

DRAFT

CCSM DATA MANGEMENT PLAN

1. Introduction

The scientific objectives of the Community Climate System Model (CCSM) project are as follows:

1.1 Develop and continuously improve a comprehensive climate modeling system that is at the forefront of international efforts to understand and predict the behavior of Earth's climate.

1.2 Use this modeling system to investigate and understand the mechanisms that lead to interdecadal, interannual, and seasonal variability in Earth's climate.

1.3 Explore the history of Earth's climate through the application of versions of the CCSM suitable for paleoclimate simulations.

1.4 Apply this modeling system to estimate the likely future of Earth's environment to provide decision-support information required by governments in support of local, state, national, and international policy determination.

The project is organized and conducted so as to involve a large community of scientists and stakeholders in the scientific and technical design, evaluation, and use of the modeling system and its data outputs.

The present policy that supports and governs the access and use of the outputs from the project is displayed on the CCSM Web site at <http://www.cesm.ucar.edu/experiments/datapolicy.html>. That policy mainly covers the timing of sharing of data from CCSM runs and data access, but it does also cover issues of data format and responsibility for data quality assurance. This paper is intended to provide further documentation on the implementation of the CCSM Data Policy and some standards for the management of the CCSM data that are not necessarily covered in that policy. There will need to be an integration of the topics covered in this paper and those on the Web site.

2. CCSM Data User Community

Users of CCSM data span a wide range of interests. An incomplete list includes:

- Scientists at universities, federal laboratories, and NCAR
- CCSM developers at universities, NCAR, and national laboratories
- CCSM working groups performing production integrations
- CCSM working groups performing development integrations
- Impact analysts
- National Assessment program
- Intergovernmental Panel on Climate Change (IPCC) Data Distribution Center

- Coupled Model Intercomparison Project/Paleoclimate Model Intercomparison Project (CMIP/PMIP)
- Policymakers
- Other modeling groups using CCSM data as forcing input to their models (e.g., regional climate models)
- Industry

The broad common needs of these users are ready access to the data, diagnostics of CCSM performance (scientifically and computationally), and various types of analysis and analysis tools.

3. Elements of the CCSM Data Management Plan

The CCSM Data Management Plan documents the policy and procedures for the management of model data produced by the CCSM. The overall goal of CCSM data management is to provide the best possible access and ease-of-use of high-quality CCSM data to diverse users over an extended period of time within the constraints of available resources.

Some of the key elements of a CCSM Data Management Plan include:

- Definition of the different categories of CCSM data sets
- Definition of ownership rights and responsibilities of the different categories of CCSM data sets
- Requirements for archiving and accessing the CCSM data
- Coordination of the distributed CCSM data archives across NSF, Department of Energy (DOE), and National Aeronautics and Space Administration (NASA) data centers
- Enabling fast and easy Web access to CCSM data
- Registering and auditing the characteristics of CCSM data users
- Definition of a life cycle for CCSM data and provision for the long-term stewardship and stability of CCSM data
- Metadata requirements and standards for CCSM model data sets

Other issues, such as coordination of CCSM data and metadata formats and providing tools for translating CCSM data into other popular formats, need to be addressed. For each element in the list above, policies will be advanced.

3.1 Categories of CCSM Data Sets

In the pursuit of the scientific goals of the CCSM project, many different types of runs are carried out. For the purposes of the CCSM Data Management Plan, categories of runs are defined on the basis of such attributes as the length of the run, changes introduced into the model, and the purpose of the run.

CCSM Control Runs are very long integrations (hundreds to thousands of years) with no changes to the model to establish the baseline climate of the model.

CCSM Experimental Simulations are long integrations (hundreds of years) made with modifications introduced into the baseline CCSM model or the model boundary data sets to conduct a scientific experiment or policy scenario.

CCSM Validation Runs are short runs (weeks to years) made to examine specific model behavior, such as response to changes in the representation of physical processes.

CCSM Test Runs are very short runs (days to weeks) made to test specific model software behavior, such as an exact restart capability or proper time coordination of the components.

Further descriptions of the characteristics of these runs are provided in Appendix A.

