Open Environmental Data Cube Europe (http://EcoDataCube.eu)

Analysis-ready datasets produced through ensemble machine learning

Leandro Parente

leandro.parente@opengeohubb.org

https://opengeohub.org

OpenGeoHUB
Connect • Create • Share • Repeat

EcoDataCube

Outline

- Introduction & context
- EO data preparation (Landsat and Sentinel-2)
- Ensemble Machine Learning workflow
- Mapping products
- Processing workflow
- Data portal and catalog (STAC)

EcoDataCube

(http://EcoDataCube.eu)

Introduction & context

Geo-harmonizer: EU-wide automated mapping system for harmonization of Open Data based on FOSS4G and Machine Learning

Programme:
CEF Telecom

Call year:
2018

Location of the Action:
Croatia, Czech Republic, Germany, Netherlands, Romania

Implementation schedule:
September 2019 to June 2022

Maximum EU contribution:
€1,423,864

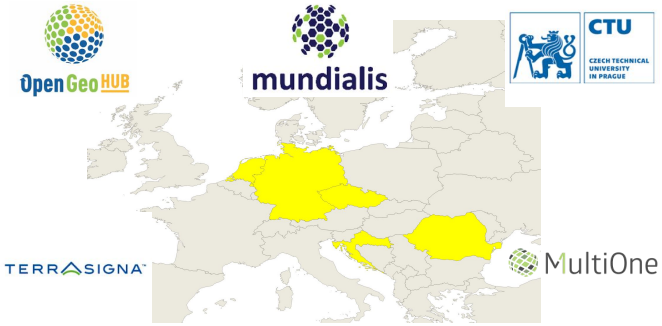
Total eligible costs:
€1,898,485

Percentage of EU support:
75%

2018-EU-IA-0095

The overall objective of the Action is to develop an original, web-based, scalable and modular system ("Geo-harmonizer") for hosting and accessing various thematic geospatial data layers (vector and raster GIS layers) to support cross-border services over the entire continental Europe.

The beneficiaries will create a data portal and a software suite extending a wide variety of free and open source software solutions for geospatial data (FOSS4G) in combination with state-of-the-art Machine Learning Algorithms, and will be made available within EU-supported High Performance



CEF programme

European Commission

Introduction & context

100 m resolution only
partially harmonized
inter-periods missing

Sentinel 2 time-series
Landsat 7/8 time-series
TANDEM-x topographic data
ALOS radar images

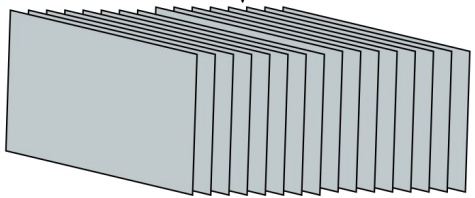
30 m resolution
full space-time coverage

cellular automata,
urban growth models...

Model training
(Machine Learning based)

Spatiotemporal model of land cover dynamics

Complete, harmonized
time-series of land cover images
2000, 2001, 2002, ... 2020



30 m resolution
full space-time coverage
fully harmonized / seamless

Help improve **land cover mapping** (CLC) by automate predictions at higher spatial resolution (10—30m)

Introduction & context

100 m resolution only
partially harmonized
inter-periods missing

Sentinel 2 time-series
Landsat 7/8 time-series
TANDEM-x topographic data
ALOS radar images

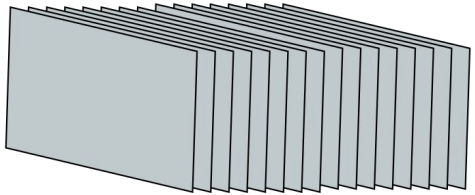
30 m resolution
full space-time coverage

cellular automata,
urban growth models...

Model training
(Machine Learning based)

Spatiotemporal model of land cover dynamics

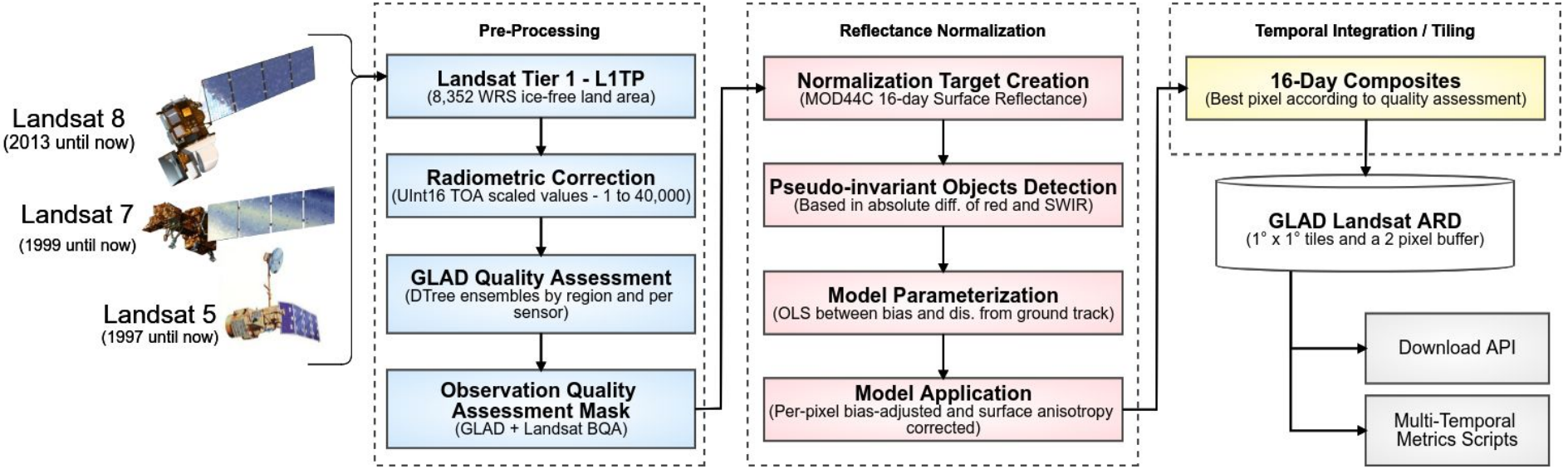
Complete, harmonized
time-series of land cover images
2000, 2001, 2002, ... 2020



30 m resolution
full space-time coverage
fully harmonized / seamless

Help improve **land cover mapping** (CLC) by automate predictions at higher spatial resolution (10—30m)

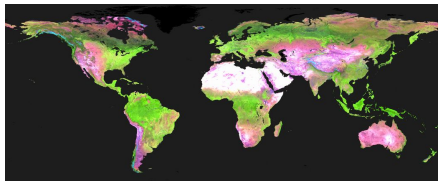
GLAD Landsat ARD



Limitations - It's not suitable for**:**

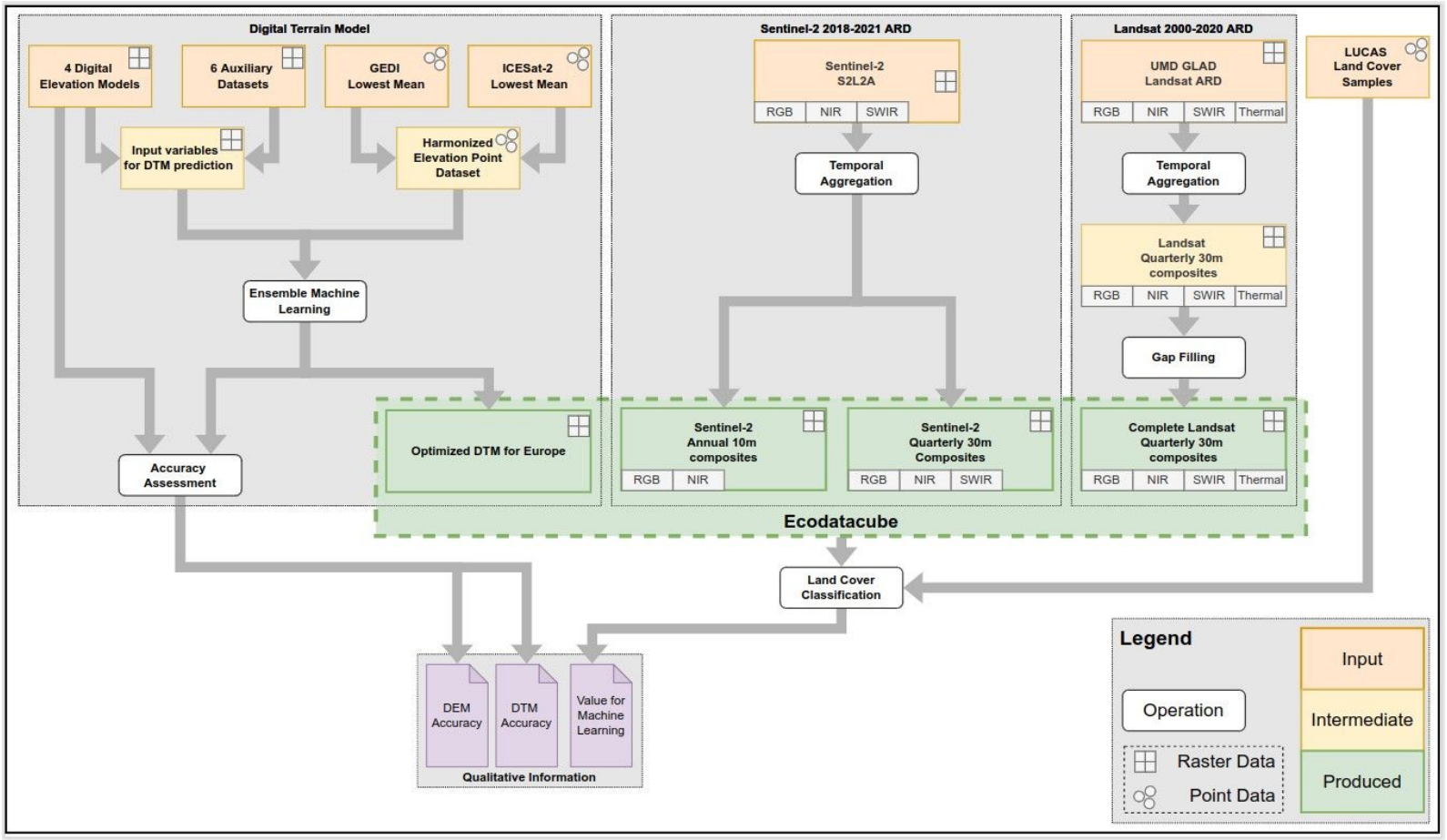
- Real-time land cover monitoring (1-month delay)
- Winter time image processing above 30N and below 45S Latitude
- Precise analysis of land surface reflectance
- Water quality assessment or any other hydrology applications beyond surface water extent mapping

