Jet Propulsion Laboratory
California Institute of Technology

Cloud-based Data Match-Up Service (CDMS)

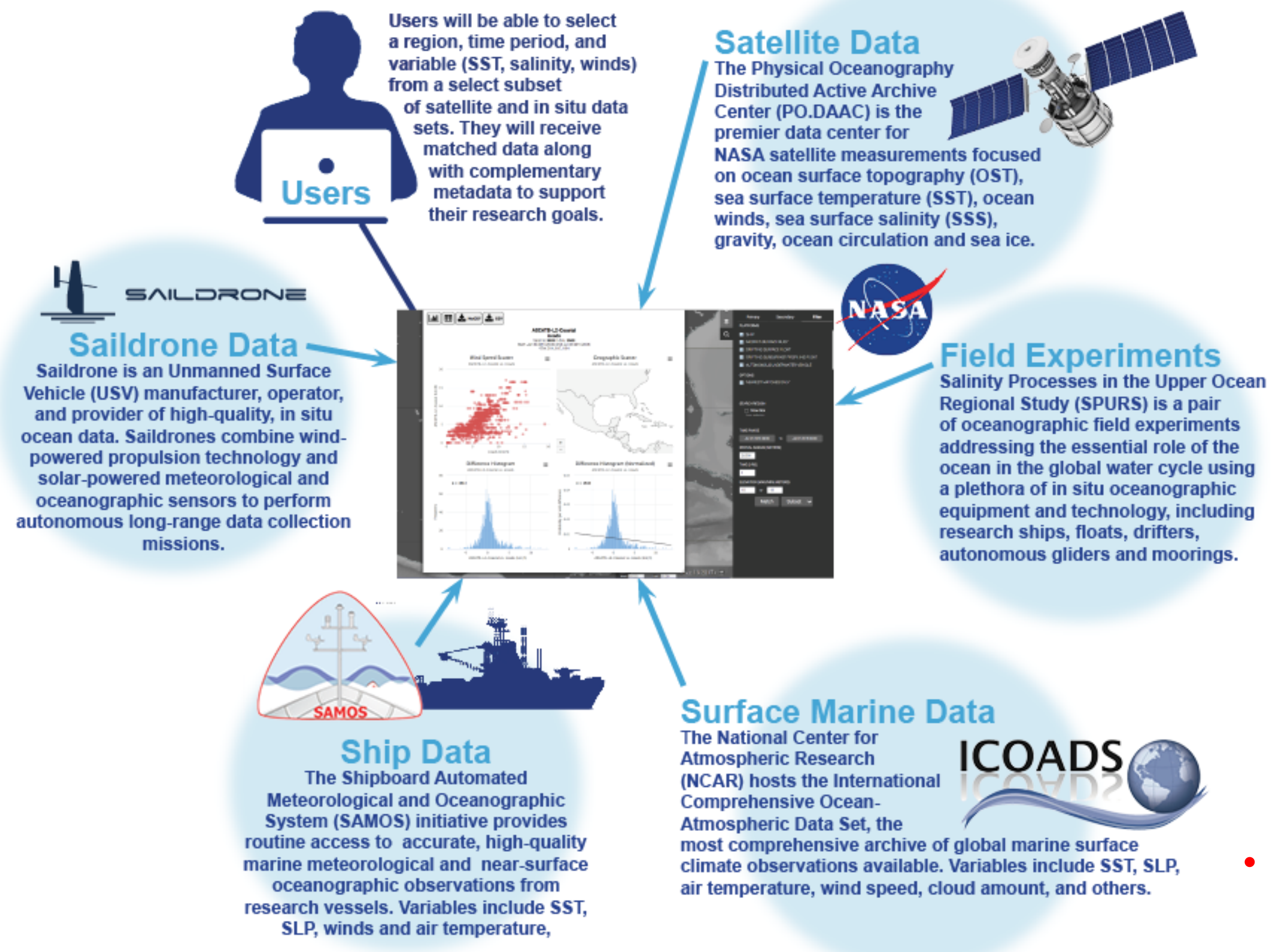
Nga Chung[1], Thomas Huang[1], Vardis M. Tsontos[1], Stepheny Perez[1], Wai Phyo[1], Joshua Rodriguez[1], Riley Kuttruff[1], Shawn R. Smith[2], Jordan Gethers[2], Thomas Cram[3], Zaihua Ji[3], Kimberly Sparling[4]

[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.



