A Webservice Platform for Big Ocean Data Science

Edward M. Armstrong[1], Mark A. Bourassa[2], **Thomas Huang**[1], Joseph Jacob[1], Yongyao Jiang[4], Yun Li[4], Nga Quach[1], Shawn Smith[2], Vardis Tsontos[1], Brian Wilson[1], Steve J .Worley[3], and Chaowei (Phil) Yang[4]

[1] NASA Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

[2] Center for Ocean-Atmospheric Prediction Studies, 2000 Levy Avenue, Building A, Suite 292, Tallahassee, FL 32306-2741, USA

[3] National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000, USA

[4] George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

[CL#17-2300]

OceanWorks: Ocean Science Platform

This presentation provides an overview of OceanWorks, the webservice platform for big ocean data science at the NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC), and to discuss the open source big data analytic solutions that OceanWorks uses to enable fast analysis of Sea Surface Temperature (SST) data. Funded through the NASA's Advance Information System Technology (AIST) Program and developed collaboratively between JPL, FSU, NCAR, and GMU, OceanWorks will be the platform for the next generation of PO.DAAC data solutions. OceanWorks is an orchestration of several previous funded NASA big ocean data solutions using cloud computing technology, which include on-the-fly data analysis (NEXUS), anomaly detection (OceanXtremes), matchup (DOMS), quality-screened subsetting (VQSS), search relevancy (MUDROD), and web-based visualization (Common Mapping Client).

With increasing global temperature, warming of the ocean, and melting ice sheets and glaciers, the impacts can be observed from changes in anomalous ocean temperature and circulation patterns, to increasing extreme weather events and super hurricanes, sea level rise and storm surges affecting coastlines, and may involve drastic changes and shifts in marine ecosystems. Ocean science communities are relying on data distributed through data centers such as the PO.DAAC to conduct their research. In typical investigations, oceanographers follow a traditional workflow for using datasets: search, evaluate, download, and apply tools and algorithms to look for trends. While this workflow has been working very well historically for the oceanographic community, it cannot scale if the research involves massive amount of data. NASA's Surface Water and Ocean Topography (SWOT) mission, scheduled to launch in April of 2021, is expected to generate over 20PB data for a nominal 3-year mission. This will challenge all existing NASA Earth Science data archival/distribution paradigms. It will no longer be feasible for Earth scientists to download and analyze such volumes of data. OceanWorks will enable fast, web-accessible fast data analysis directly on our physical ocean archive to minimize data movement and provide access, including subset, only to the relevant data.

https://oceanxtremes.jpl.nasa.gov

OceanXtremes: Data-Intensive Anomaly Detection

Hurricane Katrina Study – Using OceanXtremes

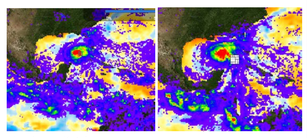
Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 deg C that lingers for several days. Following this was a short intense ocean chlorophyll bloom a few days later. The ocean may have been "preconditioned" by a cool core eddy and low sea surface height.

The SST drop is correlated to both wind and precipitation data. The Chl-A data is lagged by about 3 days to the other observations like SST, wind and precipitation.

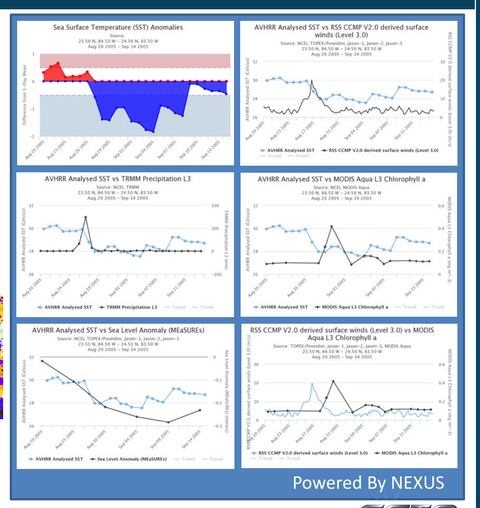
A study of a Hurricane Katrina–induced phytoplankton bloom using satellite observations and model simulations
Xiaoming Liu, Menghua Wang, and Wei Shi

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 114, C03023, doi:10.1029/2008JC004934, 2009

