

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California



Improving search relevancy for oceanographic data discovery

Ed Armstrong¹, Chaowei Yang², David Moroni¹, Thomas Huang¹, Lewis McGibney¹,
Frank Greguska¹, Yongyao Jiang², Yun Li², Christopher Finch¹

¹Physical Oceanographic DAAC
NASA Jet Propulsion Laboratory, Pasadena, CA
²George Mason University, Fairfax, VA

19th GHRSSST Science Team Meeting
Darmstadt, Germany
7 June 2018

© 2018 All rights reserved

- Dataset Ranking is a long-standing problem in geospatial data discovery...data diversity and heterogeneity, user search intent
- Determining best and more relevant dataset is important
 - Saves time and less dataset exploration
 - Improve research results
 - Less need to leverage human resources
 - Improve machine to machine interfaces, ontology performance
- Historically it has been driven by community word of mouth, publication references, and even trial and error

- 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 relevancy search precision
 - E.g., the search will often return good number of datasets (reasonable recall) but a low number of 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

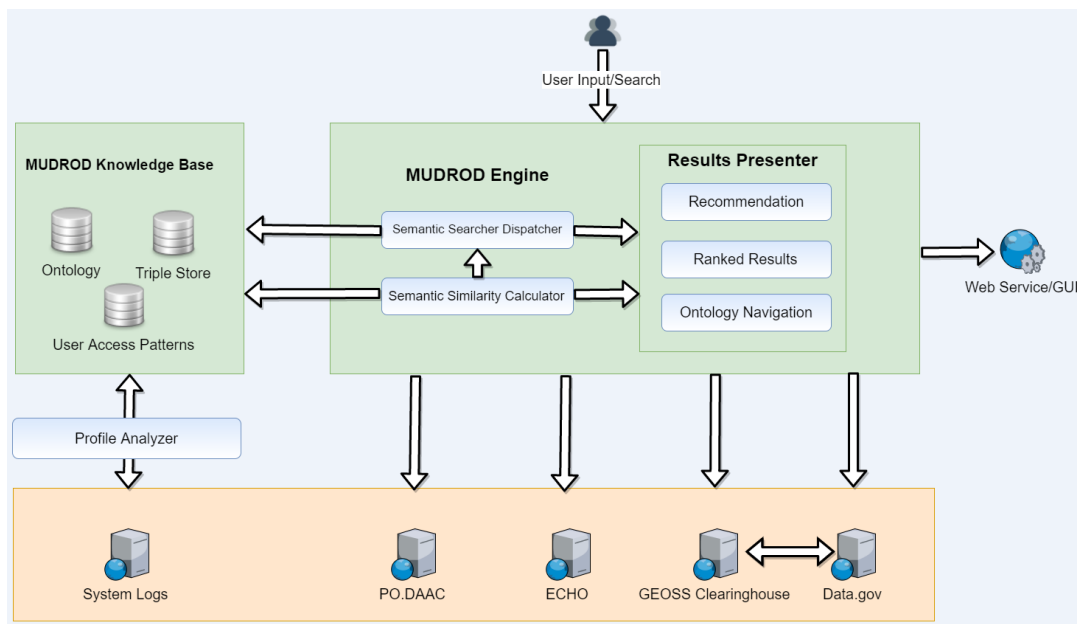
- **Mining and Utilizing Dataset Relevancy from Oceanographic Data (MUDROD)**
 - 2014 funded NASA Advanced Information Systems Technology (AIST) project
 - Technology Readiness Level development from an approximately Level 4 to Level 6-7
 - Specifically targeted to improve search relevance for earth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

- 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 vs real time applications users
 - High spatial resolution vs low spatial resolution

MUDROD search relevance methodology and technical objectives

Methodology

- Analyze **web/ftp logs and metadata** to discover user knowledge (query and data relationships)
- Construct **knowledge base** by combining semantics and profile analyzer
- Implement Machine Learning on a large training set of factors
- Improve data discovery by 1) better **ranking**; 2) **recommendation**; 3) **ontology navigation**



Technology (four technological modules)

