Restructuring Big Data to Improve Data Access and Performance in Analytic Services Making Research More Efficient for the Study of Extreme Weather Events and Application User Communities

AOGS
15th Annual Meeting
June 3-8 2018

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

Dana M. Ostrenga[1,2], Suhung Shen[1,3] , Bruce E. Vollmer[1], Dave Meyer [1]

Dana.Ostrenga@nasa.gov [1]NASA Goddard Space Flight Center, [2]ADNET, [3]George Mason University

Big Data at GES DISC

The Goddard Earth Sciences Data and Information Services Center (**GES DISC**) is one of 12 NASA EOSDIS DAACS that manages, archives and distributes Earth science data as part of NASA's Earth Science Data Systems Program (ESDIS). We provide support for the archive and distribution of data for multiple satellite sensors, reanalysis models, ground measurements, and field campaigns. These include Aqua AIRS, TRMM, GPM, OCO-2, AURA/MLS/OMI, SORCE, TOMS, TOVS, , UARS, MERRA/MERRA-2, and LDAS.

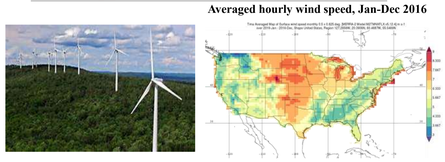
The GES DISC archives a total data volume that exceeds **1.5 Petabytes,** consisting of almost **95 million granules** covering over **2500 public and restricted collections,** and is growing every day.

The figure at left shows cumulative ESDIS archives. (Courtesy of ESDIS)

Examples of User Needs to Access Long Time-Series with MERRA-2 Data

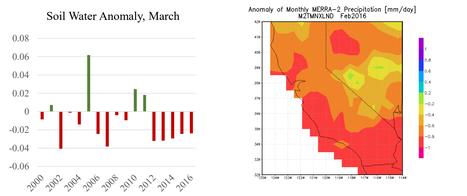
Recent activities have been focused on creating new value-added products, Data Lists, and subsetting services to support research and the following applications communities:

Application in Wind Energy



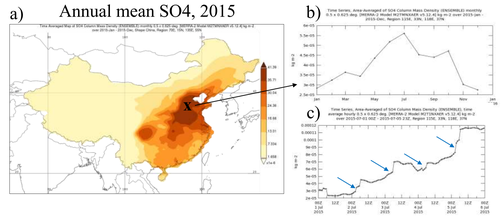
- High temporal resolution wind at 80m (the normal turbine height) for validating data and computing statistics, such as extreme, daily mean, min & max wind speed, and direction information
- Wind vertical profile in the boundary layer for studying boundary layer stability, which may affect power generation
- Temperature and moisture near the surface, for managing the on/off state of the power-grid system

Application in Drought Events



Monthly root zone water anomaly for March from 2000 to 2016 (left), and example of anomaly map of precipitation of February 2016 over California (right). The climatology base period used is 1980-01-01 to 2014-12-31.

Application in Air Quality

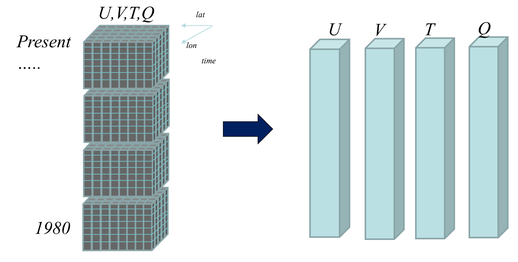


a) Year 2015 annual mean of SO4 column mass density over China; b) Area mean of 2015 monthly SO4 over East central China (33-37° N, 115-118° E), showing seasonal variations; c) hourly time series of SO4 for July 1-5, 2015 over the same area as in b), indicating diurnal variations.

User Driving Concept: Generating Data to be Accessed Easy by Users and Analysis Ready

"Big Data" is the term used when referencing large data sets that have a volume beyond the capabilities of common software tools to efficiently and effectively process, subset, analyze, or visualize in a reasonable amount of time. Data is currently archived separately, so multiple I/O's have to be performed, causing the potential of multiple issues for data access.

Scientists face challenges in dealing with the native data files; they face slow processing times, timeouts, system overloads and storage capacity issues.



