Benchmarking Earth Observation Embeddings for large scale aboveground biomass mapping

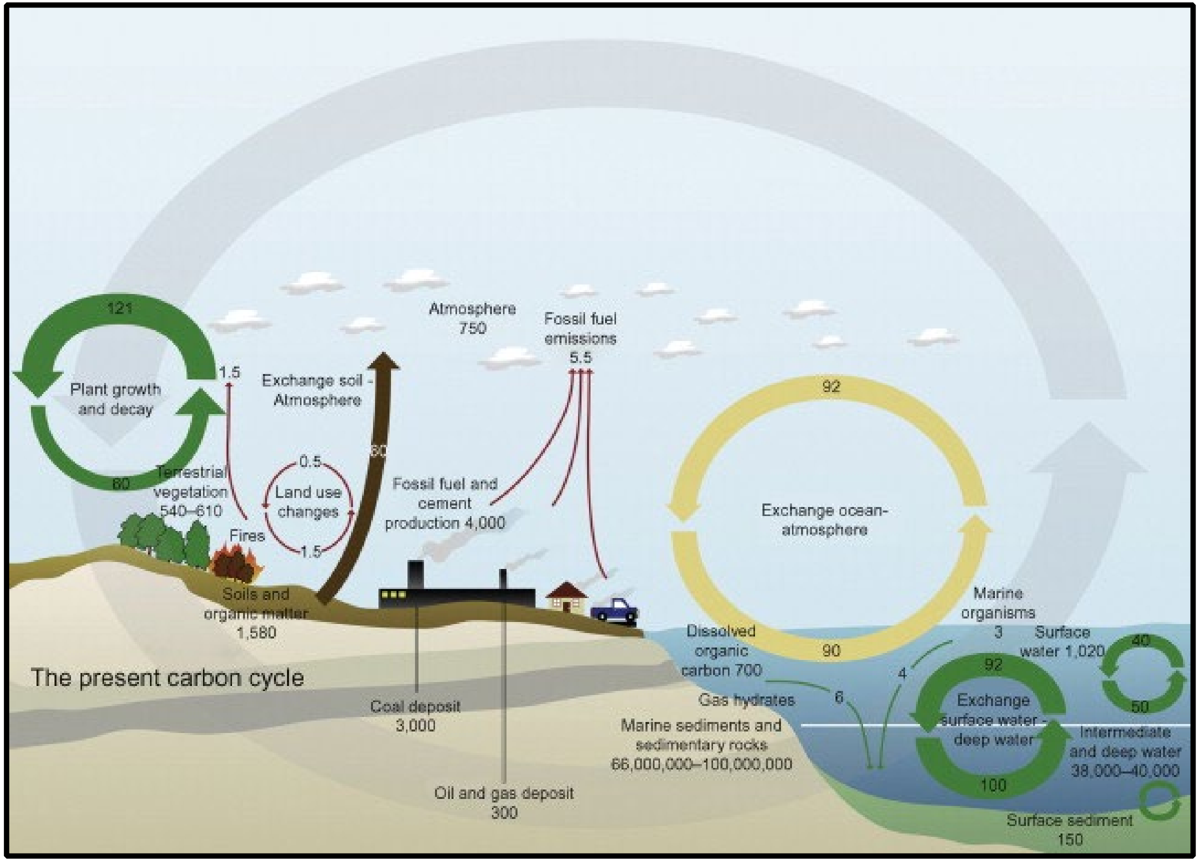
Yu-Feng Ho, Derek Karssenberg, Leandro Parente, Madlene Nussbaum

5th ECMWF-ESA Machine Learning Workshop
Bologna 15 april, 2026

This research project is sponsored by: Open Geospatial Carbon Registry
https://cordis.europa.eu/project/id/101218854

Aboveground Biomass in Carbon Cycle

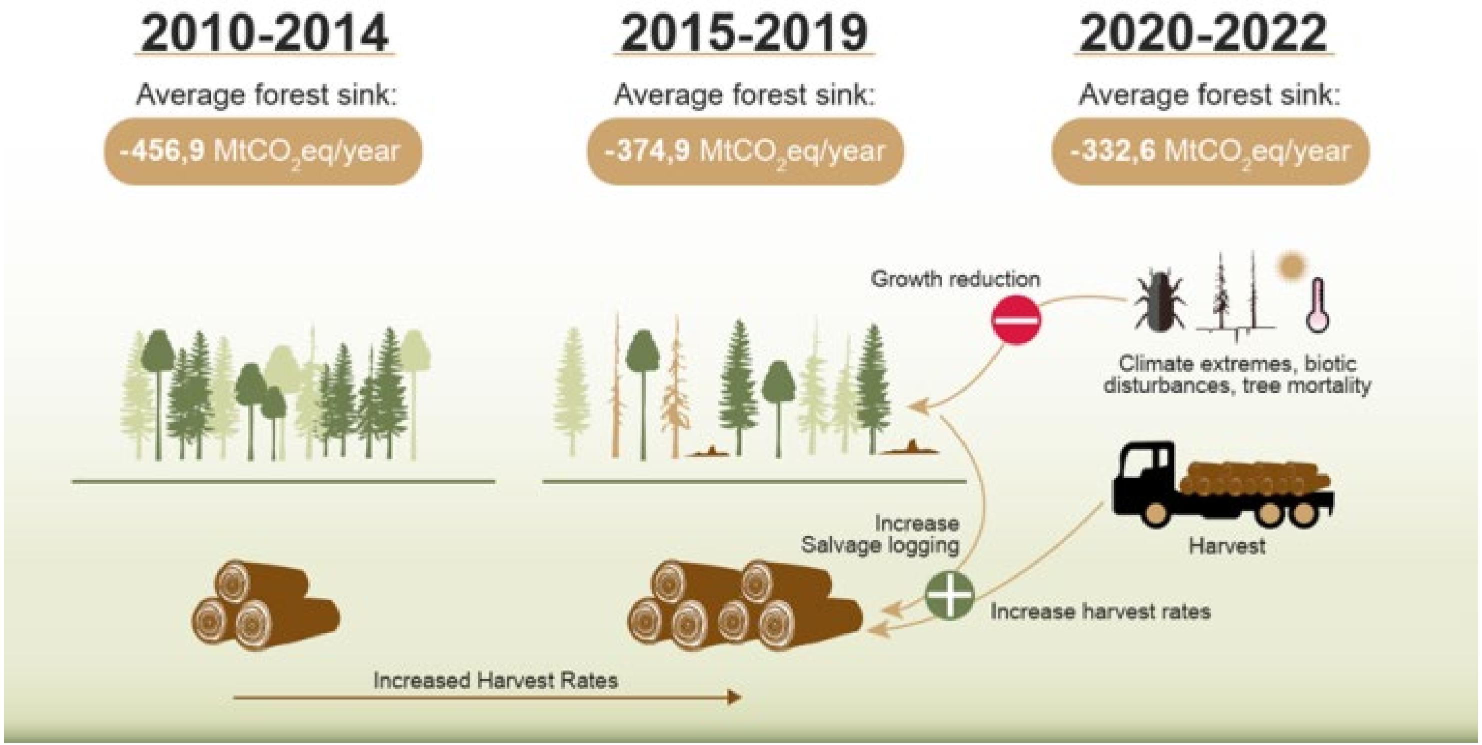
Although the major carbon stock is located in sediments under the sea, plant growth and **decay of terrestrial vegetation contribute a relatively huge portion of the carbon flux.**



(Courtesy of UNEP/GRID-Arendal)

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Aboveground Biomass in Carbon Cycle

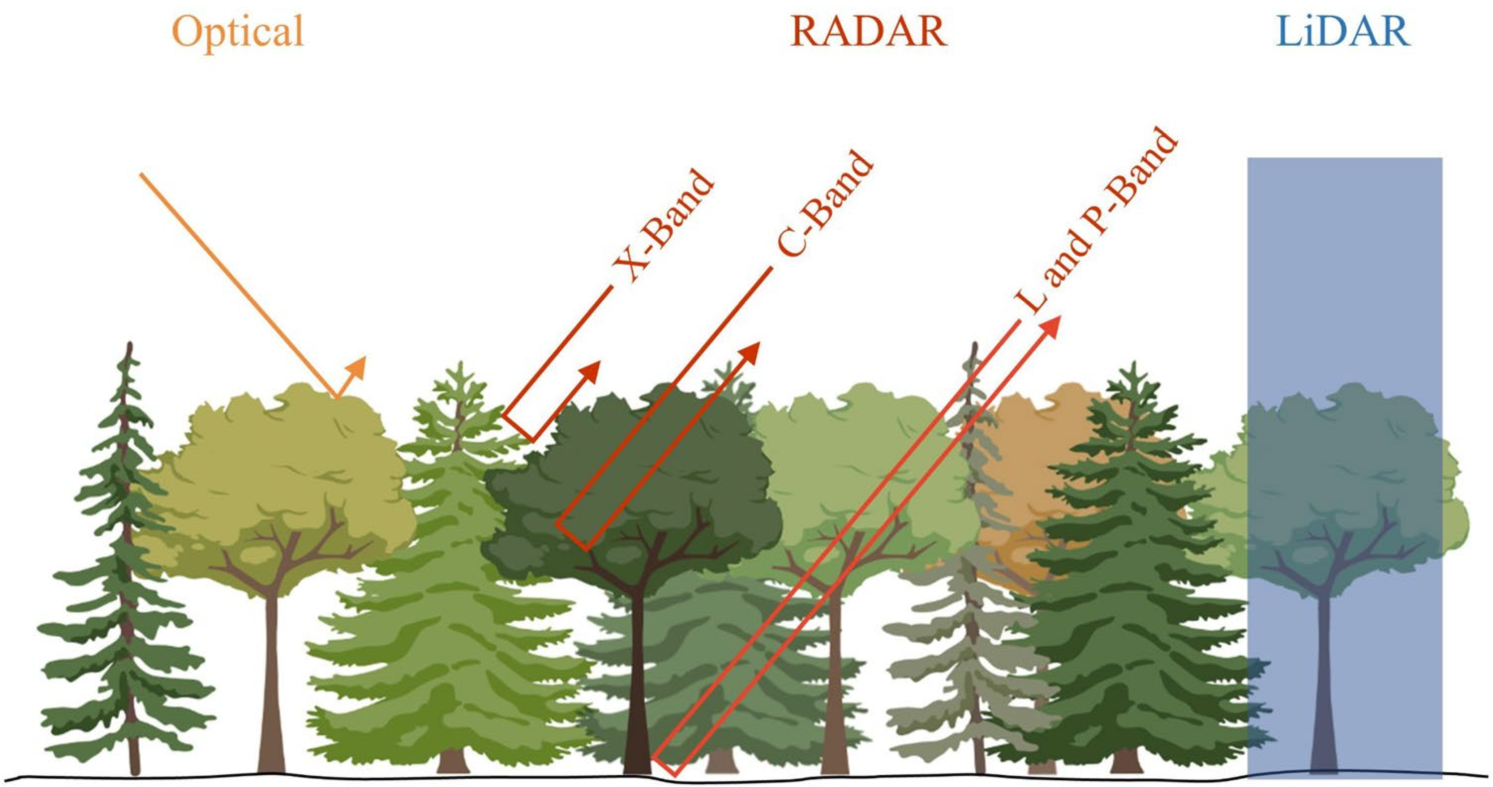


EU Forest works as **carbon sink**, but is **decreasing** and lack of **monitoring**

https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/european-forest-carbon-sink-declining-can-we-reverse-trend-2025-07-30_en

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Earth Observation for Aboveground Biomass