- PO.DAAC FTP and web log processing and session construction
- Semantic analysis of user queries & navigation, and metadata records
- Machine learning applied to search ranking training set
- Dataset recommendation engine

Objectives and algorithm factors

- Put the most desired dataset 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 (factors) by considering user behavior, query-text match and examining common geospatial metadata attributes.
 - Geospatial metadata attributes (next slide)
 - Query – metadata content overlap (spatial similarity)
 - User behavior modeling from FTP/web logs (popularity and semantic similarity)

Ranking features – Metadata attributes

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

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

Tying it all together – Machine Learning, the Rank Support Vector Machine (RankSVM)

- One of the well-recognized Machine Learning ranking algorithms
- Convert a **ranking** problem into a **classification** problem that a regular SVM algorithm can solve
 - A classifier trained to predict the ranking order of data pairs
- A ranking problem becomes a binary classification problem, where SVM is applied to find the **optimal decision boundary**
- Has the best NCDG

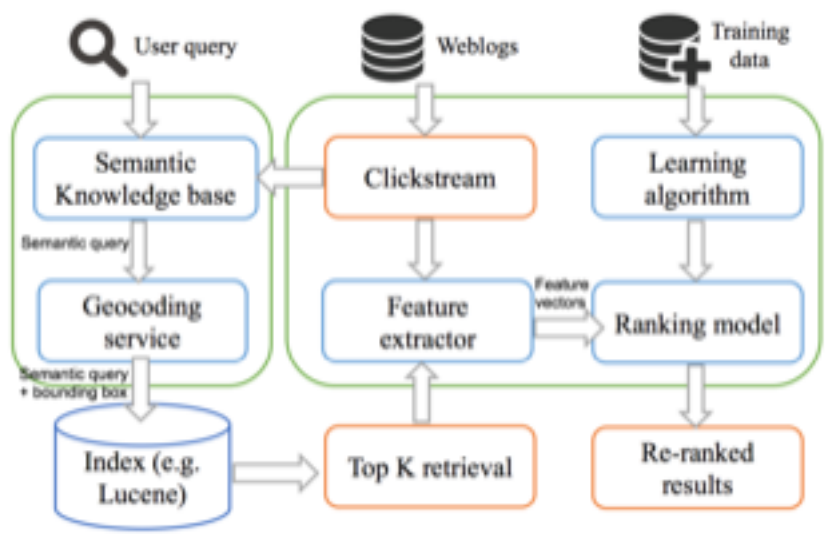


Figure 1. System workflow and architecture



MUDROD vs PO.DAAC search animation



Comparison of "VIIRS (or) SST" search results

PO.DAAC (Solr)

Data Discovery
Found 186 matching dataset(s). [Need help selecting a dataset? Visit the PO.DAAC Forum](#)

Advanced search

Free Text Search

Enter search text
VIIRS SST

Perform Search Reset

view mode:

Sort By: Popularity (All Time)

Prev 1 2 3 4 5 6 7 8 9 10 11 ... 18 19 Next

1 **GHRSSST Level 2P Global Bulk Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA-17 satellite produced by NAVO (pduO-L2P-AVHRR17_G)**
Ocean Temperature
Platform/Sensor: NOAA-17/AVHRR-3
Processing Level: 2P
Along/Across Track Resolution: 8.8 km x 8.8 km
Start/End Date: 2006-Jun-21 to 2009-Jul-6
Description: A global Group for High Resolution Sea Surface Temperature (GHRSSST) Level 2P dataset based on multi-channel sea surface temperature (SST) retrievals generated in real-time from the ... [more](#)

2 **GHRSSST Level 2P Regional Bulk Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA-18 satellite produced by NAVO (NAV0-L2P-AVHRR18_G)**
Ocean Temperature
Platform/Sensor: NOAA-18/AVHRR-3
Processing Level: 2P
Along/Across Track Resolution: 2.2 km x 2.2 km
Start/End Date: 2006-Jan-25 to 2009-Sep-9
Description: A regional Group for High Resolution Sea Surface Temperature (GHRSSST) Level 2P dataset based on multi-channel sea surface temperature (SST) retrievals generated in real-time from the ... [more](#)

