Cloud-based Data Match-Up Service (CDMS)

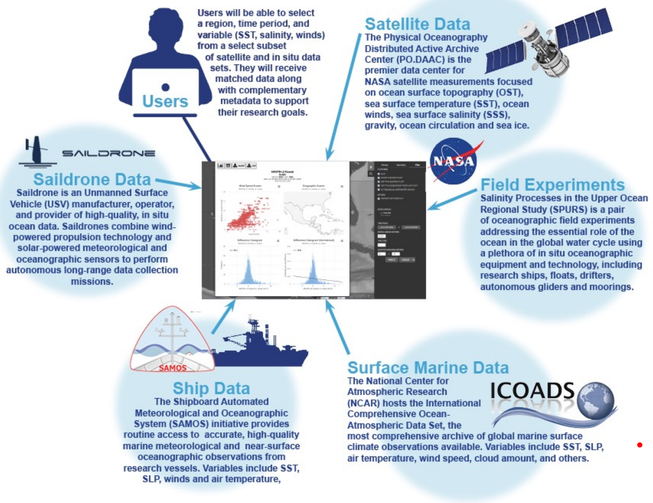
Nga Chung[1], Thomas Huang[1], Vardis M. Tsontos[1], Stepheny Perez[1], Wai Phyo[1], Shawn R. Smith[2], Homer McMillan[2], Thomas Cram[3], Zaihua Ji[3], Kimberly Sparling[4]

[1]Jet Propulsion Laboratory, California Institute of Technology, [2]Center for Ocean-Atmospheric Prediction Studies, [3]National Center for Atmospheric Research, [4]Saildrone

Overview

The **Cloud-based Data Match-Up Service (CDMS)** is a collaborative effort between NASA JPL, COAPS, NCAR, and Saildrone. CDMS is an extension of the Distributed Oceanographic Match-Up Service (DOMS) which was funded by the NASA AIST program. CDMS will provide a mechanism for users to input a series of geospatial references for satellite observations and receive the in situ or satellite observations that are matched to the primary satellite data within selectable temporal and spatial search domains.

The software stack that enables CDMS match-up capability is available via the Apache Science Data Analytics Platform (SDAP), which is an Apache incubator project. Under the ACCESS program, the team plans to deliver a production-ready match-up capability that fully leverages cloud-native services.

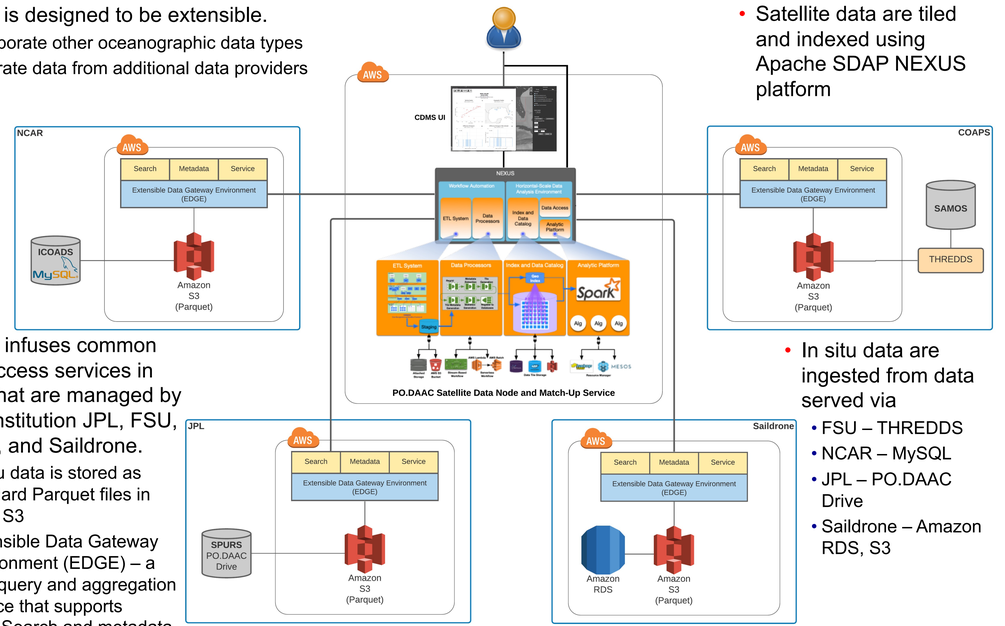


Why CDMS is Needed?

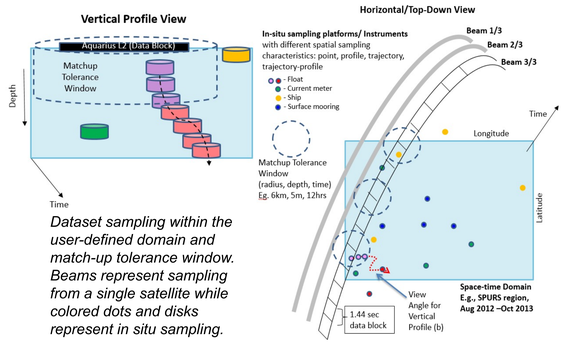
- There is a need for a generalized match-up capability that is publicly accessible and provides flexibility and reproducibility for calibration/validation (cal/val), mission data processing, and science use cases including, but not limited to:
  - Iterative cal/val of satellite retrieval algorithms
  - Decision support for designing and implementing field campaigns
  - Scientific investigations (e.g., developing blended satellite-in situ products, process studies)
  - Quality control of surface marine observations
  - Acquire colocated swath data for a key variable used in the satellite geophysical model retrieval algorithm from ancillary datasets
- CDMS eliminates the need for one-off match-up programs that require satellite and in situ data to be housed on one's local computer.

