



#AIforEOWS #ml4esop

@ECMWF | Reading | 14-17 Nov. 2022



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

Analysis-ready datasets produced through ensemble machine learning



Leandro Parente



leandro.parente@opengeohub.org



<https://opengeohub.org>



Open Geo HUB

Connect • Create • Share • Repeat



EcoDataCube



- 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

2018-EU-IA-0095

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:

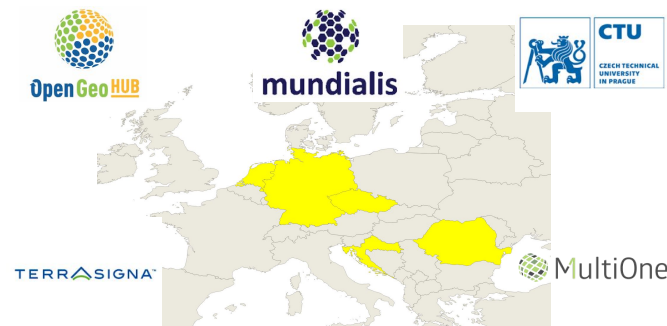
€1,898,485

Percentage of EU support:

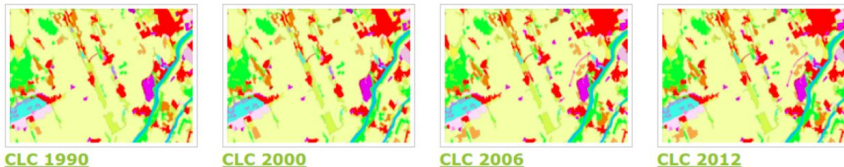
75%

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



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...

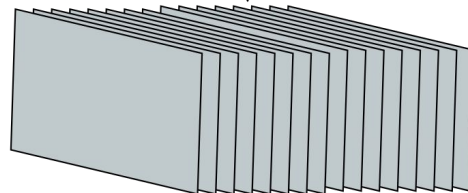


Model training
(Machine Learning based)

Spatiotemporal model
of land cover dynamics

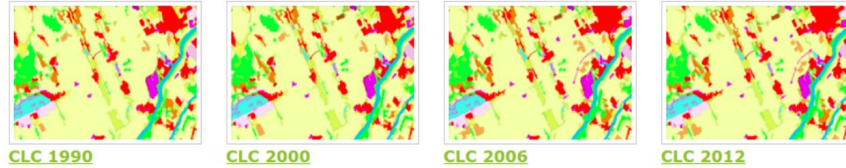
Help improve **land cover mapping (CLC)** by automate predictions at higher spatial resolution (10–30m)

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



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

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...

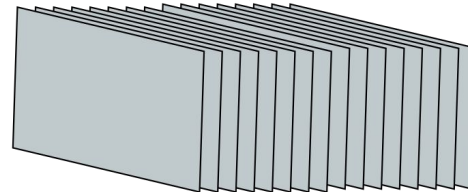


Model training
(Machine Learning based)

Spatiotemporal model
of land cover dynamics

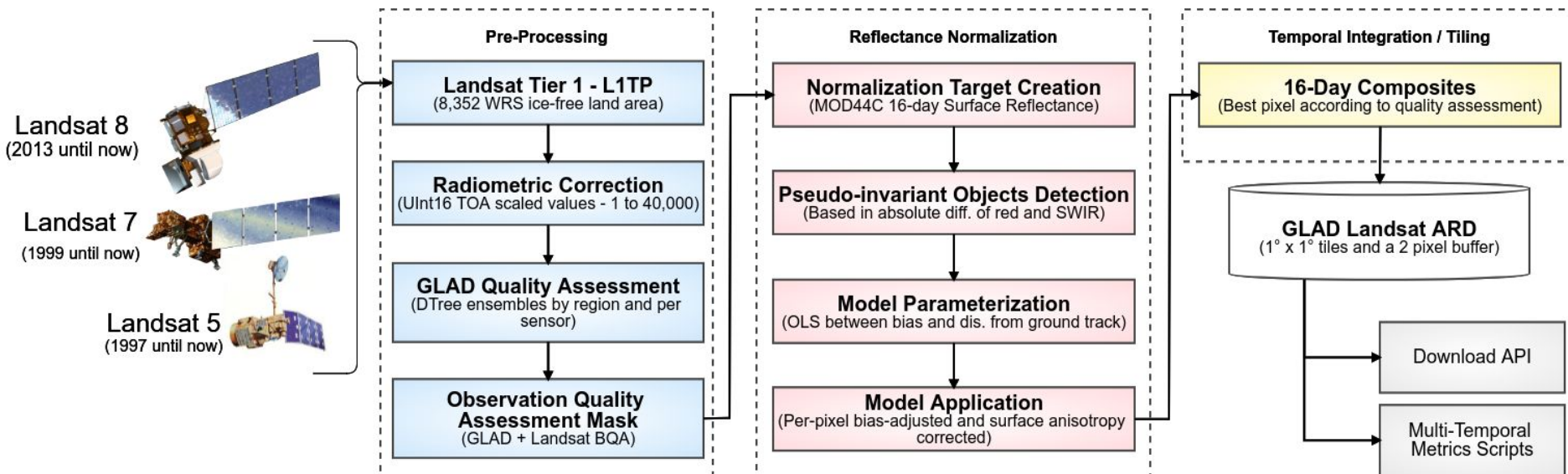
Help improve **land cover mapping (CLC)** by automate predictions at higher spatial resolution (10–30m)

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



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

GLAD Landsat ARD

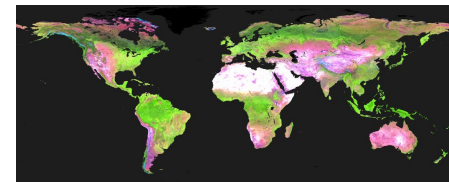


Limitations - It's not suitable for:

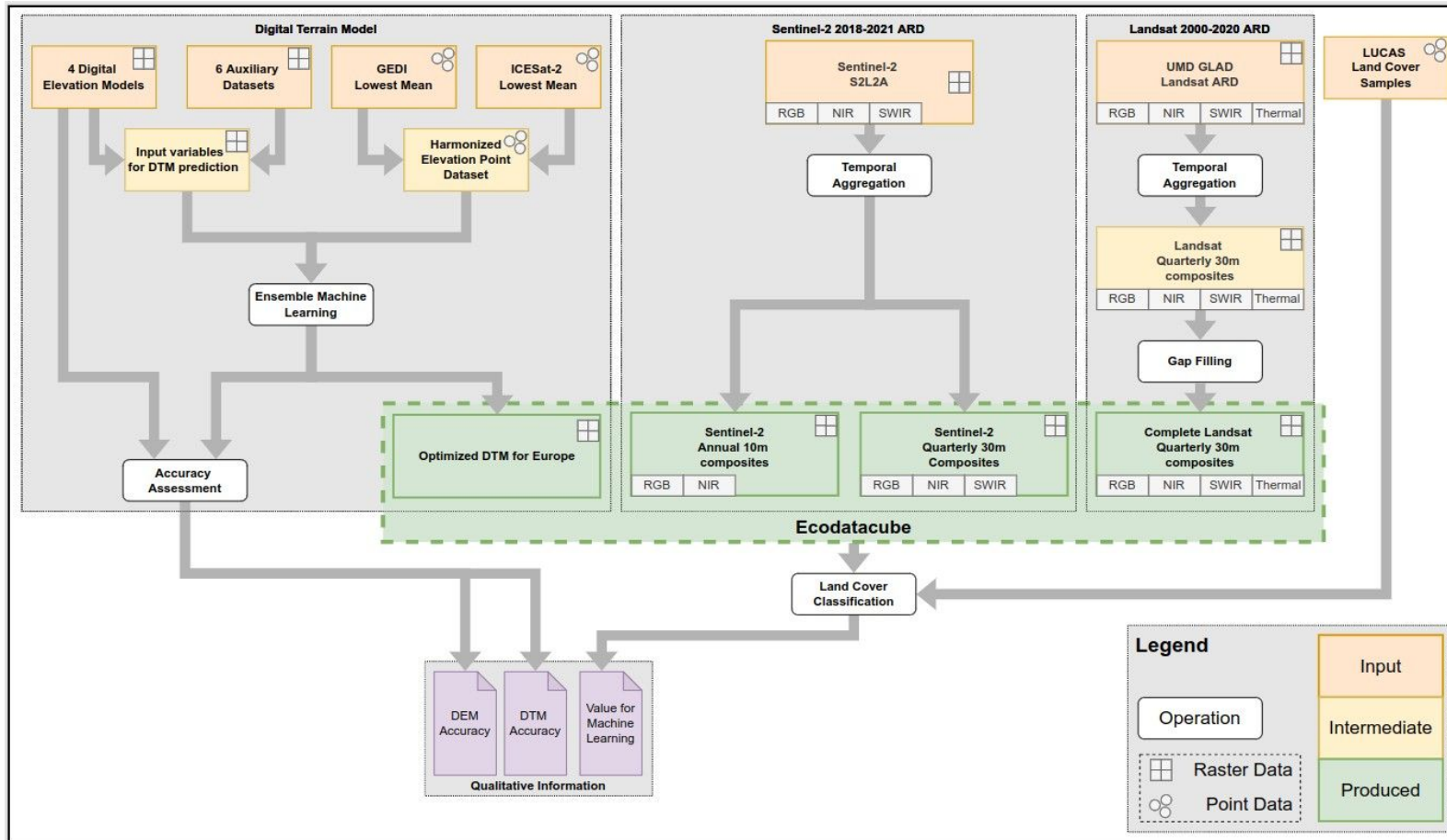
- Real-time land cover monitoring (1-month delay)
- Winter time image processing above 30N and below 45S Latitude
- Precise analysis of land surface reflectance
- Water quality assessment or any other hydrology applications beyond surface water extent mapping



Global Land
Analysis & Discovery



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

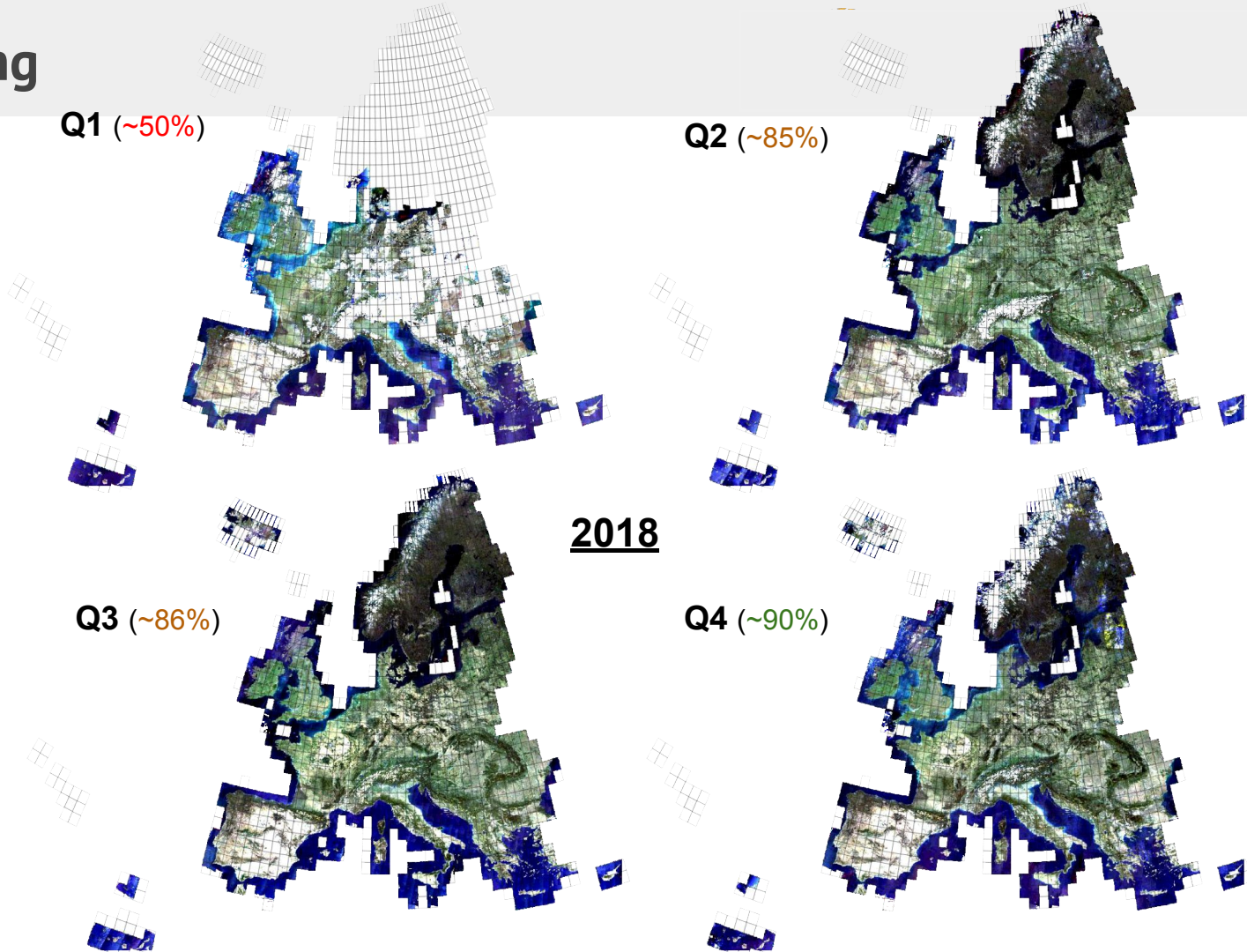
Q1 (~50%)

Q2 (~85%)

