulations are analogues of the “real” world that may provide insights on physical causal relationships, while measured data are the observed symptoms of these relationships.
6. Simulations are generated by either deterministic or statistical models (or a combination of both). Such modelling activity does not generate definitive knowledge. Models are continuously developed and hopefully (but not necessarily) provide improved or more adequate representations of the physical system as time progresses. This is to be contrasted with measurements of the earth system, which by definition, cannot be repeated with the system in the same state and are therefore unique in a rather different way to simulated data.
7. Simulated data is usually produced by individuals, teams, or projects, and may have limited applicability, and/or potential for exploitation, in the wider community. However, the role for data management is not limited to making data more widely available, there is also a recognised role for data management to minimise duplication of activities between individuals, teams and projects, and to facilitate research programmes and collaboration. It is therefore important to develop criteria by which the scope for programme facilitation or wider applicability or exploitability can be recognised.

#### **Criteria for Selecting Simulated Data for Management**

8. If the answer to one or more of the following questions is yes, then simulated data are candidates for professional data management beyond that provided by the investigating team responsible for producing the data.

---

<sup>1</sup> The word “data” is often claimed by experimental scientists to exclude simulated information, however, most reputable dictionaries include simulated products within the definition.

- a) Is there — or is there likely to be in the future — a community of potential users who might use the data without having one of the original team involved as co-investigators (or authors)?
  - b) Does some particular simulation have some historical, legal or scientific importance that is likely to persist? (Some simulations may become landmarks, in some way, along the route of scientific knowledge. They may also have been quoted to make a statement that might be challenged – either scientifically or legally – and should therefore be kept for evidential reasons.)
  - c) Is the management of the data by a project team likely to be onerous or result in duplication of effort with other NERC funded activities?
  - d) Is it likely that the simulation will be included in future inter-comparisons?
  - e) Does the simulation integrate observational data in a manner that adds value to the observations?
9. If the answer to any of the following questions is yes, then the simulated data should not be archived.
- a) Is the data produced by a trivial algorithm that could be easily regenerated from a published algorithm description?
  - b) Is the data unlikely to ever be used in a peer-reviewed publication, or as evidence to support any public assertions about the environment?
  - c) Is the data known to be of poor quality or to have had no scientific validity?
  - d) Is it impossible to adequately document the methodology used to produce the data?
10. If the answer to any of the following questions is yes, then value judgements will need to be made about how much of the simulated data should be archived. Guidelines to assist in this situation appear below.
- a) Would storage of the data be prohibitively expensive?
  - b) Would storage of statistical summaries rather than individual data items provide adequate evidential information about the simulation? (e.g. while it might normally be desirable to store all ensemble members, would ensemble and/or temporal means be adequate in a situation where storage of the individual members at full time resolution might be prohibitively expensive).

### **Guidelines for Archiving Simulated Data**

11. In some cases, datasets may be archived by the investigating team at a national facility, rather than at a NERC designated data centre.
- a) This is most likely to occur when the longevity of the dataset is in some doubt, and the added value of using a designated data centre is not clear.
  - b) Where datasets will initially have restricted access (see para 16) it should normally be the case that the data archive is held at a designated data centre where procedures are already in place for providing secure access to data.
  - c) Alternative archives should not be established where the result will be that academic staff will be spending significant amounts of time carrying out professional data management which should be carried out within institutions with more appropriate career structures.
12. Where the intention is that a dataset be held outside of a NERC designated data centre, procedures should be in place to ensure that the data holder (or holders) conform to all the following requirements. It should also be ensured that funding is in place to move the data within a designated data centre when the holder (or holding facility) is no longer able to archive and distribute the data. Such datasets will still be

the responsibility of a designated data centre, but those responsible for the remote archives will be responsible for keeping all metadata required by the designated data centre up to date, and communicating the results of internal reviews (especially those which might involve removing or superseding data holdings).

