CC BY SA

Time Series Machine Learning Predictions

Moving from modeling to production environment

Leandro Parente

leandro.parente@opengeohub.org

https://opengeohub.org

https://shorturl.at/CIIsT

Outline

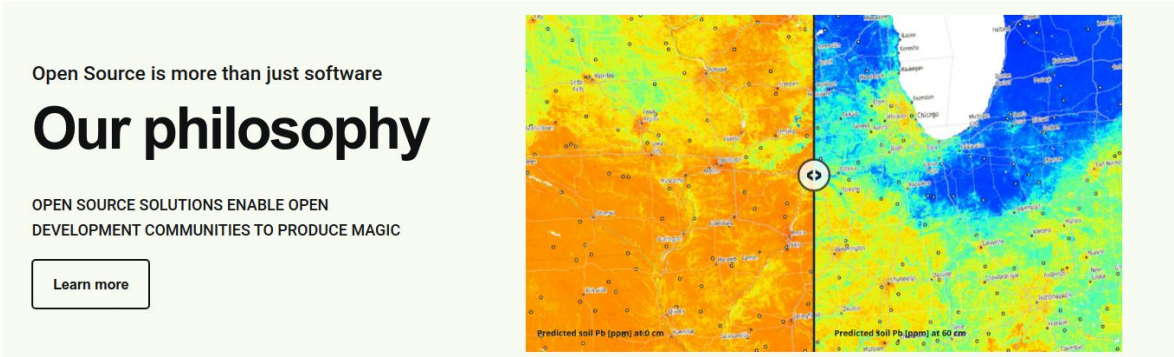
- OpenGeoHub Foundation & core activities
- ML pipeline for spatial prediction
- Analysis-Ready & Cloud-Optimized (ARCO) data
- Landsat cloud-free aggregates
- Feature engineering & selection
- Spatial prediction challenges
- High-throughput data processing infrastructure
- Global Pasture Watch mapping
  (real world application)

Who we are?

OpenGeoHub
Connect • Create • Share • Repeat

Open data culture for open development communities

OpenGeoHub is an independent, not-for-profit research foundation that utilizes machine learning and geo-computations to produce and share open geographical and geo-scientific data, develop open source software, and provide training and support to researchers and decision-makers



Who we are?

OpenGeoHub
Connect • Create • Share • Repeat

Open data culture for open development communities

OpenGeoHub is an independent, not-for-profit research foundation that utilizes machine learning and geo-computations to produce and share open geographical and geo-scientific data, develop open source software, and provide training and support to researchers and decision-makers

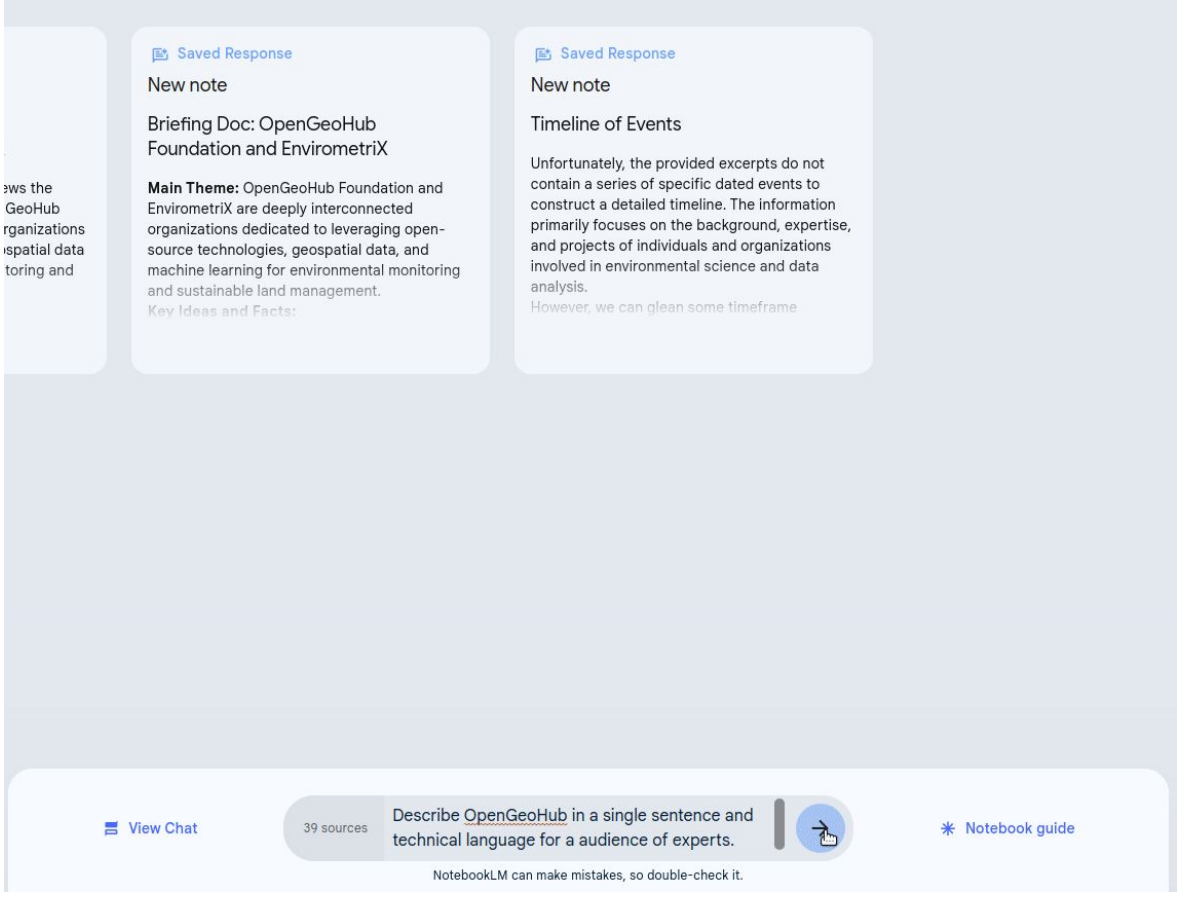
Dr. Tomislav Hengl

Dr. Ichsani Wheeler

Prof. Dr. Edzer Pebesma

Who we are?

OpenGeoHUB
Connect • Create • Share • Repeat



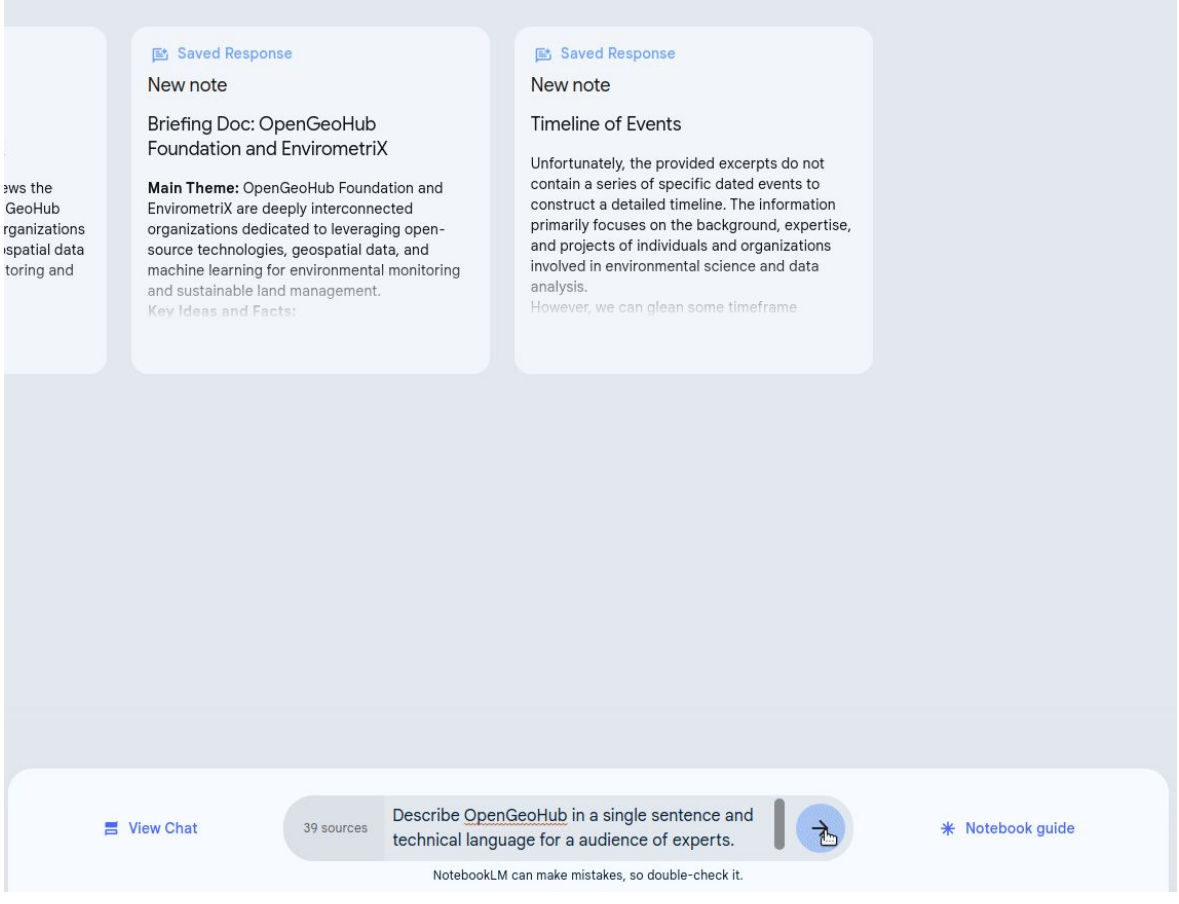
NotebookLM

Who we are?

[Deep Dive Podcast](#)

(15 min.)

NotebookLM



Which are the on-going project?

AI 4 Soil Health

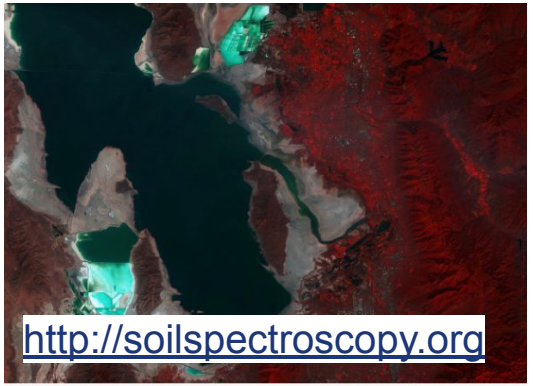
https://ai4soilhealth.eu

Global Pasture Watch

Mapping & monitoring Global Grasslands and Livestock

https://landcarbonlab.org

http://earthmonitor.org

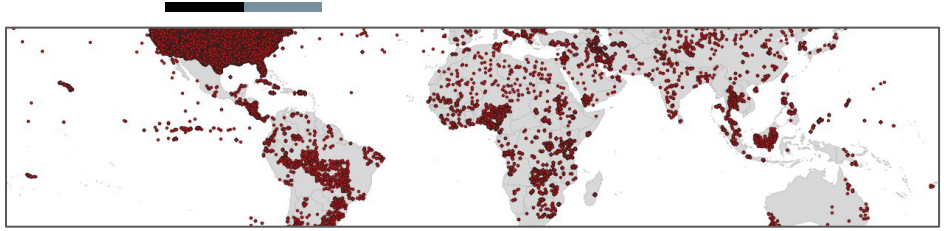


http://soilspectroscopy.org

Soil Spectroscopy for Global Good

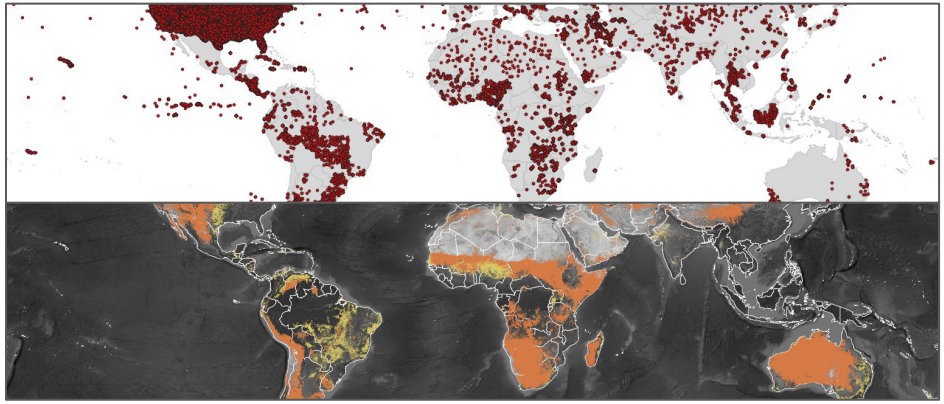
GLOBAL

Core Activities



Data organization & harmonization
(raster maps & vector points/polygons)

Core Activities



Data organization & harmonization
(raster maps & vector points/polygons)

Data processing & production
(mostly raster maps)

Core Activities

Data organization & harmonization
(raster maps & vector points/polygons)

Data processing & production
(mostly raster maps)

Open source/Open data solutions
(STAC, packages & computational notebooks)

Core Activities

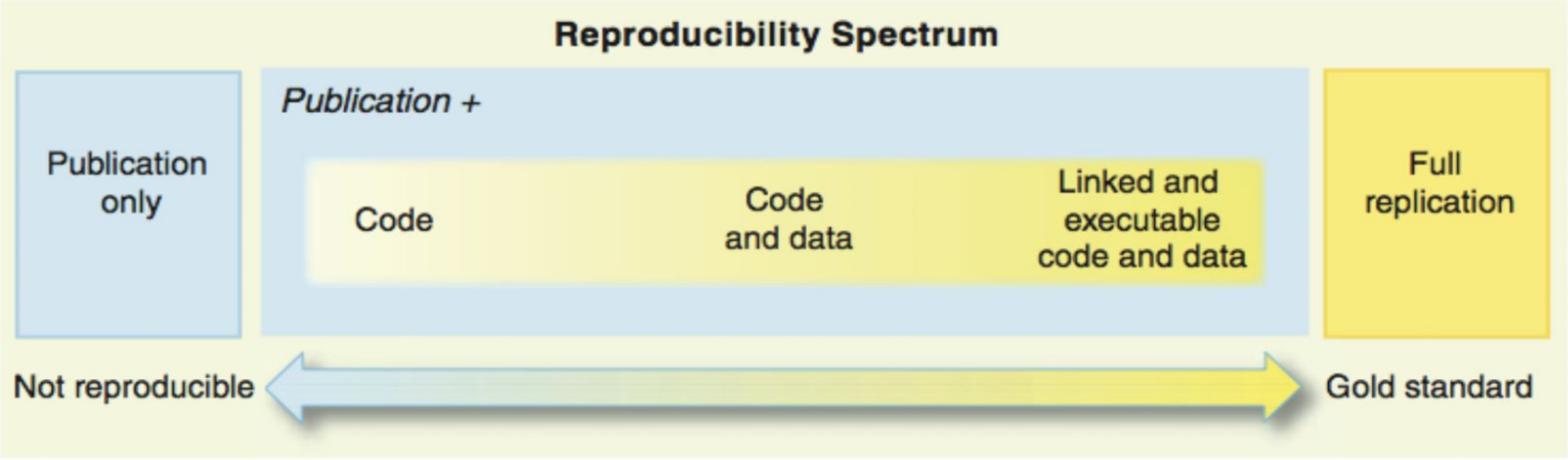
Data organization & harmonization
(raster maps & vector points/polygons)

Data processing & production
(mostly raster maps)

Open source/Open data solutions
(STAC, packages & computational notebooks)

Communication & dissemination
(Video recording, hackathons & summer schools)

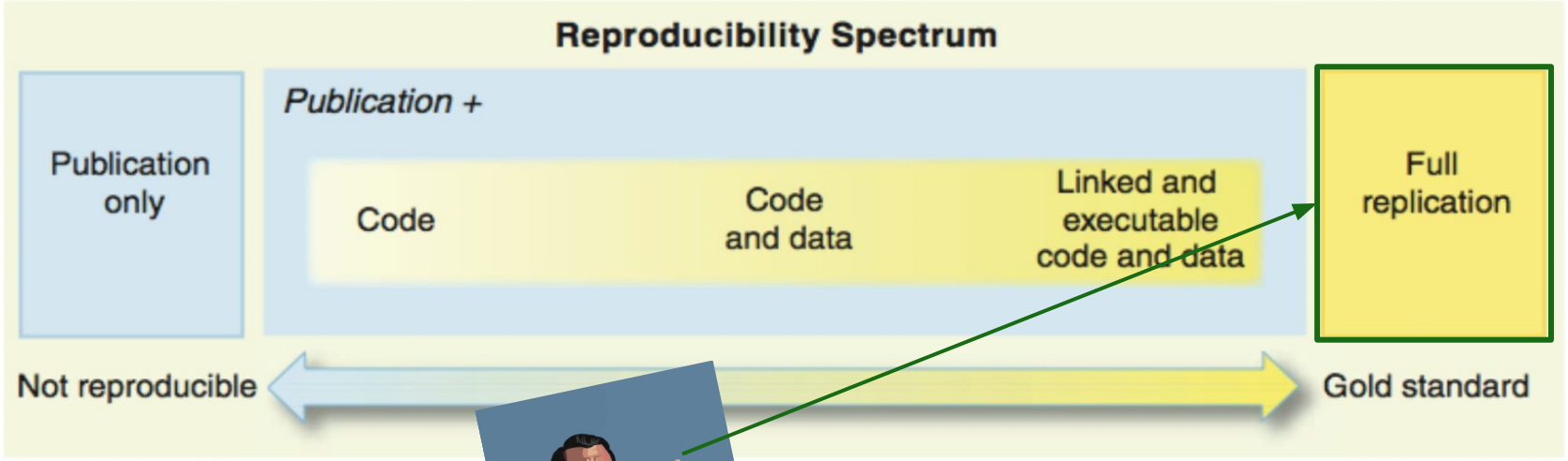
Aiming full replication



[PENG R., 2011]

Cloud-optimized and analysis-ready (ARCO),
Open catalog (STAC),
Computational notebooks, and
Software libraries

Aiming full replication



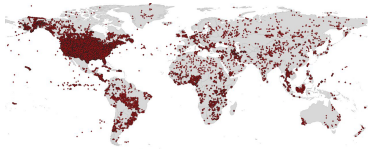
[PENG R., 2011]

Cloud-optimized and analysis-ready (ARCO),
Open catalog (STAC),
Computational notebooks, and
Software libraries

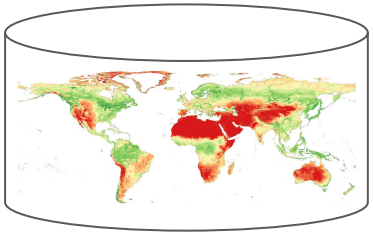
So what?

