Modern open source solutions for HPC and open science

June 24th, 2024

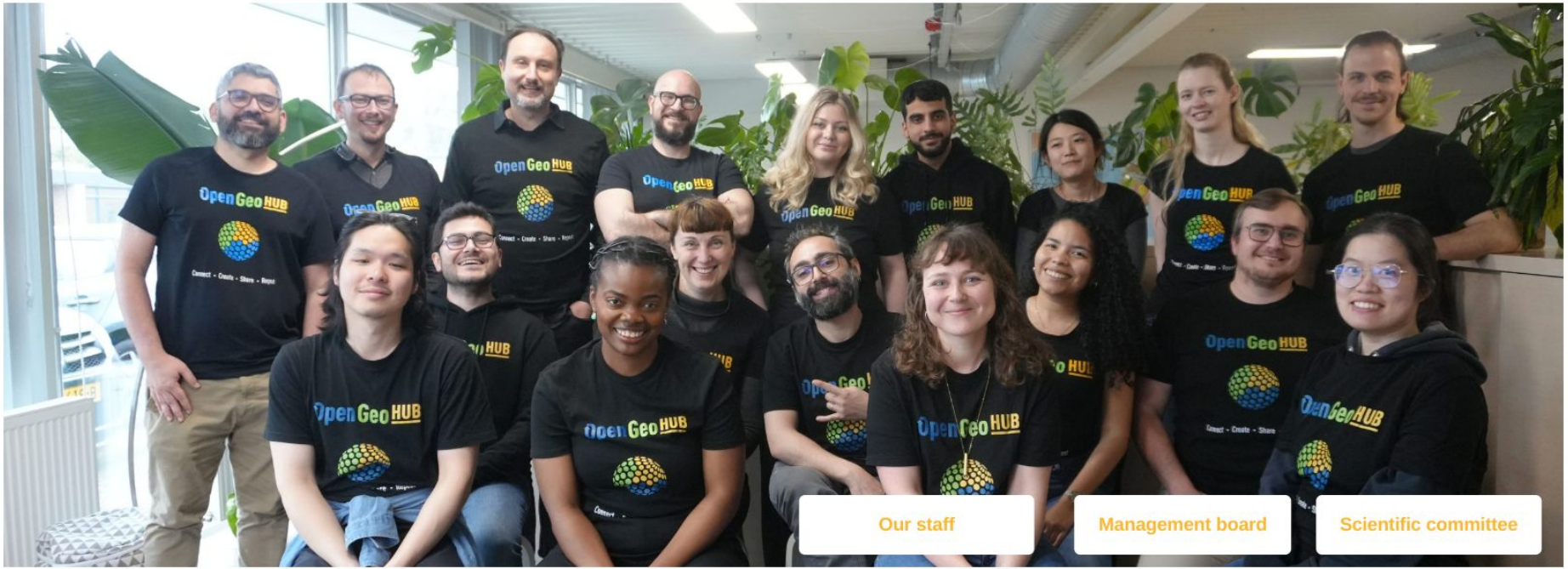
tom.hengl@opengeohub.org

https://fosstodon.org/@tomhengl

thengl

https://opengeohub.org / https://envirometrix.net

Meet the team!



OpenGeoHub Foundation

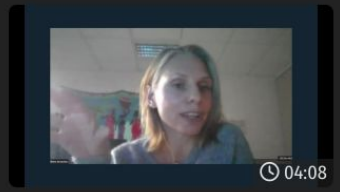
319 | 18 | 2018–2024 | 14,238 | 10 days 10 hours

Open this publisher as search result | Share publisher

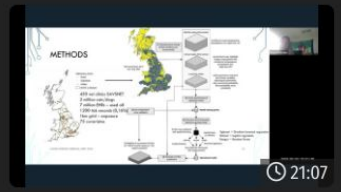
Newest Videos

Show all 319 results

Interview with William Wint



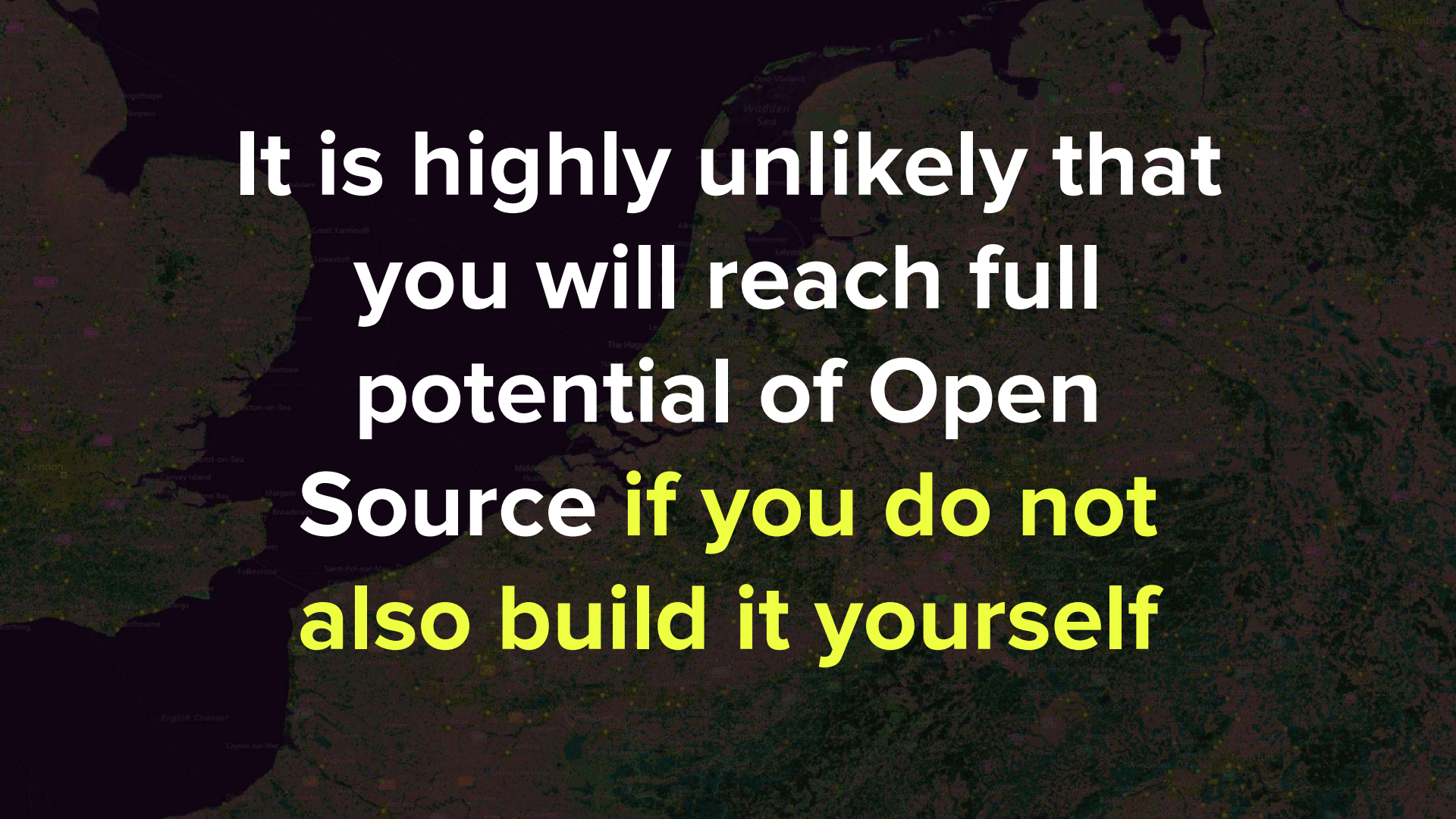
Interview with Elena Arsevska



Risk factors for tick attachment in companion animals in Great Britain: a spatiotemporal analysi...

The SERVIR program, from complex data to actionable information

Unveiling the location of cocoa farms across the pan-tropics


It is highly unlikely that
you will reach full
potential of Open
Source if you do not
also build it yourself


OpenGeo HUB
Connect • Create • Share • Repeat

https://EarthMonitor.org

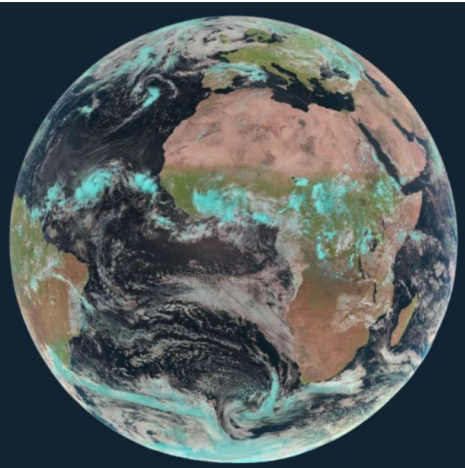
Funded by the European Union

OPEN EARTH MONITOR

Home 01 About 02 Events 03

A cyberinfrastructure to accelerate uptake of environmental information and help build user communities at European and global levels

Open-Earth-Monitor



Light Dark

Follow Us

The mission of Open-Earth-Monitor is to accelerate uptake of environmental information to guide current and future users in research, decision-making and citizens toward the most sustainable solutions.

European Green Deal

European Commission

Destination Earth initiative

European Data Science Cloud

EUROPEAN OPEN SCIENCE CLOUD

opernicus Europe's eyes on Earth

JRC esa

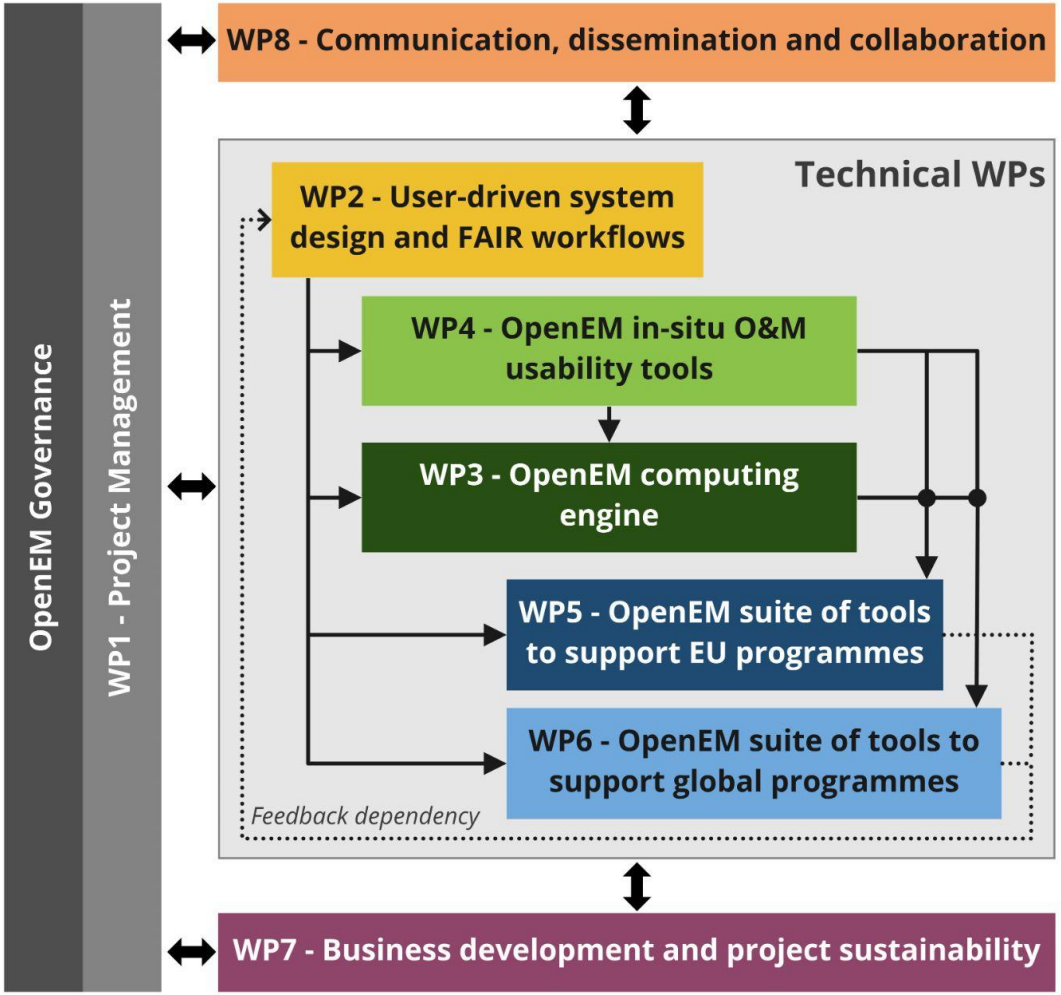
United Nations Convention to Combat Desertification

Food and Agriculture Organization of the United Nations

GEO GROUP ON EARTH OBSERVATIONS

OSGeo

Stakeholder Committee



Open-Earth-Monitor project outputs

01	Data (project Outputs)	Published via https://zenodo.org/communities/oemc-project and the Copernicus Data Space Ecosystem (CDSE)
02	Code (Tier 1)	https://github.com/Open-Earth-Monitor
03	Most impactful data from WP4, 5, and 6	A selection of most impactful data directly serving the use-cases: https://App.EarthMonitor.org
04	Global datasets	https://earthmonitor.org/knowledge-hub/; global data sets will also be reviewed via https://openlandmap.github.io/book/ and similar
05	Peer-reviewed articles	https://earthmonitor.org/scientific-publications/

https://medium.com/@opengeohub

Workshop 2023: key takeaways and what to expect in 2024

Prepared by: Tom Hengl (OpenGeoHub), Gilberto Camara (OEMC project Stakeholder Committee lead), Leandro Parente...