Global Land Analysis & Discovery



http://dx.doi.org/10.3390/rs12030426

EO data preparation



Landsat gapfilling

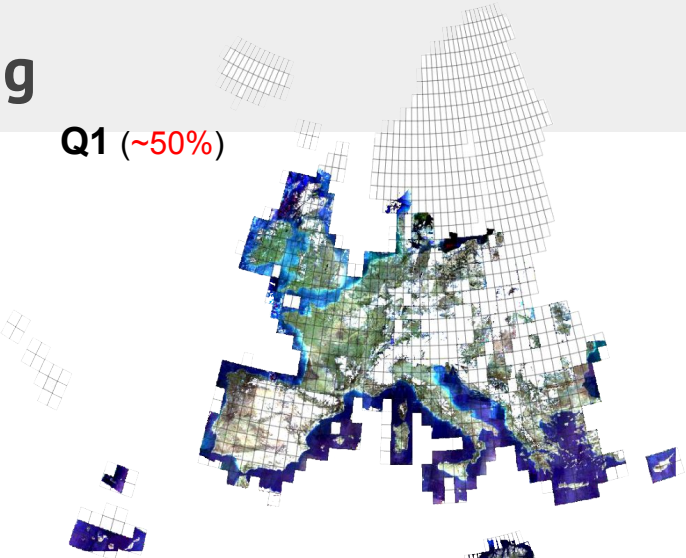
Time span
2000 — 2020

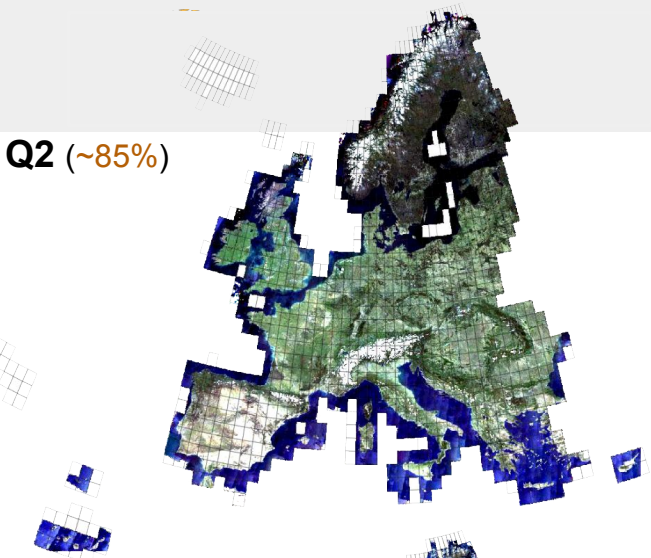
Q1: Dec. 2 — Mar. 20

Q2: Mar. 21 — Jun. 24

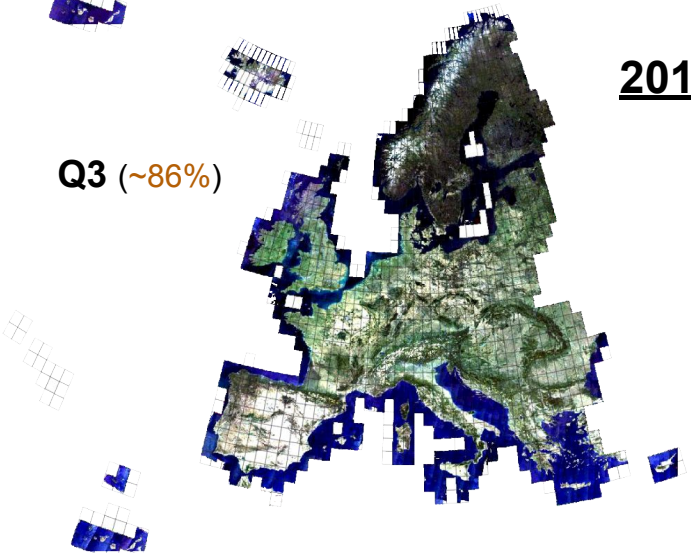
Q3: Jun. 25 — Sep. 12

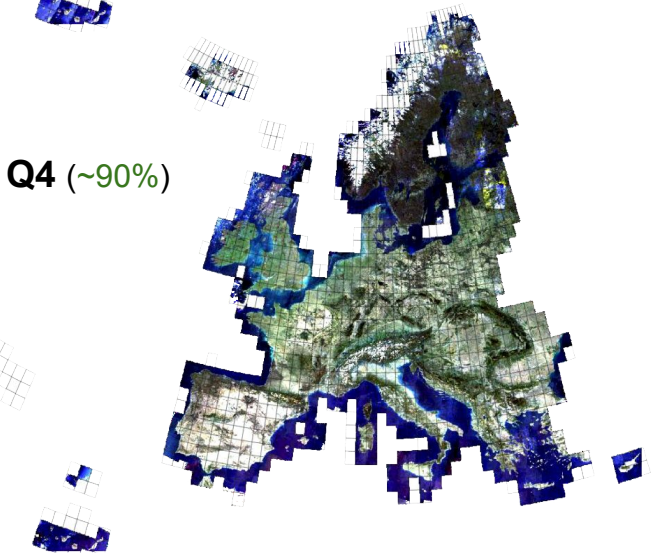
Q4: Sep. 13 — Dec. 1



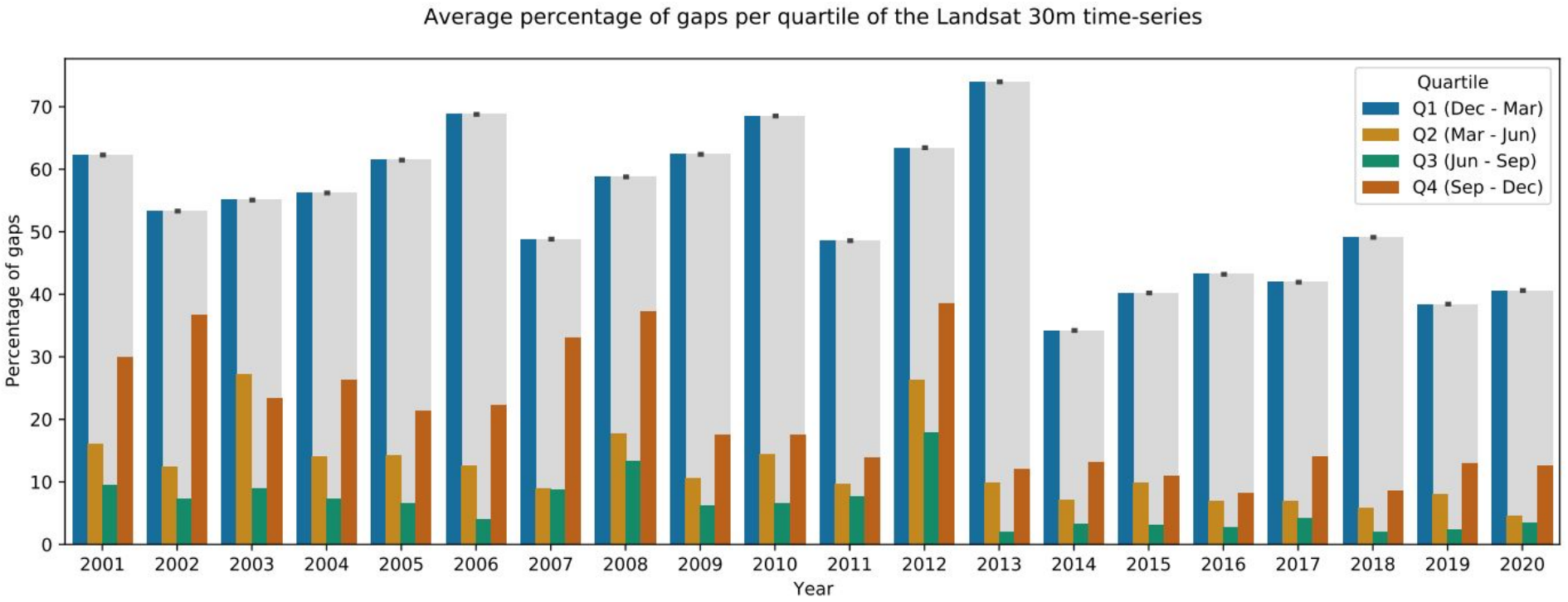


2018





Landsat gapfilling



EUMAP LIBRARY

Temporal Moving Window Median (TMWM)

https://eumap.readthedocs.io/en/latest/_autosummary/eumap.gapfiller.TMWM.htm

Landsat gapfilling - Dec. 2 — Mar. 20 / 2018