How to implement it in practices?

ML pipeline for spatial prediction



Training samples
(in-situ points)

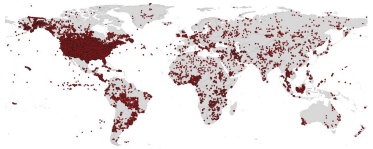


ARCO datasets
(harmonized, gapfilled
consistent layers)

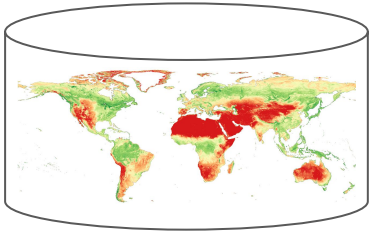
Spatial / Spacetime
overlay

Classification /
Regression matrix

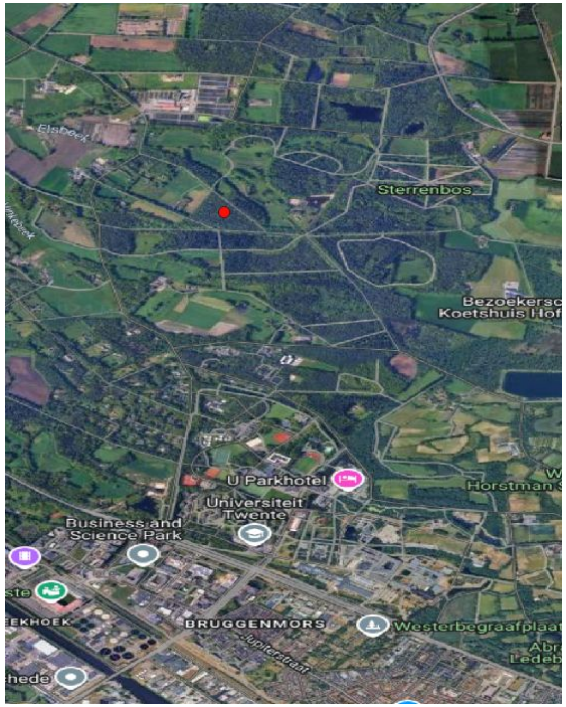
ML pipeline for spatial prediction

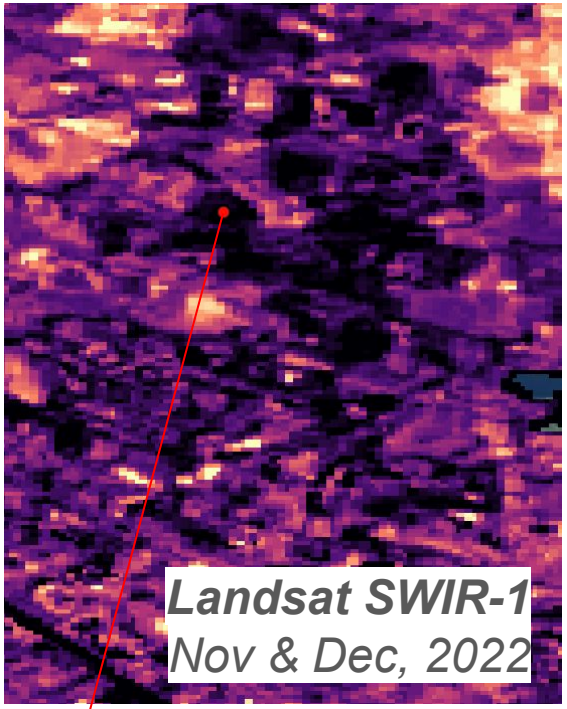


Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)



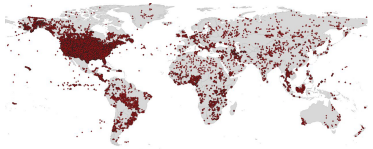


Pixel value: **0.108**

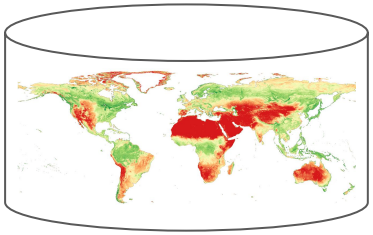
Spatial / Spacetime overlay

Classification / Regression matrix

ML pipeline for spatial prediction



Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)

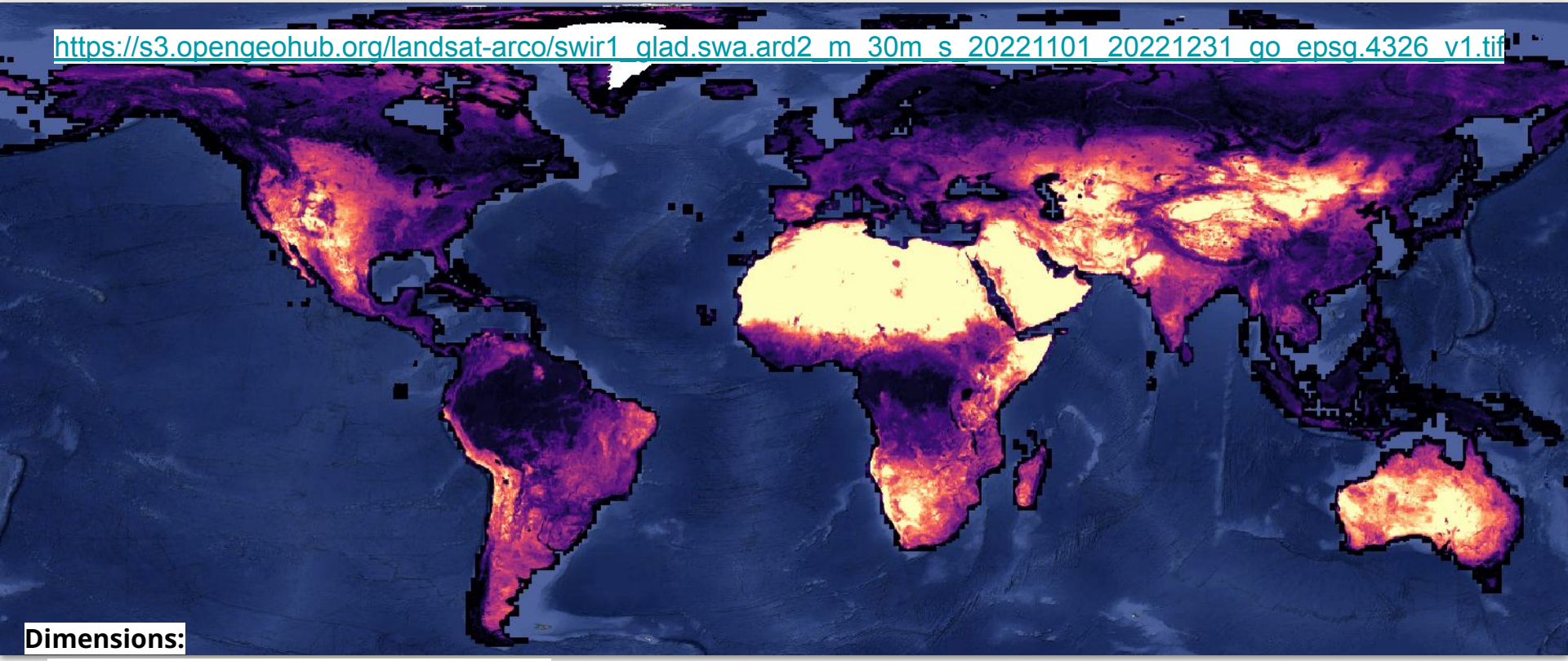
	lat_20m	lon_20m	h_canopy	year	swir1_glad.SeasConv.ard2_m_30m_s_0101_0228_go_epsg.4326_v20230908
38	46.768604	0.744858	11.478378	2019.0	53.0
51	10.877570	-0.049387	7.083252	2019.0	87.0
55	14.221970	-0.457440	5.034332	2019.0	101.0
75	34.839863	-0.818398	3.491333	2019.0	48.0
87	47.375198	-0.033251	11.574158	2019.0	55.0
...	...	...	...	...	...
38218267	5.002793	47.892220	5.887756	2022.0	88.0
38218275	5.134330	47.905411	3.733047	2022.0	101.0
38218283	5.024637	47.086670	2.669090	2022.0	121.0
38218291	5.114291	47.932446	2.675583	2022.0	101.0
38218299	5.207586	47.912750	NaN	2022.0	117.0

Spatial / Spacetime overlay

Classification / Regression matrix

Raster maps become **columns** & individual samples **rows**

Analysis-Ready & Cloud-Optimized (ARCO)

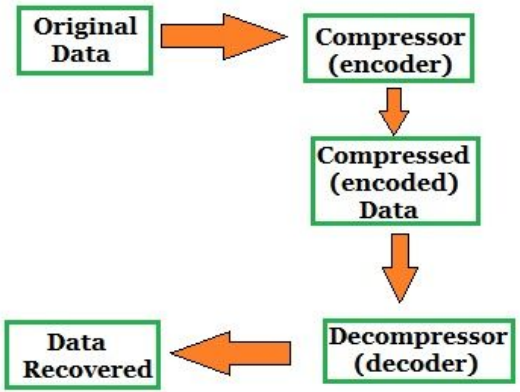


Dimensions:

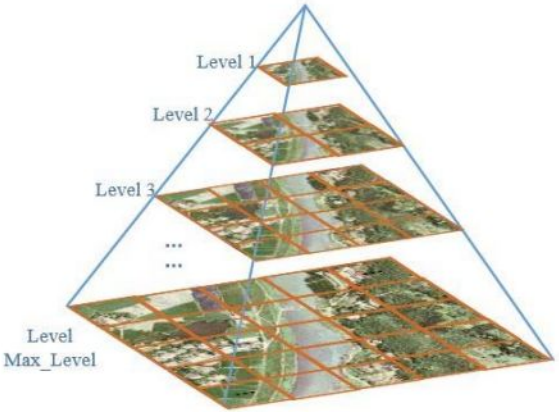
- Image size: 1,440,004 (H), 560,004 (V)
- Filesize: 134 GB (with compression)
- Format: Cloud-Optimized GeoTIFF (COG)

Cloud-Optimized GeoTIFF (COG)

A COG is a regular GeoTIFF file, aimed at being hosted on a HTTP file server, with an internal organization that enables more efficient workflows on the cloud.



Lossless compression
(DEFLATE, LZW)



Tilling and overviews
(pyramid structure)

HTTP Get Range requests
(Accept-Ranges: bytes)

Source: Cloud Optimized GeoTIFF in depth and Guo et al., 2016

Cloud-Optimized GeoTIFF (COG)

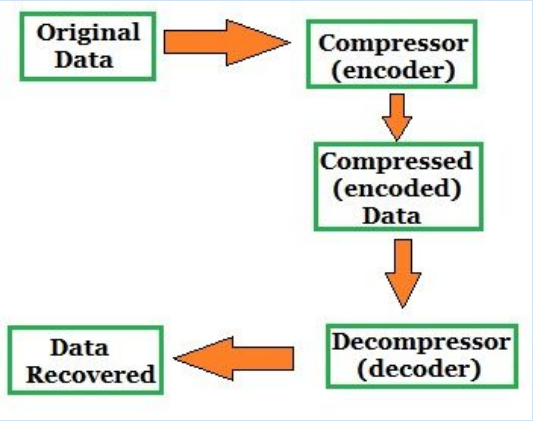
OpenGeoHUB
Connect • Create • Share • Repeat

Cloud Object Storage

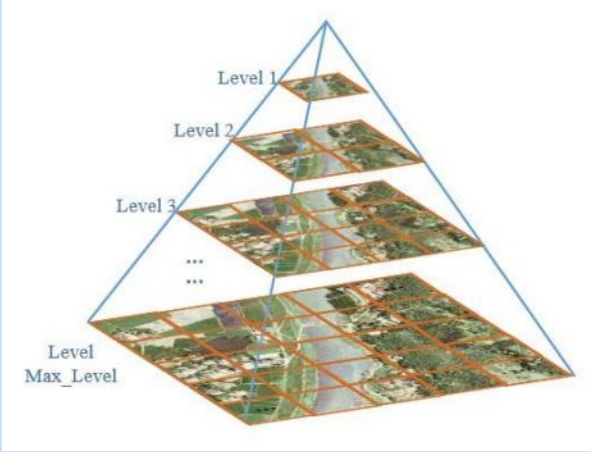
Unique Id identifier (URL)

amazon S3

Google Cloud Storage



Lossless compression
(DEFLATE, LZW)



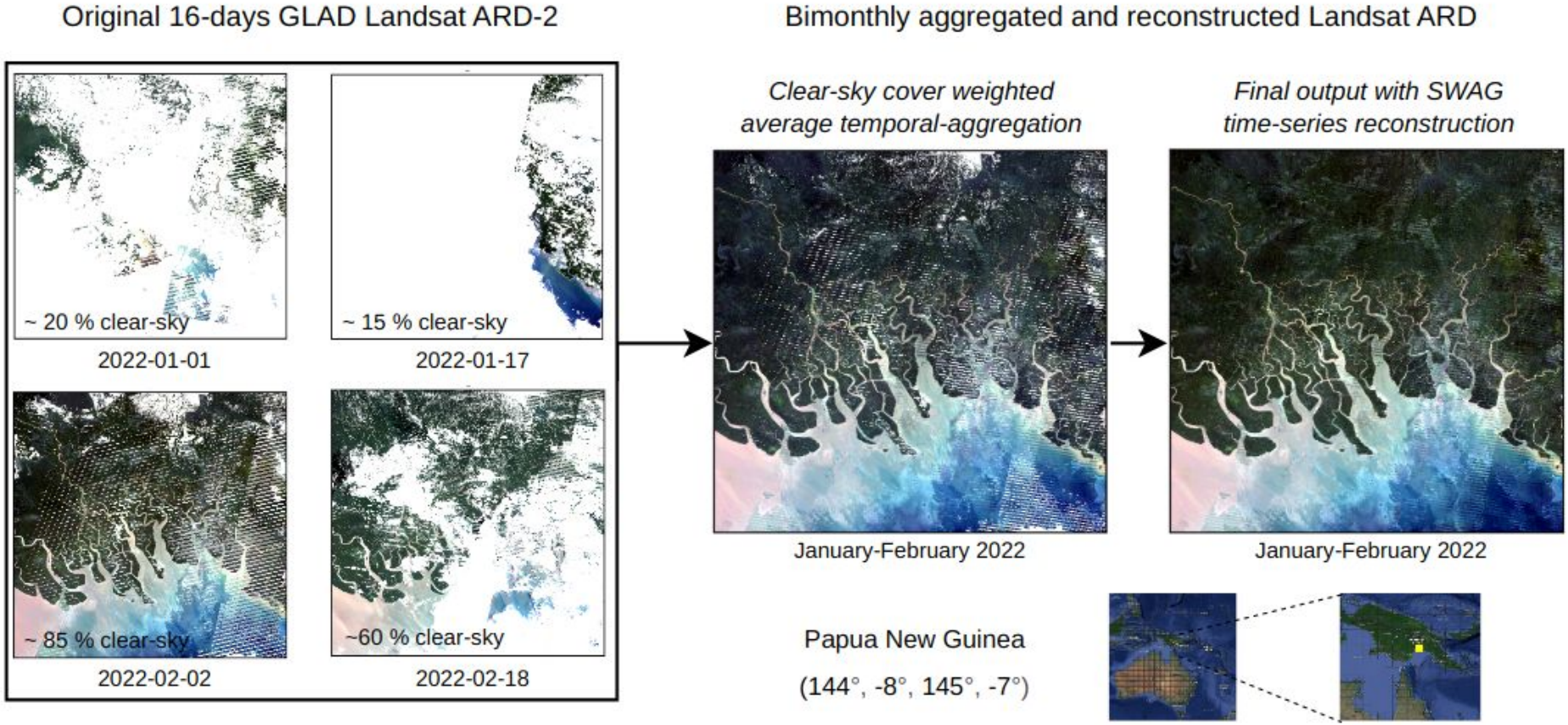
Tilling and overviews
(pyramid structure)

