Easy, Scalable Subsetting of GEDI Point Clouds

Charles Daniels[1], Alex Mandel[1], Jamison French[1], Aimee Barciauskas[1], Brian Freitag[2]

[1]Development Seed, [2]NASA

AGU23, Fall Meeting, Dec 11-15, 2023

IMPACT
Interagency Implementation and Advanced Concepts Team

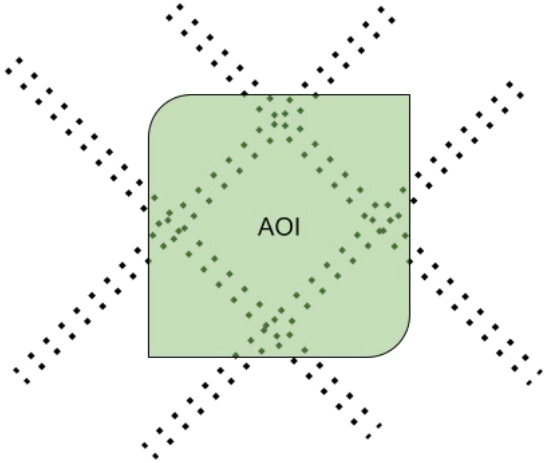
NASA

Outline

1. Motivation
2. Challenges
3. Solution
4. Results
5. Future work
6. Resources

Using GEDI Data for Earthdata Research

- **GEDI: LIDAR** instrument on International Space Station generating high-resolution **observations of 3D features of Earth**, such as:
  - forest canopy height
  - canopy vertical structure
  - surface elevation
- **Point data** collected along **orbital tracks**
- **Challenge:** efficiently subset orbital data by area of interest (AOI)



Challenges to Research Using a Large Dataset

- Datasets too large for locally downloading all required data
- Subsetting data may be tedious and time-consuming
- Ability to scale computation may be limited
- Requires significant programming effort diverting time away from science work
- Example GEDI L4A data over Equatorial Guinea (AOI)
  - Input size: 1134 Granules in bbox, 277 intersecting AOI (~78 GB)
  - Processing time: Originally 9 hours, now down to 1 hour
  - Results: 1.1 GB GeoPackage, 4,141,356 points

MAAP: Cloud System for Collaborative Research

- The **Multi-Mission Algorithm and Analysis Platform** (MAAP): collaborative project between NASA & ESA (European Space Agency) for collaborative research
- Designed to **combine data, algorithms, and computational abilities** for the processing and sharing of data related to **NASA's GEDI, ESA's BIOMASS, and NASA/ISRO's NISAR missions**.
- Addresses **data storage, processing, and sharing** challenges of **high-volume, heterogeneous data** from these missions (collected from satellites, aircraft, and ground stations at various resolutions, coverages, and processing levels)
- An **algorithm development environment** (ADE) to create repeatable, shareable science tools for the research community
- **Open source software**; adheres to ESA's and NASA's commitment to open data.

GEDI Subsetter: A MAAP Algorithm

Simplifies and speeds GEDI data subsetting by allowing users to specify:

- A **GEDI collection** (available collections: L1A, L2A, L2B, L4A)
- An **area of interest** (AOI), and a **temporal range** to limit granules found by searching **NASA's Common Metadata Repository** (CMR).
- **One or more measurements of interest**, such as *agbd* (above-ground biomass density) to limit the number of columns in the result.
- A **query expression** to select only rows that match the expression, such as a quality value or sensitivity minimum.

The subsetter subsets each granule file (HDF5) in parallel (limited by CPU capacity) appending all results into a *single* GeoPackage (GPKG) file for analysis.

GEDI Subsetter (continued)

