



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California



Building a better search engine for earth science data

Ed Armstrong, Chaowei Yang, David Moroni, Thomas Huang, Lewis Mcgibney,
Frank Greguska, Yongyao Jiang, Yun Li, Christopher Finch

Physical Oceanographic DAAC

NASA Jet Propulsion Laboratory, Pasadena, CA

George Mason University, Fairfax, VA

2017 Fall AGU

IN13D-06

New Orleans, LA

11 Dec 2017

© 2017 All rights reserved



NASA Missions & Projects

Seasat, TOPEX/Poseidon, Jason-1, NSCAT, SeaWinds on ADEOS-II, QuikSCAT, GRACE, GHRSSST, SPURS, MEaSUREs, Aquarius, CYGNSS, GRACE-FO (2017)

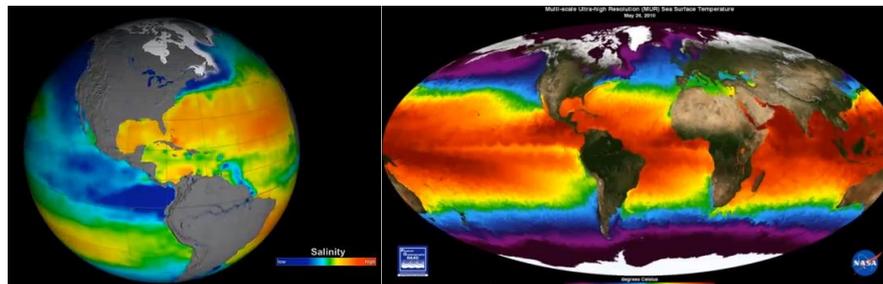
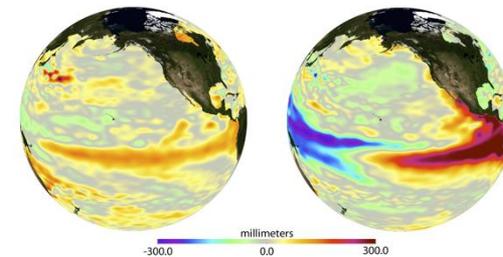
Upcoming: COWVR, AirSWOT, SWOT, GRACE-2



Ocean & Climate Community Driven

Value-added datasets in support of NASA programs

- Gravity
- Ocean Circulation & Currents
- Ocean Surface Salinity
- Ocean Surface Topography
- Ocean Vector Winds
- Sea Surface Temperature
- Hydrology*
- Ocean Color*
- Sea Ice*





Select Filter

- Processing Levels**
 - Any processing level
 - Level-2 (Swath) (2)
 - Level-3 (Grid) (10)
 - Level-4 (Blended) (6)
- Swath Spatial Resolution**
 - Any swath spatial resolution
 - 8 km (2)
- Grid Spatial Resolution**
 - Any grid spatial resolution
 - 0.01 degree(s) (2)
 - 0.25 degree(s) (13)
 - 2.66 degree(s) (1)
- Temporal Resolution**
 - Any temporal resolution
 - 1 Month (1)
 - 10 day (1)
 - 10 day re (1)
 - 10-hourly (1)
 - 5 Days (1)
 - 6 Hourly (1)
 - Daily (2)
- Parameter**
 - Any param
 - Ocean Te
 - Ocean W
 - Sea Surfa
- Collection**
 - Any collec
 - MEASURE
- Platform**
 - Any platf
 - ADEOS-II
 - AQUA (8)
 - Coniste (8)
 - DMSP-F0
 - DMSP-F1
 - DMSP-F1
 - DMSP-F1
- Sensor**
 - Any sens
 - AMR (3)
 - AMSR-E
 - AVHRR-3
 - JMR (3)
 - MODIS (2)
 - POSEIDON
 - POSEIDON

Dataset Discovery

Found 18 matching dataset(s).

Need help selecting a dataset?
Contact a PO.DAAC Data Engineer

Advanced search

View mode: [List] [Grid]

Sort By: Popularity (All Time)

Prev 1 2 Next

1

GHRSSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (JPL-L4JHfnc-GLOB-MUR)
Ocean Temperature
Platform/Sensor: AQUA/AMSR-E, AQUA/MODIS, NOAA-18/AVHRR-3 ... more
Processing Level: 4

Dataset Information Page

- * Information
 - * Dataset Metadata
- * Data Access
 - * Direct Access
 - * Tools and Services
 - * Read Software
- * Documentation
 - * Known Issues
- * Granule (File) Listing
- * Citation

Dataset Discovery

- Faceted Browsing
- Multi-level filtering
- **Keyword search**
- Dataset Information Page/DOI Landing Pages
- Granule browsing through date tree

