



**National Aeronautics and  
Space Administration**

**Jet Propulsion Laboratory**  
California Institute of Technology  
Pasadena, California

## **Big Data Analytics and Visualization for Ocean Sciences**

**Thomas Huang**

Data Scientist | Principal Investigator | Technologist | Architect  
thomas.huang@jpl.nasa.gov

Jet Propulsion Laboratory  
California Institute of Technology  
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

[CL #]



## Principal Investigator

NASA AIST OceanWorks – Ocean Science Platform on Cloud

## Project Technologist

NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)

## Co-Investigator and Architect

NASA Sea Level Change Portal

## Architect

CEOS Ocean Variables Enabling Research and Application for GEOS (COVERAGE)

## Architect

Tactical Data Science Framework for Naval Research

## Cluster Chair

Federation of Earth Science Information Partners (ESIP) Cloud Computing

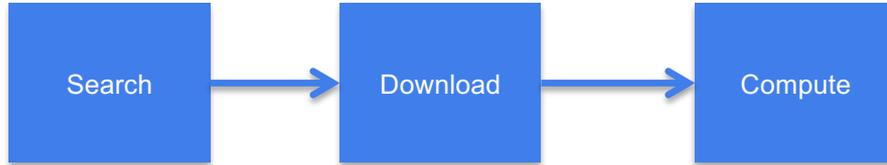
## Previously Principal Investigator / Co-Investigator

Several NASA-funded Big Data Analytic Projects – Big Data Analytics on the Cloud, Anomaly Detection, In Situ and Satellite Matchup, Search Relevancy, and Quality Screening

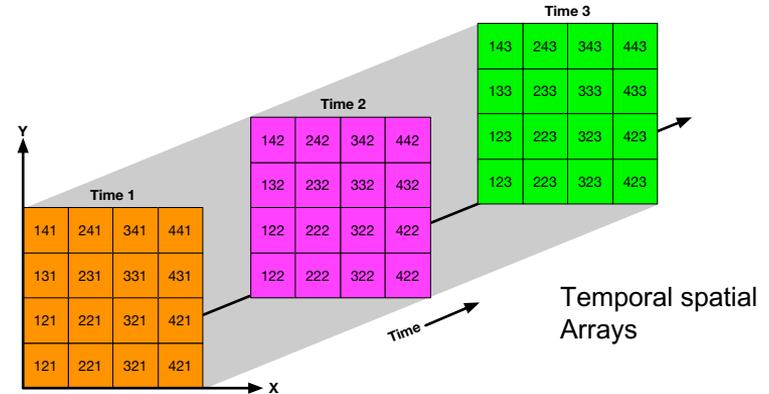
- **NASA has historically focused on systematic capture and stewardship of data for observational Systems**
- **With large amount of observational and modeling data, finding and downloading is becoming inefficient**
- **Reality with large amount of observational and modeling data**
  - Downloading to local machine is becoming inefficient
  - Search has gotten a lot faster. Too many matches.
  - Finding the relevant measurement has becoming a very time consuming process "*Which SST dataset I should use?*"
  - Analyze decades of regional measurement is labor-intensive and costly
- **Increasing “big data” era is driving needs to**
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences
  - Shift towards integrated data analytics
  - Apply computational and data science across the lifecycle
- **Scalable Data Management**
  - Capture well-architected and curated data repositories based on well-defined data/information architectures
  - Architecting automated pipelines for data capture
- **Scalable Data Analytics**
  - Access and integration of highly distributed, heterogeneous data
  - Novel statistical approaches for data integration and fusion
  - Computation applied at the data sources
  - Algorithms for identifying and extracting interesting features and patterns

- **Mainly focus on archives and distributions**
- **With additional services**
  - Better searches – faceted, spatial, keyword, ranking, etc.
  - Data subsetting – home grown, OPeNDAP, etc.
  - Visualization – visual discovery, PO.DAAC's SOTO, NASA Worldview, etc.
- **Limitations**
  - Little to no interoperability between tools and services: metadata standard, keyword, spatial coverage (0-360 or -180..180), temporal representation, etc.
  - Making sure the most relevant measurements return first
  - Visualization is nice, but it doesn't provide enough information about the event/phenomenon captured in the image.
  - With large amount of observational data, data centers need to do more than just storing bits
    - “Is the red blob in the middle of Pacific normal this time of the year?”
    - “Any relevant news and publications relate to what I am looking at?”
    - “What other measurements, phenomena, news, publications relate to the period and location I am looking at?”
    - “I can see the observation from satellite, are there any relevant in situ data I can look at?”

# Traditional Method for Analyze Satellite Measurements

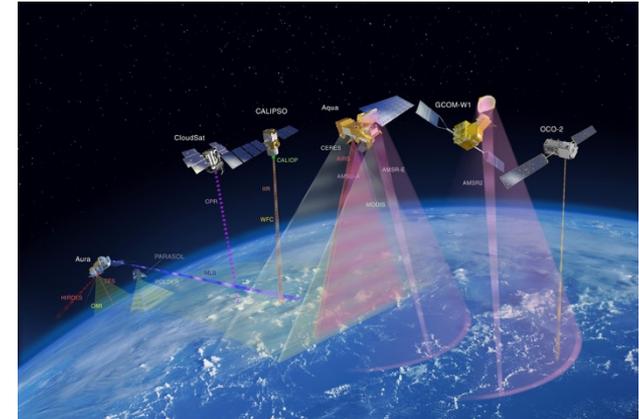
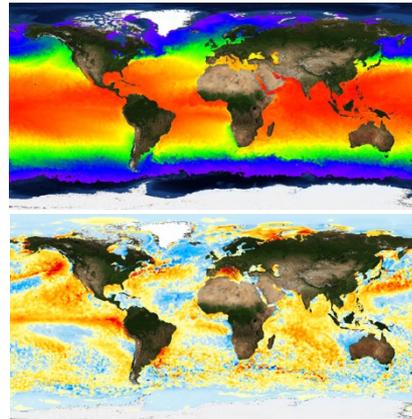


- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced, purge downloaded files



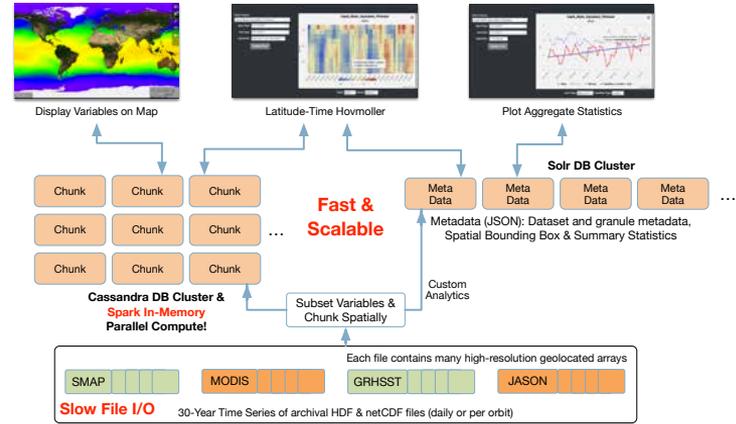
## Observation

- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be free from file I/O bottleneck

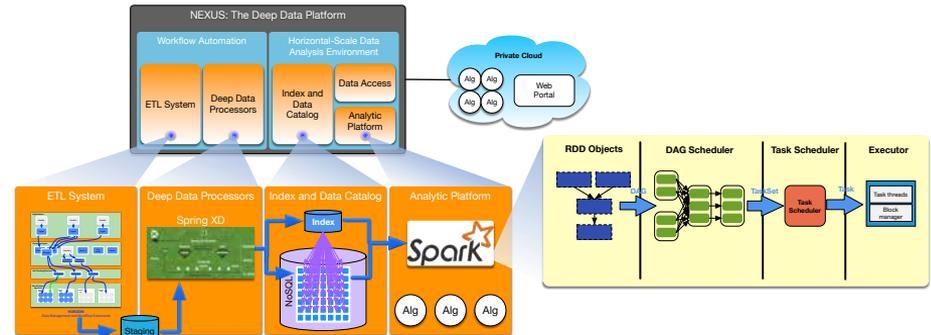


# NEXUS: Scalable Data Analytic Solution

- NEXUS is a data-intensive analysis solution using a new approach for handling science data to enable large-scale data analysis
- Streaming architecture for horizontal scale data ingestion
- Scales horizontally to handle massive amount of data in parallel
- Provides high-performance geospatial and indexed search solution
- Provides tiled data storage architecture to eliminate file I/O overhead
- A growing collection of science analysis webservices using Apache Spark: parallel compute, in-memory map-reduce framework
- Pre-Chunk and Summarize Key Variables
  - Easy statistics instantly (milliseconds)
  - Harder statistics on-demand using Spark (in seconds)
  - Visualize original data (layers) on a map quickly (Cassandra store)
- **Algorithms** – Time Series | Latitude/Time Hovmöller| Longitude/Time Hovmöller| Latitude/Longitude Time Average | Area Averaged Time Series | Time Averaged Map | Climatological Map | Correlation Map | Daily Difference Average