Users
Users will be able to select a region, time period, and variable (SST, salinity, winds) from a select subset of satellite and in situ data sets. They will receive matched data along with complementary metadata to support their research goals.

Satellite Data
The Physical Oceanography Distributed Active Archive Center (PO.DAAC) is the premier data center for NASA satellite measurements focused on ocean surface topography (OST), sea surface temperature (SST), ocean winds, sea surface salinity (SSS), gravity, ocean circulation and sea ice.

Saildrone Data
Saildrone is an Unmanned Surface Vehicle (USV) manufacturer, operator, and provider of high-quality, in situ ocean data. Saildrones combine wind-powered propulsion technology and solar-powered meteorological and oceanographic sensors to perform autonomous long-range data collection missions.

Field Experiments
Salinity Processes in the Upper Ocean Regional Study (SPURS) is a pair of oceanographic field experiments addressing the essential role of the ocean in the global water cycle using a plethora of in situ oceanographic equipment and technology, including research ships, floats, drifters, autonomous gliders and moorings.

Ship Data
The Shipboard Automated Meteorological and Oceanographic System (SAMOS) initiative provides routine access to accurate, high-quality marine meteorological and near-surface oceanographic observations from research vessels. Variables include SST, SLP, winds and air temperature,

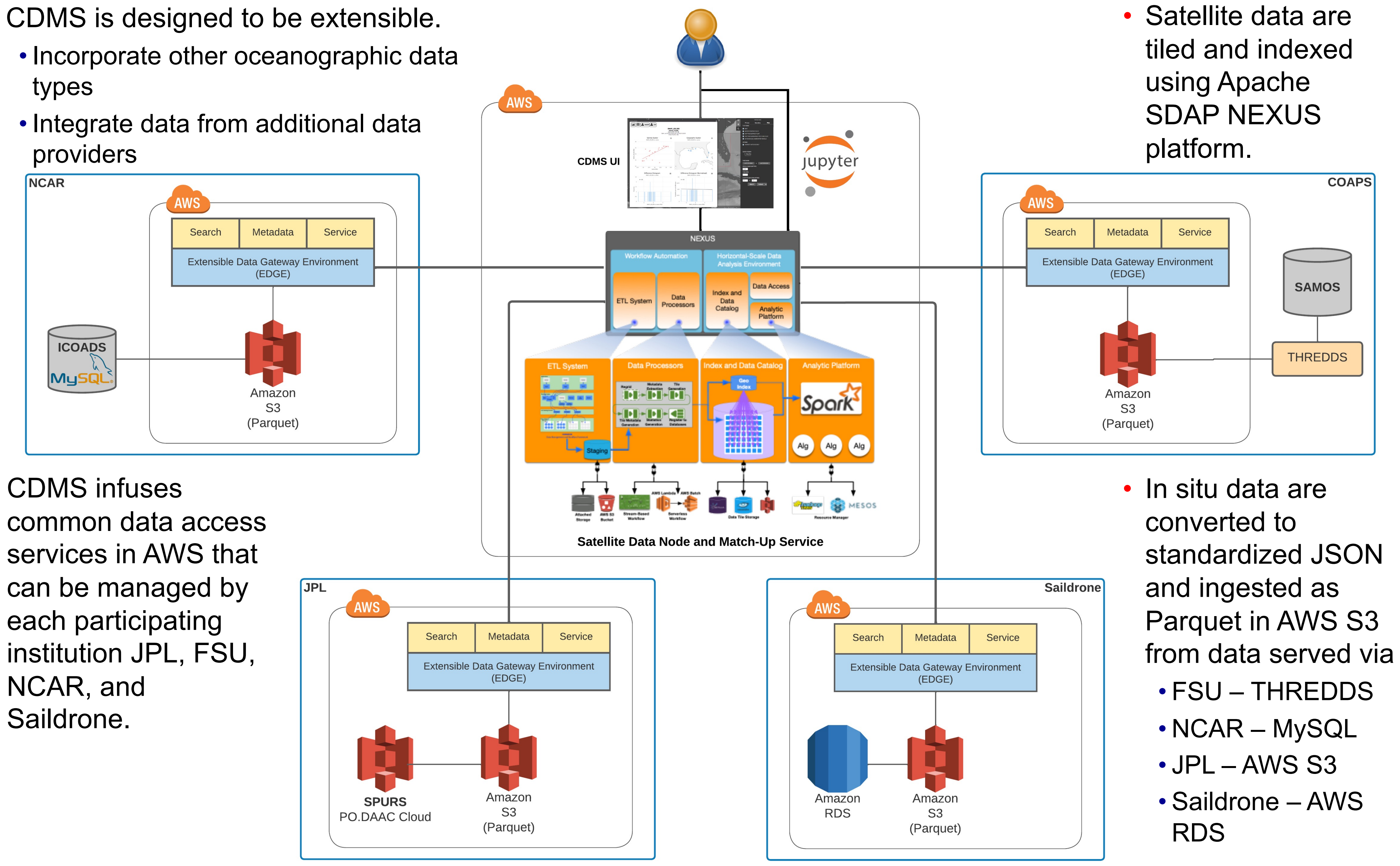
Surface Marine Data
The National Center for Atmospheric Research (NCAR) hosts the International Comprehensive Ocean-Atmospheric Data Set, the most comprehensive archive of global marine surface climate observations available. Variables include SST, SLP, air temperature, wind speed, cloud amount, and others.

