NASA Remote Sensing Data in Earth Sciences: Processing, Archiving, Distribution, Applications at the GES DISC

G. Leptoukh

NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA – Gregory.Leptoukh@nasa.gov

Abstract – The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) is one of the major Distributed Active Archive Centers (DAACs) archiving and distributing remote sensing data from the NASA's Earth Observing System. The GES DISC has developed various value-adding processing services. A particularly useful service is data processing at the DISC (i.e., close to the input data) with the users' algorithms. Moving from tape to disk archives has provided additional flexibility to extend online Web-based services. Shifting processing and data management burden from users to the GES DISC, allows scientists to concentrate on science, while the GES DISC handles the data management and data processing at a lower cost.

Keywords: Remote sensing, Data distribution, Data tools, Value-added services

1. INTRODUCTION

With the launch in 1999 of the first Earth Observing System (EOS) satellite Terra, the National Aeronautics and Space Administration (NASA) began the first satellite-based observing system to offer integrated measurements of the Earth's processes. It consists of a science component and a data system supporting a coordinated series of polar-orbiting and low-inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans. The largest scientific data system in the world -- the Earth Observing System Data and Information System -- currently collects environmental measurements from more than 30 satellites, such as NASA's Terra, Aqua, and Aura. The volume of data about our environment collected by these constant observers is truly amazing. The NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) is home of the GES Distributed Active Archive Center (DAAC), one of eight NASA DAACs that offer Earth science data, information and services to research scientists, applications scientists, applications users, and students.

Since the GES DISC DAAC archives data from various Earth observing satellite missions along with data from various field campaigns it processes more than 1 TB of data a day. Overall archive volume is currently 1.75 PB (PetaByte = 10**15 Byte). Peak daily numbers are: 3.3 TB for processing, 1.7 TB for archiving (we don't archive some intermediate products), with daily distribution to the public at peak 1.75 TB. Our record distribution to the public in number of granules is 62000. Total archive volume at the GES DISC is 1.75 PB.

This unprecedented volume creates new never before observed challenges with EOS data usability and has forced the GES DISC to developed various value-added Services. We distinguish between services provided by the GES DISC (running on GES DISC computers) and tools users run on their computers. In the former case, all the data processing is done next to the archive, while in the latter case, users have to retrieve data to their computers and use tools provided by the GES DISC or their own code to process data on their computers.
In the current article, we mostly describe services - processing options next to the input data: Algorithm running within the main production system, the Simple, Scalable, Script-based Science Processor for Missions (S4PM), Data mining within Near-line Archive Data Mining (NADM) system, On-demand subsetting (S4PM-based), On-the-fly subsetting (FTP-based), and gridded data on-line visualization and analysis (Giovanni). At first, we describe processing options from user perspective. Next, we describe processing systems supporting these options. Then, we describe archiving and distribution options recently employed or planned at the GES DISC.

2. SERVICES: PROCESSING OPTIONS

In order to assist users in the full utilization and manipulation of data and products the GES DISC has developed and implemented several applications that manipulate data on DISC servers. Below is brief explanation to some of the main services (processing options) and the Web access addresses for further information:

http://disc.gsfc.nasa.gov/services/

Depending on user algorithm maturity, input data volume, network requirements, users may choose different options below.

2.1 Main production system (S4PM)

This highest throughput option is to run algorithms within the main GES DISC production system, powered by Simple, Scalable, Script-based Science Processing for Missions (S4PM) system. In addition to generating standard products, the GES DISC routinely runs various spatial, channel or parameter subsets for calibration & validation purposes through the S4PM production system. S4PM make it possible to generate inputs for other processing systems, for other instrument teams or Principal Investigators.

In order for an algorithm to run within the main production system, it needs to be well defined and tested. Code is either delivered or developed locally, integrated through the standard Science Software Integration and Testing (SSI&T) process, and run as a Product Generation Executables (PGE) within forward and reprocessing data flows (no retrieval from tapes!)

http://disc.gsfc.nasa.gov/techlab/s4pm/index.shtml

2.2 Data mining via NADM

The NADM is the GES DISC web portal to the EOS data pool, a disk cache that holds EOS data for an extended period of time. The NADM web portal enables registered users to submit and execute data mining algorithm codes on the EOS data in the data pool. The generated mined data products can then be transferred via the network to the user.

The NADM is used for experimental (non-standard) data mining algorithm development and testing environment. The main functions include automated algorithm upload, building of algorithm code and testing environment, automated generation of data mining processing strings and interactive processing of data files limited to data and products hold online.

http://disc.gsfc.nasa.gov/services/nadm/index.shtml

2.3 On-Demand Subsetting (ODS)

ODS is for users in need for specific channels/parameters or spatial regions. Subsetting criteria are entered through the GES DISC web interface (a.k.a. WHOM) for granules selected. The order is submitted to the archive system and the selected data granules are retrieved from tapes and staged for subsetting within the S4PM system. The resulting subset outputs are delivered through standard means within 24-48 hours. The ODS system requires users to process output subsets themselves and are limited by a throughput of the subsetting system.