2018

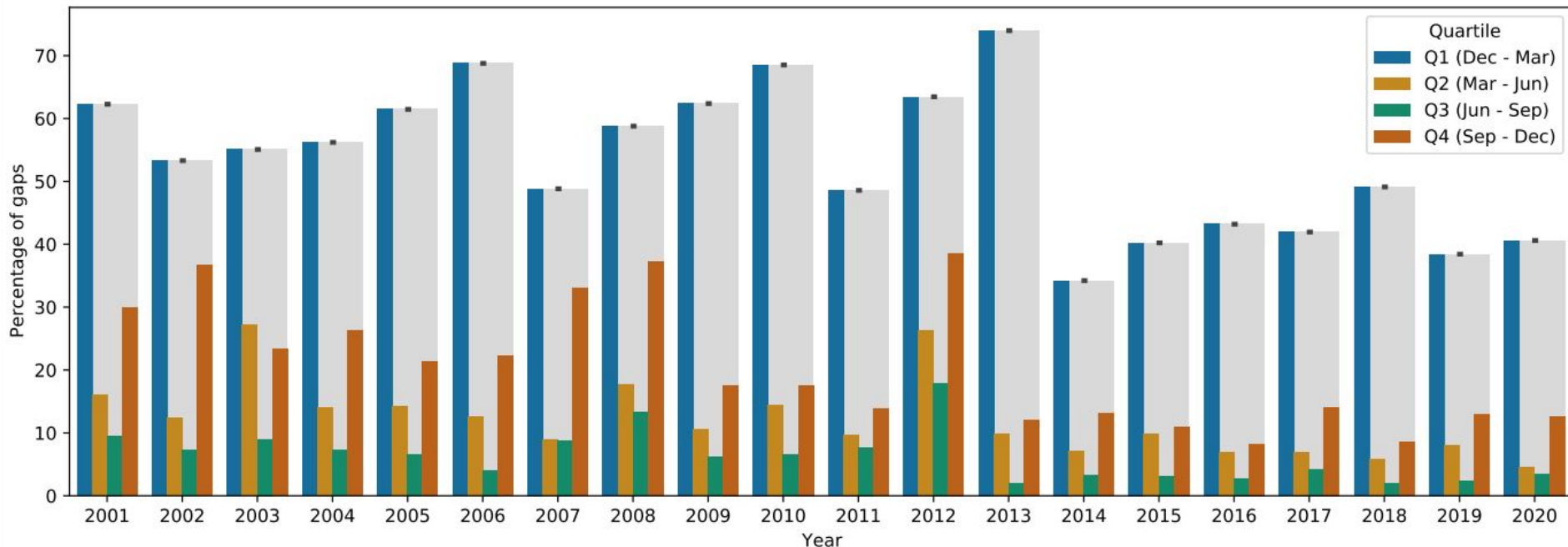
Q3 (~86%)

Q4 (~90%)



Landsat gapfilling

Average percentage of gaps per quartile of the Landsat 30m time-series



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





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>.

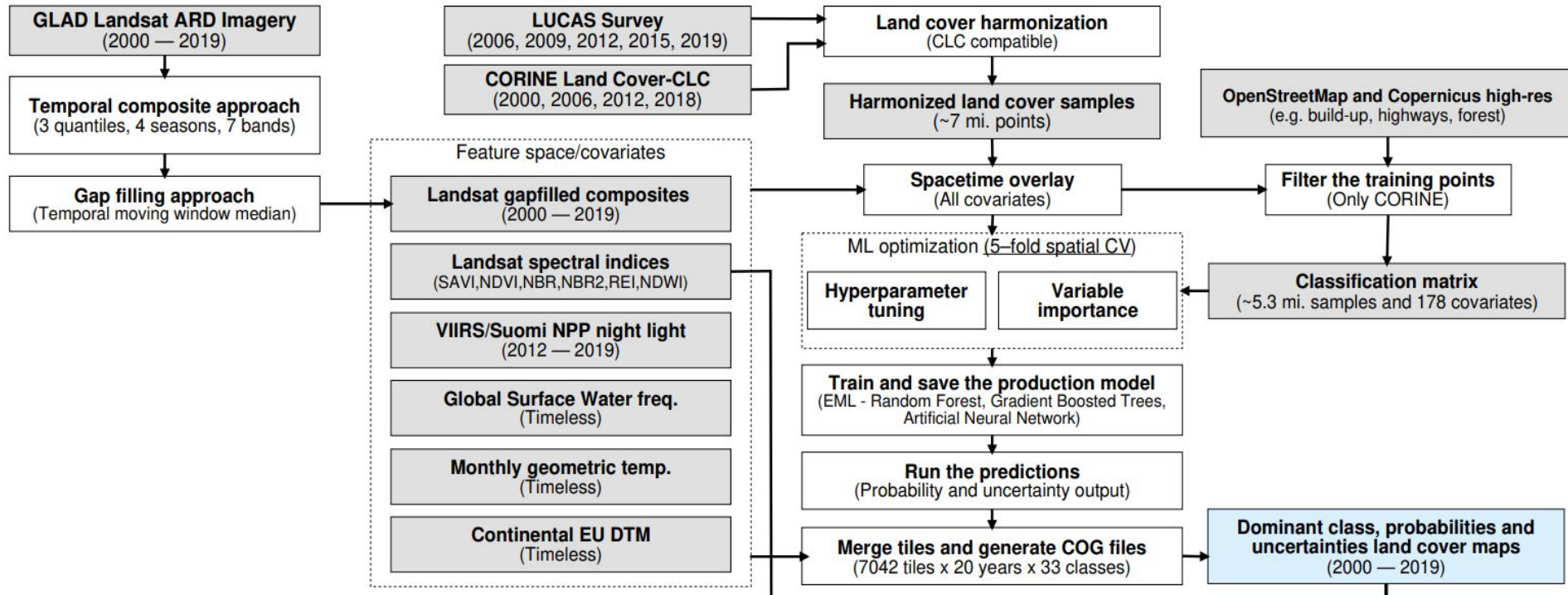
Landsat and Sentinel data cube (consistent & gapfilled)

10m resolution mosaics (120GB per image)



Ensemble Machine Learning workflow

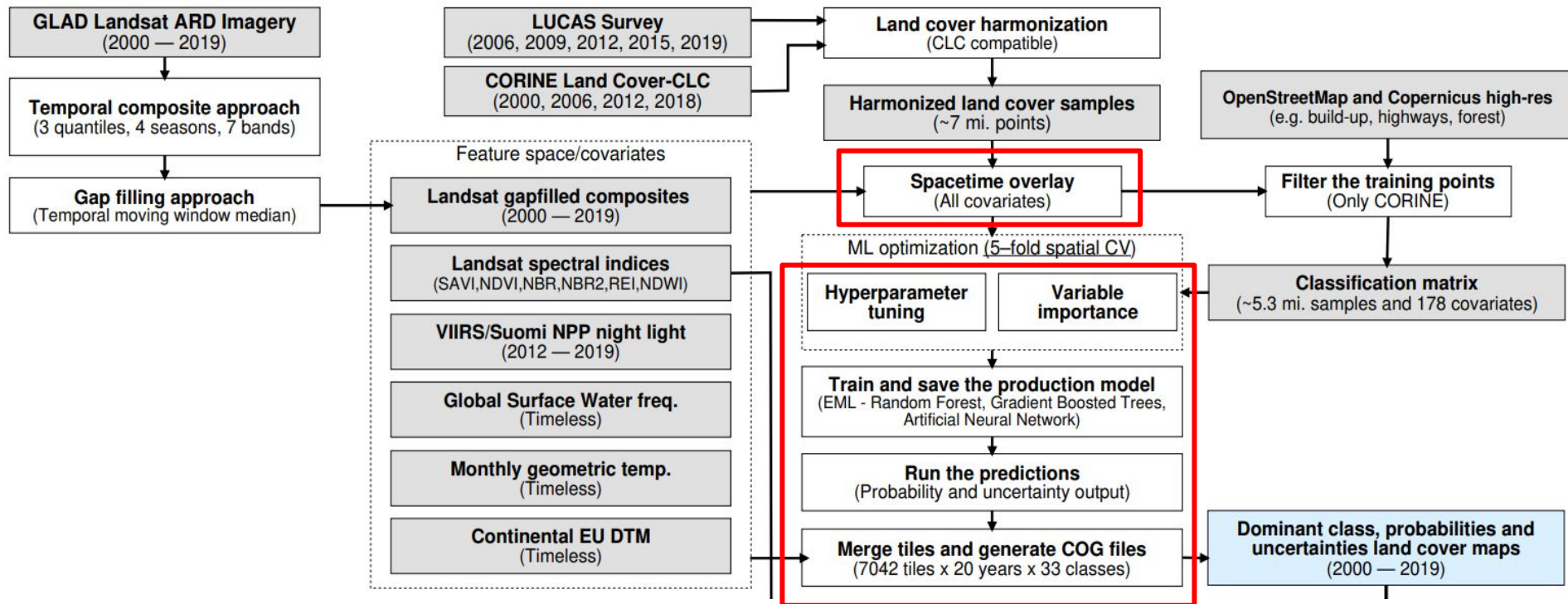
Input/Intermediate Data Methods Output Data