Dec 1, 2023

Pinned

OpenGeoHub

AI technology: what it is and what it's not, and how it can (potentially) help us solve the climate...

Prepared by: Tom Hengl (OpenGeoHub), Davide Consoli (OpenGeoHub), Marina Bagić (FER), Luca Brocca (CNR) and...

Aug 25, 2023 17

Pinned

OpenGeoHub in Nerd For Tech

Earth Observation and Machine Learning as the key technologies to track implementation of the Green...

Prepared by: Tom Hengl (OpenGeoHub), Carson Ross (OpenGeoHub) and Valentina Delconte (OpenGeoHub)

OpenGeoHub

310 Followers

Not-for-profit research foundation that promotes open geographical and geo-scientific data and develops open source software.

Edit profile

Following

- Gamze Ç.
- Samie Dorgham
- she++
- Towards Data Science
- OneSoil

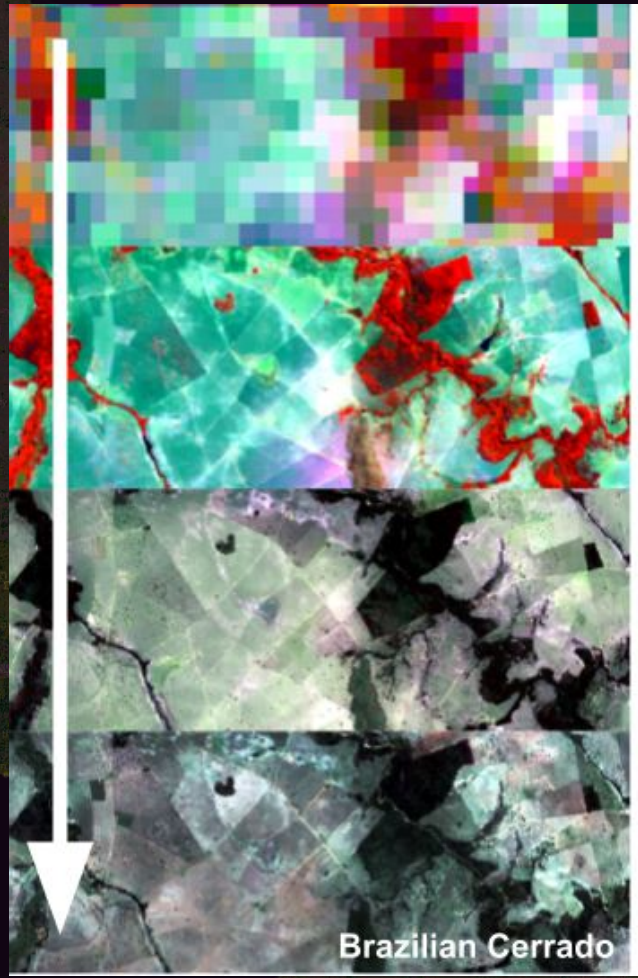
See all (7)

Help Status About Careers Press Blog Privacy Terms Text to speech Teams

Some key takeaways

#6: The EO industry has limited interest in sharing analysis-ready data — is this the business model that we need?

Today many groups unknowingly run more or less the same analysis over and over again, so that the industry can profit *N* times even where the clients are from the same organization. Is it efficient, is it even moral, to have multiple groups pay exactly the same processing to produce data used for land restoration and nature conservation projects?

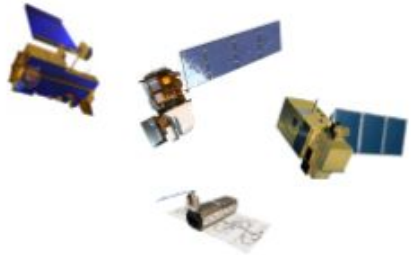


1x
MODIS
250m - *Spatial Res.*
2 days - *Temporal Res.*

64x
Landsat
30m - *Spatial Res.*
16 days - *Temporal Res.*

625x
Sentinel
10m - *Spatial Res.*
5 days - *Temporal Res.*

6,889x
PlanetScope
3m - *Spatial Res.*
1 day - *Temporal Res.*



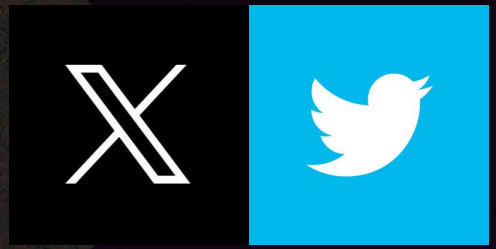
OpenGeoHUB
Connect · Create · Share · Repeat

Today we have multiple sources for Earth Observation data...

...however, most part of this data is not (real) analysis ready:

- Artifact and cloud free
- Gapfilled
- Fully accessible (COG and STAC)

The Problem With Satellite Data Is That It Is Not A Commodity



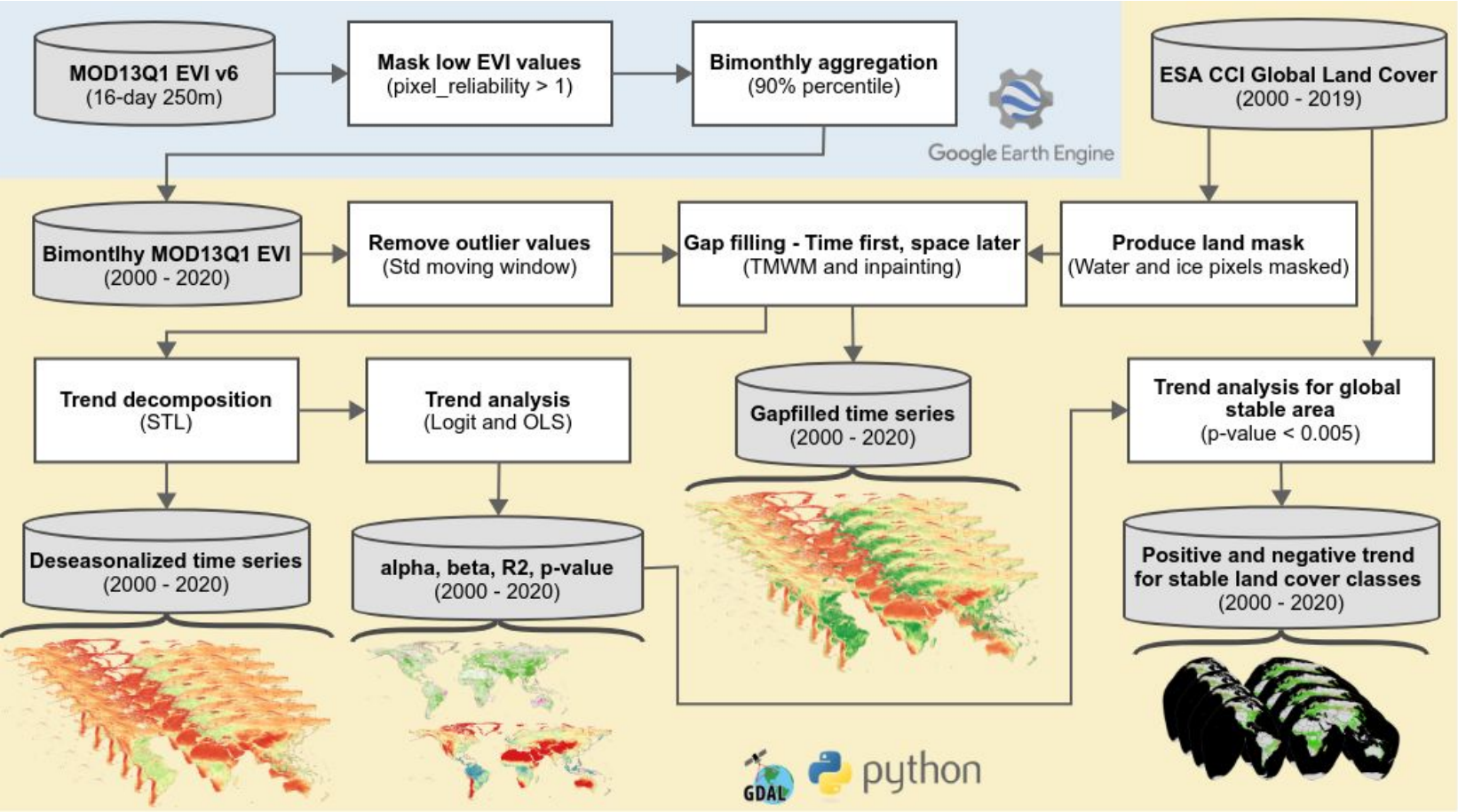
- Commercial model (currently based on WeChat?)
- Centralized;
- Content is property of Twitter?

