Geophysical Research Abstracts
Vol. 12, EGU2010-8357-1, 2010
EGU General Assembly 2010

The CMIP5 archive architecture: A system for petabyte-scale distributed archival of climate model data

Stephen Pascoe (1), Luca Cinquini (2), and Bryan Lawrence (1)

(1) British Atmospheric Data Centre (Stephen.Pascoe@stfc.ac.uk), Rutherford Appleton Laboratory, UK, (2) National Center for Atmospheric Research, Boulder, USA

The Phase 5 Coupled Model Intercomparison Project (CMIP5) will produce a petabyte scale archive of climate data relevant to future international assessments of climate science (e.g., the IPCC's 5th Assessment Report scheduled for publication in 2013). The infrastructure for the CMIP5 archive must meet many challenges to support this ambitious international project. We describe here the distributed software architecture being deployed worldwide to meet these challenges.

The CMIP5 architecture extends the Earth System Grid (ESG) distributed architecture of Datanodes, providing data access and visualisation services, and Gateways providing the user interface including registration, search and browse services. Additional features developed for CMIP5 include a publication workflow incorporating quality control and metadata submission, data replication, version control, update notification and production of citable metadata records. Implementation of these features have been driven by the requirements of reliable global access to over 1Pb of data and consistent citability of data and metadata.

Central to the implementation is the concept of Atomic Datasets that are identifiable through a Data Reference Syntax (DRS). Atomic Datasets are immutable to allow them to be replicated and tracked whilst maintaining data consistency. However, since occasional errors in data production and processing is inevitable, new versions can be published and users notified of these updates. As deprecated datasets may be the target of existing citations they can remain visible in the system.

Replication of Atomic Datasets is designed to improve regional access and provide fault tolerance. Several datanodes in the system are designated replicating nodes and hold replicas of a portion of the archive expected to be of broad interest to the community. Gateways provide a system-wide interface to users where they can track the version history and location of replicas to select the most appropriate location for download.

In addition to meeting the immediate needs of CMIP5 this architecture provides a basis for the Earth System Modeling e-infrastructure being further developed within the EU FP7 IS-ENES project.