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.

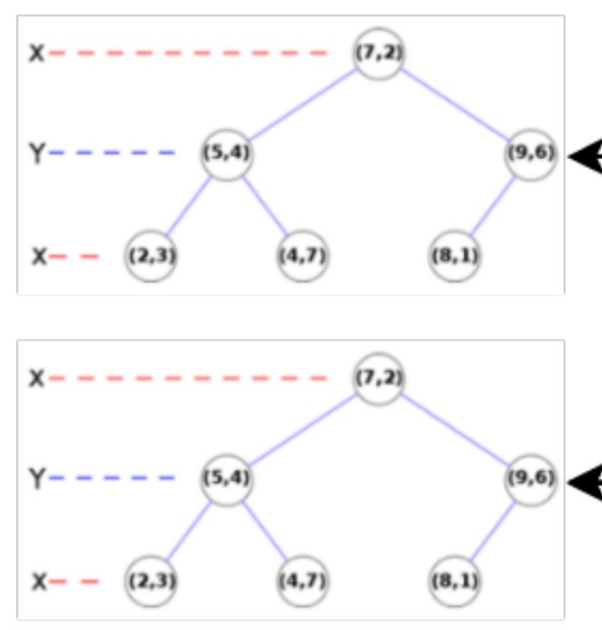
CDMS Architecture

- 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 can be managed by each participating institution JPL, FSU, NCAR, and Saildrone.
- Satellite data are tiled and indexed using Apache SDAP NEXUS platform.
- In situ data are converted to standardized JSON and ingested as Parquet in AWS S3 from data served via
  - FSU – THREDDS
  - NCAR – MySQL
  - JPL – AWS S3
  - Saildrone – AWS RDS