VS

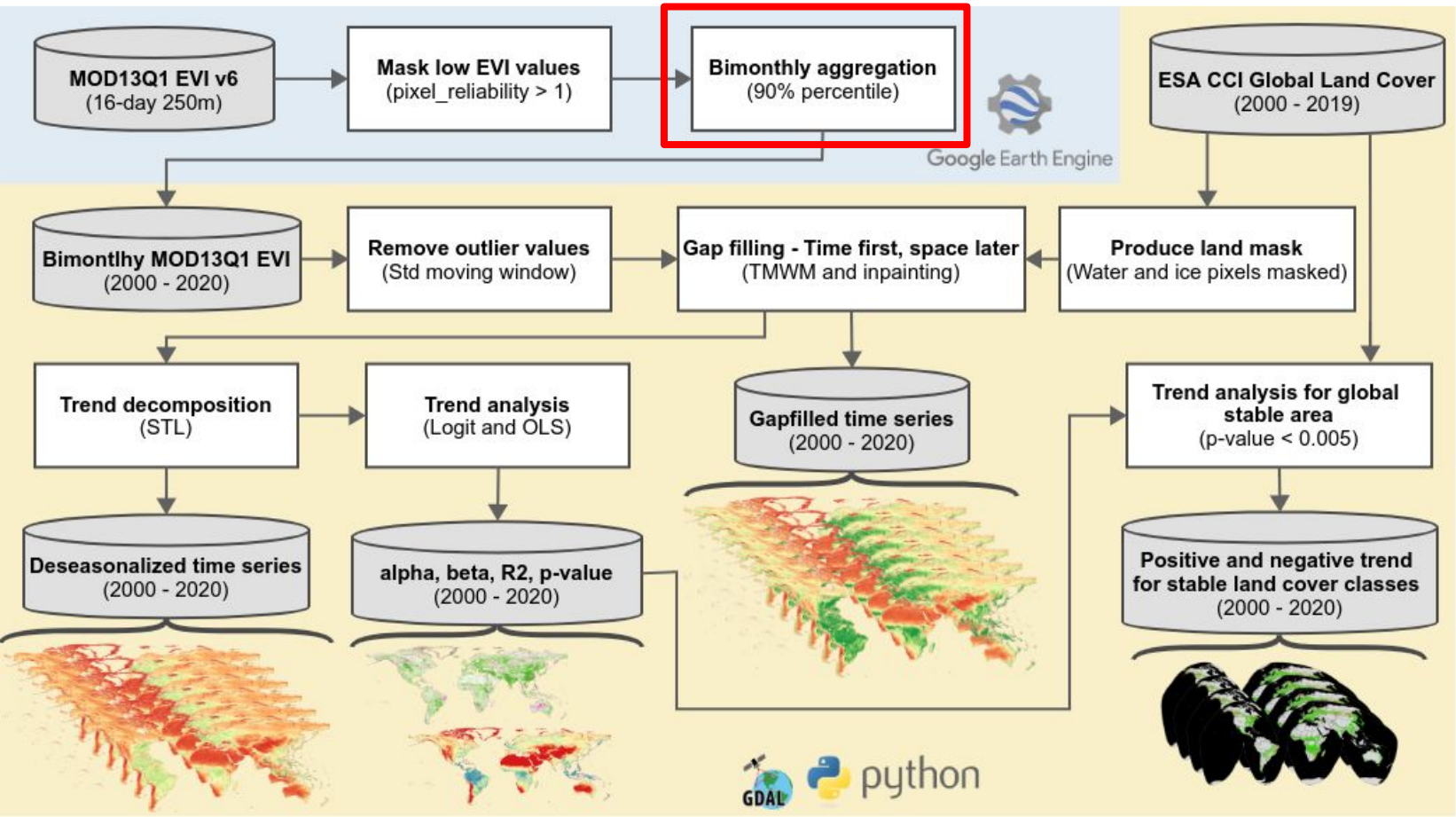
Fediverse; open source


HPC back in 2020
Global data and models at 250-m

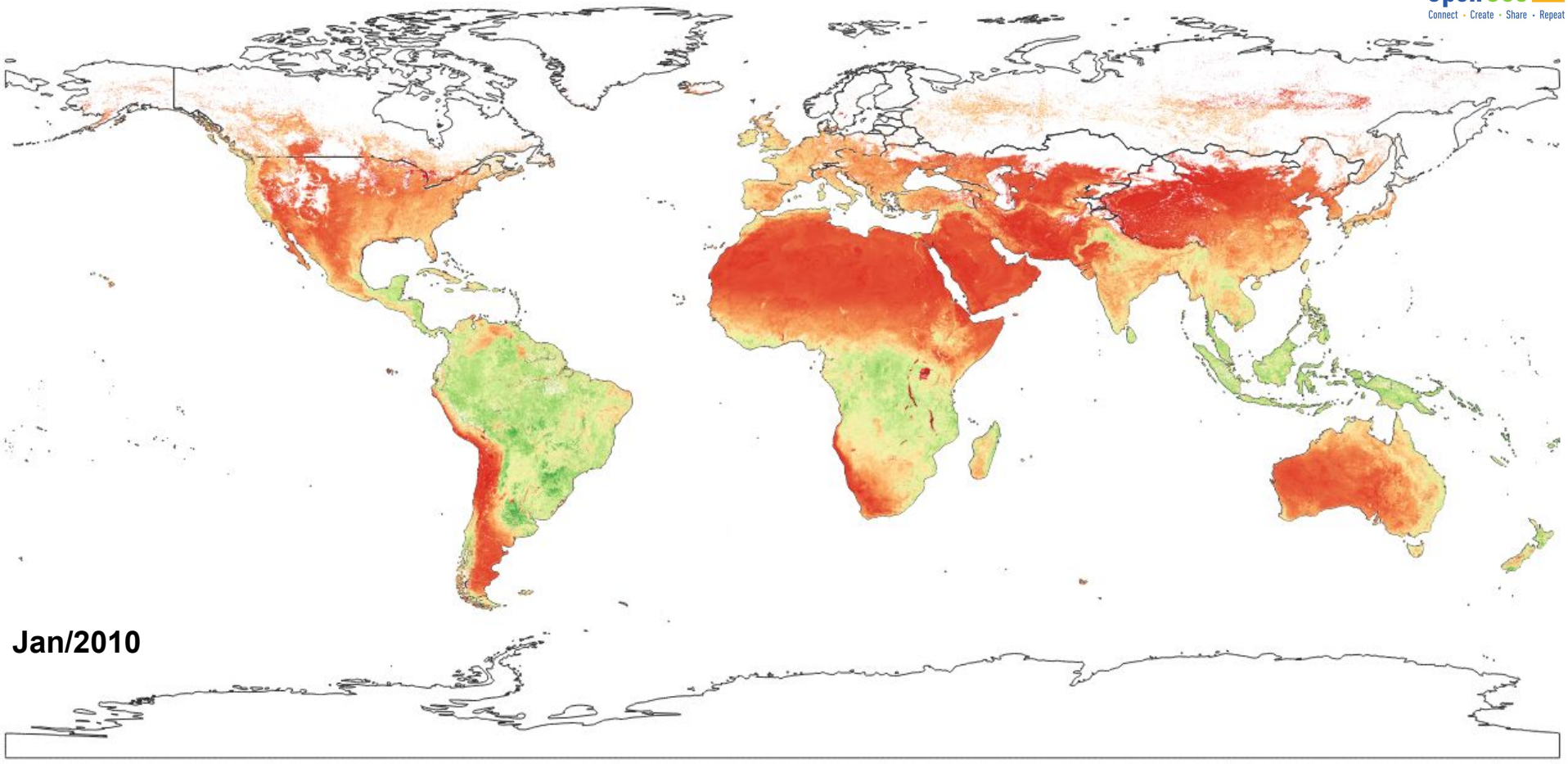
Crunching big EO data

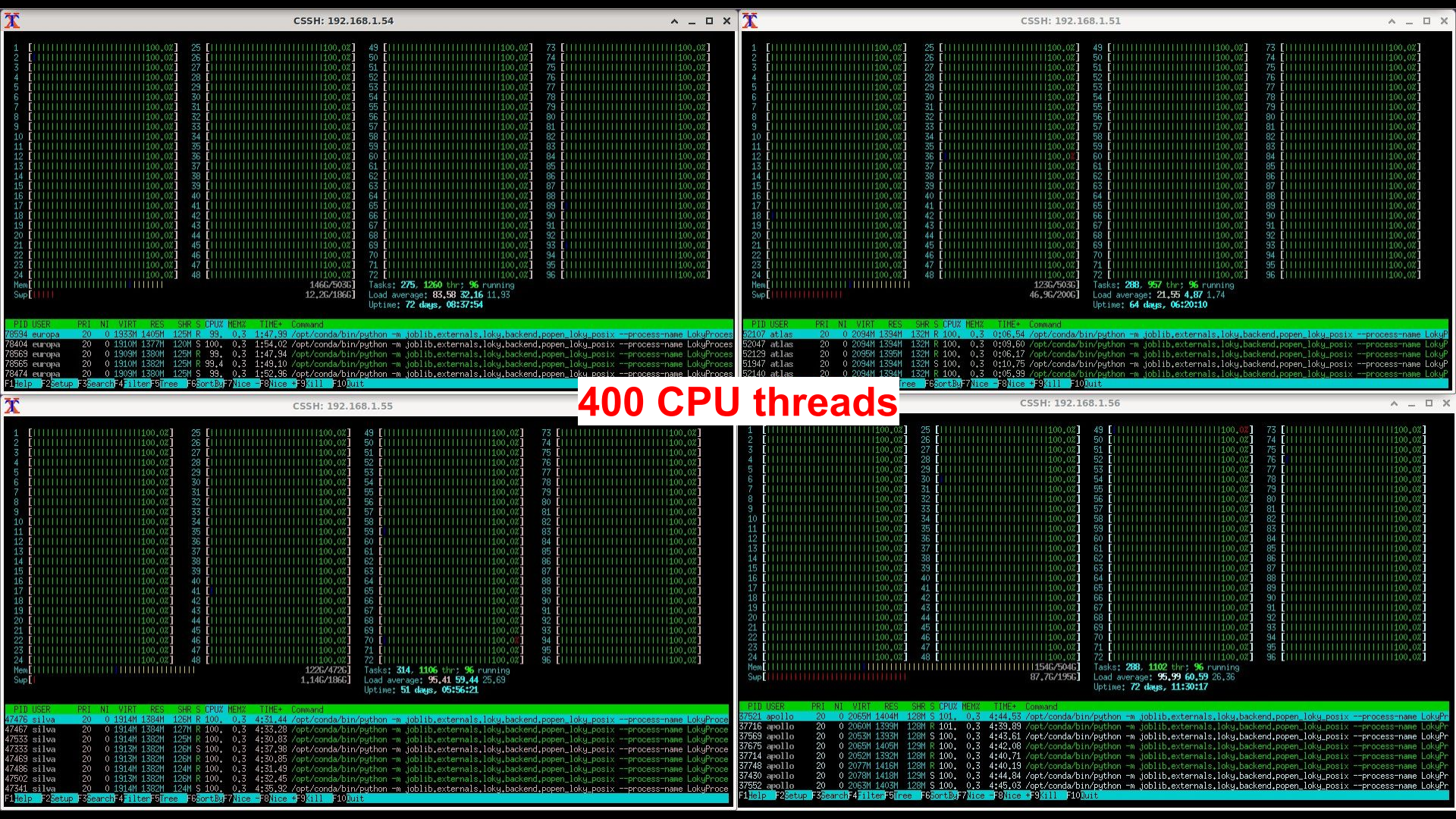


Crunching big EO data

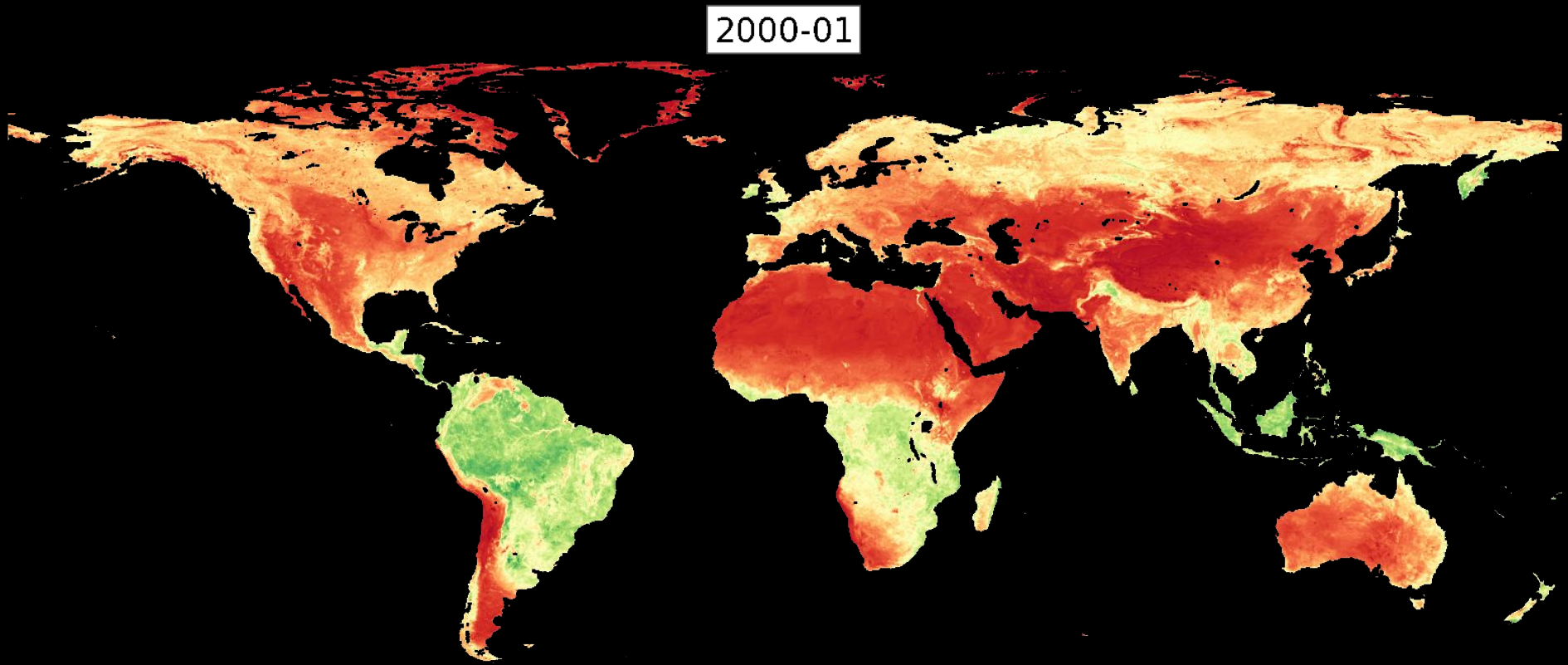


MOD13Q1 EVI — Aggregated (2 months)




CSSH: 192.168.1.54
CSSH: 192.168.1.51
CSSH: 192.168.1.55
CSSH: 192.168.1.56
400 CPU threads

MOD13Q1 EVI — Aggregated (2 months) and gap-filled



0 0.8

https://stac.openlandmap.org/

126 dates **x 160,300** columns **x 65,200** rows

Land potential assessment and trend-analysis using 2000–2021 FAPAR monthly time-series at 250 m spatial resolution

PeerJ Life & Environment

Related research

Share

Research article | Ecosystem Science | Data Mining and Machine Learning | Data Science | Environmental Impacts | Spatial and Geographic Information Science

Julia Hackländer[1,2], Leandro Parente[1], Yu-Feng Ho[1], Tomislav Hengl[1], Rolf Simoes[1], Davide Consoli[1], Murat Şahin[1], Xuemeng Tian[1,2], Martin Jung[3], Martin Herold[2,4], Gregory Duveiller[5], Melanie Weynants[5], Ichsani Wheeler[1] Post to Authors on X

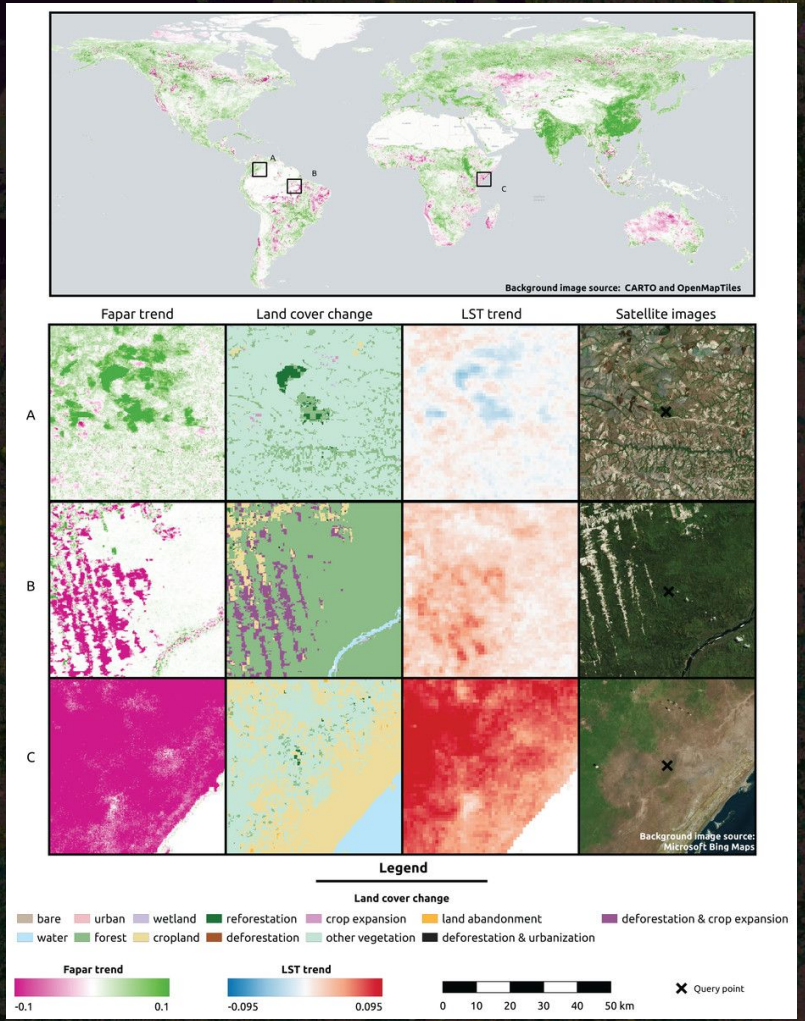
Published March 13, 2024

Read the peer review reports

Author and article information

Abstract

The article presents results of using remote sensing images and machine learning to map and assess land potential based on time-series of potential Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) composites. Land potential here refers to the potential vegetation productivity in the hypothetical absence of short–term anthropogenic influence, such as intensive agriculture and urbanization. Knowledge on this