3 **GHRSSST Level 2P Global Bulk Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA-18 satellite produced by NAVO (pduO-L2P-AVHRR18_G)**
Ocean Temperature
Platform/Sensor: NOAA-18/AVHRR-3
Processing Level: 2P
Along/Across Track Resolution: 8.8 km x 8.8 km
Start/End Date: 2006-Jan-25 to 2015-Feb-3
Description: A global Group for High Resolution Sea Surface Temperature (GHRSSST) Level 2P dataset based on multi-channel sea surface temperature (SST) retrievals generated in real-time from the ... [more](#)

MUDROD

MUDROD Home

Showing 10 of 97 total match(es) [Default Metadata Listing Settings](#)

First Previous 1 2 3 4 5 6 7 8 9 10 Next Last

Name: VIIRS_MPP-NAVO-L2P-v1.0
Long Name: GHRSSST Level 2P 1-m Depth Global Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite (GDS version 2)
Topic: Sea Surface Temperature
Platform/Sensors: VIIRS
Processing Level: 2P
Start/End Date: 02/25/2014 - Present
Description:
A global Group for High Resolution Sea Surface Temperature (GHRSSST) Level 2P dataset based on retrievals from the Visible Infrared Imaging Radiometer ... [View](#)

Name: VIIRS_MPP-ACDPO-L2P-v1.0
Long Name: GHRSSST v2 Level 2P Global Bulk Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite created by the NOAA Advanced Clear Sky Processor for Ocean (ACDPO)
Topic: Sea Surface Temperature
Platform/Sensors: VIIRS
Processing Level: 2P
Start/End Date: 05/19/2014 - Present
Description:
The ACDPO VIIRS (L2) Level 2 (Unfiltered) product is a global version of the ACDPO VIIRS L2P product available here <http://podbac.jpl.nasa.gov/abstract...> [View](#)

Name: VIIRS_MPP-ACDPO-L2P-v1.1
Long Name: GHRSSST v2 Level 2P Global Bulk Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite created by the NOAA Advanced Clear Sky Processor for Ocean (ACDPO)
Topic: Sea Surface Temperature
Platform/Sensors: VIIRS
Processing Level: 2P
Start/End Date: 05/19/2014 - Present
Description:
The Joint Polar Satellite System (JPSS), starting with JPSS launched on 20 October 2017, is the new generation of the US Polar Operational Environmental ... [View](#)

Name: VIIRS_MPP-ACDPO-L2P-v1.2
Long Name: GHRSSST v2 Level 2P Global Bulk Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite created by the NOAA Advanced Clear Sky Processor for Ocean (ACDPO)
Topic: Sea Surface Temperature
Platform/Sensors: VIIRS
Processing Level: 2P
Start/End Date: 05/19/2014 - 05/19/2015
Description:
Joint Polar Satellite System (JPSS), starting with JPSS launched on 20 October 2017, is the new generation of the US Polar Operational Environmental ... [View](#)

Name: VIIRS_MPP-NAVO-L2P-v1.1
Long Name: GHRSSST Level 2P 1-m Depth Global Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite (GDS version 2)
Topic: Sea Surface Temperature

Not Relevant !! (AVHRR SST datasets)

- MUDROD results:
 - All relevant VIIRS
 - Ordered by version
 - Improved precision



VIIRS L2P



VIIRS_NPP-NAVO-L2P-v2.0

Long Name: GHRSST Level 2P 1 m Depth Global Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite (EOS version 2)

Landing Page: https://podarc.jpl.nasa.gov/dataset/VIIRS_NPP-NAVO-L2P-v2.0

DOI: 10.5061/D04VRS-2PM21

Measurement: Oceans > Ocean Temperature > Sea Surface Temperature > Skin Sea Surface Temperature

Version: 2.0

Description: A global Group for High Resolution Sea Surface Temperature (GHRSST) Level 2P dataset based on retrievals from the Visible Infrared Imaging Radiometer Suite (VIIRS). This sensor resides on the Suomi National Polar-orbiting Operational Environmental Satellite System (NPOESS) Preparatory Project (NPP) satellite launched on 28 October 2011. The VIIRS instrument is a 22-band, multi-spectral scanning radiometer with a 3040-km swath width that builds on the heritage of the MODIS, AVHRR and SeaWiFS sensors for sea surface temperature (SST) and ocean color. For the infrared bands for SST the effective pixel size is 740 meters at nadir and the pixel size variation across the swath is constrained to no more than 1000 meters at the edge of the swath. However, the processing of this dataset aggregates two pixels into one so the resolution is 1500 meters at nadir. This dataset adheres to the GHRSST Data Processing Specification (DOS) version 2 format specifications.

