Geophysical Research Abstracts
Vol. 14, EGU2012-10569, 2012
EGU General Assembly 2012

A Collaborative Data Management Infrastructure for Climate Data Analysis

S. Kindermann, F. Schintke, and B. Fritzsch
German Climate Computing Centre (DKRZ), Hamburg, Germany (kindermann@dkrz.de)

The volume of climate data, especially data generated by climate models is growing exponentially. Typical climate data analysis workflows have to access large portions of this volume e.g. to make statements across multiple climate model runs.
These portions are no longer manageable by and for individual scientists at their home institutions. This is already true for the petabytes of CMIP5 data available for data analysis activities e.g. in the course of the preparation of the next IPCC report.
Thus there is a strong need for collaborative infrastructures supporting users in finding, accessing, sharing and processing climate data for data analysis activities.
Beyond basic data management such an infrastructure can provide higher level services for data life cycle management such as metadata generation, persistent data product identification, data description and sharing.
The Collaborative Climate Community Data and Processing Grid infrastructure (C3Grid) provides a national infrastructure for climate data access and processing. It is now evolving from a prototype to a production infrastructure providing valuable services to users.
To enable data analysis workflows based on data stored in international data federations like the CMIP5 / ESGF data federation and the European IS-ENES data federation, the C3Grid data management infrastructure was extended.
We describe the architecture and implementation of the new C3Grid infrastructure with special emphasis on generic infrastructure services supporting users with respect to climate data life cycle management:
data discovery, (parallel) data access to high volume climate archives, data pre-processing near the data center, persistent data product identification, data description and publication.
These services are essential for users to be able to manage their high-volume data analysis activities and support data provenance tracking.
We will put special emphasis on those aspects supporting users to manage complex multi model analysis tasks of e.g. CMIP5 data and sharing of their analysis results.