Open Source Science for ESO Mission Processing Study

Identify a system architecture that meets the ESO mission processing objectives, supports open science, enables system efficiencies, and promotes earth-system science.

Workshop #1
October 19-20, 2021

NISAR Science Perspective

Paul Rosen, Susan Owen, Gerald Bawden

NISAR Mission Science Objectives and Level 1 Requirements

Mission Need	
Data rate	35 Tbits/day (4 Gb/s downlink)
Raw Data volume	5 Pbytes over 3 year nominal mission
Product Data Volume	100 Pbytes over 3 years
Latency	1-2 days nominal 5 hours urgent response
L0-L2 Data products Project-Generated Globally Generated	L0 raw data L1 Single Look Complex (radar coords) L1 Interferogram L2 Single Look Complex (geocoded) L2 Polarimetric Covariance (geocoded) L2 Interferogram unwrapped (geocoded)
L3 Science Products (Only generated at validation sites; not global)	Ground deformation time series Ice-sheet velocity time series Sea-ice deformation time series Biomass Disturbance Time series Wetland Inundation time series Agricultural area change time series

NISAR Mission Science Objectives and Level 1 Requirements

Plan for Algorithm Development, Analysis, Cal/Val

- Project Algorithm Development Team (ADT) develops L0-L2 product *open source* algorithms in consultation with Science Team
- Science Team develops L3 product *open source* algorithms in consultation with ADT
- Science Team develops validation *open source* algorithms and runs code during mission to validate requirements, supported by project team and computing infrastructure
- Cloud-based production system run by project
- US Science team and Cal/Val partners use project production system for validation throughout the development and operations cycle
  - Presently the system operates within the JPL firewall, which is limiting for science team and partners
  - Cal/Val data base is also on the cloud inside JPL firewall
- ISRO science team uses ISRO systems – no plan for common production platform, but will have a data sharing portal

Plan for groups to collaborate and share information and code

- ROSES PIs, Agency partners will have access to all software on github/lab and mirrored cal/val data. Computational resources will be through other NASA cloud assets co-located with the data, e.g. Alaska Satellite Facility's OpenSARLab, or hyp3

Supporting Earth System Science

- How will L1+ Data Products be generated? For example, based on past mission experience, for L4 model enhanced products – what kind of algorithms / models / ancillary data do they rely?
  - L0-L1 uses straightforward backprojection image focusing, standard algorithms for interferograms and new area-projection backscatter correction
  - L2 is created by straightforward geocoding L1 products
- Who is expected to generate each of these L1, L2, L3, L4+ products?
  - See previous slides. L0-L2 are project derived globally. L3 are science team derived at validation sites
- What resources would product-generators need/use? Data, software, AWS, Supercomputers, etc?
  - NISAR is using AWS, custom c++ code and python, and standard opensource tools like gdal, numpy, eigen in containerized conda environment instances
  - Data products will be stored in the ASF AWS DAAC by flipping a switch from production to archive (effectively)
- How portable/open are any existing workflows? Are there license/sharing/building issues?
  - All workflows are open source.
  - Portability through conda environments is straightforward
  - Portability in other build environments has been a challenge
    - versioning of different OS and build systems for open source software is not always easy
    - Compatibility of hdf5 with particular OS and python seems fragile
- How does any L3+ ESO products support Earth System science? Through data assimilation? Derived data products? Legacy record (e.g., PoR)?
  - NISAR products will be used for boundary conditions on a range of models, extending and improving on legacy records from SAR and GPS
    - Ice sheet velocities to constrain ice sheet models
    - Sea ice motion/deformation to constrain ocean/atmosphere interaction models
    - Surface deformation to constrain fault, magma chamber, reservoir, landslide models to assess risks, coastal subsidence/uplift
    - Biomass/biomass change, disturbance, inundation, agriculture time series to constrain carbon flux models
- What are the pain points for working with the data/software/computer access?
  - NISAR is all-in on cloud computing. We are just scratching the surface on tools to best exploit cloud systems
  - Community is still convinced it is cheaper to buy their own systems and download the data
  - The largest problems are not being tackled because of data availability now, and data access/cost in the future

Supporting Open Science

- Thinking about accessibility, at what point do you consider data products (L1+) open access?
  - There are varying opinions on this, but the project scientists feel that instant open access with appropriate quality metadata is most conducive to scientific progress
- What are your project plans for growing applications / additional products based on the community's past mission experience? Extended mission activities? Applications? Science Team? Agency partnerships? Future ROSES solicitations?
  - For Applications related to US Government agencies, we look to the Satellite Needs Working Group
  - Extensive community engagement with Application Community workshops (10 and two more in the works), Envoy program, etc
  - For Science, we look to Measures, other ROSES solicitations, NSF science partners
- How are you authenticating a user for producing a product (especially L3+)? Is there an expectation they would be on-boarded/must be within the project/ mission-supported?
  - Algorithms for L3 products can be contributed without vetting (contrib area). Science Team adopted algorithms would be adopted/vetted and included in the gitlab for use by the community (project area). Vetting process TBD
- What are the criteria for an algorithm to be viable and what is the process for cal/val of that algorithm?
  - New algorithms would need to be run through the same process as planned algorithms.
  - Demonstration would be through a validation procedure that the algorithm creates products that meet requirements. All NISAR algorithms and validation workflows will be available on gitlab for emulation to demonstrate compliance.
- How will your project support open science through documentation, community development, open communications, and increasing accessibility to knowledge?
  - All L0-L2 software comes with complete documentation and tests with a full CI process. Community developments can be evaluated for compliance and adopted as needed.
  - The isce2 development over the past 10 years is an example of how this can work.
  - Dedicated funding beyond the project for maintenance is needed, after the project sunsets.
  - L3 algorithms and validation workflows will be through jupyter notebooks, so have documentation and code intertwined.
- How could a common framework, that provided easy access to another mission's algorithms and data, amplify your project's science objectives?
  - Should L3/L4 algorithms become operational, and involve multiple data sets (as with OPERA project), unified data and algorithm catalogs that allow for intelligent searches would save scientists the work of finding relevant data and algorithms
  - Semantic ontologies should developed to aid scientists in searches
  - Collection of quality and resource utilization metrics should be standardized and databased, to be used in machine learning assisted workflow assistance and guidance
  - Visualization of algorithms (workflow graphs), data products, and science results needs heavy investment and standardization
  - Data fusion

Are there any requirements, constraints, barriers, recommendations, or opportunities that you would like the study team to be aware of?

- Currently all project workflows operation on the cloud within the JPL firewall. This is an impediment for partners because the barrier to be included in the firewall is too difficult for most to overcome
- Security background check
- Difficult process to sign up for and maintain TFT authentication, tokens, passwords
- Cloud tools are still in their infancy. The on-demand system NISAR is developing is shaping up, but will need to become more user friendly, and more complete to be easily adopted by the broader community
- Machine-learning assisted workflow management based on collected databases of quality metrics and resource utilization will be key to efficiently exploring massive data sets to deliver new science
- Discovery and low-cost access of data in physically distinct cloud regions will be a challenge
    - Require centrally-controlled, distributed computing where data reduction operations are conducted near large-volume data, and results then transferred between regions for fusion
- Software maintenance needs to continue beyond the sunset of a particular project
    - ROSES opportunities
    - Strategic ESDIS programs for software maintenance