For an example, start with:
http://disc.gsfc.nasa.gov/data/dataset/,
and follow AIRS or MODIS links down to on-the-fly subsetting options.

2.4 On-The-Fly (OTF) Subsetting

OTF is one of the most popular options for subsetting. It extracts original data granules that are in the online archive and utilizes WU-FTP processing on download. Users can enter subsetting criteria through WHOM and either download subsets one-by-one by clicking on granules selected or download an FTP script generated by the system for selected granules. This is followed by initiating an FTP session to download all the required subsets. Another option is for users to FTP directly to the online archive, and download subsets by using special extensions. Subsetting capability is limited by CPU power.

For an example, start with:
http://disc.gsfc.nasa.gov/data/datapool/,
and follow AIRS or MODIS links down to on-the-fly subsetting options.

3. PROCESSING SYSTEMS AND FEATURES

3.1 S4PM

The main advantage of S4PM is an ability to automate science processing to the extent that a single operator can monitor all of the processing in an "industrial-size" data processing center. It is also flexible enough to easily add new processing threads or new algorithms to an existing thread with a minimum of effort. Both advantages respond to the need for high usability of archived data and more automation to diminish operational costs. In addition to being scalable up to large processing systems such as the GES DISC, S4PM is also scalable down to small, special-purpose processing threads or strings, the so called S4PM “strings”. It is used to operationally process MODIS Aqua and MODIS Terra data from raw formats up to Level 2. All AIRS operational processing (up to Level 3) is also done within S4PM.

Currently, S4PM is employed in the following operational production:

1. MODIS and AIRS Standard Products (12 processing strings on 4 machines - greater than 1 TB daily):
2. Pre-defined subsets (channel, parameter, or spatial);
3. On-demand processing;
4. MODIS and AIRS subsetting;
5. Special processing;
6. MODIS Direct broadcast processing;
7. SeaWiFS Cloud coverage computation;
8. QA processing (coming soon).

The most recent version of S4PM is the Data Mining Edition.

The S4PM system has demonstrated capability processing up to 3.3 TB at the GES DISC and has been reused outside of the GES DISC several times at the following institutions:

U. South Florida (Direct Readout), NASA/LaRC Atmospheric Sciences Data Center (Mission Support), NPP In-Situ Ground System (Direct Readout). It has been accepted for deployment at Land Processes DAAC. S4PM has been released under NASA Open Source Agreement.

3.2 Near-Archive Data Mining (NADM)

NADM has the following features:

1. Access 70 TB of online data: MODIS, AIRS, Aura (coming);
2. Upload your algorithm using NADM interface;
3. Algorithm language (C, FORTRAN or IDL);
4. Web interface: http://g0dug03u.ecs.nasa.gov/OPS/www/nadm;
5. The most efficient solution is a single system that combines the algorithm upload capabilities of the NADM system onto the S4PM system, the so-called S4PM- Data Mining (S4PM-DM).

The following subsetting algorithms and paradigms have been implemented in through NADM:

1. Custom MODIS L1B channel (on-demand);
2. Custom MODIS ocean L2 parameter (on-demand);
3. Custom MODIS ocean L3 spatial (on-the-fly);
4. HSE-based AIRS L1B and L2 parameter (on-demand and on-the-fly);
5. HDFLook (the original MODIS L3 spatial);
6. Custom MODIS spatial “cookie-cutter”;
7. Custom MODIS subsetter for CERES;
8. HEW (UA subsetter) MODIS L3 ocean (on-demand);
9. Aura spatial subsetter for OMI, MLS, HIRDLS) for the Aura Validation Data Center (AVDC);
10. MODIS L1B (all resolutions) – spatial subsets for AERONET sites;
11. MODIS, AIRS, OMI, MLS subsets for the A-Train cross-processing (coming soon).

3.3 Online visualization and analysis (Giovanni)

Answering to EOS data accessibility issues, the GES DISC has developed the GES-DISC Interactive Online Visualization and Analysis Infrastructure (Giovanni), the underlying infrastructure for a family of Web interfaces for online data analysis. It is very useful to and popular by modelers, global and regional trends researchers, teachers, students, etc.

Giovanni makes gridded data available in a format that anyone can learn to use within minutes and put to work productively for research or applications. With Giovanni and a few mouse clicks, one can easily obtain various remote sensing and model information from around the world. Users can explore and analyze gridded data interactively online without having to download any data. There is no need anymore to learn different complicated data formats, to retrieve and process data. Everything is done via a regular Web browser and intuitive user-friendly interfaces customized for various disciplines.

Giovanni goals:

1. Study various phenomena interactively, ask what-if questions and get back answers to stimulate further investigations;
2. Try various combinations of parameters measured by different instruments;
3. Generate graphs suitable for a publication. One caution: Giovanni is an exploration tool, so users should be aware of data preparation issues and all the caveats of statistical analysis.

