An Introduction To

OceanWorks

Ocean Science Platform

Thomas Huang
Jet Propulsion Laboratory
California Institute of Technology

JPL Team
Ed Armstrong, Joseph Jacob, Nga Quach, Vardis Tsontos, and Brian Wilson

Florida State University Team
Shawn Smith, and Mark A. Bourassa

National Center for Atmospheric Research Team
Steve J. Worley

George Mason University Team
Chaowei (Phil) Yang, Yongyao Jiang, and Yun Li

THUANG/JPL. © 2017. All rights reserved.

Big Data Reality

Reality

- With large amount of observational and modeling data, downloading to local machine is becoming inefficient
- Data centers are starting to provide additional services
  - Better searches – faceted, spatial, keyword, relevancy, etc.
  - Data subsetting – data reduction
  - Visualization – visual discovery

2015 NASA ESTO/AIST Big Data Study Roadmap: Moving from Data Archiving to Data Analytics

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

- Capturing well-architected and curated data repositories based on well-defined data/information architectures
- Architecting automated pipelines for data capture

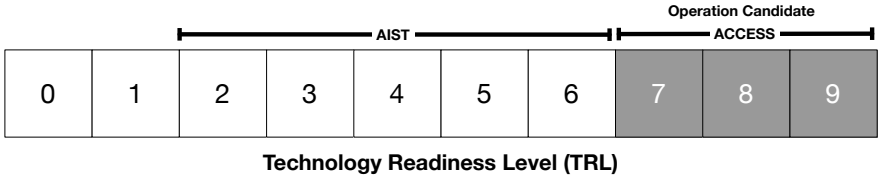
Scalable Data Analytics

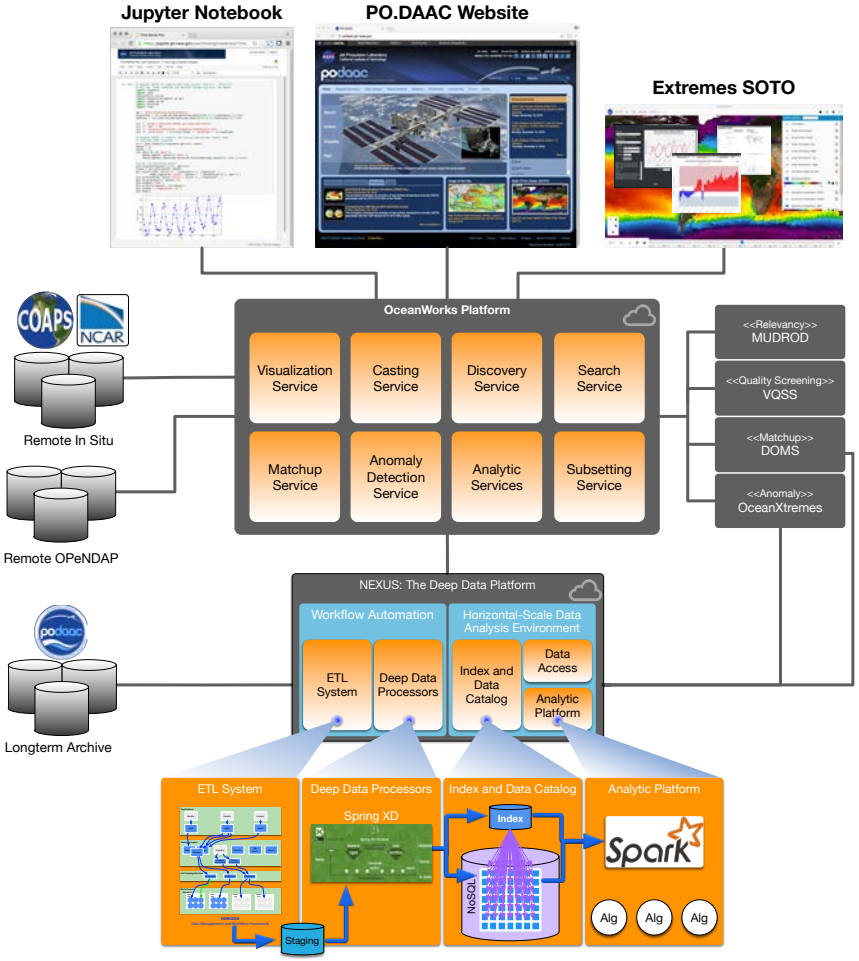
- Access and integration of highly distributed, heterogeneous data
- Novel statistical approaches for data integration and fusion
- Computation applied at the data sources
- Algorithms for identifying and extracting interesting features and patterns

THUANG/JPL. © 2017. All rights reserved.

Overview

- **OceanWorks** is an invited submission to the AIST-16 Adjunct Review as a Big Ocean Science technology integration, advancement and maturity effort
- **Collaboration between JPL, FSU, NCAR, and GMU**
- **Goal**: Develop and deploy an integrated cloud-based Big Ocean Science service platform for the PO.DAAC community
- Bringing together PO.DAAC-related big data technologies
  - **AIST-14 OceanXtremes (PI: Huang/JPL) – TRL 4**
    Anomaly detection and ocean science
  - **NEXUS (PI: Huang/JPL) – TRL 6**
    Deep data analytic platform
  - **AIST-14 DOMS (PI: Smith/FSU) – TRL 4**
    Distributed in-situ to satellite matchup
  - **AIST-14 MUDROD (PI: Yang/GMU) – TRL 6**
    Search relevancy and discovery
  - **ACCESS-13 VQSS (PI: Armstrong/JPL) – TRL 7**
    Virtualized Quality Screening Service





Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

THUANG/JPL. © 2017. All rights reserved.

Objectives

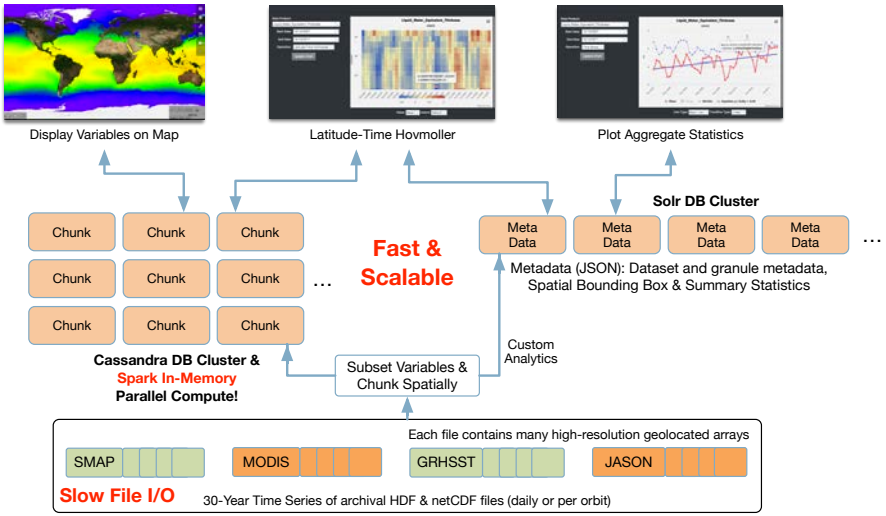
- **Improve Data Discovery**: help users to discover relevant data, anomalies and events, quality and uncertainty information, and provide multi-observing system matchup data. **PO.DAAC Search** webservice will be developed to demonstrate new data search and discover related matches and phenomenon
- **Subset and Distribute Data**: high performance data subsetting that supports quality screening
- **Identify and Catalog Ocean phenomenon**: registry of detected oceanographic phenomenon and published using datacasting technology
- **Matchup between Satellite and In-Situ Observations**: advanced matchup services that support in-situ to satellite colocation. It provides a mechanism for users to input a series of geospatial references for satellite observations and receive the in-situ observations that matched to the satellite data within a selectable temporal and spatial search domain.
- **Analyze Satellite Observations**: long time-series, correlation map, data decomposition (EOF), spectral analyses, time averaged map, climatological map, etc
- **Visualize and analyze Satellite Observation on the Web**: high performance data visualizations with linkages to actual source measurements, events and analytics. A **prototype integration** with PO.DAAC's new State of the Ocean (SOTO) web application will be developed, called Extremes SOTO, to demonstrate this objective.
- **API Integration** with Jupyter notebook to demonstrate working directly with OceanWorks' webserivce platform

THUANG/JPL. © 2017. All rights reserved.