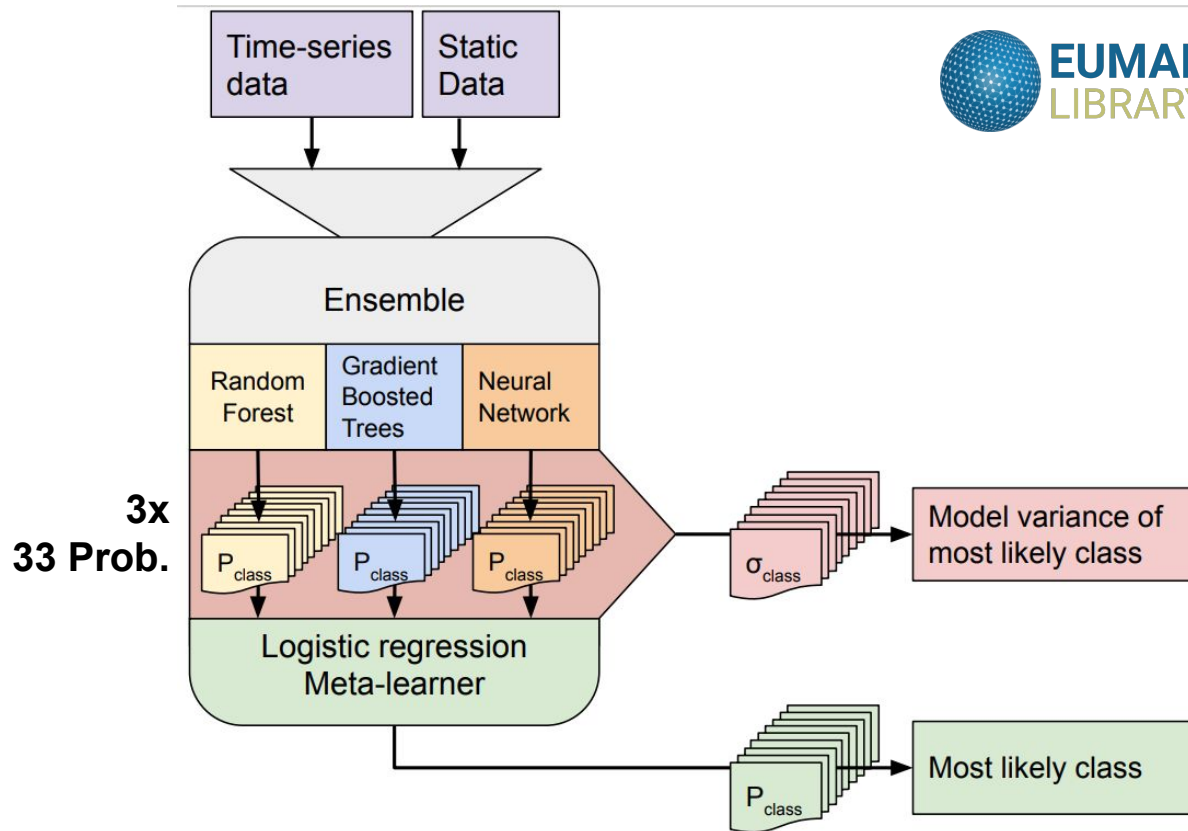
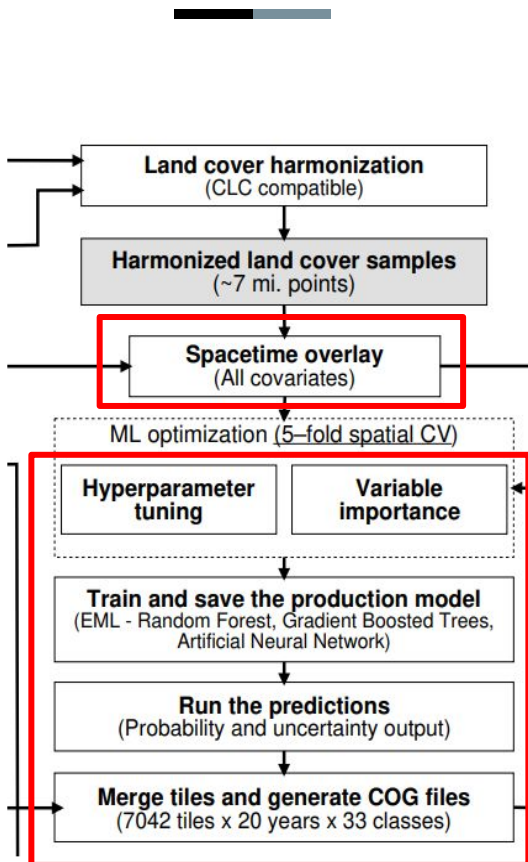
Ensemble Machine Learning workflow