HPC on steroids

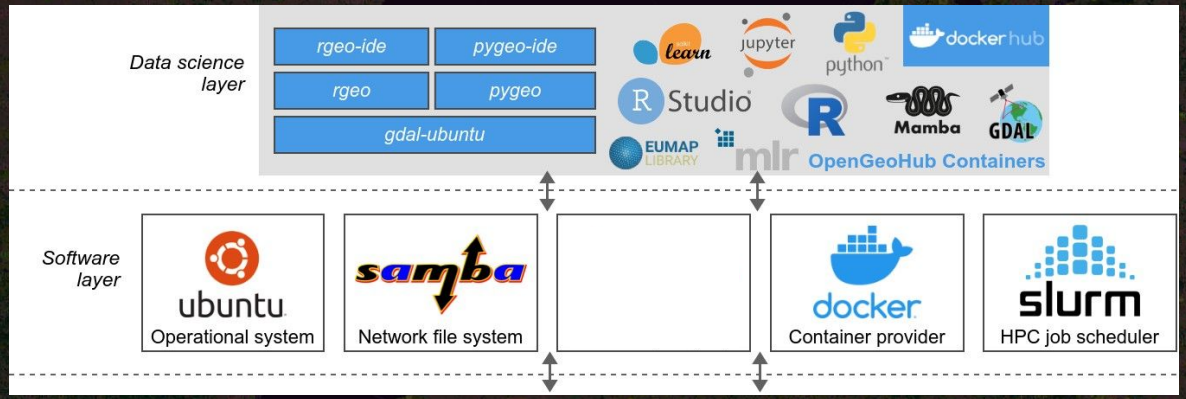
Processing the global Landsat archive.

Landsat ARD-2 — 16-days composites

Global Land Analysis & Discovery

26 years (1997–2022) x 23 composites (16–day each) => **598 images**

We looked at multiple options...



Options:

ceph

Garage

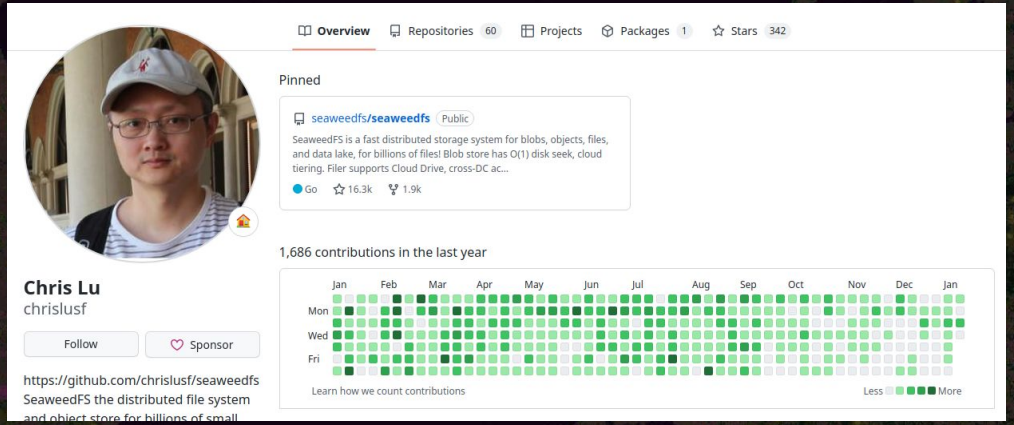
PROXMOX + ceph

GlusterFS

Seaweed

Main requirements: Storage / pool expansion without the need of data rebalancing

The github “life”



https://github.com/chrislusf

https://www.patreon.com/seaweedfs

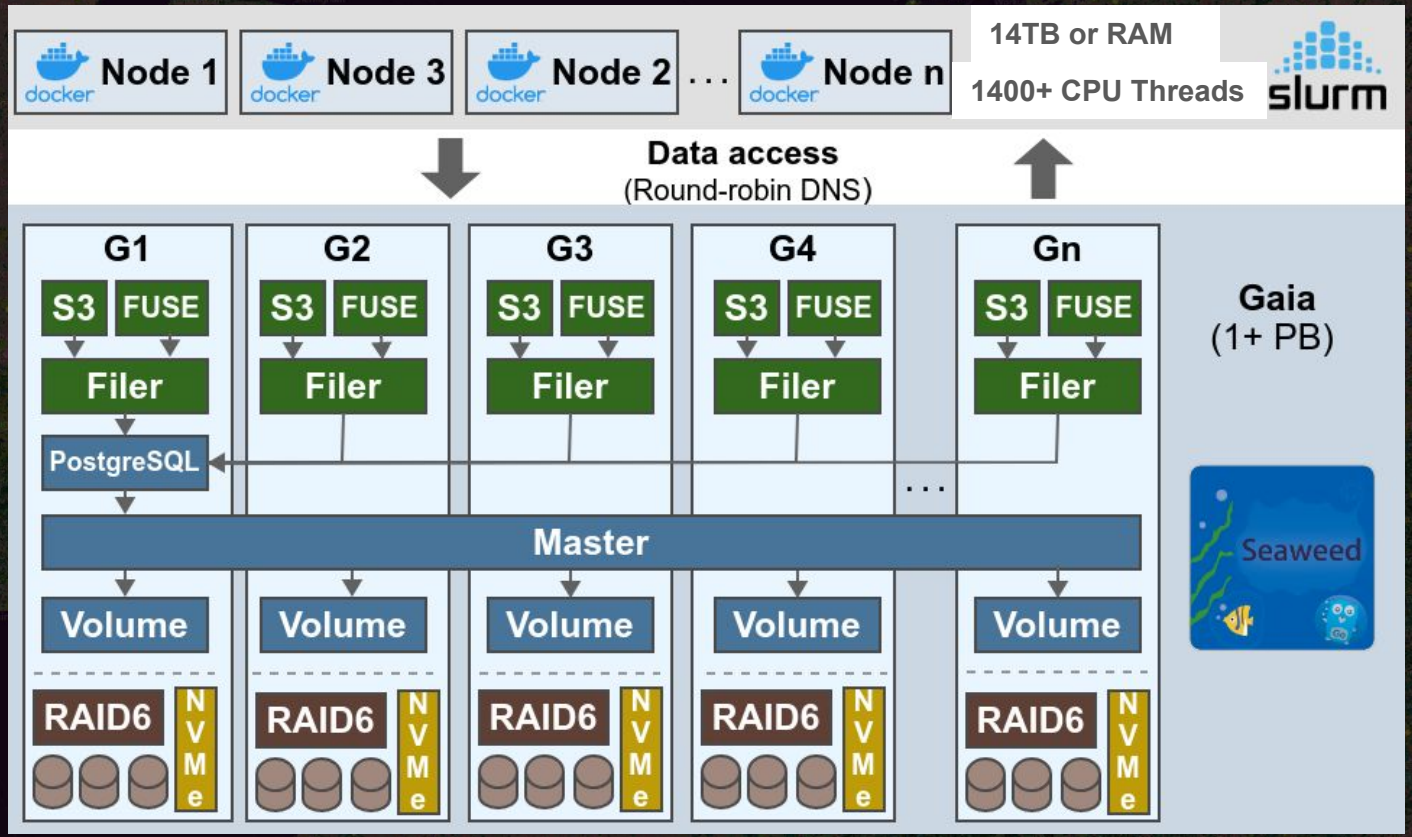
About

SeaweedFS is a fast distributed storage system for blobs, objects, files, and data lake, for billions of files! Blob store has O(1) disk seek, cloud tiering. Filer supports Cloud Drive, cross-DC active-active replication, Kubernetes, POSIX FUSE mount, S3 API, S3 Gateway, Hadoop, WebDAV, encryption, Erasure Coding.

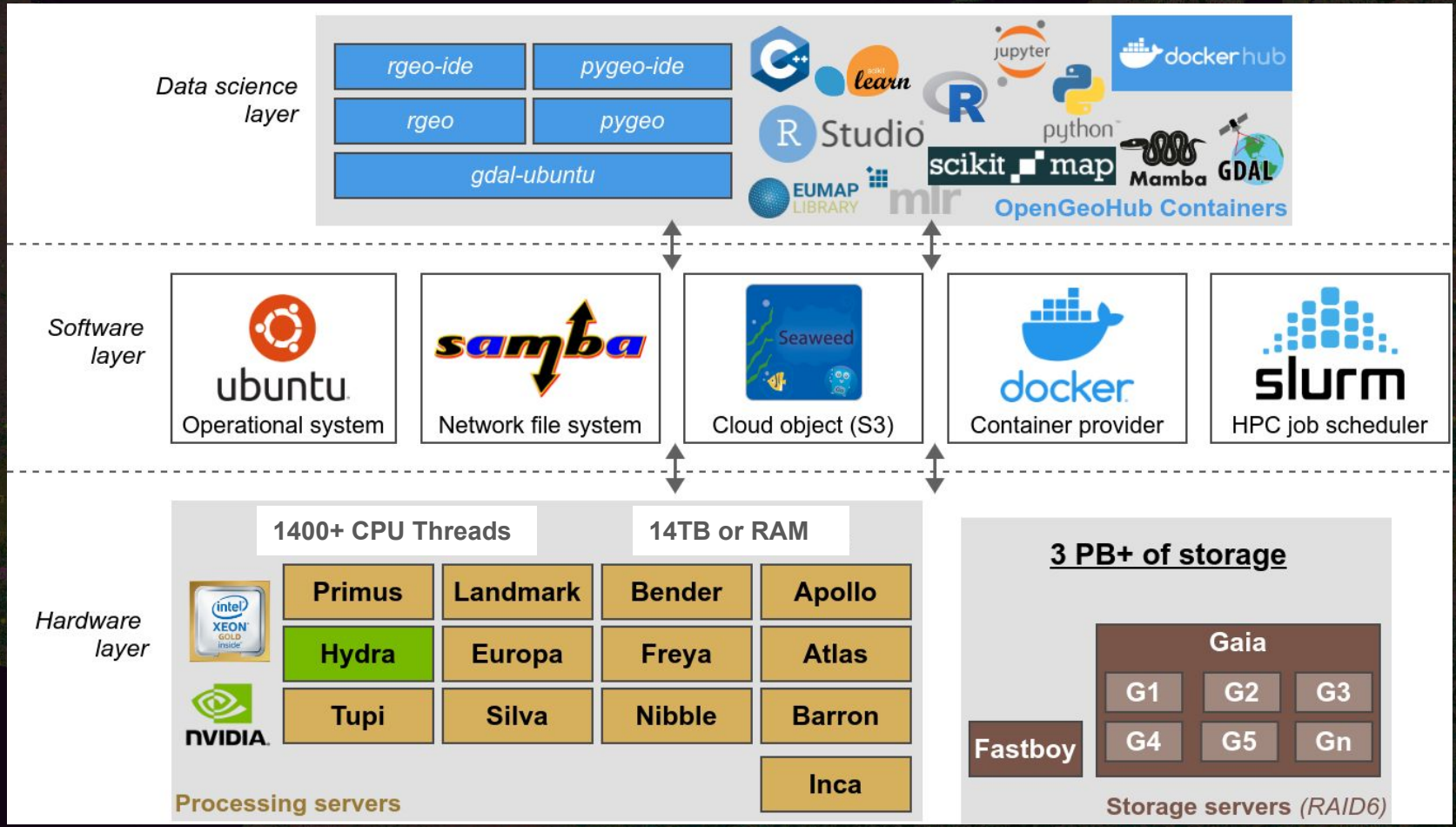
kubernetes distributed-systems fuse
replication cloud-drive s3 posix
s3-storage hdfs distributed-storage
distributed-file-system erasure-coding
object-storage blob-storage seaweedfs
hadoop-hdfs tiered-file-system