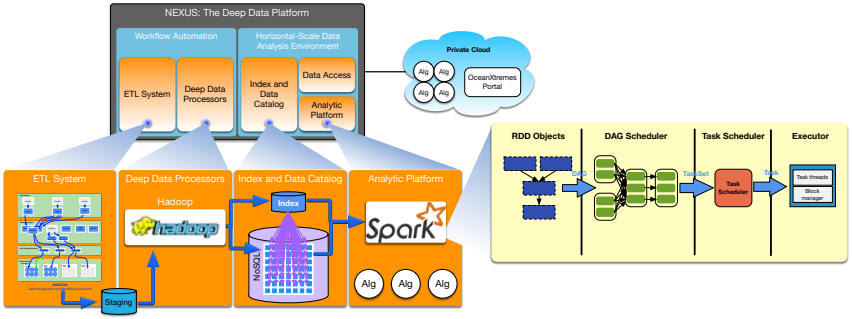
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NEXUS: Deep Data Platform

- 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 Hofmuller | Longitude/Time Hofmuller | Latitude/Longitude Time Average | Area Averaged Time Series | Time Averaged Map | Climatological Map | Correlation Map | Daily Difference Average



Two-Database Architecture



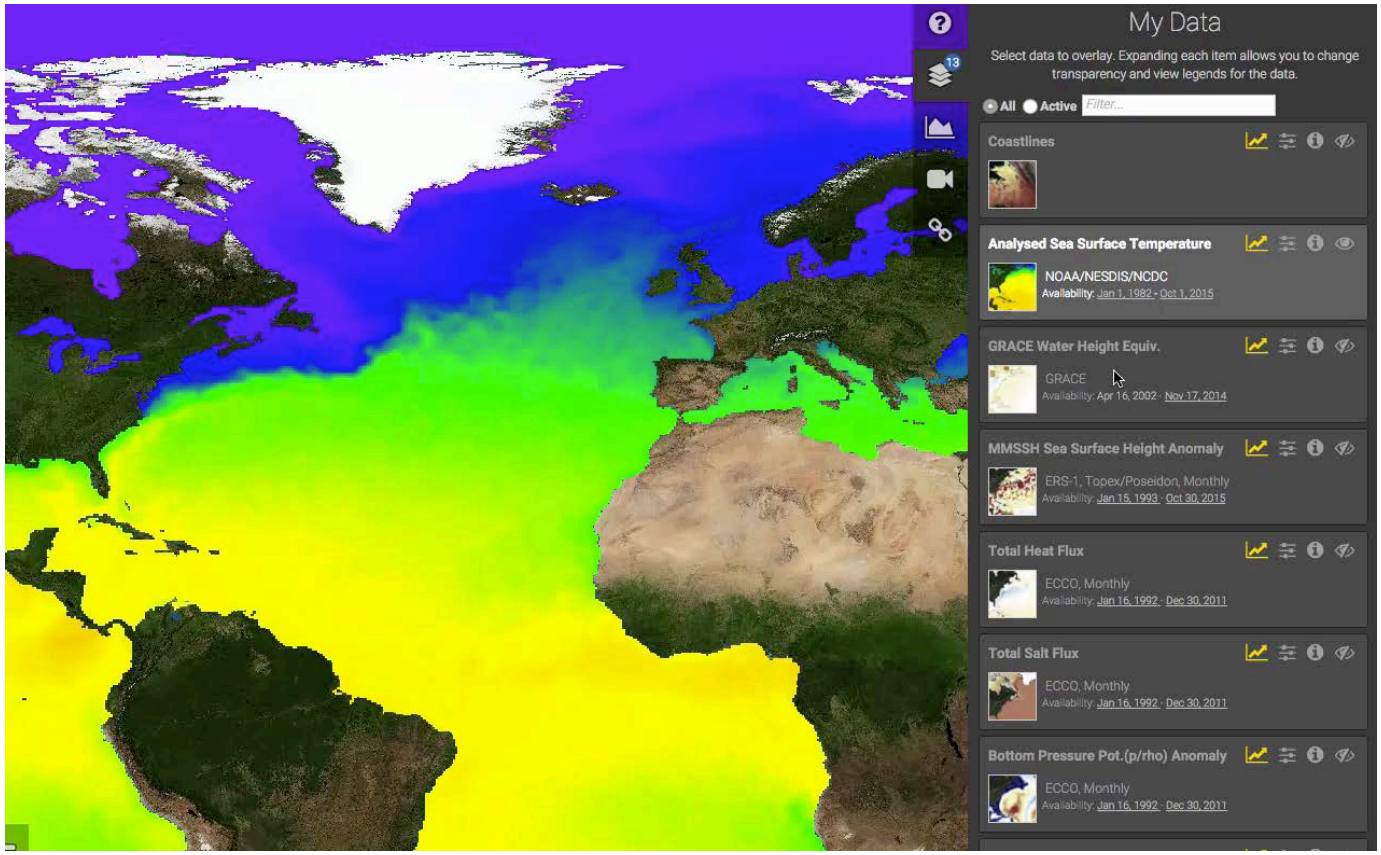
Deep Data Computing Environment (DDCE)

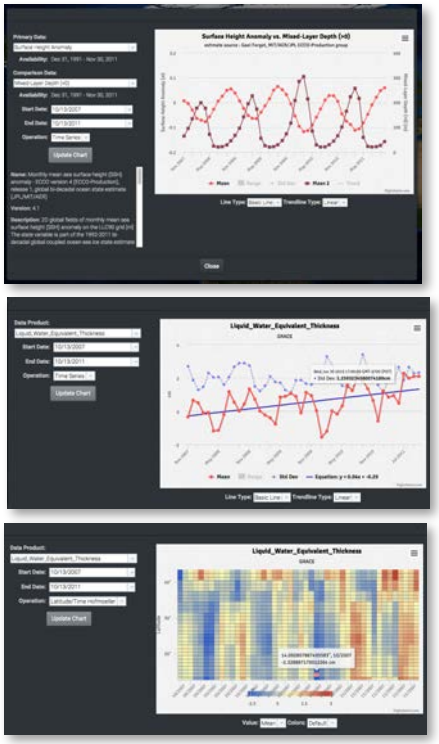
Open Source: Apache License 2
https://github.com/dataplumber/nexus

THUANG/JPL. © 2017. All rights reserved.

Earth Science Technology Office

Data Analysis Tool https://sealevel.nasa.gov





Facebook: 28K followers
Twitter: 22K followers

NASA Sea Level Change Website Offers Everything You Need To Know About Climate Change
http://www.techtimes.com/articles/147210/20160405/nasa-sea-level-change-website-offers-everything-need-know-climate.htm

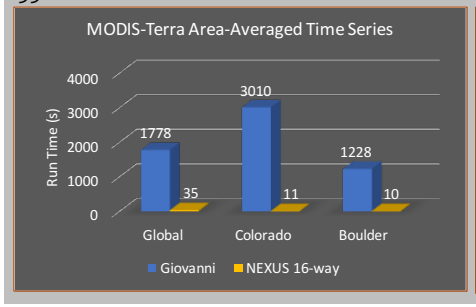
NASA's new sea level site puts climate change papers, data, and tools online
http://techcrunch.com/2016/04/04/nasas-new-sea-level-site-puts-climate-change-papers-data-and-tools-online/

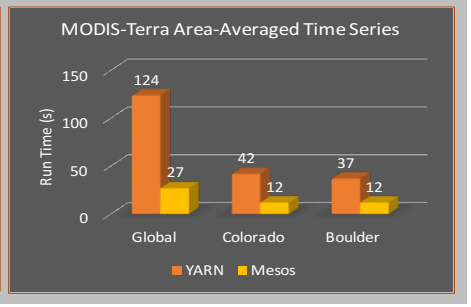
THUANG/JPL. © 2017. All rights reserved.