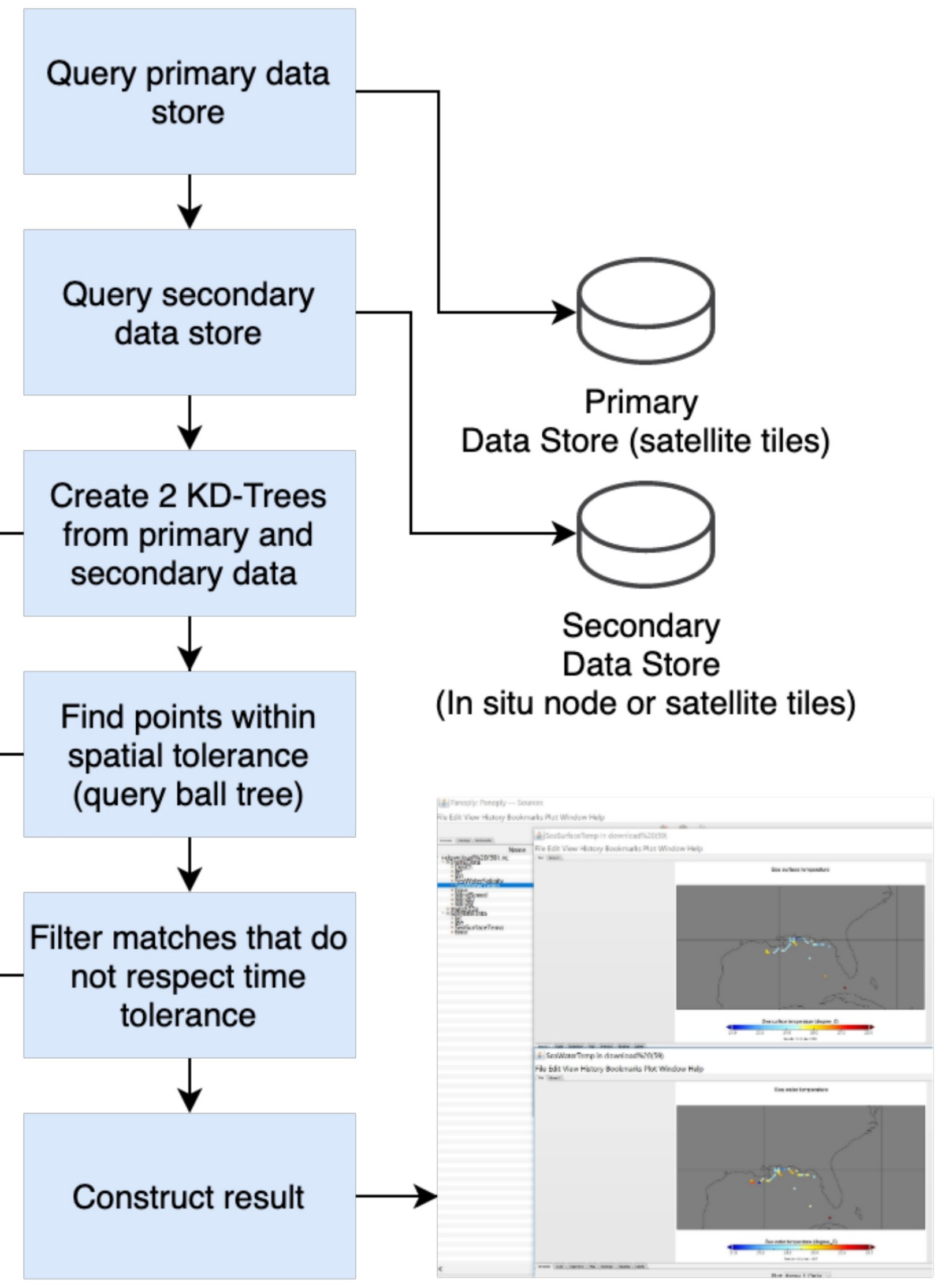


Match-Up Algorithm

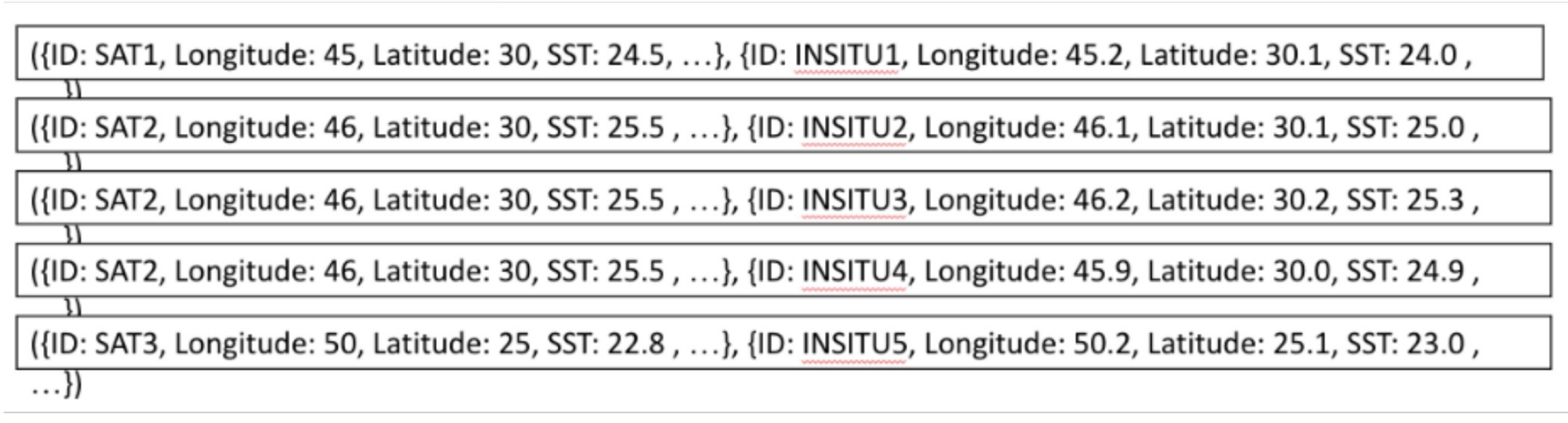
- User provided parameters
  - Primary data source - Satellite data source name
  - Secondary data source - Either satellite or in-situ data source name
  - Temporal search domain
  - Spatial search domain - Latitude/longitude bounding box
  - Platform type - ship, orbiting satellite, etc.
  - Device type - CTD, current profiler, radiometer, etc.
  - Depth min and max
  - Radius tolerance
  - Time tolerance
  - Science parameter (optional) - sst, sss, wind, etc.
  - matchOnce - if true, each primary point will only match with a single secondary point
  - Maximum result size limit



KD-Tree Data Structures



Query primary data store
Query secondary data store
Create 2 KD-Trees from primary and secondary data
Find points within spatial tolerance (query ball tree)
Filter matches that do not respect time tolerance
Construct result
Primary Data Store (satellite tiles)
Secondary Data Store (In situ node or satellite tiles)



