OpenGeoHUB
Connect • Create • Share • Repeat

Soil Type Classification across Europe: Embeddings vs Environmental Covariates

23 June 2026

Mustafa Serkan Isik

serkan.isik@opengeohub.org

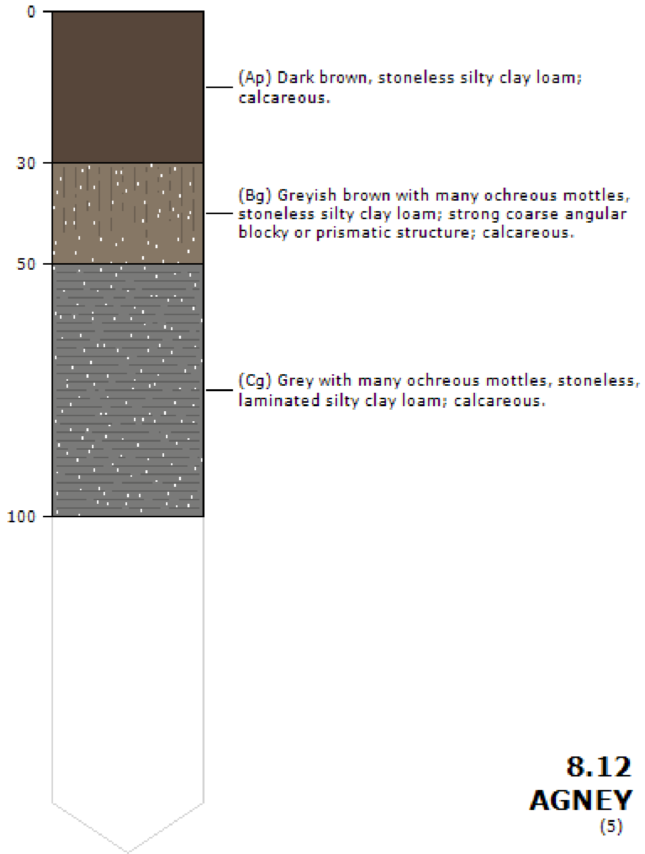
AI 4 Soil Health

OPEN EARTH MONITOR

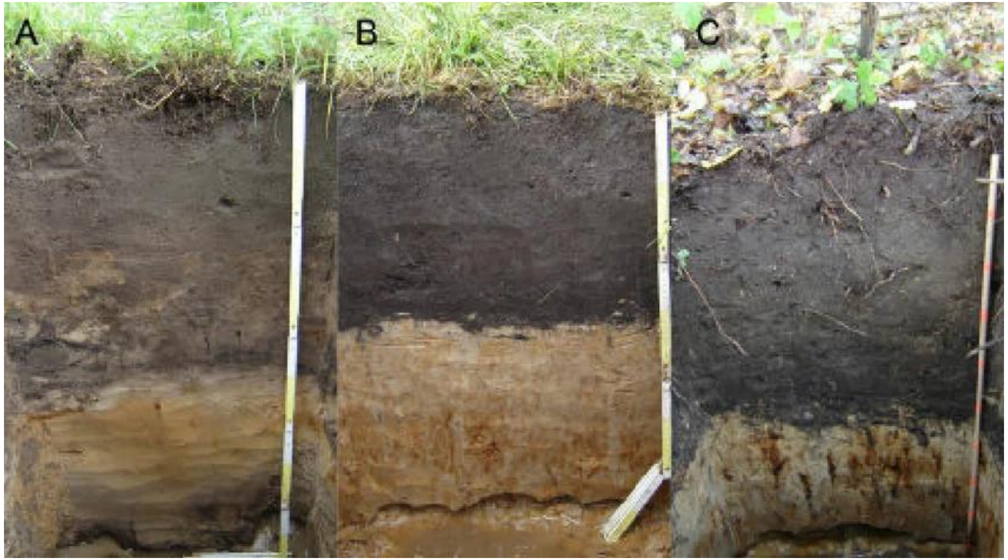
Soil Type Mapping

[WRB soil classification system] (Int. Union of Soil Science)

Example WRB soil type



Fluvic Calcaric **Gleysols**

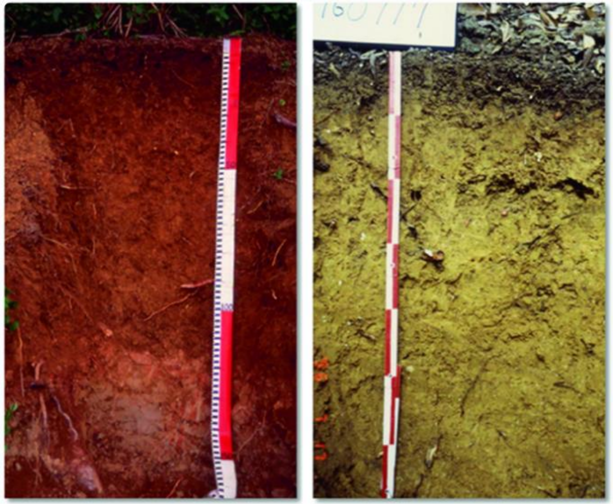


show a photo of "Fluvic Calcaric Gleysols"

Thought for 9s

Fluvic Calcaric Gleysols are a type of soil in the World Reference Base (WRB) classification. They are **Gleysols** (waterlogged soils with reducing conditions, often showing greyish/blueish mottling) that are **fluvic** (formed in alluvial/fluvial sediments with stratification) and **calcaric** (containing significant calcium carbonate/lime).

Here are representative photos of similar Gleysol profiles (exact labeled photos of "Fluvic Calcaric Gleysols" are rare online, but these show typical features):



Major Soil Types | Springer Nature Link

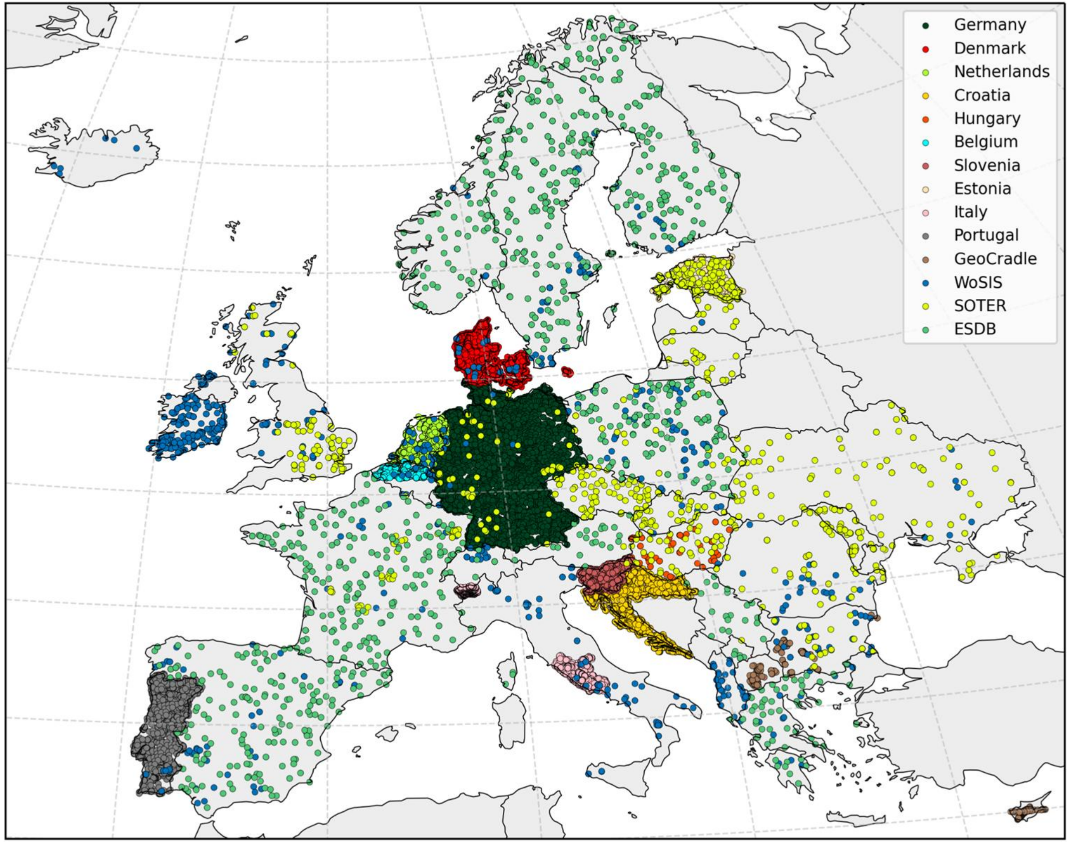
This shows soil profiles with gleyic (water-influenced) characteristics — note the color variations indicating periodic saturation.

For a more specific example, the **Newchurch soil series** in England is classified as **Clayic Fluvic Calcaric Gleysols**. You can view a real photo here: Flickr - Newchurch soil profile. Flickr

Understanding soil types is extremely useful. Soil types integrate:

- Morphological characteristics,
- Physical and chemical properties,
- Soil-forming factors and processes,
- Land use limitations and best practices,

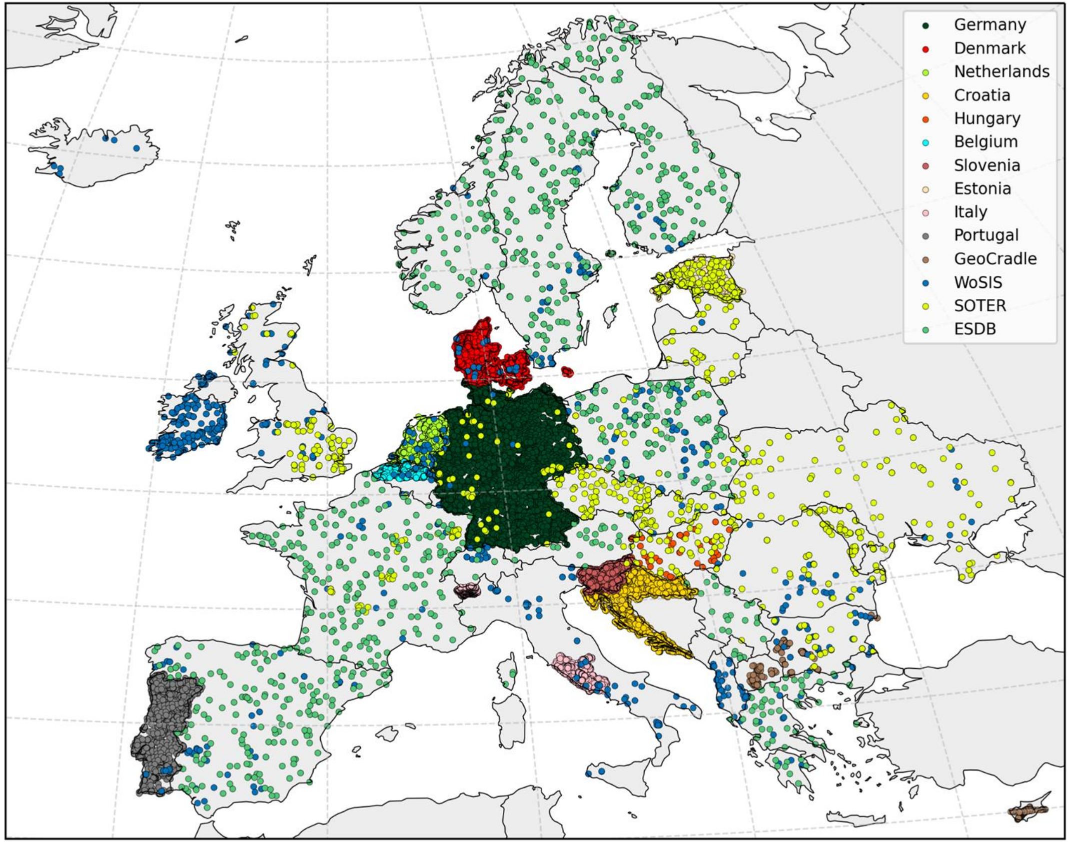
Harmonization of Soil Types Samples



Name	
Germany	Open Agrar
Belgium	AARDEWERK
Netherlands	PDOK
Slovenia	OpenLandMap
Portugal	INFOSOLO
GeoCradle	GeoCradle
SOTER	ISRIC
WoSIS	ISRIC
ESDB v2	This dataset was synthetically generated from ESDB v2 maps

