National Aeronautics and Space Administration

NASA

EXPLORE EARTH

Open Source Science Activities in the Research & Analysis Program*

Jack Kaye, Associate Director for Research
Earth Science Division
Christine Mataya, Booz Allen Hamilton

* Prepared with input from numerous R&A staff

October 14, 2021

Scope of R&A Program

- R&A program advances fundamental knowledge of the Earth system including its major components (atmosphere, ocean/hydrosphere, land surface/interior, biosphere, cryosphere), the couplings that connect them, the human-induced and naturally-occurring forcings that drive it, and the understanding of its variability in space and time – past, present, and future.
- The R&A program also supports many of the enabling capabilities that support ESD activities, including surface-based measurement networks, airborne instruments and platforms, scientific computing, global modeling, and calibration/validation infrastructure.
- The R&A program is implemented at NASA centers, universities, and laboratories of other government agencies, private sector entities, and non-profit institutions, mostly through competed individual investigator awards, with some directed funding, especially for enabling activities at NASA centers.
- The R&A program closely engages with other ESD components (Applied Sciences, Technology, Flight, Data Systems) in support of overall ESD goals.
- The components of the R&A program are closely connected with those of NASA's interagency and international partners through a variety of mechanisms – both bilateral and multilateral, including some organized under formal organizational structure and less formal coordinated plans.

Vision for Science Data Processing and Open-Science at NASA

- The R&A program has long had a policy of open data (especially for networks and field campaigns), satellite data products (jointly with flight program), and model results (including assimilation/reanalysis).
- Publication in open journals has been increasing.
- In recent years there has been more initiation of activities towards open science – not just sharing final products but by involving others in the conduct of the work itself and sharing intermediate products and codes.
- Several solicitations and programs have taken steps in this direction. In this talk only a few examples will be discussed, however. Responses were received from the following activities/programs:

ABoVE Phase 3	Computational Modeling and Cyberinfrastructure Projects
MEaSUREs	High Mountain Asia Team
Space Geodesy Program	Cryosphere Program and ICESat-2 Science Team
Ocean Biology and Biogeochemistry Program	Physical Oceanography (including ECCO, N-SLCT, plus missions)
Land Cover/Land Use Change Program	

- Additional responses pointed to flight program examples that will not be addressed here (NISAR, MAIA)

Computational Modeling and Cyberinfrastructure Program

Requirements

- Computational Modeling Algorithms and Cyberinfrastructure (CMAC) program solicited projects between 2012 to 2020 that would build and enhance community-based data analytics software and tools using Python and clouds and contribute to a jointly developed platform
  - Required open source software licenses in the solicitation
- Open source science projects not new to the NASA Earth Exchange (NEX) project: open platform that integrates data, publication, analysis tools (software), and computing for scientists to analyze data

Implementation Steps

- Platforms: ADAPT cloud at GSFC, NEX at ARC, CMDA and RCMES at JPL
- Solicitations: Computational Modeling and Advanced Cyberinfrastructure (CMAC) solicitations to support enhancements, encourage re-use, and to enforce open source requirements
- Open Source Science Initiative: integrates platforms, developments, policies, lessons learned into a new initiative supported by new technologies

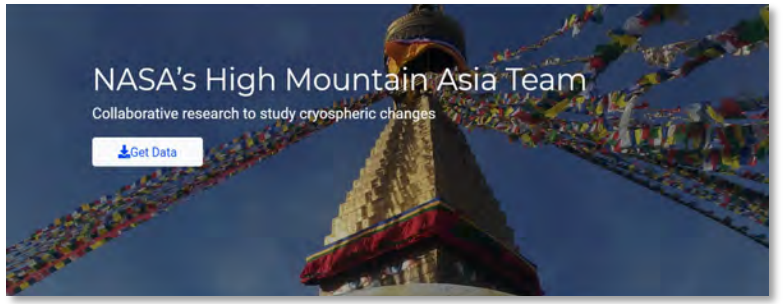
Positive Results/Challenges/Lessons Learned

- Result: initiated large scale MEaSUREs and NCA enabling tools projects to successfully create data sets such as WELD and NEX-GDDP for the research and application communities
- Lesson: just because NASA makes things (publications, data, software, and tools) open, does not mean the community will flock to the open science movement
- Challenge: research community is good at “getting there first” - to be the first to discover, to invent, or to develop
  - But, rarely wants to spend extra resources making things available to others, documenting, and answering questions. Concern is that these will slow down the core business (discover, invent, or develop new ideas)

Next Steps

- Need to change the culture and create incentives to participate in open source science
- Leverage professional societies (e.g., AGU, AMS) and partnerships (ESIP) for training, policy developments, and exchange of new ideas
- Consider inventing an “O (Open Science) -index” similar to the h-index in publications to measure the impact of PIs’ open science contributions