Readme
Apache-2.0 license
Code of conduct
16.3k stars
520 watching
1.9k forks

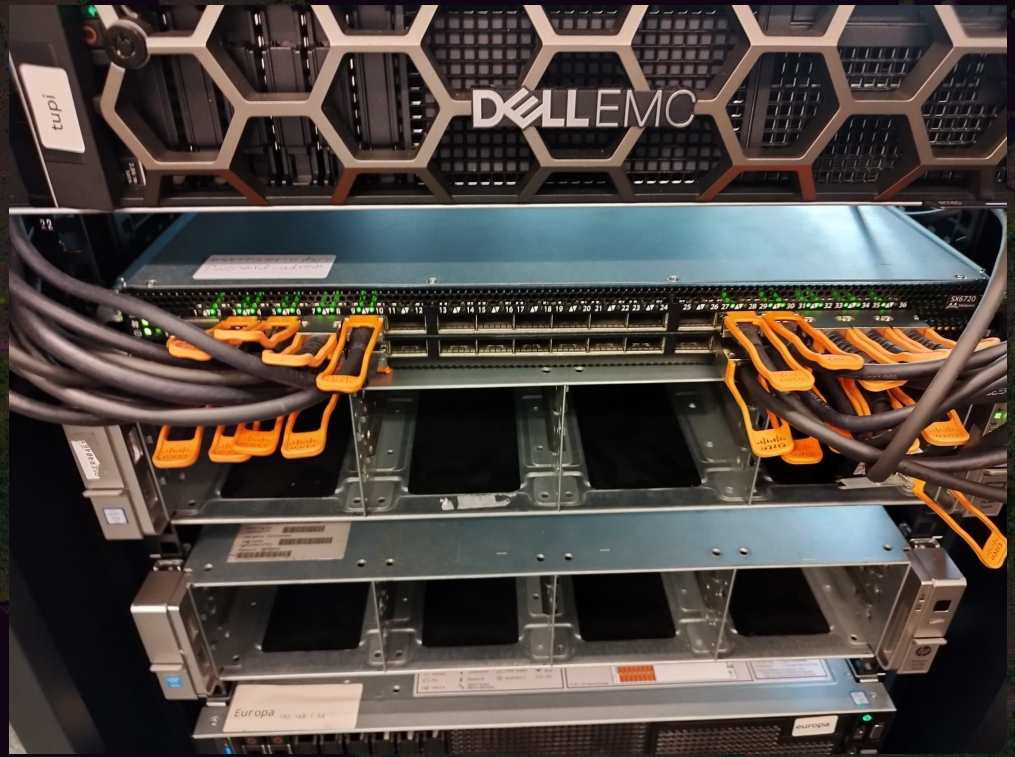
SeaweedFS Architecture



- **Load balancing** across all storage nodes (G1-n)
- S3 and file **metadata** stored in PostgreSQL
- **BLOB metadata** stored in NVMe
- **BLOB data** stored using RAID6 (HDD)
- If a storage node is **offline**, the cluster might become inconsistent


Data science layer
rgeo-ide
pygeo-ide
rgeo
pygeo
gdal-ubuntu
learn
jupyter
docker hub
R Studio
python
scikit map
Mamba
GDAL
EUMAP LIBRARY
mlr
OpenGeoHub Containers
Software layer
ubuntu
Operational system
samba
Network file system
Seaweed
Cloud object (S3)
docker
Container provider
slurm
HPC job scheduler
Hardware layer
1400+ CPU Threads
14TB or RAM
intel XEON GOLD inside
NVIDIA
Primus
Landmark
Bender
Apollo
Hydra
Europa
Freya
Atlas
Tupi
Silva
Nibble
Barron
Inca
Processing servers
3 PB+ of storage
Gaia
G1
G2
G3
G4
G5
Gn
Fastboy
Storage servers (RAID6)

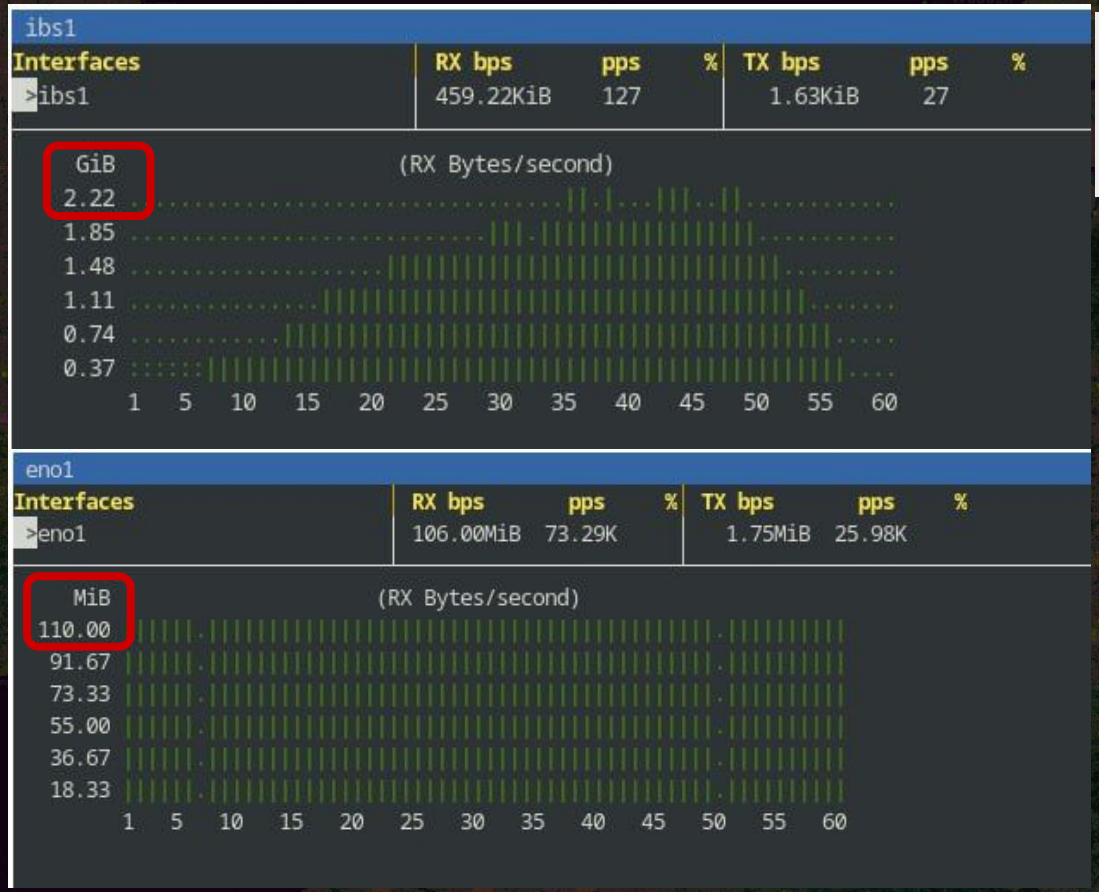
Infiniband (40 GBps)



1. Match the cable specifications with the Infiniband cards (ConnectX-3, ConnectX-3 Pro, ConnectX-5),
2. Install official Mellanox / NVIDIA driver in the Linux kernel 5.4.0-153,
3. Setup the switch and run a SM service to establish the IB connection,
4. Setup IP over Infiniband and HPC separated network (192.168.49.0/24),
5. Connect IB interface with the Docker containers.

```
root@g2:/home/ogh# iperf -c 192.168.49.35 -p 5002 -t 60 -P 10
------------------------------------------------------------
Client connecting to 192.168.49.35, TCP port 5002
TCP window size:  366 KByte (default)
------------------------------------------------------------
[ 15] local 192.168.49.31 port 40714 connected with 192.168.49.35 port 5002
[ 13] local 192.168.49.31 port 40716 connected with 192.168.49.35 port 5002
[  4] local 192.168.49.31 port 40642 connected with 192.168.49.35 port 5002
[  3] local 192.168.49.31 port 40630 connected with 192.168.49.35 port 5002
[  5] local 192.168.49.31 port 40688 connected with 192.168.49.35 port 5002
[ 10] local 192.168.49.31 port 40660 connected with 192.168.49.35 port 5002
[ 12] local 192.168.49.31 port 40676 connected with 192.168.49.35 port 5002
[  6] local 192.168.49.31 port 40662 connected with 192.168.49.35 port 5002
[  7] local 192.168.49.31 port 40656 connected with 192.168.49.35 port 5002
[  9] local 192.168.49.31 port 40700 connected with 192.168.49.35 port 5002
[ ID] Interval       Transfer     Bandwidth
[ 15]  0.0-60.0 sec  4.10 GBytes   587 Mbits/sec
[ 13]  0.0-60.0 sec  29.1 GBytes  4.17 Gbits/sec
[  4]  0.0-60.0 sec  29.3 GBytes  4.19 Gbits/sec
[  3]  0.0-60.0 sec  29.1 GBytes  4.17 Gbits/sec
[  5]  0.0-60.0 sec  29.3 GBytes  4.19 Gbits/sec
[ 10]  0.0-60.0 sec  15.5 GBytes  2.22 Gbits/sec
[ 12]  0.0-60.0 sec  18.5 GBytes  2.65 Gbits/sec
[  6]  0.0-60.0 sec  29.3 GBytes  4.19 Gbits/sec
[  7]  0.0-60.0 sec  25.2 GBytes  3.60 Gbits/sec
[  9]  0.0-60.0 sec  3.98 GBytes   570 Mbits/sec
[SUM]  0.0-60.0 sec   213 GBytes  30.5 Gbits/sec
```

Infiniband (40 GBps)




```
from scikit-map.raster import read_rasters

data, _ = read_rasters(raster_files=urls, n_jobs=len(urls), dtype='float32')
```

55 secs for reading 504 images of 4004 x 4004 => **8,080,136,064 pixels**

Consoli et al.

Research Square

Search preprints

Browse

Research Article

A computational framework for processing time-series of Earth Observation data based on discrete convolution: global-scale historical Landsat cloud-free aggregates at 30 m spatial resolution

Davide Consoli, Leandro Parente, Rolf Simoes, Murat Şahin, Xuemeng Tian, and 3 more

This is a preprint; it has not been peer reviewed by a journal.

https://doi.org/10.21203/rs.3.rs-4465582/v1

Abstract

Processing extremely large collections of Earth Observation (EO) time-series, often petabyte-sized, such as NASA's Landsat and ESA's Sentinel missions, can be computationally prohibitive and costly. Despite their name, even the Analysis Ready Data (ARD) versions of such collections can rarely be used as direct input for modeling and require additional time-series processing. Existing solutions for readily using these data are not openly available, are poor in performance, or lack flexibility. Addressing this issue, we developed SIRCLE (Signal Imputation and Refinement with Convolution Leaded Engine), a computational framework that can be used to apply diverse time-series processing techniques by simply adjusting the convolution kernel. Together with SIRCLE, this paper presents SWAG (Seasonally Weighted Average Generalization), a method for EO time-series reconstruction integrated in the

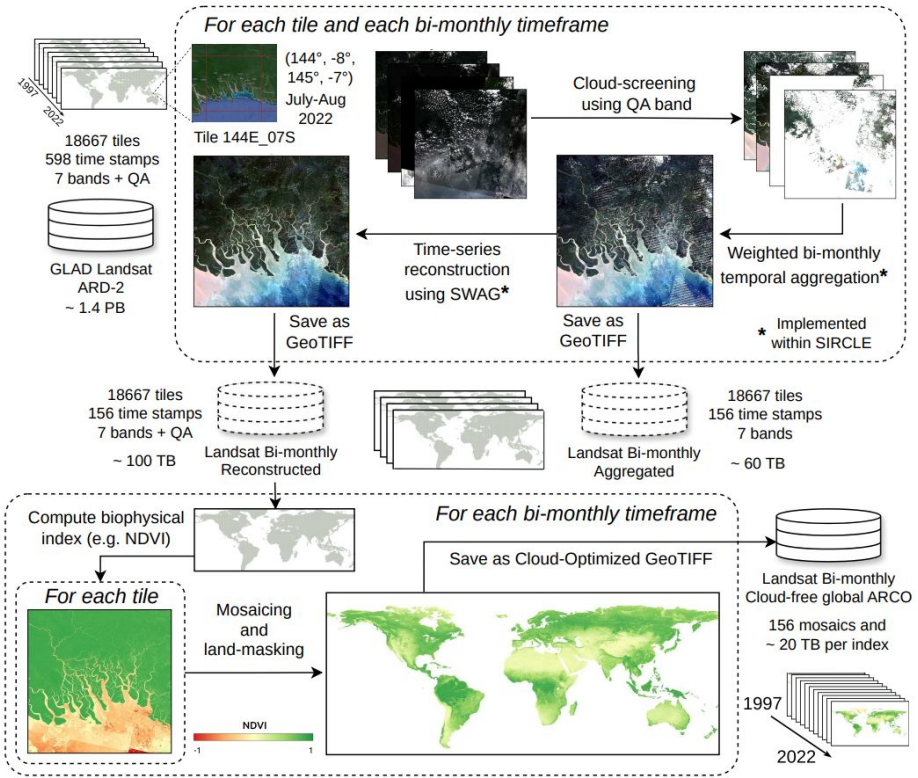
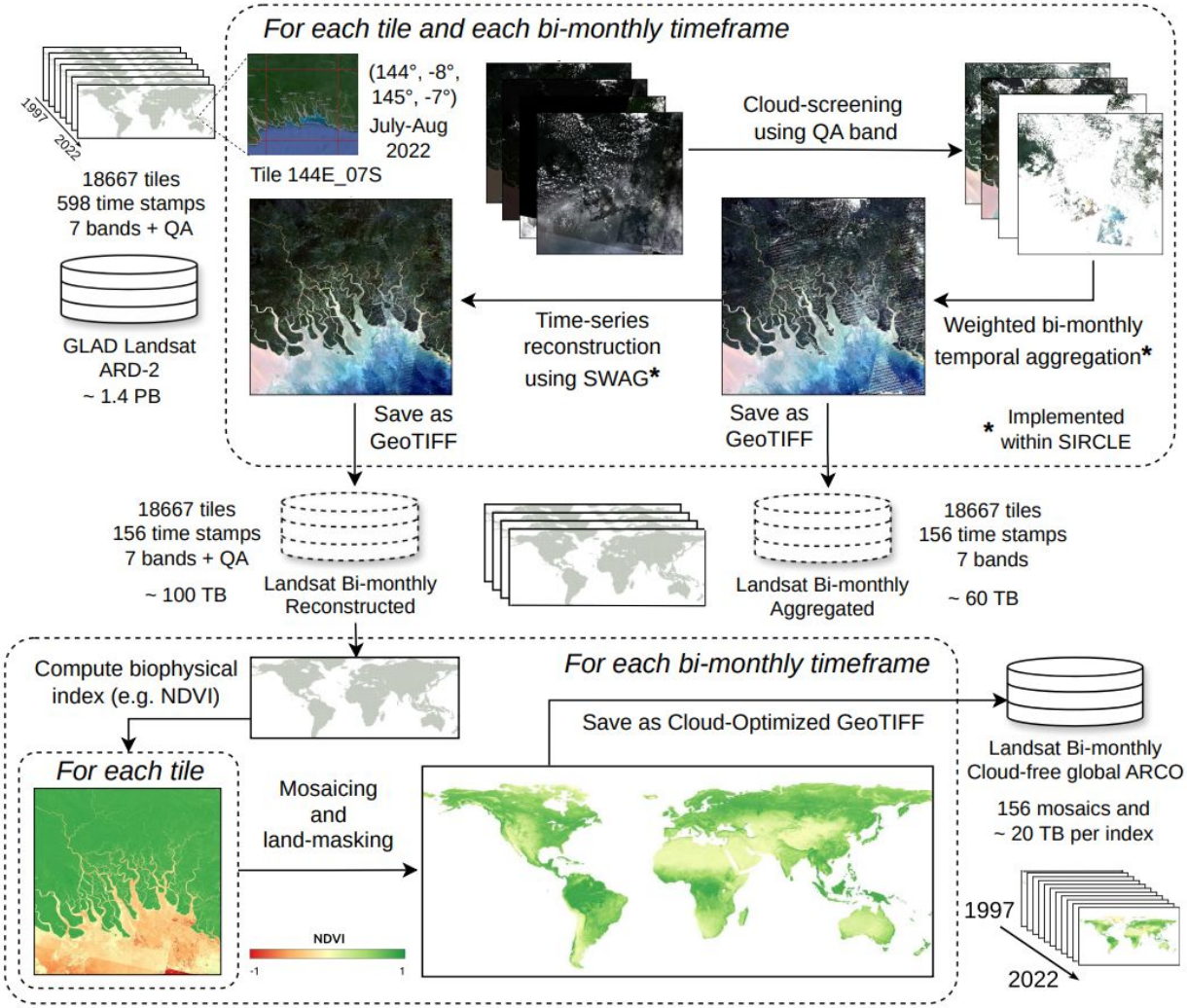