Kamp-Lintfort, Germany

Landsat gapfilling - Dec. 2 — Mar. 20 / 2018

Kamp-Lintfort, Germany

Landsat and Sentinel data cube (consistent & gapfilled)

30m resolution almost 400GB of analysis-ready **Sentinel-2** data: P25, P50 and P75 for all bands for 2016 to 2021.

30m resolution **Landsat** ARD 10TB of data available as COGs via our Wasabi service.

10m resolution **Sentinel-2** mosaics (120GB per image!) are also available via https://EcoDataCube.eu.

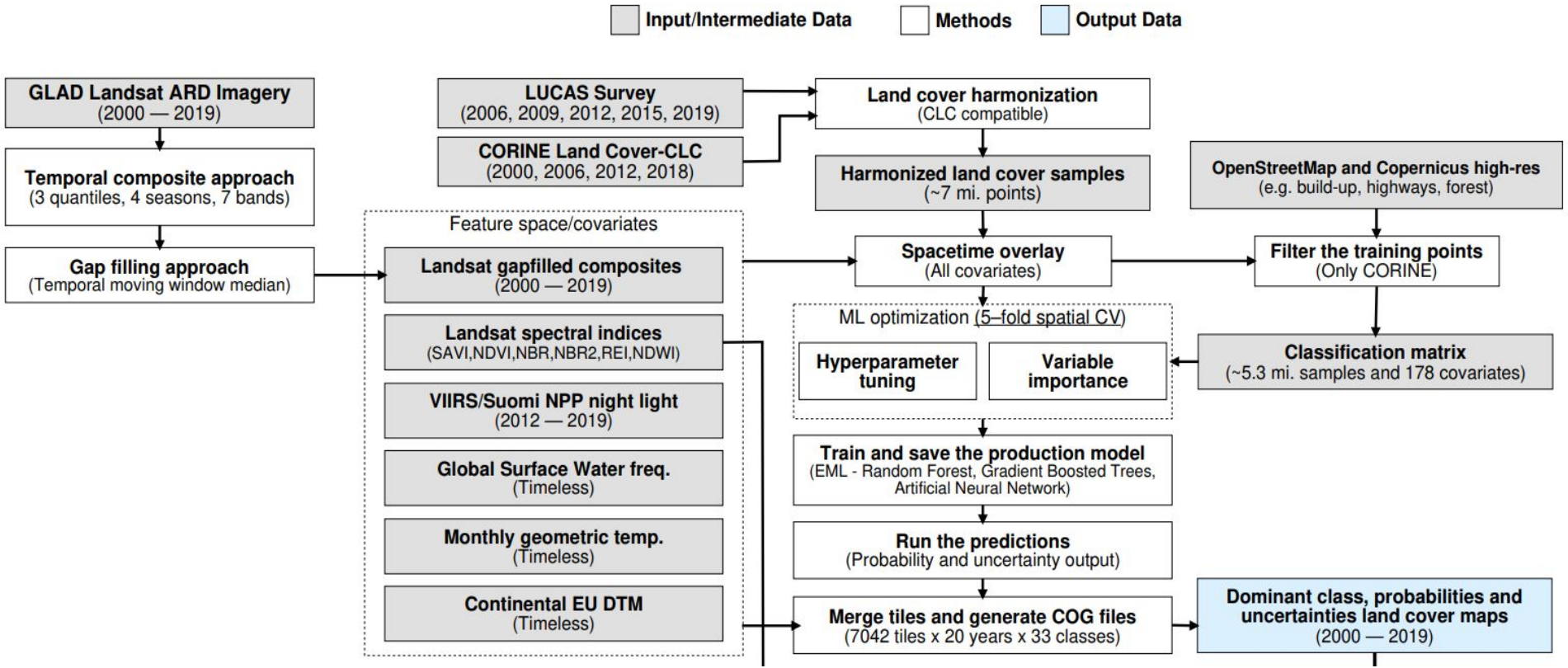
Manuscript submitted

Landsat and Sentinel data cube (consistent & gapfilled)

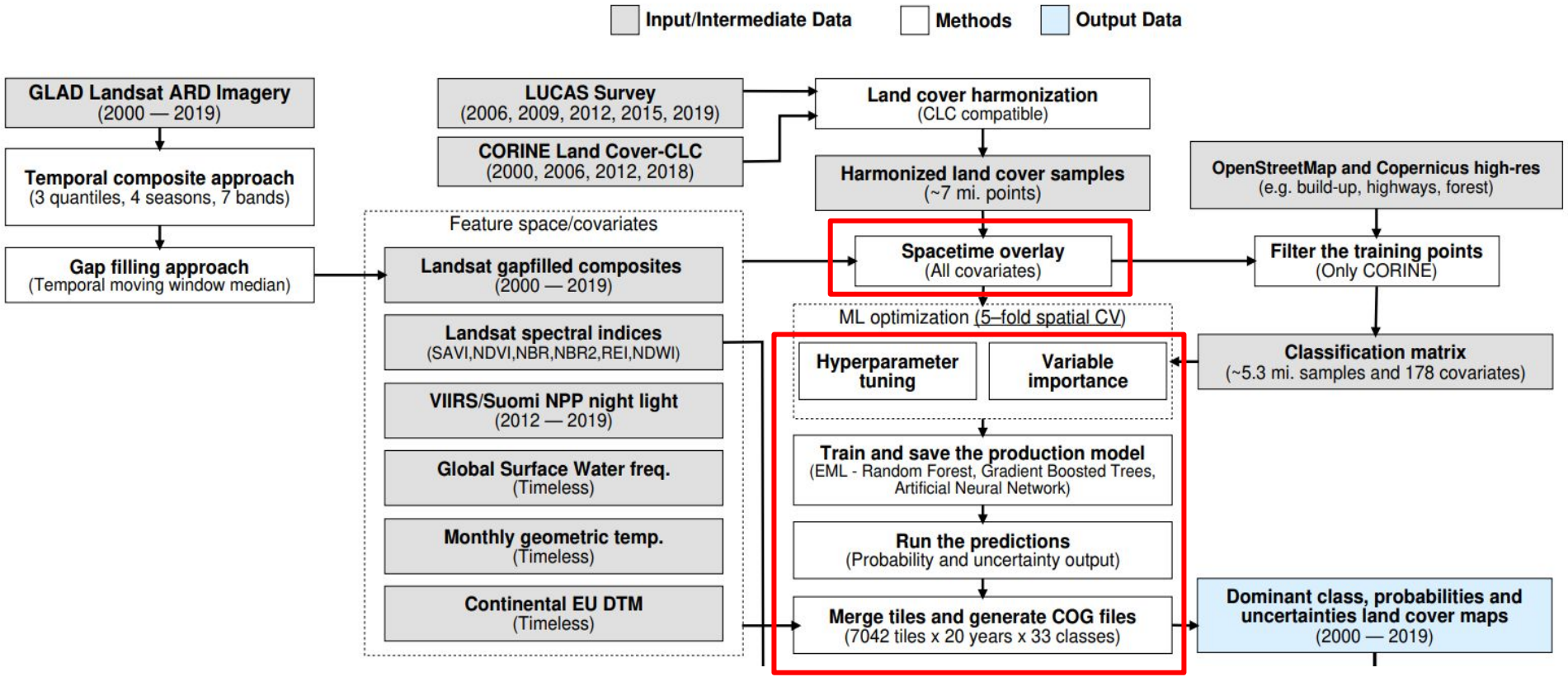
OpenGeoHUB
Connect • Create • Share • Repeat