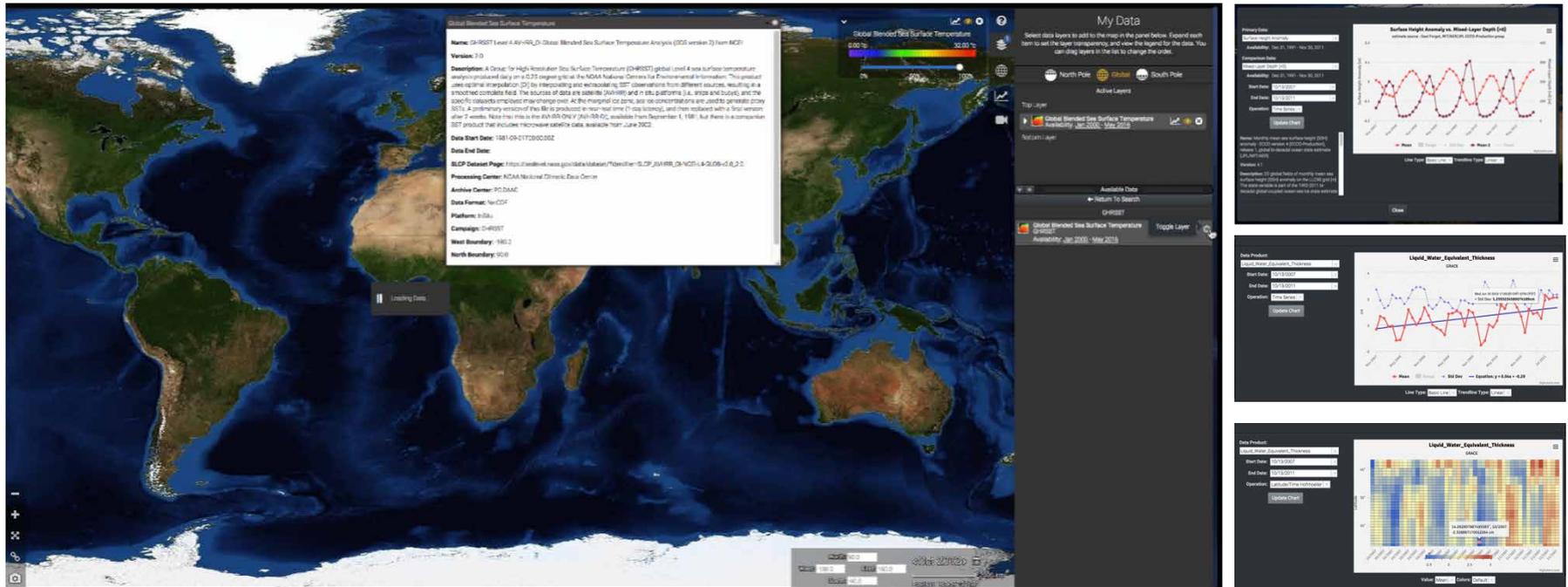


## Two-Database Architecture



**Open Source: Apache License 2**

<https://github.com/apache/incubator-sdap-nexus>



## Sea Level Change - Data Analysis Tool

Visualizations | Hydrological Basins | Time Series | Deseason | Data Comparison | Scatter Plot | Latitude/Time Hovmöller | Etc.

# NEXUS Performance: GIOVANNI vs. Custom Spark vs. AWS EMR

**Dataset:** MODIS AQUA Daily

**Name:** Aerosol Optical Depth 550 nm (Dark Target) (MYD08\_D3v6)

**File Count:** 5106

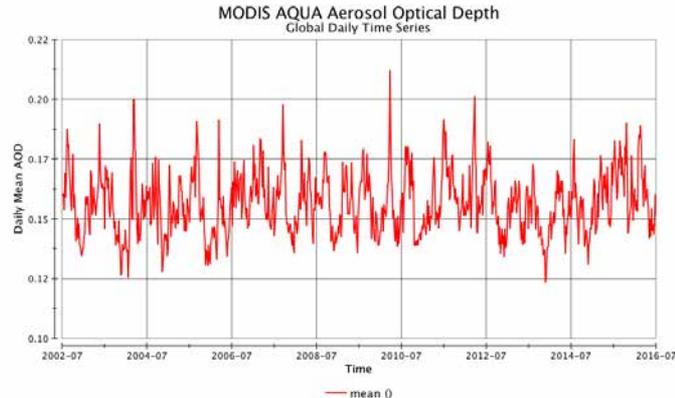
**Volume:** 2.6GB

**Time Coverage:** July 4, 2002 – July 3, 2016

**Giovanni:** A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.

- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

**AWS EMR:** Amazon's provisioned MapReduce cluster **Giovanni: 20 min**  
**NEXUS: 1.7 sec**



Area Averaged Time Series on AWS - Boulder

July 4, 2002 - July 3, 2016  
 NEXUS Performance

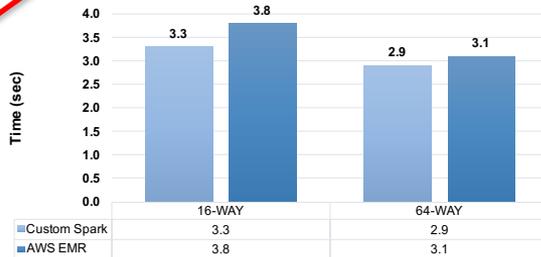
Custom Spark vs. AWS EMR  
 Ref. Speed - Giovanni: 1140.22 sec



Area Averaged Time Series on AWS - Colorado

July 4, 2002 - July 3, 2016  
 NEXUS Performance

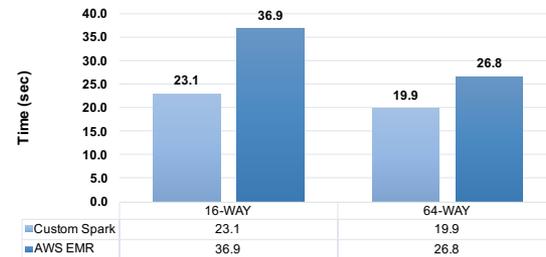
Custom Spark vs. AWS EMR  
 Ref. Speed - Giovanni: 1150.6 sec