Input/Intermediate Data Methods Output Data

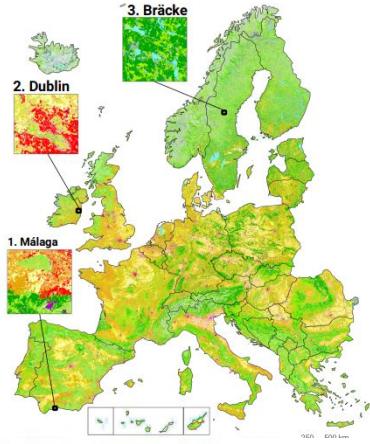


Ensemble Machine Learning workflow

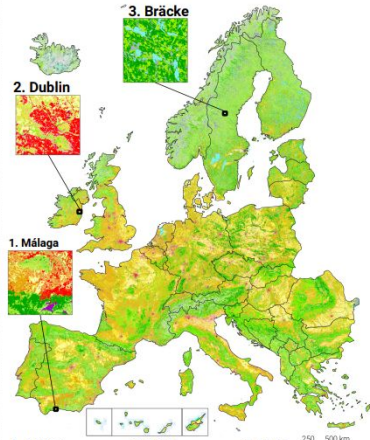


30-m Land use / land cover mapping

a. Dominant LULC - 2000

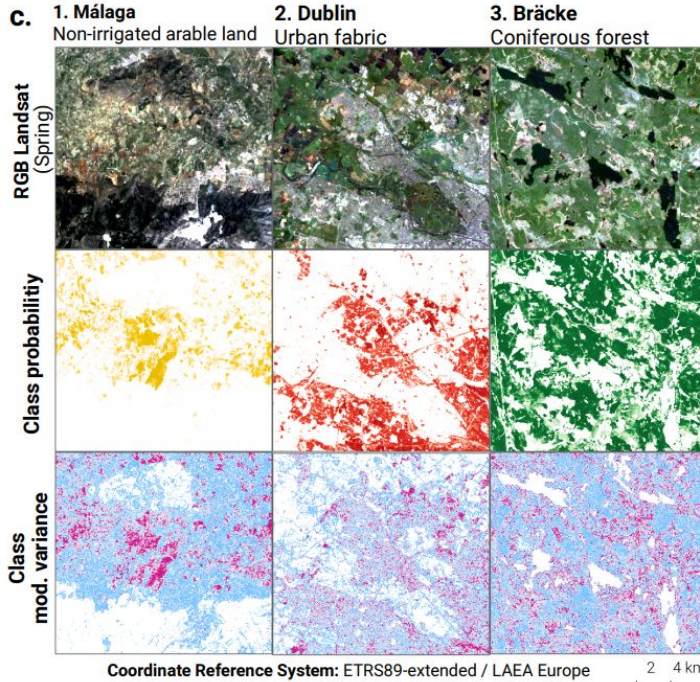


b. Dominant LULC - 2019

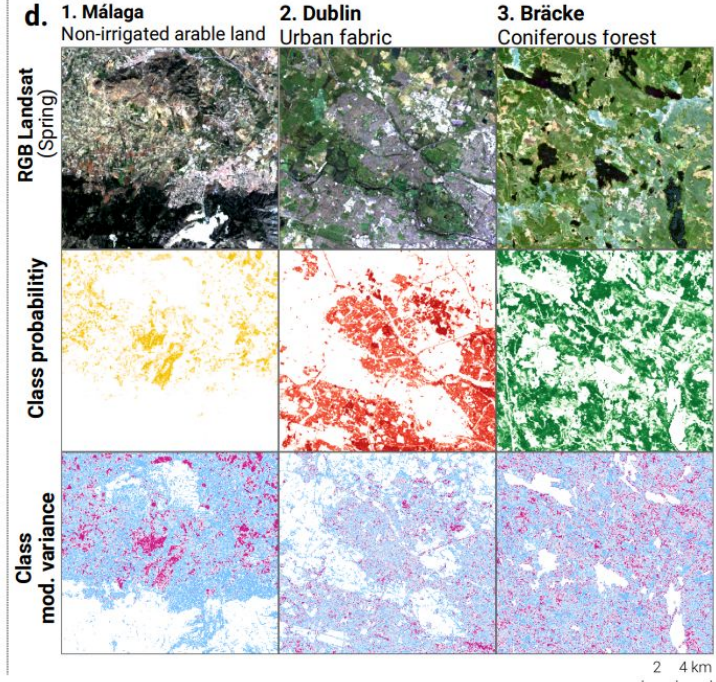


33 CLC classes (2000—2020)

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

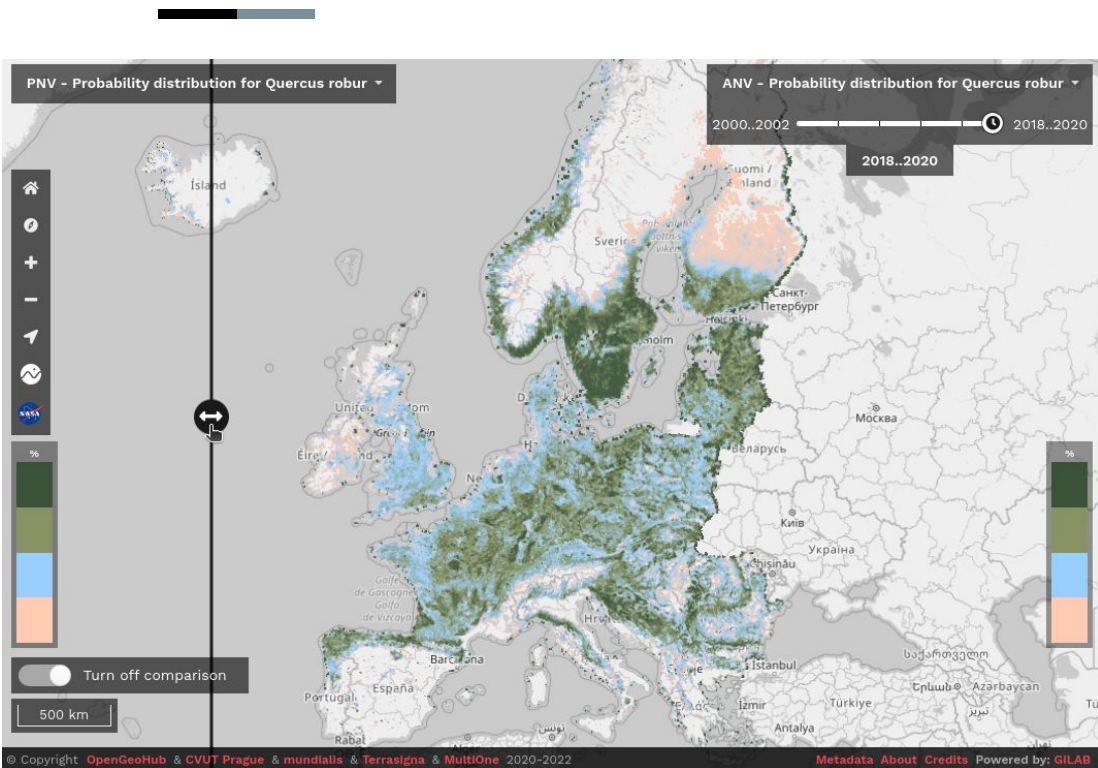


- | | | |
|--|-----------------------------------|-----------------------------|
| Urban fabric | Green urban areas | Pastures |
| Road and rail networks and associated land | Non-irrigated arable land | Broad-leaved forest |
| Port areas | Permanently irrigated arable land | Coniferous forest |
| Airports | Rice fields | Natural grasslands |
| Mineral extraction sites | Vineyards | Moors and heathland |
| Dump sites | Fruit trees and berry plantations | Sclerophyllous vegetation |
| Construction sites | Olive groves | Transitional woodland-shrub |



- | | | | |
|-----------------------------|-----------------|-------------------|------|
| Beaches, dunes, sands | Water courses | 10% | 100% |
| Bare rocks | Water bodies | Class probability | |
| Sparsely vegetated areas | Coastal lagoons | Model variance | |
| Burnt areas | Estuaries | | |
| Glaciers and perpetual snow | Sea and ocean | 1% | 10% |
| Inland wetlands | | Model variance | |
| Maritime wetlands | | | |

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

PeerJ

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

Camilo Bonaventura^{1*}, Tamás Hengl¹, Johannes Hengl¹, László Pásztor¹, Martin N. Wright¹, Martin Herold¹ and Sjoerd de Bruin¹

¹Institute of Geo-Information Science and Remote Sensing, Wageningen University and Research, Wageningen, The Netherlands
²Department of Geography, University of Vienna, Vienna, Austria
³Faculty of Forestry, University of Münster, Münster, Germany
⁴Faculty Institute for Forestry Research and Education, 1000, Vienna, Austria
⁵Centre for Forest, Wetland, and Coastal Ecosystems, 1000, Vienna, Austria
⁶Forest, 1, Vienna, Austria
⁷Forest, 1, Vienna, Austria
⁸Forest, 1, Vienna, Austria
⁹Forest, 1, Vienna, Austria
¹⁰Forest, 1, Vienna, Austria
¹¹Forest, 1, Vienna, Austria
¹²Forest, 1, Vienna, Austria
¹³Forest, 1, Vienna, Austria
¹⁴Forest, 1, Vienna, Austria
¹⁵Forest, 1, Vienna, Austria
¹⁶Forest, 1, Vienna, Austria