2016/2017 NEXUS Activities

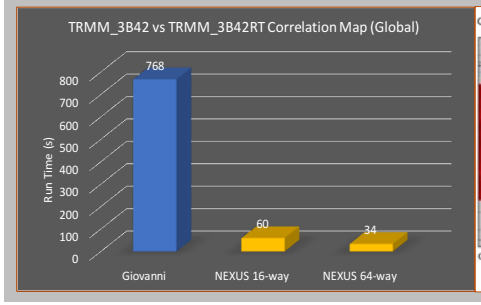
- **2016 Earthdata prototyping**
  - Twelve factor application
  - Benchmarking against Giovanni
  - Amazon cloud deployment
- **2016 ESTO Data Container Study**
  - Look at a variety of technologies for reorganizing and storing Earth science data to make them more tractable to full-scale science analysis.
  - To understand the strong points and tradeoffs of the different approaches to large-scale analysis
  - NEXUS' plug-an-play data storage
    - Apache Cassandra – Java implementation
    - ScyllaDB – C++ implementation
  - Task 1 – long time series
    - Point Time series for Boulder, CO
    - Area-averaged time series for the state of Colorado
    - Area Averaged time series for the globe
  - Task 2 – climatological map
  - Document ETL process, elapsed time, compute and storage
- **2017 Earthdata prototyping**
  - Auto ingestion
  - NASAcompilant Generation Application Platform (NGAP) Infusion
  - Cloud Analysis Toolkit to Enable Earth Science (CATEE)
  - 2017 ESIP Summer Workshops

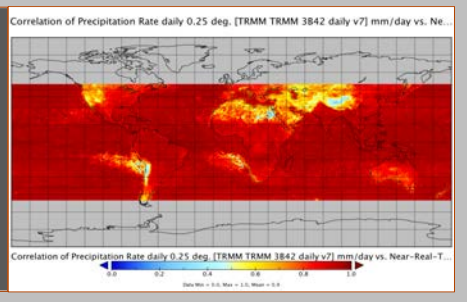
16-year area-averaged daily time series of MODIS-Terra Aerosol Optical Depth (AOD) 550 nm



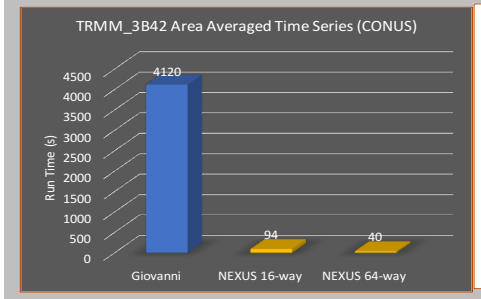


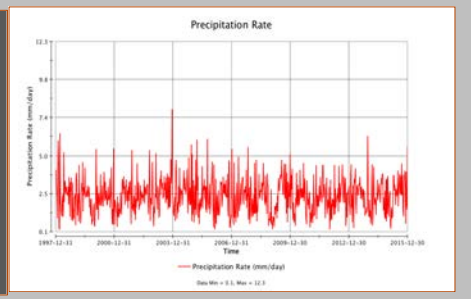
14-year global correlation map of 0.25 degree TRMM daily precipitation rate





17-year area-averaged time series over the continental United States

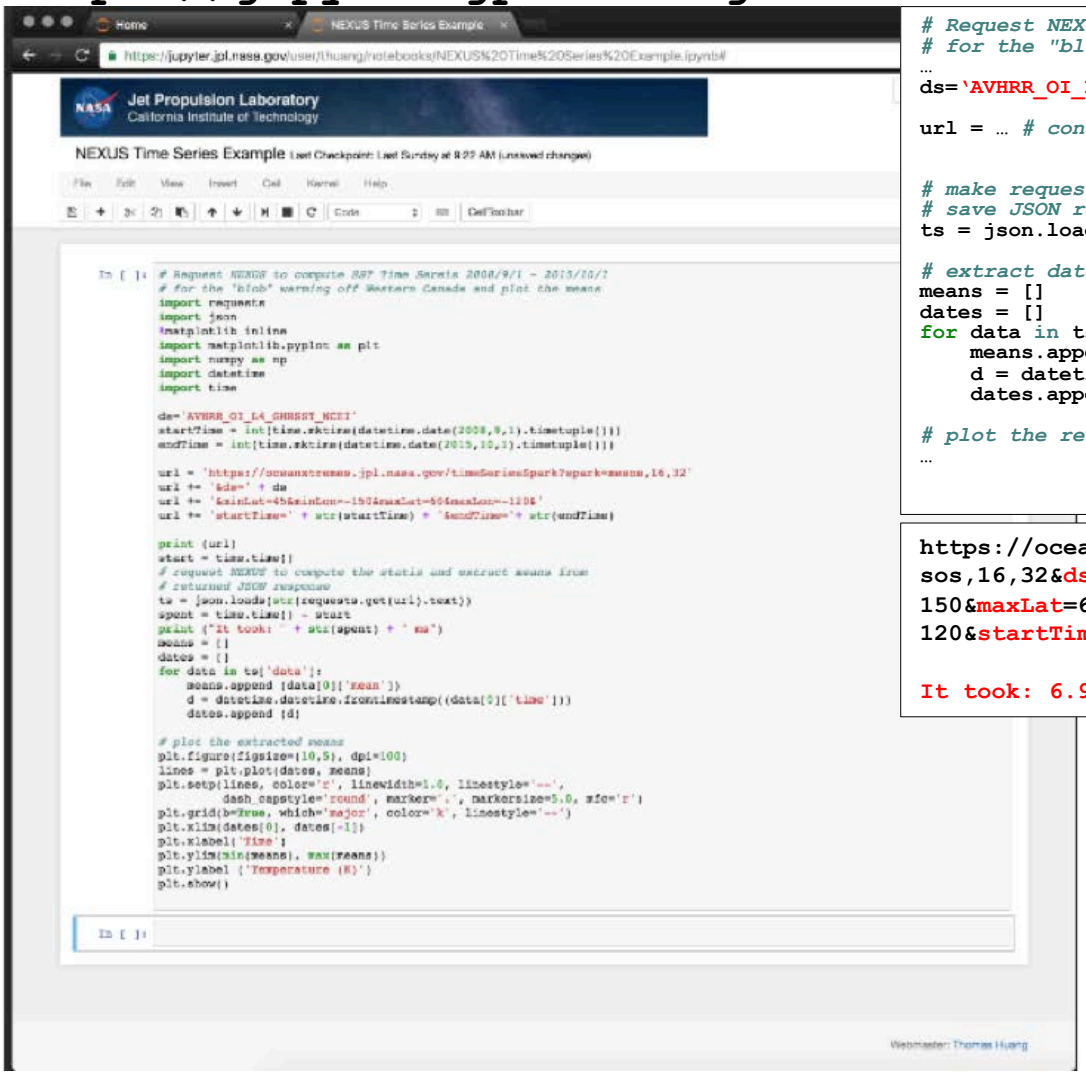




THUANG/JPL. © 2017. All rights reserved.

