CESM DATA MANAGEMENT and DATA DISTRIBUTION PLAN

Introduction

The Community Earth System Model (CESM) Data Management and Data Distribution Plan documents the procedures for the storage and distribution of data associated with the CESM project. These procedures reflect the approaches, standards, and conventions that coordinate the production, post-processing, distribution and storage of simulation data agreed upon by the CESM Scientific Steering Committee (SSC) and the different CESM Working Groups that comprise the CESM project. The overall goal of this plan is to provide the best possible access and ease-of-use of high-quality CESM data to diverse users within the constraints of available resources.

Key elements of this CESM Data Management and Data Distribution Plan include:

1. Definition of the Major Categories of CESM Data;
2. Ownership Rights and Responsibilities;
3. Data Release Timeline;
4. Retention of CESM Data;
5. Data Format Standards;
6. Metadata Requirements;
7. Case and File Naming Conventions;
8. Data Quality Assurance Procedures;
9. Access to CESM data;
10. Registering and Auditing the Characteristics of CESM Data Users;
11. Distributed CESM Data Repositories;
12. Distribution of CESM Data Products to Non-NSF Entities; and
13. Future changes to the CESM Data Management and Data Distribution Plan.

The current plan supersedes all previous Community Climate System Model (CCSM) Data Management and Data Distribution plans and documents. The procedures below are not intended to be retroactive to currently available data. This plan is also a living document that will evolve as the scope of the project and its customer base changes over time.

The CESM Data Management and Data Distribution Plan

In accordance with the National Science Foundation (NSF) data policy, the National Center for Atmospheric Research (NCAR) mission and the NCAR imperatives (see, in particular, https://ncar.ucar.edu/documents/strategic-plan/2009/ncar-imperatives/3), the CESM project is committed to the timely submission of results from CESM model runs for publication and sharing of the scientific data generated in CESM research activities. Open access to CESM data products is essential to the project. Analysis and interpretation by the broader community promotes scientific discovery, and it leads to new insights into model behavior that feedback into model development efforts. At the same time, however, the efforts of CESM developers and the architects of the scientific simulations performed with the model under the auspices of the CESM Working Groups need to be recognized by providing them a reasonable amount of time for first access to the simulations. It is the intent of this Plan to define the guidelines to meet these objectives. These procedures thus apply to all CESM data created under the auspices of the CESM Working Groups.

1. Categories of CESM Data

For the purposes of this Plan, and for consistency with proposals to the Climate Simulation Laboratory (CSL) computing facilities, CESM data are broadly categorized into Development and Production simulations. The defining characteristic of Production simulations is that they have broad appeal across the climate science community and are thus made available for community access and analysis. Examples include simulations that contribute directly to coordinated national or international modeling activities and "benchmark" simulations that document CESM components and new coupled configurations and capabilities of the model (see below). Both are examples of simulations where the project benefits directly from analysis and interpretation by the broader research community.

The outcome of community analysis can lead to development efforts, such that development and production activities are synergistic and sometimes become blurred. In contrast to Production simulations, however, Development simulations may not necessarily be made available for analysis beyond the Working Group members who produce the runs. They include, for instance: simulations to understand CESM (or component) behavior, document biases, and determine the responsible processes; efforts to improve the representation of processes; and activities to add new capabilities to CESM important for improving simulation fidelity, for new community-based science, and for future releases of the model.

Within these two general categories, CESM simulations can be further grouped as follows:

CESM Production Simulations

CESM Control integrations: Control integrations are typically hundreds to thousands of model years in length, with constant forcing over time, and use a code base that corresponds to a major community release of the CESM. Control integrations (or runs) define the basic long-term climate of the CESM. A Control run needs to be long enough in simulated time for slow climate adjustment processes, such as subsurface water in the land model, to come into balance. Some

processes, such as the deep ocean heat and salinity, may take thousands of years to reach balance. The standard data output usually corresponds to monthly averages, with perhaps daily or higher frequency output over subsets of the Control and perhaps for fewer, select variables. Higher frequency data output for special analyses or to drive regional-scale models may also be provided from Control simulations upon specific request or to meet requirements of nationally- or internationally-coordinated experimental protocols.

CESM Experiment integrations: CESM Experiments are typically tens to hundreds of model years in length (or even much longer; for instance, several millennia in paleoclimate applications) and are usually made with modifications introduced into the control version of the CESM in order to conduct a specific scientific experiment or examine a policy scenario. A CESM Experiment simulation may be a single run or a series of runs. The modifications may be either in the representation of processes in the model, the forcing data, or both. Often, a given experiment 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 imposed modifications from the Control case. In this circumstance, the ensemble of runs constitutes the data for the particular Experiment.

