Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 558–566

Procedia Computer Science

International Conference on Computational Science, ICCS 2011

Community Earth System Model Data Management: Policies and Challenges

Gary Strand*

National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000 USA

Abstract

The data management plan of the Community Earth System Model (CESM)[1] from the National Center for Atmospheric Research (NCAR) is given historical context and its policies, definitions, and features are detailed. The drivers of CESM data management are discussed, including the upcoming Coupled Model Intercomparison Project 5 (CMIP5), the ongoing Earth System Grid (ESG) project, and the strategies to address these drivers are mentioned. Future plans and strategies to address CESM data management needs and requirements are noted. The significant challenges resulting from the use of CESM output in the areas of metadata, preservation, curation, provenance, and other aspects of data management are considered.

Keywords: NCAR, CESM, CCSM, data management, climate model, ESG, CMIP

1. Introduction

The Community Earth System Model (CESM)[2, 3] from the National Center for Atmospheric Research (NCAR) is a state-of-the-art Earth system model. It comprises components representing various processes important in the climate system, such as atmospheric chemistry, radiation and dynamics, ocean biogeochemistry and dynamics, sea and land ice, dynamic vegetation, the carbon/nitrogen cycle, and others. CESM is a true community model; its development is the result of close collaboration between NCAR scientists and universities worldwide, as well as scientists from other US national laboratories and its source code is freely available. CESM and its predecessors have been used by thousands of climate researchers worldwide for their scientific analyses of the many aspects of the global climate system. CESM has also participated in national and international assessments and has been used by other scientific communities, university researchers, policymakers, government agencies, and private citizens (see Figure 1). NCAR has always strived to make the output from the simulations as publicly available as practical and as a result they have been widely used. The requirements from scientists, assessments, and this broad and global user community of CESM output present significant challenges in all areas of data management: access, distribution, metadata (information about the data), data standards, curation (who owns data and is responsible for maintenance and support), provenance (how the data were created), and more. These extensive requirements have contributed to the development of CESM through the careful consideration and application of standards, conventions, and best practices. This paper uses the terms “output” and “data” interchangeably, even if in a technical sense climate model output are not “data” because they are not measurements of observed phenomena.

 Selection and/or peer-review under responsibility of Prof. Mitsuhisa Sato and Prof. Satoshi Matsuoka
doi:10.1016/j.procs.2011.04.058

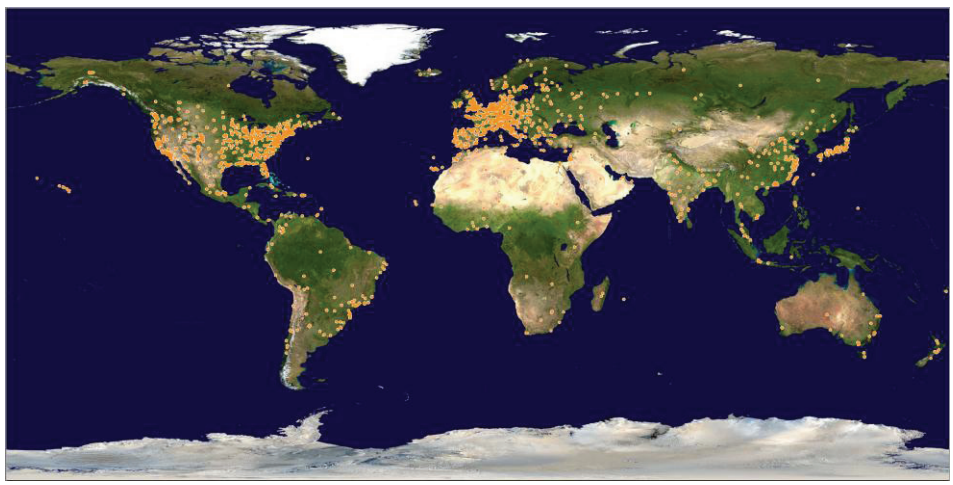

Figure 1: Geographic distribution of the CESM user community – each orange dot can represent many researchers and users of CESM data. *(Background image courtesy NASA; user locations from ESG-II metrics).*

2. Historical background

CESM is the latest in a series of NCAR-based climate models, the lineage of which extends back several decades. Until the early 1990s, the management of model output was informal, and access was limited to a handful of scientists, most located at NCAR. The ability to distribute the data was constrained. Electronic distribution via the internet did not yet exist, and the user community was mostly limited to the scientists who developed and implemented the models of the time.

Beginning in the mid-1990s, two climate models - the Parallel Climate Model (PCM), funded by the US Department of Energy (DOE), and the Climate System Model 1.0 (CSM1), funded by the US National Science Foundation (NSF), existed at NCAR. Both models required that their output be made as publicly available as practical. For PCM this was accomplished by transferring the model output from the various computing centers at which PCM was executed to the archival system at the National Energy Research Supercomputing Center (NERSC). NERSC's archival system had an open access policy, so individuals who wished access to the PCM output could request an account to access the archive. After the PCM principal investigators had examined the application, access was granted. This process resulted in the distribution of PCM data to a community of roughly 100 users between 1999 and 2004. Similarly, access to CSM1.0 output required an account at NCAR to access the archival system there. The online disk resources for even the modest output volumes of the era were not available.

For both models these distribution practices were limited because of the need for manual registration and knowledge of the software to interact with the archive system. However, these access methods did result in the first distribution of climate model output to a much broader community. These experiences also provided the initial indications that NCAR climate modeling efforts would need to formalize their data management policies, because these user communities were not intimately tied to NCAR's climate modeling efforts. The user communities needed assistance to gain understanding of the meaning, proper use, and applicability of the data. Support of the community for data issues became more important.

In the late 1990's the limitations of these techniques of data distribution became apparent as the user community grew and their needs expanded. For PCM, these changes led to the decision to capitalize on the Earth System Grid I project. ESG-I was conceived in 2000 to utilize Data Grid technologies to move and replicate terascale datasets efficiently and rapidly between computing centers. These abilities were a natural fit for PCM data. The new project was christened Earth System Grid-II[4, 5] (ESG-II) in late 2000 and was intended to provide the same abilities for PCM output. ESG-II was designed to remove the requirement that a user of PCM output needed separate accounts to download and use the data because ESG-II was web-based. With ESG-II, a registered user could browse the available data, search for the data of interest, and make a simple request to download the data. ESG-II made the