HTTP

HTTP Get Range requests
(Accept-Ranges: bytes)

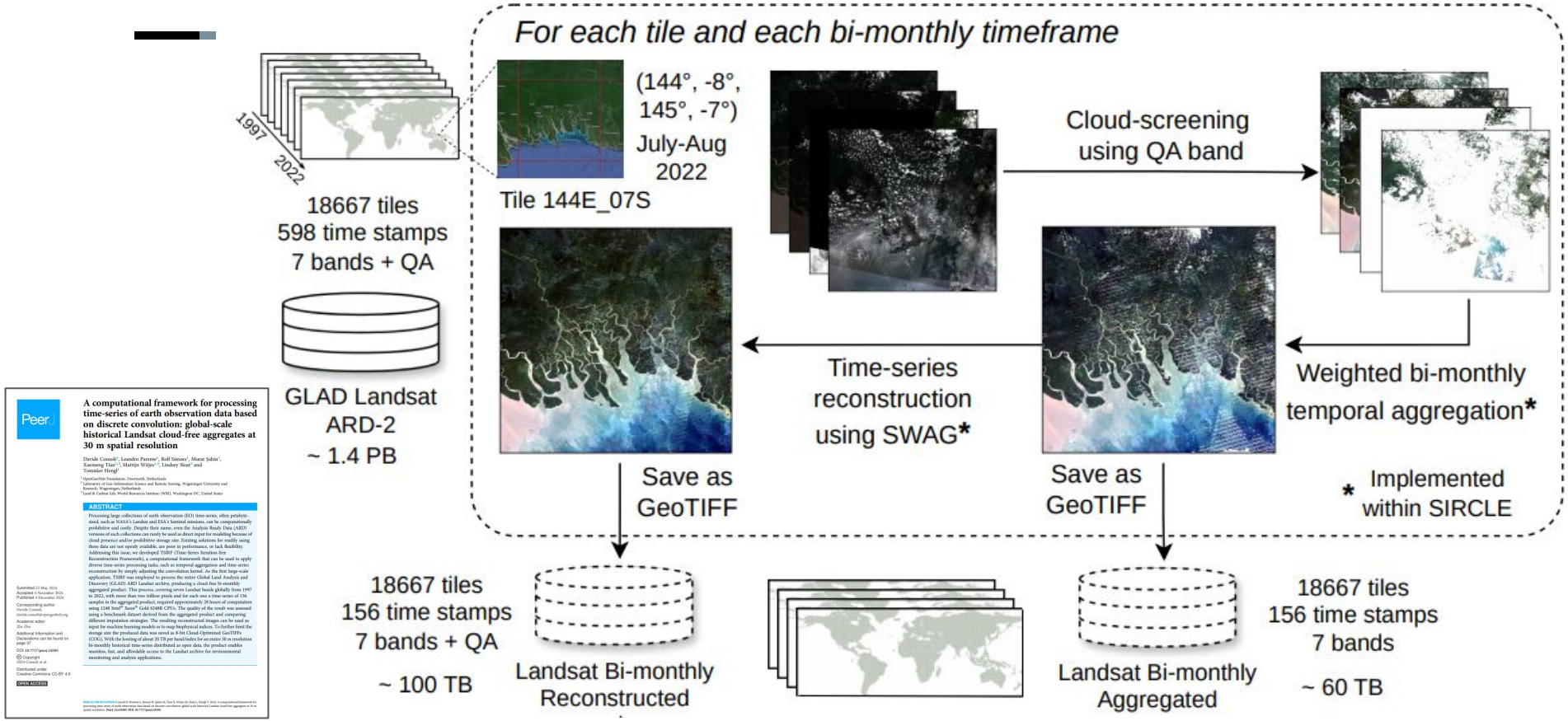
Source: Cloud Optimized GeoTIFF in depth and Guo et al., 2016

Landsat cloud-free aggregates



https://doi.org/10.7717/peerj.18585

Landsat cloud-free aggregates



PeerJ

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

ABSTRACT

https://doi.org/10.7717/peerj.18585

Landsat cloud-free aggregates

OpenGeoHUB
Connect • Create • Share • Repeat

OpenLandMap

Cloud-free reconstructed Landsat bimonthly spectral reflectance bands (1997-2022)

Source | Share | Language: English

in OpenLandMap STAC | Up | Browse

Description

Global, cloud-free, and reconstructed historical Landsat spectral bands are provided at a spatial resolution of 30 meters and a temporal resolution of bimonthly intervals (six images per year) spanning from 1997 to 2022. The dataset was generated using the GLAD Landsat Analysis Ready Data version 2 as the primary input for temporal aggregation and the imputation of missing values. Detailed information on the data processing methodology in Consoli et al., 2024. Currently the collection includes only the SWIR1 band, with plans to incorporate the other bands in future updates.

landsat | spectral bands | cloud-free | time-series

License CC-BY-4.0

Temporal Extent 1/1/1997, 12:00:00 AM UTC - 12/31/2022, 12:00:00 AM UTC

Items 156

Ascending | Descending

landsat_glad.swa.ard2_bimonthly_19970101_19970228

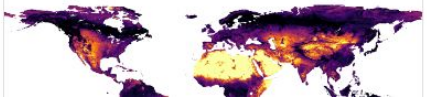
COG

1/1/1997, 12:00:00 AM UTC - 2/28/1997, 12:00:00 AM UTC

landsat_glad.swa.ard2_bimonthly_20100101_20100228

COG

1/1/2010, 12:00:00 AM UTC - 2/28/2010, 12:00:00 AM UTC



https://stac.openlandmap.org

Landsat cloud-free aggregates

QGIS Python Plugins Repository

Download latest

Oemc Plugin

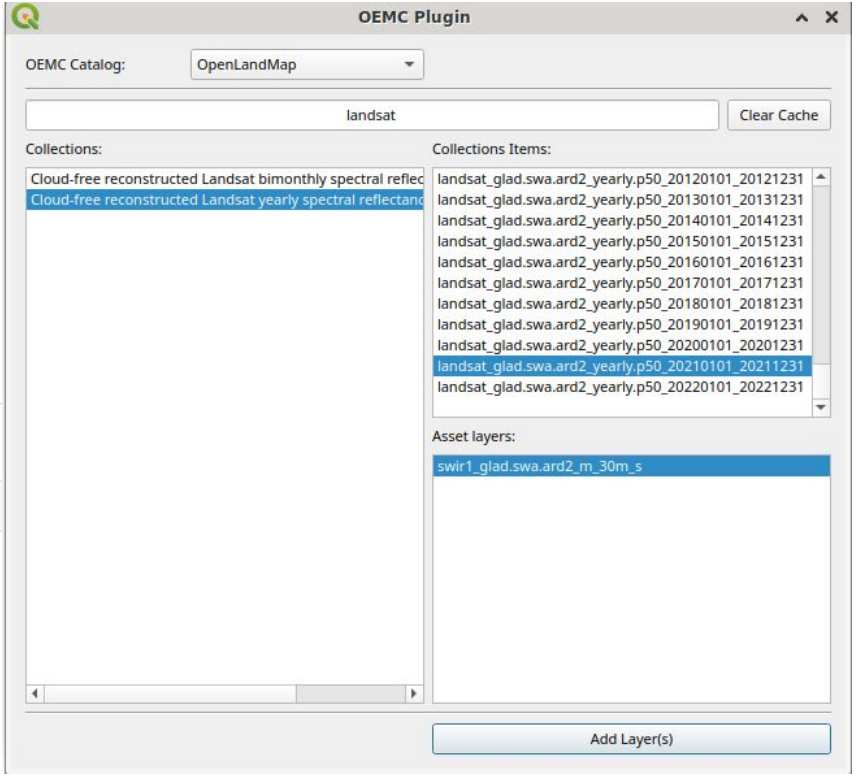
Plugin ID: 3221

(3) votes

This plugin provides easy access to OEMC STAC catalog

About | Details | Versions

Version	Experimental	Min QGIS version	Max QGIS version	Downloads	Uploaded by
0.2	no	3.0.0	3.99.0	368	muratsahin
0.1	no	3.0.0	3.99.0	649	muratsahin



https://github.com/Open-Earth-Monitor/oemc-qgis-plugin

Global land cover mapping

OpenLandMap

GLC_FCS30D annual land land-cover dynamic monitoring product

Source | Share | Language: English

in OpenLandMap STAC | Up | Browse

Description

GLC_FCS30D is the first global fine land cover dynamic product at a 30-meter resolution that adopts continuous change detection. It utilizes a refined classification system containing 35 land-cover categories and covers the time span from 1985 to 2022. Before the year 2000, the update cycle was every 5 years, while after 2000, it is updated annually. In specific, it developed by combining the continuous change detection method, local adaptive updating models and the spatiotemporal optimization algorithm from dense time-series Landsat imagery, and was validated to achieve an overall accuracy of 80.88% (±0.27%) for the basic classification system 10 major land-cover types) and 73.24% (±0.30%) for the LCCS level-1 validation system (17 LCCS land-cover types).

glc_fcs30 | global land cover | fine classification system | landsat | google earth engine

License CC-BY-4.0

Temporal Extent 1/1/1985, 12:00:00 AM UTC - 12/31/2022, 12:00:00 AM UTC

Items 26

Ascending | Descending

lc_glc.fcs30d_19850101_19851231

COG

1/1/1985, 12:00:00 AM UTC - 12/31/1985, 12:00:00 AM UTC

lc_glc.fcs30d_20100101_20101231

COG

1/1/2010, 12:00:00 AM UTC - 12/31/2010, 12:00:00 AM UTC

https://doi.org/10.5281/zenodo.8239305

Landsat cloud-free aggregates

OpenGeoHUB
Connect • Create • Share • Repeat

OPEN EARTH MONITOR

Analysis-ready in-situ data
Raw in-situ data

Open-Earth-Monitor Catalog

Source | Share | Language: English

Browse

Description

Spatio-Temporal Asset Catalog provided by Open-Earth-Monitor project.

https://stac.earthmonitor.org

Catalogs 4

Tiles | List | Ascending | Descending

Filter catalogs by title, description or keywords

Select keywords

FLUXNET: GHG fluxes dataset

Continuous GHG fluxes and quality flags. In most cases comprehensive of storage fluxes and footprint information. Most relevant variables in the dataset...

1996-01-01 0:00:00 UTC until present

OEMC Hackathon 2023: EU Land Cover Classification Dataset

Stratified sampling of the ground-truth data provided by LUCAS Survey and overlaid with 416 raster spatial layers. The target land cover considered level-3...

2006-01-01 00:00:00 UTC – 2018-12-31 00:00:00 UTC

OEMC Hackathon 2023: Global FAPAR Modeling Dataset

Monthly averaged FAPAR (Fraction of Absorbed Photosynthetically Active Radiation) values aggregated by each ground station and overlaid wit...

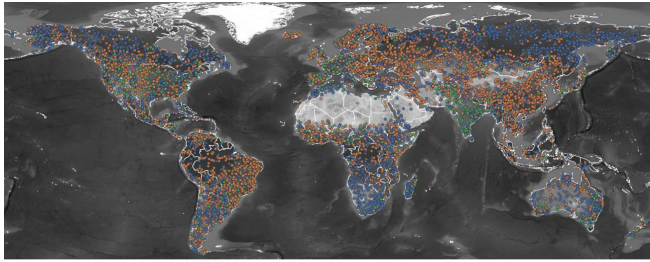
2014-01-01 00:00:00 UTC – 2020-12-31 00:00:00 UTC

Global Pasture Watch: Grassland classes reference samples

Reference samples (estabilished by Feature Space Coverage Sampling-FSCS) comprising 2.3M points visually classified using Very High Resolution image...

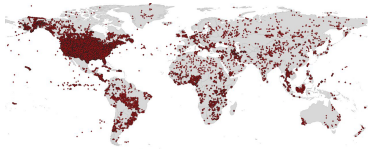
2000-04-01 00:00:00 UTC – 2022-12-31 00:00:00 UTC

Working in progress

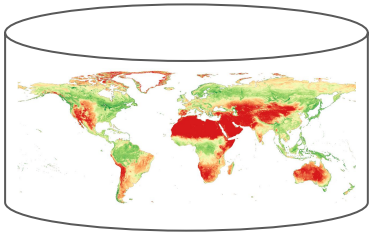


Powered by Open-Earth-Monitor Cyberinfrastructure (OEMC), a project that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement

ML pipeline for spatial prediction



Training samples
(in-situ points)

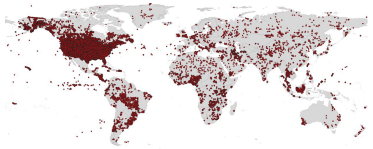


ARCO datasets
(harmonized, gapfilled
consistent layers)

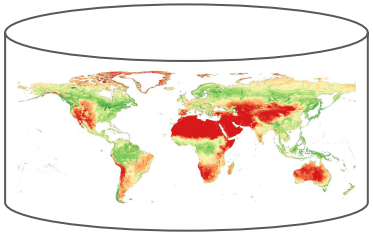
Spatial / Spacetime
overlay

Classification /
Regression matrix

ML pipeline for spatial prediction



Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)

Spatial / Spacetime overlay

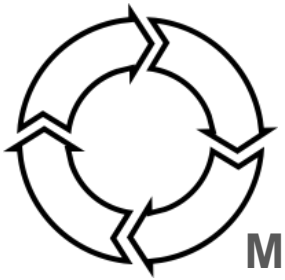
Classification / Regression matrix

Hyper-parameter tuning
(calibration & cross-validation)

Feature selection
(subsampling & cross-validation)

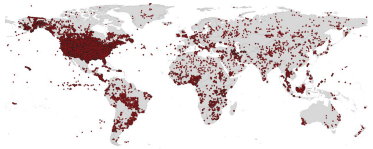
Final ML setup

ML training
(Single model or ensemble)

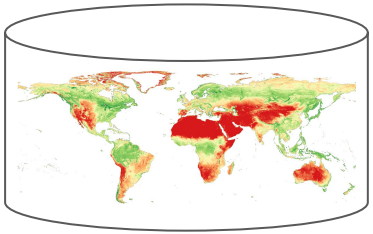


ML loop

ML pipeline for spatial prediction



Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)

Spatial / Spacetime overlay

Classification / Regression matrix

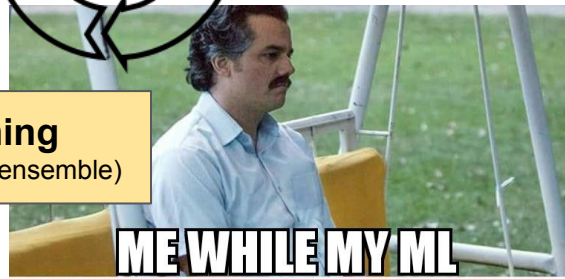
Hyper-parameter tuning
(calibration & cross-validation)

Feature selection
(subsampling & cross-validation)

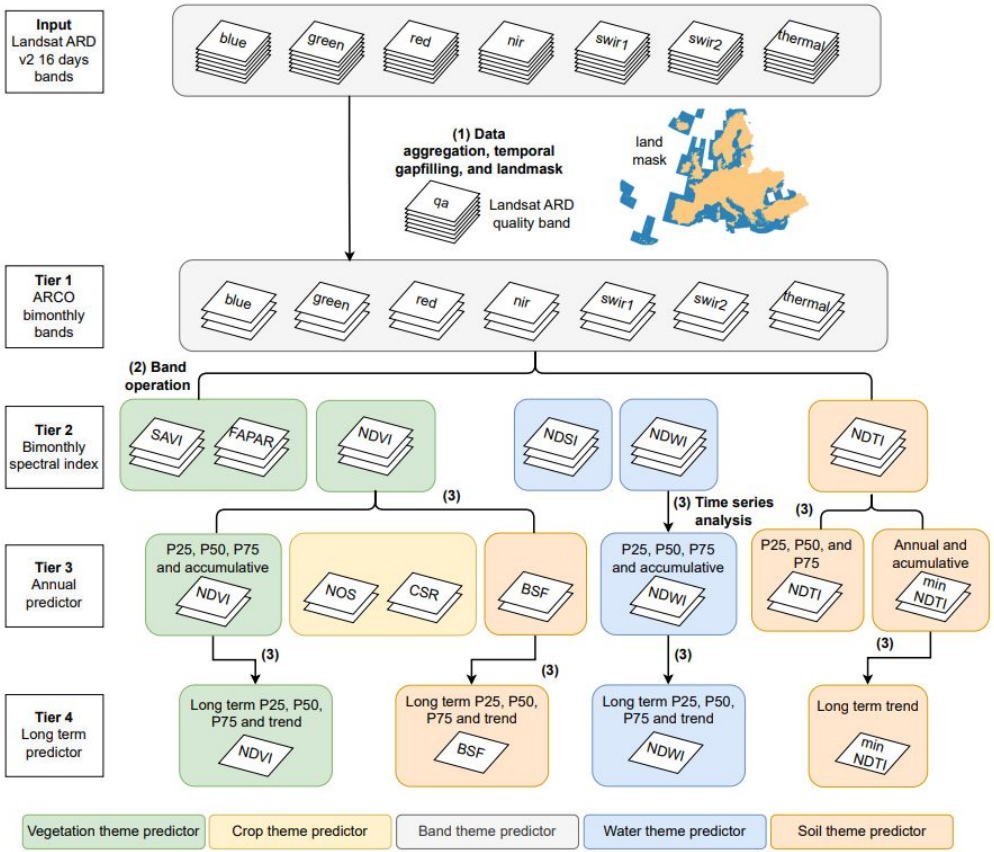
Final ML setup

ML training
(Single model or ensemble)

ML loop



Landsat cloud-free aggregates for Europe



https://doi.org/10.5194/essd-2024-266
Preprint. Discussion started: 12 September 2024
© Author(s) 2024. CC BY 4.0 License.

