podaac
Physical Oceanography Distributed Active Archive Center

PO.DAAC Cloud Data Ecosystem - Part 2: Moving Science to the Cloud

Edward M. Armstrong[1], Wen-Hao Li[1], Jinbo Wang[1], Jorge Vazquez[1], Jack McNellis[1], Catalina Oaida[1], Cassandra L Nickles[1], Michael Gangl[1],

1. NASA Jet Propulsion Laboratory, California Institute of Technology

23nd GHRSST Science Team Meeting
June 27 - July 1, 2022

These activities were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Dedicated funding for PO.DAAC activities is through a grant from NASA's ESDIS Project.

Outline

- PO.DAAC cloud data access overview
- The PO.DAAC software repository
- Cloud tutorial science stories
  - Requirements for running the notebooks:
    - NASA Earthdata Login (EDL) (free to create account)
    - Python and other modules and their associated package manager based dependencies
      - Note: these examples are written in Python due to open source nature and general community support/development for libraries and tools that work well in the cloud; however, cloud data can be accessed using tools like R or Matlab
    - Amazon Elastic Cloud Compute (EC2) account for in cloud computing if required
- Summary: Roadmap and development forward

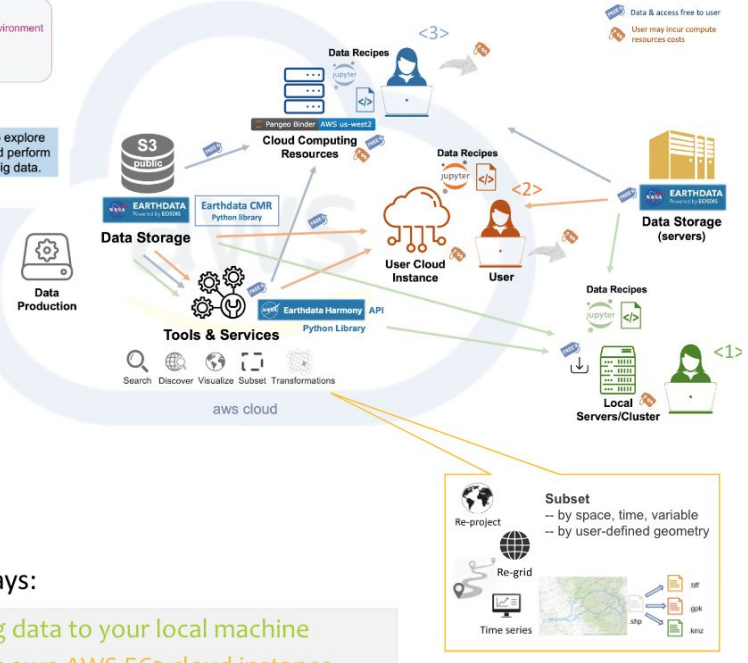
PO.DAAC Cloud Overview

- **New Cloud Paradigm**

Big Data Challenge & Solutions – EOSDIS & DAAC Goals
- Maintain DAAC level of service to user, by leveraging scalability of cloud environment
- Minimize amount of data user needs to handle
- Make data more analysis ready on behalf of user

Co-location
- Data
- Tools & Services
- Analysis

The Cloud Paradigm offers opportunity to explore new ways of using Earth Observations and perform science research and applications with big data.



Key features:

1. Data storage and tools and services are co-located in cloud
2. Migrate all datasets from 12 NASA data centers into the Earthdata cloud
3. Provide the same and higher level of service to users
4. Easy to access to datasets from different sources (12 NASA DAACs)

Three cloud data access pathways:

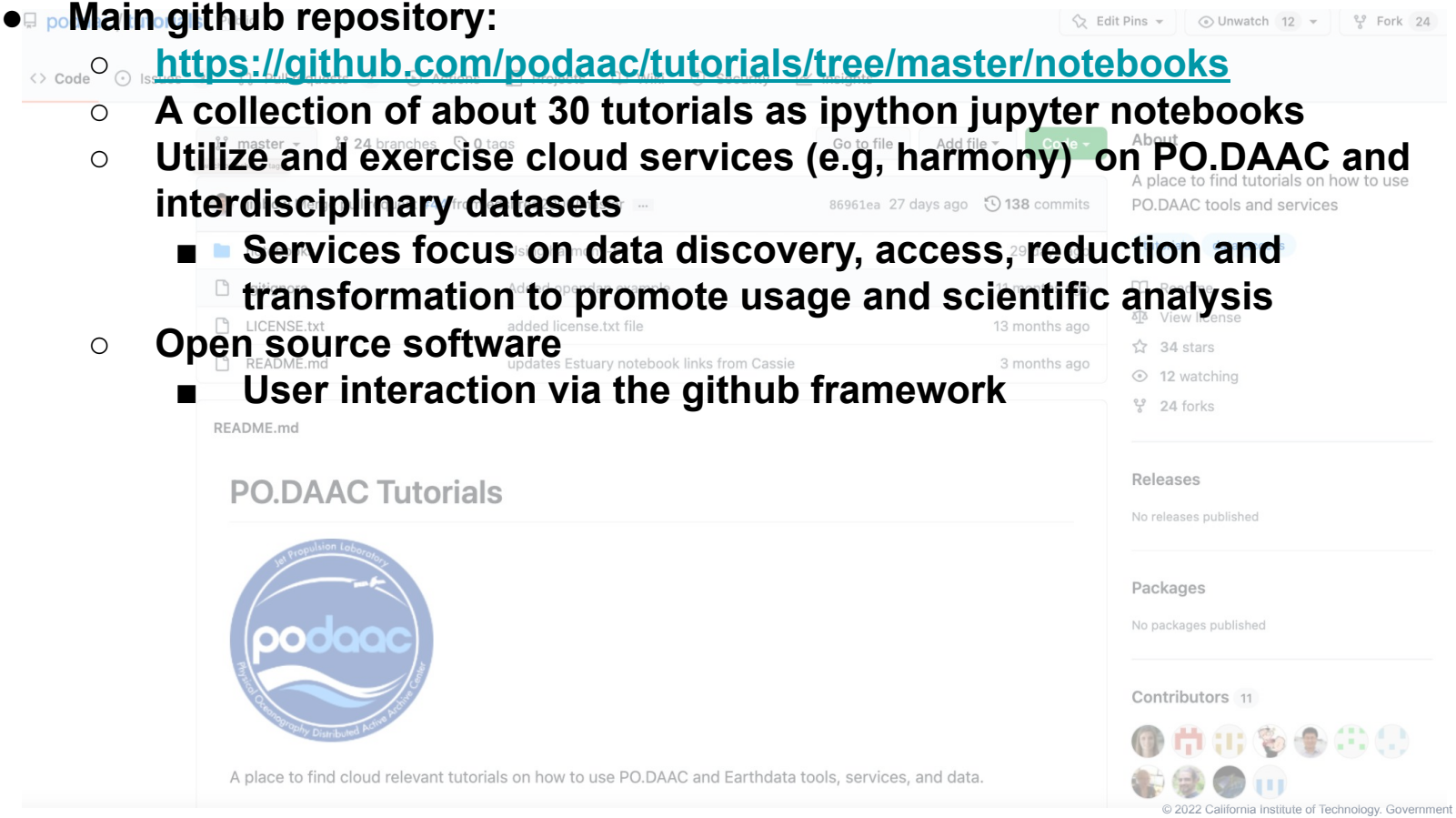
1. Working locally: downloading data to your local machine
2. Within the Cloud: Set up your own AWS EC2 cloud instance
3. Within the Cloud: Through shareable cloud environments

Earthdata Harmony APIs

- See "*PO.DAAC Cloud Data Ecosystem – Part 1: Search, Access and Services",* Wen-Hao Li et al., for additional information and context

Software Repository

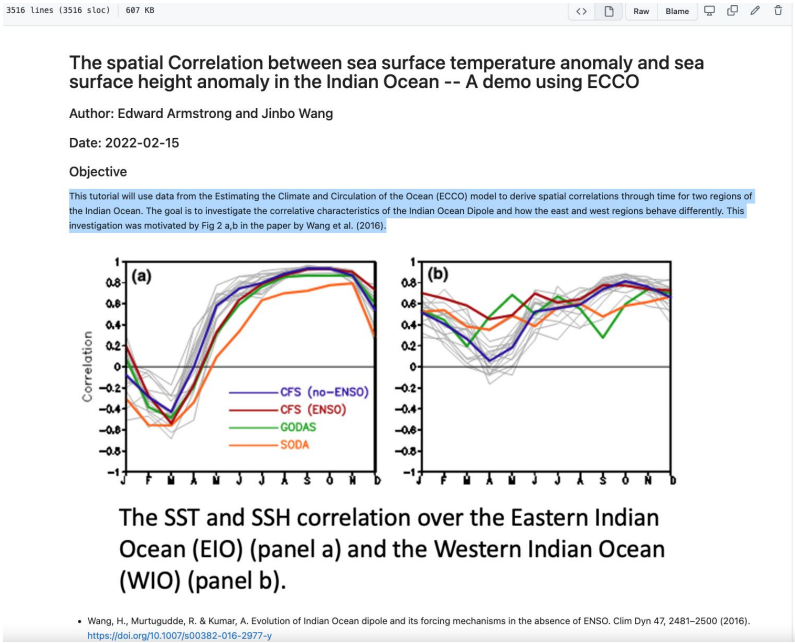
- **Main github repository:**
  - **https://github.com/podaac/tutorials/tree/master/notebooks**
  - **A collection of about 30 tutorials as ipython jupyter notebooks**
  - **Utilize and exercise cloud services (e.g, harmony) on PO.DAAC and interdisciplinary datasets**
    - **Services focus on data discovery, access, reduction and transformation to promote usage and scientific analysis**
  - **Open source software**
    - **User interaction via the github framework**