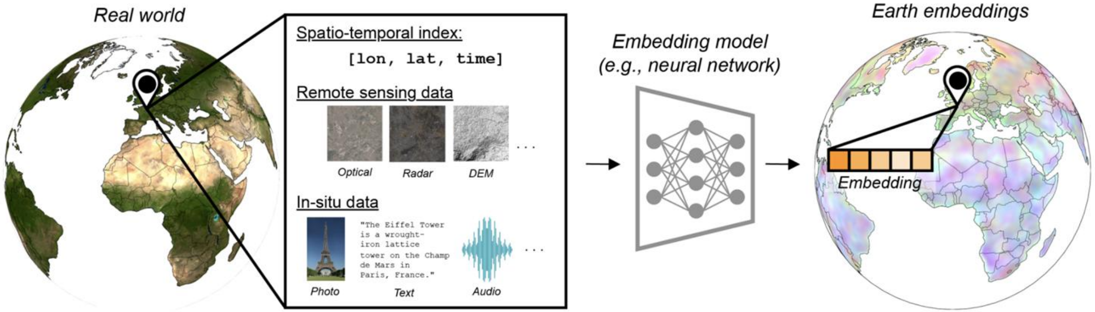
Just arrived: 85k new points from UK!

Harmonization of Soil Types Samples



- Reprojection of coordinates into the common **European reference system** (EPSG:3035)
- Alignment (translation) of soil class information to the IUSS WRB classification
- Normalization of attribute structure and metadata fields.
- Some **duplicate records** (with different WRB classification) were intentionally retained in the dataset
- **Pseudo-training samples** generated for sparse regions (notably parts of Spain, France, Poland, the Balkans, and Scandinavia)
- **Overrepresented countries** were downsampled to prevent class and spatial imbalance (e.g. NL with >200k points)
- In total ~20k training samples

AlphaEarth Embeddings



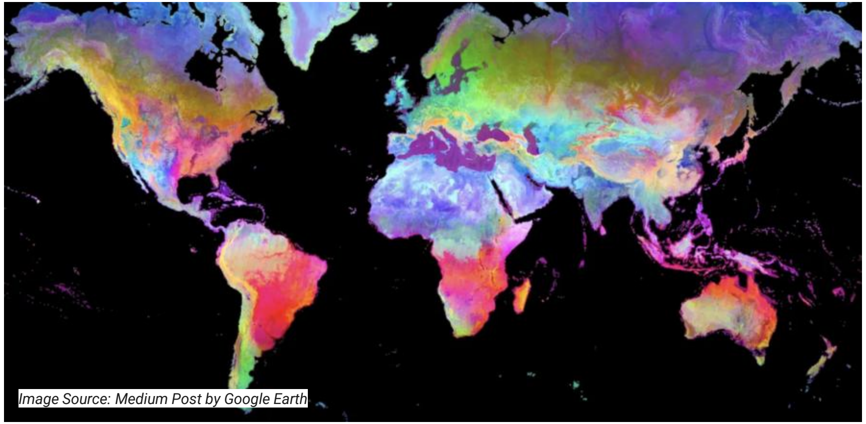


Image Source: Medium Post by Google Earth

Optical Imagery: Time-series streams from Sentinel-2 and Landsat 8/9.

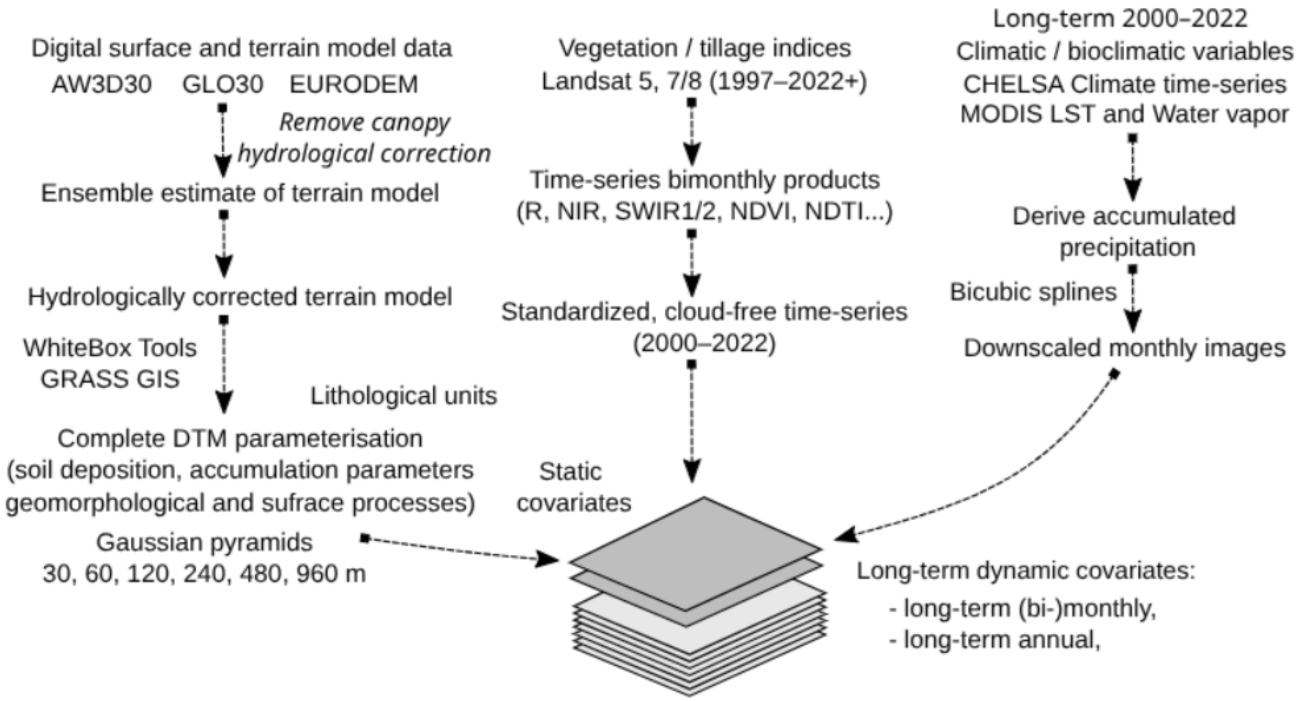
Radar (SAR): Sentinel-1 and ALOS PALSAR-2, used to penetrate cloud cover and track surface texture.

LiDAR: Canopy height and structural vegetation data from GEDI.

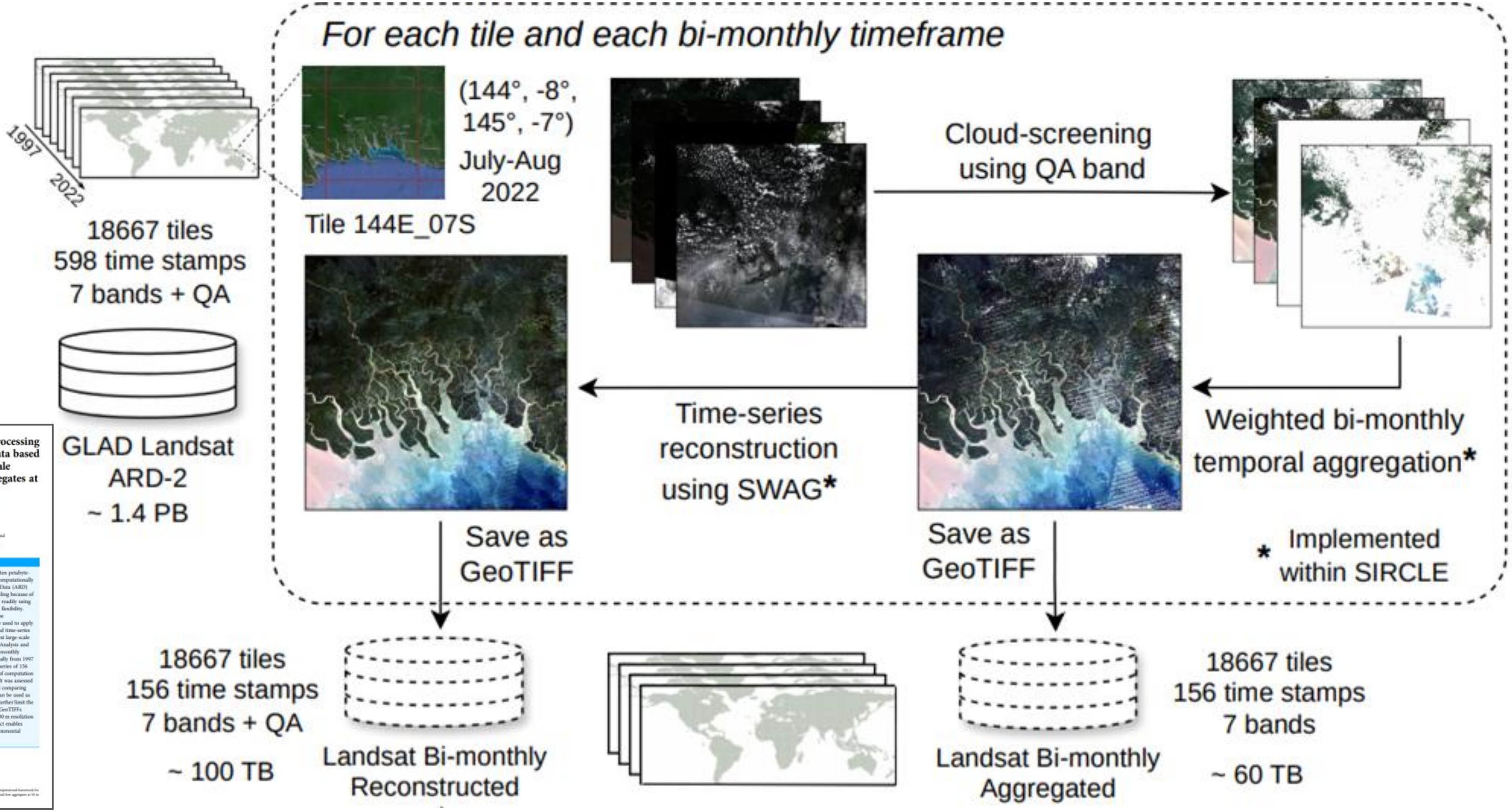
Environmental & Climate Data: ERA5-Land climate reanalysis, GLO-30 digital elevation model, and GRACE gravity fields.