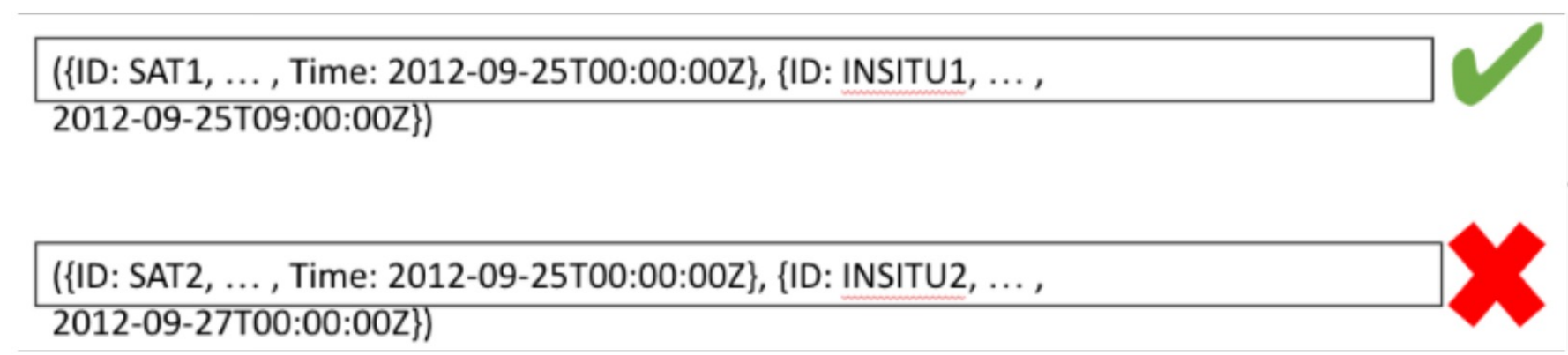
({ID: SAT1, Longitude: 45, Latitude: 30, SST: 24.5, …}, {ID: INSITU1, Longitude: 45.2, Latitude: 30.1, SST: 24.0 , …})

({ID: SAT2, Longitude: 46, Latitude: 30, SST: 25.5 , …}, {ID: INSITU2, Longitude: 46.1, Latitude: 30.1, SST: 25.0 , …})

({ID: SAT2, Longitude: 46, Latitude: 30, SST: 25.5 , …}, {ID: INSITU3, Longitude: 46.2, Latitude: 30.2, SST: 25.3 , …})

({ID: SAT2, Longitude: 46, Latitude: 30, SST: 25.5 , …}, {ID: INSITU4, Longitude: 45.9, Latitude: 30.0, SST: 24.9 , …})

({ID: SAT3, Longitude: 50, Latitude: 25, SST: 22.8 , …}, {ID: INSITU5, Longitude: 50.2, Latitude: 25.1, SST: 23.0 , …})