High Mountain Asia Team (HMA)



Requirements

- Included open source science requirements in the HMA-2 ROSES 2019 solicitation:
  - Tools should be interoperable with the understanding that they are being linked as possible into the existing and developing NASA's Glacial Melt Toolbox (GMELT)
  - Proposals must explain how research will be integrated into and/or benefit from GMELT tools, and how the tool will be documented to enable use by broader community in accordance with OSS standards
  - Solicitation sections on open science approaches, open source software, and data policies

Implementation Steps

- HMA-2 team contains PIs who were part of HMA-1, which jump started set up of sharing data/tools
- HQ program management encouraged OSS uptake

Positive Results/Challenges/Lessons Learned

- Results:
  - Helped facilitate collaboration at different levels (e.g., within and across PIs, grad students, and post-docs)
  - Encouraged PIs to write and share routines that would facilitate merging and transforming of data across domains
  - Facilitated combination of more disparate data sets
- Challenges:
  - Getting PIs used to using/understanding new tools
  - Overcoming PI concerns about ideas/research getting scooped
  - Resolving challenges with sharing large data sets (in part because of NASA rules on who can access NASA systems)
  - Handling who pays for access/space in the cloud
- Lesson learned:
  - Graduate students can facilitate adoption of new OSS tools
  - Built on previous collaborations with high trust; OSS tools helped PIs collaborate more readily and deeply

Next Steps

- Trying to find or create a data system that can be used intra-team

Cryosphere Program and ICESat-2

Requirements

- Share your Data
  - via the DAAC at NSIDC or commercial cloud providers
- Share your Knowledge
  - publish your papers as open access available to all when possible
  - promote your research
- Share your Source
  - publish and reference your code
  - treat it as your intellectual property and technical growth
  - ask for resources to maintain scientists' work

Implementation Steps

- Training the community to adopt new tools
  - Instituting ICESat-2 "Hackathons" for interested data users to develop their skills with data and share their code/tools
- Incentivizing source sharing by making open science a critical component of competitive proposal selection process

Positive Results/Challenges/Lessons Learned

- Results:
  - Code developed and shared during ICESat-2 Hackathons are still being used and expanded by attendees and other researchers
  - Leading PIs in the field of community-driven model development (CISM, ISSM, CFM)
- Challenges:
  - The competitive nature of science funding results in a certain reluctance of early code and data sharing.
  - Inconsistent guidelines for open-source science practices across disciplines

Next Steps

- Provide science/tech partnering opportunities
- Evolve OSS requirements to better serve community/customers' needs
  - Helping develop standardized OSS practices and regulations across disciplines to improve community understanding and compliance
- Support open-source library development
- Train and incentivize community to embrace new paradigm

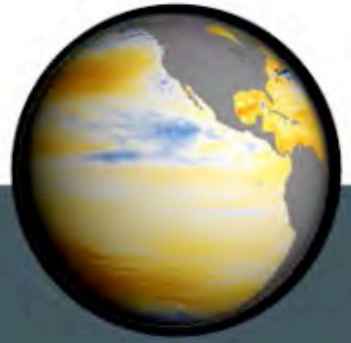
NASA Ocean Physics Open Science – progress and success stories

Sentinel-6 MF

First NASA's cloud-native mission

ECCO

First NASA direct-to-cloud ocean/ice data integration system



N-SLCT

Open and actionable sea level science

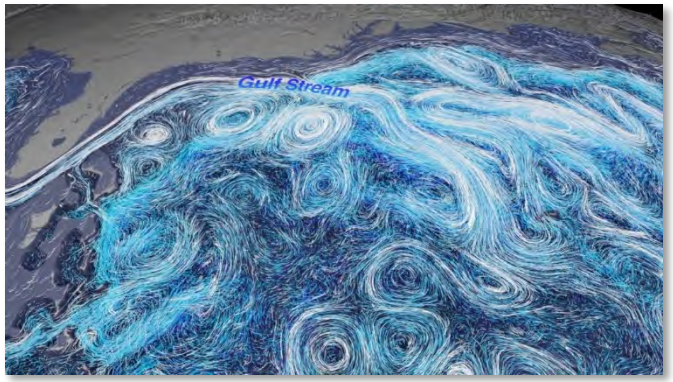
SWOT

Open Science revolution

Community

Building community of awesome PIs!