Usage

- Tutorials used in several ocean and interdisciplinary workshops and hackathons
- Tutorials are both How-To's and more end-to-end example use cases
  - Ocean Sciences 2022 NASA workshop: Transforming to Open Science and Analysis in the Cloud Using NASA Earth Science Data
    https://github.com/podaac/tutorials/tree/master/notebooks/meetings_workshops/workshop_osm_2022
  - 2022 SWOT early adopter hackweek
    https://github.com/podaac/tutorials/tree/master/notebooks/meetings_workshops/swot_ea_hackweek_2022
  - 2021 Fall AGU workshop https://nasa-openscapes.github.io/2021-Cloud-Workshop-AGU/
  - 2021 Cloud Hackathon https://nasa-openscapes.github.io/2021-Cloud-Hackathon/
  - Internal training and capacity building…..and more (GHRSST training?)

Cloud computing - ECCO SST/SSH correlation



3516 lines (3516 sloc) 607 KB Raw Blame

The spatial Correlation between sea surface temperature anomaly and sea surface height anomaly in the Indian Ocean -- A demo using ECCO

Author: Edward Armstrong and Jinbo Wang

Date: 2022-02-15

Objective

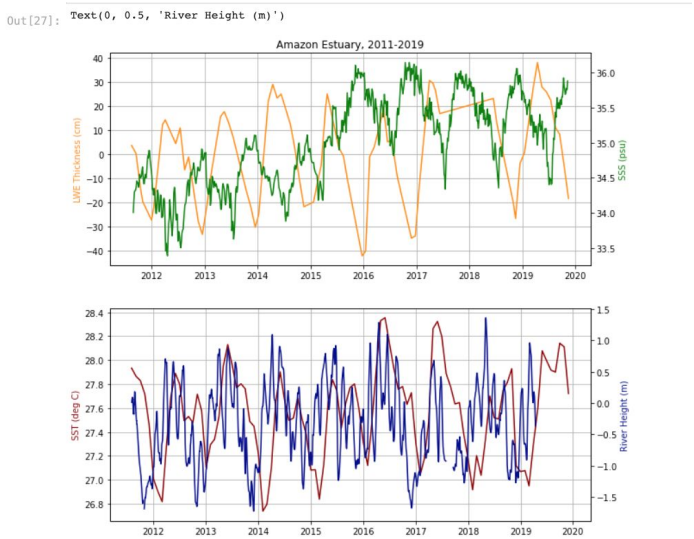
This tutorial will use data from the Estimating the Climate and Circulation of the Ocean (ECCO) model to derive spatial correlations through time for two regions of the Indian Ocean. The goal is to investigate the correlative characteristics of the Indian Ocean Dipole and how the east and west regions behave differently. This investigation was motivated by Fig 2 a,b in the paper by Wang et al. (2016).

The SST and SSH correlation over the Eastern Indian Ocean (EIO) (panel a) and the Western Indian Ocean (WIO) (panel b).

