Geophysical Research Abstracts,
Vol. 11, EGU2009-11815, 2009
EGU General Assembly 2009

Assembling Large, Multi-Sensor Climate Datasets Using the SciFlo Grid Workflow System

B. Wilson, G. Manipon, Z. Xing, and E. Fetzer
JET PROPULSION LABORATORY, 4800 Oak Grove Dr., PASADENA, United States (Brian.Wilson@jpl.nasa.gov)

NASA's Earth Observing System (EOS) is an ambitious facility for studying global climate change. The mandate now is to combine measurements from the instruments on the "A-Train" platforms (AIRS, AMSR-E, MODIS, MISR, MLS, and CloudSat) and other Earth probes to enable large-scale studies of climate change over periods of years to decades. However, moving from predominantly single-instrument studies to a multi-sensor, measurement-based model for long-duration analysis of important climate variables presents serious challenges for large-scale data mining and data fusion. For example, one might want to compare temperature and water vapor retrievals from one instrument (AIRS) to another instrument (MODIS), and to a model (ECMWF), stratify the comparisons using a classification of the "cloud scenes" from CloudSat, and repeat the entire analysis over years of AIRS data. To perform such an analysis, one must discover & access multiple datasets from remote sites, find the space/time "matchups" between instruments swaths and model grids, understand the quality flags and uncertainties for retrieved physical variables, assemble merged datasets, and compute fused products for further scientific and statistical analysis. To meet these large-scale challenges, we are utilizing a Grid computing and dataflow framework, named SciFlo, in which we are deploying a set of versatile and reusable operators for data query, access, subsetting, co-registration, mining, fusion, and advanced statistical analysis.

SciFlo is a semantically-enabled ("smart") Grid Workflow system that ties together a peer-to-peer network of computers into an efficient engine for distributed computation. The SciFlo workflow engine enables scientists to do multi-instrument Earth Science by assembling remotely-invokable Web Services (SOAP or http GET URLs), native executables, command-line scripts, and Python codes into a distributed computing flow. A scientist visually authors the graph of operation in the VizFlow GUI, or uses a text editor to modify the simple XML workflow documents. The SciFlo client & server engines optimize the execution of such distributed workflows and allow the user to transparently find and use datasets and operators without worrying about the actual location of the Grid resources. The engine transparently moves data to the operators, and moves operators to the data (on the dozen trusted SciFlo nodes). SciFlo also deploys a variety of Data Grid services to: query datasets in space and time, locate & retrieve on-line data granules, provide on-the-fly variable and spatial subsetting, perform pairwise instrument matchups for A-Train datasets, and compute fused products. These services are combined into efficient workflows to assemble the desired large-scale, merged climate datasets.

SciFlo is currently being applied in several large climate studies: comparisons of aerosol optical depth between MODIS, MISR, AERONET ground network, and U. Michigan's IMPACT aerosol transport model; characterization of long-term biases in microwave and infrared instruments (AIRS, MLS) by comparisons to GPS temperature retrievals accurate to 0.1 degrees Kelvin; and construction of a decade-long, multi-sensor water vapor climatology stratified by classified cloud scene by bringing together datasets from AIRS/AMSU, AMSR-E, MLS, MODIS, and CloudSat (NASA MEASUREs grant, Fetzer PI). The presentation will discuss the SciFlo technologies, their application in these distributed workflows, and the many challenges encountered in assembling and analyzing these massive datasets.