10m resolution mosaics (120GB per image)

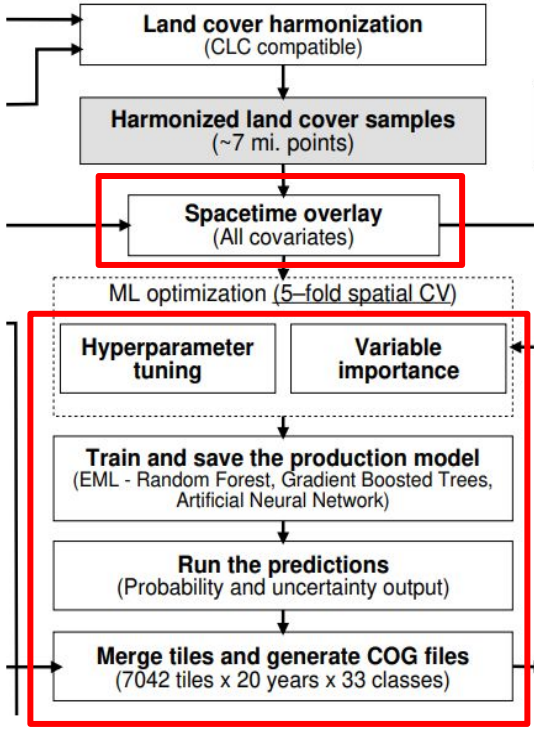
Ensemble Machine Learning workflow

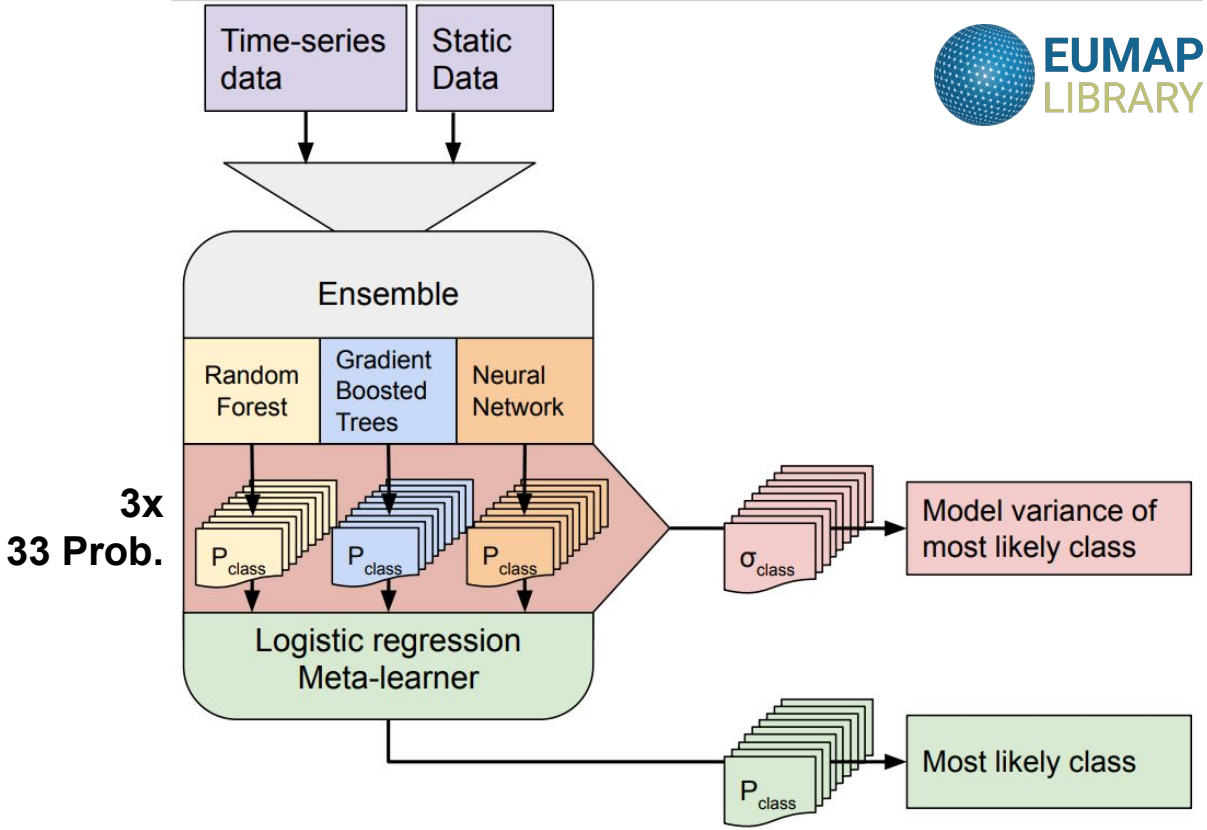


Ensemble Machine Learning workflow



Ensemble Machine Learning workflow

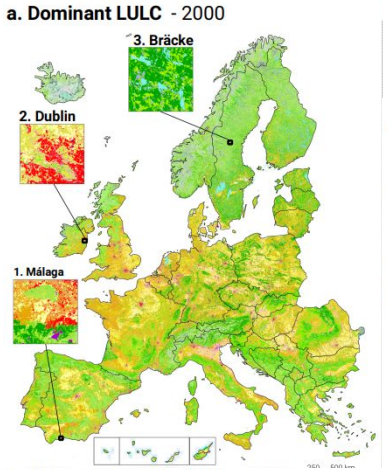


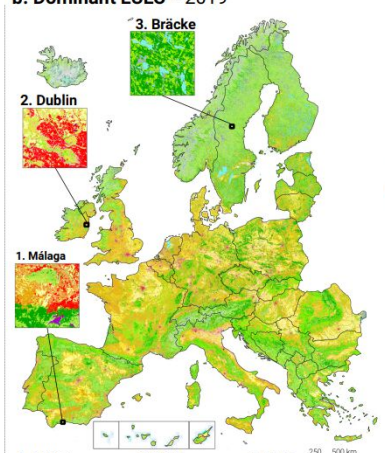


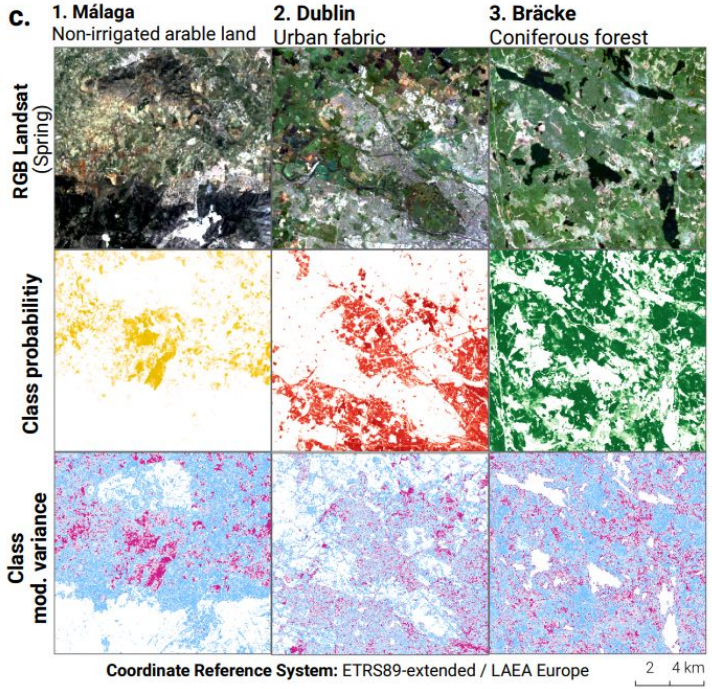
30-m Land use / land cover mapping

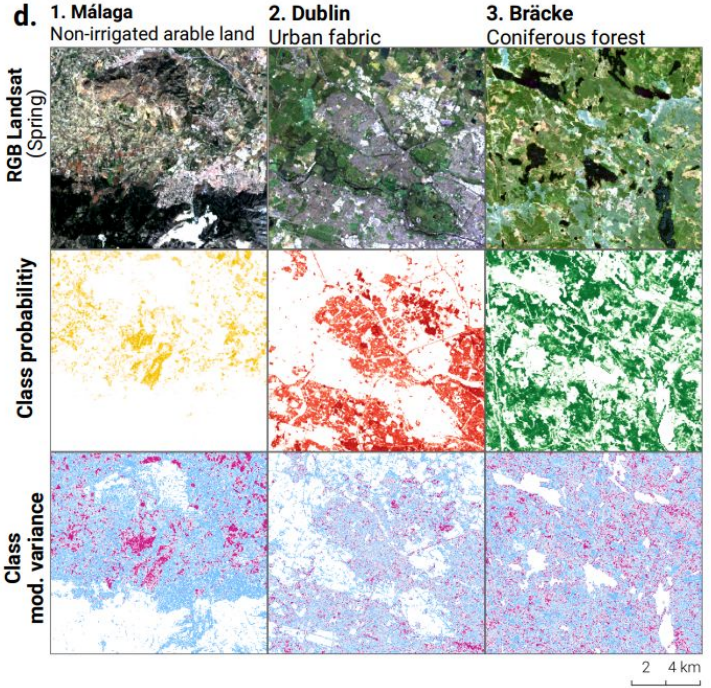
https://doi.org/10.7717/peerj.13573

33 CLC classes (2000—2020)

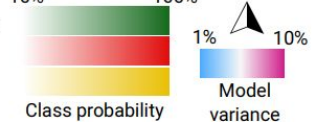




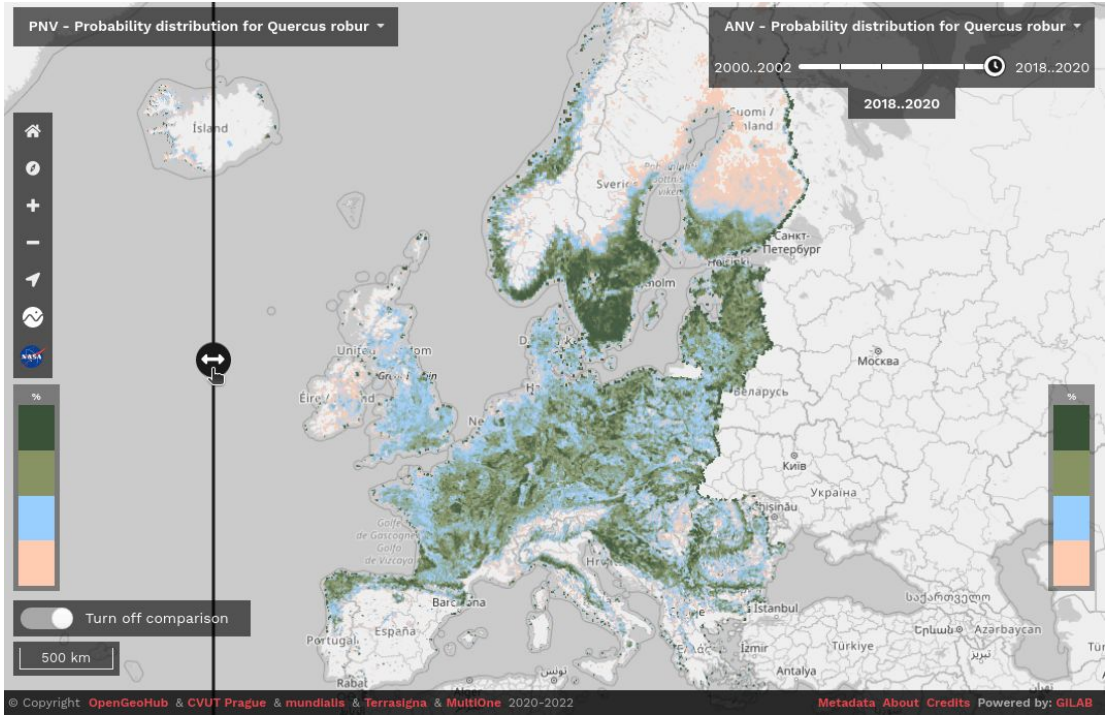




Urban fabric
Road and rail networks and associated land
Port areas
Airports
Mineral extraction sites
Dump sites
Construction sites
Green urban areas
Non-irrigated arable land
Permanently irrigated arable land
Rice fields
Vineyards
Fruit trees and berry plantations
Olive groves
Pastures
Broad-leaved forest
Coniferous forest