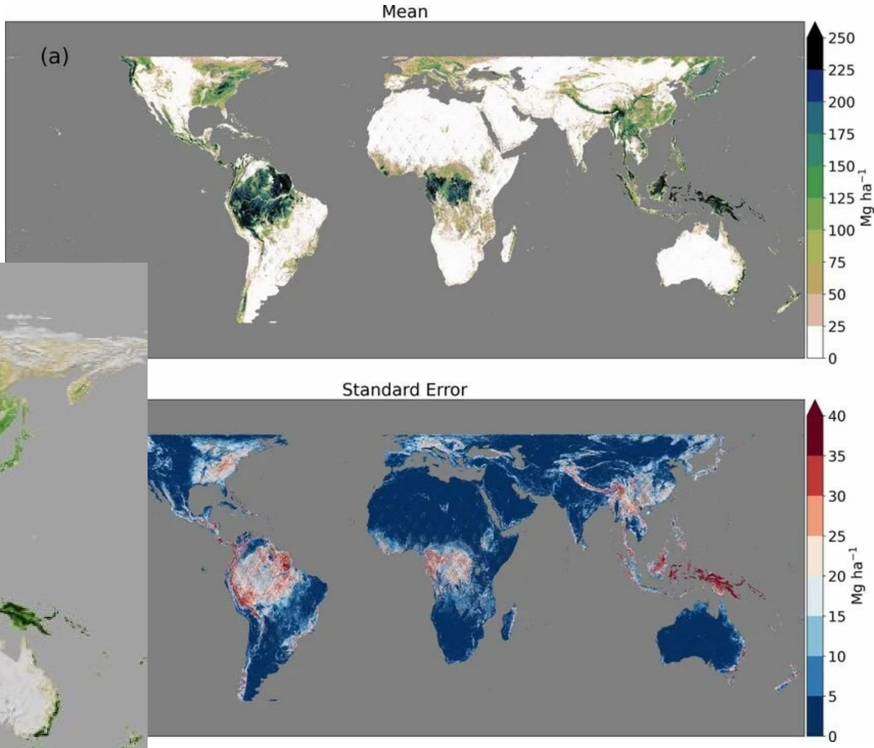


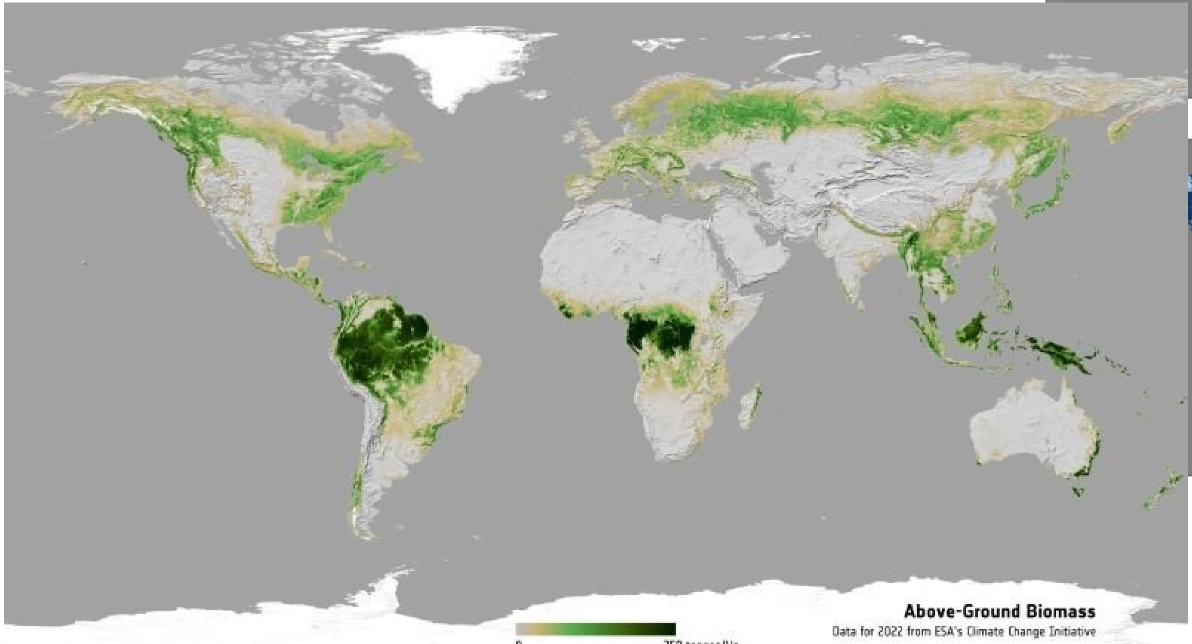
Tian, L., Wu, X., Tao, Y., Li, M., Qian, C., Liao, L., & Fu, W. (2023). Review of remote sensing-based methods for forest aboveground biomass estimation: Progress, challenges, and prospects. *Forests*, *14*(6), 1086.

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Aboveground Biomass Mapping

Santoro, M., & Cartus, O. (2023). *ESA Biomass Climate Change Initiative (Biomass_cci): Global datasets of forest above-ground biomass for the years 2010, 2017, 2018, 2019 and 2020, v4.* Centre for Environmental Data Analysis.

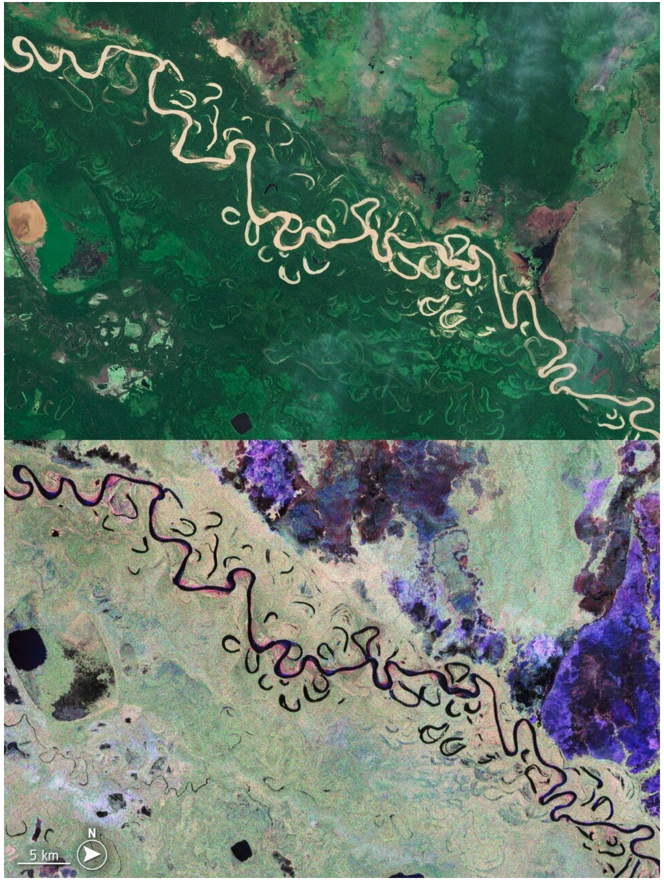




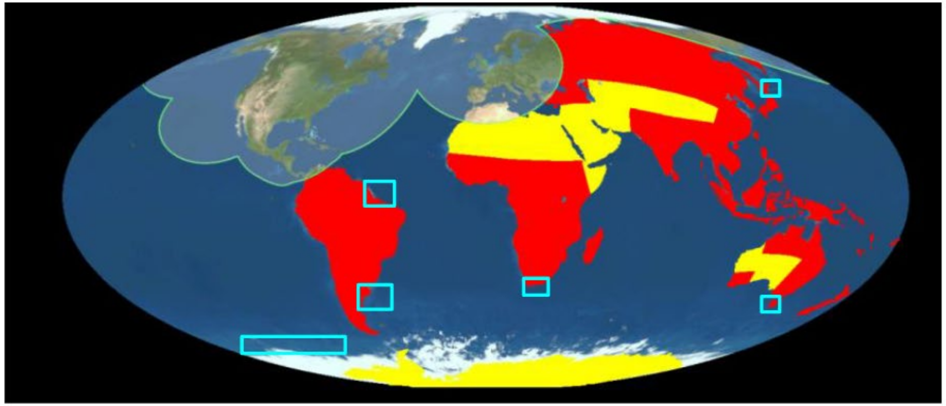
The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Dubayah, R.O., J. Armston, S.P. Healey, Z. Yang, P.L. Patterson, S. Saarela, G. Stahl, L. Duncanson, and J.R. Kellner. 2022. GEDI L4B Gridded Aboveground Biomass Density, Version 2. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/2017

Aboveground Biomass Mapping



ESA's new mission (P-band radar) to monitor forest biomass of 200m of the world, **but exclude US and EU**

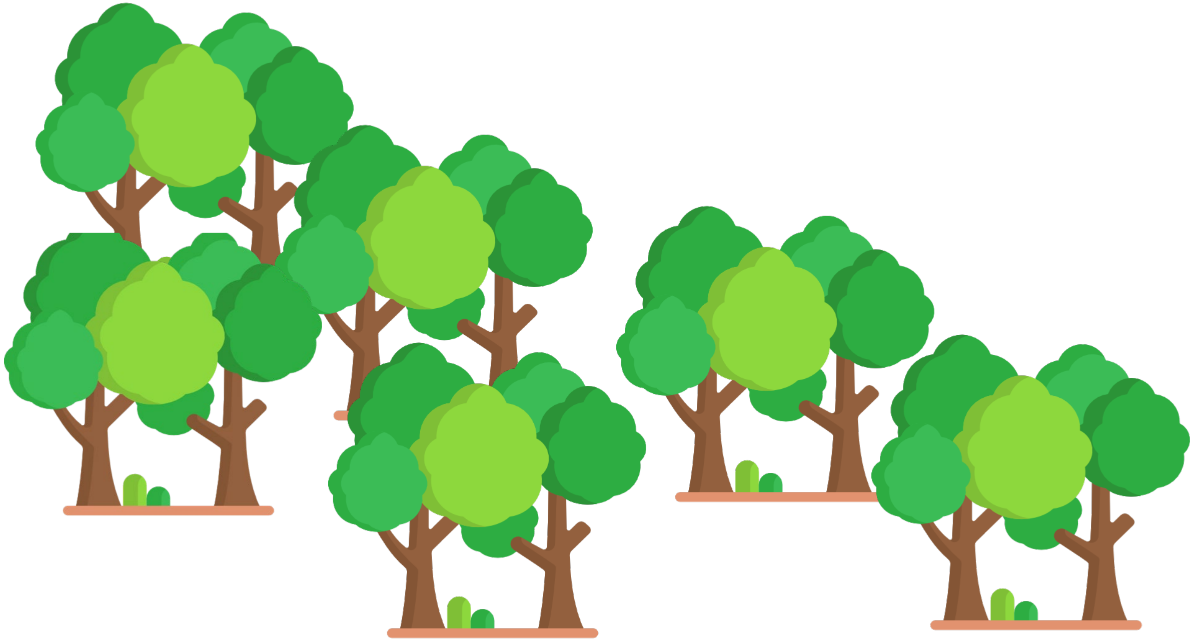


(Red = Primary objective coverage mask, Yellow = Secondary objective coverage mask)