ABSTRACT

This article describes a data-driven framework based on spatiotemporal machine learning to predict distributions maps for 16 tree species (Larix laricina Mill., Cedrus deodara Mill., Pinus sylvestris L., Picea abies (L.) Mill., Pinus peuceletii Mill., Pinus nigra (L.) Arnold, Pinus pinaster Ait., Pinus peuceletii L., Pinus sylvestris L., Quercus robur L., Quercus ilex L., Quercus pubescens L., Quercus agrifolia L., and Quercus petraea L.) at high spatial resolution (30 m). Time series data for a total of three million of pixels was used to train different algorithms: random forest, gradient boosted trees, generalized linear models, k nearest neighbors, CART and an artificial neural network. A stack of 305 coarse and high resolution covariates representing spatial heterogeneity, different topographic conditions and biotic competition was used as predictors for realized distributions, while potential distributions was modeled with environment predictors only. Light and competing tree were used to select the three best algorithms to train and train an ensemble model based on stacking such a logistic regression as a meta-classifier. An ensemble model was trained for each species probability and model uncertainty maps of realized distributions were produced for each species using a time window of 4 years for a total of six distribution maps per species, while for potential distributions each one map per species was produced. Results of model cross-validation show that the ensemble model consistently outperformed or performed as good as the best individual models in both potential distribution maps and realized distributions. The ensemble model shows higher predictive performance with potential distribution models, achieving higher predictive performance (F1S = 0.896, F1L_{max} = 0.873) than realized distribution maps on average (F1S = 0.874, F1L_{max} = 0.870). Ensemble results for Q. robur achieved the best performance in both potential (F1S = 0.906, F1L_{max} = 0.870) and realized (F1S = 0.898, F1L_{max} = 0.849) distributions, while P. sylvestris (F1S = 0.731, 0.745, F1L_{max} = 0.670, 0.670), respectively, for potential and realized distribution maps.

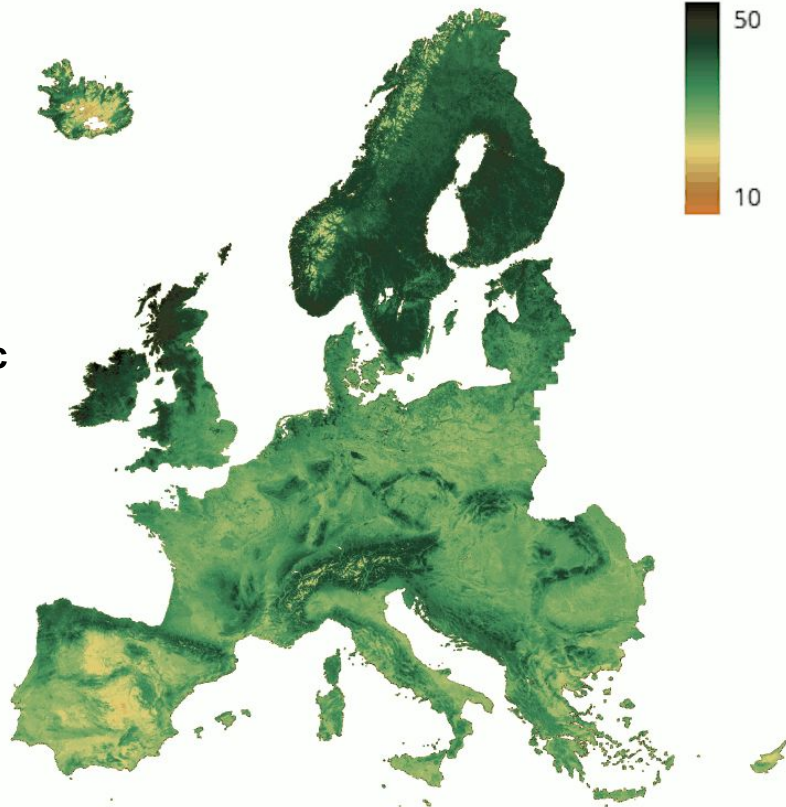
KEYWORDS

Machine Learning, Spatiotemporal, Forest, Tree Species, Distribution, Europe, 2000-2020, Mapping, Potential, Realized, Distributions, Spatiotemporal, Machine Learning, Forest, Tree Species, Distribution, Europe, 2000-2020, Mapping, Potential, Realized, Distributions, Spatiotemporal, Machine Learning

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

2020

Soil organic carbon



+ 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

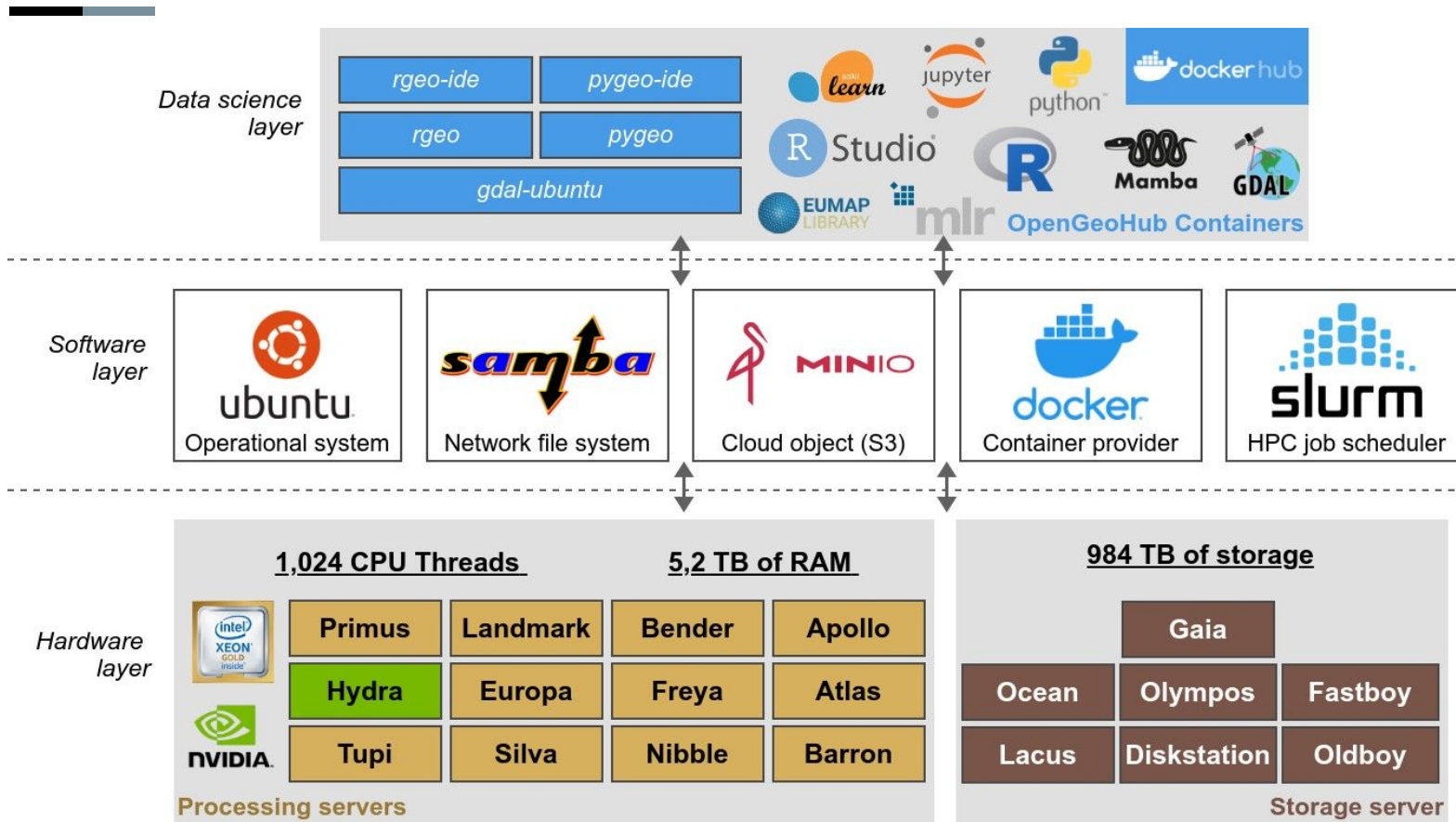
Dynamic soil information at farm scale based on Machine Learning and EO data: building an Open Soil Data Cube for Europe