Physical Oceanography Program



Requirements

- Share your Data
  - via PO.DAAC or commercial cloud providers
- Share your Knowledge
  - publish your papers as open access available to all
  - promote your research
- Share your Source
  - publish and reference your code
  - treat it as your intellectual property and technical growth
  - ask for resources to maintain scientists' work

Implementation Steps

- Making open science as a scorable metric of competitive proposal selection process
- Co-developing implementation strategies that work best for ocean physics community, with their input and needs
- Incentivizing source sharing
- Training the community to adopt new tools

Positive Results/Challenges/Lessons Learned

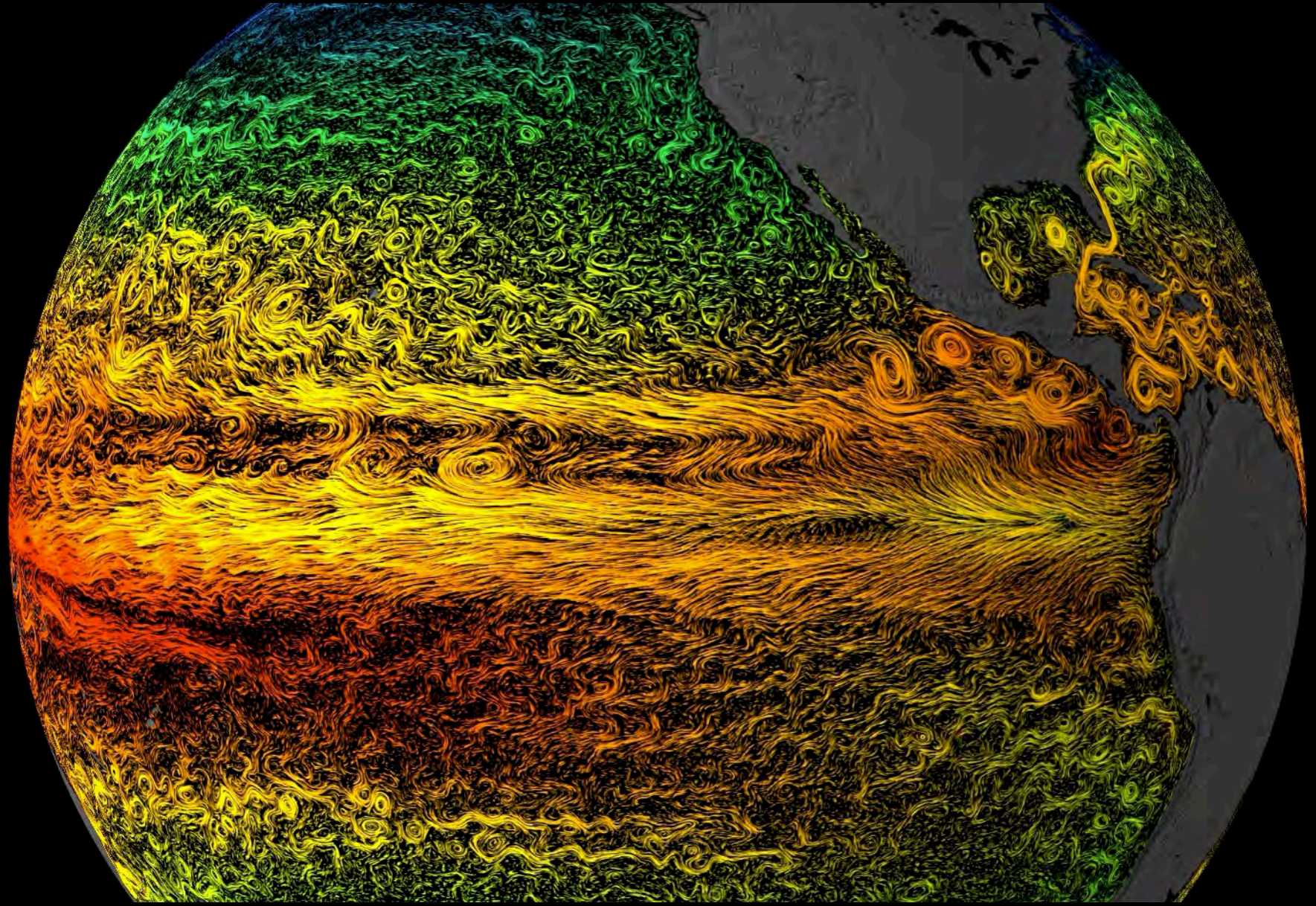
- Results:
  - First NASA missions with cloud-native data delivery (S6 MF, SWOT)
  - First NASA climate state estimate with cloud capabilities (ECCO)
  - Leading PIs in the field of open geoscience (Pangeo, MIT, JPL)
- Challenges:
  - Potentially introducing inequality in the science community
  - Marginalizing groups that are unable to maintain both discipline excellence and technological savvy
  - Disconnect between HQ requirements and realities of science business and practices