Processing Level: 2P

Coverage: Region: Global
Northernmost Latitude: 90 degrees
Southernmost Latitude: 90 degrees
Westernmost Longitude: -180 degrees
Easternmost Longitude: 180 degrees
Time Span: 2016-02-05 to Present

Spatial Resolution: 0.007 degrees (latitude) x 0.007 degrees (longitude)

Temporal Repeat: 12 Hour

Sensor: VIIRS

Project: GHRSST

Format: NETCDF

Data Access: http://data.mds.nasa.gov/cgi-bin/ghr-dsds/ghrsst/L2P/VIIRS_NPP/NAVO/2
ftp://ftp.mds.nasa.gov/pub/data.mds/ghrsst/DOS2/L2P/VIIRS_NPP/NAVO/2
https://podarc-openftp.jpl.nasa.gov/openftp/vtData/ghrsst/data/DOS2/L2P/VIIRS_NPP/NAVO/2/
ftp://podarc-ftp.jpl.nasa.gov/vtData/ghrsst/data/DOS2/L2P/VIIRS_NPP/NAVO/2

Related Datasets

[VIIRS_NPP-NAVO-L2P-v1.0](#)

[VIIRS_NPP-OSPO-L2P-v2.4](#)

[VIIRS_NPP-OSPO-L3U-v2.4](#)

[VIIRS_NPP-OSPO-L2P-v2.3](#)

[VIIRS_SST_NPP_NAR-OSISAF-L3C-v1.0](#)

[CMOS_16sq-CMC-L4-GLOB-v3.0](#)

[Gw_Polar_Blended-OSPO-L4-GLOB-v1.0](#)

[DM_O-DM-L4-GLOB-v1.0](#)

[JPL-L2P-MODIS_A](#)

[MODIS_A-JPL-L2P-v2014.0](#)

[How does MDRDD find related Datasets?](#)

Related Keyword Searches

[VIIRS_NPP_JPL_L2P_V2016.0 \(0.91\)](#)

[VIIRS \(5.8\)](#)

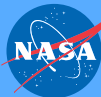
[IASI \(0.78\)](#)

[SUOMI_NPP \(0.74\)](#)

- **Dataset heterogeneity and number, and understanding user intent still represent challenges for effective earth data search**
- MUDROD demonstrated tangible improvements in the search precision over the current default PO.DAAC Solr search result
 - Results vetted by oceanographic domain experts
- MUDROD key features:
 - Implemented 11 factors derived from log mining, query analysis and metadata attributes in a Machine Learning algorithm
 - A dataset recommendation algorithm implemented to improve 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
- Deployed at: https://podaac.jpl.nasa.gov/podaac_labs and <https://mudrod.jpl.nasa.gov>
 - Publications and technical related documentation can be accessed: at <https://mudrod.oit.hub.io/>
- Will be extended in the OceanWorks Project framework
 - Support near real-time data ingestion to dynamically update knowledge base
 - Develop improved query understanding module to better interpret user's search intent (e.g. "ocean wind level 3" -> "ocean wind" AND "level 3")
 - Event based tagging, e.g. return appropriate datasets and granules for hurricanes
 - Infuse into PO.DAAC keyword text search



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





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
- **Keyword search**
- Dataset Information Page/DOI Landing Pages
- Granule browsing through date tree

Dataset Information Page

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

Platform

- ADEOS (8)
- ADEOS-II (22)
- AQUA (60)
- AQUARIUS_SAC-D (59)
- ARGO (1)
- Coriis (10)
- Cryosat-2 (1)
- DMSP-F08 (4)