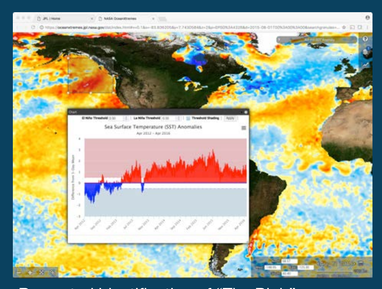
http://shoni2.princeton.edu/ftp/lyo/journals/Ocean/phybiogeochem/Liu-etal-KatrinaChlBloom-JGR2009.pdf



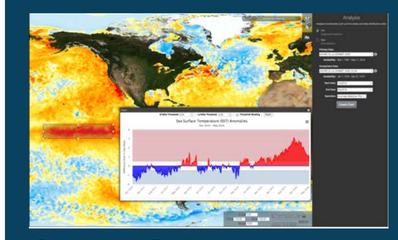
Hurricane Katrina
TRMM overlay SST Anomaly



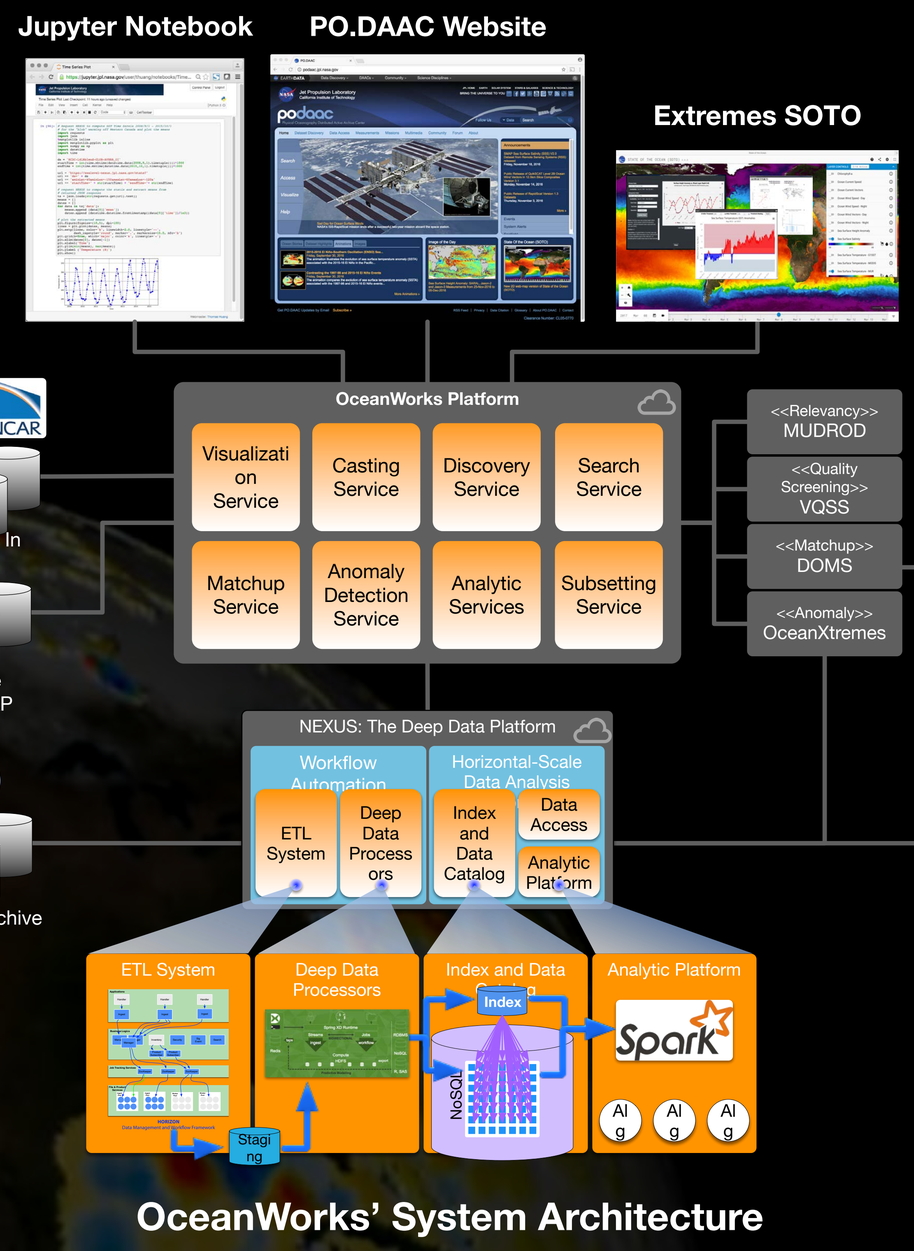
THUANG/JPL NEXUS



Recreated identification of "The Blob"