Next Steps

- Provide science/tech partnering opportunities
- Evolve OSS requirements to better serve community/customers' needs
- Support open-source library development
- Train and incentivize community to embrace new paradigm

NASA ECCO

First NASA's cloud-based multi-platform data integration & modeling framework for climate research



Estimating the Circulation and Climate of the Ocean (ECCO)

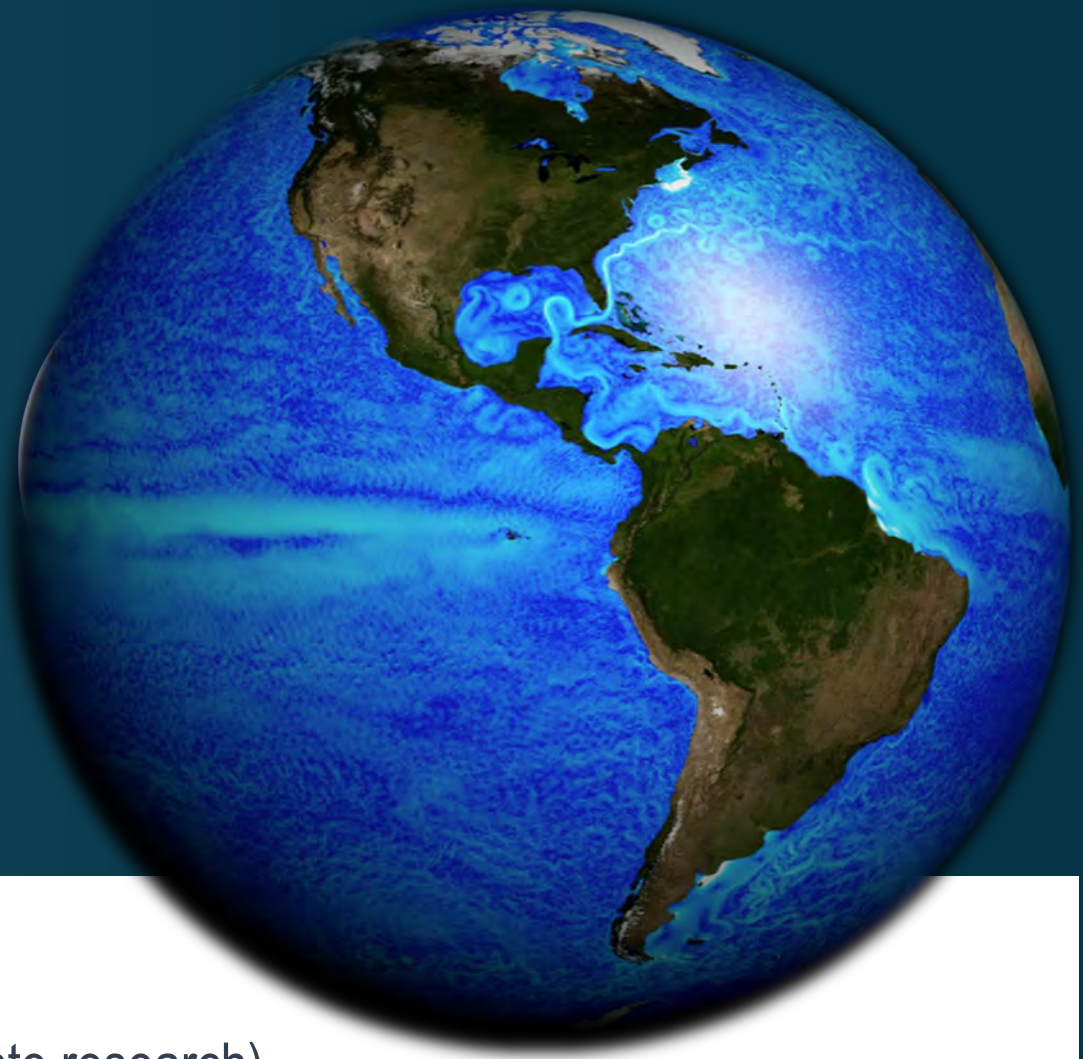
https://www.ecco-group.org/

NASA ECCO

How to reconcile various measurements into a consistent ocean state estimate?