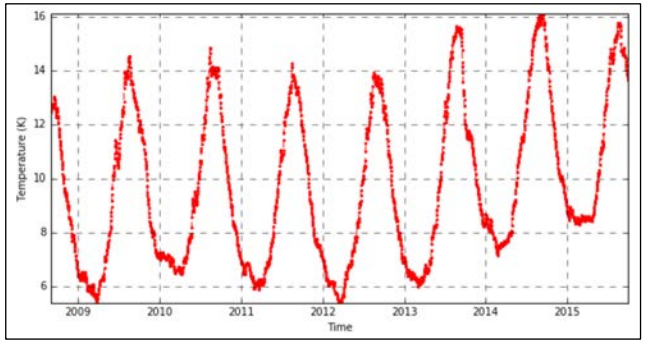
Jupyter Notebook Integration

https://jupyter.jpl.nasa.gov




```
# 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_GHRSST_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
…
```

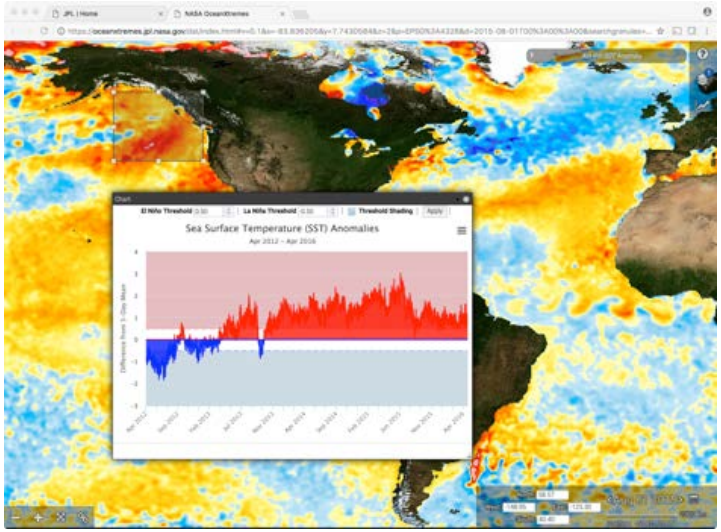
```
https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=1220227200&endTime=1443657600

It took: 6.909193992614746 ms
```



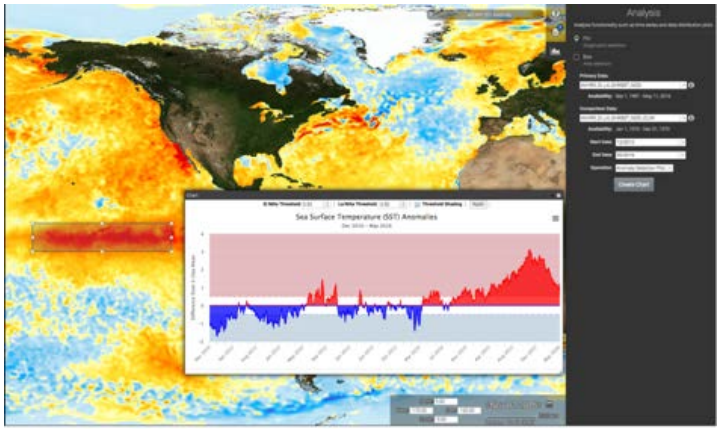
THUANG/JPL. © 2017. All rights reserved.

OceanXtremes

- https://oceanxtremes.jpl.nasa.gov
- An oceanographic data-intensive anomaly detection and analysis portal
- Cloud-based big data analytic platform for
  - Climatology generation
  - On-the-fly daily difference computation
  - Anomaly registry and publication
  - On-the-fly data analytics
- Recent highlights
  - **Recreated identification of "The Blob"**
    - **The Blob** is the name given to a large mass of relatively warm water in the Pacific ocean off the coast of North America. It was first detected in late 2013 and continued to spread throughout 2014 and 2015.
    - SST anomaly = SST – SST Climatology at each location to compare with standard deviation - Dr. Chelle Gentemann, Senior Scientist at Earth & Space Research
  - **Recreated the El Niño 3.4 regional signal**
    - **El Niño** is a phenomenon in the equatorial Pacific Ocean characterized by a five consecutive 3-month running mean of sea surface temperature (SST) anomalies in the Niño 3.4 region that is above (below) the threshold of +0.5°C (-0.5°C). This standard of measure is known as the Oceanic Niño Index (ONI).



The Blob



El Niño 3.4 regional signal

THUANG/JPL. © 2017. All rights reserved.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Hurricane Katrina Study – Using OceanXtremes

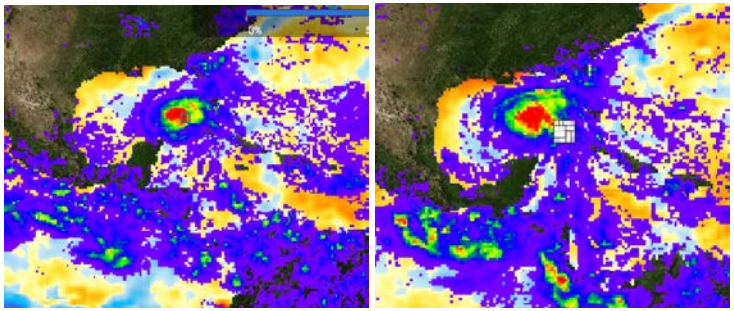
Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 deg C that lingers for several days. Following this was a short intense 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.

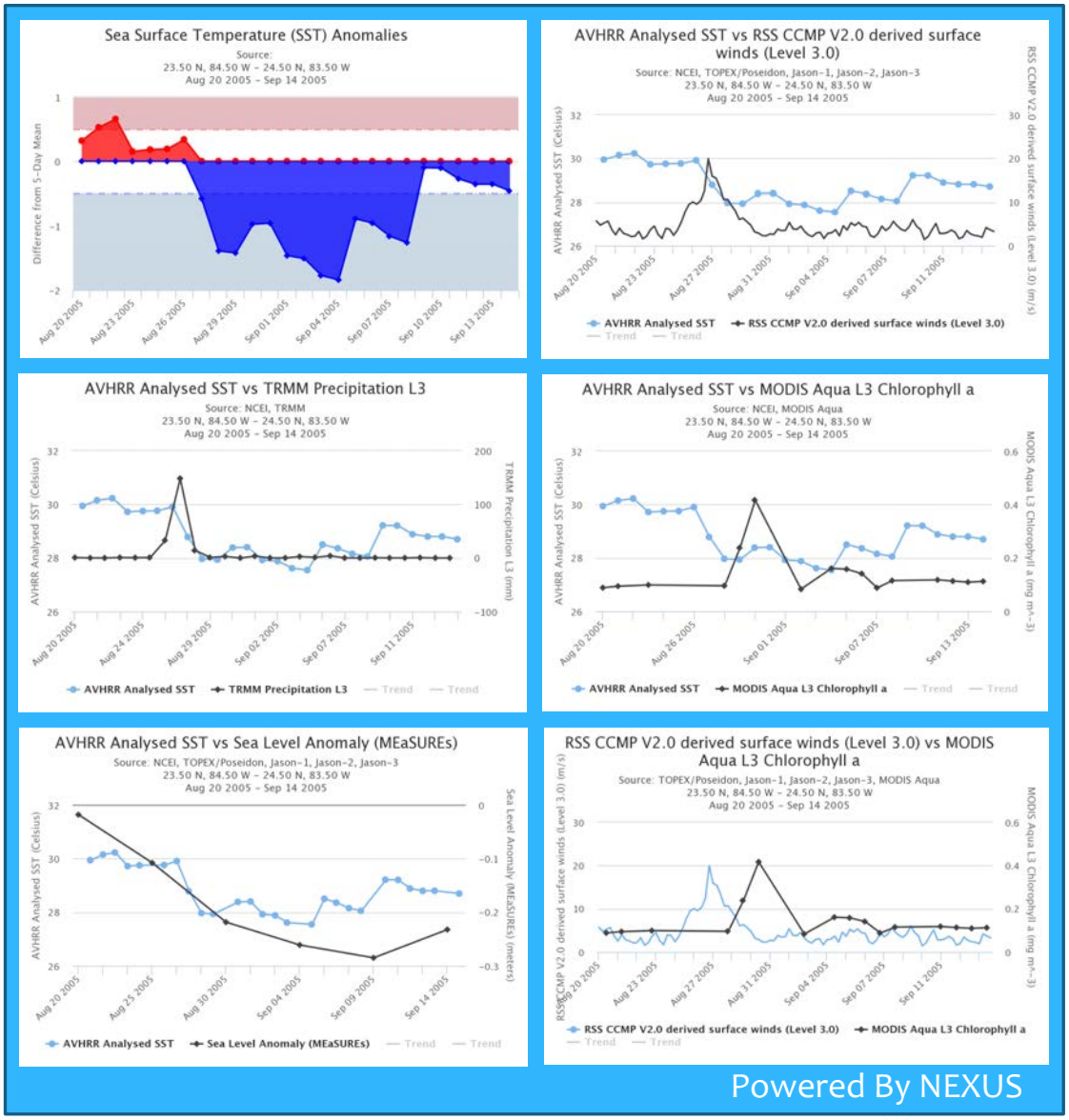
A study of a Hurricane Katrina–induced phytoplankton bloom using satellite observations and model simulations
Xiaoming Liu, Menghua Wang, and Wei Shi