CESM Development Simulations

CESM Evaluation integrations: CESM Evaluation integrations are short runs (typically model days to years) to examine specific model behavior, such as the response to changes in the representation of physical processes or boundary conditions, or to validate a port of the code to a new computing platform.

CESM Testing integrations. CESM Testing integrations are very short runs (typically model days to years) carried out to cover three traditional and well-defined tasks: verification of model functionality; performance tuning; and debugging.

These different categories of runs produce output data of various forms as defined in Appendix A. In the procedures below, the different categories of runs above will be designated simply as Control, Experiment, Evaluation, and Testing.

2. Ownership Rights and Responsibilities for CESM Data

CESM output data created with CESM project resources are considered to be owned by either the SSC or the CESM Working Group(s) that designed and carried out the runs. The Chair of the SSC and the Co-chairs of the Working Group(s) are designated as the Principal Investigators (PIs) for the simulations, acting on behalf of the SSC and the Working Group(s) respectively. The PIs are responsible for ensuring the simulation data have been checked for quality control, have been appropriately documented, and are made accessible, as appropriate, following the data release procedures documented below.

Control: These data are the property of the SSC, with the Chair of the SSC serving as the PI. CESM Control integrations are documented on the CESM Experiments and Output Data Web

page (http://www.cesm.ucar.edu/experiments/). This includes a description of the run and pointers to the validation plots and data files from the Control.

Experiment: These data are initially the property of the CESM SSC, Working Group or Working Groups that design and conduct the Experiment, with the SSC Chair and/or the Co-chairs of the Working Group(s) serving as the PIs. Examples of CESM Experiment simulations that fall under the purview of the SSC may include those performed in support of either national or international assessment activities or other coordinated model protocols. Most other examples of CESM Experiments fall under the purview of one or more CESM Working Groups. Transfer of access to a broader community of these CESM Experiment simulations takes place over time as defined by the data release procedures below. Like Control integrations, CESM Experiment integrations are documented on the CESM Experiments and Output Data Web page. This documentation includes a description of the run and pointers to the validation plots and data files.

Evaluation: The data from CESM Evaluation integrations are the property of the PIs of the relevant Working Group(s) and are considered to be non-public, internal tools. CESM Evaluation integrations are only required to be documented to the extent needed by the PIs or Working Group members.

Testing: CESM Testing integrations are treated in the same manner as CESM Evaluation runs.

Procedure

The PIs of CESM data are responsible for ensuring that the data are documented and archived in accordance with the CESM Data Release Timeline (below). This includes documenting the simulation data on the Web, keeping the data on the designated archival storage device for the specified time, and deleting the data once they have become obsolete (see section 4 below). When multiple Working Groups collaborate on Production integrations, the Co-chairs of those working groups serve as PIs, although it is acceptable for the PIs to designate only one Working Group as the primary owner of the data.

3. CESM Data Release Timeline

The intellectual investment and time committed to the design and execution of a CESM Production simulation entitles the PIs, acting on behalf of the SSC and the Working Groups, to the first benefits obtained from the resulting data. Publication of descriptive or interpretive results derived immediately and directly from the Production simulation data is the privilege and responsibility of the PIs. However, to further CESM science objectives, the PIs are encouraged to share their data with colleagues prior to the release deadlines. When the release deadlines for CESM data are reached, the data move into the public domain. In special cases, the CESM Data Release Timeline can be superseded by the SSC. Examples of this could be the involvement of the CESM project in national or international modeling or assessment activities, or other coordinated model experiment protocols, where the timely release of CESM data may be deemed necessary.

The release status of CESM data is characterized as being Protected, CESM Access, or Public. Protected data are owned by the PIs, as defined above. CESM Access data are Protected data that have been made available to all CESM Working Group members. Protected data become CESM Access data, then Public data, by permission of the PIs or through the expiration of the proprietary time period as defined by the CESM Data Release Timeline below. Public data are open to access by the broader community.

Procedure

All CESM data are initially categorized as Protected.

CESM Control data become Public once the Control integration has been validated by the Chair of the SSC.

CESM Experiment data shall be available to members of any CESM Working Group (Access data) *no later* than six months following the conclusion of the Experiment integration.

CESM Experiment data shall become Public as soon as a scientific paper on the results has been submitted by the PIs or one year after the end of the simulation, whichever comes sooner. If a scientific paper is written by a member of a CESM Working Group after the data became designated as Access data, but before the one year deadline or a paper is submitted by the PIs, the PIs are encouraged to make the data Public.