NASA uses basic physical principals and understanding of data uncertainties

$F = ma$



Advanced data assimilation and machine learning paradigm

Dynamic consistency & property conservation (essential for climate research)

1000s of publications and climate discoveries (e.g., AR6 IPCC)

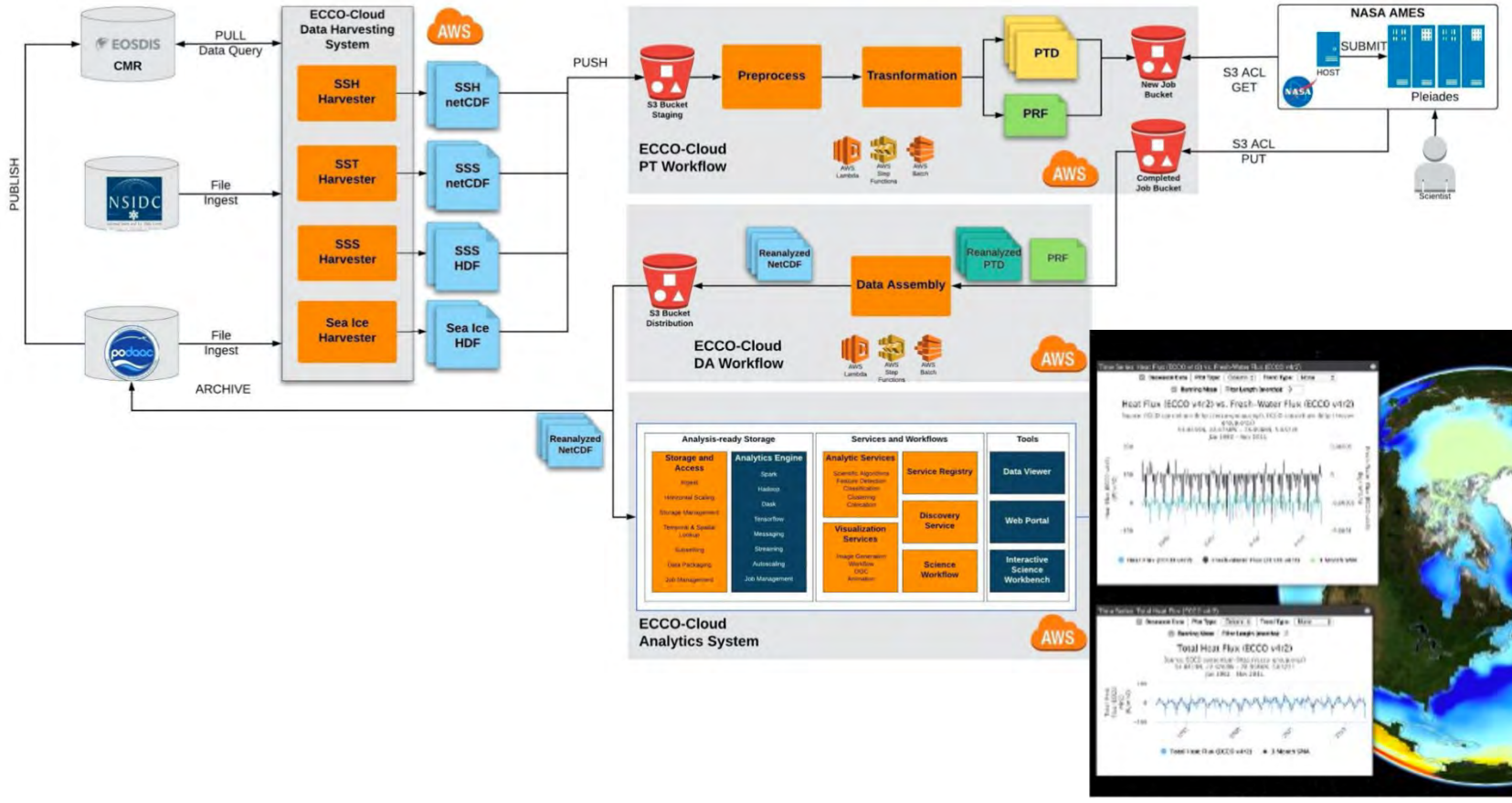
Open Science compliant with cloud-based analysis and tools

Requirements, Constraints, Recommendations, and Opportunities to Consider

- Need to be sure that scientists' contributions to open source science (e.g., sharing of algorithms, codes, etc.) gets appropriate recognition by agencies and community, including getting factored into selection, awards, & promotion.
- Need to anticipate expected demands on part of "code users" for support from originators and/or need to continue to provide support for algorithm/code used in community as technology, funding status, and/or personnel change. This may include being sure that teams providing widely shared codes, etc., are funded to provide needed support, including documentation (or community understands that they should not expect that support).
- Publication costs for publication in open journals should be covered (included in grants?).
- Need to be sure that we don't inadvertently create a dynamic where some will "sit back" and let others do the algorithm code development and then just "swoop in" to take advantage of others' hard work.
- Need to be sure that dynamic doesn't just "favor the well-supported and well-connected" who may be best positioned to utilize products developed by others – we may need to think about what we will have to do to enable a broader swath of the community to benefit from and participate in the opportunities now available. Hackathons may be helpful here because of their ability to engage large numbers of people; training can help as well.
- Need to be sure to treat different types of scientific work equivalently in terms of openness or be clear why not doing so – including both hardware-focused activities and those tied to algorithms/codes/software (note ITAR!).
- Need to be sensitive to the fact that partners (especially private sector but also international) may have different approaches to sharing information (e.g., proprietary nature).