Area Averaged Time Series on AWS - Global

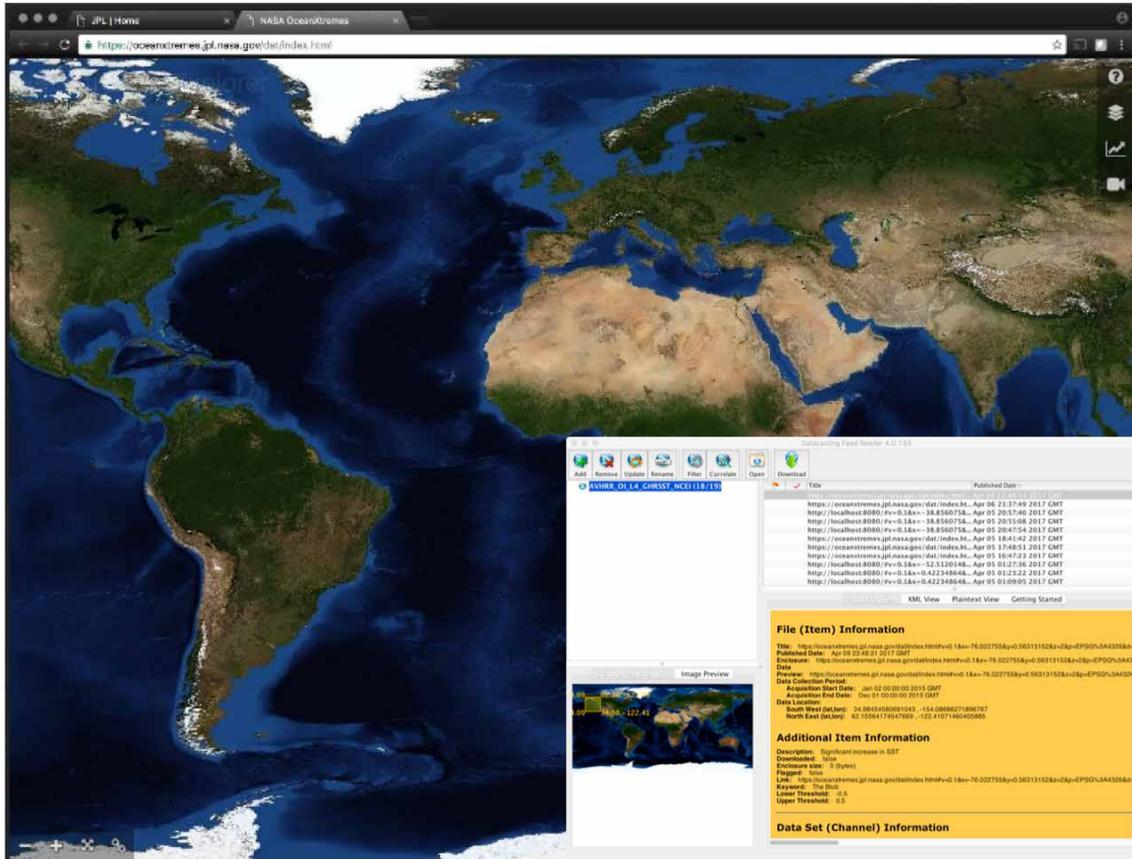
July 4, 2002 - July 3, 2016  
 NEXUS Performance

Custom Spark vs. AWS EMR  
 Ref. Speed - Giovanni: 1366.84 sec



Algorithm execution time. Excludes Giovanni's data scrubbing processing time

# Analyze Ocean Anomaly – “The Blob”



- **Visualize** parameter
- **Compute** daily differences against climatology
- **Analyze** time series area averaged differences
- **Replay** the anomaly and visualize with other measurements
- **Document** the anomaly
- **Publish** the anomaly

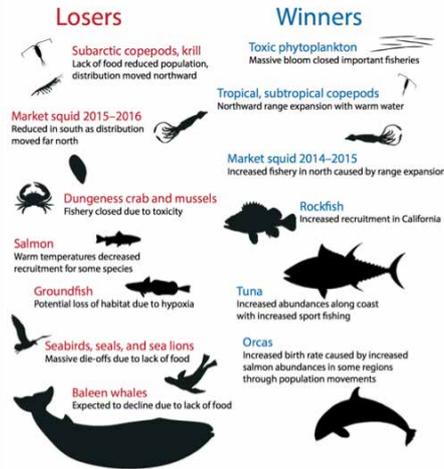


Figure from Cavole, L. M., et al. (2016). "Biological Impacts of the 2013–2015 Warm-Water Anomaly in the Northeast Pacific: Winners, Losers, and the Future." *Oceanography* 29.