Natural grasslands
Moors and heathland
Sclerophyllous vegetation
Transitional woodland-shrub
Beaches, dunes, sands
Bare rocks
Sparsely vegetated areas
Burnt areas
Glaciers and perpetual snow
Inland wetlands
Maritime wetlands
Water courses
Water bodies
Coastal lagoons
Estuaries
Sea and ocean

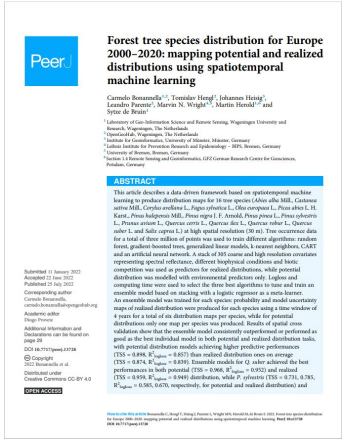


30-m Forest tree species distribution



Quercus Robur (potential and realized)

- **16 species** (potential and realized distribution)
- 2000 - 2020, with time steps:
  - 2000 - 2002
  - 2002 - 2006
  - 2006 - 2010
  - 2010 - 2014
  - 2014 - 2018
  - 2018 - 2020



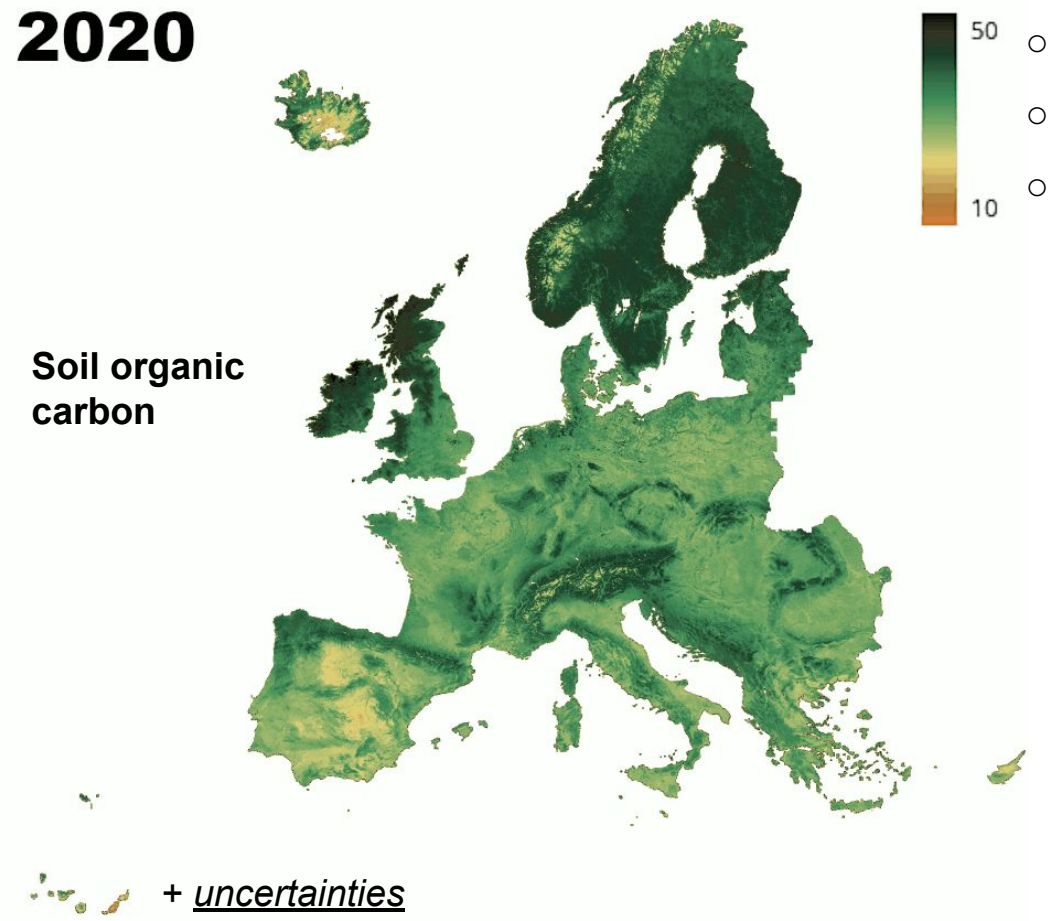
Forest tree species distribution for Europe 2000–2020: mapping potential and realized distributions using spatiotemporal machine learning

ABSTRACT

Probabilities + *uncertainties*

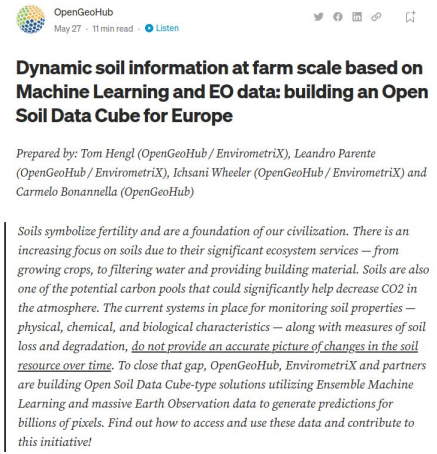
http://dx.doi.org/10.7717/peerj.13728

30-m Soil data cube (3D+t)