Select Filter

- Parameter
 - Atmospheric Electricity (3)
 - Atmospheric Water Vapor (2)
 - Geoid/GRAVITY (87)
 - Humidity Index (1)
 - Microwave (1)
 - Ocean Chemistry (1)
 - Ocean Circulation (5)
 - Ocean Heat Budget (1)
- Show More
- Collection
- Platform
- Sensor

All Products

GHRSSST Level 2P Global Skin Sea Surface Temperature from the Moderate Resolution Imaging Spectroradiometer (MODIS) on the NASA Aqua satellite

SHARE THIS PAGE
http://podaac.jpl.nasa.gov/dataset/JPL-L2P-MODIS_A

Please contact us if there are any discrepancies or inaccuracies found below.

Information | Data Access | Granule (File) Listing | Citation

DOI 10.5067/GHMDA-2PJ01

Short Name JPL-L2P-MODIS_A

Description
The Moderate-resolution Imaging Spectroradiometer (MODIS) is a scientific instrument (radiometer) launched by NASA in 2002 on board the Aqua satellite platform (a second series is on the Terra platform) to study global dynamics of the Earth's atmosphere, land and oceans. MODIS captures data in 36 spectral bands ranging in wavelength from 0.4 um to 14.4 um and at varying spatial resolutions (2 bands at 250 m, 5 bands at 500 m and 29 bands at 1 km). For the sea surface temperature (SST) products from this radiometer channels in the 4, 11 and 12 um spectrum are used. The Aqua platform is in a sun synchronous, near polar orbit at 705 km altitude and the MODIS instrument images the entire Earth every 1 to 2 days. The production of the MODIS L2P SST data as part of the Group for High Resolution Sea Surface Temperature (GHRSSST) is a joint collaboration between the NASA Jet Propulsion Laboratory (JPL), the NASA Ocean

- Driven by Solr/Lucene index of PO.DAAC metadata
 - Limited search factors: term frequency (pre-defined keyword list), inverse document frequency, and dataset popularity
 - Implements a default “OR” between keywords
 - Suffers from low precision
 - E.g., the search will often return good number of datasets (reasonable recall) but a low number relevant datasets (precision is poor)
 - “OR” syntax often returns unrelated datasets
 - Incomplete indexing. Newer versions, release date, processing levels etc. not considered
 - User popularity (unique users) is an imperfect factor



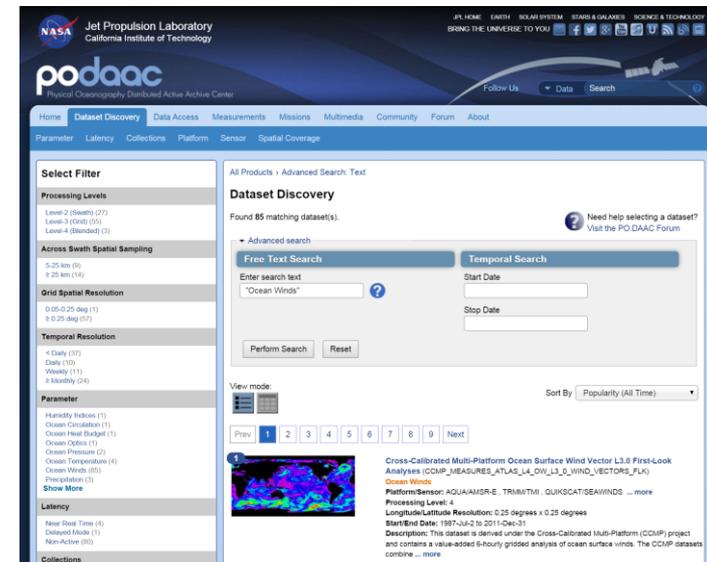
- *.....Improve search and discovery of PO.DAAC dataset via free text (.e.g., keyword) and facets*
- *.....Develop advanced search capabilities*
- Ongoing recommendation over past several years
- While faceted search provides a systematic approach to group data artifacts, facets are still static and rely on manual keywords tagging.
- Search relevance requires multi-faceted dynamic ranking of data
- Ranking is a long-standing problem in geospatial data discovery

- **Mining and Utilizing Dataset Relevancy from Oceanographic Data (MUDROD)**
 - 2014 funded NASA Advanced Information Systems Technology (AIST) project
 - Technology Readiness Level development from an approximately Level 4 to Level 6-7
 - Specifically targeted to improve search relevance for earth science data in the PO.DAAC
 - Improving the capabilities of PO.DAAC search
 - Built on services previously implemented for the hydrology community