# Enable Science without File Download

The screenshot shows a JupyterLab environment with a Python script in the top pane and a plot of SST time series data in the bottom pane. The script requests SST data from NEXUS, extracts the means, and plots the results. The plot shows a clear seasonal cycle in SST from 2008 to 2015.

```
# Request NEXUS to compute SST Time Series 2008/9/1 - 2015/10/1
# for the "blob" warming off Western Canada and plot the means
...
ds='AVHRR_OI_L4_GHRSSST_NCEI'

url = ... # construct the webservice URL request

# make request to NEXUS using URL request
# save JSON response in local variable
ts = json.loads(str(requests.get(url).text))

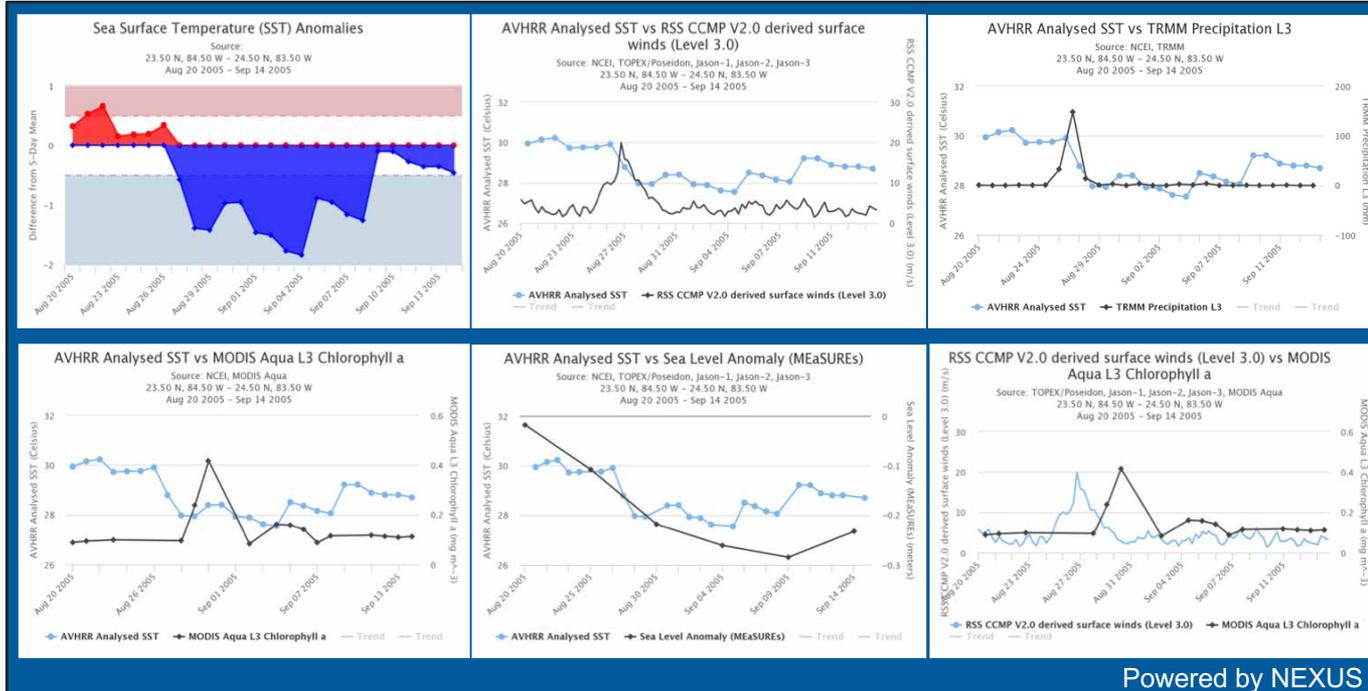
# extract dates and means from the response
means = []
dates = []
for data in ts['data']:
    means.append(data[0]['mean'])
    d = datetime.datetime.fromtimestamp((data[0]['time']))
    dates.append(d)

# plot the result
...
```

```
https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=1220227200&endTime=1443657600
```

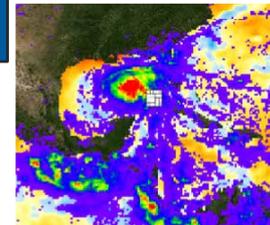
It took: 2.9428272247314453 sec

# Hurricane Katrina Study



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 °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.



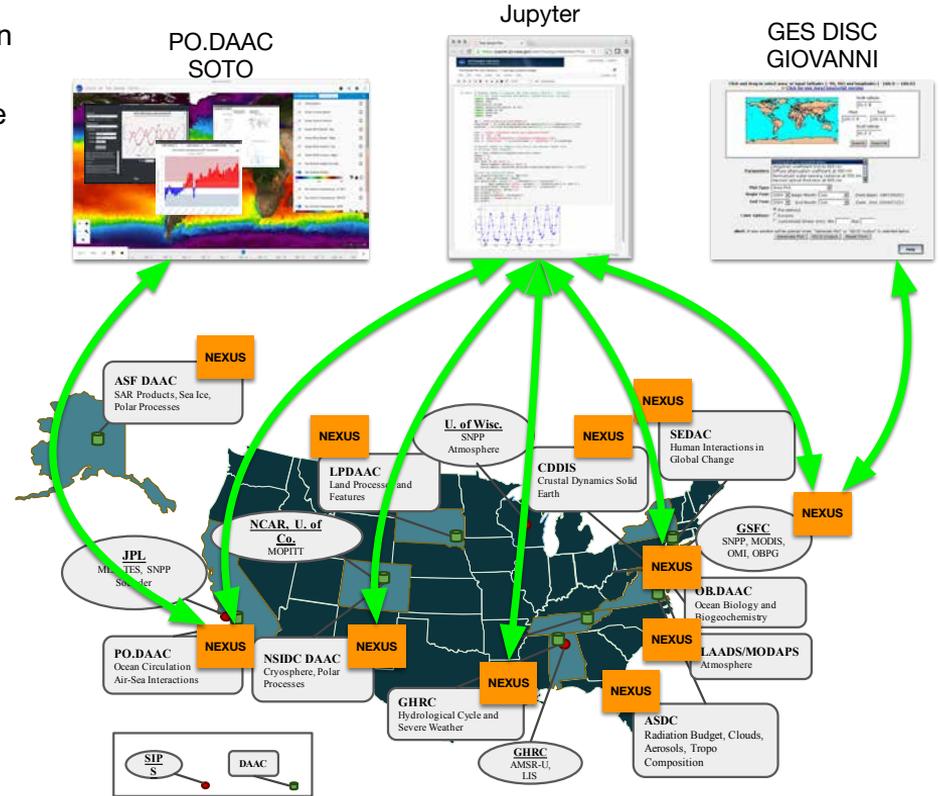
Hurricane Katrina  
 TRMM  
 overlay SST  
 Anomaly

*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

Powered by NEXUS

# Multi-Variable Analysis

- Public accessible RESTful analytic APIs where computation is next to the data
- NEXUS as the analytic engine infused and managed by the DAACs on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files
- Reduce unnecessary egress charges
- An architecture to enable next generation of scientific applications