Recreated the El Niño 3.4 regional signal

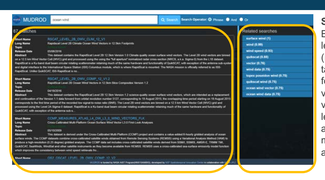


OceanWorks' System Architecture

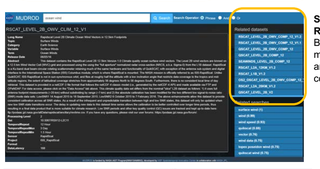
https://mudrod.jpl.nasa.gov

Search and Discovery

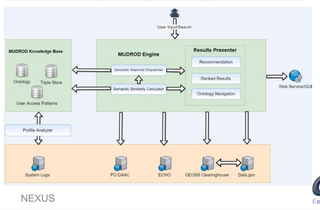
- Mining and Utilizing Dataset Relevancy from Oceanographic Dataset
- **Search** – look for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** – find something new, or in a new way
  - This is non-trivial
  - Traditional ontological method doesn't quite add up
  - The strength of semantic web is in inference
  - What happen when we have a lot of `subClassOf, equivalentClassOf, sameAs`?
  - How wide and deep should we go?
- **Relevancy**
  - It is domain-specific
  - It is personal
  - It is temporal
  - It is dynamic
- MUDROD analyzes web logs to discover user knowledge (the connections between datasets and keyword)
- Construct knowledge base by combining semantics and profile analyzer
- Improve data discovery by better ranked results



Search Ranking
Based on a machine learning model (RankSVM) which takes a number of features, such as vector space model, version, processing level, release date, all-time popularity, monthly-popularity, and user popularity.



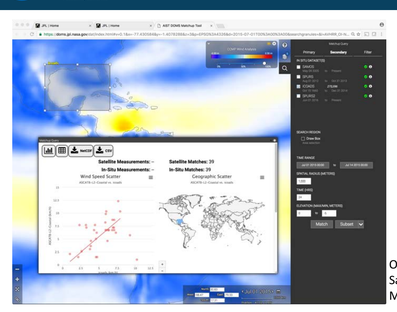
Search Recommendation
Based on dataset metadata content and web session co-occurrence



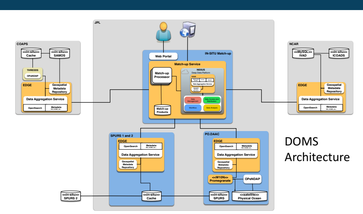
THUANG/JPL NEXUS

https://doms.jpl.nasa.gov

Distributed Oceanographic Matchup Service (DOMS)