https://earth.esa.int/eogateway/missions/biomass/description

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

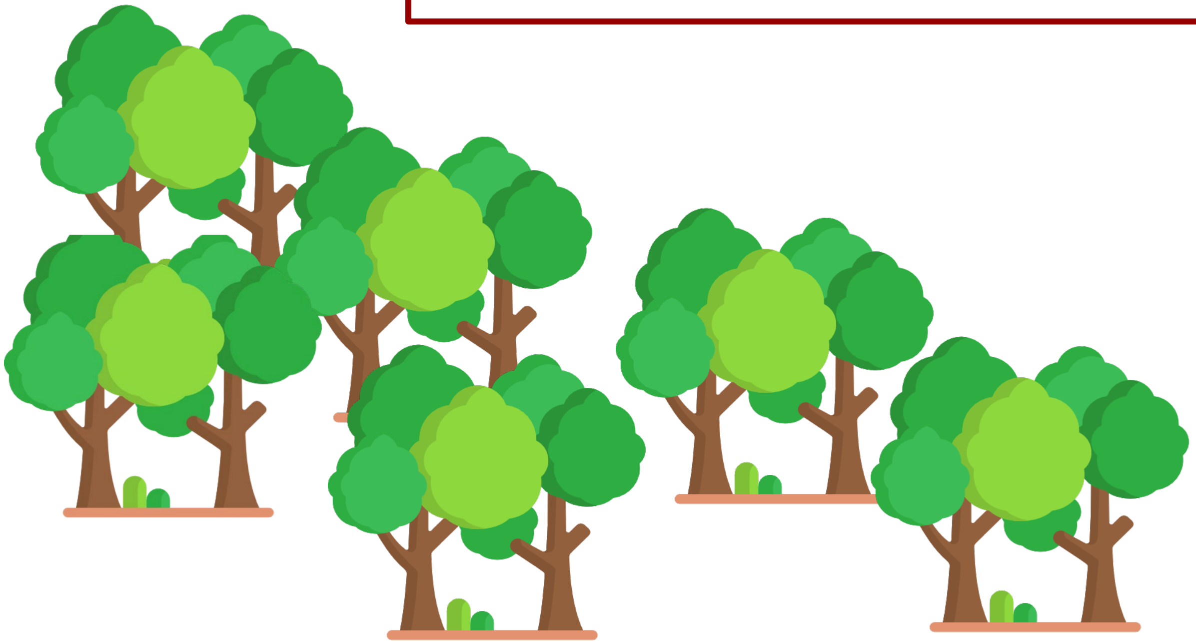
Forest monitor / AGB estimate



The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

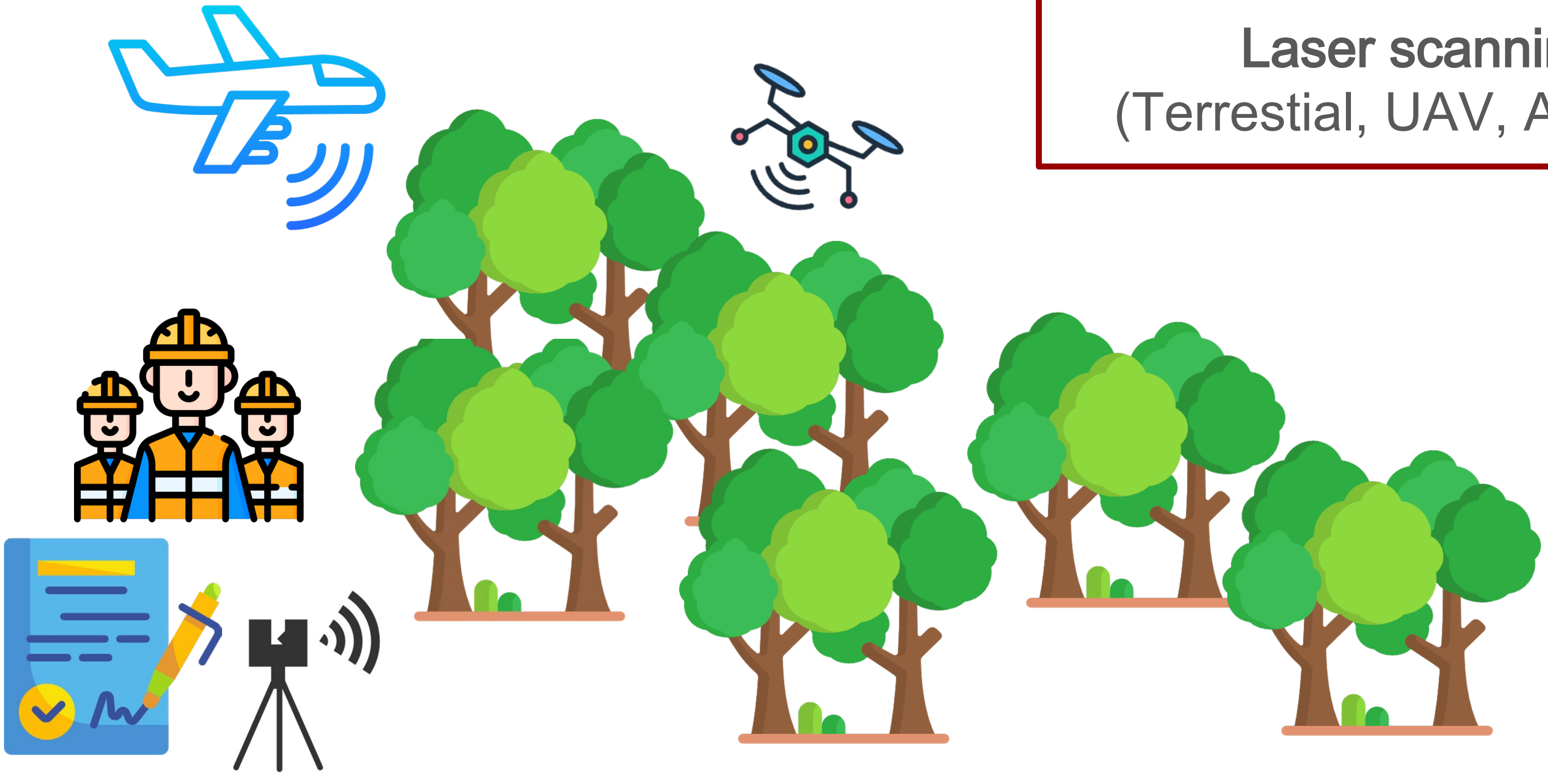
Forest monitor / AGB estimate

Forest inventory
(Field data: tree height, diameter, species)



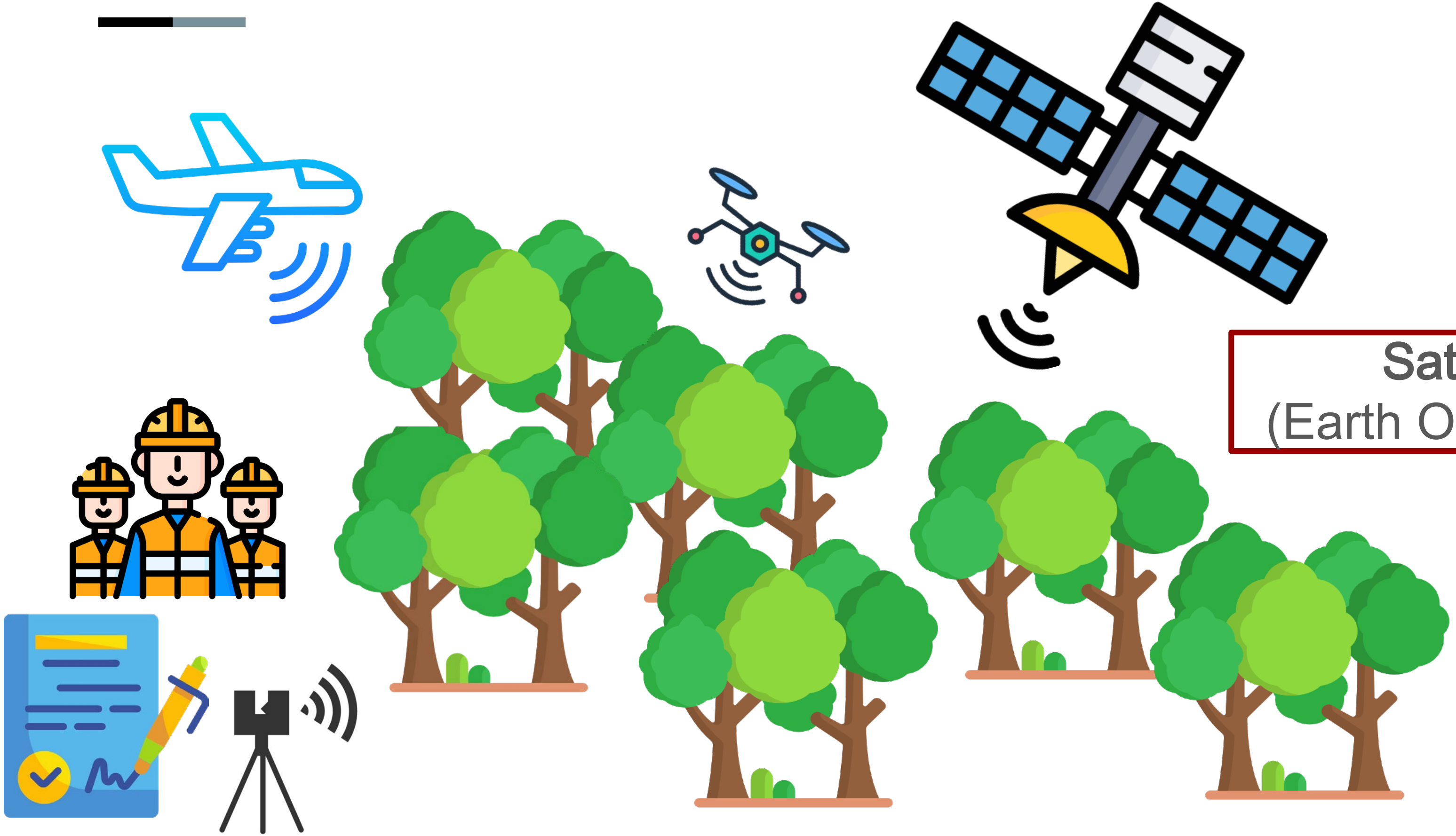
The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Forest monitor / AGB estimate



Laser scanning
(Terrestial, UAV, Airborne)

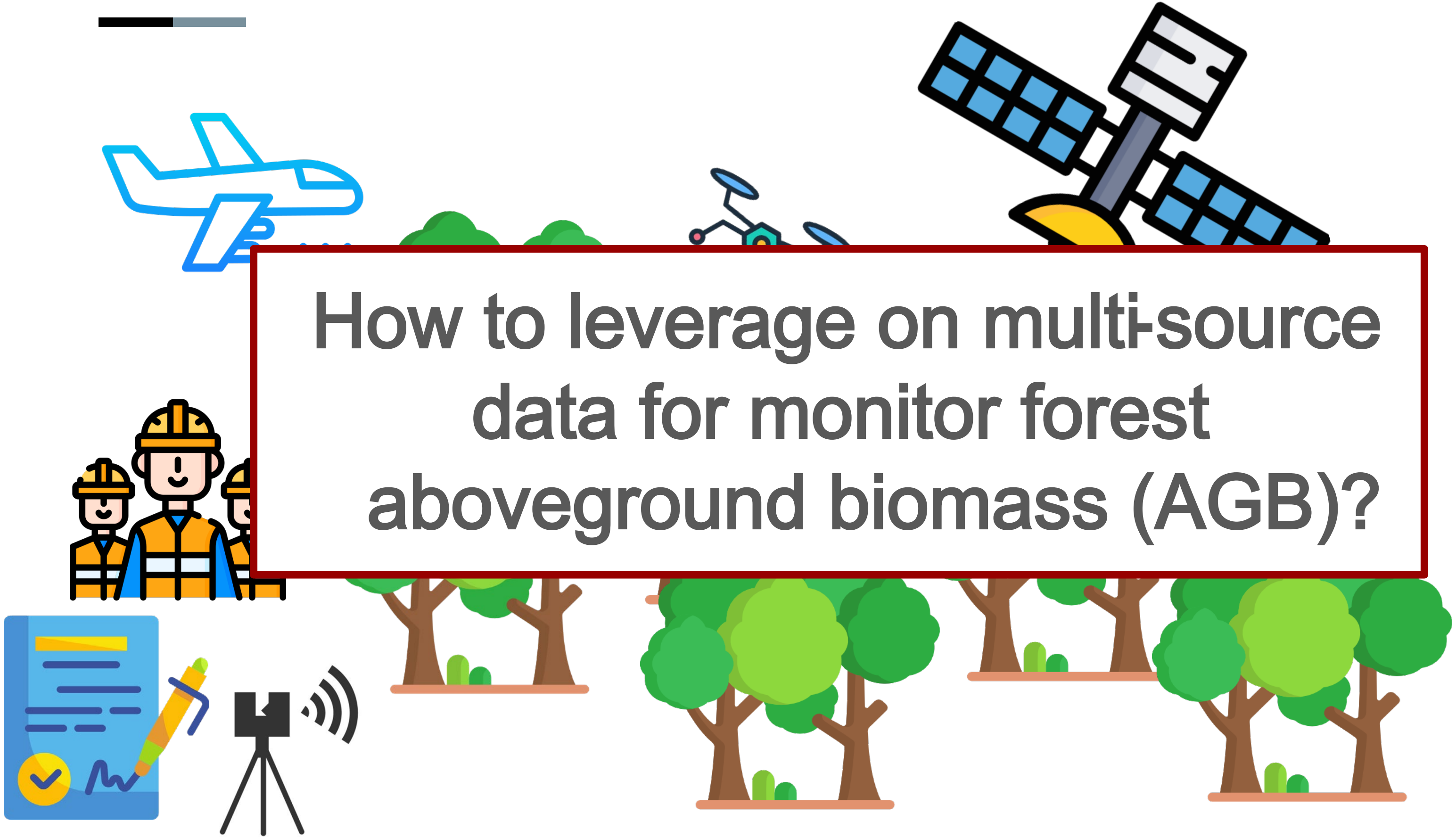
Forest monitor / AGB estimate



Satellite
(Earth Observation)