# Supported Datasets

- **Atmosphere**

- MODIS Aqua Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08\_D3v6)
- MODIS Terra Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08\_D3v6)
- MODIS Aqua Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08\_D3v6)
- MODIS Terra Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08\_D3v6)

- **Chlorophyll**

- MODIS Aqua Level 3 Global Daily Mapped 4 km Chlorophyll a

- **Estimating the Circulation and Climate of the Ocean (ECCO)**

- Monthly Mean Version 4 release 2 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Total Heat Flux, Total Salt Flux
- Monthly Mean Version 4 release 1 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Ocean Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Actual Sublimation Freshwater Flux, Total Heat Flux, Total Salt Flux

- **Gravity**

- Center for Space Research (CSR) GRACE RL05 Mascon Solutions
- JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL05M.1 CRI filtered Version 2

- **Ocean Temperature**

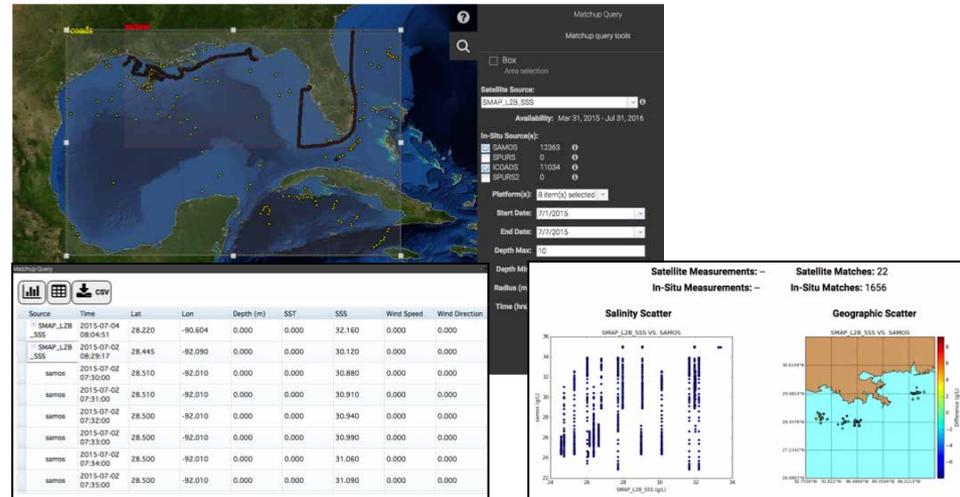
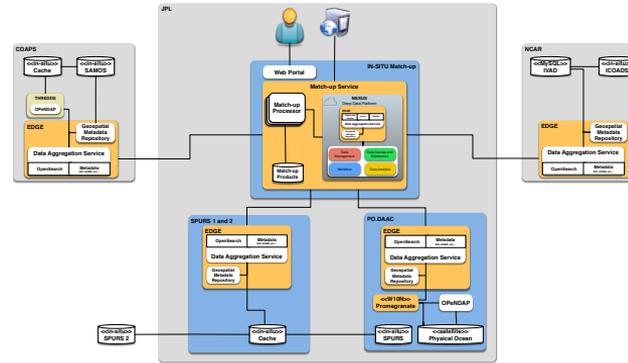
- GHRSSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (v4.1)
- GHRSSST Level 4 AVHRR\_OI Global Blended Sea Surface Temperature Analysis (GDS version 2) from NCEI
- MODIS Aqua Level 3 SST Thermal IR Daily 4km Nighttime v2014.0
- MODIS Aqua Level 3 SST Thermal IR Daily 4km Daytime v2014.0

# Supported Datasets (+)

- **Salinity**
  - JPL SMAP Level 2B CAP Sea Surface Salinity V2.0 Validated Dataset
  - JPL SMAP Level 3 CAP Sea Surface Salinity Standard Mapped Image Monthly V3.0 Validated Dataset
- **Sea Surface Height Anomalies (SSHA)**
  - JPL MEaSURES Gridded Sea Surface Height Anomalies Version 1609
- **Wind**
  - Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L3.0 First-Look Analyses
- **Precipitation (non-ocean data)**
  - TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM\_3B42\_Daily) at GES DIS
  - TRMM (TMPA-RT) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM\_3B42\_RT) at GES DISC
- **In Situ**
  - Shipboard Automated Meteorological and Oceanographic System (SAMOS)
  - International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations
  - Salinity Process in the Upper Ocean Regional Study – 1 (SPURS1)
  - Salinity Process in the Upper Ocean Regional Study – 2 (SPURS2)
  - Global gridded NetCDF Argo only dataset produced by optimal interpolation (salinity variables)
  - Global gridded NetCDF Argo only dataset produced by optimal interpolation (temperature variables)