Semantic Signals: Unstructured text, including geocoded Wikipedia articles and GBIF species observations.

Environmental Covariates



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

Davide Consoli[1], Leandro Parente[1], Rolf Simoes[1], Murat Şahin[1], Xuemeng Tian[1,2], Martijn Witjes[1,2], Lindsey Sloat[3] and Tomislav Hengl[1]

[1] OpenGeoHub Foundation, Doorwerth, Netherlands
[2] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University and Research, Wageningen, Netherlands
[3] Land & Carbon Lab, World Resources Institute (WRI), Washington DC, United States

ABSTRACT

Processing large collections of earth observation (EO) time-series, often petabyte-sized, such as NASA's Landsat and ESA's Sentinel missions, can be computationally prohibitive and costly. Despite their name, even the Analysis Ready Data (ARD) versions of such collections can rarely be used as direct input for modeling because of cloud presence and/or prohibitive storage size. Existing solutions for readily using these data are not openly available, are poor in performance, or lack flexibility. Addressing this issue, we developed TSIRF (Time-Series Iteration-free Reconstruction Framework), a computational framework that can be used to apply diverse time-series processing tasks, such as temporal aggregation and time-series reconstruction by simply adjusting the convolution kernel. As the first large-scale application, TSIRF was employed to process the entire Global Land Analysis and Discovery (GLAD) ARD Landsat archive, producing a cloud-free bi-monthly aggregated product. This process, covering seven Landsat bands globally from 1997 to 2022, with more than two trillion pixels and for each one a time-series of 156 samples in the aggregated product, required approximately 28 hours of computation using 1248 Intel® Xeon® Gold 6248R CPUs. The quality of the result was assessed using a benchmark dataset derived from the aggregated product and comparing different imputation strategies. The resulting reconstructed images can be used as input for machine learning models or to map biophysical indices. To further limit the storage size the produced data was saved as 8-bit Cloud-Optimized GeoTIFFs (COG). With the hosting of about 20 TB per band/index for an entire 30 m resolution bi-monthly historical time-series distributed as open data, the product enables seamless, fast, and affordable access to the Landsat archive for environmental monitoring and analysis applications.

Submitted 22 May 2024
Accepted 4 November 2024
Published 4 December 2024

Corresponding author
Davide Consoli,
davide.consoli@opengeohub.org

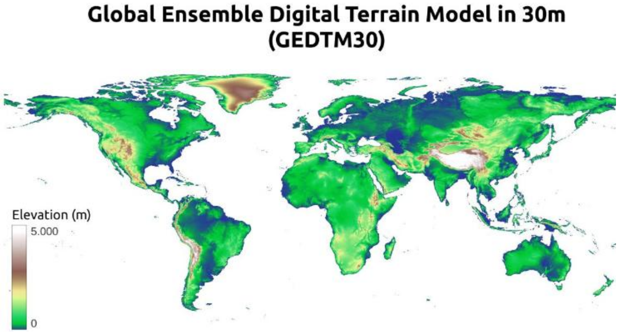
Academic editor
Zhe Zhu

Additional Information and Declarations can be found on page 37

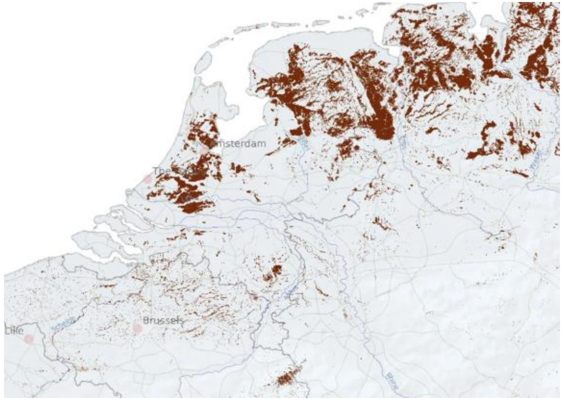
DOI 10.7717/peerj.18585

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

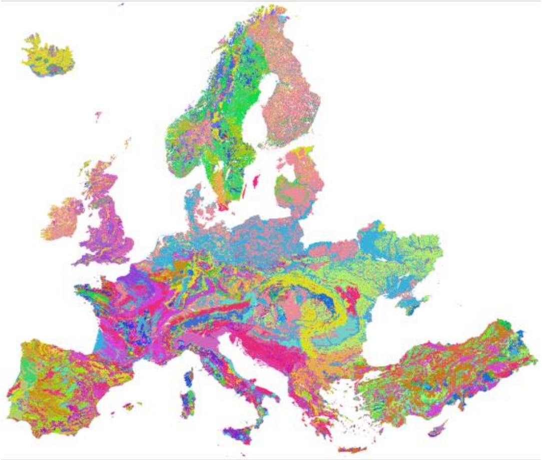
Other Covariates



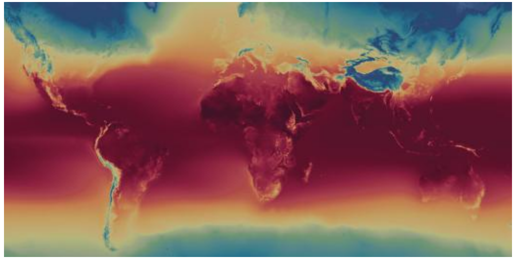
GEDTM by Ho et al. (2025) - 10.7717/peerj.19673



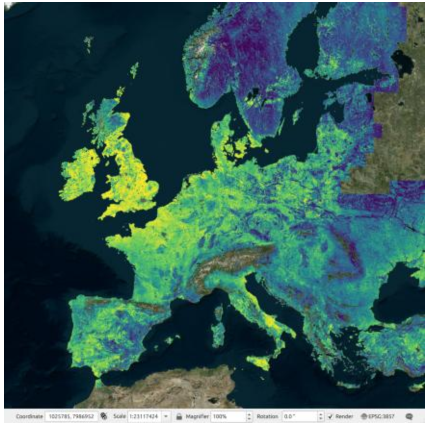
Europe Peatland Map 2025



Lithology based on EGDI / OneGeology map 10.5281/zenodo.12607973

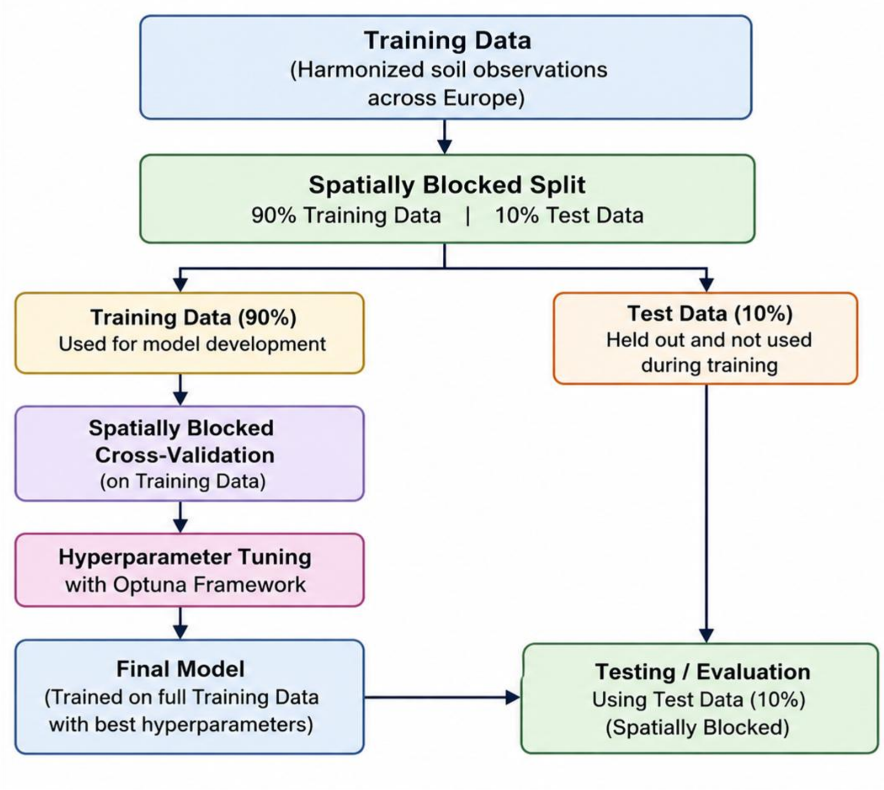


CHELSA (1km) monthly + longterm



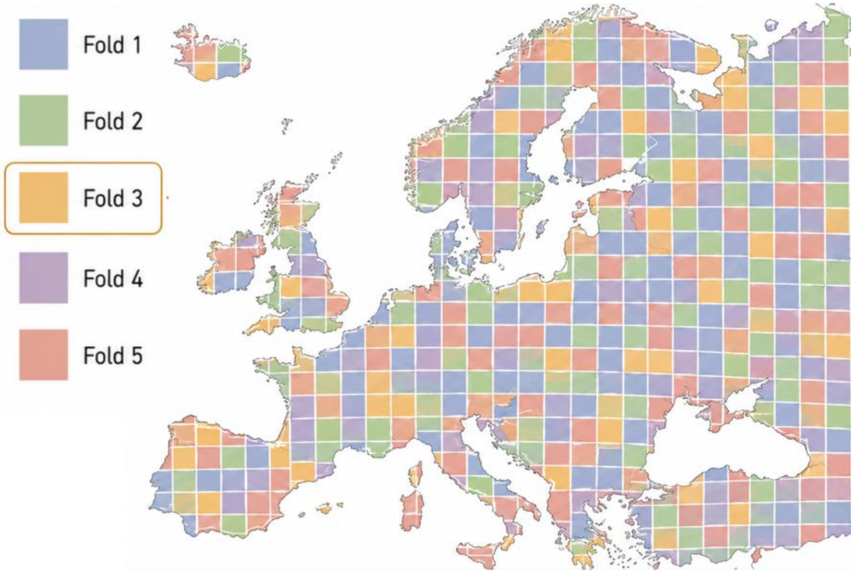
Long term soil moisture (1km) 10.5281/zenodo.14833053

