CWVC data management at the BADC

Report to the CWVC Steering Committee

Anne De Rudder, BADC
October 2002

1. Introduction

The present document reports the progress of the CWVC archive and the work accomplished at BADC during the elapsed year. As a reminder, a data management plan for the *Clouds, Water Vapour and Climate* NERC thematic programme was submitted by the BADC to the programme steering committee in September 2001. It was the result of an enquiry led by the BADC and addressed to all CWVC project principal investigators.

2. Data management plan and data protocol

The draft DMP submitted to the programme steering committee in October 2001 was discussed during the winter of 2001-2002. Several amendments were made and several versions were issued in order to take into account and conciliate, as far as possible, the wishes expressed by some PIs, the programme steering committee chairman, G. Jenkins, and the programme science coordinator, K. Bower. The final version of the DMP (see Annex 1) was agreed on in February 2002 and published on the BADC Web site. It includes a data protocol by which any scientist applying for access to CWVC data has to abide. The data protocol (see Annex 2) makes provision of rules regarding data archive locations (BADC and Met Office), accepted data formats, data submission procedure, data access restrictions, public data release and protection of publishing rights.

3. Liaison with principal investigators

A letter addressed to the third round project PIs, who were not included in the original enquiry, was sent in January 2002. Only one formal reply was received (from J. Methven), informal information was received from the GRAPE project. Four of the seven 3rd round projects are led by PIs of former round projects and are hence not likely to provide any new information, but information should be obtained regarding the entirely new projects. Note that the GRAPE project will deliver a large amount of data and the BADC will provide a large amount of support funded directly by that project. It is not yet clear what the longer term requirements for archival from that project will be (see Section 7).

4. Archive, documentation, data submission and online help

A computer archive has been set up on a BADC machine. The disk is backed up weekly on tape. Additional back-ups on both tapes and disks take place at a lesser frequency, including storage in a fire safe. Access to the archive will be temporarily restricted to the CWVC participants, as stated in the data protocol. A password-protected system has been set up. The Web-based BADC file uploader has been updated to include CWVC projects and has been linked to the CWVC home page, providing the CWVC data providers with a straightforward way to submit data files to the archive. Documentation on the projects has been made available online and guidance in data formatting, accompanying metadata, data submission and the use of the archive has been provided. The URL of the CWVC home page at BADC is http://www.badc.rl.ac.uk/data/cwvc/.

5. Acquisition of supporting data sets

The *International Satellite Cloud Climatology Project* (ISCCP) data set has been obtained and stored at the BADC. The archive of this data set must still be reorganised and documented in order to be usable (see Section 7).

6. Progress in the archive population

Four projects are finished, namely GST/02/2318 (Harries), 2321 (Kaye), 2324 (Whiteway et al.) and 2871 (Shine). No information has been received regarding the first of these projects but, according to their answers to the enquiry, the other three should have submitted data to the BADC by now. However, these projects are led by teams who have received further grants in the next rounds, so that data production, validation and submission may in some cases be considered as pertaining to the set of two or three successive projects. In any case, information and/or data should be obtained from the project leaders (see Section 7).

Robin Hogan has sent to the BADC a detailed description of processed radar data from the Chilbolton facility, that are part of the deliverables of Project GST/02/2874. The data, stored in NetCDF files, will be submitted to the BADC shortly.

7. Future work

In the near future, CWVC data management tasks will include the following.

- Chasing 3rd round PIs for information on their deliverables, in particular finding out what the long-term requirements of the GRAPE project will be (the GRAPE project will be making use of the BADC tape-library during a period while it is not being used for other projects – the longer term archival issue was not addressed by the BADC funding within the GRAPE project).
- Chasing 1st and 2nd round PIs for data submission (projects GST/02/2321, 2324, 2871).
- Making ISCCP data available.
- Monitoring archive access and releasing data to public domain when appropriate.
- Updating data documentation and Web site.
- Reporting to the CWVC SC.

8. Staff

A new staff member, Dr Helen Walker, will start to work with the BADC on October the 14th. Dr Walker is presently part of the Space Physics Division of the Space Science and Technology Department at RAL, where she will continue to work part-time (70% of her working time will be devoted to the BADC). Dr Walker will take over the data management of cloud data hosted by the BADC. She will assist A. De Rudder in CWVC data management. In particular, she will ensure the archival of new coming data sets produced by the programme and liaise with the project principal investigators, science coordinator and steering committee.