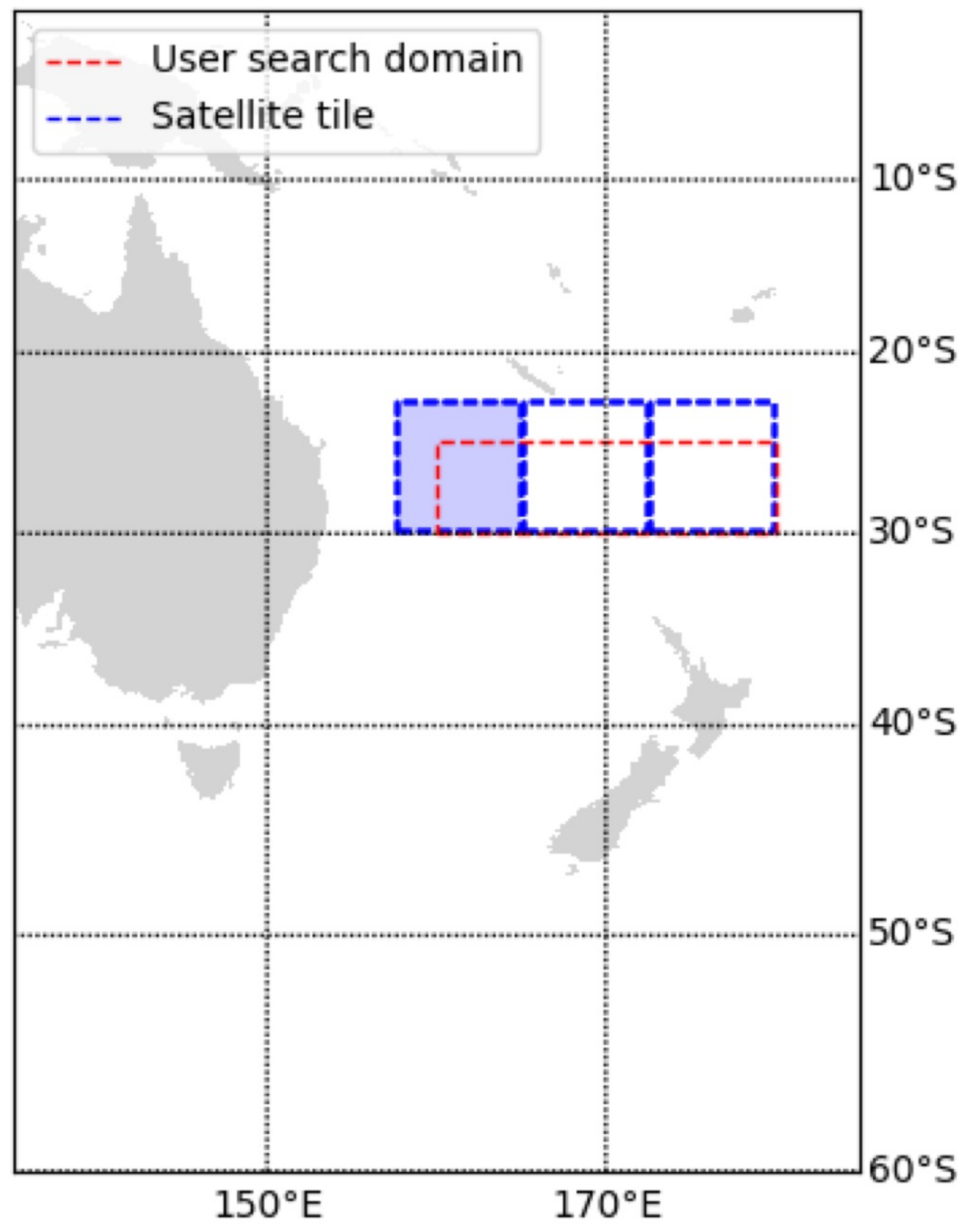


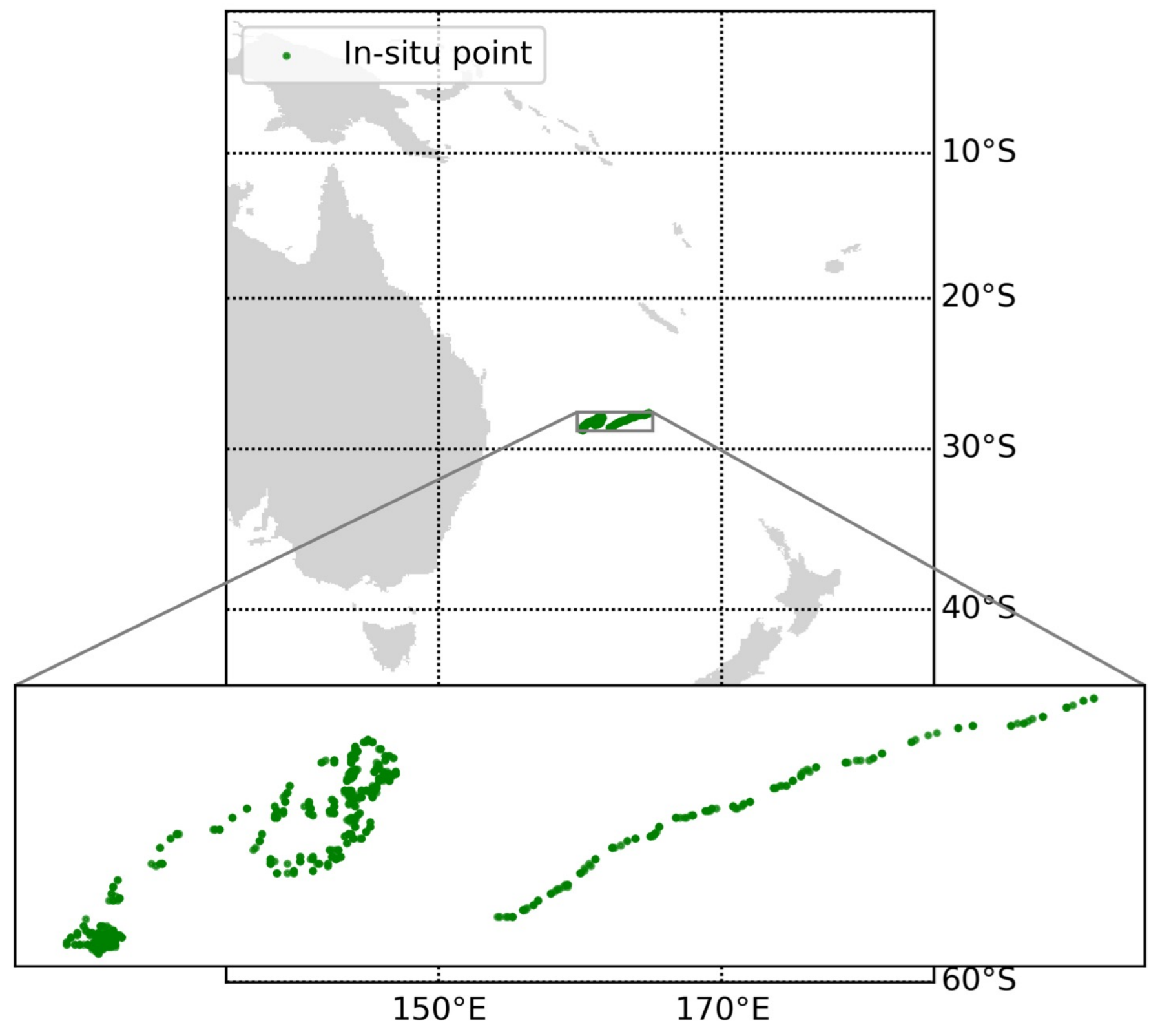
({ID: SAT1, … , Time: 2012-09-25T00:00:00Z}, {ID: INSITU1, … , 2012-09-25T09:00:00Z}) ✓

({ID: SAT2, … , Time: 2012-09-25T00:00:00Z}, {ID: INSITU2, … , 2012-09-27T00:00:00Z}) ✗

Sample matchup request

- Dataset=**MUR25-JPL-L4-GLOB-v04.2**
- Start time=**2018-09-24T00:00:00Z**
- End time=**2018-09-30T00:00:00Z**
- Bounds=**160,-30,180,-25**
- Secondary=**ICOADS Release 3.0**
- Time tolerance=**86400 seconds**
- Radius tolerance=**1000 meters**
- Platform=**drifting surface float**





1. Search for tiles using user-provided time and space bounds. 18 tiles match given the user-provided constraints and satellite dataset **MUR25-JPL-L4-GLOB-v04.2**
2. Search for in situ points within user-provided search domain. 754 in situ points match given the user-provided constraints and the secondary in situ dataset **ICOADS Release 3.0**, provider **NCAR**, and platform **drifting surface float**. Construct secondary KD tree from points
3. Construct primary KD trees for each 18 satellite primary tile from step 1
4. Find points within user-provided **1000 m** radius tolerance of one another using primary and secondary KD trees
5. Filter results that are not within user-provided +/- time tolerance (**86400 seconds**)

Goals of CDMS Under ACCESS Program

Near-Term

- Deploy publicly accessible satellite and in situ data nodes in AWS
- Publish Jupyter notebooks illustrating matchup APIs on public GitHub
- Continue validation and benchmarking efforts
- Explore cloud-optimized formats, e.g. Zarr, for satellite data
- Add support for large match-up requests
- Build CDMS web interface

Longer-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 NASA DAACs and remote data hosts
- Capture and analyze user match-up metrics to enable future data search and recommendations

Acknowledgements

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