Experimental Setup

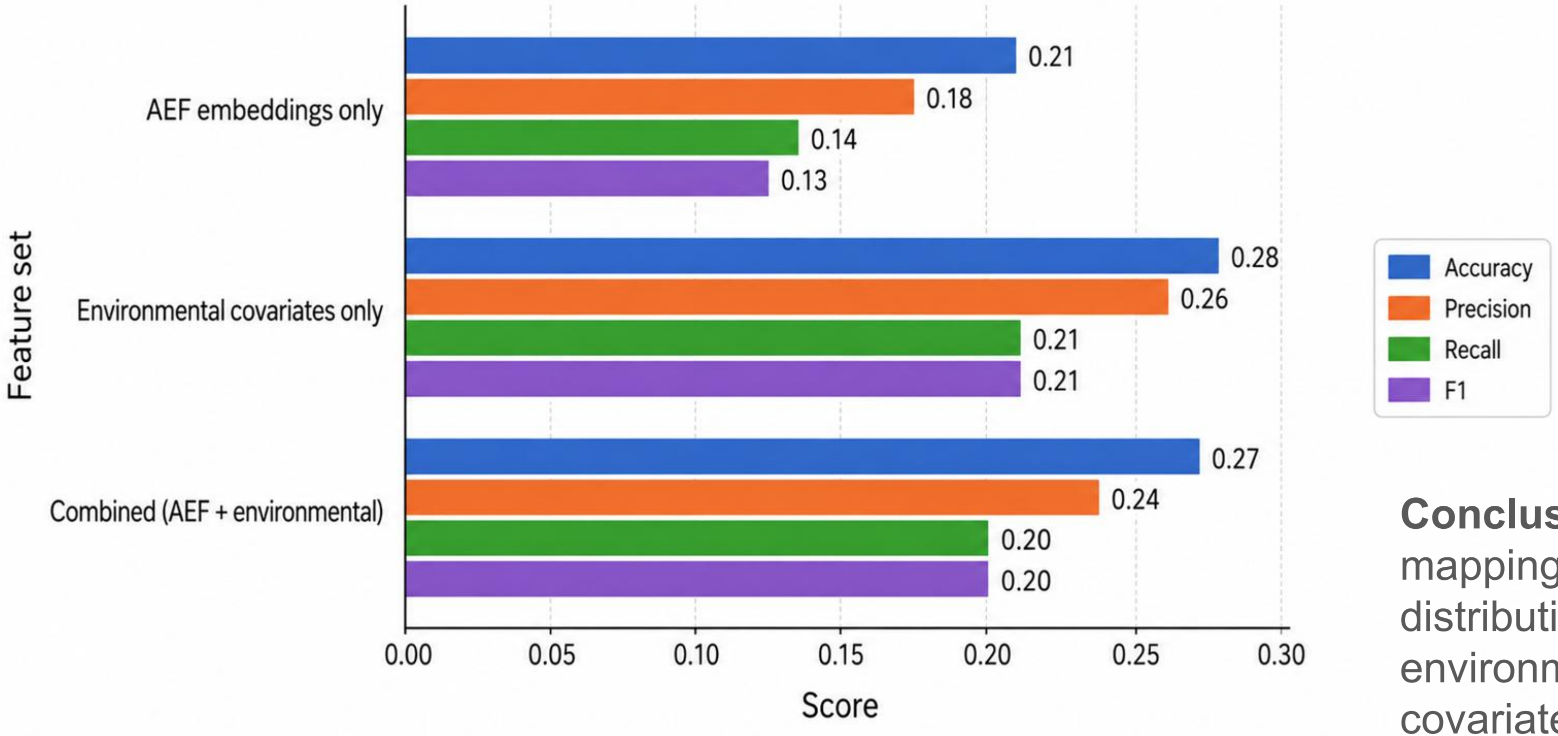


Explainable Boosting Machines (EBMs)

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

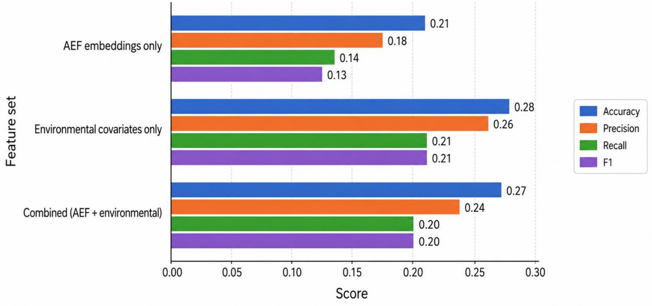


Results

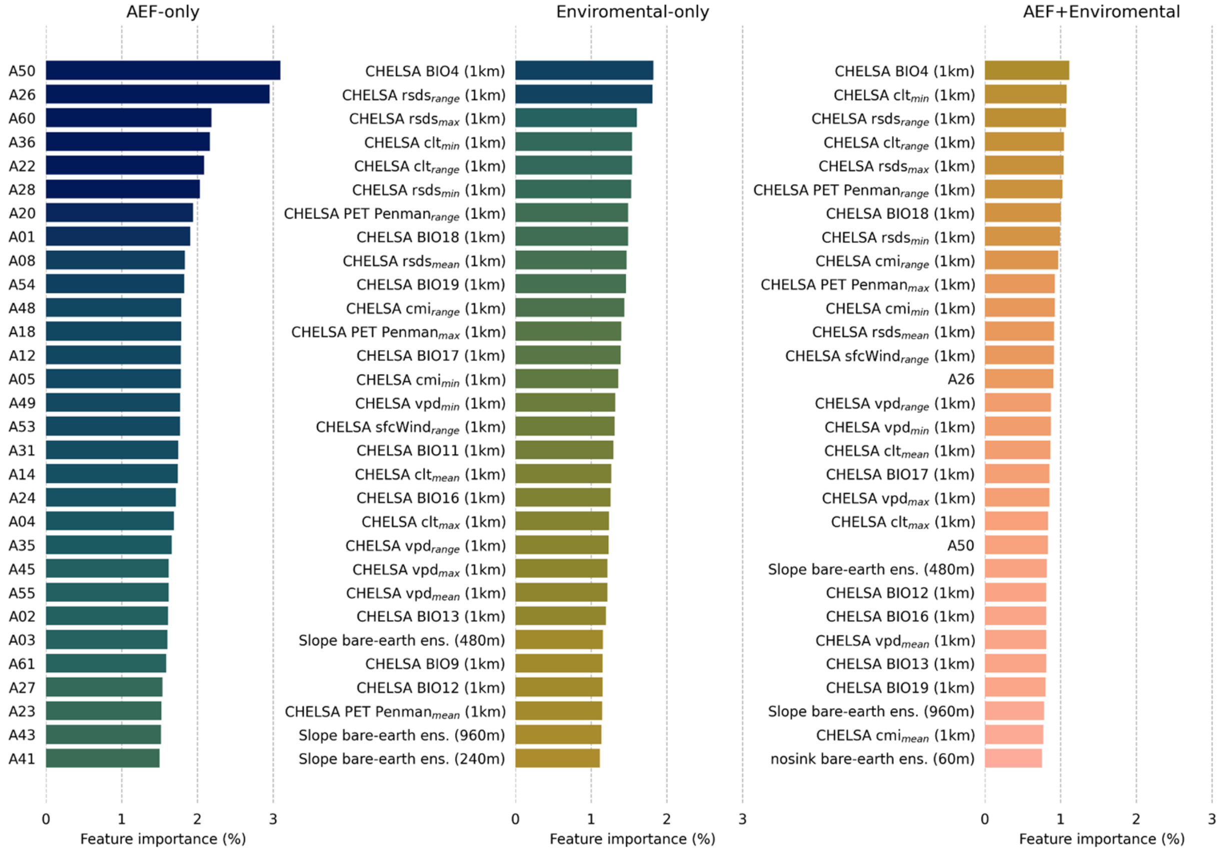


Conclusion: for mapping spatial distribution of soil types, environmental covariates (climate, terrain etc) are the most important.

Results

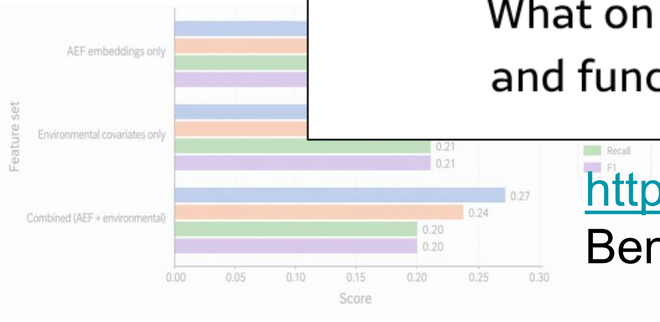


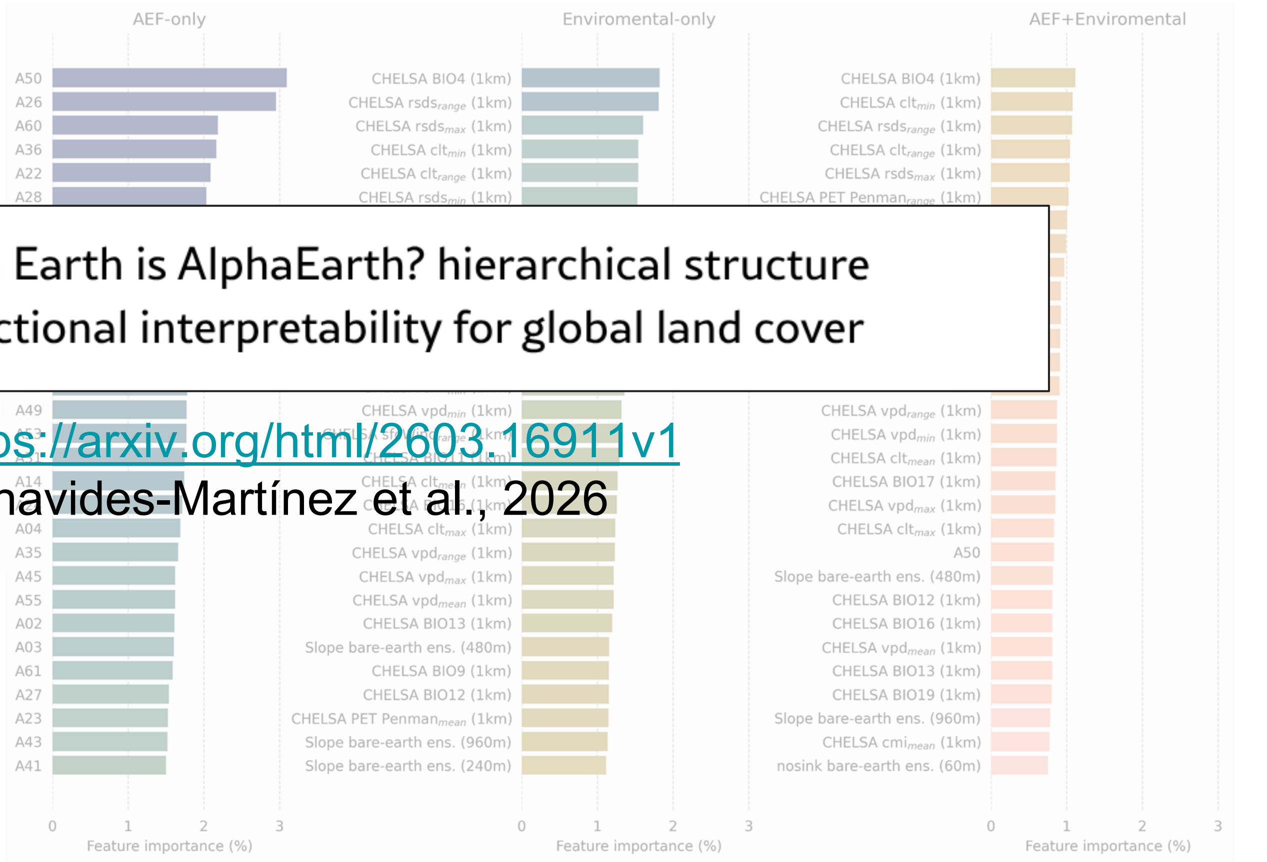
AEF embeddings only
Environmental covariates only
Combined (AEF + environmental)
Feature set
0.21
0.18
0.14
0.13
0.28
0.26
0.21
0.21
0.27
0.24
0.20
0.20
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Score
Accuracy
Precision
Recall
F1