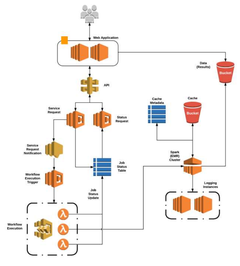
Data has to be conformable to the needs of the user:

- Downloading in different formats, time series at a single point or a small area into a single file in different data format (e.g.:ASCII, NetCDF, or flat binary) easy and fast;
- Computing basic statistics–for example, mean, minimum, maximum, and standard deviation;
- Finding extreme events, such as days of high winds, hot or cold records, drought, high aerosol load, etc.

How do we address this?

- Reconstructing data files
- Creating value-added datasets or tools for specific applications
- Integrating selected data into a cloud environment
- Packaging data based on application areas

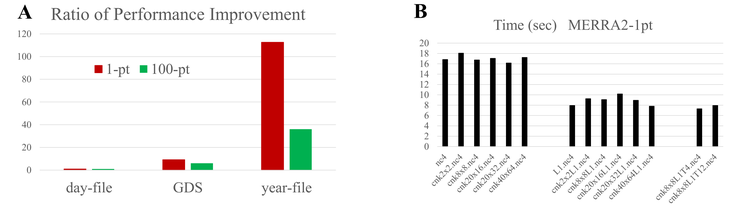
Future: In the Cloud…..



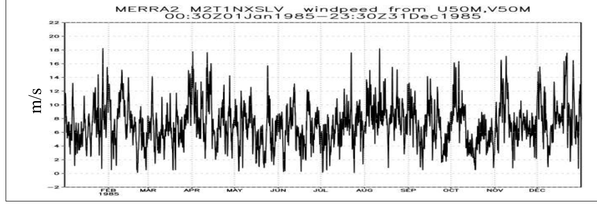
The cloud environment can provide multiple solutions for services:

- API gateway for service endpoints
- Cluster computing capabilities for high performance
- Simple Storage Service for scalable data storage and management
- Analytical services
- Provides the agility required to meet our customers' increasing and changing needs.

Performance test examples to access time series: Current archive and reconstructed data cubes



A shows how much faster it is to access the same amount of data through GDS and reconstructed data-cubes compared to the original archive. It takes ~ 30 min to extract 1-year of MERRA-2 hourly time series at a single point from current archives (day-file structure), which is reduced to ~ 8 to 16 sec from the reconstructed year-file, i.e. over 100 times faster. **B** shows the time used to extract 1-year data at a single point from year-file saved in different internal chunking and compression.



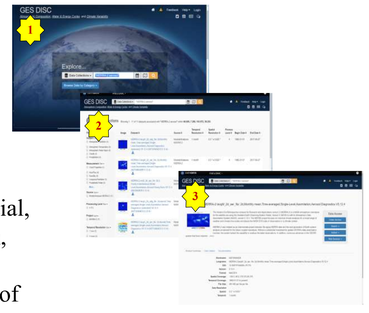
The figure at left is an example of a time- series of wind speed (m/s) at 50m above surface over (1° W 52° N) for year 1985.

Services at GES DISC to Handle Big Data

https://disc.gsfc.nasa.gov

By developing and enhancing various services and tools, the GES DISC provides users with the capability to access and visualize data, and to make comparisons of data from multiple sensor and models via a number of cross-discipline projects.

Discovering Data via Faceted Web Interface



- ✓ Web interface to data products and services
- ✓ Search and Download mechanisms
- ✓ Dataset Landing Pages

Downloading Data Basics:

- ✓ **Subset and Regridding Service** – Parameter, Spatial, Time, Vertical, Mean averaging, format conversion, and regridding for L3/L4 gridded data
- ✓ **Swath Data Subsetter** – Parameter, spatial subset of L2 /L1 data.

Accessing Data through Interoperable Services:

- ✓ **GDS** – GrADS Data Server
- ✓ **OPeNDAP** - Open-source Project for a Network Data Access Protocol
- ✓ **WMS** – OGC service
- ✓ **GIS connector** – allowing GIS tools to access data easier (coming soon)
- ✓ **HTTPS --** direct online access

Visualizing Data Online:

- ✓ **Giovanni** –Visualization and Analysis L3/L4 gridded data
- ✓ **AIRS NRT Viewer** – AIRS near-real-time
- ✓ **DQVis** – L2 data quality visualization