These different categories of runs produce output data sets of various forms as defined in Appendix B. For the purposes of the CCSM Data Management Plan, the data sets produced by the categories of runs outlined above will be designated as Control, Experimental Simulations, Validation, and Test.

The various categories of CCSM data sets are also characterized as Public, Published, or Private. All Control data sets are categorized as Public. All Experimental Simulations data sets are initially private. Data sets will evolve from Private to Published or Public according to the data access policy. Public data sets are owned by the CCSM Scientific Steering Committee (SSC), managed by the CCSM Project Office, and freely available to all. Private data sets are owned by the principal investigators (PIs) or working groups that created them, and they are managed by the working groups. These data sets become Published data sets by permission of the initial owners or through the expiration of the proprietary time period as defined by the data access policy. Published data sets are open to the public and managed by the working groups that created them. In some cases, the SSC may determine that there is sufficient broad interest that a Published data set should be designated as a Public data set.

All Validation data sets are private and remain so. All Test data sets are Public.

In general, the CCSM Data Management Plan is meant to apply to Public data sets. However, since Private data sets can be a source of Public data sets, it is expected that many of the procedures and standards will be common among categories of data sets.

3.2 Data Repositories for CCSM Data

CCSM Control and Experimental Simulations integrations are being carried out on a continuous basis at a number of computing centers around the world. Due to the large volume of data that is generated, no one center can support all this CCSM data. It will be necessary to coordinate CCSM data storage, discovery, and access policies among the various sites where CCSM data will be archived. This is particularly important for Public data sets.

Data produced by the CCSM are to be stored, managed, and distributed by the data archive center appointed by the entity sponsoring the CCSM run that produces the data. CCSM data created at NCAR under NSF support will be archived on the NCAR Mass Storage System (MSS). CCSM data generated at non-NCAR facilities should be archived at either the site of generation or its associated data archive center. CCSM data created at non-NCAR sites may be archived on the NCAR MSS if prior arrangements have been made with both CCSM and NCAR's Scientific Computing Division's management.

3.3 Online Access to CCSM Data

Web technologies allow for the efficient discovery and access of CCSM data. The CCSM working groups have been very active in establishing experimental Web portals to CCSM data subsets, both within NCAR and through DOE's collaborations. The two current systems for doing this are the Community Data Portal (CDP) and the Earth System Grid (ESG). To maximize the ease-of-access and value of the data to the scientific community, all CCSM "Public" data shall be made available via the ESG (<http://www.earthsystemgrid.org/>). The registration process through ESG will permit the assembly of information on the users and use of the CCSM Public data. CCSM Public data may also be made available through the CDP if there is sufficient demand and support for this access.

NCAR will serve as much CCSM data online as possible. Other centers archiving and serving CCSM data are encouraged to do so as well. All sites are expected to coordinate their data services.

3.4 Stewardship of CCSM Model Data

The CCSM data retention policy strikes a balance between the scientific need to retain data from older CCSM simulations, with the growing cost of doing so in a resource-limited environment. Unlike observational data, model simulation data often becomes less valuable with time as better models at higher resolutions are developed and run. Nevertheless, scientific analyses of CCSM Control Runs and Experimental Simulations Runs continue years after the data were generated. Accordingly, data from CCSM Control Runs and Experimental Simulations Runs shall be preserved for specified time periods to allow extraction of the maximum scientific content.

Data from CCSM Control and CCSM Experimental Simulations Runs that are Public will be retained for a period of 10 years.

Retention periods for CCSM test runs are:

- 5 years for public release testing
- 2 years for all other testing runs
- Retention periods for data from CCSM Validation Runs and Private data sets are left to the discretion of the PIs and the working groups.

CCSM data at NCAR will be retained under the guidelines of the data stewardship policy above. The owners of the data sets are responsible for the stewardship. A similar policy for

CCSM data sets held at other sites is encouraged. Sites holding CCSM Control or Experimental Simulations output should give the CCSM SSC the option of archiving this data at NCAR before this data is deleted.