AEF-only
A50
A26
A60
A36
A22
A28
A20
A01
A08
A54
A48
A18
A12
A05
A49
A53
A31
A14
A24
A04
A35
A45
A55
A02
A03
A61
A27
A23
A43
A41
0
1
2
3
Feature importance (%)
Enviromental-only
CHELSA BIO4 (1km)
CHELSA $rsds_{range}$ (1km)
CHELSA $rsds_{max}$ (1km)
CHELSA clt_{min} (1km)
CHELSA clt_{range} (1km)
CHELSA $rsds_{min}$ (1km)
CHELSA PET $Penman_{range}$ (1km)
CHELSA BIO18 (1km)
CHELSA $rsds_{mean}$ (1km)
CHELSA BIO19 (1km)
CHELSA cmi_{range} (1km)
CHELSA PET $Penman_{max}$ (1km)
CHELSA BIO17 (1km)
CHELSA cmi_{min} (1km)
CHELSA vpd_{min} (1km)
CHELSA $sfcWind_{range}$ (1km)
CHELSA BIO11 (1km)
CHELSA clt_{mean} (1km)
CHELSA BIO16 (1km)
CHELSA clt_{max} (1km)
CHELSA vpd_{range} (1km)
CHELSA vpd_{max} (1km)
CHELSA vpd_{mean} (1km)
CHELSA BIO13 (1km)
Slope bare-earth ens. (480m)
CHELSA BIO9 (1km)
CHELSA BIO12 (1km)
CHELSA PET $Penman_{mean}$ (1km)
Slope bare-earth ens. (960m)
Slope bare-earth ens. (240m)
0
1
2
3
Feature importance (%)
AEF+Enviromental
CHELSA BIO4 (1km)
CHELSA clt_{min} (1km)
CHELSA $rsds_{range}$ (1km)
CHELSA clt_{range} (1km)
CHELSA $rsds_{max}$ (1km)
CHELSA PET $Penman_{range}$ (1km)
CHELSA BIO18 (1km)
CHELSA $rsds_{min}$ (1km)
CHELSA cmi_{range} (1km)
CHELSA PET $Penman_{max}$ (1km)
CHELSA cmi_{min} (1km)
CHELSA $rsds_{mean}$ (1km)
CHELSA $sfcWind_{range}$ (1km)
A26
CHELSA vpd_{range} (1km)
CHELSA vpd_{min} (1km)
CHELSA clt_{mean} (1km)
CHELSA BIO17 (1km)
CHELSA vpd_{max} (1km)
CHELSA clt_{max} (1km)
A50
Slope bare-earth ens. (480m)
CHELSA BIO12 (1km)
CHELSA BIO16 (1km)
CHELSA vpd_{mean} (1km)
CHELSA BIO13 (1km)
CHELSA BIO19 (1km)
Slope bare-earth ens. (960m)
CHELSA cmi_{mean} (1km)
nosink bare-earth ens. (60m)
0
1
2
3
Feature importance (%)

Results



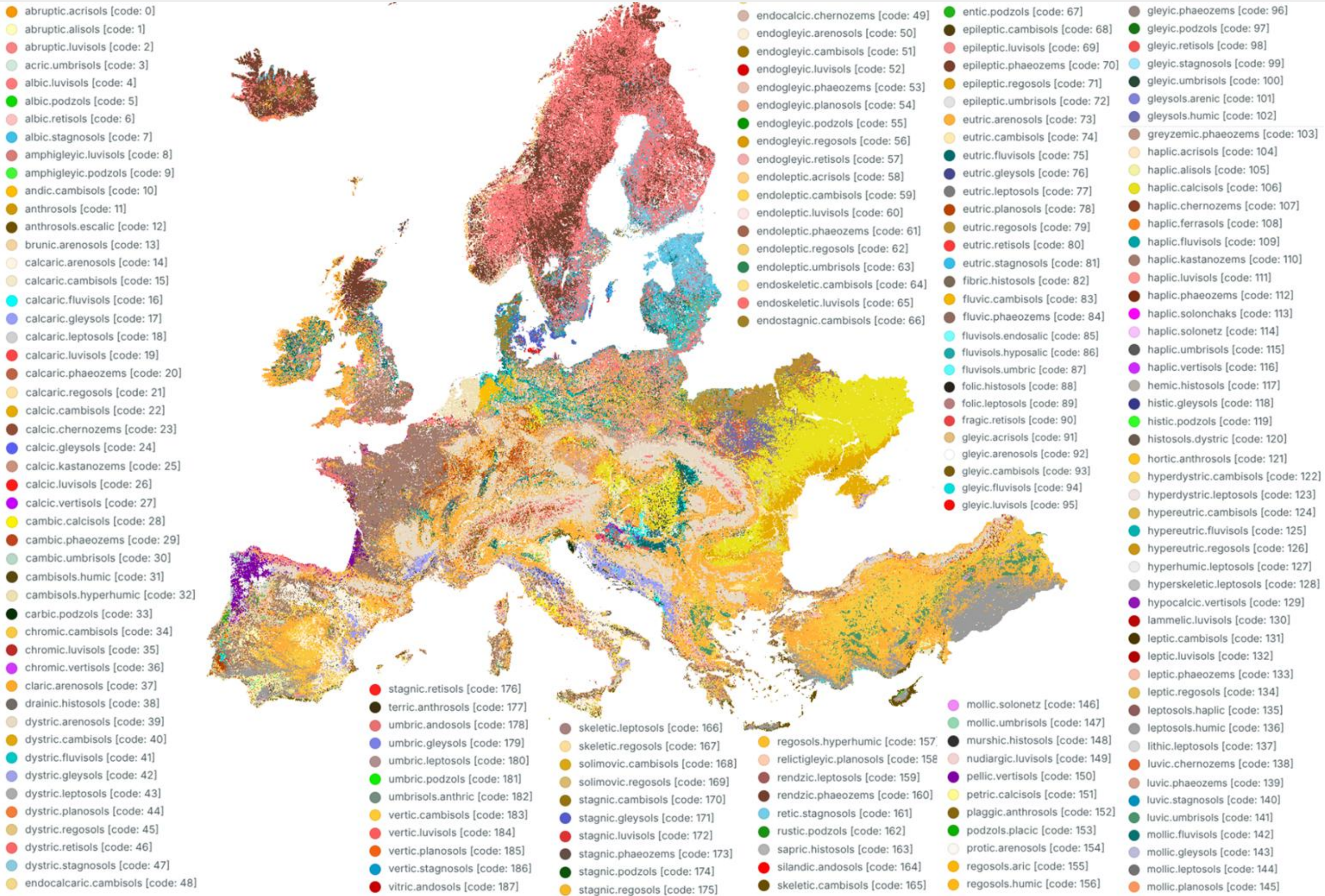


What on Earth is AlphaEarth? hierarchical structure and functional interpretability for global land cover