Additional Charts

ECCO Cloud System Architecture for Climate Model & Big Data Challenges (Acquisition, Transformation, Distribution, Discovery)



Making Earth System Data Records for Use in Research Environments (MEaSUREs)

Requirements

- Meet all applicable U.S.G-mandated standards for data products and information systems
- Regarding data products designated for distribution, comply with the Data and Information Policy for NASA's Earth Science
- Data products designated for distribution shall contain and be searchable via ISO 19115 Geographic Information - Metadata standards. Implementation details shall be worked out in interface control documents in collaboration with the project-specific DAAC and the ESDIS Project to ensure that MEaSUREs project's products ingested at EOSDIS DAACs are searchable in ways similar to other products at those DAACs
- Release designated data, along with the source code and/or other algorithm documentation, and ancillary data to the designated project DAAC, as specified in the submitted proposal
- Make public-domain products & services available on an internet-accessible web server
- Participate in relevant scientific meetings identified by the MEaSUREs Program Manager as pertinent to project goals and user communities

Implementation Steps

- Selected projects were funded via Cooperative Agreements: funded institutions held responsible to deliver to NASA data and code
- Projects were able to promote their products (websites) and sponsor community workshops and beta-testing prior to final delivery
- Project participation in NASA community meetings allowed interaction with research programs, science teams and field campaigns
- 5 year projects allowed development and testing prior to full production

Positive Results/Challenges/Lessons Learned

- MEaSUREs is well-established in the Earth science community, extremely successful
- Many data sets have been heavily used by researchers and can be found in research publications (individual projects monitor). ESDS stats on data downloads.
- Challenges in data set extension and code usability is the considerable dependency on auxiliary data sets

Next Steps

- MEaSUREs new competition in ROSES 2022 for 5 years

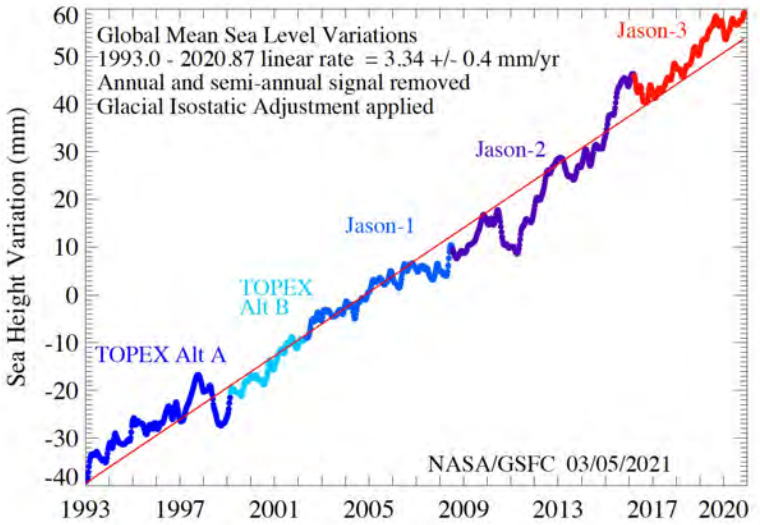
Space Geodesy Program (SGP)

Requirements

- Global-international nature drives use of open source science (OSS) approach
- Geodetic measurements openly exchanged/shared to provide global coverage

Implementation Steps

- Formation of international geodetic services for Earth rotation, GNSS, laser ranging, VLBI, and DORIS techniques (e.g., IERS, IGS, ILRS, IVS, IDS) to coordinate geodetic activities and encourage international cooperation/collaboration
- SGP plays a leading role in all the geodetic services and actively works to ensure OSS principles are maintained and implemented



The decadal sea-level change curve depends on the OSS-derived ITRF

Positive Results/Challenges/Lessons Learned

- ITRF built on OSS principles and has become foundation of precise orbit determination and geolocation that enhances scientific analysis of datasets taken by multiple instruments over different locations and times
  - Accessible: All geodetic measurements and products are made publicly available through open data centers such as NASA's CDDIS. Many geodetic analysis software packages are open source
  - Reproducible: Multiple analysis centers compare results against each other to ensure consistency and reproducibility
  - Transparent: Methodologies and standards are posted publicly on services' websites and in peer reviewed journals
  - Inclusive: All services have broad international participation and regularly solicit and welcome new members. Regular workshops and "schools" are sponsored by SGP and geodetic services to share information and train the next generation of geodesists
- IGS and contributing organizations, including SGP, helped turn GPS from military infrastructure into one of the foundations of Earth Science with diverse applications

