Complete CCSM Data Management Plan

CCSM Data Management Plan (29 Aug 2003)

Index

Note: This is a draft version. Any policies described here NOT yet been reviewed or approved by the CCSM SSC. These are strawman plans and policies, please feel free to comment on them and suggest any modification you see fit.

Lawrence Buja
Gary Strand

This document outlines issues, and provides reference material, for initiating a comprehensive plan involving the management of data produced by the CCSM. The issues cover a wide variety of topics, from the treatment of raw CCSM component output (file naming convention, storage location and organization, transfer requirements, etc.), to access by end users (access policies, user requirements, cataloging, etc.), to collaboration with cutting edge scientific data management strategies (Earth System Grid, etc.).

1. Introduction

The primary goal of the CCSM project is the development and continuous improvement of a comprehensive climate modelling system to be used to understand and predict the behavior of the Earth's climate system. The CCSM project focuses the efforts of a large scientific community to develop and

CCSM Data Management Plan (29 Aug 2003)

Index

1. Introduction
2. CCSM Data Management Plan
3. The CCSM Data User Community
    a. User Community Identification
    b. User Requirements
4. Strawman Plans and Policies
    a. CCSM Data Access
    b. Data Repositories for CCSM data
    c. Online Access to CCSM data
    d. CCSM Quality Control
    e. CCSM Data Retention
    f. Long-term Stewardship of CCSM Data
    g. CCSM data and metadata requirements

Appendices

5. CCSM Run Categories
    a. CCSM Control Runs
    b. CCSM Experiment Simulations
    c. CCSM Validation runs
    d. CCSM Test runs
6. CCSM Datasets
    a. Printed Output
    b. Restart Data
    c. Raw History Data
    d. Postprocessed history data
    e. CCSM Data Output Volume
7. CCSM Data tools

Note: This is a draft version. Any policies described here NOT yet been reviewed or approved by the CCSM SSC. These are strawman plans and policies, please feel free to comment on them and suggest any modification you see fit.

Lawrence Buja
Gary Strand

This document outlines issues, and provides reference material, for initiating a comprehensive plan involving the management of data produced by the CCSM. The issues cover a wide variety of topics, from the treatment of raw CCSM component output (file naming convention, storage location and organization, transfer requirements, etc.), to access by end users (access policies, user requirements, cataloging, etc.), to collaboration with cutting edge scientific data management strategies (Earth System Grid, etc.).

1. Introduction

The primary goal of the CCSM project is the development and continuous improvement of a comprehensive climate modelling system to be used to understand and predict the behavior of the Earth's climate system. The CCSM project focuses the efforts of a large scientific community to develop andmaintain a stable scientific modeling environment and to provide large climate datasets to the US research and climate assessment community. CCSM data products must be visible and easily accessible to these communities for the CCSM project to be effective.

The Climate System Model version 1 (CSM1) coupled together distinct models simulating each of the four major processes of the Earth's climate systems: atmosphere, ocean, sea-ice and the land surface. Each of these models had been developed separately within their respective scientific communities and the resulting output datastreams tended to be very different from each other. The need to easily intercompare data from the various components of the CSM1 runs provided the motivation to unify the formats and metadata conventions of the output data. The sole CSM1 data archive was the NCAR Mass Storage System, with modest amounts of popular CSM1 data being mirrored on public CGD web servers. The primary audience for CSM1 data products was a broad, university-based, research community.

In the same time period, the Parallel Climate Model (PCM) project developed and applied a similar coupled climate model for DOE climate assessment and prediction studies. PCM's tight focus on performance portability and the ability to tap into existing DOE data management systems made PCM a very effective climate modeling application.

Recently the CSM and PCM projects were merged. The title "Community" was added to the project name to reflect the project's broader, interagency, scope. In this new environment which spans NSF, DOE and NASA, the data management and integration needs for CCSM and PCM have evolved as well.

2. CCSM Data Management Plan

A CCSM data management plan is needed to outline issues, and provide reference material for initiating a comprehensive strategy for the management of data produced by these models. The issues involved cover a wide variety of topics, from the treatment of raw model output, to access by end users, and including collaboration with external data management strategies. Although some of these issues are already under consideration, it remains evident that CCSM should develop a more organized and comprehensive plan to address the expected array of data management issues. Thus, the first recommendation arises:
PRIMARY RECOMMENDATION: A CCSM Data Management Group (CDMG) be formed immediately to begin to identify and priorize the needs of the community who will generate, access, or use CCSM output data. The group should be composed of core CCSM scientific and software engineering staff, as well as members of the community, but should be kept to a small, manageable size. Data management policies developed by this group should be approved by the SSC.
The task of the proposed CCSM Data Management Group will be to develop and implement a data management strategy for CCSM data. One of the primary success criteria for the CCSM project is the extent to which CCSM data is used by the climate change community for research and assessment studies. Accordingly, the overall goal of CCSM Data Management is to provide the best possible access and ease of use of CCSM data to the US scientific community.
Key elements of a draft CDMG data management plan include

- coordinating the distributed CCSM data archives across the NSF, DOE and NASA d ata centers,
- coordination of timely post-processing and data presentation when a CCSM run is completed.
- enabling fast and easy web access to CCSM data
- providing procedures for quality control, analysis and validation of CCSM data ,
- definition of a lifecycle for CCSM data and provision for the long-term stew ardship and stability of CCSM data.
- registering and auditing the characteristics of CCSM data users
  Other details such as coordination of CCSM data and metadata formats and providing tools for translating CCSM data into other popular formats will be addressed. For each element in the draft data management plan, specific policies will be proposed to support the various elements.

3. The CCSM Data User Community

3.1 User Community Identification

Users of CCSM data span a wide range of interests. In many cases, these users will have unique data requirements. An incomplete list includes:

- CCSM Scientists at participating universities, federal laboratories ORNL, NERS C etc) and NCAR.
- CCSM developers
- CCSM production groups performing integrations
- Impact analysts
- The National Assessment program
- The IPCC Data Distribution Center
- CMIP
- Policy makers
- Other modelling groups using CCSM data as forcing input to their odels (region al modelling, for example)
- Others
  The activities of this set of users can be combined into a reasonably small set of common activities - access to the data, diagnostics of CCSM performance (scientifically and computationally), and various types of analysis. A goal of the CCSM Data Management Plan is to meet the needs of these users so that these activities can be accomplished reasonably easily and efficiently.

3.2 User Requirements

What are the requirements of the users?
While many of the requirements of the users are known from past experiments, these requirements not formally documented. It is possible that previous data usage from CSM1 can be used to infer user requirements. The addition of DOE funded climate change researchers within the Climate Change Working Group brings an additional set of requirements.
How should user's requirements be solicited? In November 2001 the SSC requested that more information concerning the weekly CCSM scientists