Figure 4. Block scheme of Landsat ARD-2 processing based on SIRCLE. In top left the input tiled dataset (7 bands + quaintly assessment, 30 m spatial resolution and 16-days time resolution). For each tile the whole time-series is sequence (i) cloud screened, (ii) time aggregated in bimonthly frames and (iii) reconstructed using SWAG. Time aggregation and SWAG are implemented within the SIRCLE framework, and both their result are saved in a S3 storage system. The Landsat bimonthly Reconstructed dataset is used as input to compute biophysical indices, like the normalized difference vegetation index (NDVI), land-masked and stored as global mosaiced and cloud optimize GeoTIFFs (COG) in a S3 storage system. Base map © Google Hybrid.


For each tile and each bi-monthly timeframe
1997
2022
18667 tiles
598 time stamps
7 bands + QA
GLAD Landsat
ARD-2
~ 1.4 PB
(144°, -8°, 145°, -7°)
July-Aug 2022
Tile 144E_07S
Cloud-screening using QA band
Time-series reconstruction using SWAG*
Weighted bi-monthly temporal aggregation*
Save as GeoTIFF
Save as GeoTIFF
* Implemented within SIRCLE
18667 tiles
156 time stamps
7 bands + QA
~ 100 TB
Landsat Bi-monthly Reconstructed
Landsat Bi-monthly Aggregated
18667 tiles
156 time stamps
7 bands
~ 60 TB
Compute biophysical index (e.g. NDVI)
For each bi-monthly timeframe
Save as Cloud-Optimized GeoTIFF
Landsat Bi-monthly Cloud-free global ARCO
156 mosaics and ~ 20 TB per index
For each tile
Mosaicing and land-masking
NDVI
-1
1
1997
2022

Parente et al.

Research Square

Search preprints

Browse

Data Note

Mapping global grassland dynamics 2000—2022 at 30m spatial resolution using spatiotemporal Machine Learning

Leandro Parente, Lindsey Sloat, Vinicius Mesquita, Davide Consoli, and 16 more

This is a preprint; it has not been peer reviewed by a journal.

https://doi.org/10.21203/rs.3.rs-4514820/v1

Abstract

The paper describes the production and evaluation of global grassland dynamics mapped annually for 2000-2022 at 30~m spatial resolution. The dataset showing the spatiotemporal distribution of cultivated and natural/semi-natural grassland classes was produced by using GLAD Landsat ARD-2 image archive, accompanied by climatic, landform and proximity covariates, spatiotemporal machine learning (per-class Random Forest) and over 2.3M reference samples (visually interpreted in Very High Resolution imagery). Custom probability thresholds (based on five-fold spatial cross-validation) were used to derive dominant class maps with balanced precision and recall values, 0.64 and 0.75 for cultivated and natural/semi-natural grassland, respectively. The produced maps (about 4~TB in size) are available under an open data license as Cloud-Optimized GeoTIFFs and as Google Earth Engine assets. The suggested uses of data include (1) integration with other compatible land cover products and (2) tracking the intensity and drivers of conversion of land to cultivated grasslands and from natural

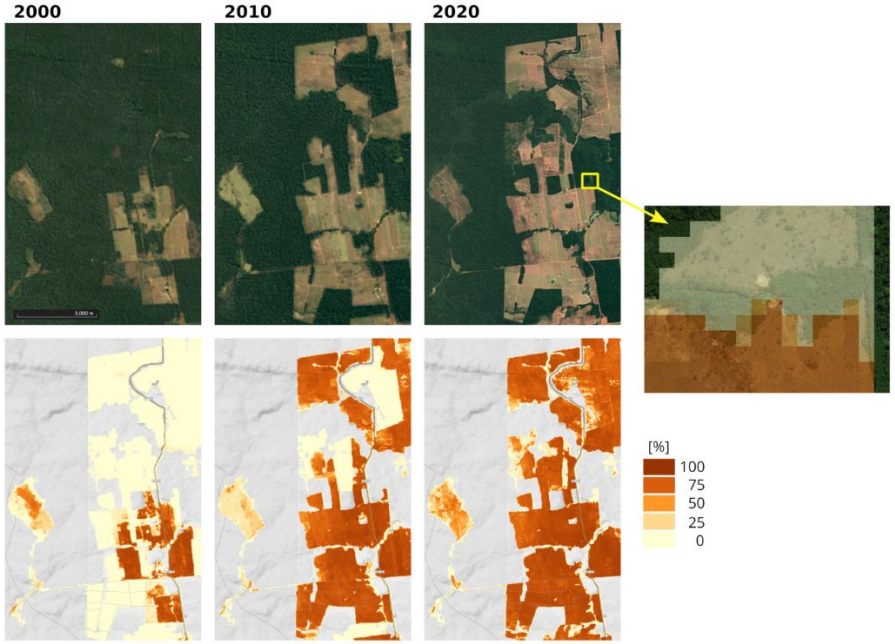
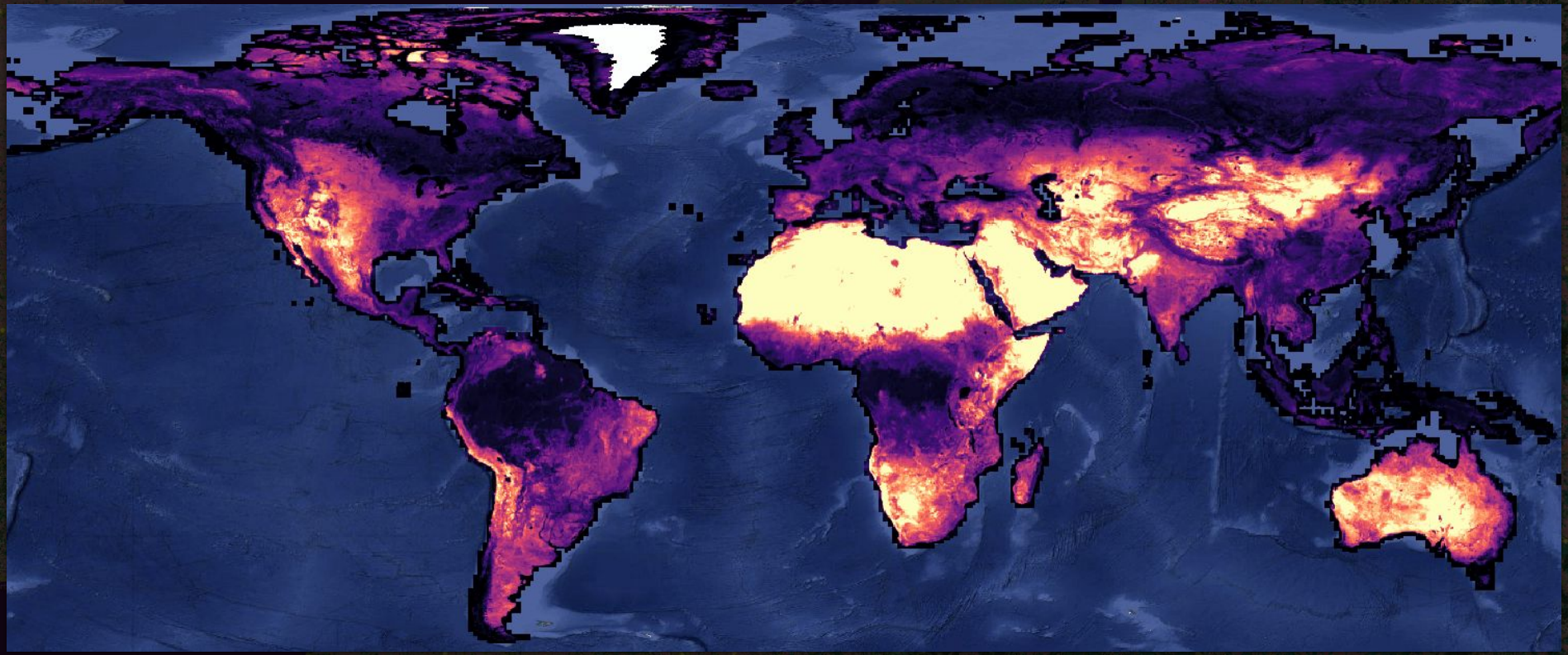
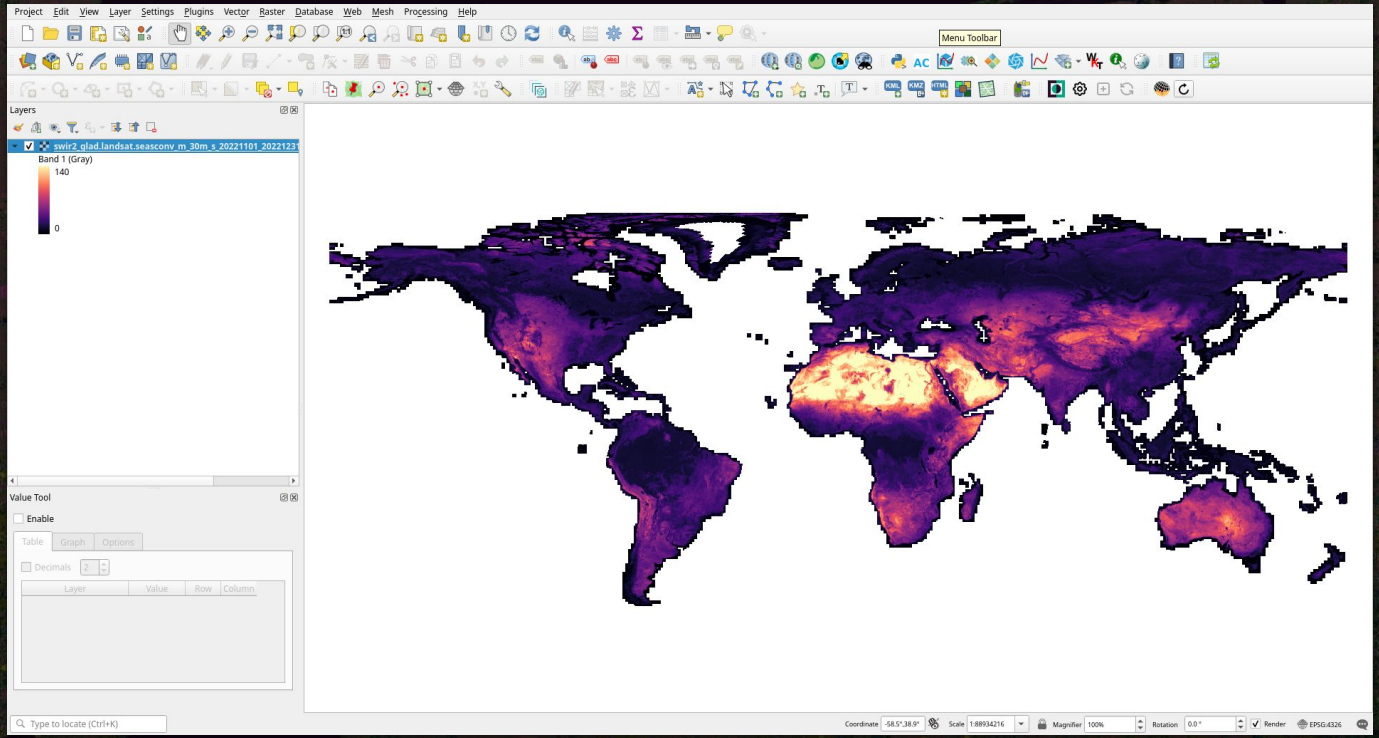


Figure 9. Our predictions of probabilities for cultivated grassland for 2000, 2010 and 2020 at 30 m spatial resolution (below) for an area in Brazil (close to Serra Morena) as compared to the Google Time lapse images (above); based on the AirbusMaxar Technologies high resolution images.