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



Using Open Source to crunch big EO data



Versioned and fully reproducible development environment



opengeohub/rgeo-ide ☆

By [opengeohub](#) • Updated a month ago

Rstudio and R ($\geq 4.1.1$) working with GDAL ($\geq 3.1.4$) and several geospatial and ML packages

Container



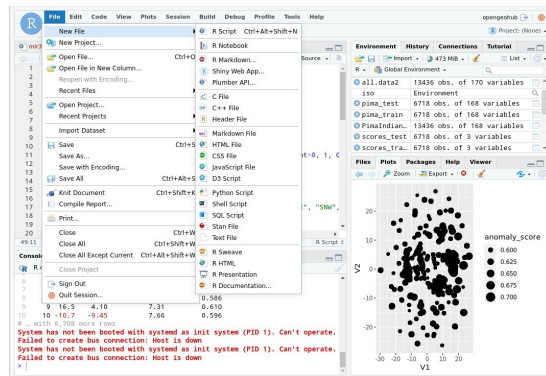
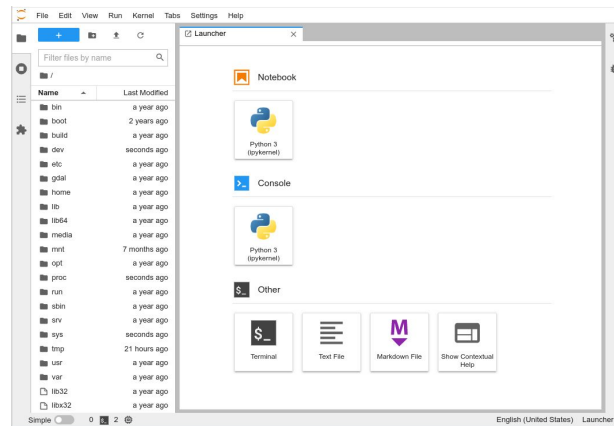
opengeohub/pygeo-ide ☆

By [opengeohub](#) • Updated a month ago

JupyterLab and Python ($\geq 3.8.6$) working with GDAL ($\geq 3.1.4$) and several geospatial and ML packages

Container

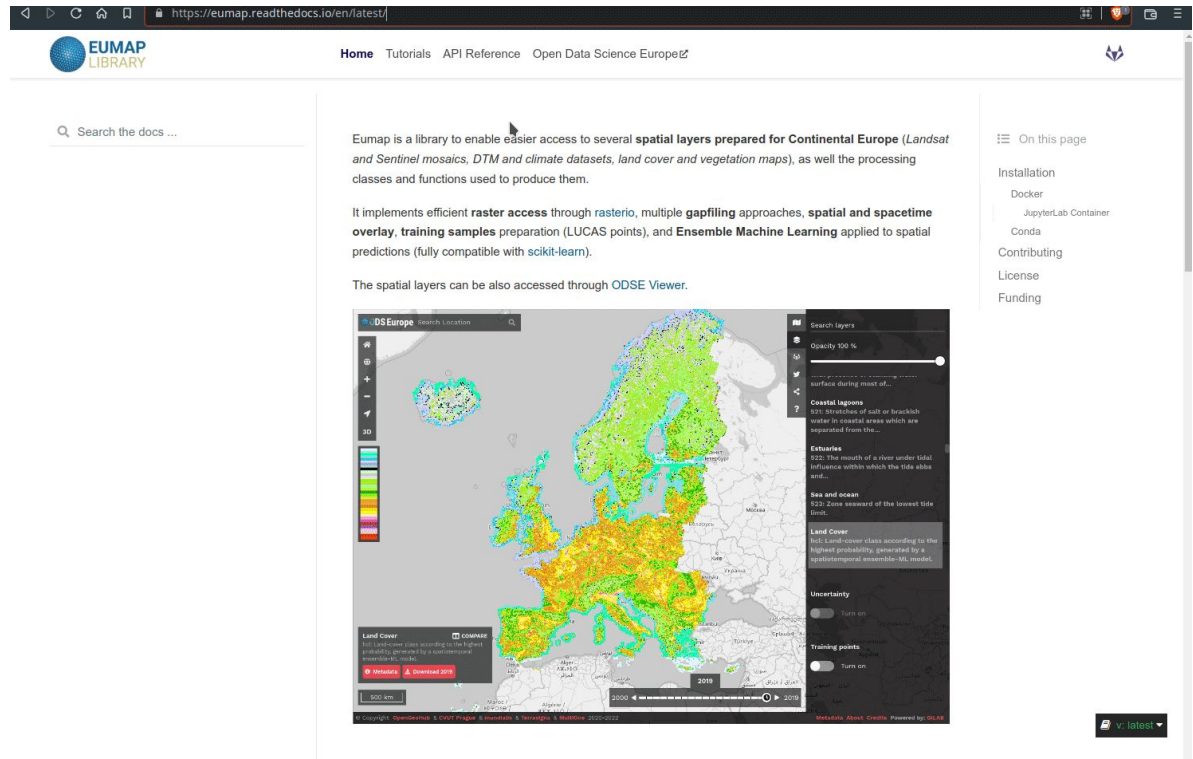
<https://hub.docker.com/r/opengeohub/>



Using Open Source to crunch big EO data

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



Home Tutorials API Reference Open Data Science Europe

Search the docs ...

Eumap is a library to enable easier access to several **spatial layers prepared for Continental Europe** (*Landsat and Sentinel mosaics, DTM and climate datasets, land cover and vegetation maps*), as well the processing classes and functions used to produce them.

It implements efficient **raster access** through rasterio, multiple **gapfilling** approaches, **spatial and spacetime overlay**, **training samples** preparation (LUCAS points), and **Ensemble Machine Learning** applied to spatial predictions (fully compatible with scikit-learn).

The spatial layers can be also accessed through ODSE Viewer.