Earth System Science Data

Time-series of Landsat-based bi-monthly and annual spectral indices for continental Europe for 2000–2022

Xuemeng Tian[1,2], Davide Consoli[1], Martijn Witjes[1], Florian Schneider[3], Leandro Parente[1], Murat Şahin[1], Yufeng Ho[1], Robert Minařík[1], and Tomislav Hengl[1]

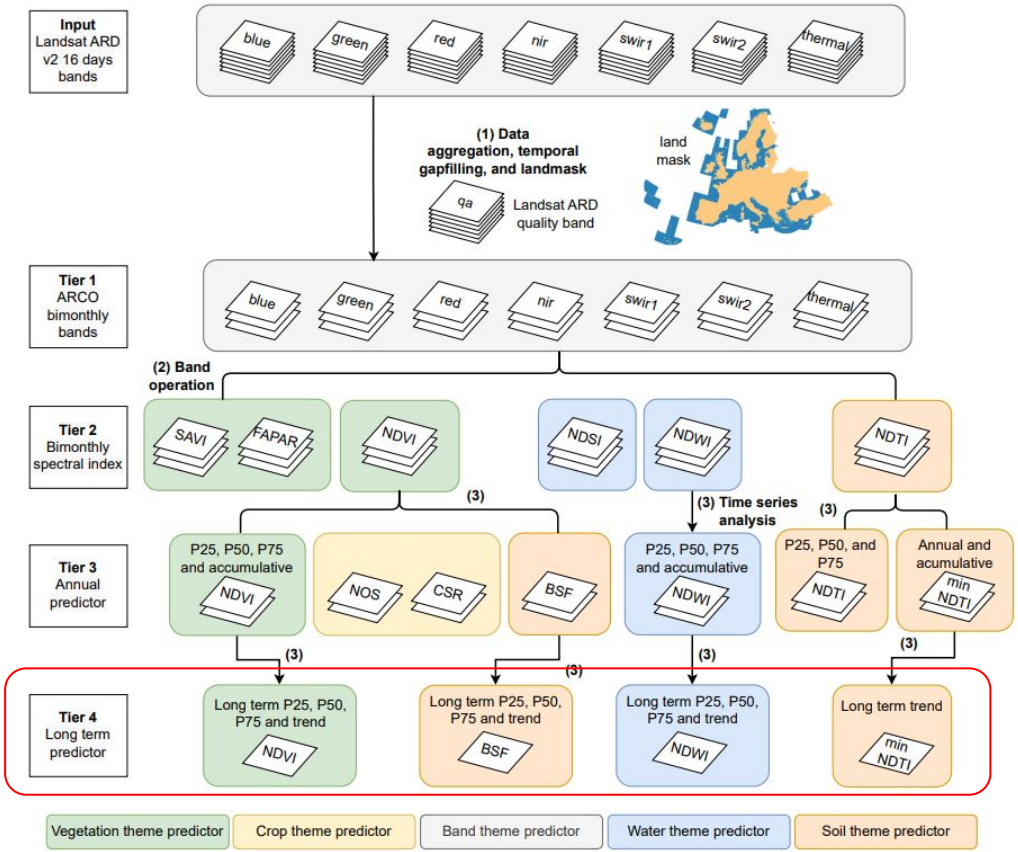
[1]OpenGeoHub, Doorwerth, The Netherlands
[2]Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen, The Netherlands
[3]Thünen Institute of Climate-Smart Agriculture, Germany

Correspondence: Xuemeng Tian (xuemeng.tian@opengeohub.org)

Abstract. The production and evaluation of the Analysis Ready and Cloud Optimized (ARCO) data cube for continental Europe (including Ukraine, the UK, and Turkey), derived from the Landsat Analysis Ready Data version 2 (ARD V2) produced by Global Land Analysis and Discovery team (GLAD) and covering the period from 2000 to 2022 is described. The data cube consists of 17TB of data at a 30–meter resolution and includes bimonthly, annual, and long-term spectral indices on various thematic topics, including: surface reflectance bands, Normalized Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Normalized Difference Snow Index (NDSI), Normalized Difference Water Index (NDWI), Normalized Difference Tillage Index (NDTI), minimum Normalized Difference Tillage Index (minNDTI), Bare Soil Fraction (BSF), Number of Seasons (NOS), and Crop Duration Ratio (CDR). The data cube was developed with the intention of providing a comprehensive feature space for environmental modeling and soil, vegetation, and land cover mapping. To evaluate its effectiveness for this purpose, the quality of the produced time series was assessed by: (1) visual examination for artifacts and inconsistencies, (2) plausibility checks with ground survey data, and (3) predictive modeling tests, examples with soil organic carbon (SOC) and land cover (LC) classification. The results of visual examination indicate that the gap-filled product is complete and consistent, except for winter periods in northern latitudes and high-altitude areas where high cloud and snow density make gap-filling complex, and hence many artifacts remain. The plausibility results further show that the indices effectively help differentiate landscapes and crop types: the BSF index showed a strong negative correlation (-0.73) with crop coverage data, effectively detecting soil exposure. The minNDTI index had a moderate positive correlation (0.57) with the Eurostat tillage practices survey data, indicating valuable information on the intensity of the tillage. The detailed temporal resolution and long-term characteristics provided by different tiers of predictors in this data cube proved to be important for both soil organic carbon regression and LC classification experiments based on the 60,723 LUCAS observations: long-term characteristics (tier 4) were particularly valuable for predictive mapping of SOC and LC coming on the top of variable importance assessment. Crop-specific indices (NOS and CDR) provided limited value for the tested applications, possibly due to noise or insufficient quantification methods. The data cube is made available under a CC-BY license and will be continuously updated.

https://doi.org/10.5194/essd-2024-266

Feature Engineering

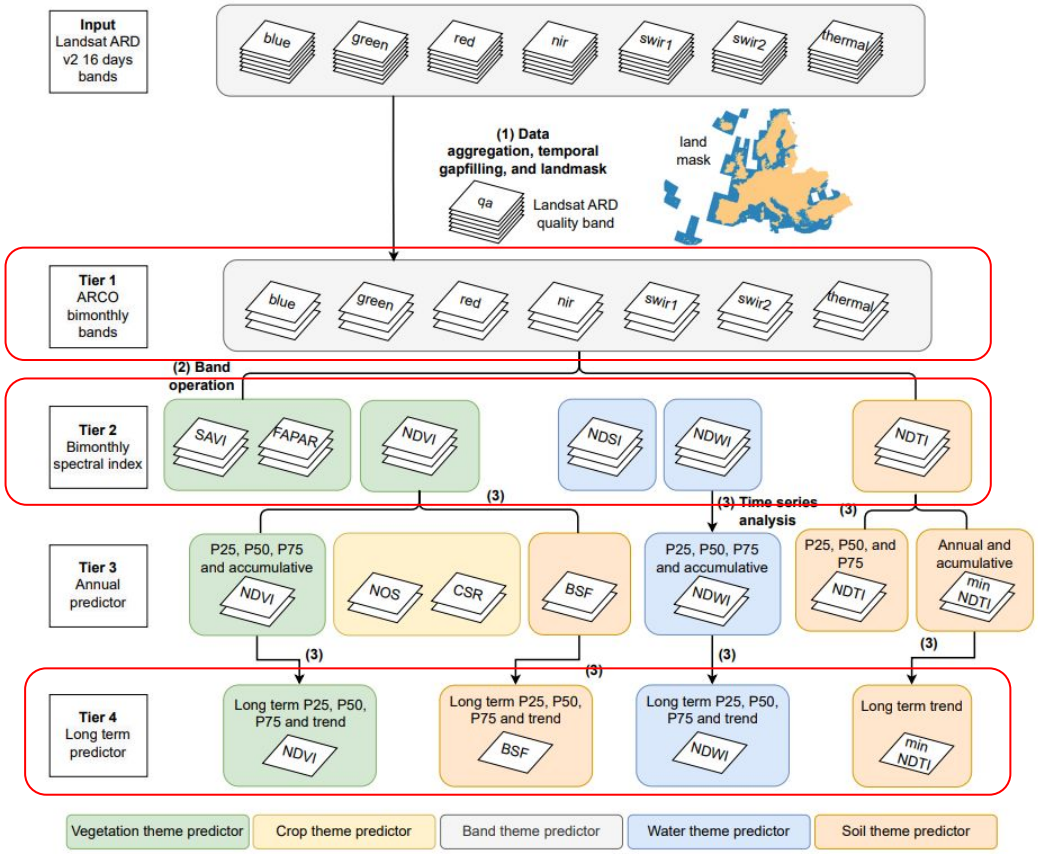


Soil organic carbon regression

Number of tiers combined	Tiers	Feature number	CCC	RMSE	R2
1	tier1	42	0.635	0.651	0.489
	tier2	36	0.623	0.660	0.476
	tier3	38	0.637	0.652	0.487
	tier4	13	**0.682**	**0.613**	**0.547**
2	tier1+tier2	78	0.647	0.642	0.503
	tier1+tier3	80	0.648	0.642	0.503
	tier1+tier4	55	0.661	0.630	0.521
	tier2+tier3	74	0.648	0.643	0.501
	tier2+tier4	49	0.644	0.643	0.502
	tier3+tier4	51	0.659	0.634	0.516
3	tier1+tier2+tier3	116	0.652	0.639	0.507
	tier1+tier2+tier4	91	0.664	0.629	0.523
	tier1+tier3+tier4	93	0.663	0.630	0.522
	tier2+tier3+tier4	87	0.662	0.631	0.520
4 (all predictors)	tier1+tier2+tier3+tier4	129	0.665	0.629	0.523

https://doi.org/10.5194/essd-2024-266

Feature Engineering

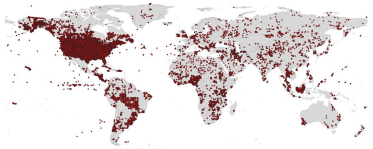


ber of tiers combined	Tiers	Feature number	F1-score	Precision (U.A.)	Recall (P.A.)
1	tier1	42	0.769	0.776	0.784
	tier2	36	0.776	0.781	0.789
	tier3	38	0.761	0.762	0.773
	tier4	13	0.785	0.782	0.791
2	tier1+tier2	78	0.782	0.787	0.794
	tier1+tier3	80	0.774	0.777	0.787
	tier1+tier4	55	0.785	0.792	0.799
	tier2+tier3	74	0.777	0.781	0.790
	tier2+tier4	49	0.785	0.791	0.797
	tier3+tier4	51	0.775	0.777	0.787
3	tier1+tier2+tier3	116	0.779	0.783	0.792
	tier1+tier2+tier4	91	**0.790**	**0.796**	**0.802**
	tier1+tier3+tier4	93	0.782	0.785	0.795
	tier2+tier3+tier4	87	0.784	0.788	0.796
4 (all predictors)	tier1+tier2+tier3+tier4	129	0.786	0.789	0.798

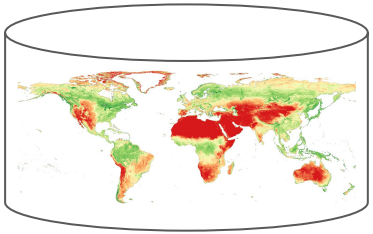
Land cover classification

https://doi.org/10.5194/essd-2024-266

ML pipeline for spatial prediction



Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)

Spatial / Spacetime overlay

Classification / Regression matrix

Hyper-parameter tuning
(calibration & cross-validation)

Feature selection
(subsampling & cross-validation)

Final ML setup

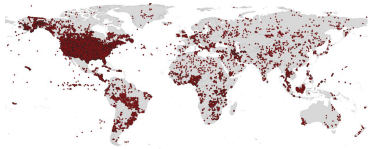
ML training
(Single model or ensemble)

- Random Forest
- Gradient boosting trees
- Artificial Neural Networks

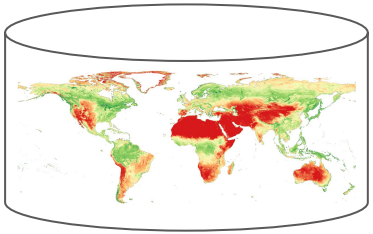
Recursive Feature Elimination (RFE)

Successive Halving Search

ML pipeline for spatial prediction



Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)

Spatial / Spacetime overlay

Classification / Regression matrix

Hyper-parameter tuning
(calibration & cross-validation)

Feature selection
(subsampling & cross-validation)

Final ML setup

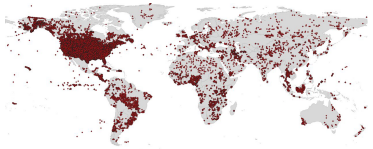
ML training
(Single model or ensemble)

Predictions

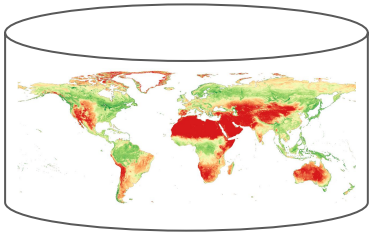
Model performance

Feature importance

ML pipeline for spatial prediction



Training samples
(in-situ points)



ARCO datasets
(harmonized, gapfilled consistent layers)

Spatial / Spacetime overlay

Classification / Regression matrix

Hyper-parameter tuning
(calibration & cross-validation)

Feature selection
(subsampling & cross-validation)

Final ML setup

ML training
(Single model or ensemble)