9. Budget

A data management budget (see Annex 3) has been submitted to the SC and has been accepted in April 2002. The budgets for Years 2001-02 and 2002-03 are of £10K and £20K respectively. It is distributed into tasks as displayed in the table below.

Year	Task	Budget (£K)	Status
2001-02	Enquiry on project deliverables + production of DMP and Data Protocol	10	Completed

2002-03	Set up of archive, Web site and uploading system	7	Completed
	Data archival, maintenance and documentation + support to data providers	8	• Chilbolton radar data in course of submission (GST/02/2874). • Support to data providers in the form of online documentation on format and CF compliant metadata
	3rd Party data.	5	ISCCP data set obtained but needs some reorganisation before being published

Annex 1

NERC Thematic Programme

'Cloud Water Vapour and Climate (CWVC)'

Data Management Plan

BADC

December 2001
Updated February 2002

Scope

The purpose of the CWVC 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.
- CWVC data are made available to CWVC collaborators in a timely fashion.
- 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 CWVC data archive is created.
- Data and documents are eventually distributed to the scientific community.

The following sections tackle the tasks to be completed during the programme development in terms of data management, namely:

(1) Preliminary enquiry, data management plan writing and adoption of a data protocol
(2) Third-party data acquisition
(3) CWVC data archival
(4) Data distribution
(5) Publication

1. Preliminary tasks

An enquiry has been conducted with the project principal investigators (PI) and some of their collaborators to determine their needs, wishes and some characteristics of their deliverables, in order to produce the present data management plan and annexed CWVC Data Protocol.

2. Third-party data

2.1 Third-party data external to the CWVC programme

Third-party data required for the development of the 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. Other data sets distributed by BADC may be of interest to the CWVC community, such as the data from the International Satellite Cloud Climatology Project (ISCCP) or from the Stratospheric Photochemistry, Aerosols and Dynamics Expedition (SPADE). 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 (possibilities include the Small Cumulus Microphysics Study data, additional Met Office data, etc.).

2.2 CWVC third-party data and model results

Data and model results generated by CWVC groups during the programme development will be made available to all other CWVC groups through the designated data centre(s) (see Section 3.1). Update of preliminary data should be announced as early as possible to collaborators from other CWVC teams. Publication issues are dealt with in Section 6.

3. CWVC data archive

3.1 Archive location

The central CWVC archive will be located at BADC. Data collected aboard the C-130 aircraft for Projects GST/02/2316 & 2874 will be archived, maintained and distributed by the Met Office (Y46 Bldg, Cody Technological Park, Farnborough GU14 0LX) provided that a letter from the principal investigator guarantees the long-term integrity of the archive. The C-130 data will be moved to the BADC in the case where the programme requirements cannot be met by the Met Office in terms of archival, maintenance and distribution. Although the C-130 archive itself does not need to be duplicated, full information on the C-130 database, including how to access the data, will be supplied to the BADC so that all the CWVC documentation can be obtained from a central location.

3.2 Archiving policy

In recognition that validated raw data (i.e. QA/QC'ed data prior to additional processing) potentially represent an invaluable source of information for the future, the programme participants will archive them in a way that guarantees longevity and accessibility. Although not necessarily located at one of the CWVC data centres, validated raw data bases and their access must be fully documented at the BADC. Processed (final) data will be archived at one of the official CWVC data centres. In addition, investigators are encouraged to submit model results which would be the basis of theoretical studies or would illustrate the model use.

3.3 Format

Spectroscopic data (Project GST/02/2871) will be stored in the format of the HITRAN database, since this format is widely used within the spectroscopists' and modellers' communities. Other data will be formatted in either NASA Ames or NetCDF. Documentation on all three formats is available from the BADC (http://www.badc.rl.ac.uk/formats/), as well as links to downloadable free software packages to produce and read NetCDF files.

3.4 Data submission

When needed by other CWVC groups, preliminary data should be made available to them as soon as possible, if possible via one of the designated data centres. Processed data and model results should be supplied to the relevant data centre 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 BADC provides an automatic Web based file uploader accessible by clicking on the *Submit Data* option in the BADC Web pages menu. Online assistance is provided. Alternatively, files can be submitted by *ftp*. Both ways are fully documented on the BADC Web site.