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



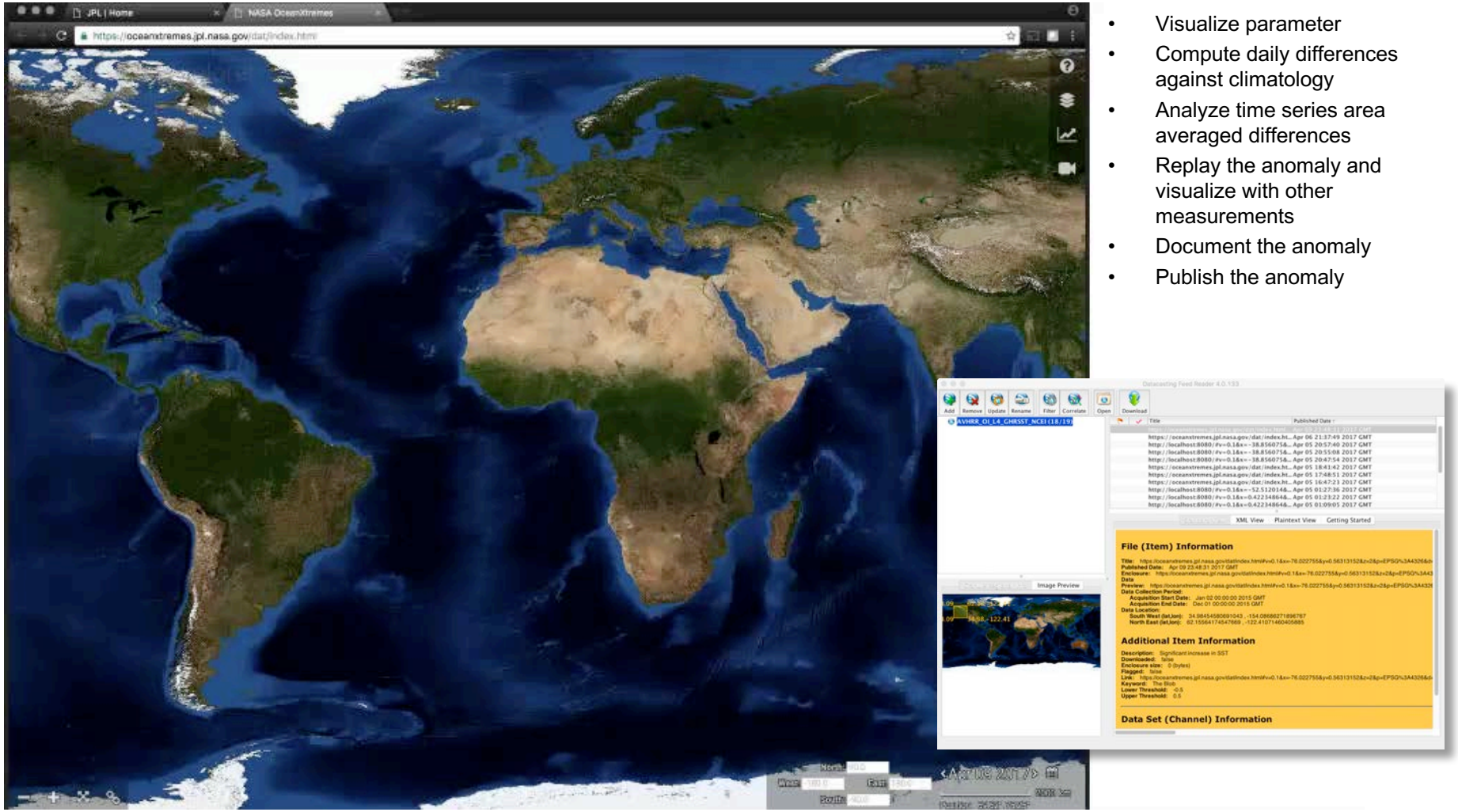
THUANG/JPL. © 2017. All rights reserved.

ESTO
Earth Science Technology Office

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OceanXtremes: Identify . Analyze . Share

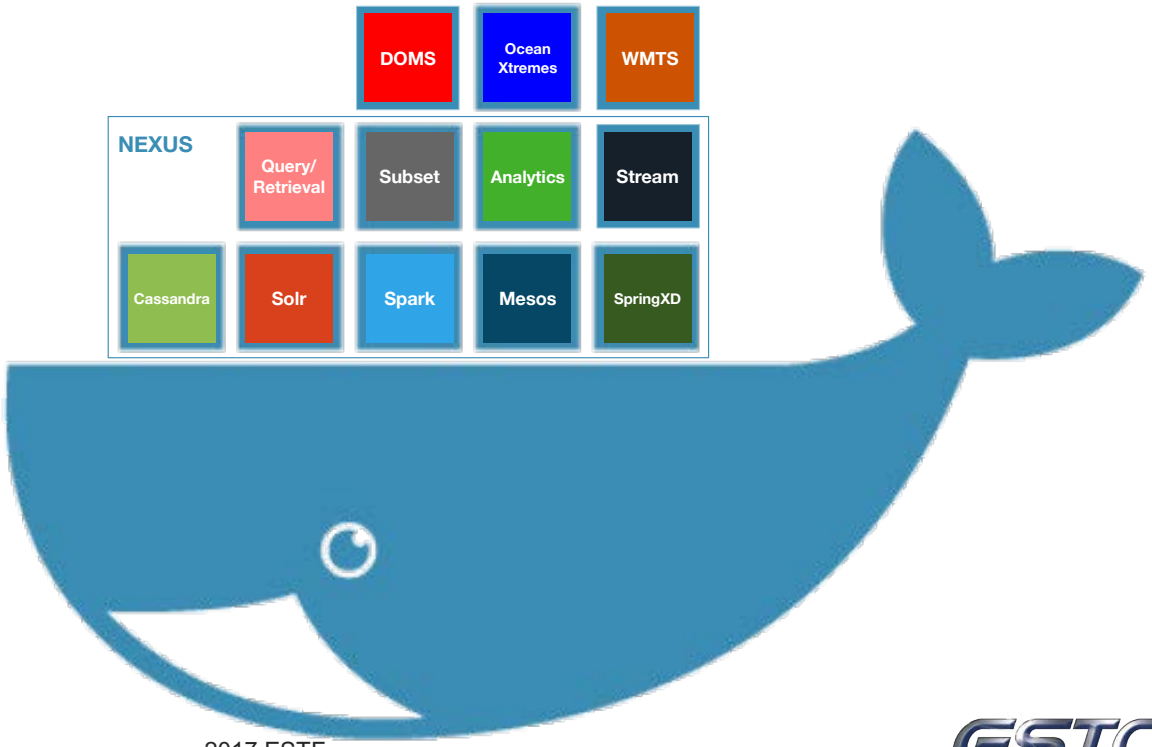


- Visualize parameter
- Compute daily differences against climatology
- Analyze time series area averaged differences
- Replay the anomaly and visualize with other measurements
- Document the anomaly
- Publish the anomaly

THUANG/JPL. © 2017. All rights reserved.

Containerization

- Dockerizing all services
- **Why?**
  - Rapid application deployment
  - Portability across machines
  - Application-centric vs machine/server-centric
  - Version control and component reuse
  - Secure due to isolation and encapsulation
  - Sharing
  - Lightweight footprint
  - Minimal overhead
  - Simplified maintenance



THUANG/JPL. © 2017. All rights reserved.