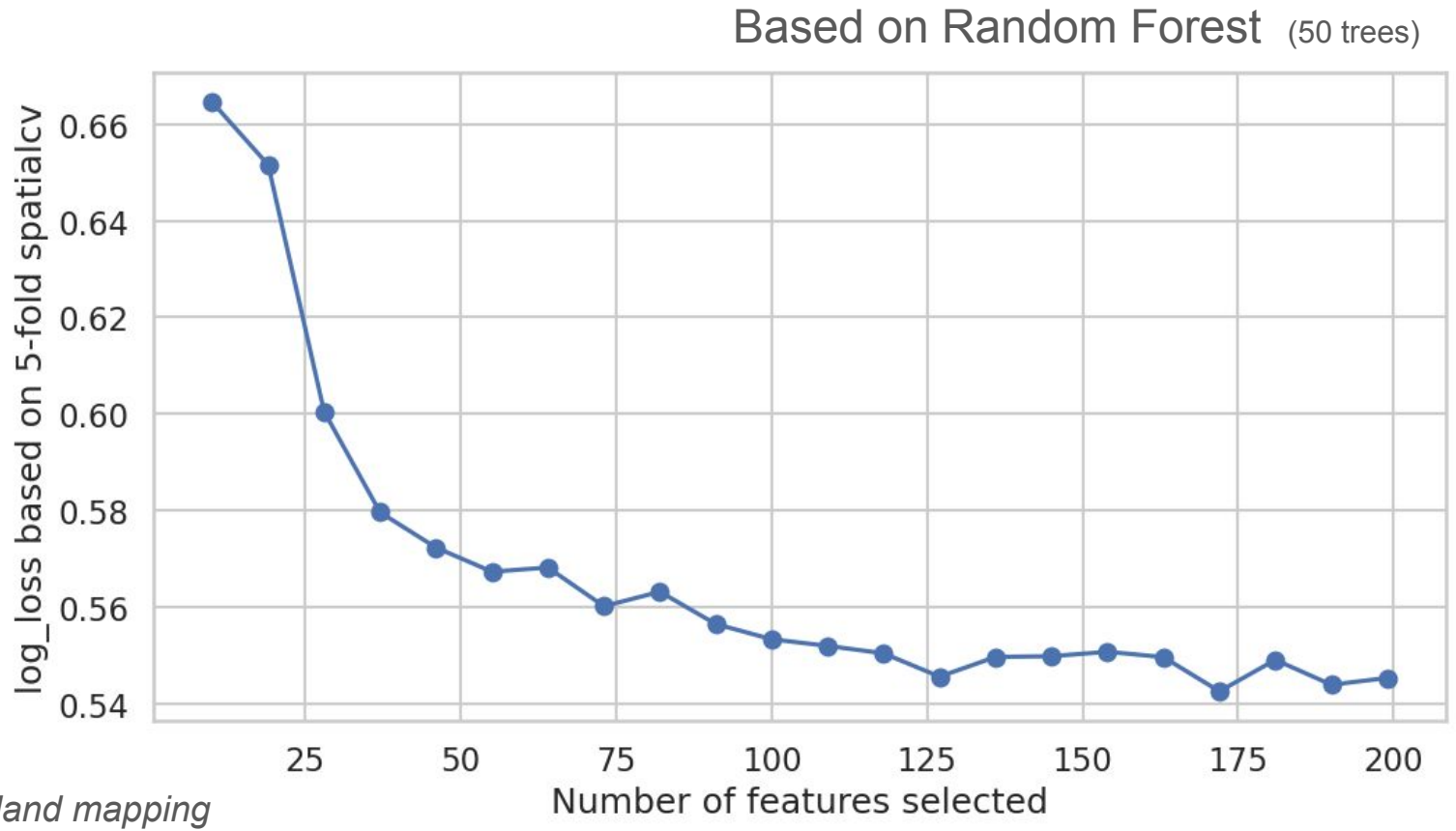
Predictions
(most of the computation cost)

Model performance

Feature importance

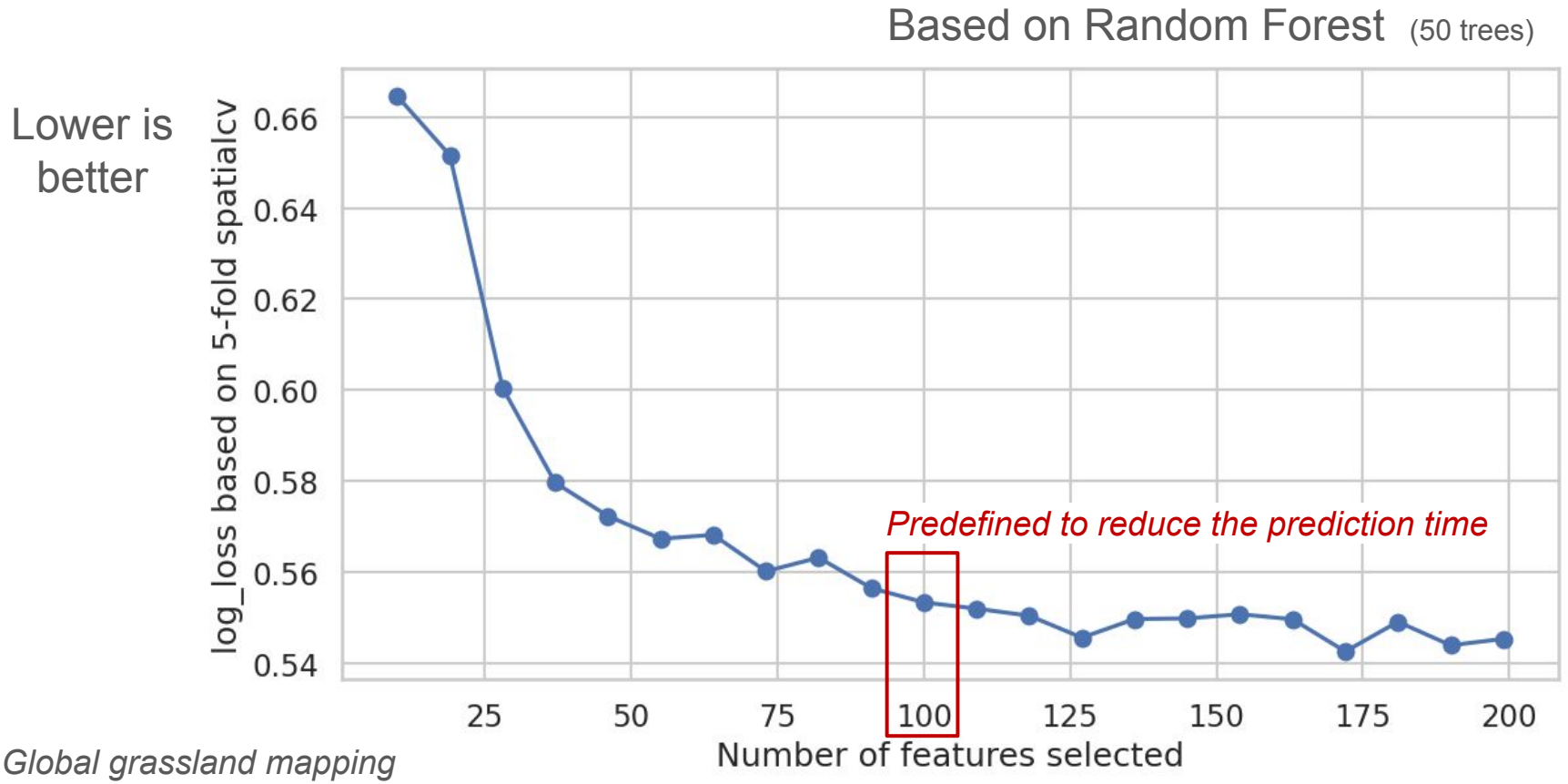
Recursive Feature Elimination example

Lower is better

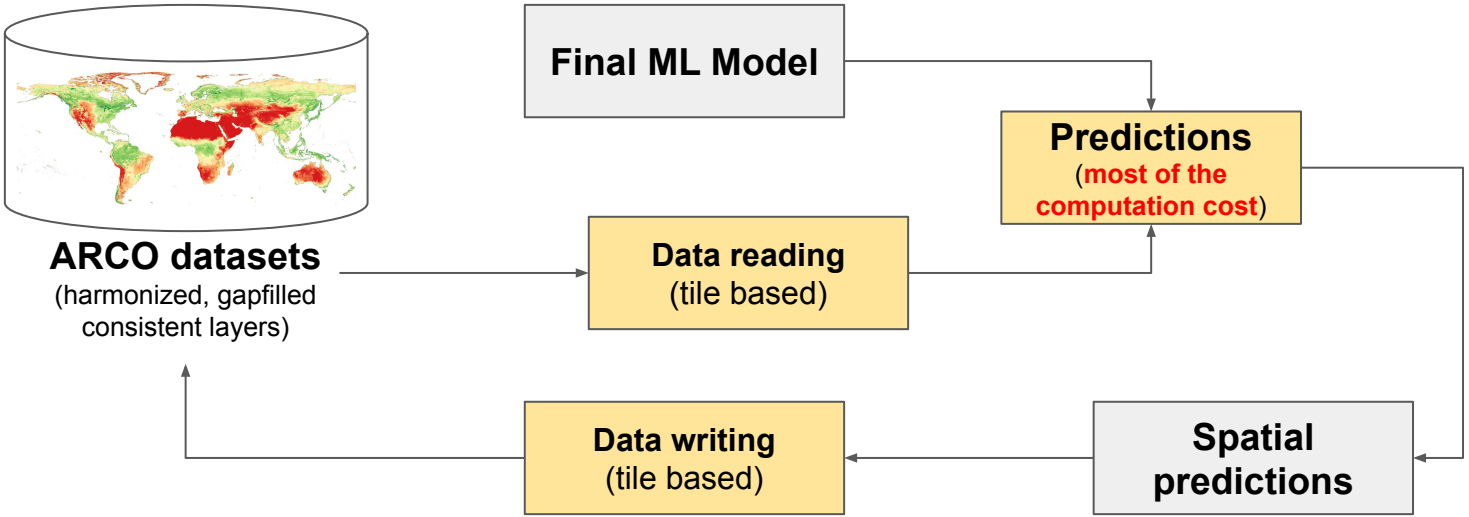


Global grassland mapping

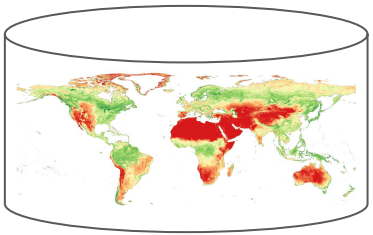
Recursive Feature Elimination example



ML pipeline for spatial prediction



ML pipeline for spatial prediction



ARCO datasets
(harmonized, gapfilled consistent layers)

Final ML Model

CPU bound

Predictions
(**most of the computation cost**)

Data reading
(tile based)

Data writing
(tile based)

Spatial predictions

IO bound
(but raster compression algorithms use CPU)

ML pipeline for spatial prediction

- **Increase data reading/saving speed:**
  - Improved hardware (SSD disks, RAM, 10 Gigabit Ethernet, InfiniBand),
  - Faster software storage solution (distributed file system, S3),
  - Efficient IO parallelization (thread based & with low overhead – basically C++),
  - Efficient memory management (pre-allocation & zero-copy),

ML pipeline for spatial prediction

- **Increase data reading/saving speed:**
  - Improved hardware (SSD disks, RAM, 10 Gigabit Ethernet, InfiniBand),
  - Faster software storage solution (distributed file system, S3),
  - Efficient IO parallelization (thread based & with low overhead – basically C++),
  - Efficient memory management (pre-allocation & zero-copy),
- **Reduce** prediction time:
  - More hardware (CPU, GPU),
  - Compiled ML models
  - Multi-year predictions (all years together)

ML pipeline for spatial prediction

- **Increase data reading/saving speed:**
  - Improved hardware (SSD disks, RAM, 10 Gigabit Ethernet, InfiniBand),
  - Faster software storage solution (distributed file system, S3),
  - Efficient IO parallelization (thread based & with low overhead – basically C++),
  - Efficient memory management (pre-allocation & zero-copy),
- **Reduce** prediction time:
  - More hardware (CPU, GPU),
  - Compiled ML models
  - Multi-year predictions (all years together)

TL2cgen
latest
https://tl2cgen.readthedocs.io

lleaves
https://lleaves.readthedocs.io

Hummingbird

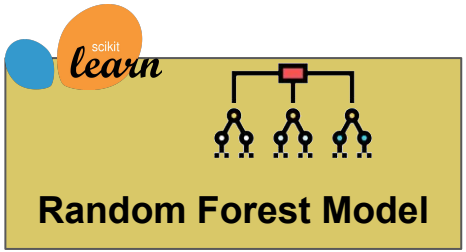
ML pipeline for spatial prediction

- **Increase data reading/saving speed:**
  - Improved hardware (SSD disks, RAM, 10 Gigabit Ethernet, InfiniBand),
  - Faster software storage solution (distributed file system, S3),
  - Efficient IO parallelization (thread based & with low overhead – basically C++),
  - Efficient memory management (pre-allocation & zero-copy),
- **Reduce** prediction time:
  - More hardware (CPU, GPU),
  - Compiled ML models
  - Multi-year predictions (all years together)

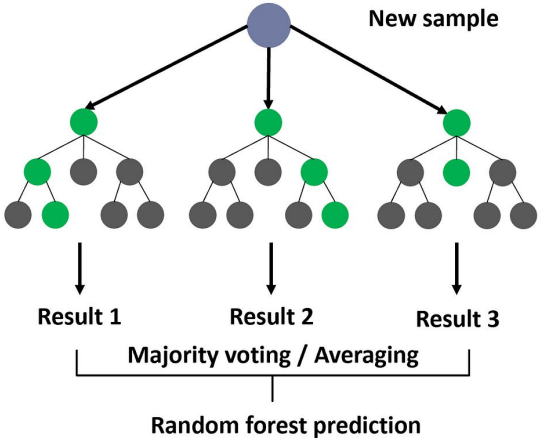
scikit map

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

ML pipeline for spatial prediction

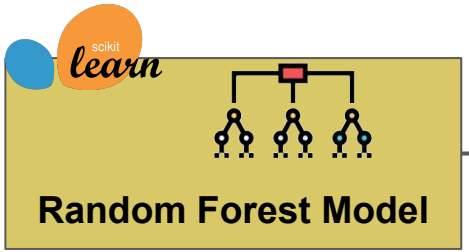


https://scikit-learn.org




```
>>> from sklearn.ensemble import RandomForestClassifier
>>> from sklearn.datasets import make_classification
>>> X, y = make_classification(n_samples=1000, n_features=4,
...                            n_informative=2, n_redundant=0,
...                            random_state=0, shuffle=False)
>>> clf = RandomForestClassifier(max_depth=2, random_state=0)
>>> clf.fit(X, y)
RandomForestClassifier(...)
>>> print(clf.predict([[0, 0, 0, 0]]))
[1]
```

ML pipeline for spatial prediction



```
#include "header.h"
#include "header.h"
#include "header.h"
void predict_margin_multiclass_unit10(union Entry* data, double* result) {
  double sum[3] = {0};
  unsigned int tmp;
  int nid, cond, fid;  /* used for folded subtrees */
  if ( UNLIKELY( !(data[3].missing != -1) || (data[3].qvalue <= 24) ) ) {
    if ( LIKELY( !(data[5].missing != -1) || (data[5].qvalue <= 146) ) ) {
      if ( LIKELY( !(data[71].missing != -1) || (data[71].qvalue <= 652) ) ) {
        if ( UNLIKELY( !(data[84].missing != -1) || (data[84].qvalue <= 224) ) ) {
          if ( LIKELY( !(data[63].missing != -1) || (data[63].qvalue <= 678) ) ) {
            if ( LIKELY( !(data[49].missing != -1) || (data[49].qvalue <= 8718) ) ) {
              if ( UNLIKELY( !(data[77].missing != -1) || (data[77].qvalue <= 448) ) ) {
                if ( UNLIKELY( !(data[93].missing != -1) || (data[93].qvalue <= 594) ) ) {
                  if ( LIKELY( !(data[25].missing != -1) || (data[25].qvalue <= 1050) ) ) {
                    if ( UNLIKELY( !(data[29].missing != -1) || (data[29].qvalue <= 2776) ) ) {
                      sum[0] += (double)0;
                      sum[1] += (double)0;
                      sum[2] += (double)1;
                    } else {
                      if ( LIKELY( !(data[33].missing != -1) || (data[33].qvalue <= 726) ) ) {
                        sum[0] += (double)0;
                        sum[1] += (double)0.5952380952380952328;
                        sum[2] += (double)0.40476190476190476672;
                      } else {
                        sum[0] += (double)0;
                        sum[1] += (double)0;
                        sum[2] += (double)1;
                      }
                    }
                  } else {
...
```

Model compiler for decision tree models

https://tl2cgen.readthedocs.io

C++ compiler

One IF-ELSE per tree node

The Performance of Decision Tree Evaluation Strategies

Global Pasture Watch

A consortium of research partners developing global grassland and livestock monitoring data for impact