On this page

- Installation
- Docker
- JupyterLab Container
- Conda
- Contributing
- License
- Funding

Land Cover

Coastal lagoons

Sea and ocean

Land Cover

Uncertainty

Training points

© 2020, CC BY-SA. EUMAP LIBRARY is licensed under a Creative Commons Attribution 4.0 International License. Website About Privacy Powered by OGD

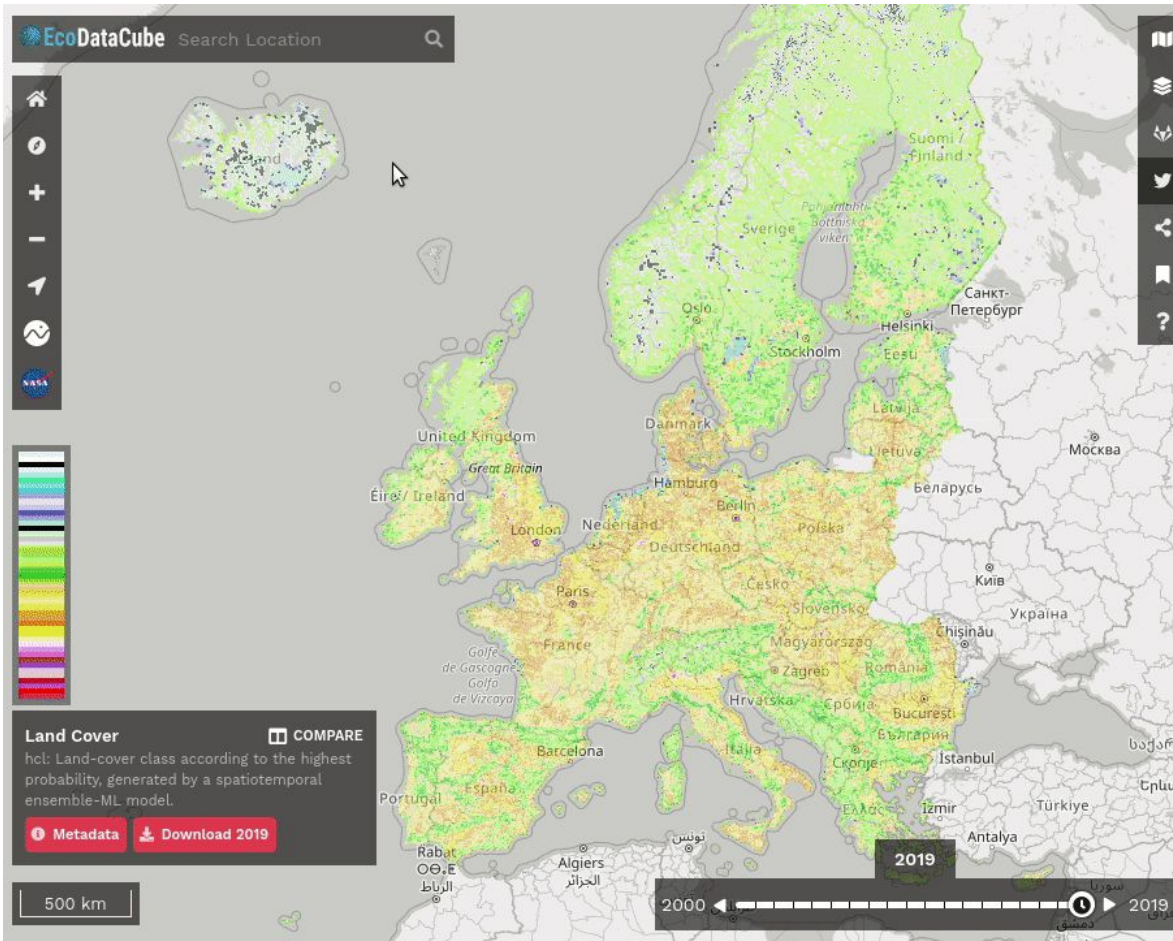
v: latest

Data portal <https://ecodatacube.eu>



Open Geo HUB

Connect • Create • Share • Repeat



Geo-harmonizer Retweeted

Mundialis GmbH & Co.
@MundialisInfo

The second ODSE conference day just starts (streamed online): opendatascience.eu/workshop-2022/

Keynotes and talks about #visualization, #VR, #opensource, #bigdata, #forest, #humanitarian, #HaDea, #airpollution, #radonhazard and more... Please tune in! @OSGeo @HarmonizerGeo #geodata

Open Data Science

WORKSHOP
13 - 16 June 2022
Magyar, Czech Republic

Jun 16, 2022

Geo-harmonizer Retweeted

Angela Baker
@bakerangela

Delighted to be at the @HarmonizerGeo #ODSE workshop today to share @EuroGeographics, #OpenMapsForEurope #OpenData @cines_eu @EU_HaDEA #CEFTelecom #ODSE

Jun 16, 2022

Geo-harmonizer Retweeted

Mundialis GmbH & Co.

A complete comprehensive raster database with all layers imported and documented

STAC catalog:

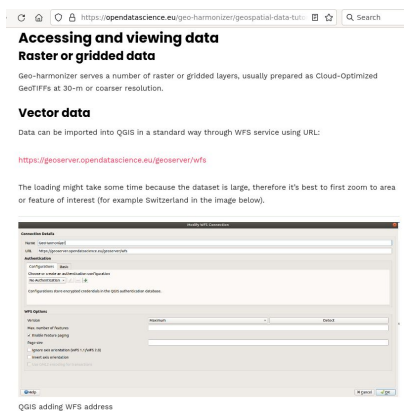
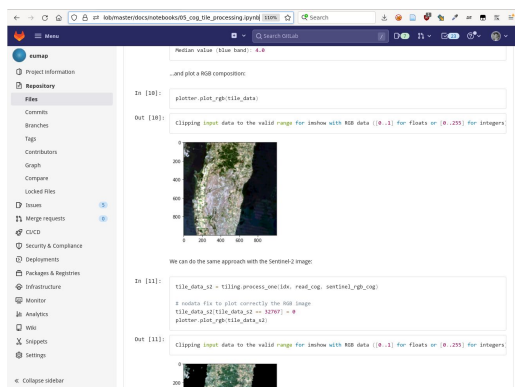
<https://stac.EcoDataCube.eu>

Online tutorials:

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

Jupyter notebooks:

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



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 establishing 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@opengeohub.org



<https://opengeohub.org>



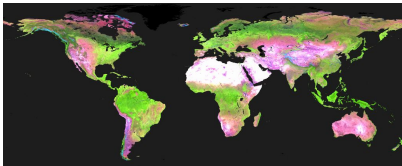
OpenGeoHUB

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 gridding capabilities
- Different strategies to generate composites (e.g. best available pixel, spectral temporal metrics)
- Free software under GNU License v.3



Working properly with Landsat data



GLAD Landsat ARD

(Level 4 product)

Level 4
(Model output)

FORCE
(All levels)

Level 3
(Temporal composites and gridded data)

Level 2
(Atmospheric correction)

Level 1
(Radiometrically calibrated and georectified data)



(Level 1 and level 2 products)

Using Open Source to crunch big EO data



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



For every tile

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



Keys to increasing computing efficiency:

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

Using Open Source to crunch big EO data



23,116 tiles

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

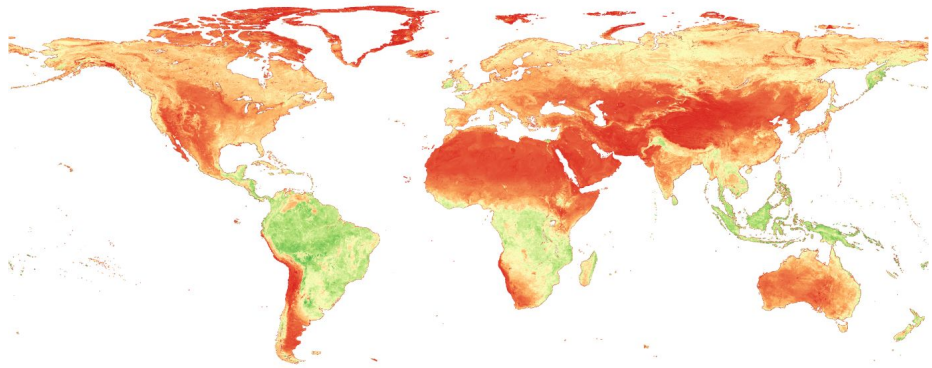


Reading (http)

Writing (S3)



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



For every tile

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