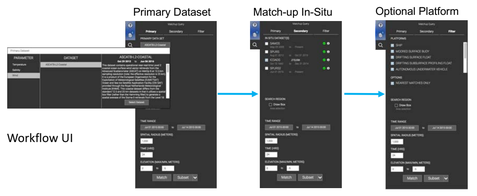


On-The-Fly Satellite to In-Situ Matchup



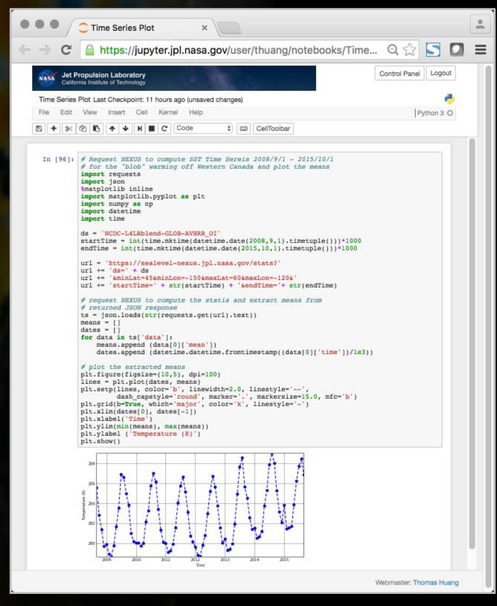
DOMS Architecture

Webservice API



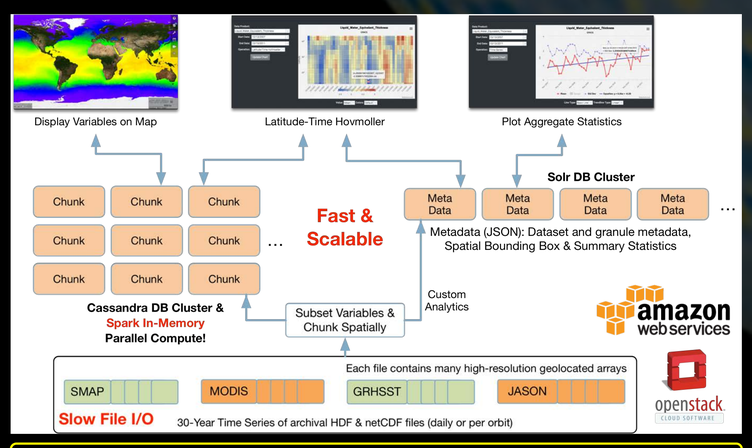
THUANG/JPL NEXUS

https://jupyter.jpl.nasa.gov



Big Data Analytics On Cloud Using NEXUS

Open Source: https://github.com/dataplumber/nexus

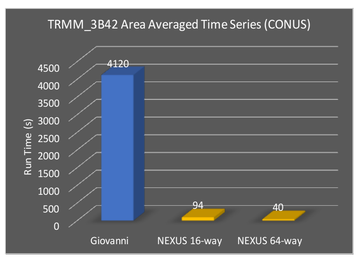


Almost all of the existing Earth science data analysis solutions are built around large archives of files. When an analysis involves large collection of files, performance suffers due to the large amount of I/O required. Common data access solutions, such as OPeNDAP and THREDDS, provide web service interface to archives of observational data. They also yield poor performance when it comes to large amount of observations, because they are still built around the notion of files. In his famous 2005 paper on Scientific Data Management in the Coming Decade, the late Jim Gray stated, "The scientific file-formats of HDF, NetCDF, and FITS can represent tabular data but they provide minimal tools for searching and analyzing tabular data." He continued to point out, "Performing this filter-then-analyze, data analysis on large datasets with conventional procedural tools runs slower and slower as data volume increases."

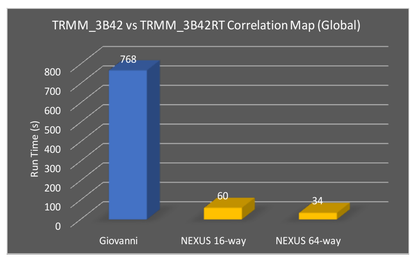
NEXUS (https://github.com/dataplumber/nexus) is an emerging, open source, data-intensive analysis framework developed with a new approach for handling science data that enables large-scale data analysis. MapReduce is a well-known paradigm for processing large amounts of data in parallel using clustering or Cloud environments. Unfortunately, this paradigm doesn't work well with temporal, geospatial array-based data. One major issue is they are packaged in files in various sizes. The size of each data file can range from tens of megabytes to several gigabytes. Depending on the user input, some analysis operations could involve hundreds to thousands of these files.

NEXUS takes on a different approach in handling file-based observational temporal, geospatial artifacts by fully leveraging the elasticity of Cloud Computing environment. Rather than performing on-the-fly file I/O, NEXUS stores tiled data in Cloud-scaled databases with high-performance spatial lookup service. NEXUS provides the bridge between science data and horizontal-scaling data analysis. This platform simplifies development of big data analysis solutions by bridging the gap between files and MapReduce solutions such as Spark. NEXUS has been integrated into the NASA Sea Level Change Portal (https://sealevel.nasa.gov) as the Big Data analytic backend for its Data Analysis Tool.