OpenGeoHUB — Connect · Create · Share · Repeat

LAPIG — Laboratório de Processamento de Imagens e Geoprocessamento

iDiv

GLAD — Global Land Analysis & Discovery

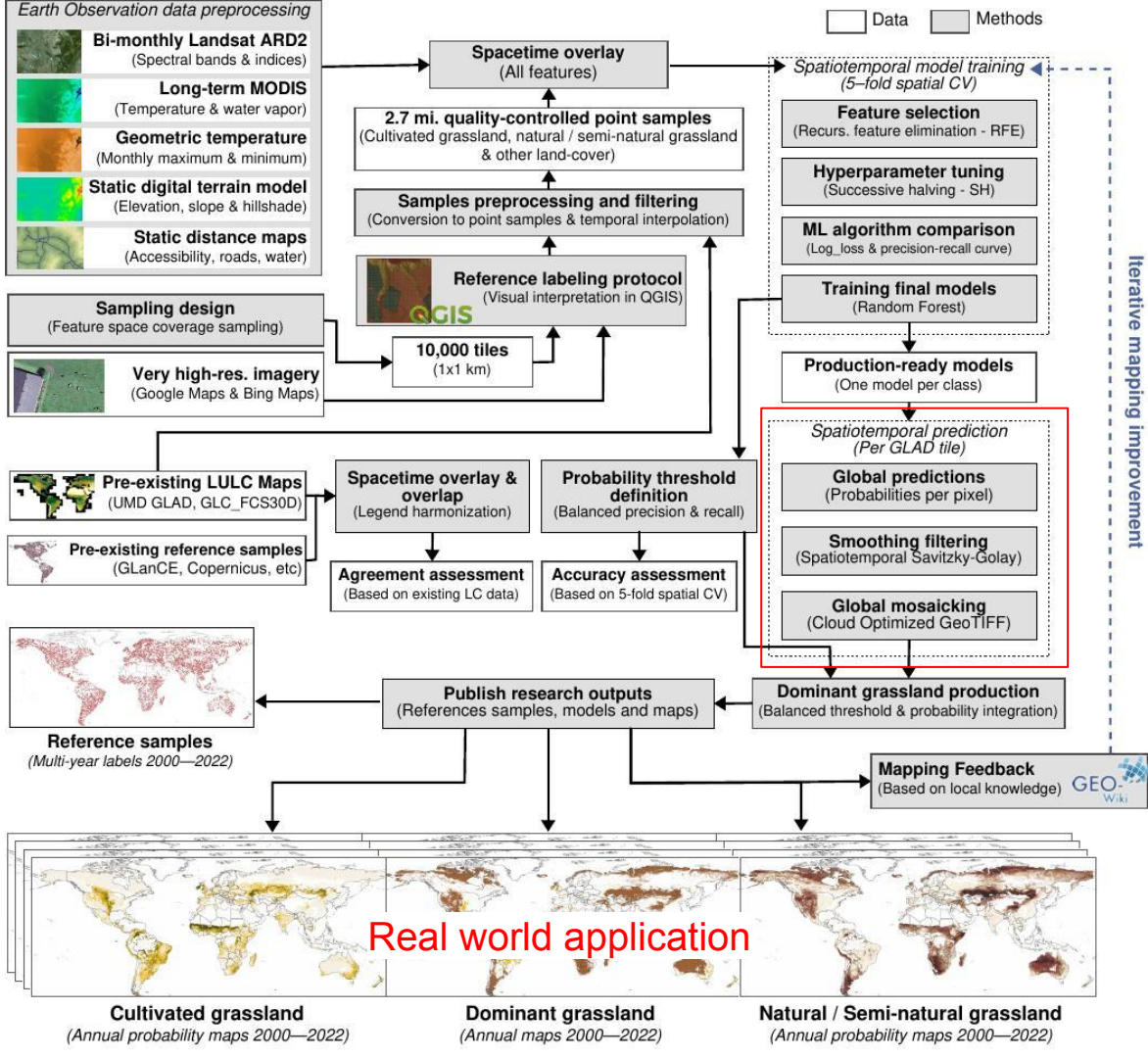
Cornell University

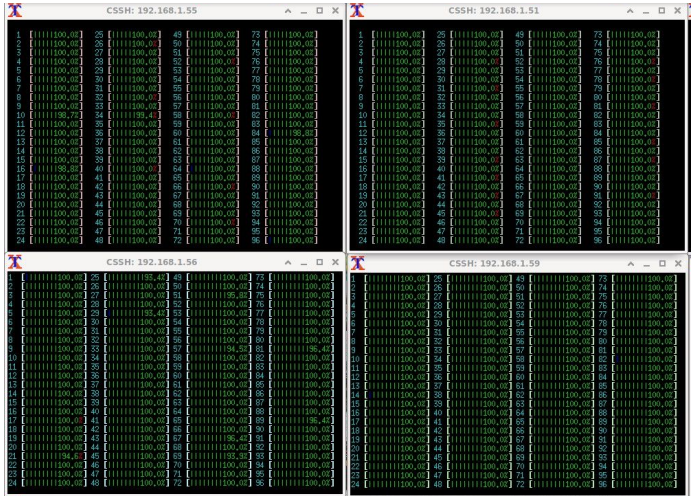
IIASA

Land & Carbon Lab

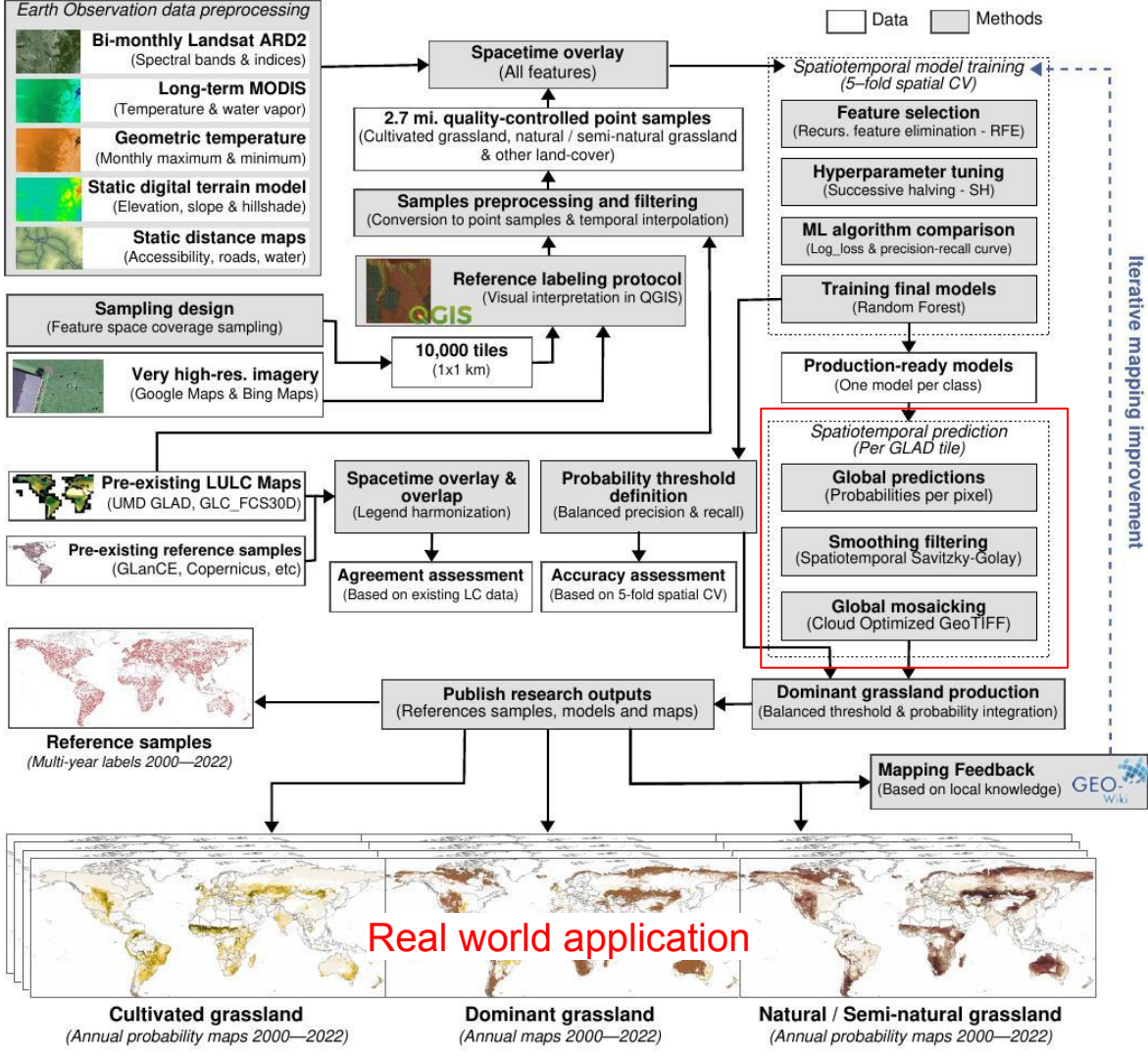
World Resources Institute

https://landcarbonlab.org/about-global-pasture-watch/



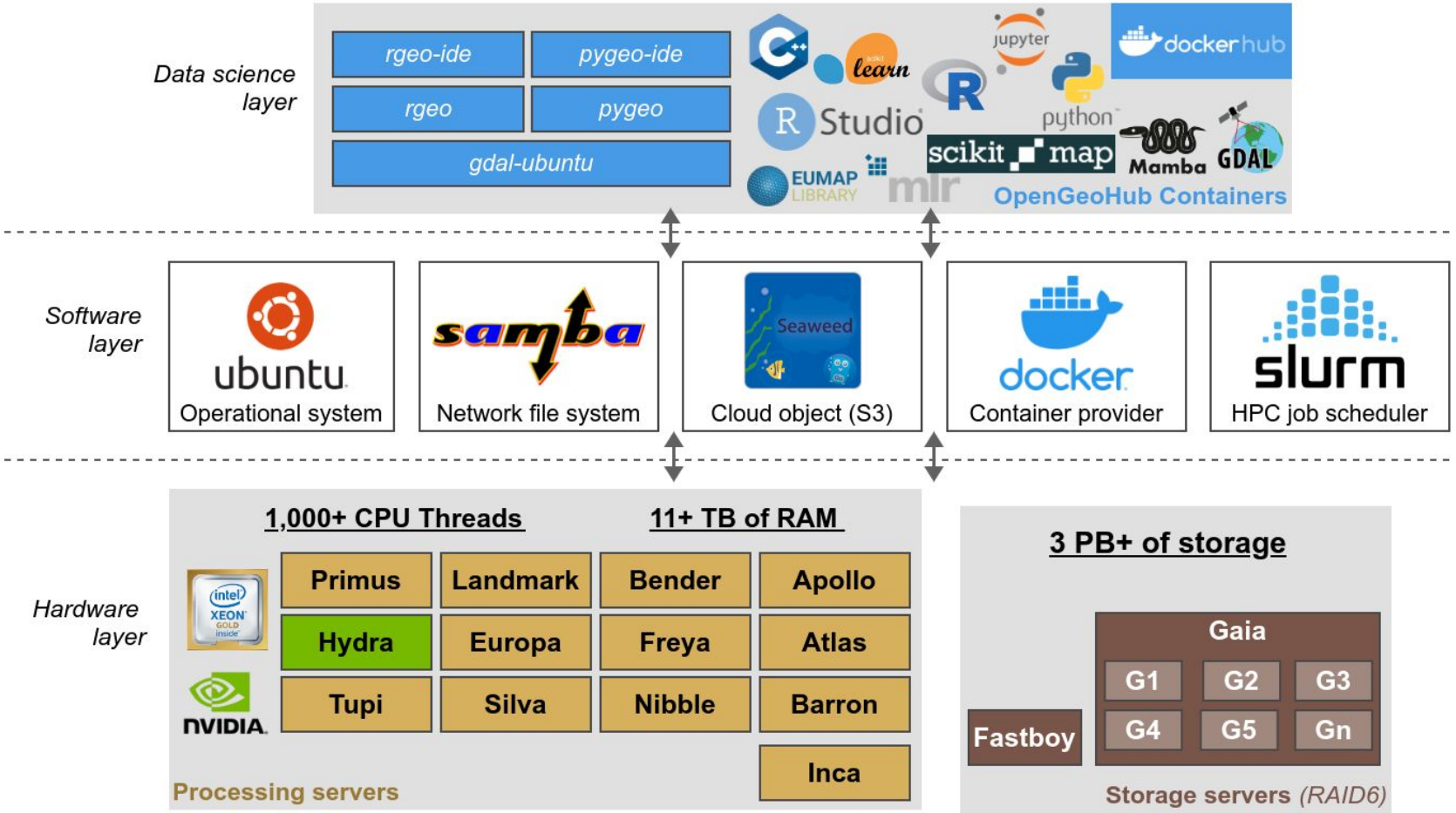


Approximately
120,960 CPU hours &
7.2 terabytes of RAM

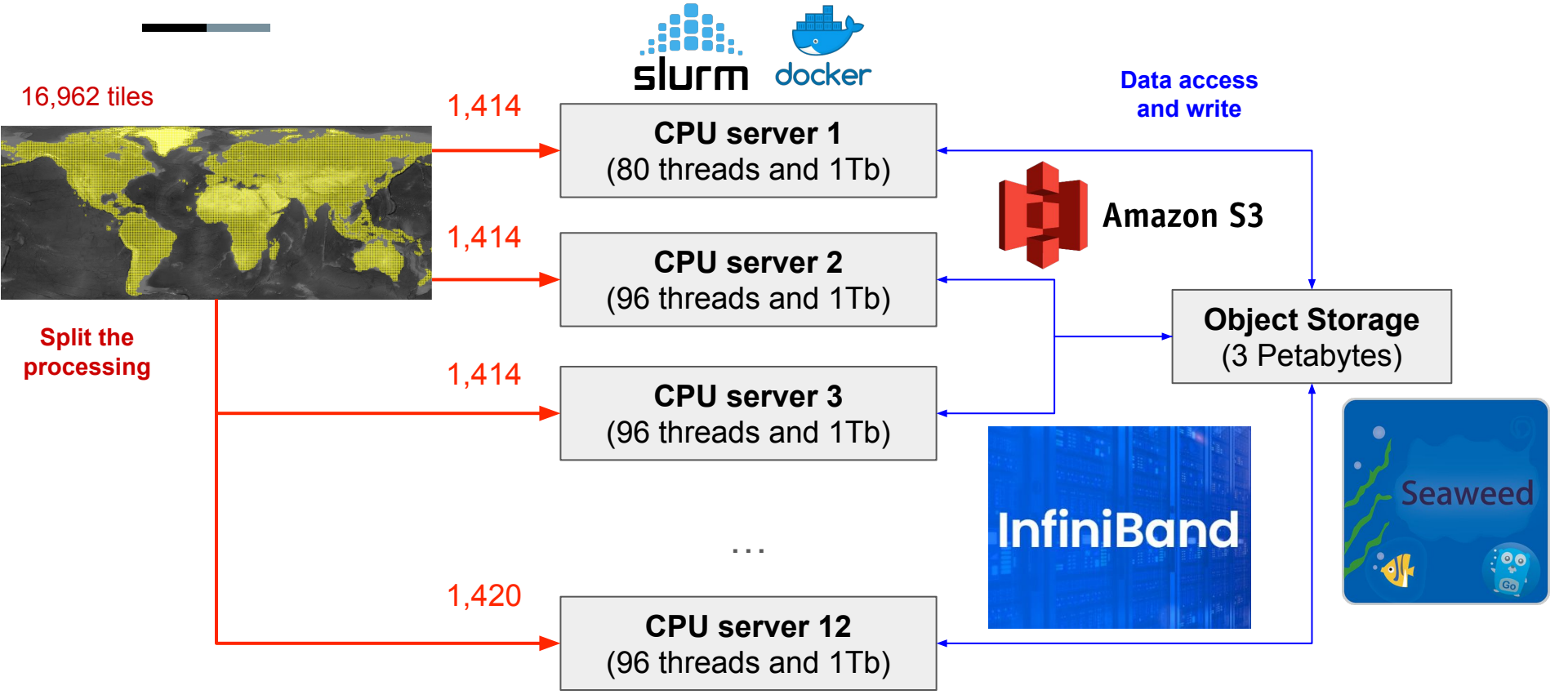


High-throughput data processing infrastructure

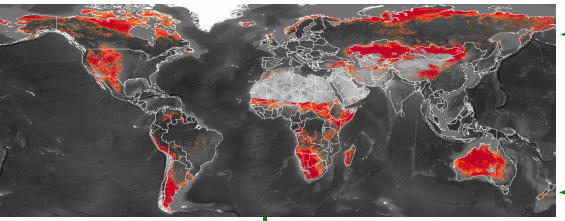
OpenGeoHUB
Connect • Create • Share • Repeat



Global ML prediction



Global ML prediction



slurm

docker

CPU server 1
(80 threads and 1Tb)

CPU server 2
(96 threads and 1Tb)

CPU server 3
(96 threads and 1Tb)

...

