

Polluted Troposphere NERC Thematic Programme

Data Management Plan

BADC
March 2004

Scope

The purpose of the Polluted Troposphere (PT) 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.
- PT data are made available to PT 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 Polluted Troposphere data archive is created.
- Data and documents are eventually distributed to the scientific community.

The following sections tackle the issues to be addressed in terms of data management, namely the provision of a data management plan and a data protocol, the determination of the programme participants' needs, the setting up of the archive, the monitoring of data access, the data distribution, the publication of results based on Polluted Troposphere data and the support offered to data providers.

1. Data management plan and data protocol

The present document is the final version of a data management plan drafted by the BADC, submitted to the Polluted Troposphere Steering Committee for discussion, amended to reflect the outcomes of the discussions, and agreed by the steering committee. A Polluted Troposphere Data Protocol has also been issued. This last document is a list of clauses stating the conditions that rule various aspects of the Polluted Troposphere data submission, access and publication. All applicants to Polluted Troposphere data will be asked to abide by this protocol. The two documents can be freely consulted on the Polluted Troposphere web page at BADC (<http://badc.nerc.ac.uk/data/polluted-trop/>).

2. Collection of information

Visits to the project principal investigators have been carried out by the BADC from May to September 2003 to determine their needs in terms of data support (provision of third-party data sets or other services). The enquiry has also addressed issues related to the data generated by the projects (nature, volume, flow, etc.). The collected information will provide the basis of a planned and coordinated data management.

3. A collaborative tool for the exchange of ideas and preliminary data

A Polluted Troposphere online workspace has been created on the BADC BSCW (Basic Support for Cooperative Work) server. The workspace is visible and accessible only by the Polluted Troposphere participants and is 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, a workshop and a temporary repository for data in the validation phase, draft papers, reports, etc.

4. Third-party data

4.1 *Third-party data external to the Polluted Troposphere 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. 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.

4.2 *Polluted Troposphere third-party data and model results*

Data and model results generated by PT groups in the course of the programme will be made available to all Polluted Troposphere groups through the BADC. Update of preliminary data should be announced as early as possible to collaborators from other Polluted Troposphere teams. Publication issues are dealt with in Section 6.

5. Polluted Troposphere archive

5.1 *Archive location*

The Polluted Troposphere archive is located at the BADC (<http://badc.nerc.ac.uk/>). All the documentation related to Polluted Troposphere data can be accessed from the BADC “Polluted Troposphere” front page at <http://badc.nerc.ac.uk/data/polluted-tropo/> or by following various links from the BADC home page.

5.2 *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 are encouraged to archive them in a way that guarantees longevity and accessibility (for example by using the BADC archive), and to document them at the BADC. Processed (final) data will be archived at the BADC, as well as model output that will be requested by other PT groups. In addition, investigators are encouraged to submit model results which will have been the basis of theoretical studies or that illustrate the model capabilities. They are welcome to archive any relevant pieces of code.

5.3 *Format*

In order to ease future data use and unless this reveals inadequate to some particular cases (such as the widespread use of a specific format by a specialised community), NetCDF (and possibly NASA Ames if requested by some investigators) will be adopted for all the processed data generated by the programme. The reasons why the BADC encourages the use of NetCDF are its generalised use within the atmospheric community (which also holds for NASA Ames), the existence of a large library of software reading, writing and converting NetCDF files, and the fact that it will ease the datasets visibility through a number of devices currently under development (BADC Live Access Server, NERC metadata gateway, NERC data grid). Documentation on these two formats is available from the BADC (<http://www.badc.rl.ac.uk/help/formats/>), as well as links to downloadable free software packages to produce and read NetCDF files.

Note on the BADC Live Access Server (LAS).

This tool will allow users to select subsets of data sets and retrieve them or visualise them on their screen. It will also allow comparison and some manipulation of the data. Although a variety of formats will eventually be readable by the LAS, NetCDF files will first benefit from this new device.

5.4 Information on the data

Metadata (i.e. information on the data) are a crucial part of any data archive since they ensure the accessibility and readability of the data. It is therefore essential that metadata be submitted at the same time as the data sets to which they pertain. Metadata pertaining to all raw Polluted Troposphere data not archived at the BADC should also be supplied to the BADC.

To guarantee the PT 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. The BADC has developed a set of metadata requirements which is available to PT participants via the BADC Web site (<http://badc.nerc.ac.uk/help/metadata/>) and is accessible via a link from the Polluted Troposphere front page. Metadata pertaining to NetCDF files should comply with the Climate Forecasting (CF) convention. Metadata inserted in NetCDF and NASA Ames files are fully documented online.

The provision of a complete and carefully written set of metadata by each investigator will ensure the swift integration of the Polluted Troposphere into the network system that will soon be under development at the BADC in the framework of the NERC Data Grid project. When operational, this system should provide searchable information about all NERC data that will have been integrated into the grid, and provide links to the NERC archives. For the future use of Polluted Troposphere results, it would be desirable that the Polluted data sets be visible through this gateway.

In addition to these minimal metadata, investigators are encouraged to archive at BADC all relevant information, including model and experiment descriptions, references, papers, reports, etc. Designated directories have been created in the PT archive for this purpose.

5.5 Data submission

If needed by other Polluted Troposphere 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 BADC provides an automatic Web based file uploader accessible by clicking on the *Submit Data* option in the BADC Web pages menu. The uploader has been set up to accept files from the Polluted Troposphere programme. Online assistance is provided. Alternatively, files can be submitted by *ftp*. Both ways are fully documented on the BADC Web site.

6. Data distribution

The access to all data submitted to the BADC will be restricted to the Polluted Troposphere participants during two years following the concerned project end date, after which they will be released into the public domain.

A password protected system has been set up at the BADC for the Polluted Troposphere participants. Whilst the data are restricted from the public domain and contingent on the adoption of a PT Data Protocol, it is suggested that participants would be prompted to agree with the Polluted Troposphere Data Protocol in order to access the PT archive (whether to submit, view or download data).

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 Polluted Troposphere programme and the data suppliers in any publication using Polluted Troposphere data. Candidate users will be asked to indicate agreement to these terms prior to being given access to the data.

Distribution of the Polluted Troposphere data held at BADC will take place via the Web. During the validation period, entitled Polluted Troposphere 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. If the SC expresses the wish of it, this facility will be extended to external collaborators who will have been personally authorised to access the data by the project PI. A Polluted Troposphere Web front page has been set up at <http://www.badc.rl.ac.uk/data/polluted-tropo/>. This will be the gateway to all Polluted Troposphere data and metadata, and to all relevant information and links.

7. Publication

Results coming out of the Polluted Troposphere 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 the PT programme are used in a PT 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 so that they can be integrated into the PT documentation.

8. Liaison between BADC and programme participants

The BADC Polluted Troposphere homepage (<http://www.badc.rl.ac.uk/data/polluted-tropo/>) and its links will be the primary source of information regarding the PT archive. The Web pages will be updated as new pieces of information are available and new data files are submitted.

The BADC will keep in touch with the PIs and their collaborators during and after the programme development. In particular, information on the submission procedure, on the relevant WWW links, on the Data Management Plan and on the population of the PT archive will be sent to PIs via this mean.

9. Support to PT participants

In addition to the provision of third-party data, the BADC will endeavour to provide services that would be required by the project PIs. PT participants will benefit from the existing BADC services (file uploader, NASA Ames file checker, online trajectory calculator, quick plot facility) and from the future developments of access technology mentioned above (format converter, LAS, NDG).