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

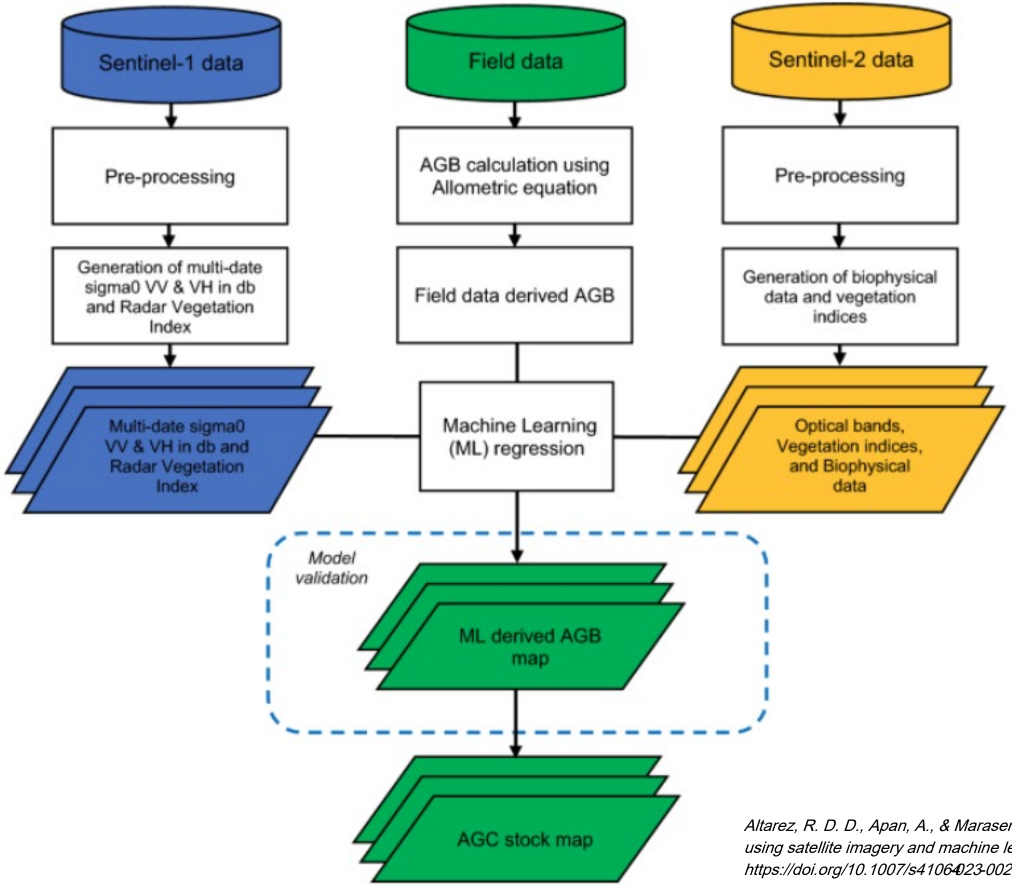
Forest monitor / AGB estimate



How to leverage on multi-source data for monitor forest aboveground biomass (AGB)?

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Machine Learning Framework for AGB



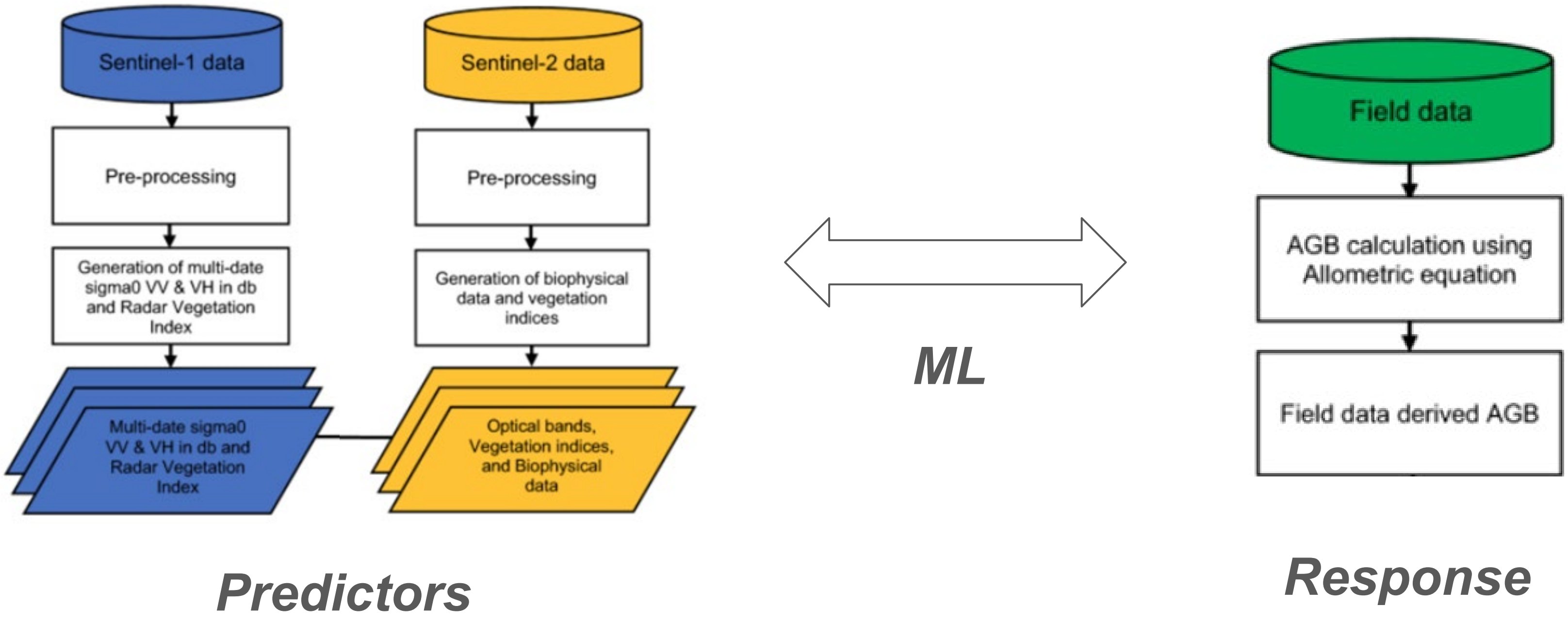
Variables

- Field data: Reference AGB
- Sentinel-1: Forest structure
- Sentinel-2: Vegetation abundance

More ref. Nuthammachot et al. (2020), Ghosh and Behera (2018)

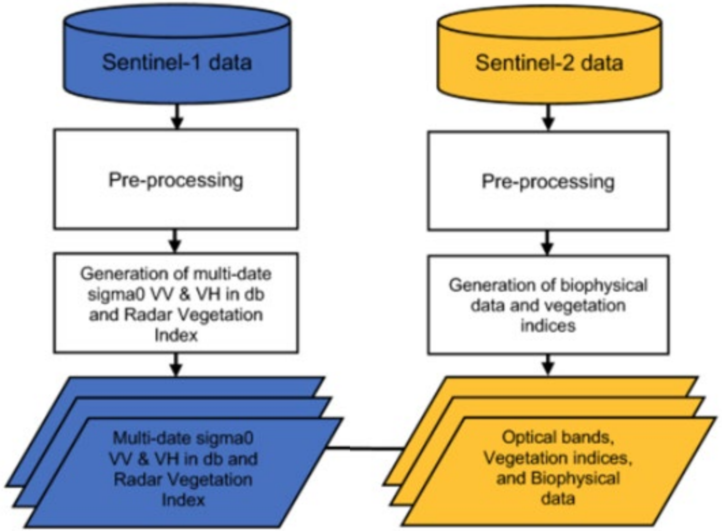
Altarez, R. D. D., Apan, A., & Maraseni, T. (2024). Uncovering the hidden carbon treasures of the Philippines' towering mountains: A synergistic exploration using satellite imagery and machine learning. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 92(1), 55–73. https://doi.org/10.1007/s41064-023-00264-w

Machine Learning Framework for AGB



The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Machine Learning Framework for AGB



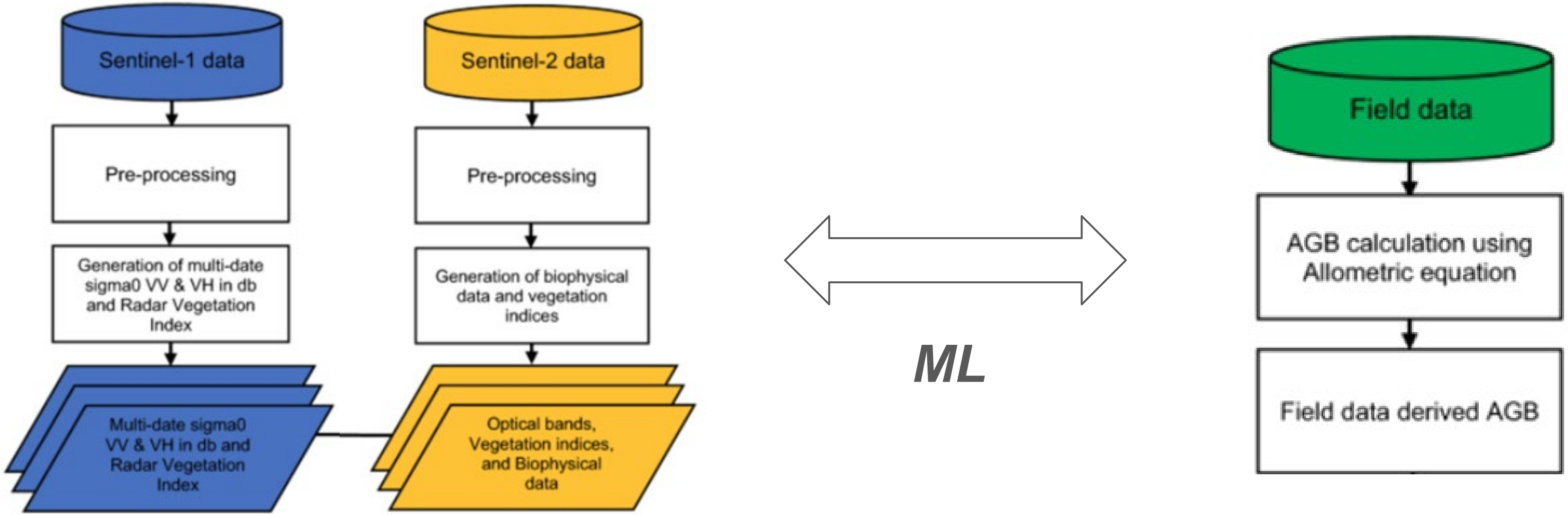
Predictors

Not an easy task for large scale mapping…

- High resolution,
- Multiple bands,
- Challenging pre-processing,
- **High frequent revisit**

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Machine Learning Framework for AGB

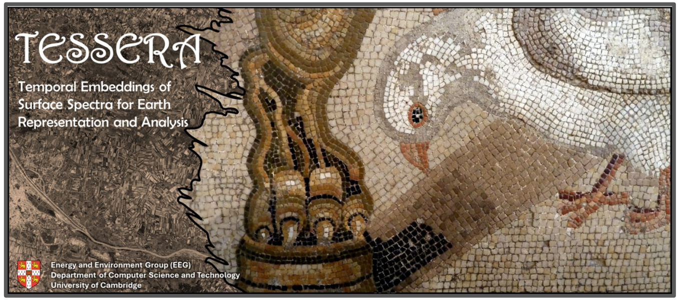


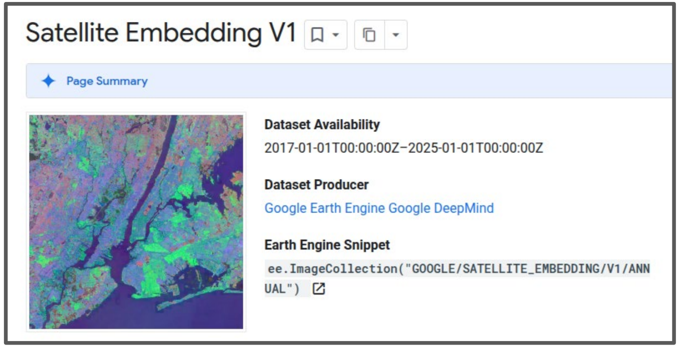
Predictors

Response

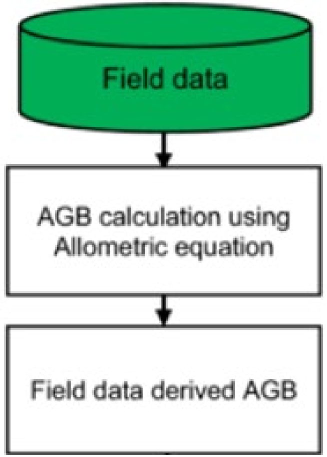
The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Machine Learning Framework for AGB