Any person wishing to make use of Experiment data before these dates should communicate directly with the PIs about access to the data. In this circumstance, it is anticipated the PIs will be offered co-authorship of any published results, if they wish.

All Evaluation and Testing data are classed as Protected and remain so unless the PIs decide otherwise.

In special cases, when more timely delivery of CESM data is necessary and/or is in the best interest of the CESM project, the CESM Data Release Timeline can be superseded by a decision of the SSC.

Information regarding CESM Public data availability, including appropriate references and acknowledgments, will appear on the CESM Experiments and Output Data Web page of the CESM Web pages.

4. Retention of CESM Data

The key to management of CESM data is to have Production and Development data stored and distributed via different strategies, with each tailored to suit the different user needs. It can no longer be assumed that storage capacity will grow much faster than the data volumes. The CESM data retention procedure below attempts to strike a balance between the scientific need to retain data from older simulations with the growing cost of doing so in a resource-limited environment. Unlike observational data, model simulation data often become less valuable with time as better

models of higher quality are developed and run. Nevertheless, publication of scientific analyses of CESM Control and Experiment integrations continue years after the data were generated. Accordingly, data from CESM Control and Experiment integrations shall be preserved for specified time periods to allow extraction of the scientific content.

CESM data at NCAR will be retained under the guidelines of this data stewardship plan. The PIs of the data are responsible for the stewardship. A similar policy for CESM data held at other sites is encouraged. Sites holding CESM Control or Experiment integration data that are Public should give the CESM SSC the option of archiving these data at NCAR before deletion. A similar courtesy should be provided to the SSC for Public data that are slated for deletion.

Procedure

Development data: Output data will be stored at NCAR (on the High Performance Storage System, or HPSS) for a period of 36 months after creation, at which point they will be removed, unless retention is requested from the relevant Working Group Co-chairs. Should storage resources become an issue, the SSC reserves the right to intervene.

Production data: Output data will be stored at NCAR for a period of four years. These data will then be gradually reduced to 50% of their initial volume over a period of three additional years, based on usage and anticipated demand. This data level will be maintained for three more years. Afterward, each CESM Working Group will determine data to be removed and at what rate, as the archived data are gradually reduced to an acceptable level, as determined by data archiving costs at the time.

A designated data manager will be responsible making sure the aforementioned procedures are implemented and followed by the PIs.

5. CESM Output Data Format Standards

Standard data and metadata formats are crucial for the automated analysis necessary to efficiently interact with large data collections. CESM uses netCDF as the standard data format for all output data. The use of netCDF makes CESM output data readily accessible to a variety of existing graphics and analysis packages.

Procedure

All CESM components will create netCDF output history data.

All post-processed CESM data will be made available in netCDF format.

6. CESM Metadata Requirements

In the broadest sense, metadata are simply “structured data about data”, describing important attributes of an information resource. Metadata for CESM data are carried in the header section of the model output netCDF files.

Procedure

All CESM netCDF output history data will comply, to the extent possible, with the Climate and Forecast (CF) metadata convention.

7. CESM Output Filename Conventions

The CESM project has adopted naming conventions for output files. CESM output files fall into two broad categories: (1) those generated by the CESM component models at run-time (i.e., model output data); and (2) those created by post-processing of the run-time files (i.e., post-processed data). The naming conventions are described on the CESM Web page: http://www.cesm.ucar.edu/models/cesm1.0/filename_conventions_cesm.html.

Procedure

CESM Production and Development simulations will conform to the output filename conventions.

8. CESM Data Quality Assurance Procedures

Primary responsibility for quality control of CESM data products lies with the PIs overseeing the model integration. Quality control of CESM data is carried out by individual component Working Groups or by collaborations of several component Working Groups.

Procedure

PIs are responsible for maintaining the quality and correctness of their CESM data. The PIs should address questions raised by the researchers using these data as quickly as possible.

9. Access to CESM Data

Web technologies allow for the efficient discovery and access of CESM data. The CESM Working Groups have been very active in establishing Web portals to CESM data subsets, both within NCAR and through Department of Energy (DOE) collaborations. Currently, the primary distribution system for CESM data is the Earth System Grid for Enabling Technologies (ESG-CET). The UCAR Graphical Information System (GIS) portal may also be another online source of CESM data.

Procedure

In general, output data from CESM Development integrations will be made available only to the Working Group members that are directly involved in the Development experiments. For Working Group members that do not have access to the NCAR HPSS, these data will be made available via the ESG-CET.

