CM-DataONE: A Framework for collaborative analysis of climate model output

Hao Xu (1), Yuqi Bai (1), Sha Li (1), Wenhao Dong (1), Wenyu Huang (1), Shiming Xu (1), Yanluan Lin (1), Bin Wang (1,2)

(1) Ministry of Education Key Laboratory for Earth System Modelling, Center for Earth System Science (CESS), Tsinghua University, Beijing, China (xuhao13@mails.tsinghua.edu.cn), (2) State Key Laboratory of Numerical Modelling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences,Beijing, China

CM-DataONE is a distributed collaborative analysis framework for climate model data which aims to break through the data access barriers of increasing file size and to accelerate research process. As data size involved in project such as the fifth Coupled Model Intercomparison Project (CMIP5) has reached petabytes, conventional methods for analysis and diagnosis of model outputs have been rather time-consuming and redundant. CM-DataONE is developed for data publishers and researchers from relevant areas. It can enable easy access to distributed data and provide extensible analysis functions based on tools such as NCAR Command Language, NetCDF Operators (NCO) and Climate Data Operators (CDO).

CM-DataONE can be easily installed, configured, and maintained. The main web application has two separate parts which communicate with each other through APIs based on HTTP protocol. The analytic server is designed to be installed in each data node while a data portal can be configured anywhere and connect to a nearest node. Functions such as data query, analytic task submission, status monitoring, visualization and product downloading are provided to end users by data portal. Data conform to CMIP5 Model Output Format in each peer node can be scanned by the server and mapped to a global information database. A scheduler included in the server is responsible for task decomposition, distribution and consolidation. Analysis functions are always executed where data locate. Analysis function package included in the server has provided commonly used functions such as EOF analysis, trend analysis and time series. Functions are coupled with data by XML descriptions and can be easily extended. Various types of results can be obtained by users for further studies.

This framework has significantly decreased the amount of data to be transmitted and improved efficiency in model intercomparison jobs by supporting online analysis and multi-node collaboration. To end users, data query is therefore accelerated and the size of data to be downloaded is reduced. Methodology can be easily shared among scientists, avoiding unnecessary replication. Currently, a prototype of CM-DataONE has been deployed on two data nodes of Tsinghua University.