ML



Predictors

Response

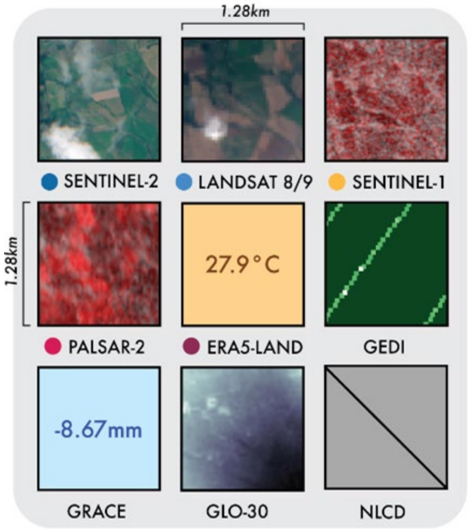
https://github.com/ucam-eo/tessera

https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL

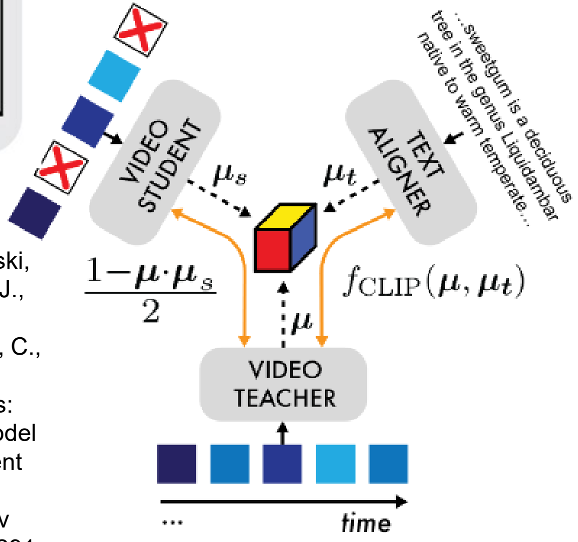
The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

AlphaEarth Foundations

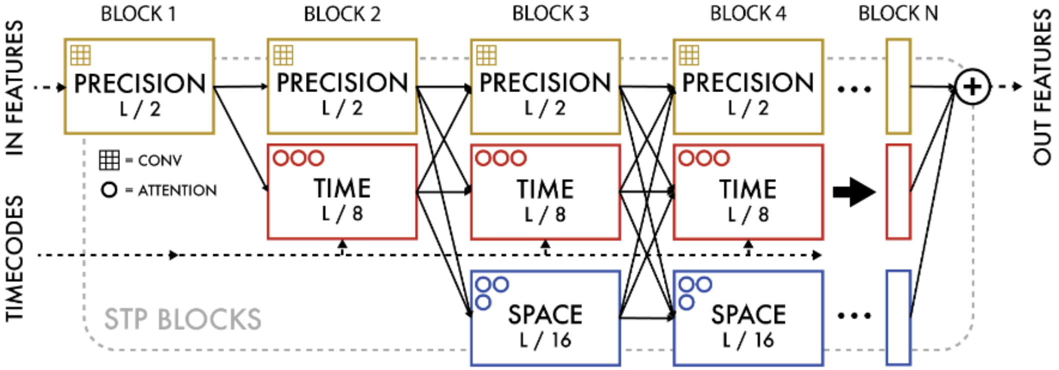
Data sources



Model concept

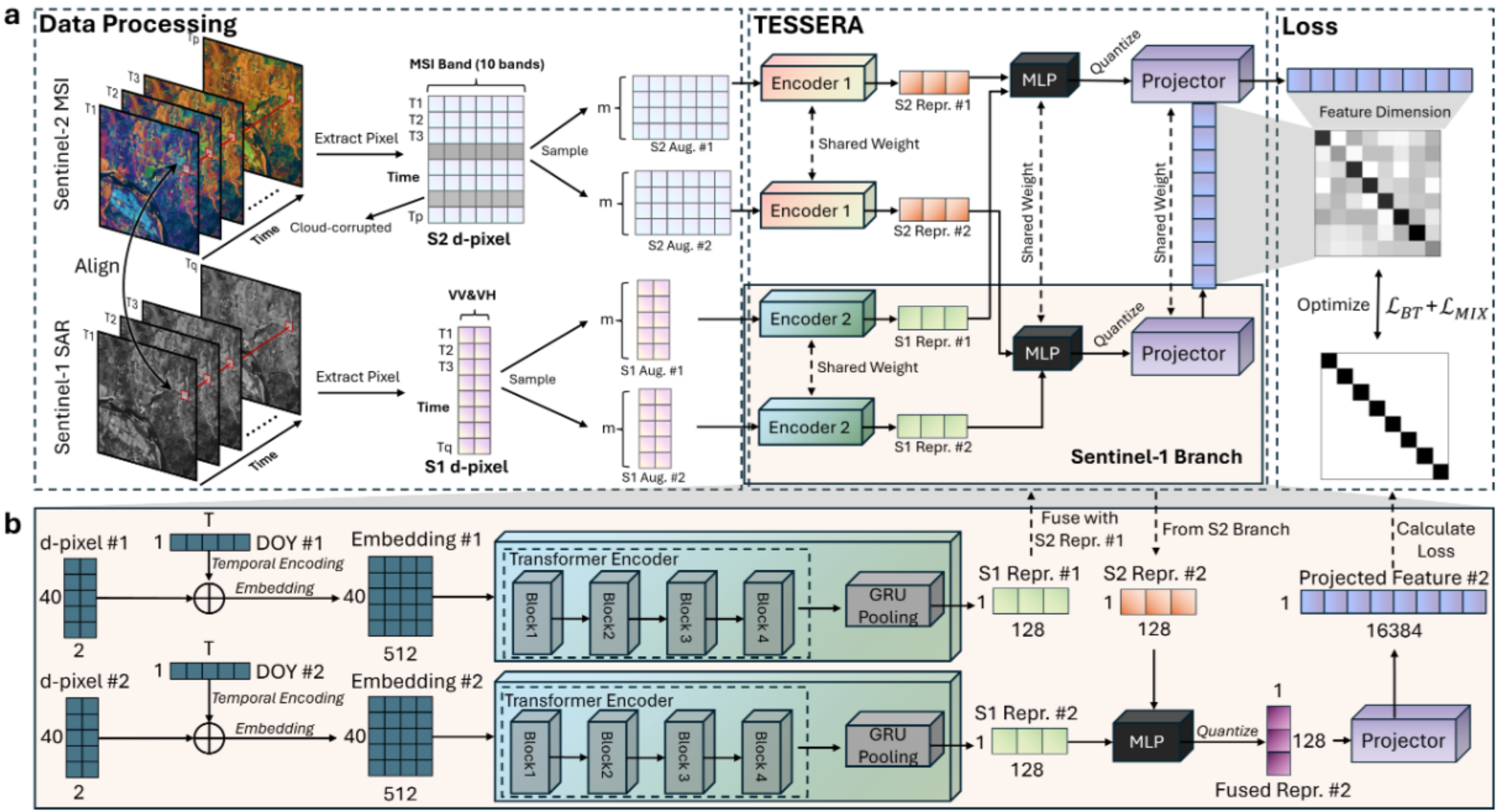


Modeling framework



Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., ... & Kohli, P. (2025). Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv preprint arXiv:2507.22291.

TESSERA

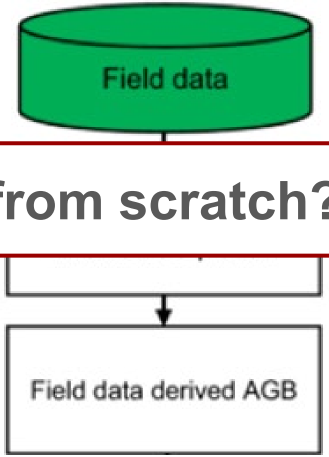


Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., ... & Keshav, S. (2025). Tessera: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*.

Machine Learning Framework for AGB

Question 1: Can EO embedding replace training from scratch?

ML



Predictors

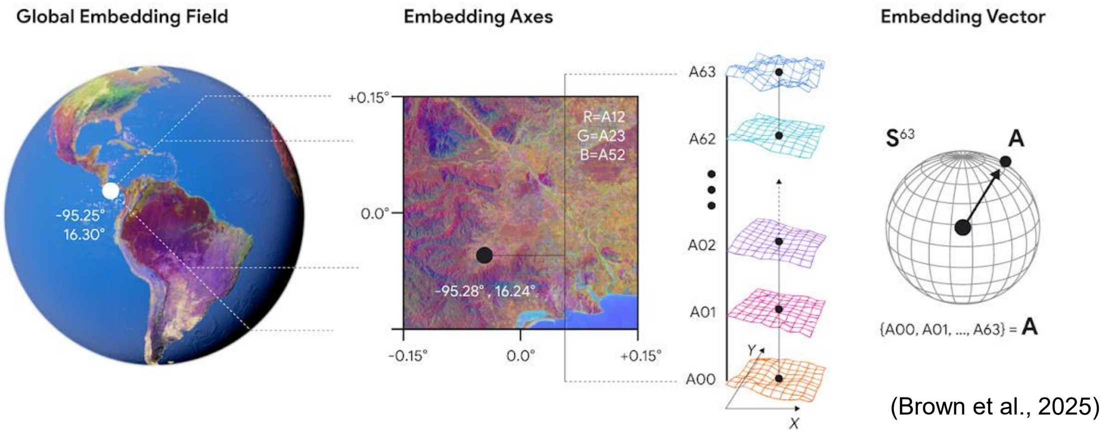
Response

https://github.com/ucam-eo/tessera

https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

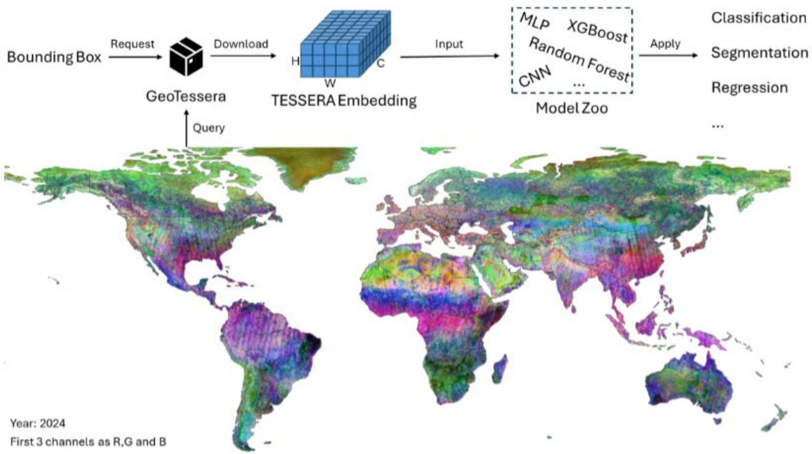
EO Embedding dataset

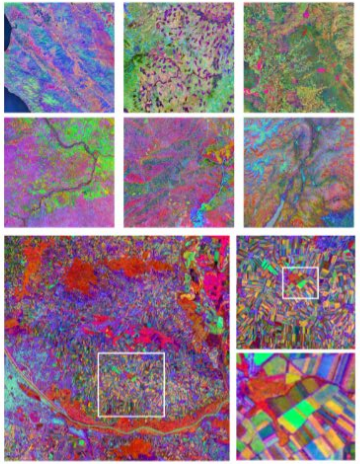


1. Numerical vectors stored as any other **raster layers**
2. Easy to build **downstream** models with **non-limit choices.**

(Brown et al., 2025)

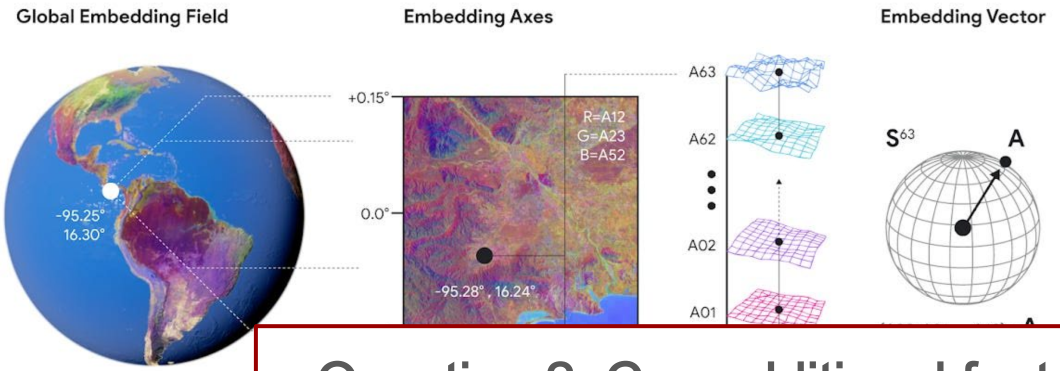
3. Flexibility to add **additional features** or do **feature engineering**





(Feng et al., 2025)

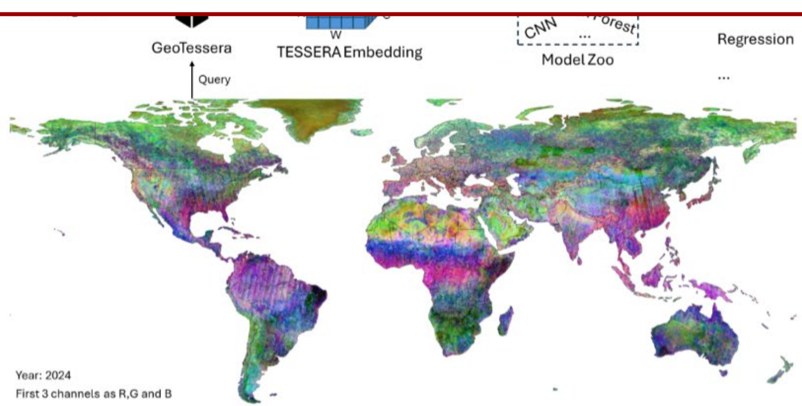
EO Embedding dataset

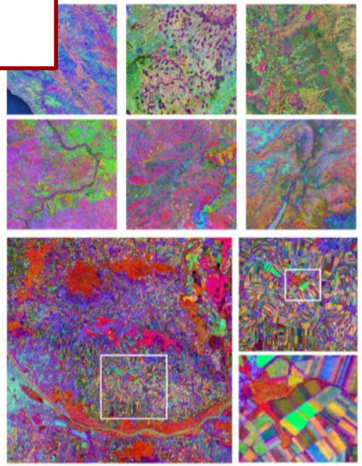


1. Numerical vectors stored as any other **raster layers**
2. Easy to build **downstream** models with **non-limit choices**

Question 2: Can additional features further increase model performance?