For Production integrations, to maximize the ease-of-access and value of the data to the scientific community, all CESM Public data shall be made available via the ESG-CET (http://www.earthsystemgrid.org/). The registration process through ESG-CET will permit the assembly of information on the users and use of the CESM Public data.

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

Restart and initial data will not be made publicly available. They may be obtained upon request from the Working Group liaison.

10. Registering and Auditing the Characteristics of CESM Data Users

Procedure

To measure the CESM project's contribution to and impact on the broader scientific community, registration information will be collected from users downloading Public CESM data. This information will describe the user's name, contact information, institutional affiliation, and intended use of the data. Summaries of this CESM registration information will be reported to the CESM SSC and the CESM planning agencies to demonstrate how the CESM project is serving the scientific community.

11. Distributed CESM Data Repositories

CESM Production integrations are 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 CESM data. It will be necessary to coordinate CESM data storage, discovery, and access policies among the various sites where CESM data will be archived. This is particularly important for Public data.

Procedure

Data produced by the CESM will be stored, managed, and distributed by the data archive center appointed by the entity sponsoring the CESM run that produces the data. CESM data created at NCAR under NSF support will be archived on the NCAR HPSS. CESM data generated at non-NCAR facilities should be archived at either the site of generation or its associated data archive center. CESM data created at non-NCAR sites may be archived on the NCAR HPSS if prior arrangements have been made with both CESM and NCAR's Computational and Information Systems Laboratory (CISL) management.

12. Distribution of CESM Data Products to Non-NSF Entities

Procedure

Data will be made available for users who are not CESM collaborators at the marginal cost of making and shipping the copies (not the full cost of data production and archive maintenance). However, for large data orders, the CESM SSC reserves the right to make special policies and perhaps ask for a data exchange that is beneficial to both sides.

13. Changes to this CESM Data Management and Data Distribution Plan

Recognizing that climate modeling is an evolving field, the CESM SSC reserves the right to update the guidelines and procedures outlined in this document. Any changes will be made with respect for the resource needs of PIs with regard to the processing and distribution of information. When changes in data policy require substantial increases in equipment, supplies, or personnel, pre-existing investigations will not be expected to comply with the changes.

APPENDIX A.

CESM Data

During the course of an integration, the CESM produces three distinct output data streams: plain-text log information, restart data, and history data. After a CESM run finishes, the raw history data are 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, raw binary, or ascii
b. Output Log Data	(small)	Plain text files
c. Output Restart Data	(medium)	netCDF and 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

Types of CESM Output Data

a. Input Initial and Boundary Condition Data

CESM runs are typically started using initial data that represent a known or idealized climate state for each CESM 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.

b. Output Log Data

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

c. Output Restart Data

The CESM restart data are primarily netCDF files with the presence of some raw binary files. The restart data contains sufficient information for the CESM to restart exactly when given a suitable model code base. Restart data are usually output every two years, although more frequent output may be necessary as the model resolution increases.

d. Output Raw History Data

Raw CESM history data are the original, high-volume data streams directly created by each CESM component during the course of a CESM integration. The history data consist 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, for instance, instantaneous values, extreme values, or average values over the output period. In total, several hundred fields are output by the CESM components.

e. Post-Processed History Data

Post-Processed CESM data are all other CESM data products. The CESM 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 do 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 CESM 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 existing data, making averages along any or all of the data dimensions, or sampling the data in different ways. These post-processed data are usually condensed into lower volume collections that are more portable and easy to use. The Post-Processed CESM data are the most useful data for climate analysis researchers and represent the desired results of CESM experiments.

APPENDIX B.

CESM Data Tools

The CESM project uses netCDF as its output data format and benefits from the large suite of software tools that support this format. Unidata (http://www.unidata.ucar.edu) has an extensive listing of software that can manipulate netCDF data. Each model component maintains its own post-processing utility suite that can be accessed from the release code repository. Component post-processing utilities are currently provided only as a service to the community, so only informal community support is provided via the CESM bulletin board. See http://www.cesm.ucar.edu/models/cesm1.0/model_diagnostics/ for more information.

APPENDIX C.

The CESM Data User Community

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

- Scientists at universities, federal laboratories, and NCAR;
- CESM Working Groups performing Production and Development integrations;
- Impact analysts;
- National assessment programs;
- Intergovernmental Panel on Climate Change (IPCC) Data Distribution Center;
- Coupled Model Intercomparison Project/Paleoclimate Model Intercomparison Project (CMIP/PMIP);
- Policymakers;
- Other modeling groups using CESM data as forcing input to their models (e.g., regional climate models); and
- Industry.

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