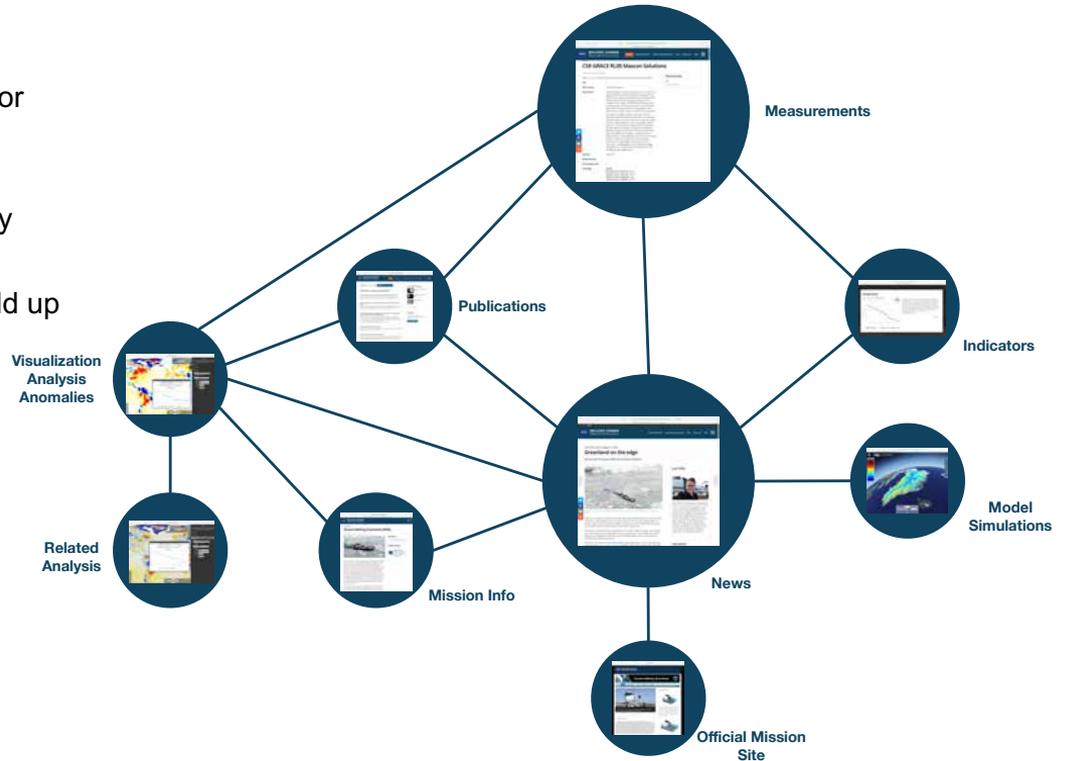
# In Situ to Satellite Matchup

- Distributed Oceanographic Matchup Service (DOMS)
- Typically data matching is done using one-off programs developed at multiple institutions
- A primary advantage of DOMS is the reduction in duplicate development and man hours required to match satellite/in situ data
  - Removes the need for satellite and in situ data to be collocated on a single server
  - Systematically recreate matchups if either in situ or satellite products are re-processed (new versions), i.e., matchup archives are always up-to-date.
- In situ data nodes at JPL, NCAR, and FSU operational.
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Plugin architecture for in situ data source using EDGE, an open source implementation of Open Search



# Tackling Information Discovery

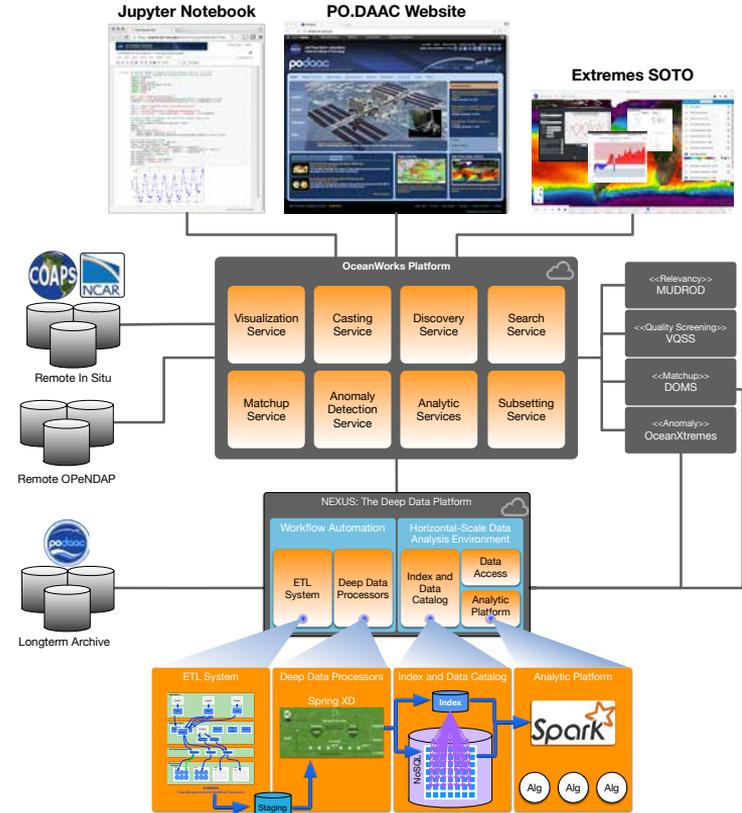
- **Search** is looking for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** is finding something new, or in a new way
  - This is non-trivial
  - Traditional ontological method doesn't quite add up
  - The strength of semantic web is in inference
  - Need method involves
    - Dynamic data ranking
    - Dynamic update to the ontology
    - Mining user interaction and news outlets
- **Relevancy** is
  - Domain-specific
  - Personal
  - Temporal
  - Dynamic



# AIST OceanWorks

PI: Thomas Huang

- **OceanWorks** is to establish an **Integrated Data Analytic Center** at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) for Big Ocean Science
- Focuses on technology integration, advancement and maturity
- Collaboration between JPL, FSU, NCAR, and GMU
- Bringing together PO.DAAC-related big data technologies
  - Anomaly detection and ocean science
  - Big data analytic platform
  - Distributed in-situ to satellite matchup
  - Search relevancy and discovery – linking datasets, services, and anomalies through recommendations
  - Metadata translation and services aggregation
  - Fast data subsetting
  - Virtualized Quality Screening Service