CPU server 12
(96 threads and 1Tb)

Data access

Amazon S3

Object Storage
(3 Petabytes)

InfiniBand

Seaweed

GDAL

Cloud Optimized

GeoTIFF generation

gdalbuildvrt

gdal_translate

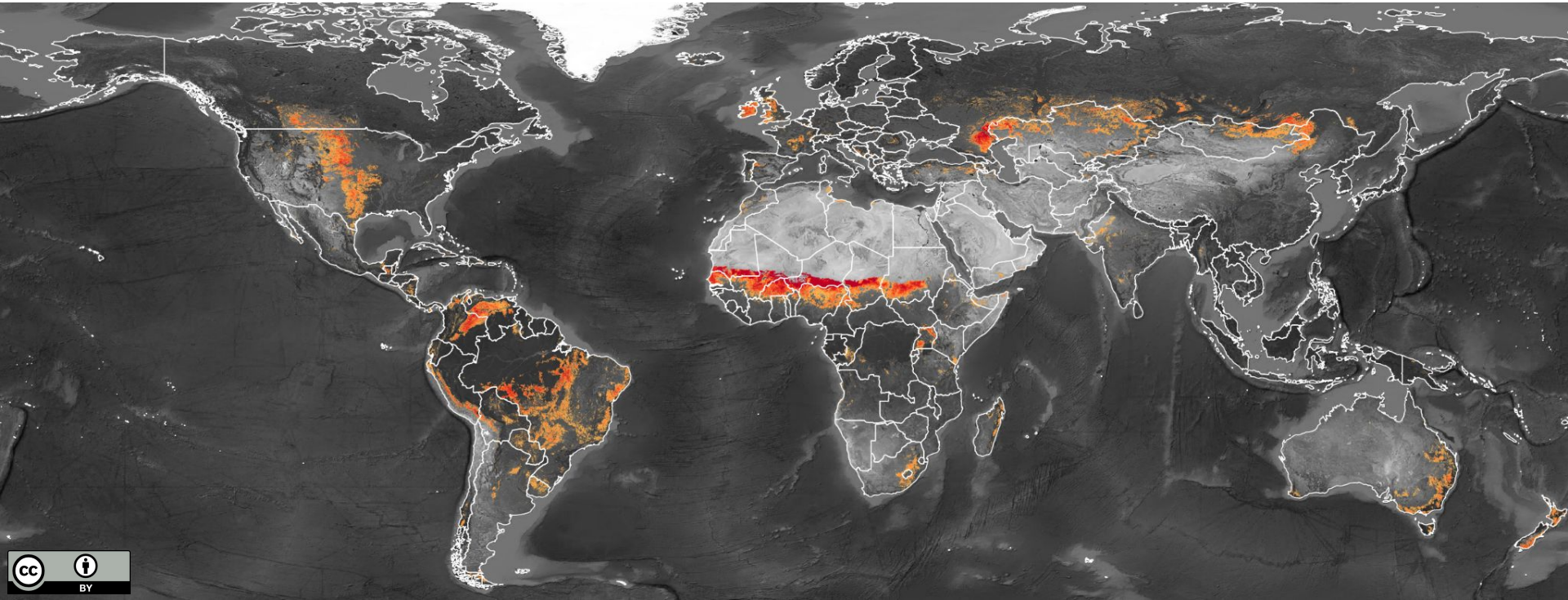
Global Pasture Watch mapping

Cultivated grassland - 2022

0.32

0.15 — 0.75

Probability threshold



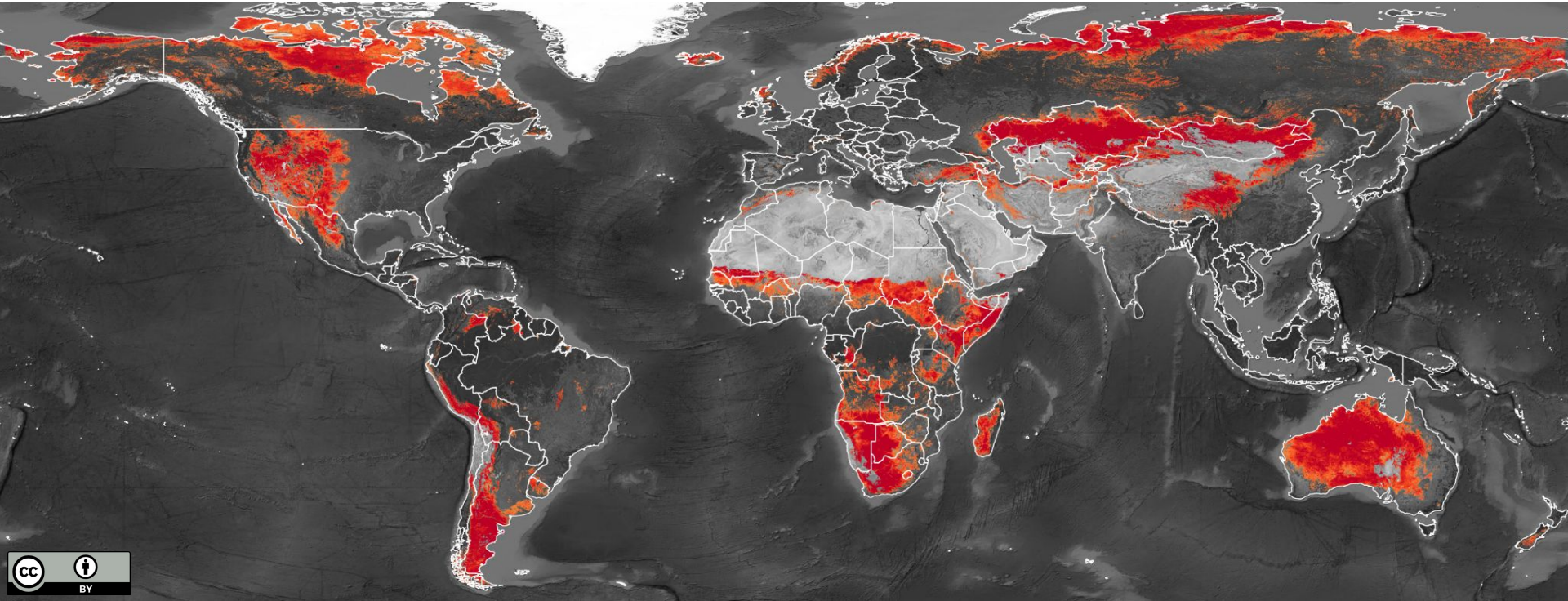
CC BY

Global Pasture Watch mapping

OpenGeoHUB
Connect • Create • Share • Repeat

Natural / Semi-natural grassland - 2022

0.42
0.15 0.75
Probability threshold



CC BY

Global Pasture Watch mapping

OpenGeoHUB
Connect • Create • Share • Repeat

	Cultivated grassland			Natural/Semi-natural grassland		
ML algorithm	$R^2_{logloss}$	Balanced probability threshold	F1 score	$R^2_{logloss}$	Balanced probability threshold	F1 score
Random Forest - RF	0.924	0.328	0.644	0.773	0.428	0.759
Gradient boosting trees - GBT	0.924	0.162	0.653	0.767	0.352	0.760
Artificial Neural Network - ANN	0.916	0.380	0.607	0.697	0.468	0.720

		Expected			Recall (Producer's acc.)
		Cultivated grassland	Other LC	Total	
Predicted	Cultivated grassland	**0.062**	0.034	0.096	0.643
	Other LC	0.034	**0.869**	0.904	0.962
	Total	0.096	0.904	**1.000**	
Precision (User's acc.)		0.644	0.962		

		Expected			Recall (Producer's acc.)
		Natural/Semi-natural grassland	Other LC	Total	
Predicted	Natural/Semi-natural grass	**0.202**	0.064	0.266	0.758
	Other LC	0.064	**0.670**	0.734	0.913
	Total	0.266	0.734	**1.000**	
Precision (User's acc.)		0.759	0.913		

2.1 millions reference samples & 5-fold spatial blocking CV

Global Pasture Watch mapping

OpenGeoHUB
Connect • Create • Share • Repeat

	Cultivated grassland			Natural/Semi-natural grassland		
ML algorithm	$R^2_{logloss}$	Balanced probability threshold	F1 score	$R^2_{logloss}$	Balanced probability threshold	F1 score
Random Forest - RF	0.924	0.328	0.644	0.773	0.428	0.759
Gradient boosting trees - GBT	0.924	0.162	0.653	0.767	0.352	0.760
Artificial Neural Network - ANN	0.916	0.380	0.607	0.697	0.468	0.720

		Expected			Recall (Producer's acc.)
		Cultivated grassland	Other LC	Total	
Predicted	Cultivated grassland	**0.062**	0.034	0.096	0.643
	Other LC	0.034	**0.869**	0.904	0.962
	Total	0.096	0.904	**1.000**	
Precision (User's acc.)		0.644	0.962		

		Expected			Recall (Producer's acc.)
		Natural/Semi-natural grassland	Other LC	Total	
Predicted	Natural/Semi-natural grass	**0.202**	0.064	0.266	0.758
	Other LC	0.064	**0.670**	0.734	0.913
	Total	0.266	0.734	**1.000**	
Precision (User's acc.)		0.759	0.913		

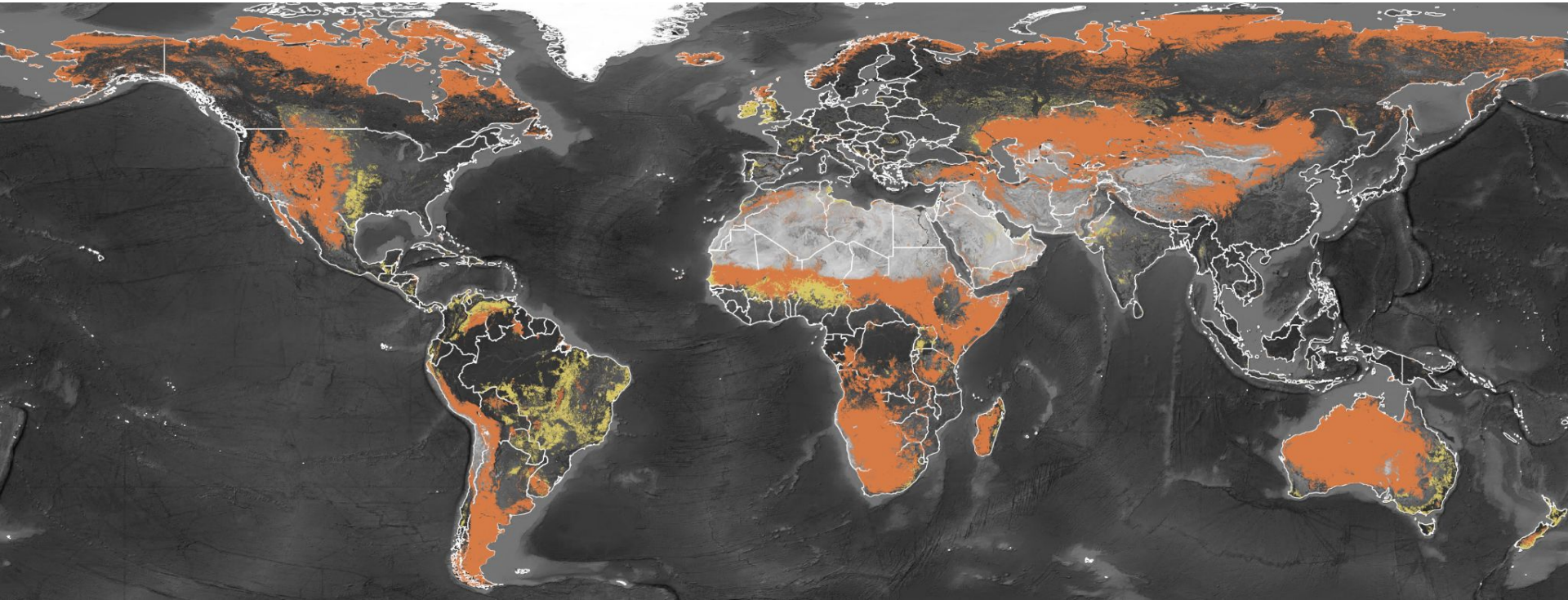
2.1 millions reference samples & 5-fold spatial blocking CV