Internally, Giovanni provides access to data from multiple locations, server-side processing and support for multiple input data formats: HDF, HDF-EOS, netCDF, GRIB, and binary. For a single geophysical parameter, Giovanni supports multiple output plot types including area, time, Hovmoller, and image animation, along with very popular ASCII output. Giovanni is easily configurable to support customized portals for measurements-based projects or disciplines.

Beyond the basics, Giovanni provides the following parameter intercomparison:

1. Area plot of time averaged parameters - geographical intercomparison between two parameters;
2. Area plot of time averaged parameters difference;
3. Time plot of area averaged parameters - an X-Y time series plot for several parameters;
4. Scatter plot of parameters in selected area and time period - relationship between two parameters geographically;
5. Scatter plot of area averaged parameters - regional (i.e., spatially averaged) relationship between two parameters;
6. Temporal correlation map - relationship between two parameters at each grid point in the selected spatial area;
7. Temporal correlation of area averaged parameters - a single value of the correlation coefficient of a pair of selected parameters;
8. A single file ASCII output with all selected parameters in a format suitable for importing spreadsheets and other programs for off-line analysis.

Giovanni family consists now of the following instances:

1. TOVAS - TRMM and other gridded precipitation data: http://lake.nascom.nasa.gov/tovas;
2. MOVAS - MODIS aerosol related remote sensing and model data: http://lake.nascom.nasa.gov/movas;
3. Ocean-color - SeaWiFS and MODIS Aqua: http://reason.gsfc.nasa.gov/Giovanni;
4. Atmospheric chemistry - starting with TOMS, moving to HALOE and AIRS, later – Aura instruments: http://reason.gsfc.nasa.gov/Giovanni_toms.

The following features will be added to Giovanni:

1. Vertical profile presentation – 2D slices through 3D data;
2. Enhanced parameter intercomparison;
3. Full support of output formats suitable for Geographic Information Systems (GIS), for example GeoTIFF;
4. Lagged temporal correlations;
5. Better support for multi- instrument analyses with smart handling of multiple grids.
6. Errors representation due to missing data and data quality in meaningful ways.

http://disc.gsfc.nasa.gov/techlab/giovanni/index.shtml

4. TOOLS

Tools are applications that users run on their local machines. These tools have been researched, tested, and enhanced if needed to handle the HDF and HDF-EOS data formats. Most of these software tools are freely available and can be downloaded via the Web from the GES DISC Web site. Below is brief explanation to some of the main tools and the Web access addresses for further information:

http://disc.gsfc.nasa.gov/tools/

5. ACHIVING AND DISTRIBUTION

Until recently, archiving at the GES DISC has been handled by robotic tape archives, or silos. Tape silos are expensive to deploy and operate but have the advantage of scaling well for large data volumes. However, the viability of disk based archives have been enhanced by the recent NASA trend toward smaller data systems that service specific, focused communities rather than the general public. GES DISC has developed a disk-based science archive, based on its long experience in archiving requirements, design and operations. The GES DISC's successful implementation of the Simple, Scalable Script-based Science Processor (S4P) points the path as a demonstration of the utility of Radical Simplification in implementing inexpensive, robust, scalable systems. Already all "heritage" data from pre-EOS missions and all field campaigns have been moved to disk archives, powered by another "incarnation" of S4P idea, the S4PA (A stands for Archive).

As part of the transition to S4PA, we are also downsizing our data distribution on media (tapes, CDs, DVDs). With the progress of the Internet, larger volumes of data can be transmitted by networks. Instead of sending TB of data on media, this transition allows us to concentrate our resources on value-added services, in order to reduce volume of data going to users. Users are getting only those bits of data or information needed for their research or

applications, not the original mountains of data they have to mine through to extract what they need.

Another potential option to improve distribution is to establish mirror distribution nodes in different regions. This way, large volumes of data covering specific regions go only to few selected regional centers with robust network connection, while individual users get data from these centers. For example, remote sensing satellite data over Russia can be distributed to few Russian centers for consequent dissemination to Russian institutions.

6. CONCLUSIONS

This paper describes a series of services provided by the NASA GES DISC DAAC to assist in the use and distribution of the immense data/product archived from the NASA's Earth Observing System. In addition to providing data and products, the GES DISC/DAAC has developed various sophisticated value-adding processing services including a configuration-managed algorithm within the main processing stream next to the on-line data storage, and a build-it-yourself data mining system that allow users to build and run their on algorithm in the DISC servers. The center has also made available to user an on-the-fly analysis with simple algorithms embedded into the web-based tools that avoid the user need of downloading unnecessary all the data. Besides, the existing data management infrastructure at the GES DISC supports a wide spectrum of options allowing the shifting processing and data management burden from users to the GES DISC, allowing scientists to concentrate on science, while the GES DISC handles the data management and data processing at a lower cost.

Our main goal is to make it more cost-effective to users the utilization of the existing data management and processing infrastructure at the GES DISC for their own data needs. That is possible because the GES DISC has the capability, infrastructure and personnel that supports a wide spectrum of options, from simple data support to sophisticated online analysis tools.