AMCE Onboarding

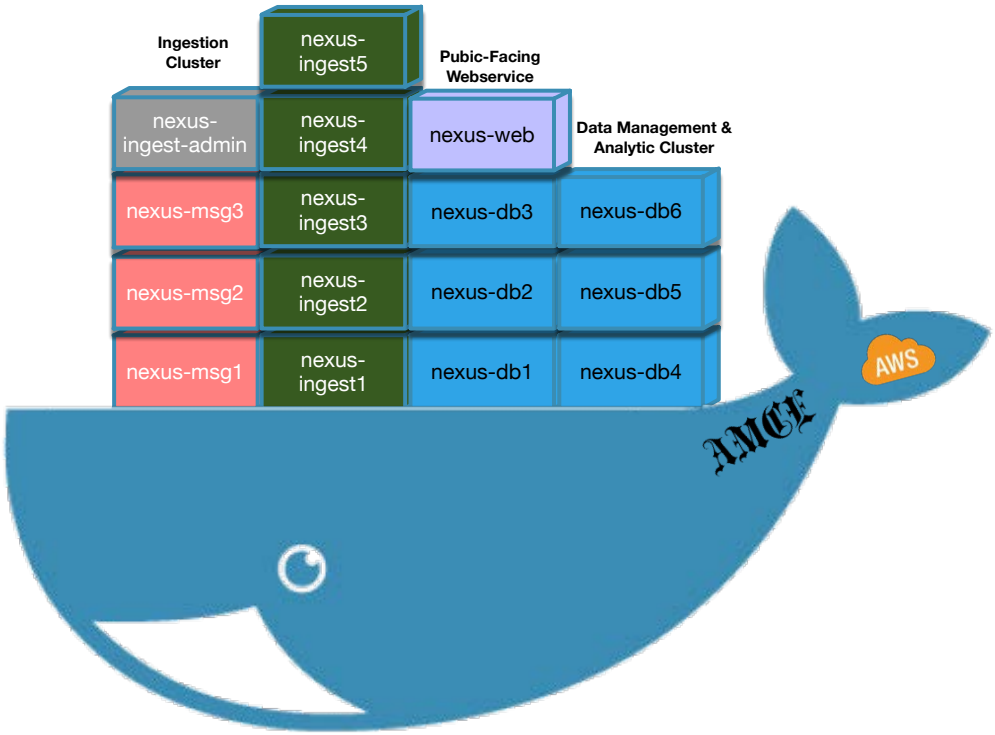
AMCE deployment

Multi-container, multi-cluster deployment

Leverage public DockerHub

Working toward dev-test pipeline automation

Deployed under 16 Amazon instances. Needs testing

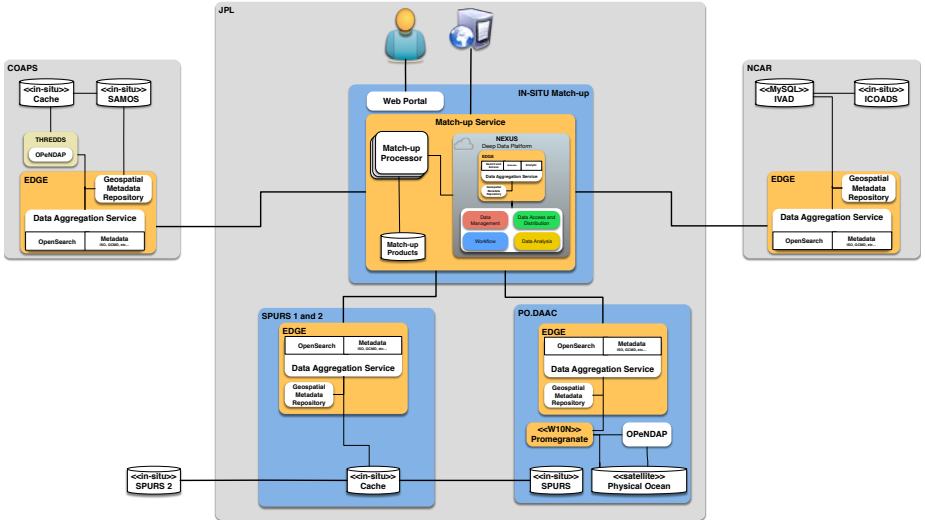


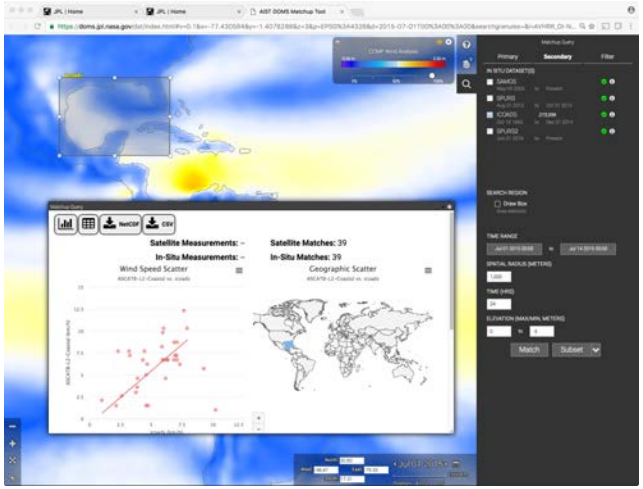
Name	Environment	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)
nexus-db1		i-06d7e7340bf32c990	r4.4xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-193-67-220.compute-1.a
nexus-db2		i-0fc92d0b326c3c452	r4.4xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-196-189-54.compute-1.a
nexus-db4		i-00a9f1c7b7f80017d	r4.4xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-198-153-52.compute-1.a
nexus-ingest-admin		i-03eae7f858cc52ee0	t2.medium	us-east-1a	running	2/2 checks passed	None	ec2-52-54-3-52.compute-1.amaz
nexus-db3		i-05533f514e80fe79e	r4.4xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-197-79-203.compute-1.a
nexus-db5		i-0ed67a7ad6f25a90e	r4.4xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-199-25-89.compute-1.a
nexus-db6		i-093edbad62b90b553	r4.4xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-199-51-137.compute-1.a
nexus-ingest-msg1		i-0b051ce88a6dae59d	r4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-203-43-74.compute-1.a
nexus-ingest-msg2		i-06b032315130b87816	r4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-203-131-208.compute-1
nexus-ingest-msg3		i-0cf40b22572e55768	r4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-203-72-84.compute-1.a
nexus-ingest2		i-06f48714c55e294bd	m4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-205-79-138.compute-1.a
nexus-ingest3		i-07e36cfb94fda197c	m4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-207-20-109.compute-1.a
nexus-ingest4		i-0a5feea6febe93fe8	m4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-239-234-252.compute-1
nexus-ingest5		i-01f71638c6011f6805	m4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-239-244-209.compute-1
nexus-ingest1		i-0a387d62e9e1e19a4	m4.xlarge	us-east-1a	running	2/2 checks passed	None	ec2-34-203-86-57.compute-1.a
nexus-web		i-03f2a2b73e1afd07d	t2.medium	us-east-1a	running	2/2 checks passed	None	ec2-34-205-91-196.compute-1.a

THUANG/JPL. © 2017. All rights reserved.

DOMS

- https://doms.jpl.nasa.gov
- Distributed Oceanographic Matchup Service
- 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.
- Data nodes at JPL, NCAR, and FSU operational.
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Prototype graphical user interface (UI) and APIs accessible for external users.
- Plugin architecture for In-situ data source using EDGE
  - Extensible Data Gateway Environment is an Apache License 2 open source technology
  - https://github.com/dataplumber/edge
- Defined specification for packaging matchup results. Working with Unidata and ESDSWG's data interoperability and standard groups



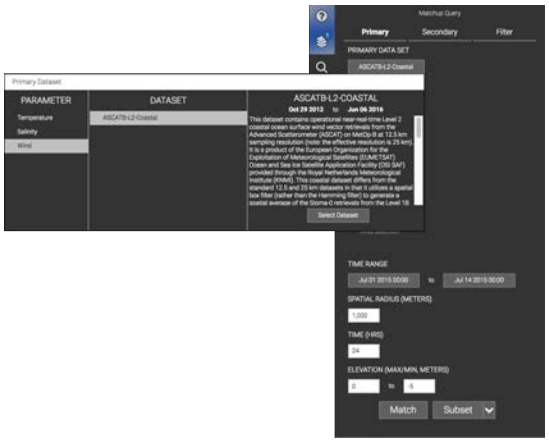


THUANG/JPL. © 2017. All rights reserved.