Global Pasture Watch mapping

Dominant grassland - 2022

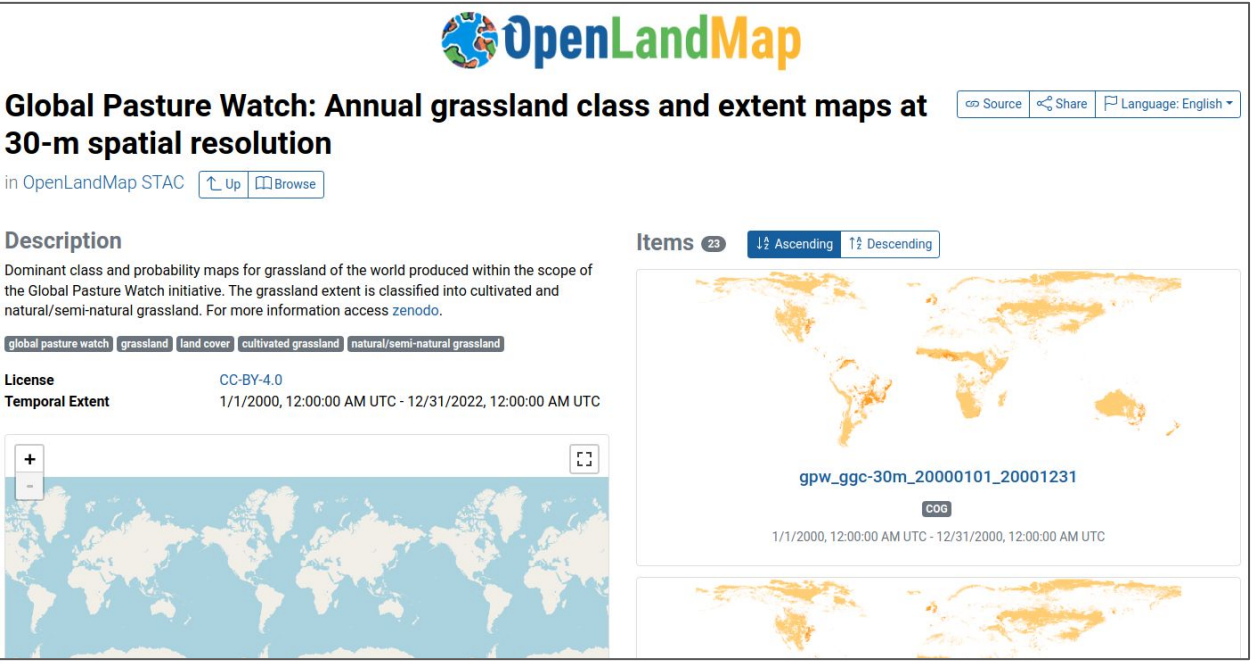
Cultivated

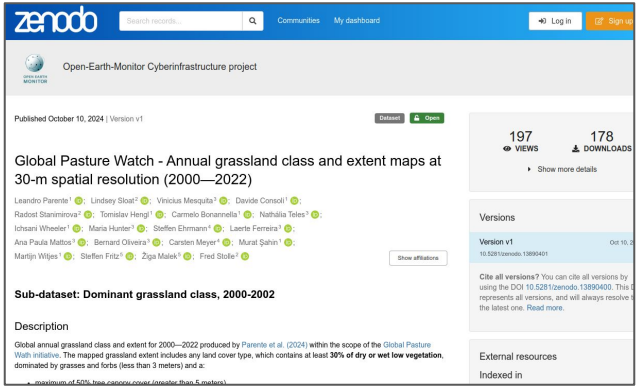
Natural / Semi-natural



Global Pasture Watch mapping

COG files accessible via STAC & Zenodo



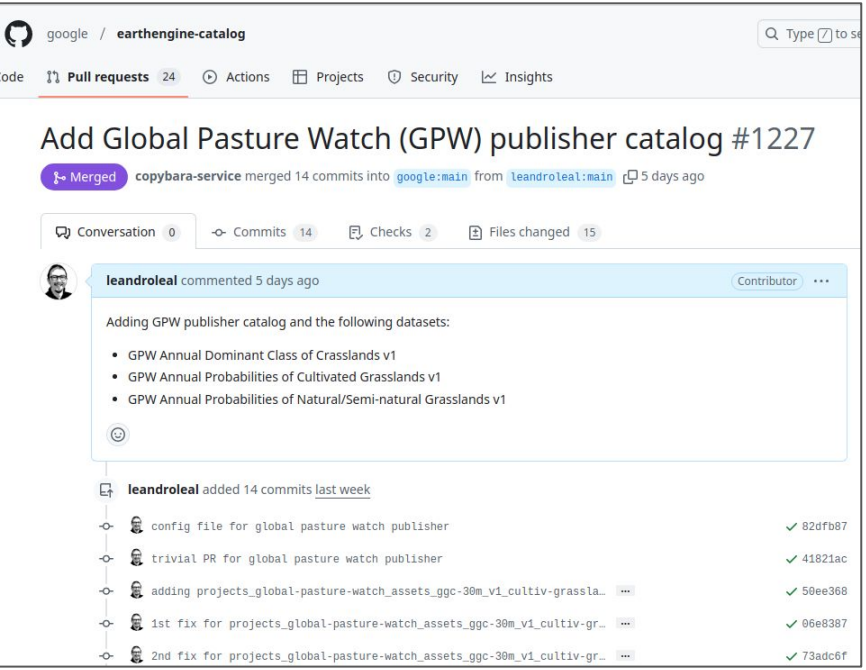


https://stac.openlandmap.org/gpw_ggc-30m/collection.json

https://doi.org/10.5281/zenodo.13890401

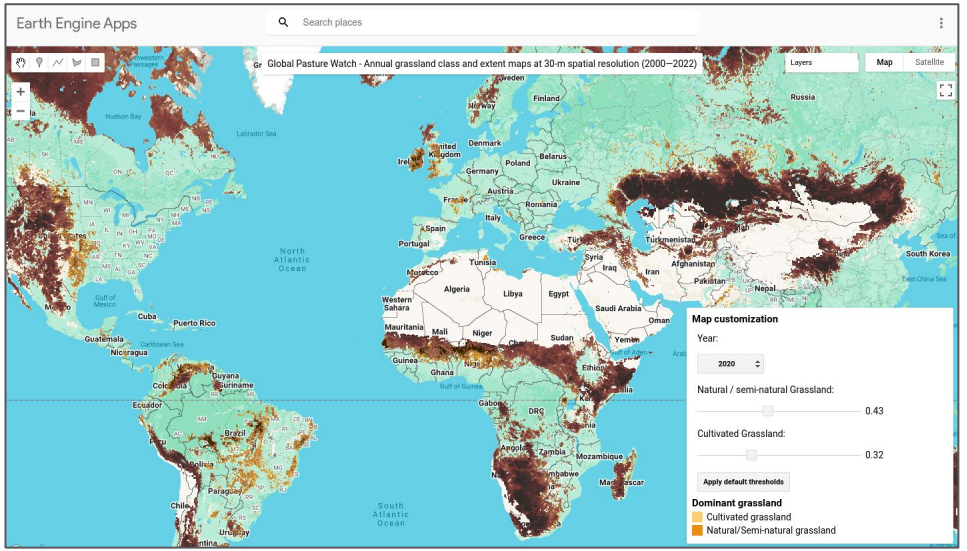
Global Pasture Watch mapping

Dataset submission to GEE catalog



https://developers.google.com/earth-engine/datasets/publisher

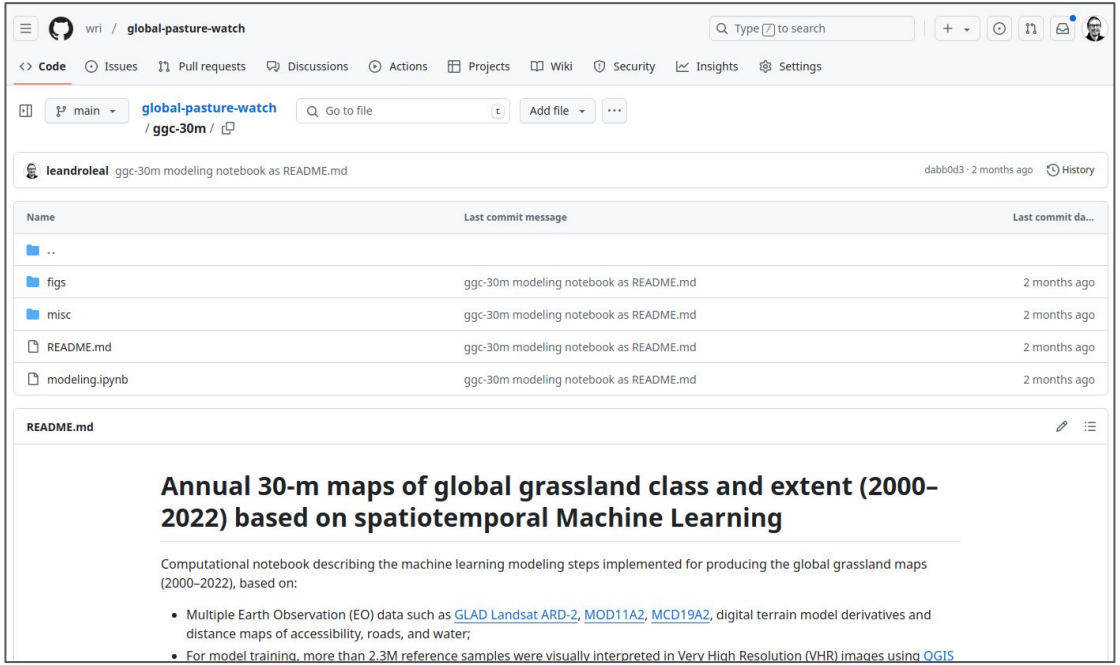
GEE seamless accessing global COG file hosted in GCS



https://global-pasture-watch.projects.earthengine.app/view/ggc-30m

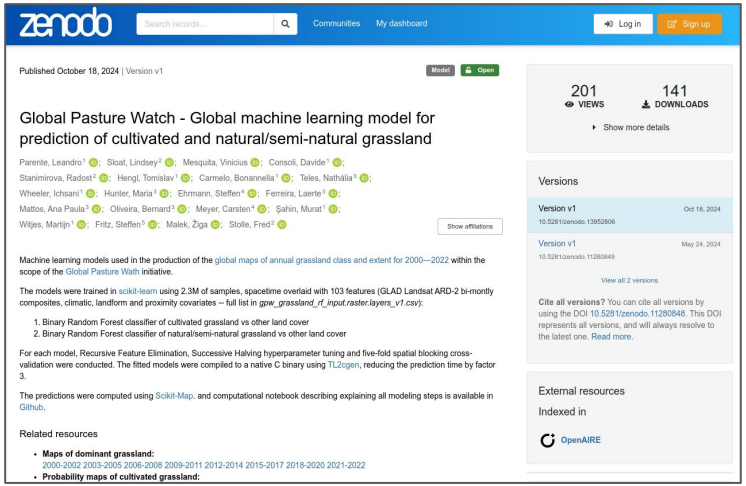
Global Pasture Watch mapping

Modeling pipeline as computational notebook



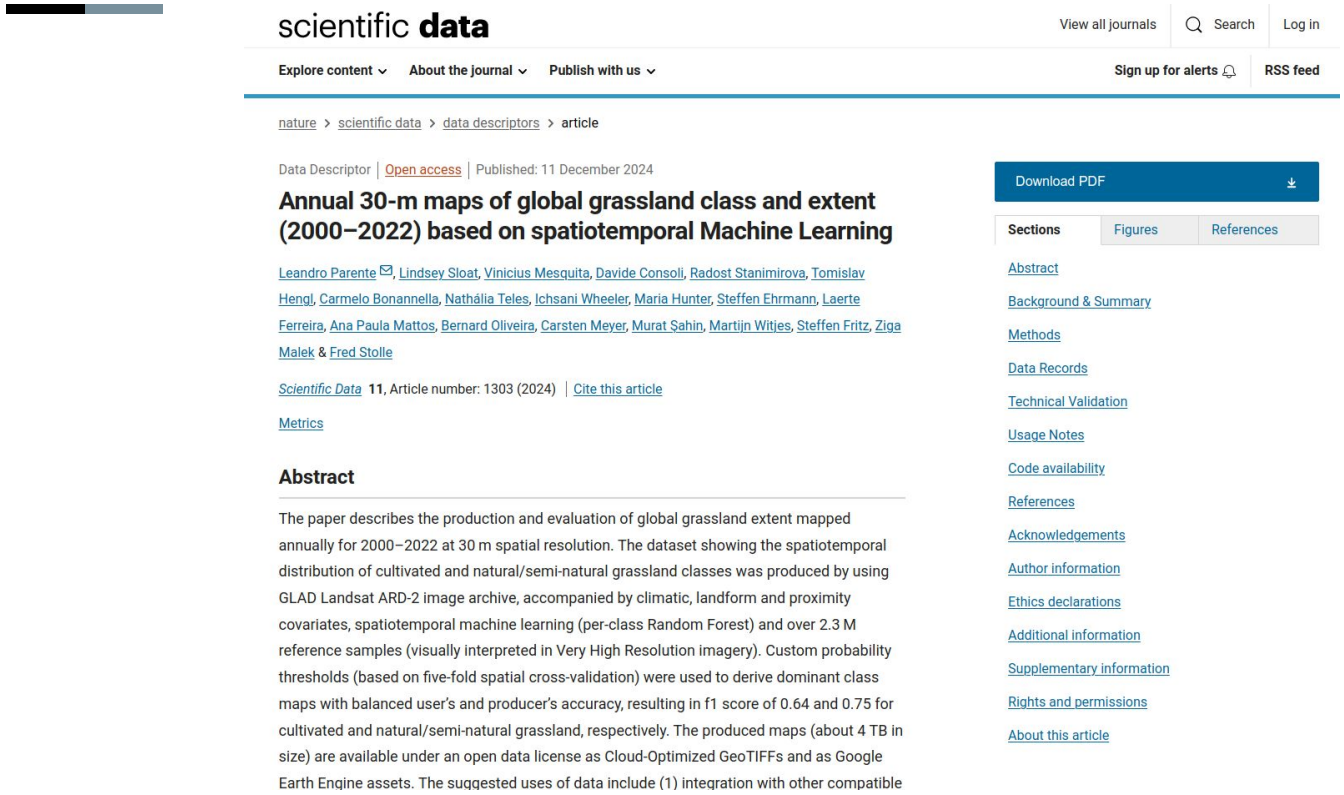
https://github.com/wri/global-pasture-watch/tree/main/ggc-30m

Classification matrix & reference samples



https://doi.org/10.5281/zenodo.13952806

Global Pasture Watch mapping



scientific data

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Sign up for alerts | RSS feed

nature > scientific data > data descriptors > article

Data Descriptor | Open access | Published: 11 December 2024

Annual 30-m maps of global grassland class and extent (2000–2022) based on spatiotemporal Machine Learning

Leandro Parente, Lindsey Sloat, Vinicius Mesquita, Davide Consoli, Radost Stanimirova, Tomislav Hengl, Carmelo Bonannella, Nathália Teles, Ichsani Wheeler, Maria Hunter, Steffen Ehrmann, Laerte Ferreira, Ana Paula Mattos, Bernard Oliveira, Carsten Meyer, Murat Şahin, Martijn Witjes, Steffen Fritz, Ziga Malek & Fred Stolle

Scientific Data **11**, Article number: 1303 (2024) | Cite this article

Metrics

Abstract

The paper describes the production and evaluation of global grassland extent 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 2.3 M 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 user's and producer's accuracy, resulting in f1 score of 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

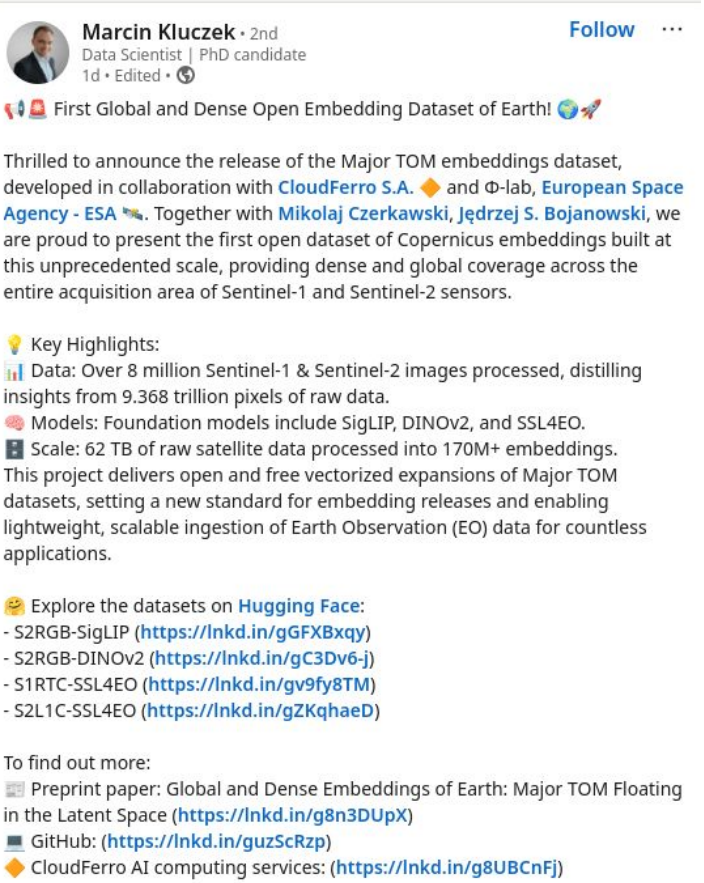
Download PDF

Sections | Figures | References

Abstract
Background & Summary
Methods
Data Records
Technical Validation
Usage Notes
Code availability
References
Acknowledgements
Author information
Ethics declarations
Additional information
Supplementary information
Rights and permissions
About this article

https://www.nature.com/articles/s41597-024-04139-6

EO Data Foundation Model



Marcin Kluczek • 2nd
Data Scientist | PhD candidate
1d • Edited •

Follow ⋯

📢🚨 First Global and Dense Open Embedding Dataset of Earth! 🌍🚀

Thrilled to announce the release of the Major TOM embeddings dataset, developed in collaboration with CloudFerro S.A. 🔶 and Φ-lab, European Space Agency - ESA 🛰️. Together with Mikolaj Czerkawski, Jędrzej S. Bojanowski, we are proud to present the first open dataset of Copernicus embeddings built at this unprecedented scale, providing dense and global coverage across the entire acquisition area of Sentinel-1 and Sentinel-2 sensors.

💡 Key Highlights:
📊 Data: Over 8 million Sentinel-1 & Sentinel-2 images processed, distilling insights from 9.368 trillion pixels of raw data.
🧠 Models: Foundation models include SigLIP, DINOv2, and SSL4EO.
🗄️ Scale: 62 TB of raw satellite data processed into 170M+ embeddings.
This project delivers open and free vectorized expansions of Major TOM datasets, setting a new standard for embedding releases and enabling lightweight, scalable ingestion of Earth Observation (EO) data for countless applications.

🤗 Explore the datasets on Hugging Face:
- S2RGB-SigLIP (https://lnkd.in/gGFXBxqy)
- S2RGB-DINOv2 (https://lnkd.in/gC3Dv6-j)
- S1RTC-SSL4EO (https://lnkd.in/gv9fy8TM)
- S2L1C-SSL4EO (https://lnkd.in/gZKqhaeD)

To find out more:
📰 Preprint paper: Global and Dense Embeddings of Earth: Major TOM Floating in the Latent Space (https://lnkd.in/g8n3DUpX)
💻 GitHub: (https://lnkd.in/guzScRzp)
🔶 CloudFerro AI computing services: (https://lnkd.in/g8UBCnFj)

clay

https://github.com/ESA-PhiLab/Major-TOM

https://clay-foundation.github.io

Main takeaways

OpenGeoHUB
Connect • Create • Share • Repeat

- There are a lot **EO data** out there. However this data is not ready to be used and integrated into environmental applications, requiring a comprehensive work to transform it into data cubes / **ARCO** mosaics,

Main takeaways

OpenGeoHUB
Connect • Create • Share • Repeat

- There are a lot **EO data** out there. However this data is not ready to be used and integrated into environmental applications, requiring a comprehensive work to transform it into data cubes / **ARCO** mosaics,
- **Cloud-native formats / protocols** (COG, zarr, MVT, STAC) came to stay. They take more advantage of new technologies, are maintained for user communities and become a "de-facto" standards,

Main takeaways

OpenGeoHUB
Connect • Create • Share • Repeat

- There are a lot **EO data** out there. However this data is not ready to be used and integrated into environmental applications, requiring a comprehensive work to transform it into data cubes / **ARCO** mosaics,
- **Cloud-native formats / protocols** (COG, zarr, MVT, STAC) came to stay. They take more advantage of new technologies, are maintained for user communities and become a "de-facto" standards,
- There are multiple **ML open-source frameworks** out there. However a proper integration of these frameworks into processing workflows and applications requires customized code and additional libraries (scikit-map, sits),

Main takeaways

OpenGeoHUB
Connect • Create • Share • Repeat

- There are a lot **EO data** out there. However this data is not ready to be used and integrated into environmental applications, requiring a comprehensive work to transform it into data cubes / **ARCO** mosaics,
- **Cloud-native formats / protocols** (COG, zarr, MVT, STAC) came to stay. They take more advantage of new technologies, are maintained for user communities and become a "de-facto" standards,
- There are multiple **ML open-source frameworks** out there. However a proper integration of these frameworks into processing workflows and applications requires customized code and additional libraries (scikit-map, sits),
- OpenGeoHub is providing **terabytes** of ARCO datasets that can be easily accessed and integrated into **ML pipelines** for different applications (including GEE).

Main takeaways

OpenGeoHUB
Connect • Create • Share • Repeat

- There are a lot **EO data** out there. However this data is not ready to be used and integrated into environmental applications, requiring a comprehensive work to transform it into data cubes / **ARCO** mosaics,
- **Cloud-native formats / protocols** (COG, zarr, MVT, STAC) came to stay. They take more advantage of new technologies, are maintained for user communities and become a "de-facto" standards,
- There are multiple **ML open-source frameworks** out there. However a proper integration of these frameworks into processing workflows and applications requires customized code and additional libraries (scikit-map, sits),
- OpenGeoHub is providing **terabytes** of ARCO datasets that can be easily accessed and integrated into **ML pipelines** for different applications (including GEE).
- The full reproducibility of this pipeline depends on have the **input EO data** publicly accessible, **container with all libraries** used and **better code abstraction** for the optimizations implemented.

Leandro Parente

leandro.parente@opengeohub.org

https://opengeohub.org

OpenGeoHUB

Connect • Create • Share • Repeat