13. All simulated datasets will be subject to regular lifetime review (described below).
14. Given that a simulation dataset is to be archived, what is involved in archiving such a dataset?
  - a) The simulated data itself should be archived in a format that is supported by the designated data centre community (whether or not the data is to be initially archived in a designated data centre. It is recognised that in taking on data, potentially in perpetuity, every new format is a significant ongoing cost.)
  - b) Any non-self-describing parameter codes (e.g. stash codes) included within the data should be fully documented.
  - c) Discovery metadata conforming to appropriate standards and conventions<sup>2</sup> should be supplied for all datasets to the responsible designated data centre.
  - d) Where possible, documented computer codes and parameter selections should also be provided (e.g. the actual Fortran, and full descriptions of any parameter settings chosen<sup>3</sup>).
  - e) Where initial conditions and boundary conditions are themselves ancillary datasets, these too should be archived and documented.
  - f) Estimates of the difficulty (both practically and financially) of recreating the simulation. (This will be needed to inform the lifetime review).
  - g) Where special tools (e.g. diagnostic software codes) are available to help interpret the simulation, these tools themselves should be archived if possible.
  - h) All documents and information (“further metadata”) should conform to appropriate archival standards (published open formats, suitable metadata structures etc)..
15. Where only a subset of the simulation is to be archived, the following considerations should be assessed in making decisions:
  - a) Potential usage (e.g. if the climate impacts community are involved appropriate parameters might include daily min/max temperatures, whereas instantaneous values are more likely to be useful if the simulation is to be used to generate initial conditions for other runs).
  - b) Illustrative value (where a simulation is being archived because of its scientific importance, those parameters relative to the scientific thesis should be the most important).
  - c) Physical Relevance (e.g. case studies, one might only store those parameters necessary to make the relevant points, but there are obvious risks in retrospectively identifying key parameters).
  - d) Volume and cost of storage.
  - e) Standard Parameters used in model-intercomparison exercises. Where possible and appropriate datasets should always seek to keep these, and the designated data centre community will provide guidance on current standard lists of parameters.

---

<sup>2</sup> In October 2005 this would be NASA GCMD DIF documents with the Numerical Simulation Extensions

<sup>3</sup> It is hoped that in the near future, the Earley Suite being developed at the University of Reading will provide an appropriate formalism for Unified Model Simulations.

- f) Can the temporal or spatial resolution be decremented without losing impact
16. When simulated data is initially archived, it may be possible for access to be embargoed in some way for a defined period<sup>4</sup>. When this occurs the following issues need to be addressed:
    - a) To which community should it be restricted and for how long?
    - b) Should conditions of use apply to the data during and/or after the retention period (e.g. communication with investigators, offer of co-authorship, acknowledgement in publications)?
  17. Where it is known a priori that simulation data will be archived, they should normally be archived at the time they are produced. Where multiple versions are expected within a project, and no other groups are expecting access to the data before a final version is produced, early simulations need not be archived. It should never be assumed that any part of a dataset would be archived after the end of the originating project.

### **Archive Lifetime**

18. As described in the introduction, continuous model improvement/development may make obsolete datasets made with previous versions. All simulated datasets should be subject to more frequent review procedures than measured datasets.
19. Where a dataset is being held for legal reasons, or because of historical interest, such a dataset might be kept indefinitely.
20. Where a dataset has been formally cited and formally published, it should be kept indefinitely, unless it is not possible to migrate the format to future media.
21. A suitable timescale for review of simulation datasets held at designated data centres would be at four-year intervals. Four years should give time for work to be published and follow-up work to be performed, and for an initial assessment of the likely longevity of datasets to be established. Most international programmes (e.g. IPCC) should have exploited datasets on a timescale of eight years, and again, further longevity could then be assessed. More frequent reviews may be appropriate where datasets are held elsewhere.
22. Reviews should involve at the minimum: the data supplier (if available), the custodians (especially if not held inside a designated data centre), representatives of the user community (if it exists), and an external referee.
23. Reviews may recommend removing subsets of a dataset.
24. Reviews may recommend acquiring new datasets to supersede existing datasets (and to keep multiple versions).
25. Reviews should consider the availability of tools to manipulate datasets.
26. In all cases metadata should be kept for datasets which have been removed.

### **Custodial Responsibilities**

27. The custodial responsibilities of designated data centres are described elsewhere. These points are here to provide guidance for the minimum responsibilities of facilities formally archiving simulation data on behalf of one or more designated data centres.

---

<sup>4</sup> The Freedom of Information Act (2000) and the Environmental Information Regulations (2004) stipulate that an embargo, if any, can only apply for some limited amount of time, to allow for “work in progress”.

28. All archived data will be duplicated, either in a formal backup archive, or by complete archive duplication at multiple sites (in which case the remote sites must support all the same metadata structures, and they must advise the designated data centre should they consider removing their copy).
29. All cataloguing and metadata required by the designated data centre must be provided and kept up to date.
30. User support must be provided to include help with any access control, on how to view and interpret the metadata, and on how to obtain and use the data in the archive.
31. Formal dataset reviews must be carried out.
32. Adequate bandwidth to the data holdings must exist.
33. Appropriate tools to use and manipulate the data must be provided.