CDMS Architecture

- CDMS is designed to be extensible.
  - Incorporate other oceanographic data types
  - Integrate data from additional data providers
- CDMS infuses common data access services in AWS that are managed by each institution JPL, FSU, NCAR, and Saildrone.
  - In situ data is stored as standard Parquet files in AWS S3
  - Extensible Data Gateway Environment (EDGE) – a data query and aggregation service that supports OpenSearch and metadata export
- Satellite data are tiled and indexed using Apache SDAP NEXUS platform
- In situ data are ingested from data served via
  - FSU – THREDDS
  - NCAR – MySQL
  - JPL – PO.DAAC Drive
  - Saildrone – Amazon RDS, S3



Search Domain & Match-Up Tolerances

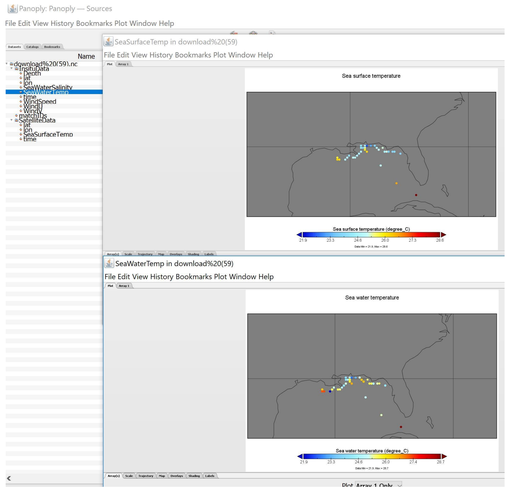


Dataset sampling within the user-defined domain and match-up tolerance window. Beams represent sampling from a single satellite while colored dots and disks represent in situ sampling.

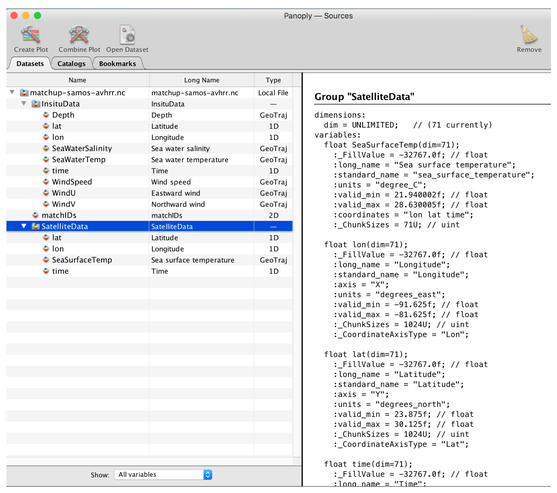
- Queries will be facilitated by indexing the following in CDMS
  - Parameter – salinity, sea temperature, winds, etc.
  - Temporal search domain – ISO 8601 UTC
  - Horizontal search domain – latitude/longitude box
  - Vertical search domain above/below sea level
  - Data source
    - Satellite: SMAP, ASCAT MetOp-B, VIIRS NPP, etc.
    - In situ: ICOADS Release 3.0, SAMOS, SPURS-1, 2, Saildrone
  - Platform type (ship, orbiting satellite, etc.)
  - Device type (CTD, current profiler, radiometer, etc.)
  - Provider (Saildrone, NCAR, FSU)
  - Collection (Mission name, project e.g., ICOADS, SAMOS)
  - Data quality flag - Mapped to IODE standard
- Users also specify spatial and temporal match-up tolerances for locating a match (e.g., within 1 hours and 30 km)

Standardized Output for Match-Up Data

- CDMS supports both netCDF4 and CSV output formats and implements a consistent set of relevant CF attributes but also CDMS specific metadata used to document query parameters fully.



- Since CDMS Group variables leverage existing CF standards, an off the shelf tool like Panoply that is CF-metadata aware can automatically interpret and map in situ and satellite group data independently just as if they were given in their separate CF-compliant source files.



- Metadata is also included at the Group variable level. Both coordinate and measurement variable attributes such as valid_max/min, __fillvalue, and units are included consistent with CF. A reader is provided to reconstruct the matches in each group using a join operation based on match IDs.

Goals of CDMS Under ACCESS Program

Long-Term

- Deliver a production-ready near real-time and delayed-mode match-up service in the cloud to address cal/val and science use cases
- Integrate interactive match-up capability with a visualization platform
- Formalize architecture and information model for in situ and satellite data nodes to efficiently onboard additional datasets via PO.DAAC and remote data hosts
- Capture and analyze user match-up metrics to enable future data search and recommendations

Near-Term

- Update match-up algorithm to:
  - Remove restriction that both source datasets must contain the same parameter. Identify matches based only on latitude-longitude position and time.
  - Support in situ to L2 satellite swath data matches.
  - Support satellite-to-satellite data matches (L2 and gridded L3, L4).
- Enhance in situ data node architecture to take advantage of the expected performance improvements in having in situ and satellite data co-located in the cloud
  - Extend translation specification to include additional meteorological and oceanographic parameters and QC flags
  - Convert in situ data from various sources (MySQL, THREDDS, NetCDF-4, Parquet) into a schematized JSON exchange format for ingesting into CDMS. JSON schema maps to existing standards such as CF metadata conventions where appropriate.
  - Implement in situ subsetting API that uses Spark SQL to query Parquet files

Acknowledgements

The CDMS project is funded by NASA via the Advancing Collaborative Connections for Earth System Science (ACCESS) program.