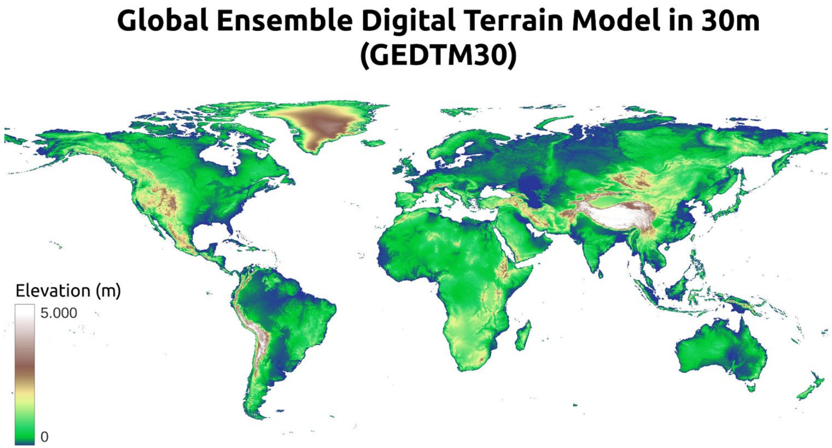
3. Flexibility to add **additional features** or do **feature engineering**





(Feng et al., 2025)

Topographic attributes as additional features



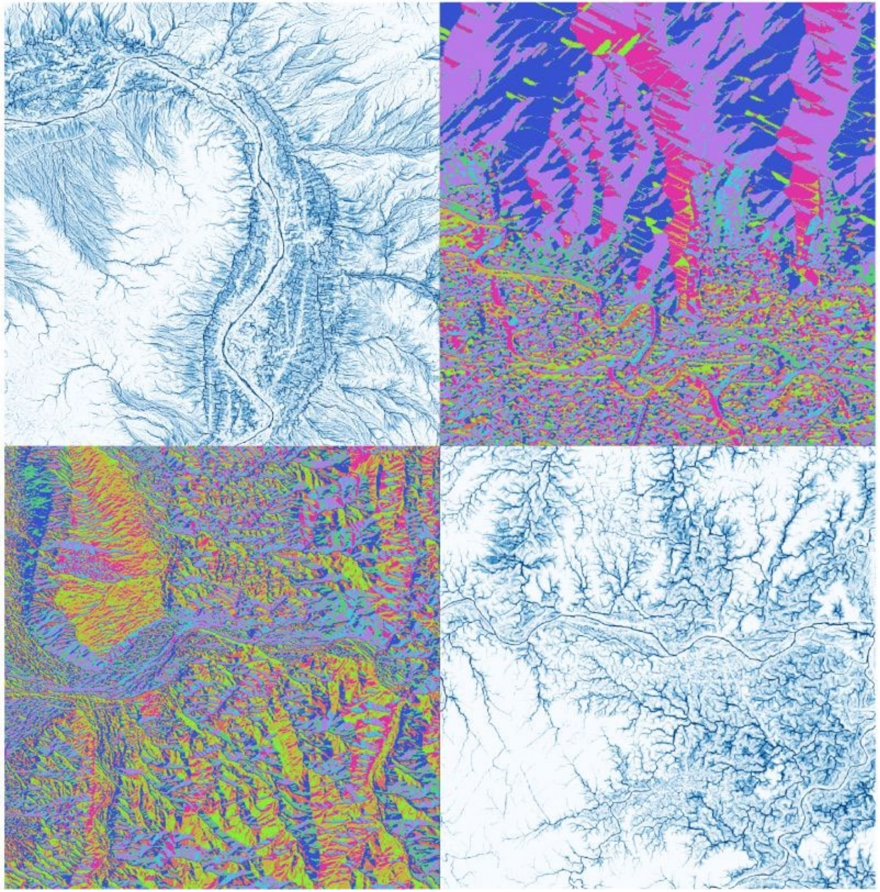
‹ ARTIFICIAL INTELLIGENCE, COMPUTER VISION AND NATURAL LANGUAGE PROCESSING

GEDTM30: global ensemble digital terrain model at 30 m and derived multiscale terrain variables

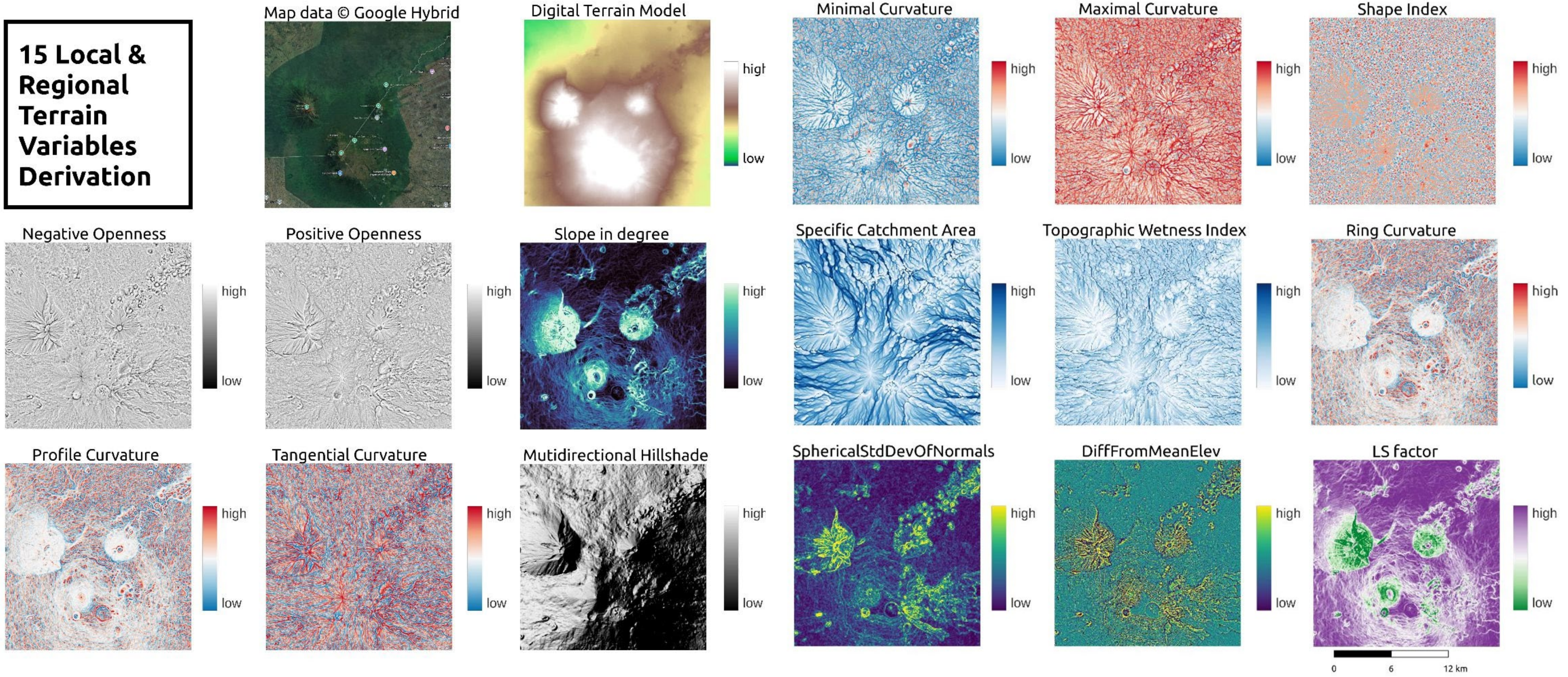
Research Article | Spatial and Geographic Information Science

Yu-Feng Ho [1], Carlos H. Grohmann [2], John Lindsay [3], Hannes I. Reuter [4,5], Leandro Parente [1], Martijn Witjes [1], Tomislav Hengl [1] Post to Authors on X

Published July 24, 2025



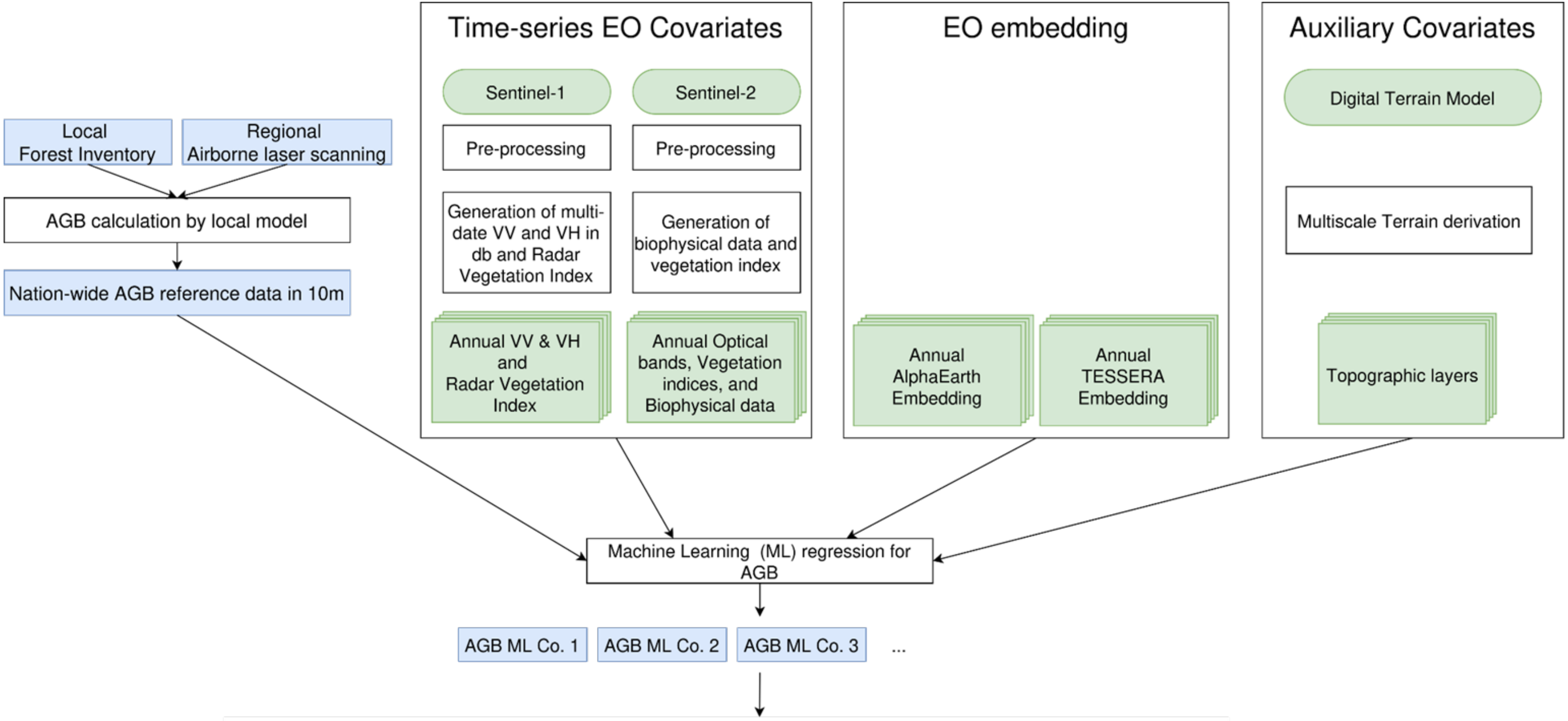
Topographic attributes as additional features



Topographic features underlie the natural process...