3.5 CCSM Data and Metadata Requirements

Standard data and metadata formats are essential for the automated analysis necessary to efficiently interact with large data collections. In its broadest sense, metadata are simply “structured data about data,” describing important attributes of an information resource. Scientific metadata examples include descriptions of telescope images or the header files describing gridded CCSM model output. Metadata can be conceptually classed into two general types: discovery and use. Discovery metadata addresses the information necessary to find a data collection and determine its availability and appropriateness for the intended application. Use metadata provides the technical information necessary to actually use the data in the collection. Of the two types, use metadata are more mature due to the creators and consumers of geodata, converging in the last decade to a modest number of data storage formats containing reasonably well-defined data descriptions. Discovery metadata has only recently become an issue as operational and science centers have begun to move from static, in-house, data archives to dynamic, online, data services.

a. netCDF and the CF Conventions

CCSM uses netCDF as the standard data format for CCSM-related data sets. All CCSM models either create netCDF history files or provide a filter to convert files into netCDF. The use of netCDF makes CCSM output data readily accessible to a variety of existing graphics and analysis packages. In addition, CCSM3.0 uses the CF1.0 netCDF metadata convention, which is designed for the representation of gridded geophysical data. CF1.0 is based on, and very similar to, the Cooperative Ocean/Atmosphere Research Data Service (COARDS) conventions. This policy should be periodically re-evaluated in light of changing CCSM needs, data storage costs, and the emergence of new data formats. For example, netCDF lacks both a multi-threaded output capability and a good compression method. CCSM3 netCDF data sets comply with the [Climate and Forecast \(CF\)](#) metadata conventions. The convention is designed for the representation of gridded geophysical data. The CCSM netCDF convention follows the [COARDS conventions](#), with a few exceptions and additions to meet CCSM requirements. Translations of CF metadata into other metadata conventions, such as Dublin Core (<http://www.xml.com/pub/a/2000/10/25/dublincore/>), ISO (<http://www.fgdc.gov/publications/documents/metadata/nimapaper.html>), and FGDC metadata standards (http://www.sci.gc.ca/metadata/b4_a01_en.html), will be pursued through CDP and ESG collaborations.

b. Case and File Naming Conventions

The CCSM project has Case and File Name conventions to help keep track of the numerous simulations and their output data. The CCSM case naming conventions are outlined in the Web page at <http://www.cgd.ucar.edu/csm/experiments/csm1/names.html>. The CCSM file

naming conventions are outlined in the Web page at <http://www.cgd.ucar.edu/~njin01/ccsm/draft.html>.

An automated system will be put in place to assure compliance with the CF metadata standard as part of the quality control process in the conduct of all runs. The CCSM file naming conventions as outlined in the URL above should also apply to post-processed data sets.

APPENDIX A

More Detailed Characterization of Run Categories

1. CCSM Control Runs

CCSM Control Runs define the basic long-term climate of the CCSM. A control run needs to run long enough for slow adjustment processes, such as subsurface water in the land model, to come into balance. Some of these, such as the deep ocean heat and salinity, may take thousands of years to reach balance. The standard data output frequencies are monthly averages, with daily averages for a few select variables. Higher frequency data output for special analyses or to drive regional scale models may also be provided upon specific request, but are usually created for limited periods of the integration. CCSM Control Runs are documented on the CCSM experiments and data Web page at <http://www.cesm.ucar.edu/experiments/>. This documentation includes a description of the run and pointers to the validation plots and data files from the control run. The data from the control runs are the property of the SSC and managed by the CCSM Project Office.

2. CCSM Experimental Simulations

A CCSM Experimental Simulation is a run or a series of runs made with some modifications made to the control version of the CCSM. The modifications may be either in the representations of the processes in the model, the boundary data sets, or both. Experimental Simulations are usually a few model centuries but may extend to millennia. Usually, a given experimental design will include the production of an ensemble of runs to provide an estimate of the range and significance of the model response to the changes. The ensemble of runs constitutes the data set for the particular Experimental Simulations.

CCSM Experimental Simulations are documented on the CCSM experiments and data Web page at <http://www.cesm.ucar.edu/experiments/>. This documentation includes a description of the run and pointers to the validation plots and data files from the control run. The data set from an Experimental Simulation is initially the property of the PIs or working groups who have designed and conducted the experiment. Transfer of ownership to a broader community takes place over time as defined by the data access policy.