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

- Dataset Ranking is a long-standing problem in geospatial data discovery...data diversity and heterogeneity, user search intent
- UWG recommendations over past several years
- *.....Improve search and discovery of PO.DAAC dataset via free text (.e.g., keyword) and facets*
- *.....Develop advanced search capabilities*
- 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-dimensional dynamic ranking of data

NASA Missions & Projects

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

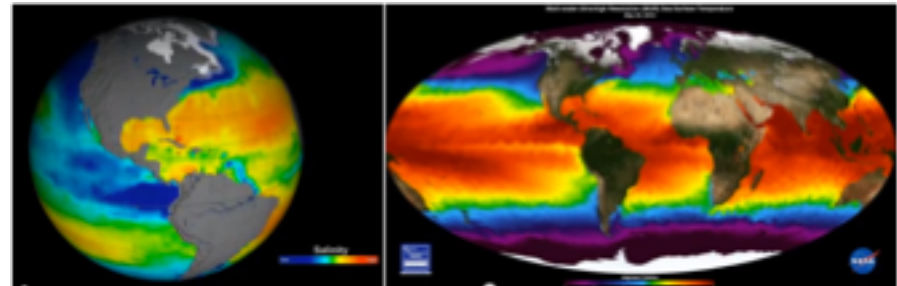
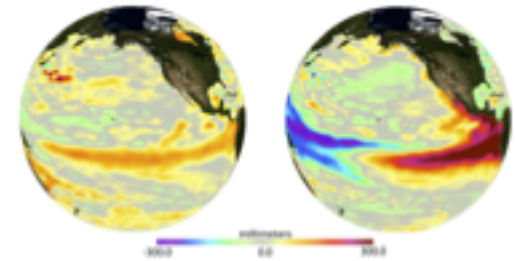


Upcoming: COWR, AirSWOT, SWOT, GRACE-FO

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*



Additional potential future improvements

- 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
- Leverage advanced computing techniques to speed up the process

- Dataset relevancy and selection factors
 - Data quality, accuracy
 - Documentation, interoperability
 - Time series length, latency
 - Images
 - Resolution
 - Application
- PO.DAAC facets address some of these but there is lots to improve

- Objectives:
 - Dataset relevancy mining, Dataset similarity calculation, Recommendations based on metadata attributes and user workflow patterns
- JPL contributions:
 - Developing use cases for search, discovery and utilization of earth science data focusing on datasets residing the JPL PO.DAAC
 - Ontology development for improving semantic searching of MUDROD
 - Implementation in PO.DAAC labs
 - Leverage ESDSWG Data Quality recommendations on distinguishability

• MUDROD

- *(Mining and Utilizing Dataset Relevancy from Oceanographic Dataset)*
- MUDROD funded by NASA AIST project intends to improve search relevancy and ranking for NASA earth science products from a user perspective with the Metadata, Usage Metrics, and User Feedback information.
- **Key features:**
 - ❑ *Goals: (upper figure)*
 1. Analyze *web logs* to discover user knowledge
 2. Construct knowledge base by combining semantics and profile analyzer
 3. Improve data discovery by 1) better ranking; 2) recommendation; 3) ontology navigation
 - ❑ *Architecture (bottom figure)*
- The flowchart of constructing MUDROD smart search engine from user query to the ranked results, indicating a complicate process of user driven data search.

Comparison of "ocean OR wind" search results



PO.DAAC (Solr)

MUDROD

Dataset Discovery
 Found 382 matching dataset(s)

Need help selecting a dataset? Visit the PO.DAAC Forum

Advanced search

Free Text Search: Enter search text: ocean wind

Temporal Search: Start Date, Stop Date

Perform Search, Reset

View mode: List, Grid

Sort By: Popularity (All Time)

1. GHRSST Level 4 Q1SST Global Foundation Sea Surface Temperature Analysis (JPL_OUROCEAN-L4Q1SST-GLOB-Q1SST)
 Ocean Temperature
 Platform/Sensor: AQUAMSR-E, AQUAMODIS, IHSuVnSiv ... more
 Processing Level: 4
 Longitude/Latitude Resolution: 0.01 degree x 0.01 degree
 Start/End Date: 2010-Jun-9 to Present
 Description: A Group for High Resolution Sea Surface Temperature (GHRSST) Level 4 sea surface temperature analysis produced daily on an operational basis at the JPL OurOcean group using a multi-scale ... more