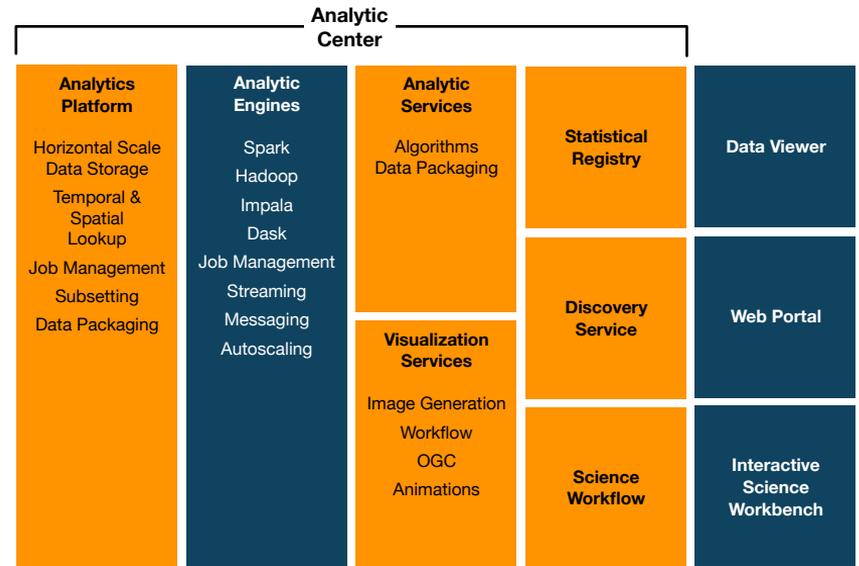




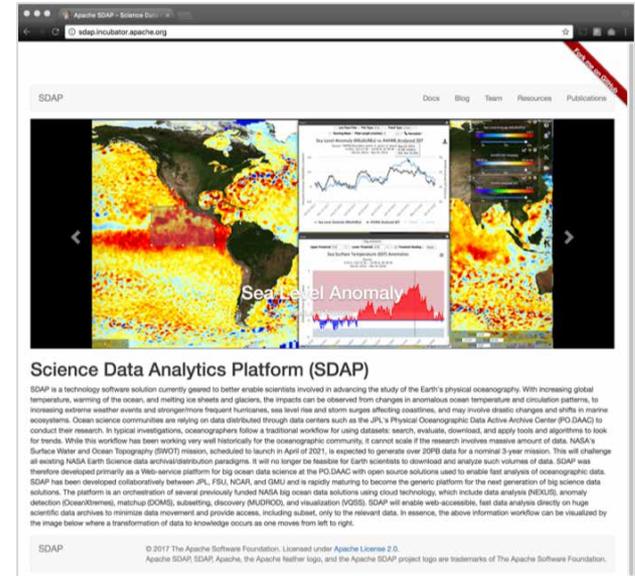


# Integrated Data Analytics Center

- An **Integrated Data Analytics Center**: an environment for conducting a Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (ocean, atmospheric, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the research community to focus on the investigation
  - Reduce the data preparation time to something tolerable
  - Catalog of optional resources
  - Semantic-enabled catalog of resources
  - Relevant publications
  - Provide established training data sets of varying resolution
  - Provide effective project confidentiality, integrity and availability
  - Single sign-on and unified financial tracking



- Technology sharing through Free and Open Source Software (FOSS)
- Further technology evolution that is restricted by projects / missions
- **Science Data Analytic Platform (SDAP)**, the implementation of **OceanWorks**, in **Apache Incubator**
  - Cloud platform
  - Analyzing satellite and model data
  - In situ data analysis and colocation with satellite measurements
  - Fast data subsetting
  - Mining of user interactions and data to enable discovery and recommendations
  - Streamline deployment through container technology



<http://sdap.incubator.apache.org>



# In Summary

- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- Think beyond archive and file downloads
- Connected information enables discovery
- Community developed solution through open sourcing
- Thanks to the NASA ESTO/AIST and Sea Level Rise programs, and the NASA ESDIS project
- Investment in data and computational sciences
- Data Centers might want to be in the business of Enabling Science!
- OceanWorks infusion 2018 – 2019
  - Watch for changes to the Sea Level Change Portal
    - Even faster analysis capabilities
    - More variety of measurements – satellites, in situ, and models
    - Event more relevant recommendations
  - NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)
    - More than just pretty pictures. SOTO will have new analytic capabilities.
- Lead Editor: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**



**National Aeronautics and  
Space Administration**

**Jet Propulsion Laboratory**  
California Institute of Technology  
Pasadena, California



**Thomas Huang**

Jet Propulsion Laboratory  
California Institute of Technology

**JPL Team**

Ed Armstrong, Frank Greguska, Joseph Jacob, Lewis McGibbney,  
Nga Quach, Vardis Tsonos, and Brian Wilson

**Florida State University Team**

Shawn Smith, Mark A. Bourassa, Jocelyn Elya

**National Center for Atmospheric Research Team**

Steve J. Worley, Tom Cram, Zaihua Ji

**George Mason University Team**

Chaowei (Phil) Yang, Yongyao Jiang, and Yun Li