3. CCSM Validation Runs

A CCSM Validation Run is made under the direction of PIs or working groups for the purpose of measuring the model's response to some change made in the representation of processes or boundary conditions. Validation runs are usually of short duration (model weeks to

years). The data from such runs are the property of the PIs or working groups and are meant to be internal tools.

4. CCSM Test Runs

CCSM Test Runs verify that the CCSM successfully meets the requirements defined in the CCSM Requirements document at <http://www.cesm.ucar.edu/>. Test Runs are very short duration (days to weeks). Different test suites exist corresponding to the different CCSM release levels, and they are described in the CCSM Testing Plan at <http://www.cesm.ucar.edu/>.

APPENDIX B CCSM Data Sets

During the course of an integration, the CCSM produces three distinct output data streams: printed, restart, and history data. After a CCSM run finishes, the raw history data is post-processed into more useful collections referred to as post-processed history data.

Description	Volume	Data Format/Convention
a. Input Initial/Boundary data	(small)	netCDF or raw binary
b. Output Printed output	(small)	Plain text files
c. Output Restart data	(small)	Raw binary
d. Output Raw History data	(large)	netCDF compliant with CF convention
e. Post-processed History data	(large)	netCDF/CF, JPG images, HTML pages

1. Types of CCSM Output Data

1.1 Input Initial and Boundary Condition Data

CCSM runs are typically started using initial data sets that represent a known or idealized climate state for each CCSM component. Boundary condition files may also be used to prescribe time varying values of variables that are not predicted, such as the annual cycle of ozone in the atmosphere or emission profiles for future climate change scenarios.

1.2 Printed Output

The printed output contains diagnostic messages written by the various CCSM components during the course of a run. This includes a printed log file for the entire system, as well as printed log files from each of the CCSM components. The printed output's primary importance is for archiving details about the model run, how long it ran, and when it stopped and restarted. While the printed output contains little information useful for detailed model diagnostics, it provides a convenient method for displaying "quick look" diagnostics.

1.3 Output Restart Data

The CCSM restart data sets are raw binary files containing sufficient information for the CCSM to restart exactly. Restart data are usually output at monthly, half year, or yearly intervals. As the integration progresses, most old restart data are deleted to save disk space. The accepted practice is to retain restart data at decadal intervals.

1.4 Output Raw History Data

The raw history data contains the model data from each component of the CCSM. The history data consists of grid point representations of three-dimensional (latitude, longitude, time) and four-dimensional (latitude, longitude, height/depth, time) model fields. These fields include such variables as surface temperature, precipitation, and ocean salinity. Output frequencies can range from minutes to months or years, and the data can represent either instantaneous values or averages over the output period. In total, several hundred fields are output by the CCSM components.

1.5 Post-processed History data

The CCSM is a collection of distinct component models optimized for very high speed multi-processor computing. This results in raw output data streams from each component that does not present the data in the most coordinated or user-friendly manner. While raw history data can be analyzed, the raw data packages have not allowed for easy time series analysis. For example, the atmosphere component puts all the requested variables into one large file at each requested output period. While this allows for very fast model execution, this makes it impossible to analyze time series of individual variables without having to access the entire data volume. The process of transforming the raw CCSM history output into data collections more useful for analysis is called post-processing. This step may involve reformatting the data, deriving new fields from the set of existing data, making averages along any or all of the data dimensions, or sampling the data in different ways. Post-processed history data sets are the main actual CCSM "product."

APPENDIX C

CCSM Data Tools

The CCSM project uses netCDF as its data format and benefits from the large suite of software tools that support this format.

NCAR's Climate and Global Dynamics division hosts the CCSM Support Network at <http://www.cesm.ucar.edu/support> for CCSM data analysis. Instructional workshops are available.

Other tools that have been found to be particularly useful for analysis and visualization of CCSM data are:

Public domain tools:

NCO: The netCDF operators

NCL: The NCAR Command Language
VCDAT: DOE analysis and visualization tools
ncview: A simple netCDF display tool
GrADS: A visualization and analysis tool
FERRET: A netCDF visualization tool

Commercial tools:

IDL

MatLab

etc.