2. TOPEX/Poseidon L2 Ocean Surface Topography Merged Geophysical Record Crossover v1.0 (TOPEX_L2_OST_MGRD_CROSSOVER)
 Ocean Waves, Sea Surface Topography
 Platform/Sensor: TOPEX/POSEIDON/TOPEX.ALTIMETER, TOPEX/POSEIDON/POSEIDON.ALTIMETER, TOPEX/POSEIDON/TOPEX.MICROWAVE.RADIOMETER
 Processing Level: 2
 Along/Across Track Resolution: 11.2 km x 5.1 km
 Start/End Date: 1995-Apr-24 to 1999-Jun-26
 Description: This dataset contains the crossover points from TOPEX/Ocean Topography Experiment/Poseidon Merged Geophysical Data Record version 1.0 (MGRD-B). The MGRD-B contains measurements from ... more

3. Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L2.0 First Look SSM/I-F14 Microwave Analysis (COMP_MEASURES_ATLAS_L3_OR_L2_S_SSM_F14_WIND_VECTORS_F14)
 Ocean Winds
 Platform/Sensor: DMSP-F14SSM/I
 Processing Level: 3
 Longitude/Latitude Resolution: 0.25 degree x 0.25 degree
 Start/End Date: 1997-May-7 to 2008-Aug-8
 Description: This dataset is derived under the Cross-Calibrated Multi-Platform (CCMP) project and contains value-added Special Sensor Microwave Imager (SSM/I) ocean surface winds from the Defense ... more

4. GHRSST Level 2P Global Skin Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on the MetOp-A satellite produced

MUDROD Home

Showing 10 of 471 total match(es)

1. Name: MOCAT_LEVEL_2B_OCN_CLM_10_Y1
 Long Name: Tropical Level 2B Climate-Ocean Wind Vectors in 10 Bin Footprints
 Topic: Surface Winds
 Platform/Sensor: Replicator
 Processing Level: 2
 Start/End Date: 1955-0014 - 2014-0014
 Description: This dataset contains the Tropical Level 2B 10 Bin Version 1.0 Climate-quality ocean surface wind vectors. The Level 2B wind vectors are derived on a ... More

2. Name: ALIS_L2_OST_JRS016_V1
 Long Name: ALIS Jason-2 Coastal Altimetry Version 1
 Topic: Sea Surface Height, Significant Wave Height
 Platform/Sensor: TRSA, POSEIDON-S, AMR
 Processing Level: 2
 Start/End Date: 2014-0000 - Present
 Description: Adaptive Loading Edge Subswath (ALIS) provides high resolution ocean altimetry measurements by applying a specialized refractor to Jason-2 data. ... More

3. Name: QRSAT_L1C_NONSPINNING_SIGMA0
 Long Name: QuikSCAT Level 1C Nonspinning Sigma0
 Topic: Sigma-Rough, Surface Winds
 Platform/Sensor: SEAWINDS
 Processing Level: 1C
 Start/End Date: 2016-0010 - Present
 Description: This dataset contains geo-located and averaged Level 1C Sigma-0 measurements and wind-retrievals from the SeaWinds on QRSAT platform, related to ... More

4. Name: MOCAT-L3-COASTAL
 Long Name: MetOp-B MOCAT Level 3 Coastal Ocean Surface Wind Vectors Optimized for Coastal Ocean
 Topic: Surface Winds
 Platform/Sensor: ASCAT
 Processing Level: 3

Related Searches:
 SURFACE WIND (1)
 WIND SPEED (48)
 WIND DIR (48)
 WIND (67)
 VECTOR (67)
 OCEAN WIND VECTOR (67)
 OCEAN CURRENT (67)
 QUIKSCAT (48)
 SCATTERMETER (48)
 WIND VELOCITY (48)



Not Relevant !! (SST or SSH altimeter datasets)

- MUDROD results:
 - Recall similar
 - Precision improved !