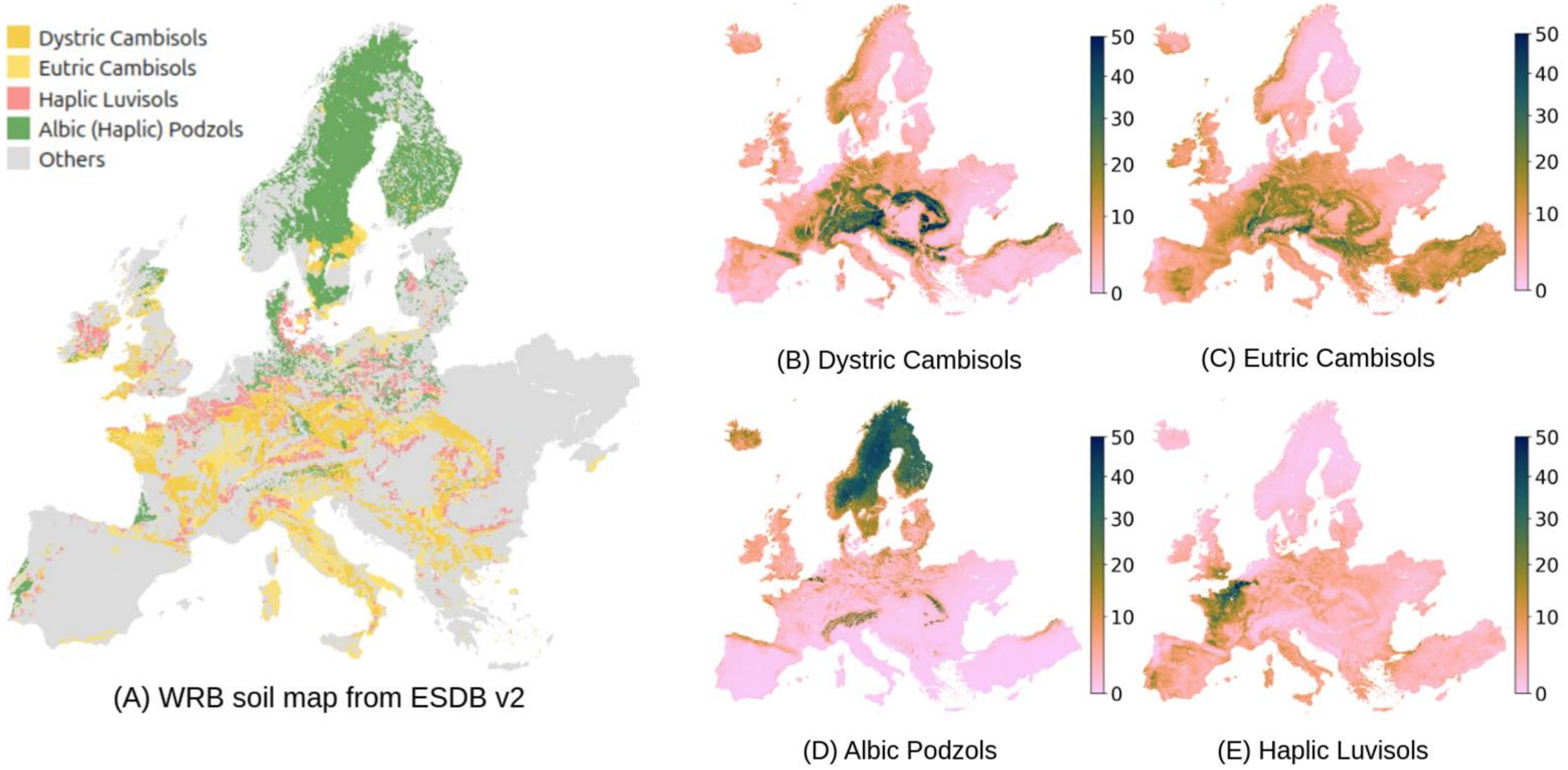
https://arxiv.org/html/2603.16911v1

Benavides-Martínez et al., 2026

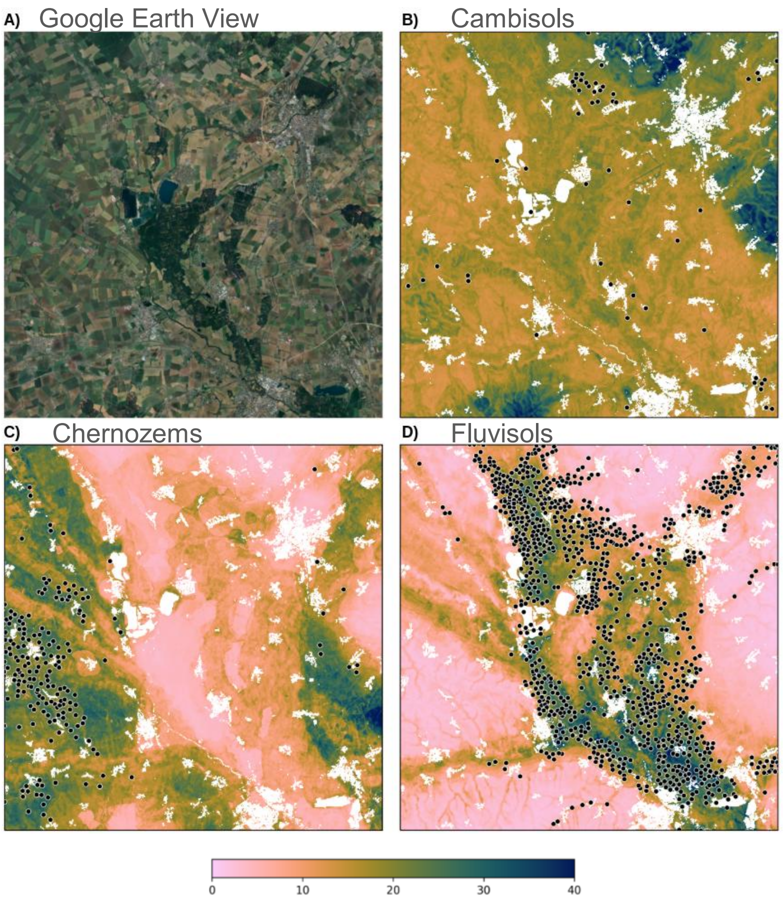
Soil Types Map — dominant class (188)



Soil Types Map — probabilities

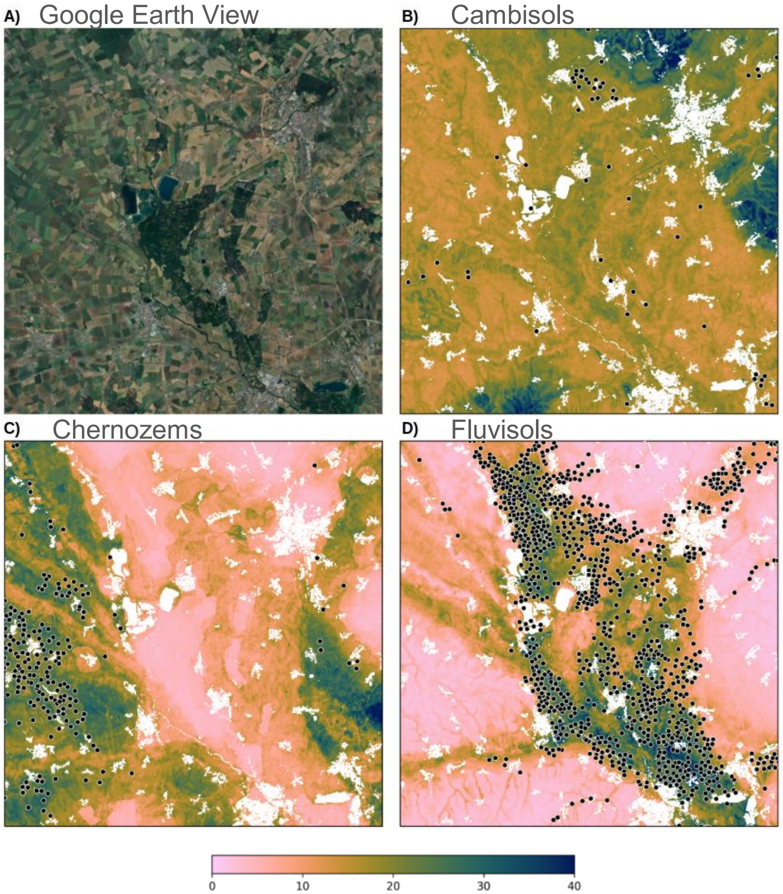


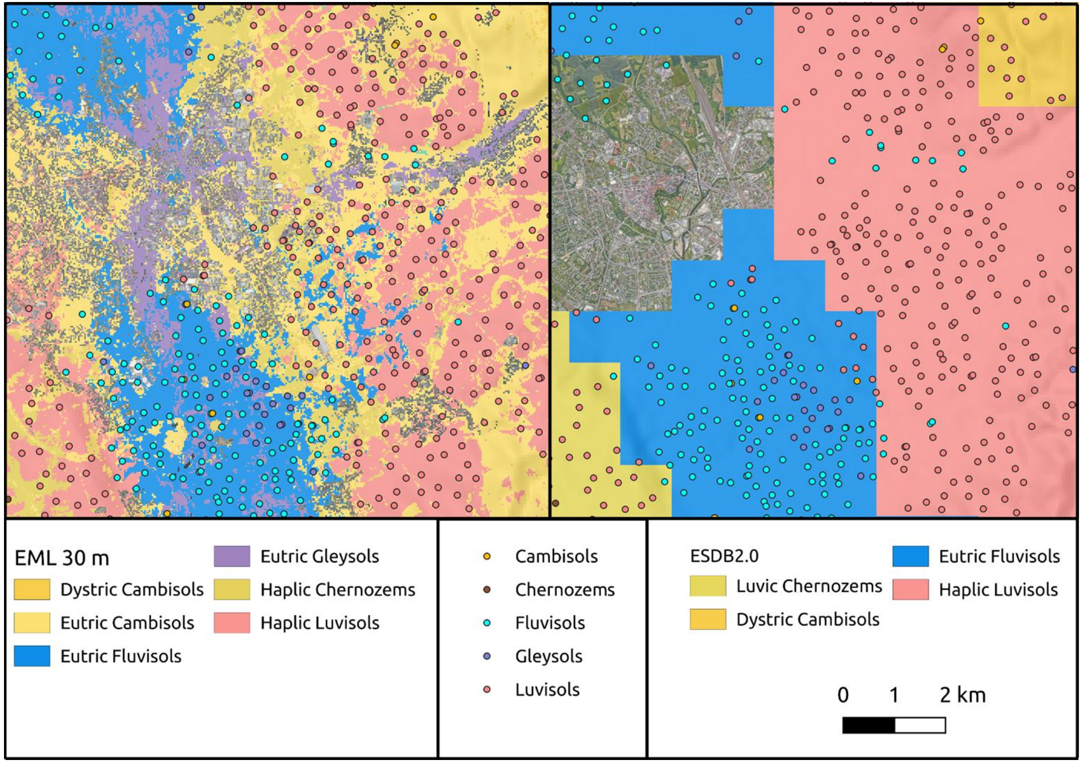
Comparisons



Comparisons

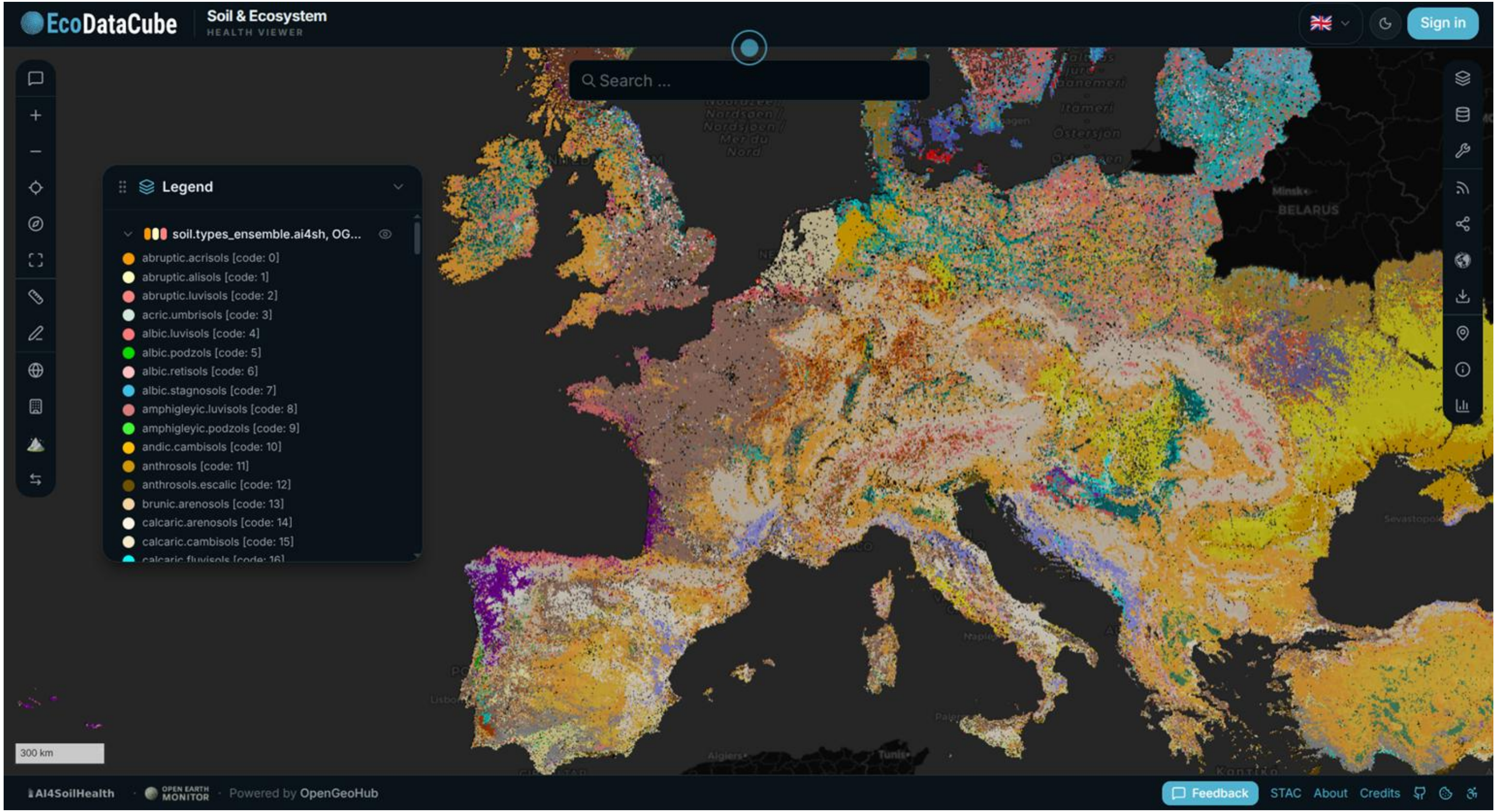
OpenGeoHUB
Connect · Create · Share · Repeat





Comparison with ESDB v2

EcoDataCube.eu



https://EcoDataCube.eu

OpenGeoHUB
Connect • Create • Share • Repeat

EcoDataCube

EcoDataCube

Browse

Description

Spatio-Temporal Asset Catalog for European-wide layers provided by EcoDataCube.