Next Steps

- Continue to build the community by encouraging new participants and contributors to the international geodetic services through international forums such as the United Nations Global Geospatial Information Management (UN-GGIM)
- Continue to develop bilateral partnerships to expand access and production of geodetic data from underrepresented regions such as South America and Africa
- Continue to play a leading role in the international geodetic services to integrate and improve scientific collaboration

Ocean Biology and Biogeochemistry (OBB) Program

Requirements

- All OBB solicitations request open data sharing to increase data access, scientific repeatability, and transparency
- OBB data sharing policies require immediate access (with no period of exclusivity), as per NASA policies
- Encourage and support open access publications from all funded projects
- OBB participates in science meetings and community workshops to share programmatic information, activities, data, and applications of OBB data
- Participate in NASA outreach and hackathon activities to demonstrate (1) ease of discoverability and accessibility of OBB data products and collections and to improve (2) the language explaining OBB data holdings

Implementation Steps

- Hyperspectral In Situ Support for PACE (HyperInSPACE) was released via NASA GitHub as an open-source project for community development
- OB.DAAC's SeaBASS element is currently expanding the number and types of variables it is ingesting, and developing more comprehensive metadata to enhance data useability
- The SeaDAS software provides open-source processing code and data analysis tools for NASA Ocean Color data holdings.

Positive Results/Challenges/Lessons Learned

- Results:
  - Most PIs request funding for open access publications
  - Data collected during campaigns NAAMES, EXPORTS, and CORAL were archived and made publicly available to non-science team research communities to support new science and discovery; data are being used by others not directly involved in the efforts
  - Investments in SOCCOM project and novel in-water optical profiling floats and data are being used by the broader community
- Challenge: Sharing, archiving, and hosting scientific and data processing code; modeling code is often shared through PI websites, but it is unclear how/where NASA can host and archive code

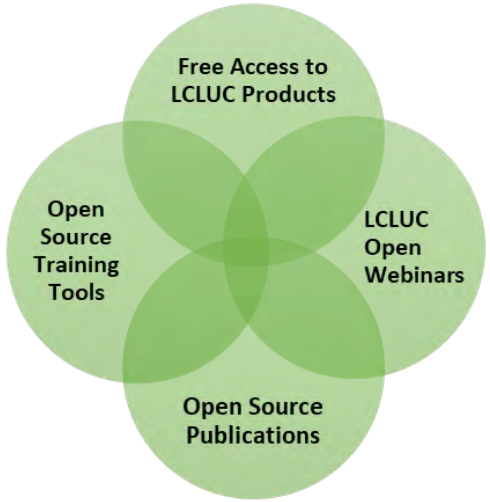
Next Steps

- Continue to promote increased awareness, discoverability, and accessibility of OBB data collections, products, and holdings
- Encourage further adoption of open access publications
- Encourage the use/development of open-source code/libraries
- Continue to integrate open science principles in OBB programmatic calls, in accordance with NASA policies

Land-Cover and Land-Use Change (LCLUC) Program

Requirements

- Free Access to LCLUC Products
- Open Source Tools in Training Events
- Open Webinars on LCLUC Science
- Open Source Publications (*optional*)



Positive Results/Challenges/Lessons Learned

- Lessons learned and related challenges:
  - Slow response from PIs as they would like to publish data and associated algorithms in journals first before sharing products freely
  - Unlike commercial software, open source tools can lack robust functionalities. One has to look for multiple open source tools (not a single one) to do the same job. Also, wherever coding is involved using open source tools, there is a learning curve (compared to clicking a button in commercial software to achieve the same result. Example, “R” statistical software). Thus, taking more time to train people.
- Challenge: Although beneficial to larger community, open source journals charge publication fees (~$1500-$2500/article); PIs are paying this from funded projects; PIs might ask for more funding to cover charges

Implementation Steps

- Sharing its PI-generated products through the LCLUC website. A dedicated metadata page serves multiple researchers worldwide
- Conducting ~3-4 training events in different countries to promote NASA data, LCLUC algorithms, and approaches. In all training events, use open source code and open source software
  - Examples include free and open source Cloud computing - Google Earth Engine; Geospatial tools – QGIS; Scripting tools - Python, Julia; Statistical software - "R"
- Conducting frequent research exchange webinars involving LCLUC PIs, which are open to all researchers. Every year 120-150 researchers typically attend these webinars. Also, videos are posted after the webinars for free access
- Promoting open source publications by leading and organizing special issues in Open Access journals on various topics. Articles published are freely accessible (*see notes for details*)