+ *uncertainties*

- **4 depths** (0, 30, 60 and 100 cm)
- **SOC, pH, clay, sand and bulk density**
- 2000 - 2020, with time steps:
  - 2000 - 2002
  - 2002 - 2006
  - 2006 - 2010
  - 2010 - 2014
  - 2014 - 2018
  - 2018 - 2020



OpenGeoHub
May 27 · 11 min read · Listen

Dynamic soil information at farm scale based on Machine Learning and EO data: building an Open Soil Data Cube for Europe

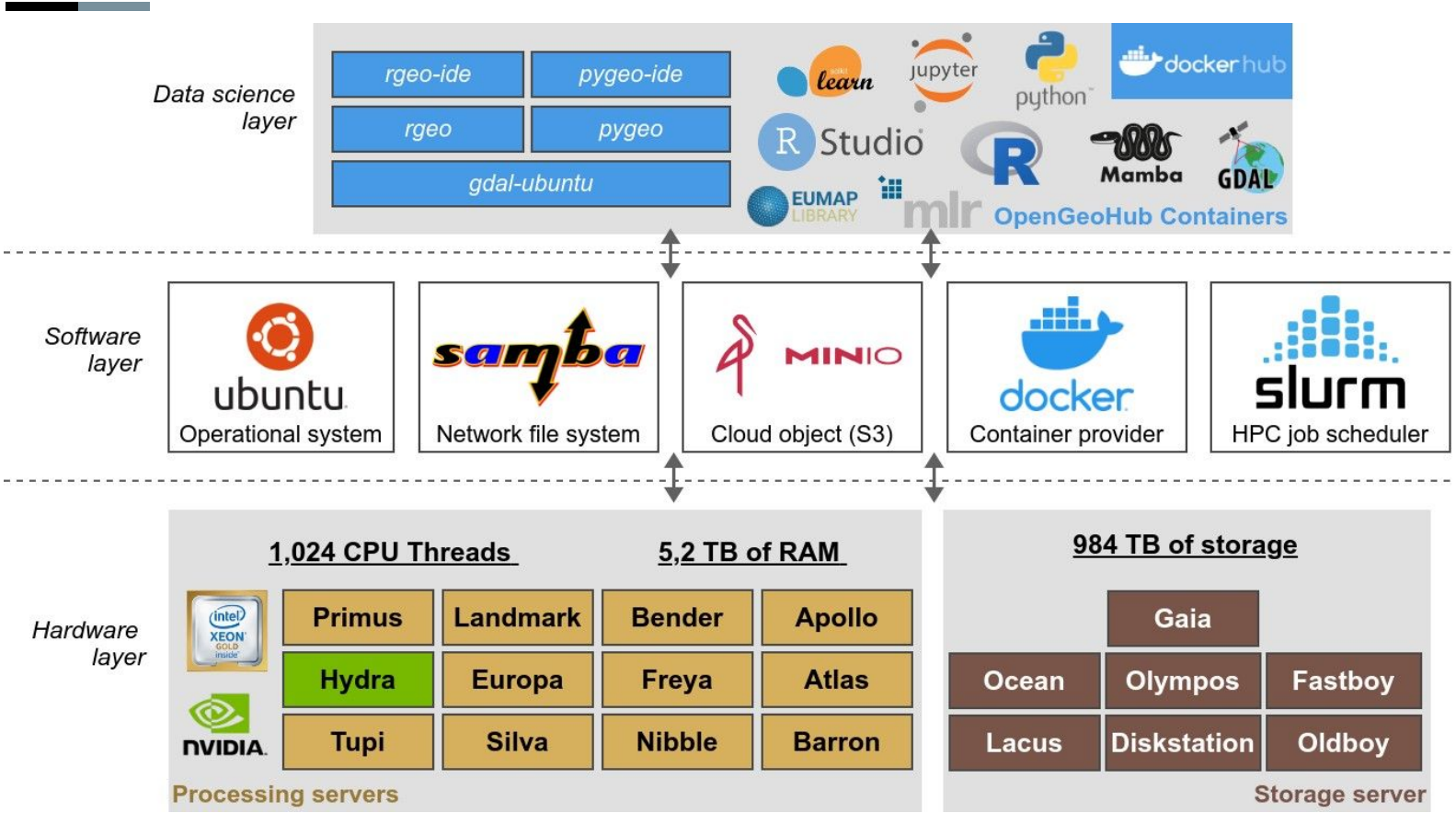
Prepared by: Tom Hengl (OpenGeoHub / EnvirometriX), Leandro Parente (OpenGeoHub / EnvirometriX), Ichsani Wheeler (OpenGeoHub / EnvirometriX) and Carmelo Bonannella (OpenGeoHub)

Soils symbolize fertility and are a foundation of our civilization. There is an increasing focus on soils due to their significant ecosystem services — from growing crops, to filtering water and providing building material. Soils are also one of the potential carbon pools that could significantly help decrease CO2 in the atmosphere. The current systems in place for monitoring soil properties — physical, chemical, and biological characteristics — along with measures of soil loss and degradation, do not provide an accurate picture of changes in the soil resource over time. To close that gap, OpenGeoHub, EnvirometriX and partners are building Open Soil Data Cube-type solutions utilizing Ensemble Machine Learning and massive Earth Observation data to generate predictions for billions of pixels. Find out how to access and use these data and contribute to this initiative!

Manuscript in preparation

Using Open Source to crunch big EO data

OpenGeoHUB
Connect • Create • Share • Repeat



Using Open Source to crunch big EO data

OpenGeoHUB
Connect • Create • Share • Repeat

Versioned and fully reproducible development environment

opengeohub/rgeo-ide ☆

By opengeohub • Updated a month ago

Rstudio and R (>= 4.1.1) working with GDAL (>= 3.1.4) and several geospatial and ML packages

Container

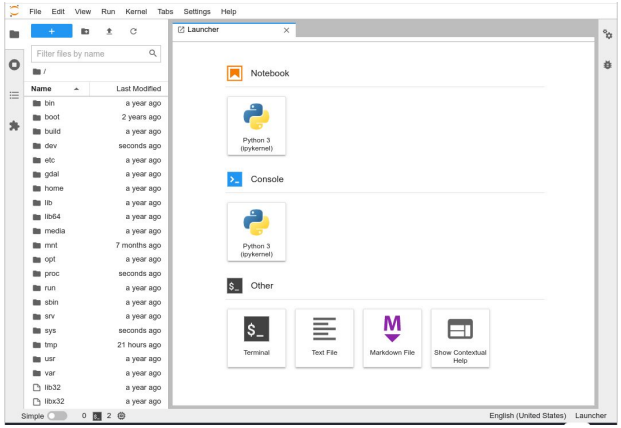
opengeohub/pygeo-ide ☆

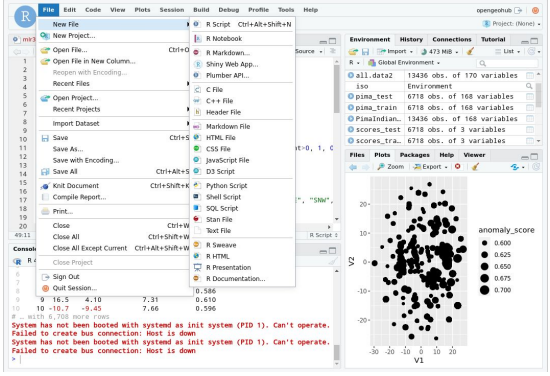
By opengeohub • Updated a month ago

JupyterLab and Python (>= 3.8.6) working with GDAL (>= 3.1.4) and several geospatial and ML packages