Ho, Y. F., Grohmann, C. H., Lindsay, J., Reuter, H. I., Parente, L., Witjes, M., & Hengl, T. (2025). GEDTM30: Global ensemble digital terrain model at 30 m and derived multiscale terrain variables. PeerJ, 13, e19673.

General Framework



What is the optimal setup for large scale aboveground biomass mapping?

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Evaluation by Machine Learning Method

Q: which EO embedding setup has better performance?
- High Accuracy
- Lower Uncertainty

Quantile Random Forest Regressor
- Scalable
- Per-pixel uncertainty
- Easy to implement (Scikit-learn)

To produce per-pixel uncertainty, we use the TB-QRF (Meinshausen, 2006), where the output of each single tree in the RF has been used to derive the prediction intervals:

$$
\begin{aligned}
\widehat{y1}_{\text{stb}_{1,1}} &= [P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2}] \\
\widehat{y2}_{\text{stb}_{1,1}} &= [P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2}] \\
\widehat{y3}_{\text{stb}_{1,1}} &= [P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2}] \\
&\cdots \\
\widehat{yN}_{\text{stb}_{1,1}} &= [P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2}]
\end{aligned}
\tag{13}
$$

where $\widehat{y1}_{\text{stb}_{1,1}}$ is the mean prediction from the individual tree (out of a total of N trees). These are then used to derive the lower and upper 68 % probability quantiles:

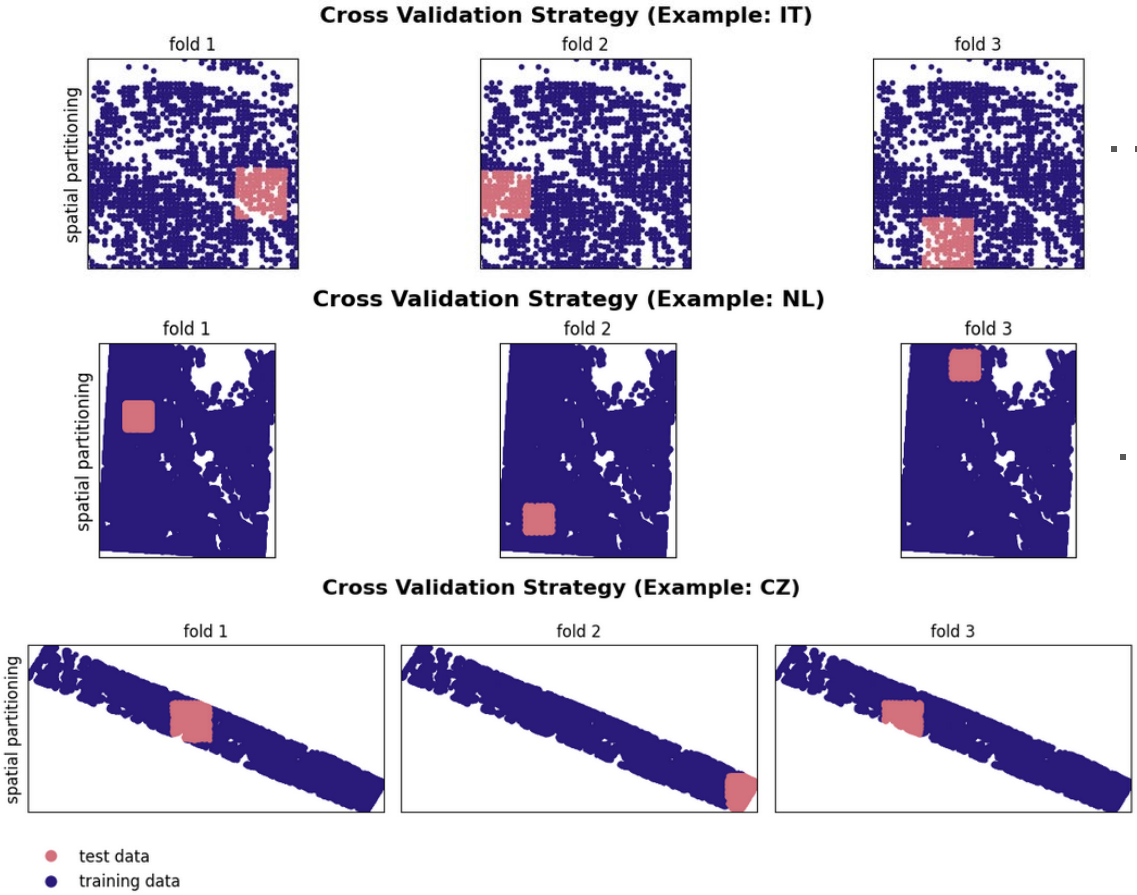
$$\widehat{y}_{\text{lower}} = Q_{0.16}\left(\left\{\widehat{y1}_{\text{stb}_{1,1}}, \widehat{y2}_{\text{stb}_{1,1}}, \cdots, \widehat{yN}_{\text{stb}_{1,1}}\right\}\right), \tag{14}$$

$$\widehat{y}_{\text{upper}} = Q_{0.84}\left(\left\{\widehat{y1}_{\text{stb}_{1,1}}, \widehat{y2}_{\text{stb}_{1,1}}, \cdots, \widehat{yN}_{\text{stb}_{1,1}}\right\}\right). \tag{15}$$

Hengl, T., Consoli, D., Tian, X., Nauman, T. W., Nussbaum, M., Isik, M. S., ... & Harris, N. (2026). OpenLandMap-soildb: global soil information at 30 m Spatial resolution for 2000–2022+ based on Spatiotemporal machine learning and harmonized legacy soil samples and observations. Earth System Science Data, 18(2), 999-1036.

Cross validation- Spatial Blocking

```
Automatic Spatial Range Detected: 115.95 meters
Remaining samples after filtering zeros: 1359
```



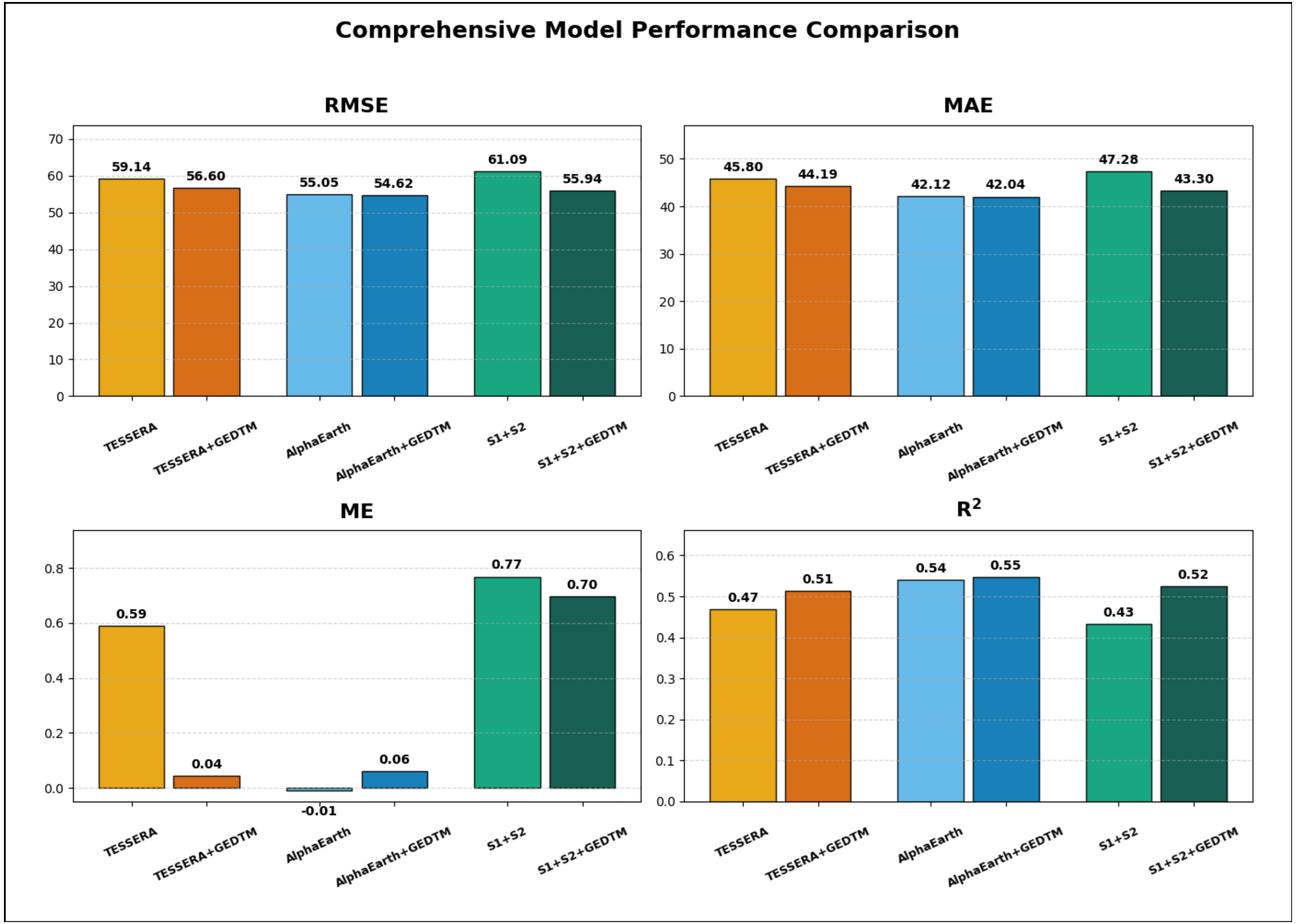
. . .

. . .

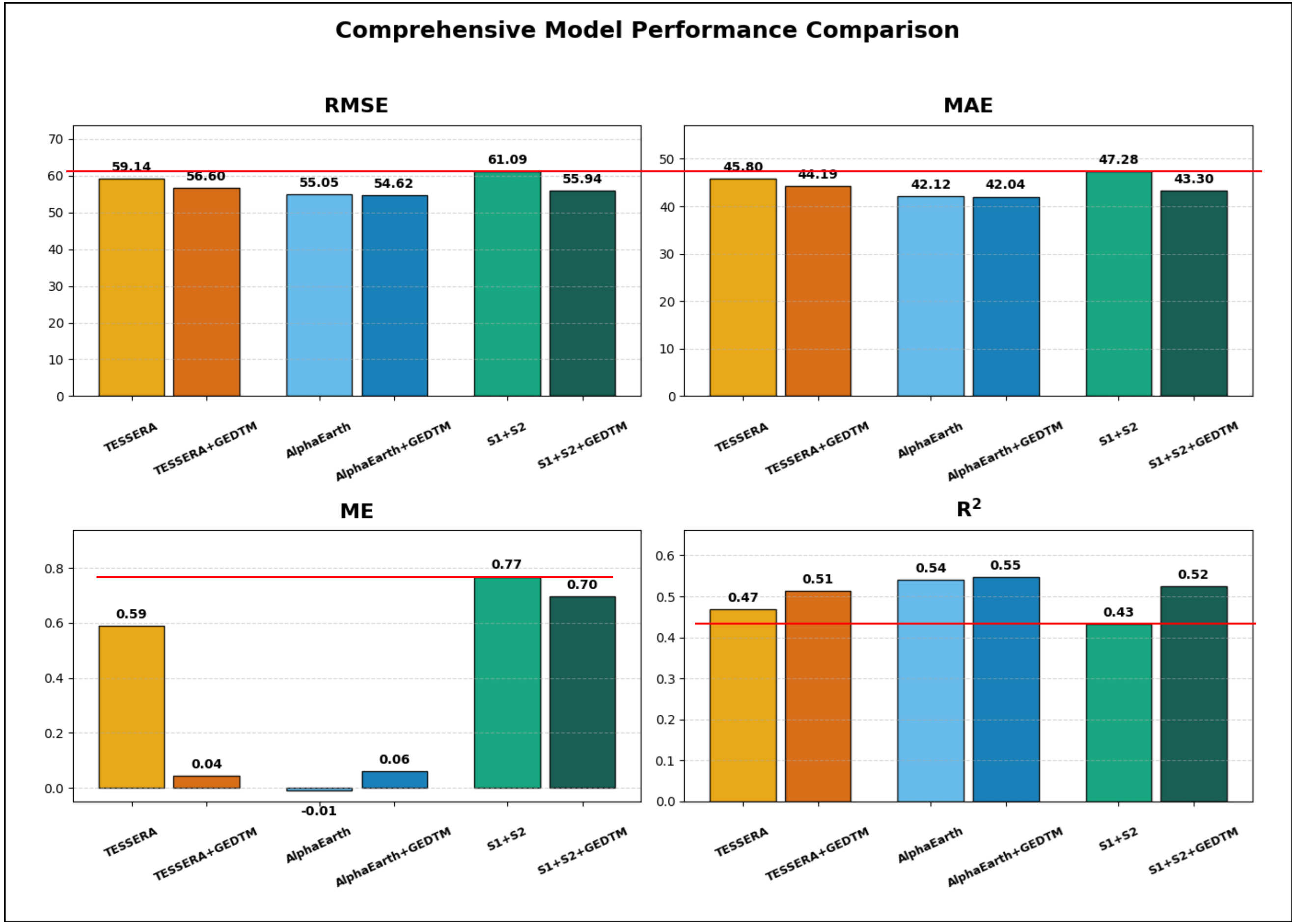
. . .