Next Steps

- Continue to encourage PIs to follow open science protocols to share algorithms, products, publications, expertise (as trainers)
- Integrate open science principles in LCLUC solicitations following NASA policies

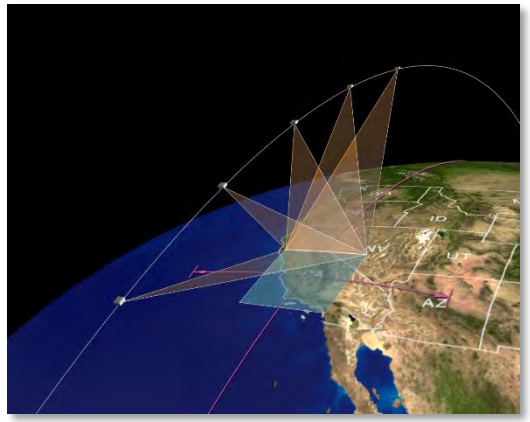
Snapshots of Open Source Science (OSS) in the Research & Analysis (R&A) Program

- Called out OSS and the OSS Policy in the **ABoVE Phase 3** and the **EVI-6** solicitations
  - Asked PIs to explain how their proposal would advance OSS at NASA.
  - Will do the same for the next CMS solicitation
- Expect to fund one SOSS-20 proposal for OSS managed through Terrestrial Ecology and another managed through MEaSUREs
- **SNWG** products make access to higher level products easier as data from multiple missions used to generate consistent product with greater temporal frequency; most products in the cloud (e.g., HLS)
- **NISAR** is emphasizing cloud-based activities/computing and using Jupyter Notebooks for the ATBDs such that people can run and work with the algorithms that NISAR will be using
- **GRFN** provided an understanding of what is possible in the cloud environment for a wide range of SAR/InSAR sciences, while **OpenSARlab** is 'sandbox'-like environment that helps researchers begin transitioning to cloud computing/processing. Both are helping to provide motivation (financial, scientific benefit, time savings) and tools that help people processing data on the computers and will make OSS successful

Snapshots of Open Source Science (OSS) in the Research & Analysis (R&A) Program

- **ACTIVATE** is archiving data for public use following other field campaign protocols for data dissemination; archiving open-source software and data analysis tools for public use; making publications openly available; delivering presentations to public audiences about data and findings; engaging diverse communities to share the science undertaken (includes outreach with the public and open data workshops)
- **Hackathons**: **SnowEx** held one in summer 2021; **SWOT** held an Early Adopters Hackathon in May 2020 and March 2021
  - Introduce new groups to the data
  - Can maximize readiness of data by a broad range of user communities after a launch/deployment
- **MEaSUREs** has rigorous OSS requirements in its solicitations (not limited to the following):
  - Data products designated for distribution to comply with the Data and Information Policy for NASA's Earth Science
  - Data products designated for distribution shall contain and be searchable via ISO 19115 Geographic Information - Metadata standards. MEaSUREs project's products ingested at EOSDIS DAACs are to be searchable in ways similar to other products at those DAACs
  - Release designated data, along with the source code and/or other algorithm documentation, and ancillary data to the designated project DAAC
  - Make public-domain products and services available on an internet-accessible web server

Multi-Angle Imager for Aerosols (MAIA) Instrument



Requirements

- Some key open source science elements in MAIA's PLRA Science Data Management and Science Data Requirements sections include:
  - Public release of data shall conform to Earth Science Data and Information (ESD&I) Policy
  - No period of exclusive access
  - Algorithm Specification Documents shall be delivered to the DAACs at the time of deliveries associated with Level 1 and higher products
  - Publicly available data products shall be distributed in standardized formats in conformance with Data System Standards
  - For standard data products that can be represented as images, the project shall generate full-resolution browse products
  - Project shall transfer to the designated DAACs, all documentation required for long-term preservation of knowledge about the resulting products

Implementation Steps

- Instrument is in development
- With objective of optimizing the data for various user communities, MAIA project and Applied Sciences Program partner on a MAIA Early Adopters Program, which invites the community of potential users to give feedback on planned products and provides user resources pre-launch to streamline incorporation into workflows

Positive Results/Challenges/Lessons Learned

- TBD as instrument is in development

Next Steps

- PM data, currently planned to be NASA's first operational PM products, will be freely available through the ASDC
- Partnering with ASDC to improve user experience through data visualization and download tools; synergies with the other air quality missions that ASDC supports, including TEMPO; and incorporation into NASA-wide tools such as Worldview and Earthdata Search