Catalogs 2/163

Tiles | List | Ascending | Descending

soil ty

Select keywords

AI4SoilHealth: Soil type dominant class

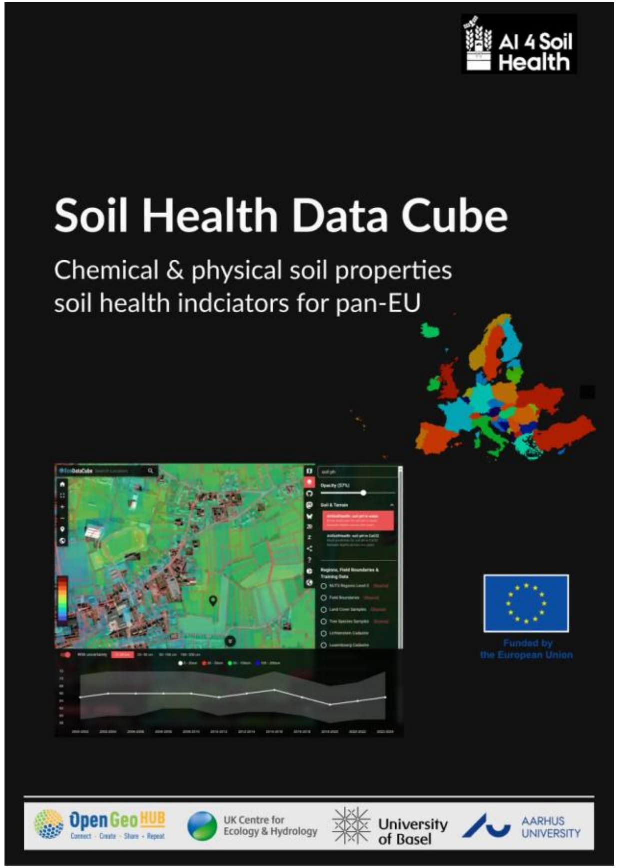
Ensemble prediction of hard classification soil type maps, stable over time

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

AI4SoilHealth: Soil type probability

Ensemble prediction of the probability of a specific soil type, stable over time

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



Conclusions & Future Work

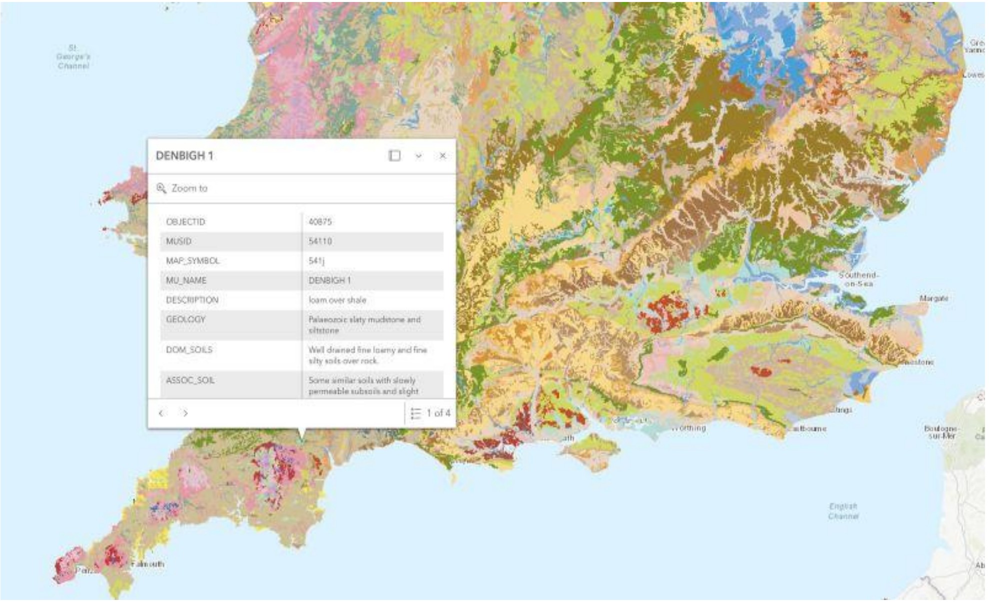
- Environmental covariates performed best. For soil type-mapping, AlphaEarth embeddings alone were weaker.
- Adding embeddings did not improve accuracy. They mainly duplicated broad environmental signals.
- Current embeddings should complement, not replace, interpretable covariates for certain tasks
- Alternative foundation-model representations and their combinations (*see van der Plas et al. Better Together: Evaluating the Complementarity of Earth Embedding Models*)
- Task-specific adaptation.
- Class- and region-specific analysis.
- Additional baseline comparisons needed in other applications and with other state-of-art models.

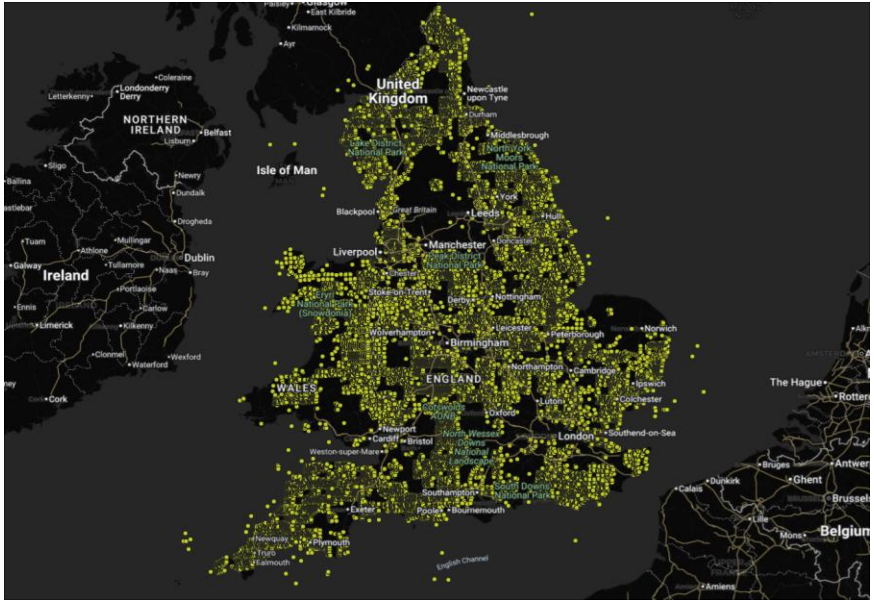
Conclusions & Future Work

Just arrived: 85k new points from UK!

UK has recently open soil database! (thanks to Cranfield University)

1) UK soil samples are the highest quality WRB training points for EU
2) New version of Pan-EU Soil Types map will be available soon!





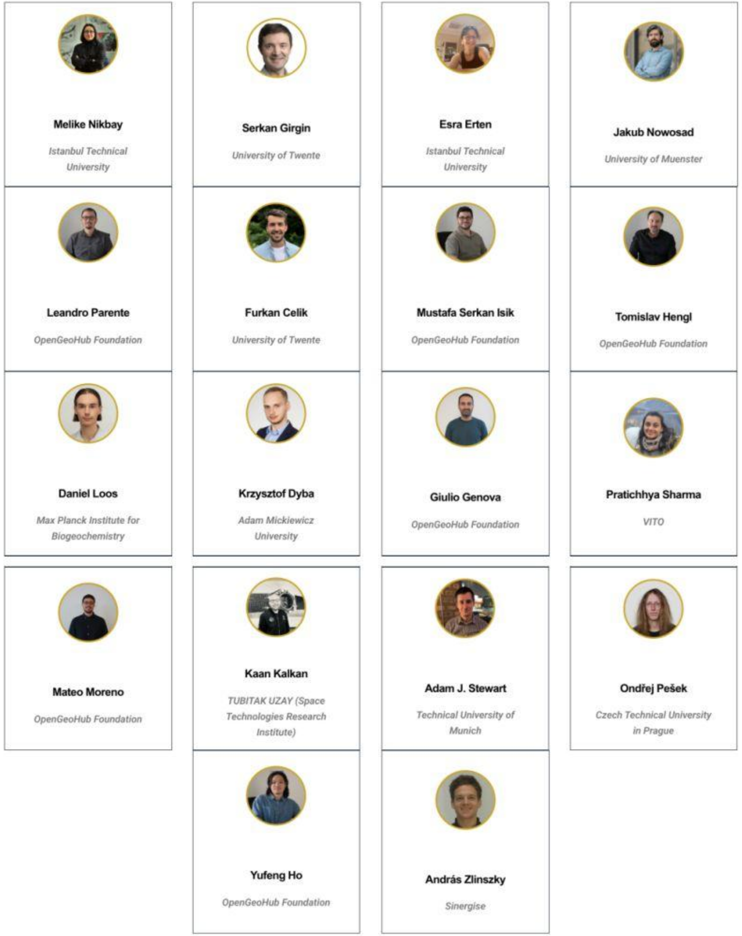
Upcoming event

Find more information about it on

https://opengeohub.org/2026/01/29/earth-observation-summer-school-2026/

Register directly via

https://pretix.eu/opengeohubfoundation/EOSummerSchool2026/



Funded by the European Union

Thank you!