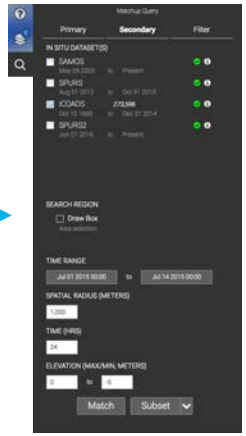
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

More than just matchup

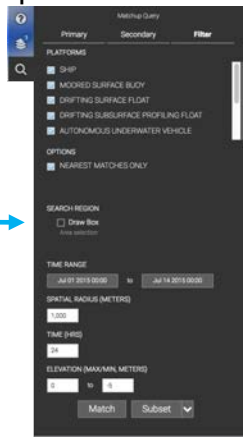
Primary Dataset



Match-up In-Situ



Optional Platform



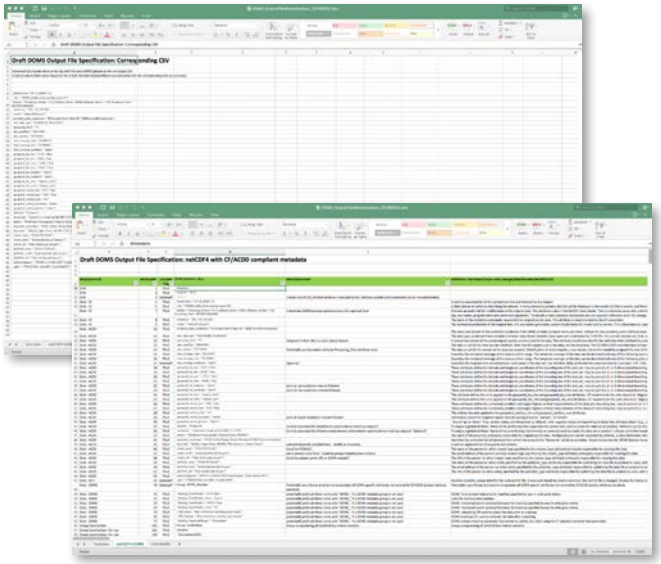
- **Workflow driven**
- **Fast on-the-fly subseting satellite and in-situ**
- **Open webservice API**
- **ON-the-fly data analysis using NEXUS**
- **Cloud ready**

PO.DAAC Satellite Data

- AVHRR_OI_L4_GHRSST_NCEI
- JPL-L4-GHRSST-SSTfnd-MUR-GLOB-v02.0fv04.1
- SMAP_L2B_SSS
- ASCATB-L2-Costal

In-situ Data

- JPL SPURS 1 and 2
- FSU SAMOS
- NCAR ICOADS



Drafted and implemented matchup specification

- Promote CF and ACDD standards
- Self-contained
- Simple to use by other tools and services

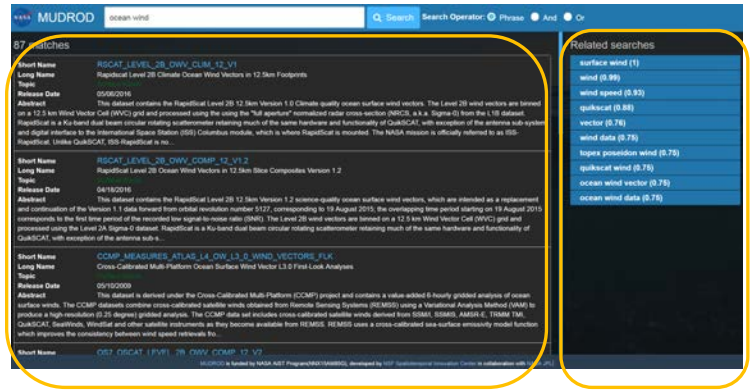
THUANG/JPL. © 2017. All rights reserved.