- Minimal blocksize: semivariogram
- Crosscountry modeling:
  1. block aggregation
  2. 5-fold cross validation through radom blocks

Result - accuracy assement

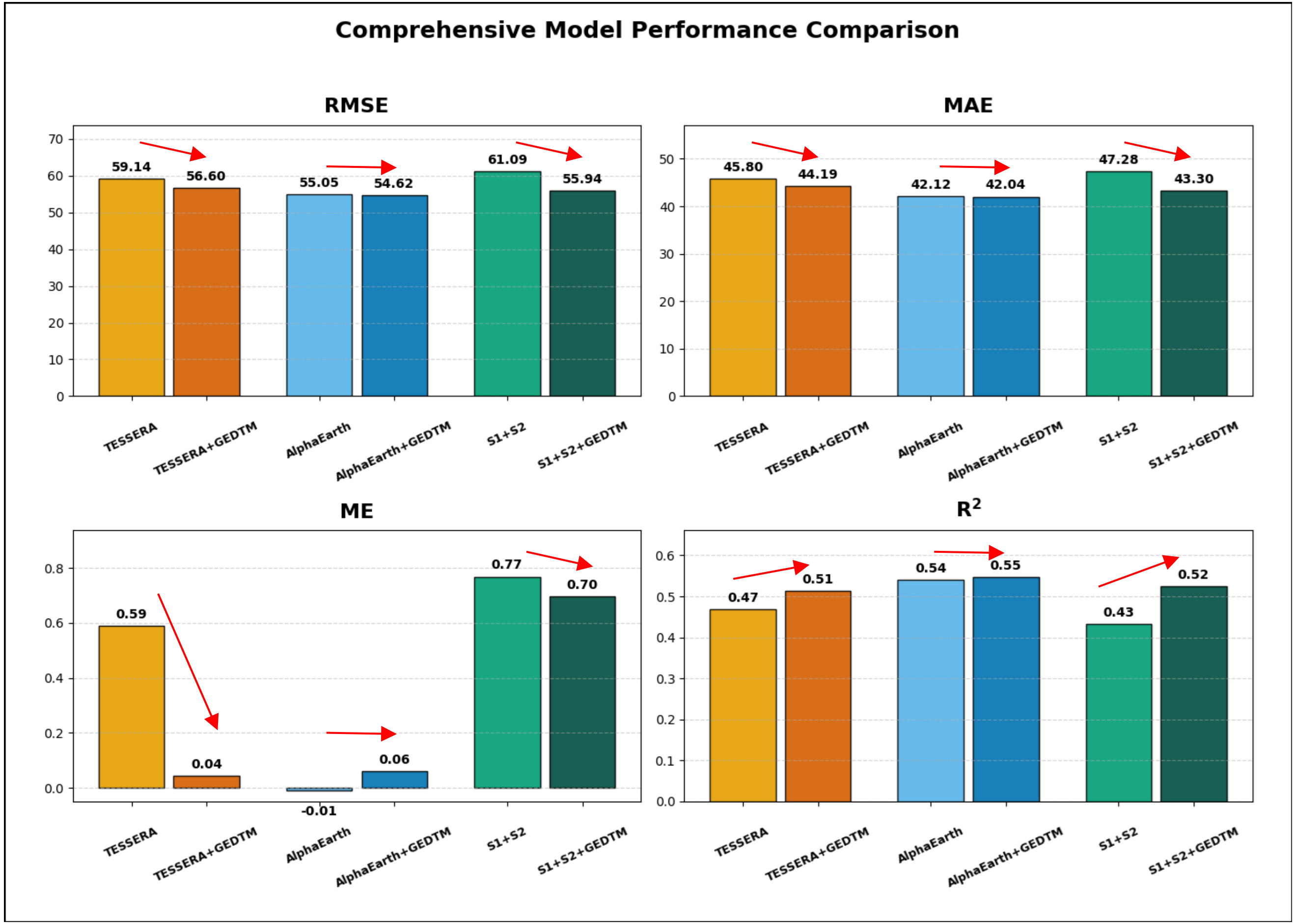


Result - accuracy assement



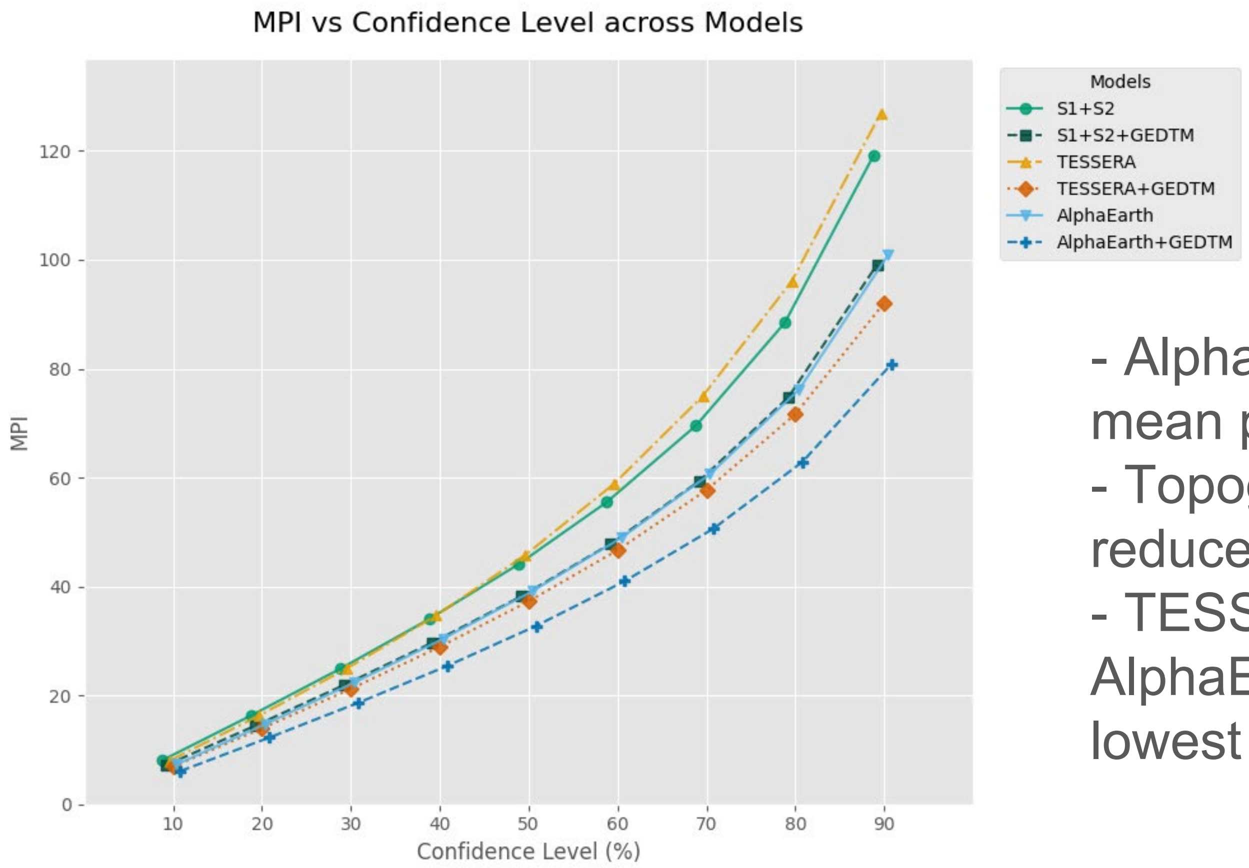
- TESSERA & AlphaEarth performs better than S1+S2
- AlphaEarth performs slightly better than TESSERA

Result - accuracy assement



- Topographic attributes contribute to prediction accuracy
- Except for AlphaEarth

Result - uncertainty assessment



- AlphaEarth has the lower mean prediction interval (MPI)
- Topographic attributes reduces MPI
- TESSERA+GEDTM & AlphaEarth+GEDTM has the lowest MPI

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Discussion and Conclusion

- EO embedding has **comparable** model performance to baseline Sentinel-1/2 model.
- Topographic attributes improve both **prediction accuracy** and **uncertainty**, except for AlphaEarth's point estimate accuracy.
- AlphaEarth integrates GEDI and COPDEM in training modeling. The self-supervised learning model has learnt the topographic attributes.
- More probablisitc deep learning model should be included in this benchmark (e.g. Bayesian Neural Network).
- Expansion to **climatic datasets** could reinforce the benefit of additional features.

The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Thank you for your attention!

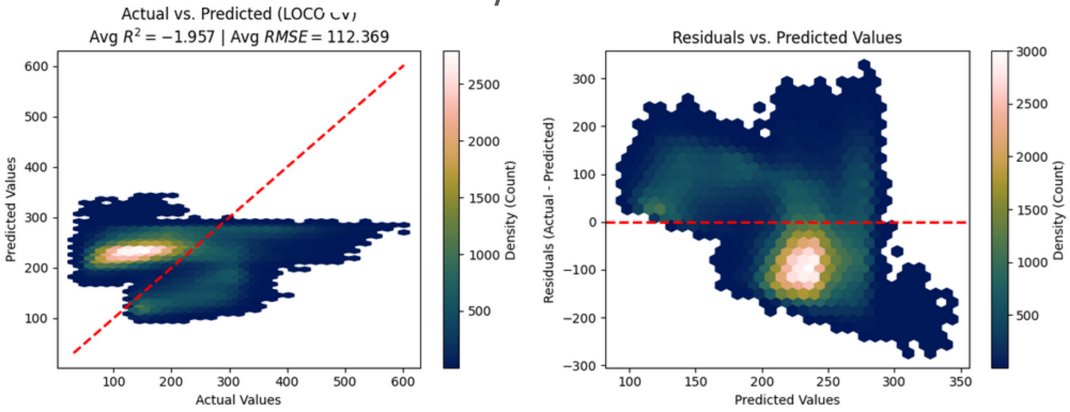
Open Geospatial Carbon Registry

Partnership with...

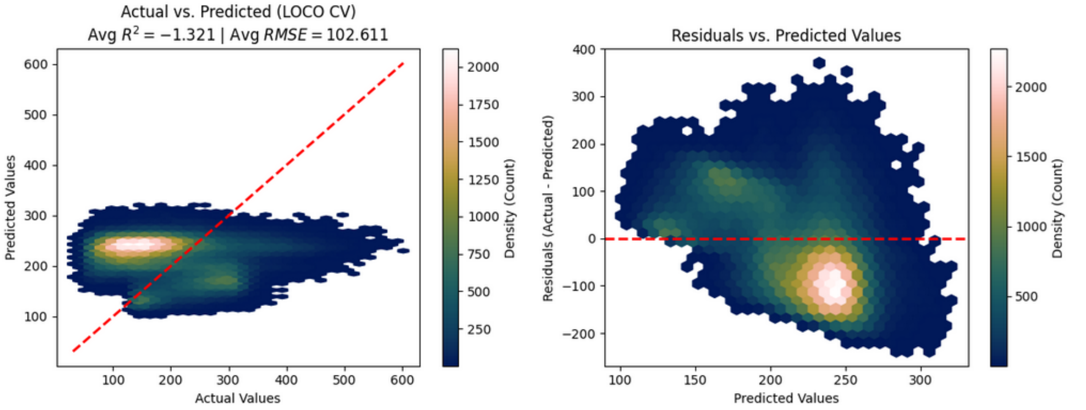
The Intergenerational Open Geospatial Carbon Registry (OGCR). Grant agreement ID: 101218854.

Can we do pan-EU mapping now?

AlphaEarth



TESSERA



Leave-one-country-out (LOCO) was performed. Models were trained on 2 countries and were validated by the third one.

It shows a poor performance, indicates the extrapolation to other countries is currently impossible