Data Discovery Problems

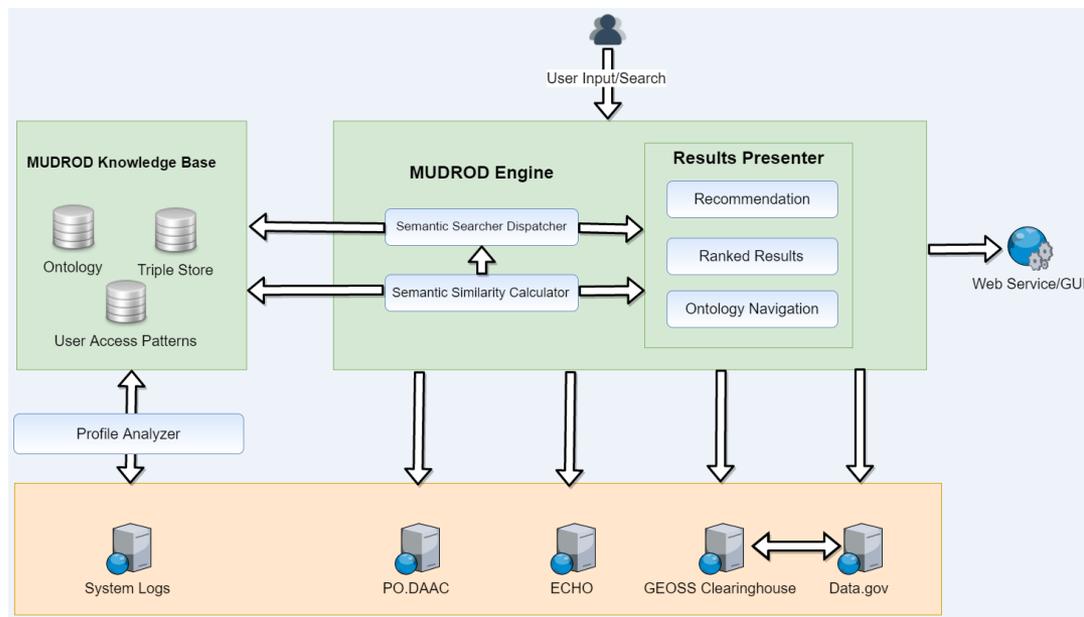
- Keyword-based matching (traditional search engines)
 - User query: **ocean wind**
 - Final query: ocean AND wind
- Reveal the real intent of user query
 - ocean wind = “**ocean wind**” OR “**greco**” OR “**surface wind**” OR “**mackerel breeze**” ...
- PO.DAAC UWG Recommendation 2014-07
- NASA ESDSWG Search Relevance Recommendations 2016 & 2017



- Rank most recent versions of datasets higher
- Rank new mission dataset higher
- Allow user choice of “AND” vs “OR” or phrase keyword syntax
- Improve search across different ocean variables
- Find (and rank) related PO.DAAC datasets
- Prioritize datasets that have been vetted by “domain experts”
- Consider user search intent, e.g.
 - Climate users v.s. real time applications users
 - High spatial resolution v.s. low spatial resolution

Objectives

- Analyze **web logs** to discover user knowledge (query and data relationships)
- Construct **knowledge base** by combining semantics and profile analyzer
- Improve data discovery by 1) better **ranking**; 2) **recommendation**; 3) **ontology navigation**



Tech Advance (four technological modules)

- Web log preprocessing
- Semantic analysis of user queries & Navigation
- Machine learning based search ranking
- Data Recommendation

- Put the most desired data to the top of the result list
- What **features** can represent users' search preferences for geospatial data?
- How can the ranking function reach a **balance** of all these features?
- Identified eleven features from
 - Geospatial metadata attributes
 - Query – metadata content overlap
 - User behavior from web logs



Ranking features – Metadata attributes

Features	Description
Release date	The date when the data was published
Processing level (PL)	The processing level of image products, ranging from level 0 to level 4.
Version number	The publish version of the data
Spatial resolution	The spatial resolution of the data
Temporal resolution	The temporal resolution of the data

- Five metadata features
- Verified by domains experts
- Query-independent: static, depends on the data itself, won't change with the query

RankSVM

- One of the well-recognized ML ranking algorithm
- Convert a **ranking** problem into a **classification** problem that a regular SVM algorithm can solve
- 3 main steps
 - 1) Standardize: mean = 0, std = 1
 - SVM is not scale invariant
 - Over-optimized
 - Longer to train
 - 2) For any pair of training data, calculate the difference
 - 3) A ranking problem becomes a binary classification problem, where SVM is applied to find the **optimal decision boundary**

- Add more features (e.g., temporal similarity)
- Create training data from web logs for RankSVM
- Develop a query understanding module to better interpret user's search intent (e.g. "ocean wind level 3" -> "ocean wind" AND "level 3")
- Support near real-time data ingestion to dynamically update knowledge base
- Integration with DOMS and OceanXtremes to support an ocean science analytics center
- Leverage advanced computing techniques to speed up the process