ESTO Earth Science Technology Office

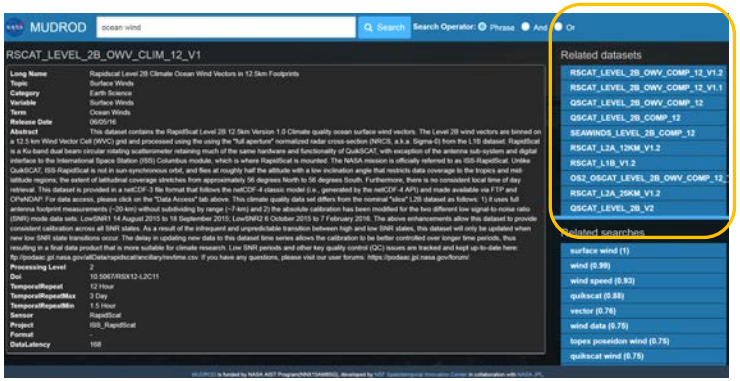
MUDROD

- 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

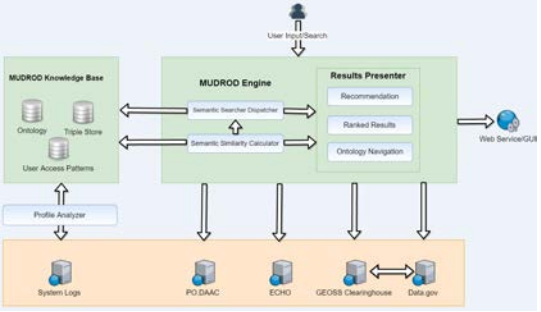
https://mudrod.jpl.nasa.gov



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

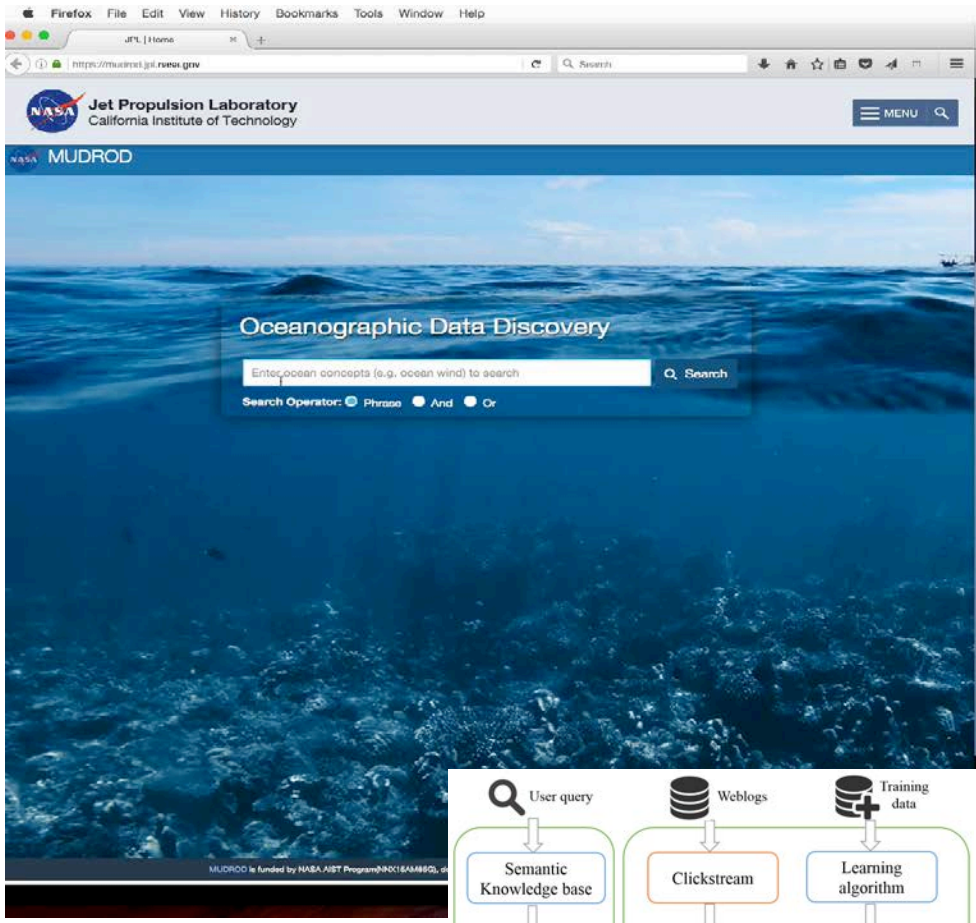


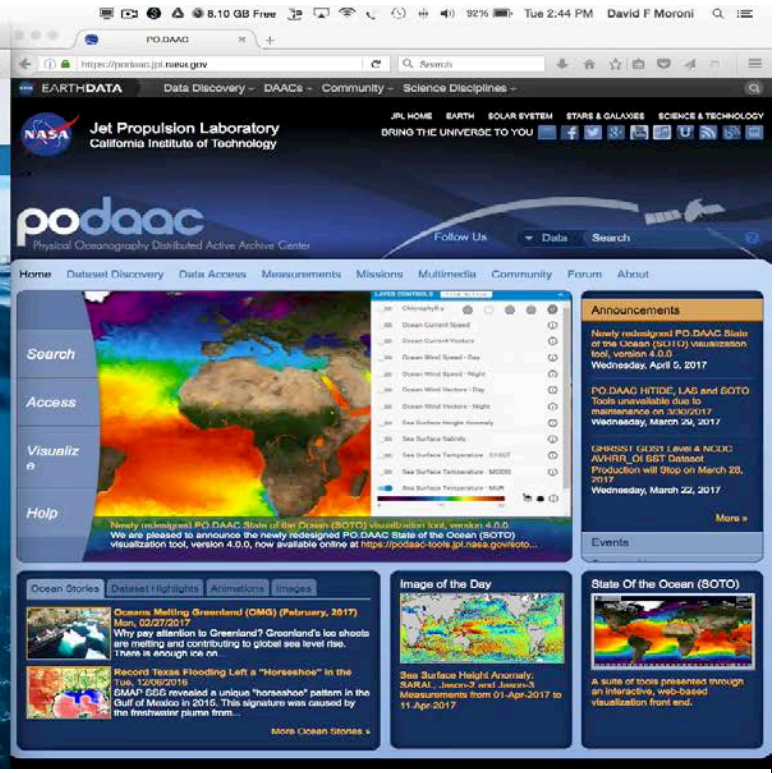
AWS

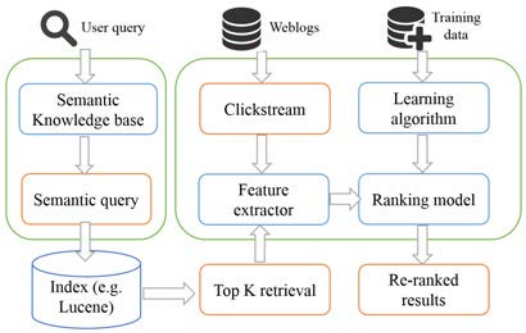
THUANG/JPL. © 2017. All rights reserved.

ESTO Earth Science Technology Office

Speed . Relevant . Discovery







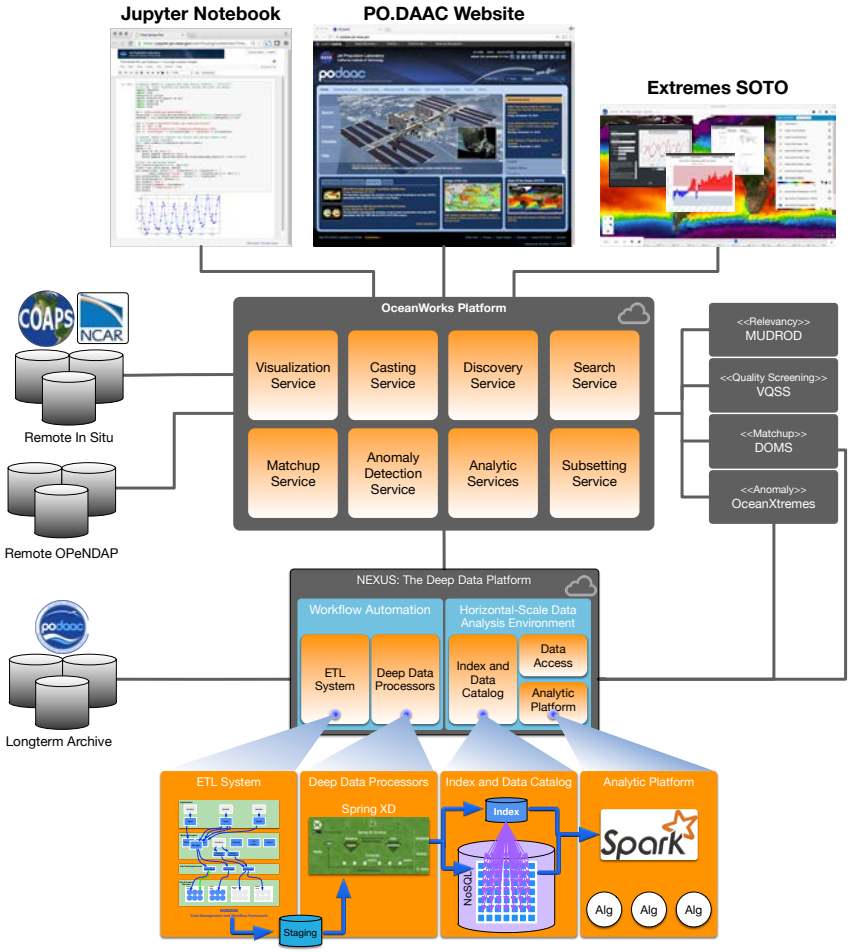
Features

- Fast web log ingestion and processing using Apache Spark, in-memory MapReduce
- Session reconstruction
- Vocabulary semantic relationship extraction
- Machine Learning Search ranking
- Integration with SWEET Ontologies for semantic-driven search and recommendations
- Recommendation

THUANG/JPL. © 2017. All rights reserved.

The Game Changer in NASA Ocean Science

- Lives on the Cloud
- Auto-Scaling
- One-The-Fly multi-parameter data analysis
- Access and matchup with in-situ measurements
- Smart subsetting
- Anomaly detection and registration
- Sharing of analytic results
- Lighting speed searches
- Discover relevant data, services, news and publications like never before
- Fully Open Source



THUANG/JPL. © 2017. All rights reserved.

OceanWorks Schedule – Tentative

- **Goal**: Develop and deploy an integrated cloud-based Big Ocean Science service platform for the PO.DAAC community
- OceanWorks will leverage the AIST Managed Cloud Environment (AMCE) for development – the AIST-provisioned Amazon Cloud environment
  - NEXUS/OceanXtremes/DOMS – by end of May 2017
  - MUDROD – already deployed

• System design and dataset selection	09/17
• Initial integration	12/17
• Extremes SOTO initial integration	12/17
• **PO.DAAC UWG review**	**04/18**
• PO.DAAC Search	09/18
• Subsetting service	09/18
• Performance optimization	09/18
• Jupyter notebook integration	12/18
• Extremes SOTO update	12/18
• **PO.DAAC UWG review**	**04/19**
• PO.DAAC UAT deployment	04/19

THUANG/JPL. © 2017. All rights reserved.

Credits

NASA
Mike Little
Chris Lynnes
Kevin Murphy

FSU COAPS
Mark Bourassa
Jocelyn Elya
Shawn Smith
Adam Stallard

NCAR
Steve Worley
Ji Zaihua

GMU
Yongyao Jiang
Chaowei (Phil) Yang

JPL NEXUS Engineers
Stewart (Parker) Abercrombie
Kevin Gill
Frank Greguska
Joseph Jacob
Nga Quach
Brian Wilson

JPL Science Contributors
Ed Armstrong
Andrew Bingham
Carmen Boening
Mike Chin
Michelle Gierach
Ben Holt
Tony Lee
David Moroni
Rob Toaz
Vardis Tsontos
Victor Zlotnicki

Questions, and more information

Thomas.Huang@jpl.nasa.gov

THUANG/JPL. © 2017. All rights reserved.

WIP: AGU Wiley Book

Thomas Huang, JPL

Phil Yang, GMU

Chris Lynnes, NASA

Tiffany Vance, NOAA

THUANG/JPL. © 2017. All rights reserved.