Container

https://hub.docker.com/r/opengeohub/





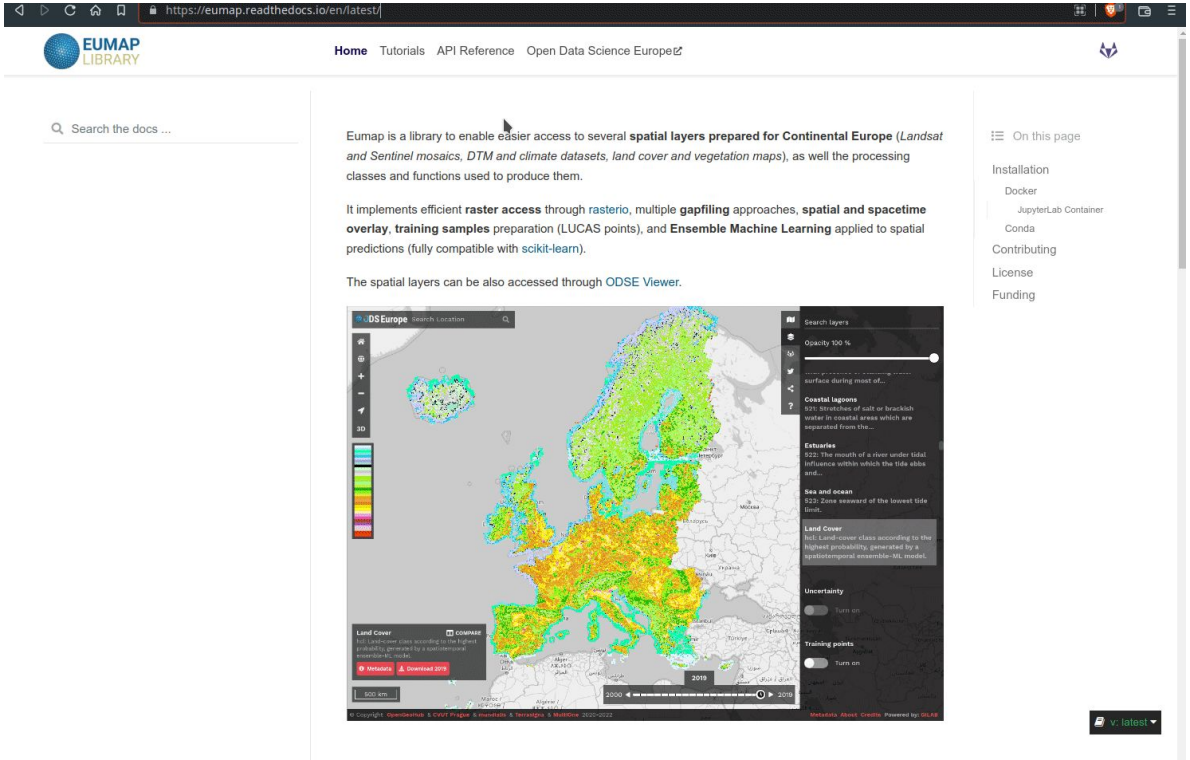
docker

Using Open Source to crunch big EO data

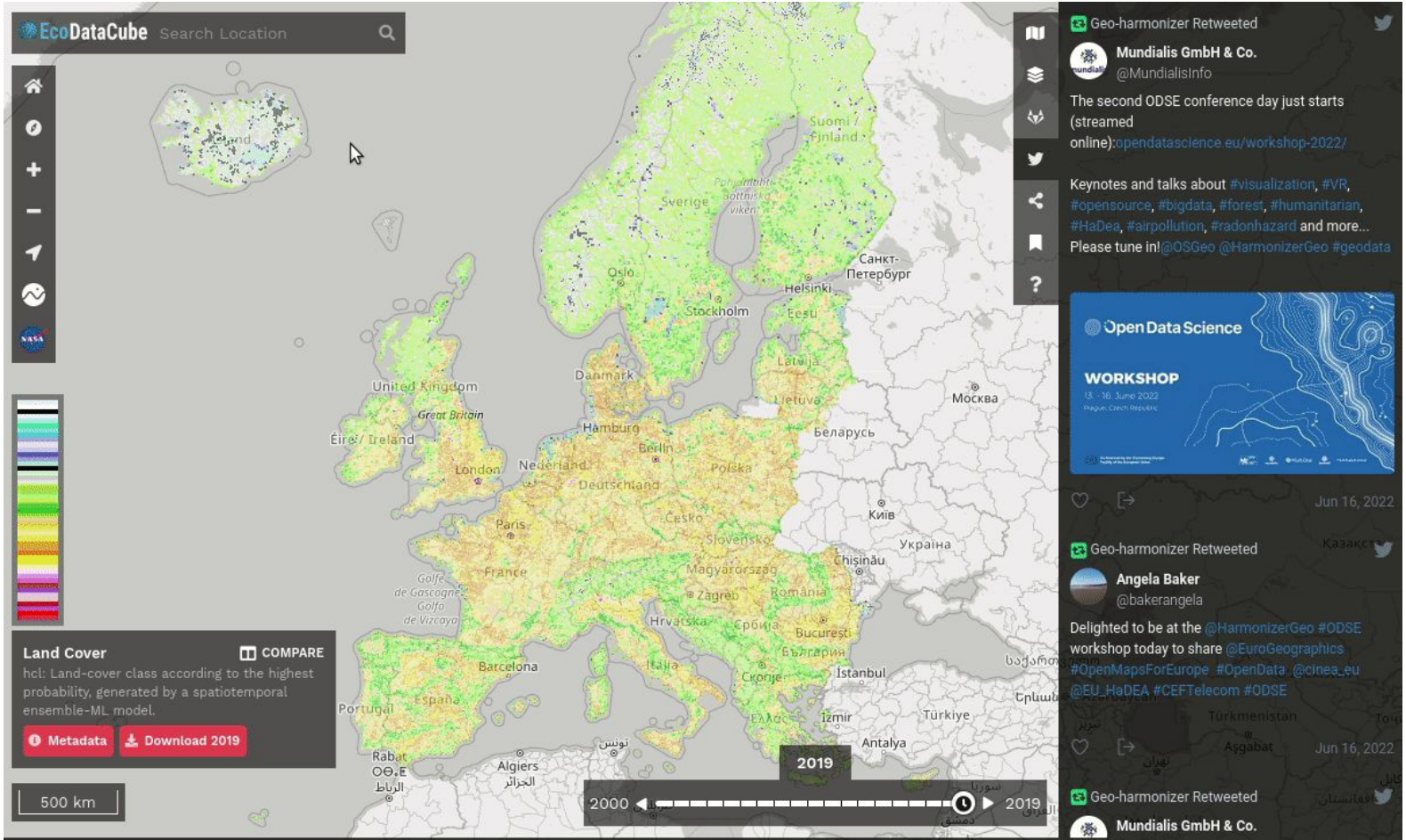
OpenGeoHUB
Connect • Create • Share • Repeat

eumap Python library (http://eumap.readthedocs.io/)

- Extensive utility suite for working with EO and spatial data in general
- Was heavily used for production of OGH datasets (e.g. LandMapper ML automation)
- Complete, auto generated documentation



Data portal https://ecodatacube.eu



Data catalog

A complete comprehensive raster database with all layers imported and documented

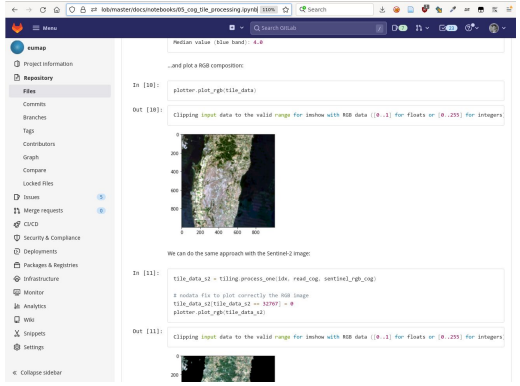
STAC catalog:
https://stac.EcoDataCube.eu

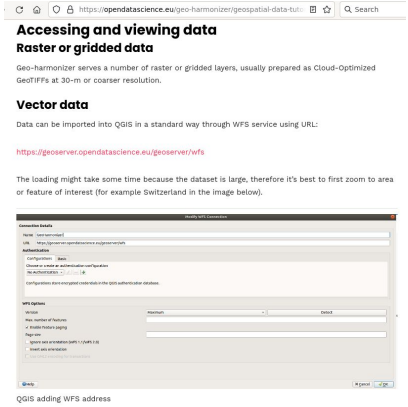
Online tutorials:

https://opendatascience.eu/geo-harmonizer/geospatial-data-tutorial/

Jupyter notebooks:

https://eumap.readthedocs.io/en/latest/notebooks/10_stac.html





QGIS adding WFS address

EcoDataCube

Open Environmental Data Cube Europe

Source Share

Browse

Description

Spatio-Temporal Asset Catalog for European-wide layers provided by Open Environmental Data Cube Europe.

