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

MC Science Perspective

David Wiese
Caltech/Jet Propulsion Laboratory

MC Mission Science Objectives and Level 1 Requirements

- What are your key and driving requirements?
  - What are the data rates, data volume, processing needs, latency?
    - Data Rates and Data Volume: low (~380 MB/day)
    - Processing Needs: Supercomputing resources required; large amounts of intermediate data storage required (e.g. current local compute cluster is 500 TB - mostly full)
    - Latency: Still TBD on final architecture, but not considered to be a driving requirement.
      - GRACE-FO currently produces “quick-look” fields with 1-day latency in processing time.
  - What are the expected data products L1+?
    - L1, L2, and L3 – responsibility of Project
    - L4 data products not a requirement, but carry high value
  - What is the plan for Algorithm Development, Analysis, Cal/Val?
    - Algorithm Development and Analysis: Reliance upon 20 years of mature science data processing. Some minor complementary algorithm development/evolution may be explored dependent upon final architecture
    - Cal/Val: See next slide
  - Who needs access and when? (Science, Cal/Val partners/ ROSES PIs/ Agency Partners/etc)
    - Project/Instrument/Flight Ops: immediate access to L0 data.
    - Science: L1 = 14 days; L2 = 40 days (typical); quick-look fields = 1 day
  - What is the plan for groups to collaborate and share information and code?
    - GRACE/GRACE-FO SDS Processing centers have rich history of tight collaboration on intellectual topics; L1+ code not shared since each SDS has their own unique software packages. This diversity in software/approaches is of great value to test parameterizations and implementations, contributes to V&V; past G/GFO experience has leveraged the diversity to improve products

Calibration/Validation

- The long spatial wavelength of the observable (the gravity field) limits the calibration and validation strategy to space observations at similar scales
- Mass change is a uniquely observed Earth system variable
- The exact cal/val approach will continue to be assessed.
  - Overlap with GRACE-FO is the best form of cal/val
- External Validation Options (all deficient in some way; long history of using these with GRACE/GRACE-FO)
  - Satellite Laser Ranging
  - GNSS data from Precise Orbit Determination of satellites in LEO (e.g. SWARM)
  - Ocean Bottom Pressure Recorders
  - GNSS data (Earth surface deformation)
  - Extrapolation of existing mass change data record
  - Known variations in regions of low variability
  - Satellite altimetry and ARGO over the ocean
  - Satellite altimetry over the ice sheets
  - Terrestrial Water Storage Reconstruction products
  - Models of mass variations (hydrology and ocean)
- The above observables also capture areas of synergy; ripe for addressing Earth system science questions when viewed in a common ecosystem

Supporting Earth System Science

- How will L1+ Data Products be generated?
  - Based on past mission experiences and successes (see next talk for details)
- Who is expected to generate each of these L1, L2, L3, L4+ products?
  - L1: JPL
  - L2: JPL/CSR/GSFC/Partner
  - L3: JPL + user community
  - L4+: user community (not responsibility of project, but project can contribute meaningfully)
- What resources would product-generators need/use? Data, software, AWS, Supercomputers, etc?
  - L1-4: Appropriate data and software
  - L2: Supercomputer; large amounts of intermediate data storage
- How portable/open are any existing workflows? Are there license/sharing/building issues?
  - Existing workflows have limited portability and openness. Portability is easy to address with software engineering practices.
  - Licensing issues with software; For GRACE-FO Caltech considered code developed at JPL proprietary
  - The software suites used to process L2 data have been developed over decades; consist of hundreds of thousands of lines of code
  - No clear need identified over 20 years of G/GFO to make L1/L2 code open
- How does any L3+ ESO products support Earth System science? Through data assimilation? Derived data products? Legacy record (e.g., PoR)?
  - L3+ products significantly support Earth System science. Data assimilation is one very important way, but not the only way.
  - Creation of mass change climate data record at L3+ level is critical for Earth System science.
- What are the pain points for working with the data/software/computer access?
  - Complexity of data processing at L1-L2 requires users with significant expertise.
  - Expertise requirements are orbital dynamics, estimation theory, large linear systems, geodesy; not software/data science
  - Significant computation/data storage (intermediate) needs at L2;
  - Resources: Ability to reprocess 20+ years in an efficient manner, repeatedly, for necessary experiments.

Supporting Open Science

- Thinking about accessibility, at what point do you consider data products (L1+) open access?
  - Intellectually: all Levels of data products are open and accessible, and methods/algorithms are transparent
- 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?
  - Involving Science Team in advancing methods, applications via ROSES Science Team Calls, but also many other calls (MEaSUREs)
  - European partners have additional funding sources through special programs – valuable from GRACE/GRACE-FO experience
- 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?
  - Project would typically not promote products from individual users – anyone can produce L3/L4 products
  - Past experience shows that too many flavors of data products can be detrimental to expanded uptake by community by causing confusion with new users. Less can be more.
- What are the criteria for an algorithm to be viable and what is the process for cal/val of that algorithm?
  - Algorithm is tested in a forward numerical simulation run in a controlled environment to test viability and performance
  - Real data processing: cal/val methods (previous slide) can be used to help assess algorithm viability
- How will your project support open science through documentation, community development, open communications, and increasing accessibility to knowledge?
  - Documentation of all algorithms to generate L1, L2, and L3 data products, just as was done in G/GFO
  - Open Science Team Meetings, AGU/EGU town halls, user workshops
- How could a common framework, that provided easy access to another mission's algorithms and data, amplify your project's science objectives?
  - MC observations are unique; little overlap with other missions' SDS
  - Unclear benefits of a common framework to access another mission's algorithms and data for SDS
  - Common framework to access L3/L4 datasets from other missions has benefits for Cal/Val; promotes synergistic Earth system science

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

Constraint: Proprietary software to do L1 and L2 processing

Barrier: MC Paradigm: “The satellite is the instrument”. L1/L2 processing requires an expert user with sophisticated knowledge of the entire spacecraft system.

Recommendation: There is limited value in making L1/L2 data processing completely open as the target audience is narrow and the benefit is unclear and most likely negligible. Value increases significantly with L3/L4 data products, in particular when viewed concurrently with other observations. MC data to address Earth system science happens primarily at the L3 and L4 level where users have the most access and control.

Recommendation: Expert skill is needed for L1/L2 SDS; it is not only a software and data science problem