3.5 Documentation

Metadata are a crucial part of any data archive since they ensure the readability of the data. It is therefore essential that metadata are submitted at the same time as the data sets to which they pertain. Metadata pertaining to all CWVC data archived at the Met Office or elsewhere must also be supplied to the BADC.

To guarantee the CWVC data archive quality, full documentation on all validated raw and processed data, as well as on models and model results, must be provided to the BADC. Metadata standards for each of these cases and for the three data formats will be available online.

Standard metadata will be archived within the NASA Ames and NetCDF data files. Standard metadata pertaining to HITRAN data files and to models will be submitted as text files.

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

4. Data distribution

The access to all data submitted to the designated data centres will be restricted to the CWVC participants during one year following the concerned project end date, after which they will be released into the public domain

A password protected system will be set up at BADC to reflect actual access permissions. Whilst the data are restricted from the public domain, participants will be prompted to agree with the CWVC Data Protocol (see Annex) in order to access the CWVC archive.

After release of the data to the public domain, anonymous users will be requested not to use the data for commercial purposes. They will be asked to contact the relevant data providers before using the data and to acknowledge the CWVC programme and the data suppliers in any publication using CWVC data. Users will be asked to indicate agreement to these terms prior to being given access to the data.

Distribution of the CWVC data held at BADC will take place via the Web. During the validation period, entitled CWVC participants who will have applied for access to the data will be allocated an account at BADC that will allow 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 the project PI. A CWVC Web front page has been set up at http://www.badc.rl.ac.uk/data/cwvc/. This will be the gateway to all CWVC data and metadata, and to all relevant information and links.

CWVC data held at BADC will benefit from future development of access technology. Facilities currently under development include a metadata catalogue and gateway, and a *live access server* allowing data subsetting, visualisation, conversion, etc.

5. Publication

Results coming out of the CWVC research projects will be published in the usual way. During the data validation period, each investigator will have the right to refuse the use of his results in a publication or a presentation prior to the investigator's own publication of that work. If measurements or model results from other groups within CWVC are used in a CWVC participant's publication during or after the programme, joint authorship must 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 will be communicated to the BADC.

Annex 2

CWVC Data Protocol

The aims of the Data Protocol are

- to encourage rapid dissemination of scientific results from CWVC;
- to protect the rights of the individual scientists funded by CWVC;
- to have all the involved researchers treated equitably;
- to ensure the quality of the data in the CWVC 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 CWVC programme. Therefore to try to meet these aims, all PIs involved in CWVC, in accordance with and on behalf of their co-investigators, must agree to abide by the following conditions:

1. CWVC data and model results produced during the programme will be made available to all CWVC participants, and to CWVC 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 CWVC data centres are the Met Office for data collected aboard the C-130 aircraft and the BADC for all other data.
3. The longevity of validated raw data must be ensured in a secure archive, if possible at one of the designated data centres. 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.
4. When relevant, preliminary data must be made available to CWVC collaborators as soon as possible. Any corrections or amendments to the preliminary data should be announced as soon as possible.
5. Validated processed data (i.e. data sets in their final form) must be archived at one of the designated CWVC data centres. Archival must take place no later than the end of the concerned project.
6. Results of model studies feeding other CWVC projects or using data acquired during CWVC can be made available via the BADC.
7. Data submitted to the BADC must be in the data format agreed between CWVC principal investigators and the BADC. All agreed metadata describing data, models and model results, regardless of their archival location, must be supplied to BADC. Format and metadata are documented at 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 research groups within CWVC are used in a publication by a CWVC 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.
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 CWVC Scientific Steering Committee.

Annex 3

CWVC Data Management budget

Year	Task	Personnel (person.year)	Estimated cost (K£)	Yearly cost (K£)
2001-02	• Enquiry on project deliverables + production of DMP and Data Protocol	0.17	10	10
2002-03	• Archive, Web site, uploading system set up • Data archival, maintenance and documentation + support to data providers • Data purchase (5K)	0.12 0.14 /	7 8 5	20
2003-04	• Data archival, maintenance and documentation + support to data providers	0.17	10	10
2004-05	• Data archival, maintenance and documentation + support to data providers	0.17	10	10
2005-06	• Data archival, maintenance and documentation + support to data providers • Data distribution and integration of data sets into Live Access Server (LAS)	0.35 0.17	20 10	30
Total	• Work • Data purchase	1.29 /	75 5	**80**