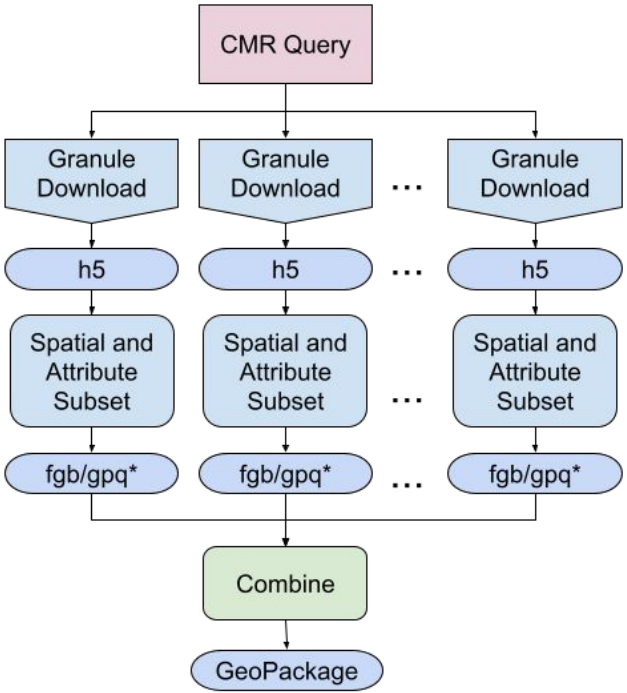
```
from maap.maap import MAAP

inputs = dict(
    aoi="https://.../GNQ-ADM0.geojson",
    doi="L4A",
    lat="lat_lowestmode",
    lon="lon_lowestmode",
    columns="agbd,agbd_se,geolocation/sensitivity_a2",
    query="l2_quality_flag == 1 and `geolocation/sensitivity_a2` > 0.95",
)

maap = MAAP(maap_host='api.maap-project.org')
maap.submitJob(algo_id="gedi-subset", version="0.6.1", ...,
     queue="maap-dps-worker-32gb", **inputs)

# Output: gedi_subset.gpkg (within user-specific job location)
```

GEDI Subsetter (continued)

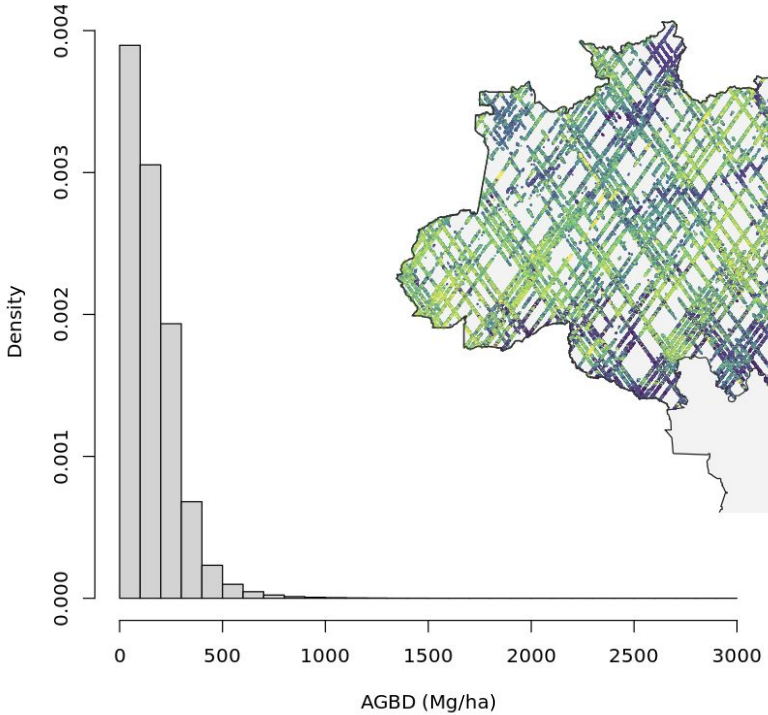


Timings of different workers (Equatorial Guinea)

- 2 CPUs, 8GB RAM, ~9 hours
- 4 CPUs, 32GB RAM, ~4.5 hours
- 16 CPUs, 32GB RAM, ~0.93 hours (56m)

Takeaway: The more CPUs we have, the more granules we can run concurrently, reducing the amount of time required.

GEDI Subsetter (continued)



Credit: Laura Duncanson, University of Maryland (to confirm)

Conclusions

- GEDI Subsetter MAAP algorithm allows researchers to easily **specify an area of interest (AOI) and measurements of interest** to obtain only the GEDI data required for their research **with no coding required** to produce the data.
- Parallelization of data subsetting yields **roughly an order of magnitude speed improvement**, depending on resource selection.
- MAAP's job scheduler allows researchers to run algorithms **without concern for storage and compute resources**.
- Source code (repo link) available for adoption or as inspiration for other data processing

Future Work

- Read data directly from AWS S3
- Allow users to combine results from multiple jobs, perhaps across multiple GEDI collections.
- Allow user to choose alternative output formats (other than GeoPackage)
- Explore scaling horizontally rather than (or in addition to) scaling vertically
- Eliminate dependency on MAAP
- Query metadata by geometry rather than bounding box

Resources

- NASA ESDS MAAP
- MAAP Project
- NASA Common Metadata Repository (CMR)
- Global Ecosystem Dynamics Investigation (GEDI)