This work was supported by the European Union's Horizon Europe Research and Innovation Programme through:

- ***AI4SoilHealth*** (No. 101086179)
- ***Open-Earth-Monitor Cyberinfrastructure*** (No. 101059548)

OpenGeoHUB
Connect • Create • Share • Repeat

AI 4 Soil Health

OPEN EARTH MONITOR

Remote Sensing

Special Issue

Multimodal Remote Sensing and Big Data Analytics for Earth Observation

Guest Editors
Dr. Gizem Senel
Dr. Mehmet Furkan Celik
Dr. Mustafa Serkan Isik

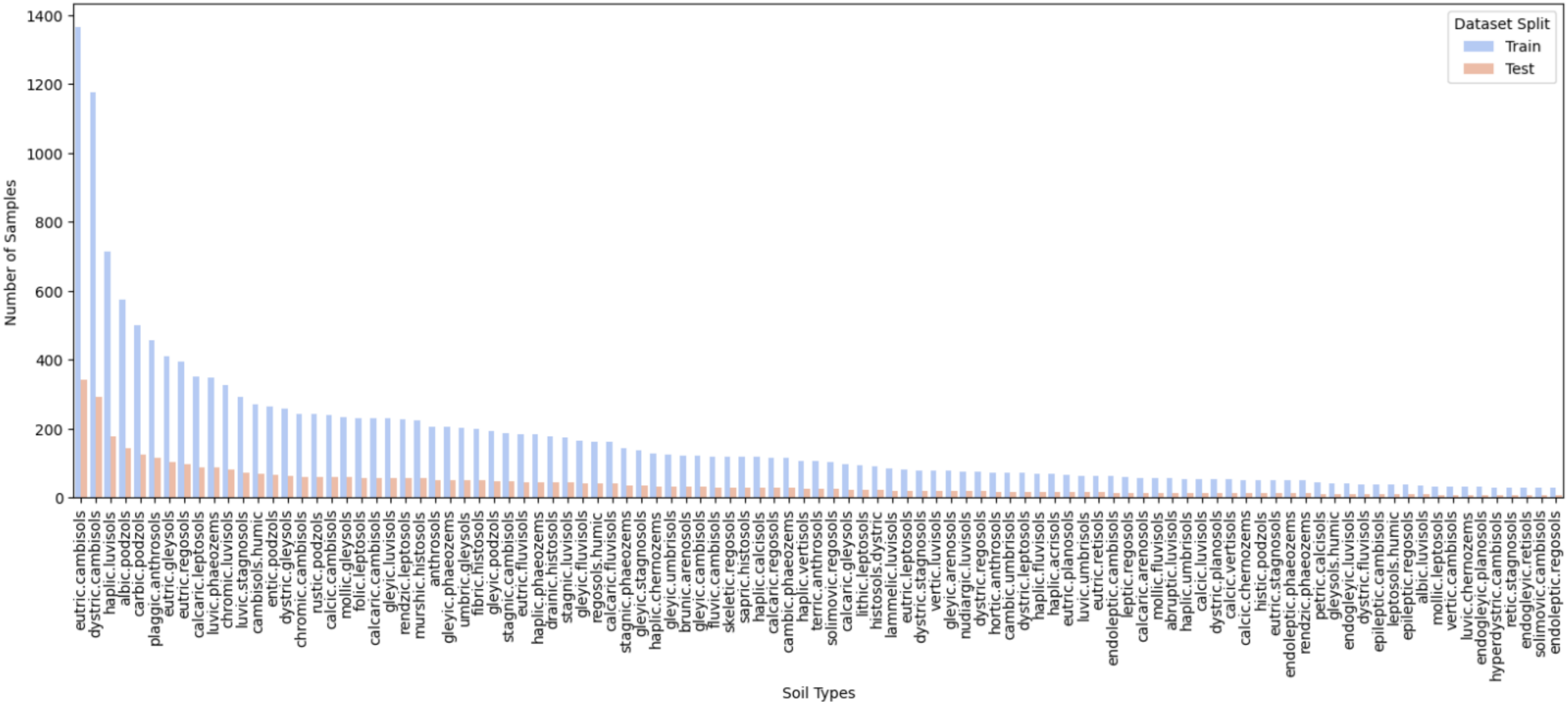
Deadline
30 November 2026

mdpi.com/si/280583

Remote Sensing
Editorial Office
MDPI, Grosspeteranlage 5
4052 Basel, Switzerland
Tel: +41 61 683 77 34
remotesensing@mdpi.com

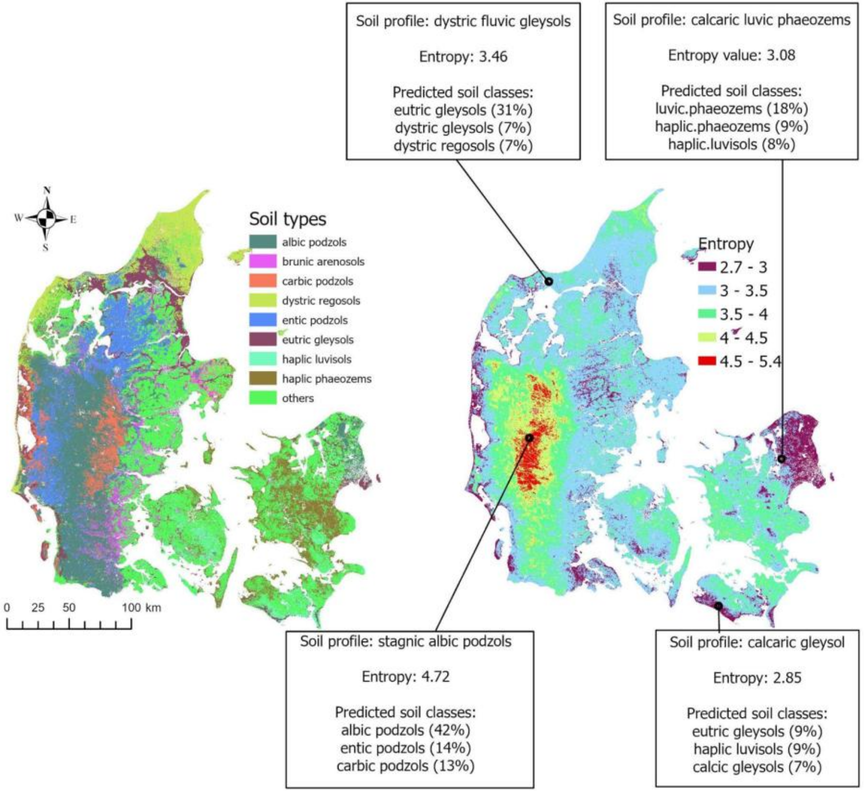
mdpi.com/journal/remotesensing

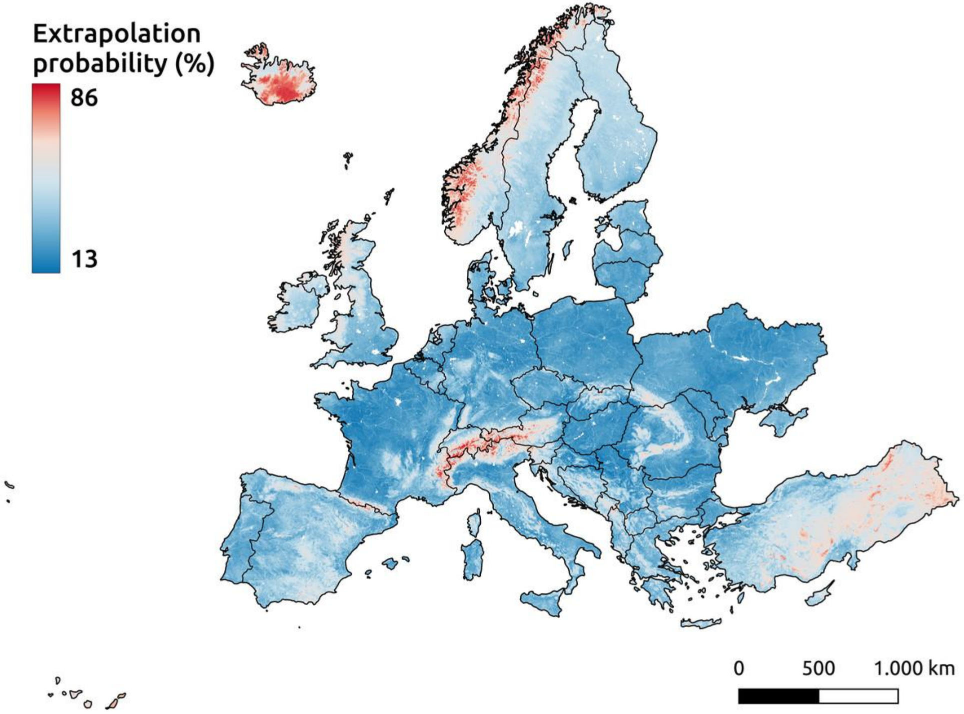
Split



Dataset Split
Train
Test
Number of Samples
0
200
400
600
800
1000
1200
1400
Soil Types
eutric.cambisols
dystric.cambisols
haplic.luvisols
albic.podzols
carbic.podzols
plaggic.anthrosols
eutric.gleysols
eutric.regosols
calcaric.leptosols
luvic.phaeozems
chromic.luvisols
luvic.stagnosols
cambisols.humic
entic.podzols
dystric.gleysols
chromic.cambisols
rustic.podzols
calcic.cambisols
mollic.gleysols
folic.leptosols
calcaric.cambisols
gleyic.luvisols
rendzic.leptosols
murshic.histosols
anthrosols
gleyic.phaeozems
umbric.gleysols
fibric.histosols
gleyic.podzols
stagnic.cambisols
eutric.fluvisols
haplic.phaeozems
drainic.histosols
stagnic.luvisols
gleyic.fluvisols
regosols.humic
calcaric.fluvisols
stagnic.phaeozems
gleyic.stagnosols
haplic.chernozems
gleyic.umbrisols
brunic.arenosols
gleyic.cambisols
fluvic.cambisols
skeletic.regosols
sapric.histosols
haplic.calcisols
calcaric.regosols
cambic.phaeozems
haplic.vertisols
terric.anthrosols
solimovic.regosols
calcaric.gleysols
lithic.leptosols
histosols.dystric
lammelic.luvisols
eutric.leptosols
dystric.stagnosols
vertic.luvisols
gleyic.arenosols
nudiargic.luvisols
dystric.regosols
hortic.anthrosols
cambic.umbrisols
dystric.leptosols
haplic.fluvisols
haplic.acrisols
eutric.planosols
luvic.umbrisols
eutric.retisols
endoleptic.cambisols
leptic.regosols
calcaric.arenosols
mollic.fluvisols
abruptic.luvisols
haplic.umbrisols
calcic.luvisols
dystric.planosols
calcic.vertisols
calcic.chernozems
histic.podzols
eutric.stagnosols
endoleptic.phaeozems
rendzic.phaeozems
petric.calcisols
gleysols.humic
endogleyic.luvisols
dystric.fluvisols
epileptic.cambisols
leptosols.humic
epileptic.regosols
albic.luvisols
mollic.leptosols
vertic.cambisols
luvic.chernozems
endogleyic.planosols
hyperdystric.cambisols
retic.stagnosols
endogleyic.retisols
solimovic.cambisols
endoleptic.regosols

Entropy


Soil profile: dystric fluvic gleysols
Entropy: 3.46
Predicted soil classes:
eutric gleysols (31%)
dystric gleysols (7%)
dystric regosols (7%)
Soil profile: calcaric luvic phaeozems
Entropy value: 3.08
Predicted soil classes:
luvic.phaeozems (18%)
haplic.phaeozems (9%)
haplic.luvisols (8%)
Soil types
albic podzols
brunic arenosols
carbic podzols
dystric regosols
entic podzols
eutric gleysols
haplic luvisols
haplic phaeozems
others
Entropy
2.7 - 3
3 - 3.5
3.5 - 4
4 - 4.5
4.5 - 5.4
0 25 50 100 km
Soil profile: stagnic albic podzols
Entropy: 4.72
Predicted soil classes:
albic podzols (42%)
entic podzols (14%)
carbic podzols (13%)
Soil profile: calcaric gleysol
Entropy: 2.85
Predicted soil classes:
eutric gleysols (9%)
haplic luvisols (9%)
calcic gleysols (7%)


Extrapolation probability (%)
86
13
0 500 1.000 km