4C ARCO =
Complete Consistent
Current Correct





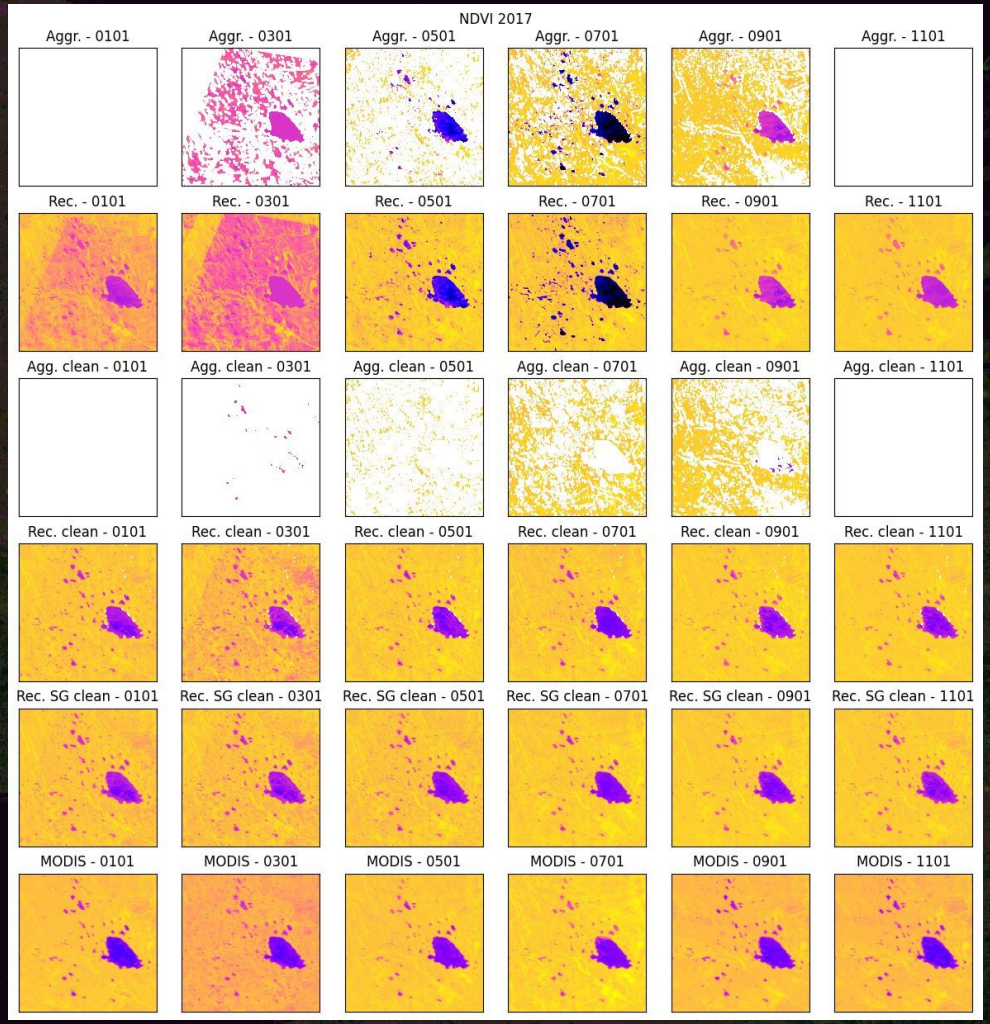
The biggest 2 bottlenecks of this project are:

(1) the **storage** problem (we need about 2PB of storage to host all open data)

(2) **sustainability** problem (we need to think of new commercial services post 2024) that could pay the production costs.

Dimensions:

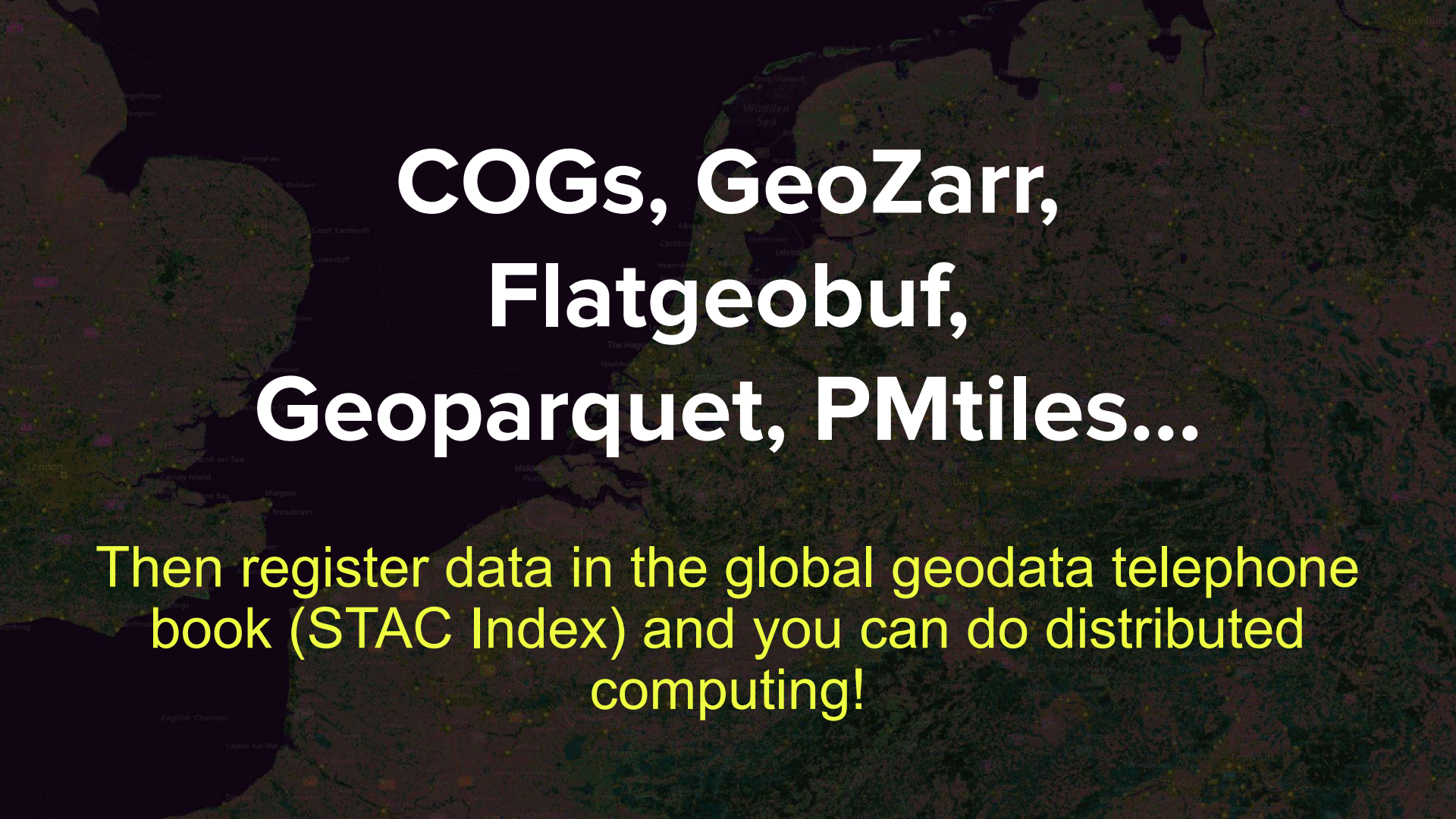
- Image size: 1,440,004 (H), 560,004 (V)
- Filesize: 134 GB (with compression)
- Format: Cloud-Optimized GeoTIFF (COG)



The main objective at the moment is to try to reconstruct the Landsat bands and 100% gap-filled them using ALL data available:

- MODIS monthly time-series (250-m) 2000–2023+ (MOD13Q1);
- Savitzky-Golay filter;

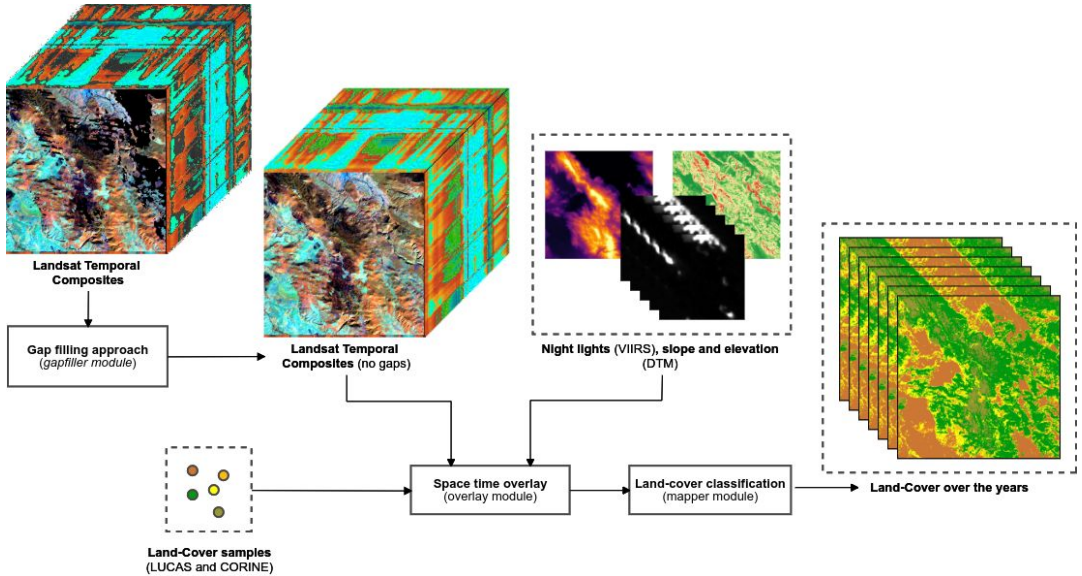
This way we could potentially reduce Landsat archive to max 300TB of data (but all 4C ARCO)


COGs, GeoZarr,
Flatgeobuf,
Geoparquet, PMtiles...
Then register data in the global geodata telephone book (STAC Index) and you can do distributed computing!