- Wang, H., Murtugudde, R. & Kumar, A. Evolution of Indian Ocean dipole and its forcing mechanisms in the absence of ENSO. Clim Dyn 47, 2481–2500 (2016). https://doi.org/10.1007/s00382-016-2977-y

- https://github.com/podaac/tutorials/blob/master/notebooks/meetings_workshops/workshop_osm_2022/ECCO_ssh_sst_corr.ipynb
- Uses the harmony-py and netcdf-to-zarr services
  - Simplified harmony service (API) calls for granule discovery, temporal subsetting and transformation to Zarr (cloud optimized format):
    - NASA Harmony netcdf-to-Zarr service and harmony-py package
    - xarray, requests, json, pandas, numpy, matplotlib, s3fs
  - The harmony request
    - ecco_request = Request(collection=ecco_collection, temporal=time_range, format='application/x-zarr', concatenate='False')
  - 100% in Cloud Amazon EC2 !! Next to the data

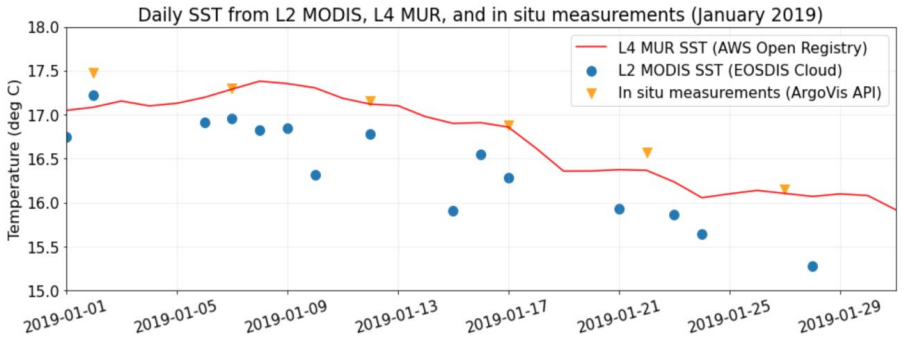
Cloud computing - Amazon River exploration



- https://github.com/podaac/tutorials/blob/master/notebooks/meetings_workshops/workshop_osm_2022/CloudAWS_AmazonRiver_Estuary_Exploration.ipyn
- Explores the relationships between river height, land water equivalent thickness, sea surface salinity, and sea surface temperature in the Amazon River estuary and coastal region from multiple datasets including GRACE-FO, MODIS, SMAP, pre-SWOT hydro
- Uses Cloud Amazon EC2, without any data being downloaded

Cloud computing - Satellite/in-situ matchups

- https://github.com/podaac/tutorials/blob/master/notebooks/SWOT-EA-2021/Colocate_satellite_insitu_ocean.ipynb
- Uses the cloud data download model and number of different repositories and APIs
- Exploration of cloud and non-cloud data in a single workflow
- Main Steps in Workflow
  - Define study region and period of time of interest: e.g., Atlantic Ocean west of Portugal and Morocco, January 2019
  - Get in-situ Argo floats using the Argo API
  - Get coincident SST observed by the MODIS satellite, from the NASA Earthdata Cloud (in AWS)
  - Quality control the MODIS data with daytime and quality flag filters
  - Plot time series comparing the in-situ and satellite data at in-situ location(s)
  - Validate with a third dataset, MUR L4 SST (version stored in the *AWS Registry of Open Data* - public data access)



Moving science to the cloud - the path forward

- For now, cloud data download to local computer is probably the best option for most
  - See *podaac-data-subscriber.py* and *podaac-data-downloader.py* programs (on Github) for easy to use local command line executables
  - Leverage existing cloud data download PO.DAAC jupyter notebook tutorials
- But scaling science justifies eventually moving code and processing to the cloud
  - Example: To download all (future) SWOT data would take nearly 50 years at current projected internet speeds (20PB @ 13.8 MegaBytes/Second)
  - Interdisciplinary use cases more easily supported in the cloud (all NASA Earth science data will be in the cloud!)
  - Promotes open and repeatable science
  - Open potential new ways of working with Earth Observation data using Machine Learning and Artificial Intelligence
- Yet bottlenecks remain….
  - Learning curve (but leverage existing and free activities)
  - Cost management may be an issue (but really not that bad)
  - Amazon cloud services provide lots of choices (overwhelming)
    - Recommendation: start with EC2 (start with a free Amazon account)
  - Amazon not the only game in town…also Google Cloud, Microsoft Azure …and HPCs etc.
    - How to collaborate across cloud providers and HPCs?

Summary

- Amazon Cloud not just for storage and access; provides an ecosystem for in cloud data processing and computing
- NASA family of cloud search/subset/transform services (called harmony) streamline the experience even for local data download (faster time to science, data reduction)
- PO.DAAC tutorials provide data processing and science use case examples for both in cloud and local download in an open source framework
- These PO.DAAC cloud focused resources are there to be leveraged by GHRSST users, and future GHRSST training and capacity building
- This technology is rapidly evolving
    - We encourage users to learn and explore this new cloud paradigm and stand ready to assist !