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[1] 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.

[1] The word "data" is often claimed by experimental scientists to exclude simulated information, however, most reputable dictionaries include simulated products within the definition.

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.

    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[2] 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[3]).

    e) Where initial conditions and boundary conditions are themselves ancillary datasets, these too should be archived and documented.

---

[2] In October 2005 this would be NASA GCMD DIF documents with the Numerical Simulation Extensions.

[3] 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) 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 it's 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.

    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[4]. 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.

[4] 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".

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.

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.