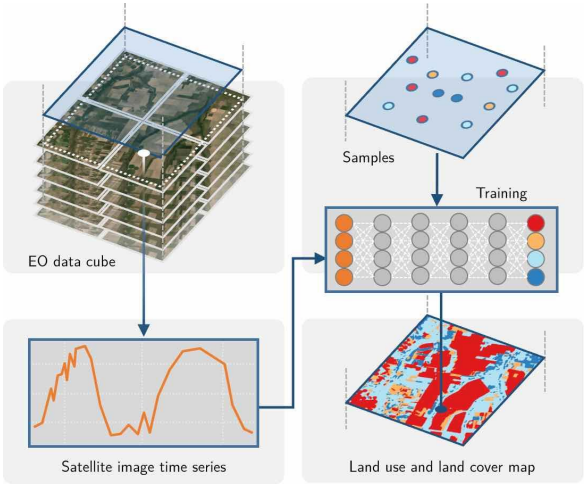
Integrating ARCO into ML

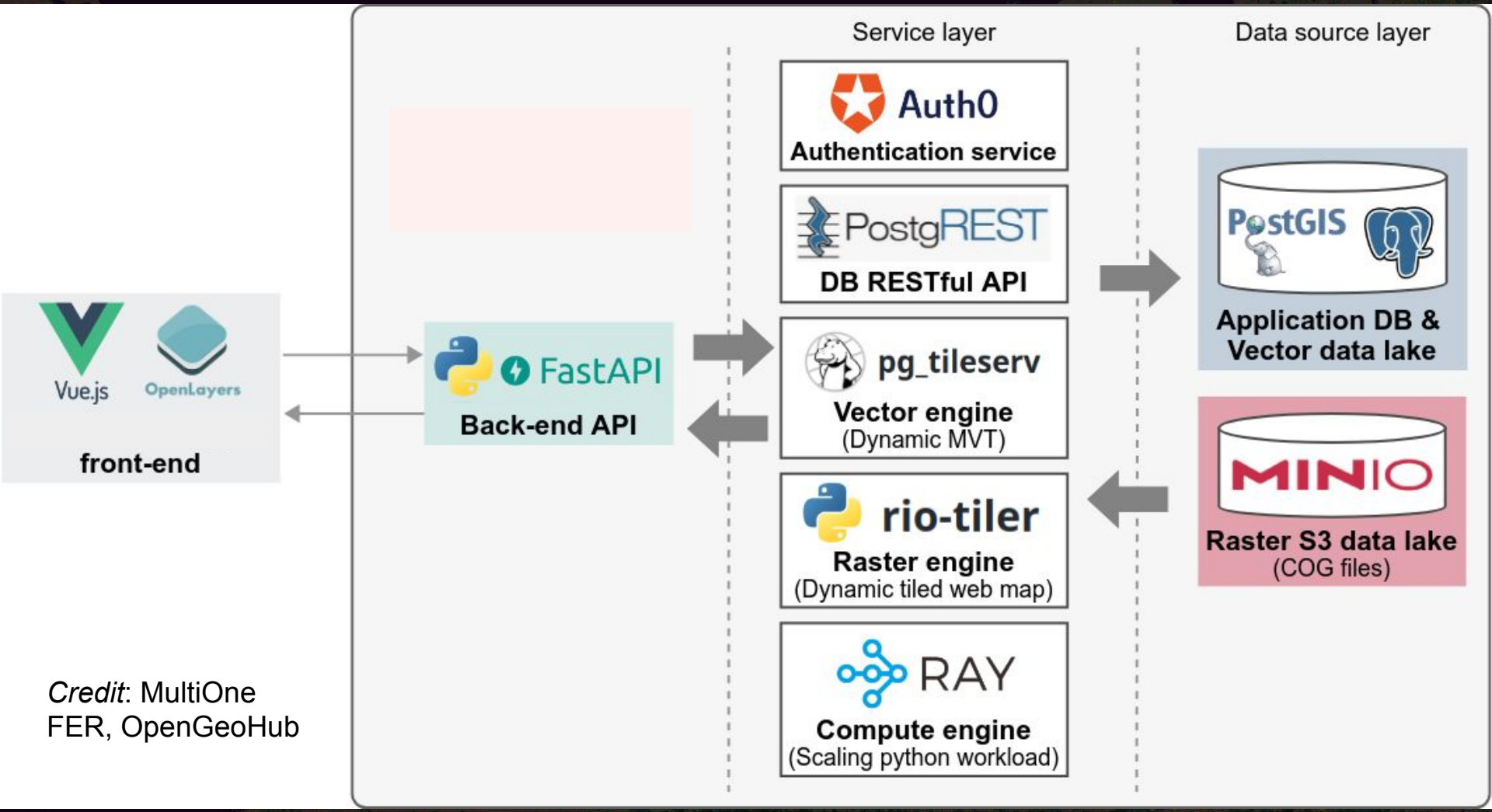
scikit map

https://github.com/openlandmap/scikit-map



https://github.com/e-sensing/sits



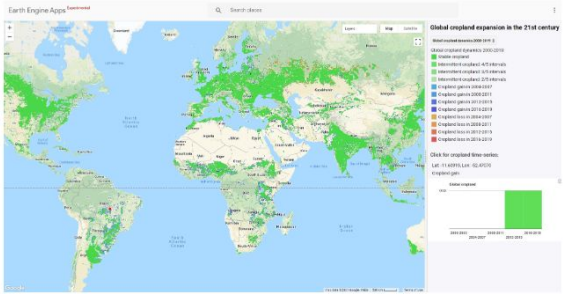


Credit: MultiOne
FER, OpenGeoHub

Dataset Reference

P. Potapov, S. Turubanova, M.C. Hansen, A. Tyukavina, V. Zalles, A. Khan, X.-P. Song, A. Pickens, Q. Shen, J. Cortez. (2021) Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. Nature Food. https://doi.org/10.1038/s43016-021-00429-z

Data visualization using Google Earth Engine Apps



https://glad.earthengine.app/view/global-cropland-dynamics

Dataset Access in Google Earth Engine

Image collection IDs for each cropland layer:

2003 "*users/potapovpeter/Global_cropland_2003*"
2007 "*users/potapovpeter/Global_cropland_2007*"
2011 "*users/potapovpeter/Global_cropland_2011*"
2015 "*users/potapovpeter/Global_cropland_2015*"
2019 "*users/potapovpeter/Global_cropland_2019*"

Data values: *0 – no croplands or no data; 1 – croplands.*

Data Download

Map data provided in the geographic coordinates using the WGS84 reference system.
Data format: 8-bit unsigned LZW-compressed GeoTiff. Pixel size is 0.00025 × 0.00025 degree (~30 m × 30 m at Equator). Data aggregated into quadrant mosaics.
Pixel values: *0 – no croplands or no data; 1 – croplands.*

Global_cropland_NE_2003.tif [1.8 GB]
Global_cropland_NE_2007.tif [1.8 GB]
Global_cropland_NE_2011.tif [1.8 GB]
Global_cropland_NE_2015.tif [1.8 GB]
Global_cropland_NE_2019.tif [1.8 GB]

Global_cropland_NW_2003.tif [0.9 GB]
Global_cropland_NW_2007.tif [0.9 GB]
Global_cropland_NW_2011.tif [0.9 GB]
Global_cropland_NW_2015.tif [0.9 GB]

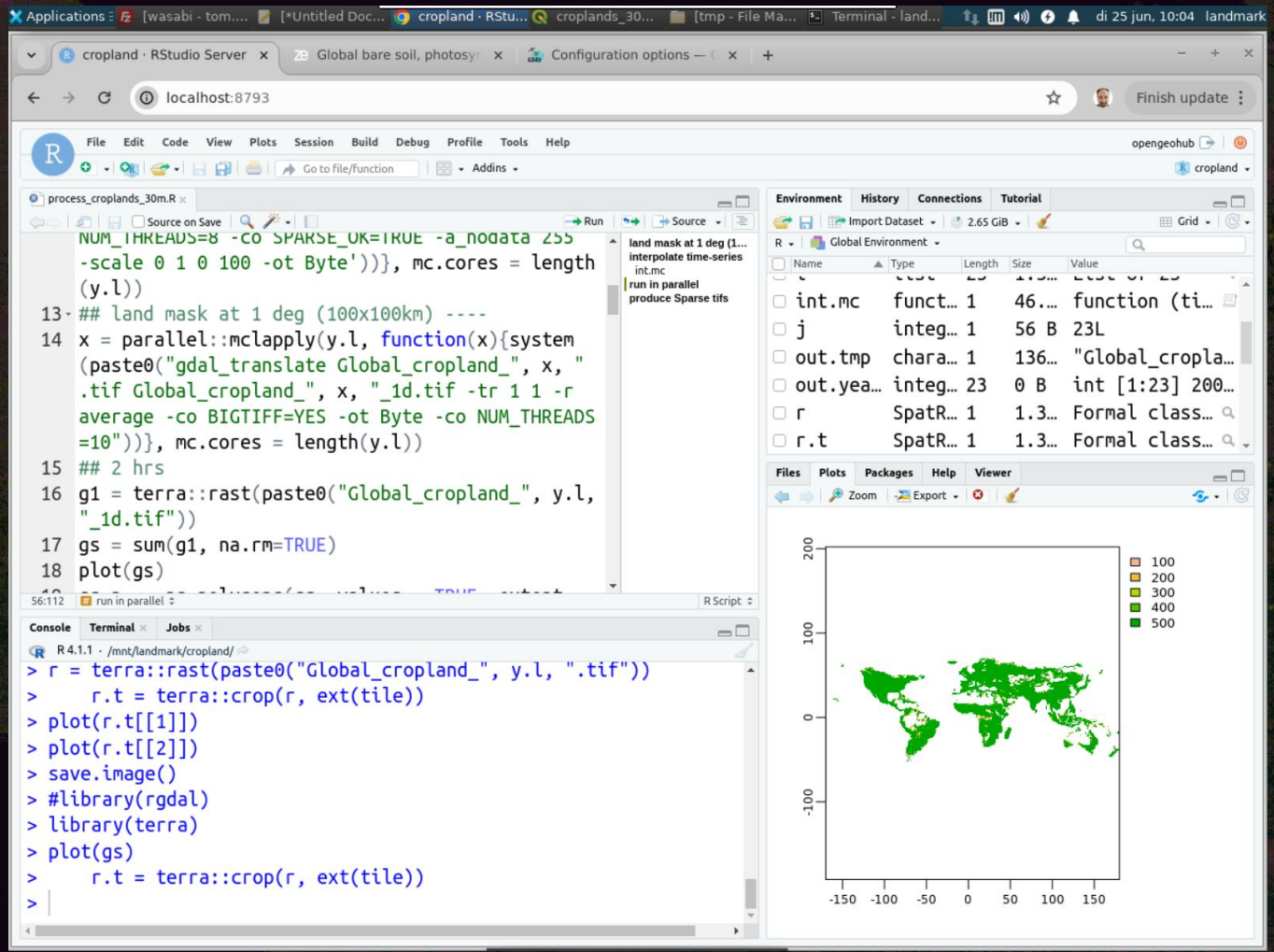
The data is available as TIFs split in blocks.

Only years 2003, 2007, 2015, 2019 are available.

Only occurrence pixels.

How to make this data more ARD?

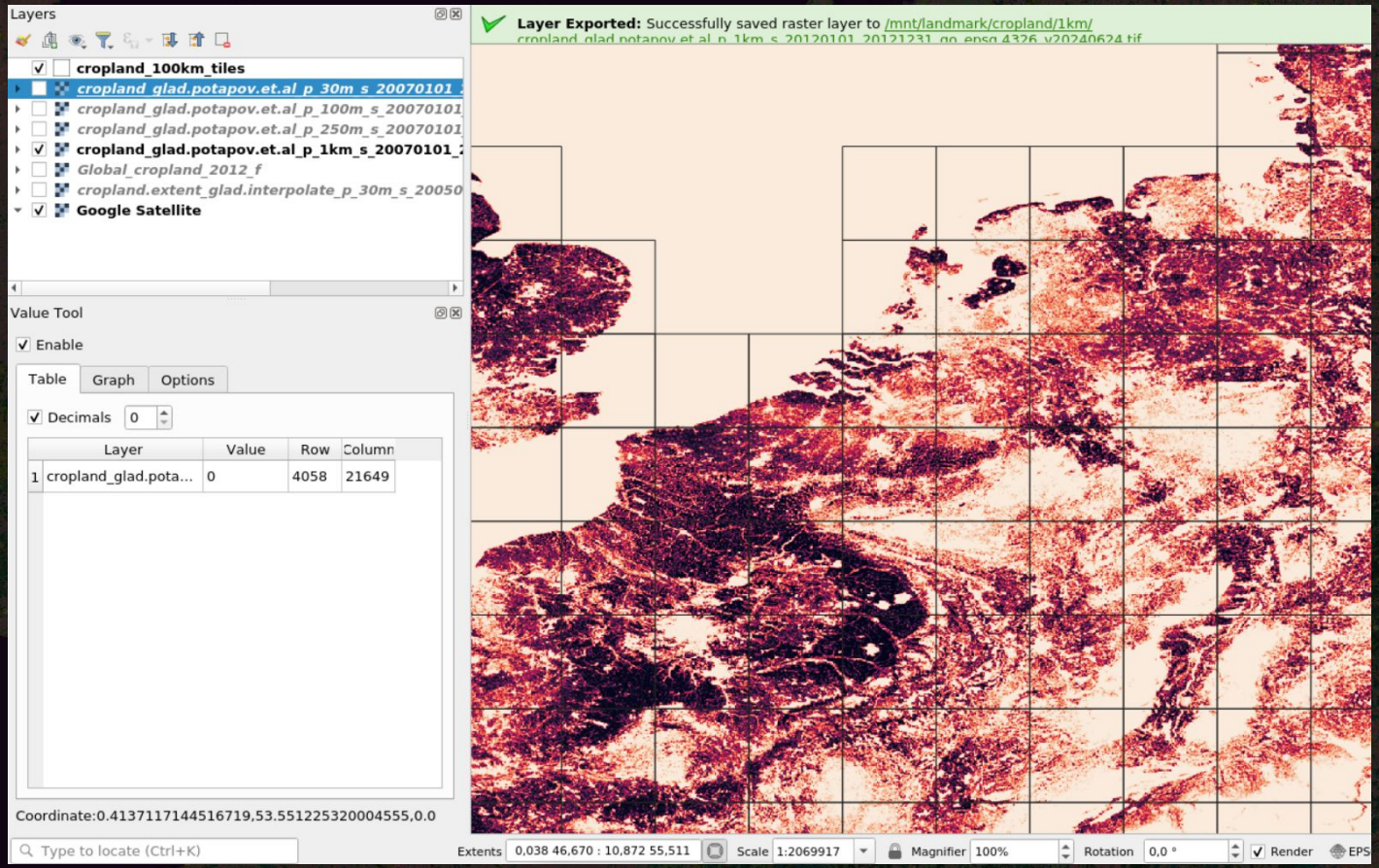
The terra package



Processing steps:

1. Prepare all data / prepare a tiling system (1deg blocks; 10,600 tiles with values)
2. Prepare and test a function to fit a temporal interpolation (approx function)
3. Run in parallel on 10,600 tiles.
4. Build global mosaics.

Outputs at 30m, 100m, 250m and 1km



https://rspatial.github.io/terra/reference/terra-package.html

The terra package introduces the following new classes:

- **SpatRaster**: for raster data and limits memory usage in comparison to the raster package data models
- **SpatVector**: represent vector-based point, line, or polygon features and their associated attributes.
- **SpatExtent**: for spatial extent information derived from a SpatRaster or SpatVector or manually defined.

The value of data is in its use

If you plan to profit from selling basic data, this might be the worst case scenario.
If you make data that is used with passion and with happy customers providing feedback, you might have a chance!

REGISTRATIONS

Sign up to attend!

Open-Earth-Monitor
GLOBAL
WORKSHOP
2024

https://earthmonitor.org/global-workshop-2024/#register-here

https://landcarbonlab.org

Global Pasture Watch

Mapping & monitoring Global Grasslands and Livestock

WORLD RESOURCES INSTITUTE

Land & Carbon Lab

LAPIG Laboratório de Processamento de Imagens e Geoprocessamento

BEZOS EARTH FUND

GLAD Global Land Analysis & Discovery

CORNELL UNIVERSITY FOUNDED A.D. 1865

OpenGeoHUB Connect · Create · Share · Repeat

IIASA International Institute for Applied Systems Analysis