Catalogs (142) Tiles List Ascending Descending

Continental Europe land mask
Overview: Land mask estabilishing the mapping area of the project according to European Economic Area (EEA)...
1/1/2014 - 12/31/2016

Annual moors and heathland at 30 m (2000–2020)
Overview: Moors and heathland Pastures for continental Europe based on Ensemble Machine Learning (EM...
1/1/2000 - 12/31/2020

PNV - Probability distribution for Salix caprea (2000–2020)
Overview: Potential Natural Vegetation (PNV): potential probability of occurrence for the Goat willow...
1/1/2018 - 12/31/2020

Quarterly blue band of GLAD landsat ARD (2000–2020)
Overview: The temporal composites of blue band based on GLAD Landsat ARD, considering four quarterly per...
12/2/1999 - 12/1/2020

Annual sclerophyllous vegetation at 30 m (2000–2020)
Overview: Sclerophyllous vegetation Pastures for continental Europe based on Ensemble Machine Learning (EM...
1/1/2000 - 12/31/2020

ERA5 Land precipitation daily sum (2000–2020)
Overview: Precipitation daily sums from 2000 to 2020 resampled with CHELSA to 1 km resolution...
1/1/2000 - 12/31/2020

Quarterly green band of GLAD landsat ARD (2000–2020)
Overview: The temporal composites of green band based on GLAD

Annual transitional woodland-shrub at 30 m (2000–2020)

ERA5 Daily land air temperature (2000–2020)
Overview: Air temperature daily

Leandro Parente

leandro.parente@opengeohubb.org

https://opengeohub.org

OpenGeoHUB

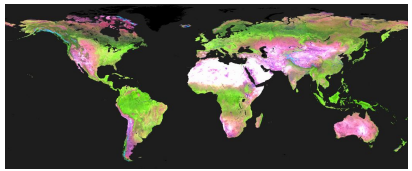
Connect • Create • Share • Repeat

Working properly with Landsat data

OpenGeoHUB
Connect • Create • Share • Repeat

GLAD Landsat ARD

- Globally consistent analysis ready data (ARD) for multi-decadal LCLU monitoring
- 16-day time-series composites from Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Per-pixel observation quality flag
- MODIS (MOD44C) surface reflectance calibrated
- Product organized by 1 × 1 degree tiles
- Automatically download through HTTP API
- Product under Creative Commons Attribution License



FORCE

- An all-in-one remote sensing processing framework for Sentinel-2 A/B MSI and Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Advanced cloud and cloud shadow detection
- Integrated atmospheric, topographic and BRDF correction
- Reprojection and grinding capabilities
- Different strategies to generate composites (e.g. best available pixel, spectral temporal metrics)
- Free software under GNU License v.3

FORCE
Framework for Operational Radiometric Correction for Environmental Monitoring

Working properly with Landsat data

OpenGeoHUB
Connect • Create • Share • Repeat

GLAD Landsat ARD
(Level 4 product)

Level 4
(Model output)

FORCE
(All levels)

Level 3
(Temporal composites and gridded data)

(Level 1 and level 2 products)

Level 2
(Atmospheric correction)

Level 1
(Radiometrically calibrated and georectified data)

Asrar, G.; Greenstone, R. (Eds.) MTPE EOS Reference Handbook; NASA/Goddard Space Flight Center: Greenbelt, MD, USA, 1995; p. 281.

Using Open Source to crunch big EO data

1. **Split tiles across the nodes** (Slurm array id)

slurm

docker Node 1 | docker Node 2

docker Node 3 | docker Node n

For every tile

2) **Read input data**
3) **Execute data analysis** (in parallel)
4) **Write output data** (individual files)

Reading (http) ↑ ↓ **Writing** (S3)

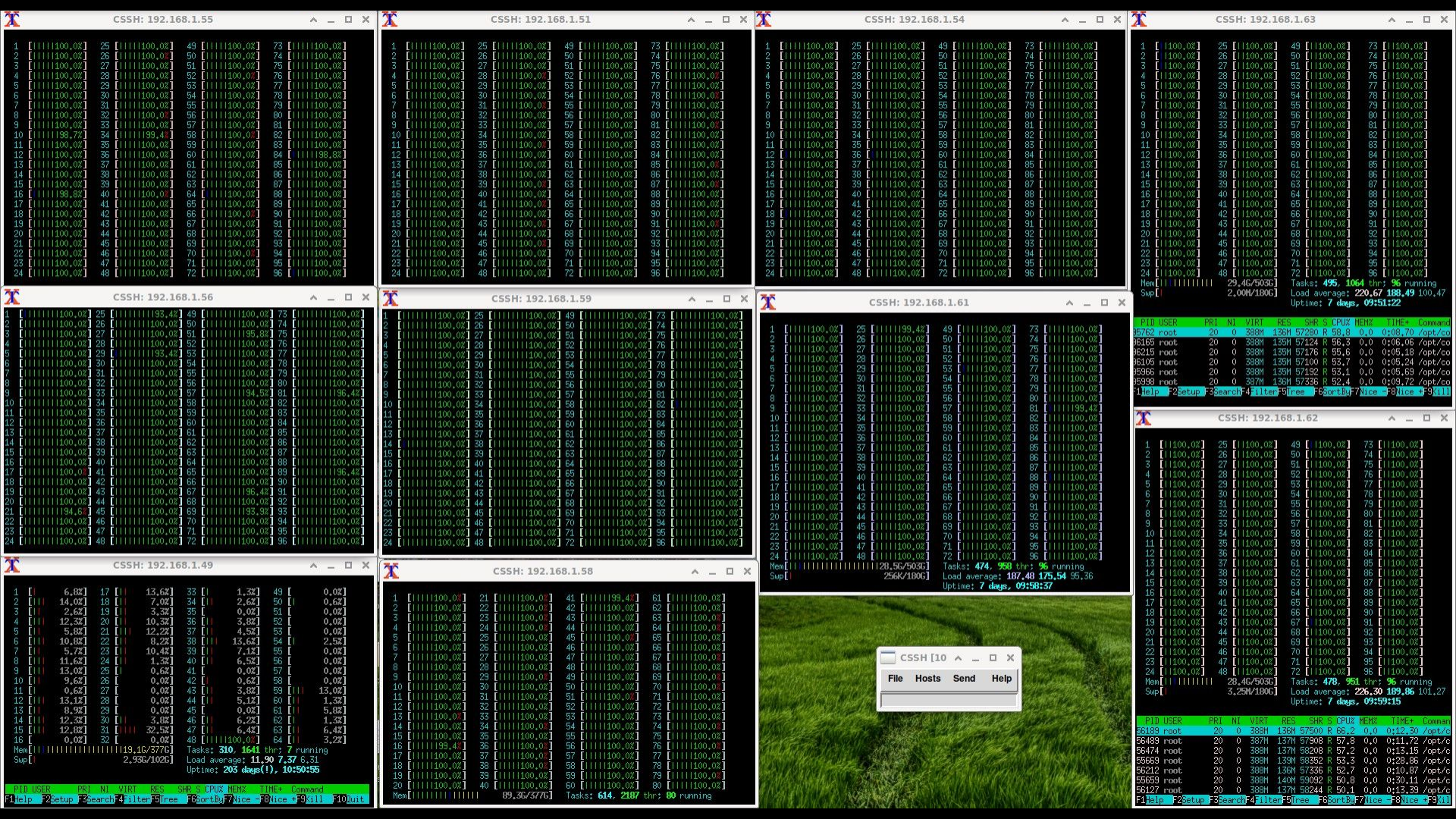
MINIO

Node 1 | Node 2

Node 3 | Node n

Keys to increasing computing efficiency:

- ❑ **Efficient parallel computing,**
- ❑ **Efficient data access RW (S3),**
- ❑ **Efficient storage (COGs),**


CSSH: 192.168.1.55
CSSH: 192.168.1.51
CSSH: 192.168.1.54
CSSH: 192.168.1.63
CSSH: 192.168.1.56
CSSH: 192.168.1.59
CSSH: 192.168.1.61
CSSH: 192.168.1.62
CSSH: 192.168.1.49
CSSH: 192.168.1.58
CSSH [10
File
Hosts
Send
Help

Using Open Source to crunch big EO data

OpenGeoHUB
Connect • Create • Share • Repeat

1. **Split tiles across the nodes** (Slurm array id)

slurm

docker Node 1 | docker Node 2
docker Node 3 | docker Node n

For every tile

2) **Read input data**
3) **Execute data analysis** (in parallel)
4) **Write output data** (individual files)

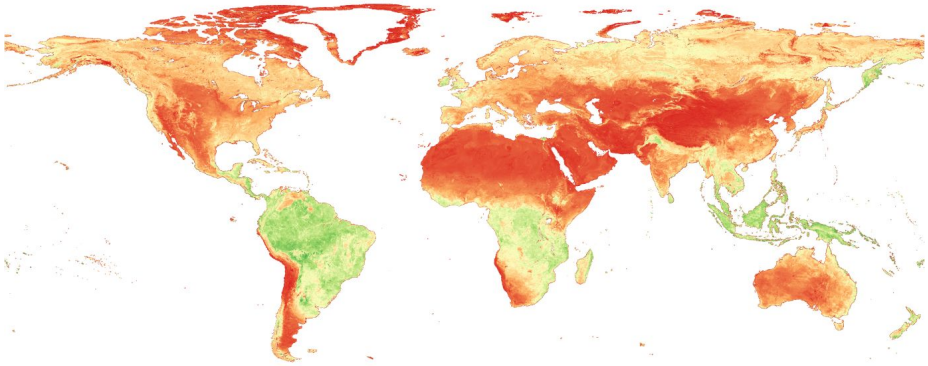
Reading (http) ↑ ↓ **Writing** (S3)

MINIO

Node 1 | Node 2
Node 3 | Node n

GDAL

5) Cloud-optimized geotiff production
(gdalbuildvrt + gdal_translate / gdalwarp)



S3 service

wasabi