MUDROD deployed in PO.DAAC Labs

- https://podaac.jpl.nasa.gov/podaac_labs

Explore New Ideas, Prototypes and Tools.
Offer feedback directly to the engineers who developed them

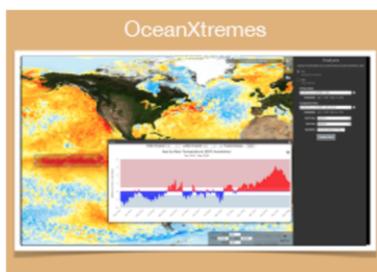
For questions or to provide feedback, please visit the [PO.DAAC Forum](#)



Mining and Utilizing Dataset Relevancy from Oceanographic Datasets to Improve Data Discovery (MUDROD)

MUDROD is collaborative effort between George Mason University and NASA JPL to improve the search and relevancy ranking of oceanographic data via a simple search interface and powerful backend services. MUDROD has mined and utilized the combination of Earth Science dataset metadata, usage metrics, and user feedback to objectively extract relevance for improved data discovery and access at the Physical Oceanographic Distributed Active Archive Center (PO.DAAC). In addition to improved dataset relevance and ranking, the MUDROD search engine also returns recommendations to related datasets and related user queries.

For questions or to provide feedback, please visit the [MUDROD Forum](#).



OceanXtremes: Oceanographic Data-Intensive Anomaly Detection and Analysis Portal

OceanXtremes is a computational platform powered by an intelligent, Cloud-based analytic service backend that enables execution of domain-specific, multi-scale anomaly and feature detection algorithms across the entire archive of ocean science datasets. Using this platform scientists can efficiently search for anomalies or ocean phenomena, compute data metrics for events or over time-series of ocean variables, and efficiently find and access all of the data relevant to their study (and then download only that data).

For questions or to provide feedback, please visit the [OceanXtremes Forum](#).

- Log mining enables a data portal integrating implicit user preferences
- Word similarity retrieved by data mining tasks expands any given query to improve search recall and precision.
- The rich set of ranking features and the ML algorithm provide substantial advantages over using other ranking methods
- The recommendation algorithm can discover latent data relevancy
- The proposed architecture enables the loosely coupled software structure of a data portal and avoids the cost of replacing the existing system



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



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Backups





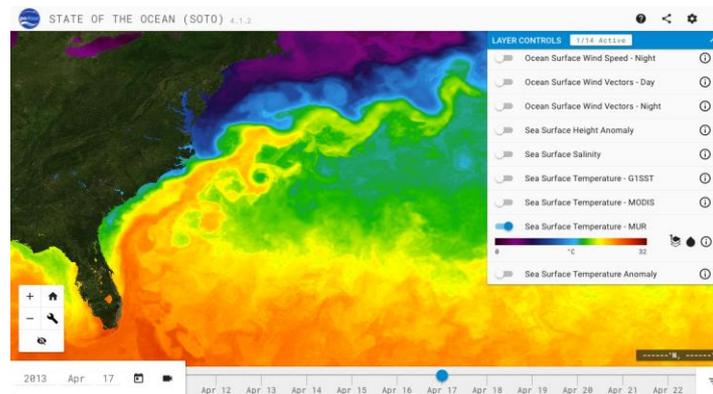
Data Management & Stewardship

Preserve NASA's data for the benefit of future generations



Data Access

Provide intuitive services to discover, select, extract and utilize data



Science Information Services

Provide a knowledgebase to help a broad user community understand and interpret satellite ocean data and related information

HELP

Questions? Answers.

Ask PO.DAAC. Once click access to:

ABOUT US ACRONYMS GLOSSARY DATA TERMINOL & FORMAT



Parameter	Retired/Retiring	Ongoing	New	Future
SSS	Aquarius	SPURS I	SPURS II SMAP SSS	
SST	GHRSSST (GDS)	GHRSSST (GDS2) MODIS GOES VIIRS	Sentinel-3a	Sentinel-3b VIIRS/JPSS-1
Gravity	GOES-3	GRACE		GRACE-FO
OST	TOPEX/Poseidon Jason-1	Jason-2/OSTM	Jason-3 Sentinel-3a OMG	SWOT Sentinel-3b Sentinel-6/Jason-CS
OVW	SeaSat NSCAT AMSR-E Seawinds on ADEOS-2	WindSat QuikSCAT RapidScat ASCAT	CYGNSS	MetOp-C COWVR
Ocean Circulation		OSCAR		



PO.DAAC Web, Tools and Services