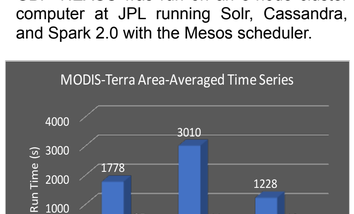
SPEED



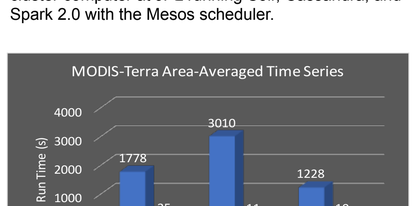
Comparison of Giovanni and NEXUS run times to compute area-averaged time series over the continental United States of 0.25 degree TRMM daily precipitation rate for 1/1/1998 - 12/31/2015. These runs included 6,574 global data granules (between 50 deg. south and 50 deg. north latitude) totaling 26 GB. NEXUS was run on an 8-node cluster computer at JPL running Solr, Cassandra, and Spark 2.0 with the Mesos scheduler.



Comparison of Giovanni and NEXUS run times to compute global correlation map of 0.25 degree TRMM daily precipitation rate vs. TRMM real-time daily precipitation for 1/1/2001 - 12/31/2014. These runs included 5,113 global data granule pairs (between 50 deg. south and 50 deg. north latitude) totaling 40 GB. NEXUS was run on an 8-node cluster computer at JPL running Solr, Cassandra, and Spark 2.0 with the Mesos scheduler.



Comparison of Giovanni and NEXUS run times to compute 16-year area-averaged daily time series of MODIS-Terra Aerosol Optical Depth (AOD) 550 nm dark target at 1 degree resolution for the indicated spatial bounds (Global, Colorado, or Boulder). NEXUS was run on 6 Amazon Web Services (AWS) Cloud instances of type "i2.4xlarge" with Solr, Cassandra, Spark 2.0, and the Mesos scheduler. NEXUS has superior performance for subsetting operations, as indicated by the ~300x speedup for the Colorado subset.

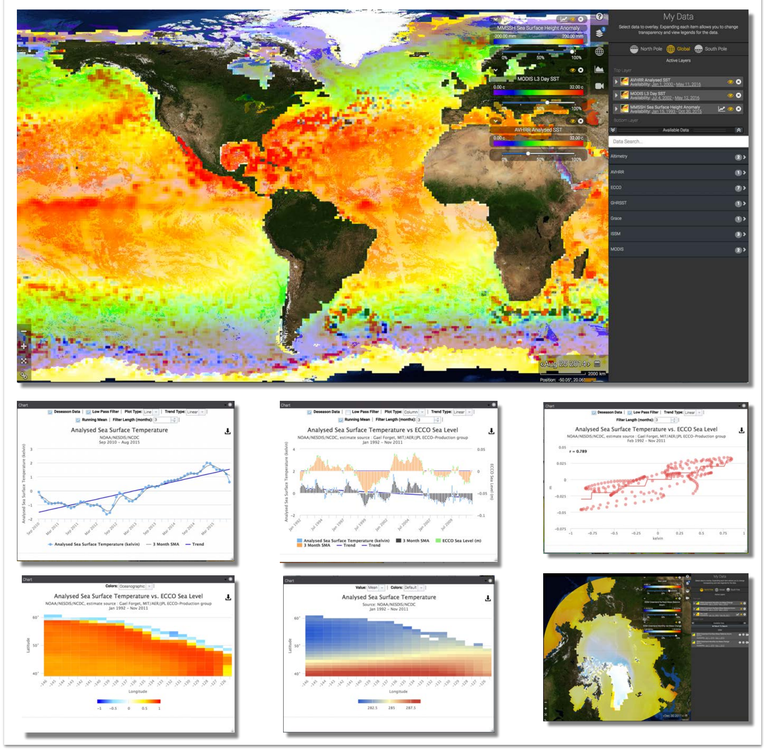


Comparison of Giovanni and NEXUS run times to compute 16-year area-averaged daily time series of MODIS-Terra Aerosol Optical Depth (AOD) 550 nm dark target at 1 degree resolution for the indicated spatial bounds (Global, Colorado, or Boulder). NEXUS was run on 6 Amazon Web Services (AWS) Cloud instances of type "i2.4xlarge" with Solr, Cassandra, Spark 2.0, and the Mesos scheduler. NEXUS has superior performance for subsetting operations, as indicated by the ~300x speedup for the Colorado subset.

https://sealevel.nasa.gov

Data Analysis Tool

Parameter Visualization | Polar Projection | Dataset Info | Data Sequencer
Suite of Analytic Functions | Download Plots and Results



podaac
ESTO Earth Science Technology Office