FAST-EO: Transforming Earth Observation Through Multi-Modal Foundation Models

Image Credit: J. Jakubik, B. Blumenstiel

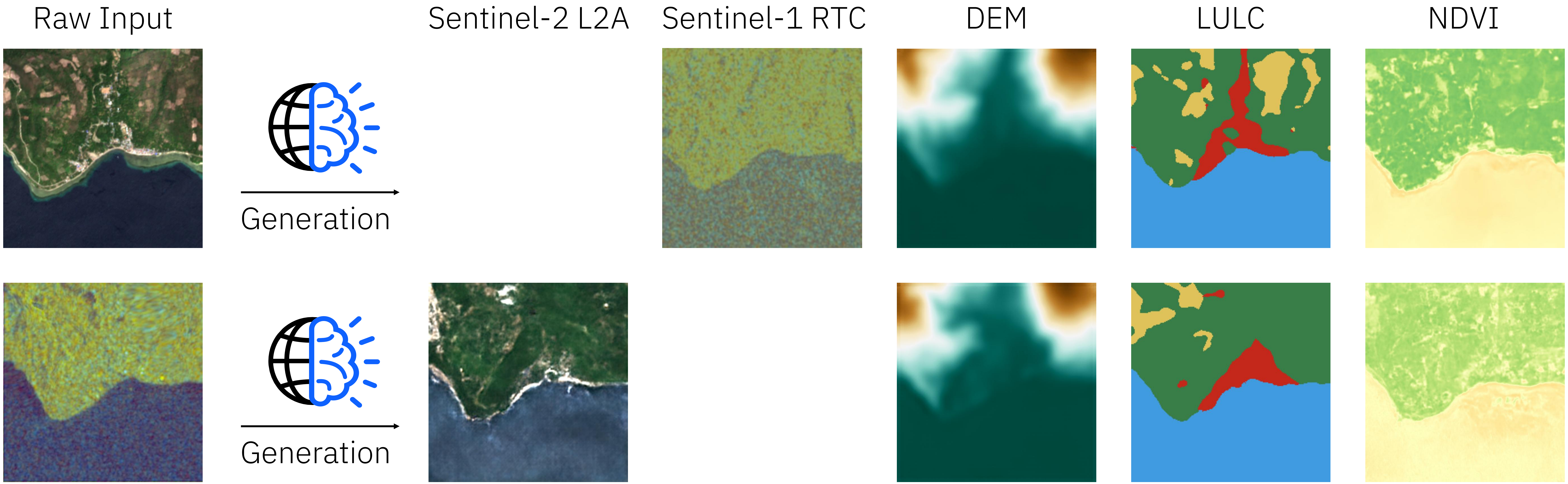
R. S. Kuzu, T. Brunschwiler, G. Cavallaro, J. Nalepa, 5, C. O. Dumitru, A. Zappacosta, D. E. Molina, R. Kienzler, J. Jakubik, B. Blumenstiel, Paolo Fraccaro, Felix Yang, R. Sedona, S. Maurogiovanni, E. Scheurer, A. Wijata, L. Tulczyjew, D. Marek, J. Sadel, S. Ofori-Ampofo , N. Dionelis, N. Longepe

FAST-EO
Project Overview

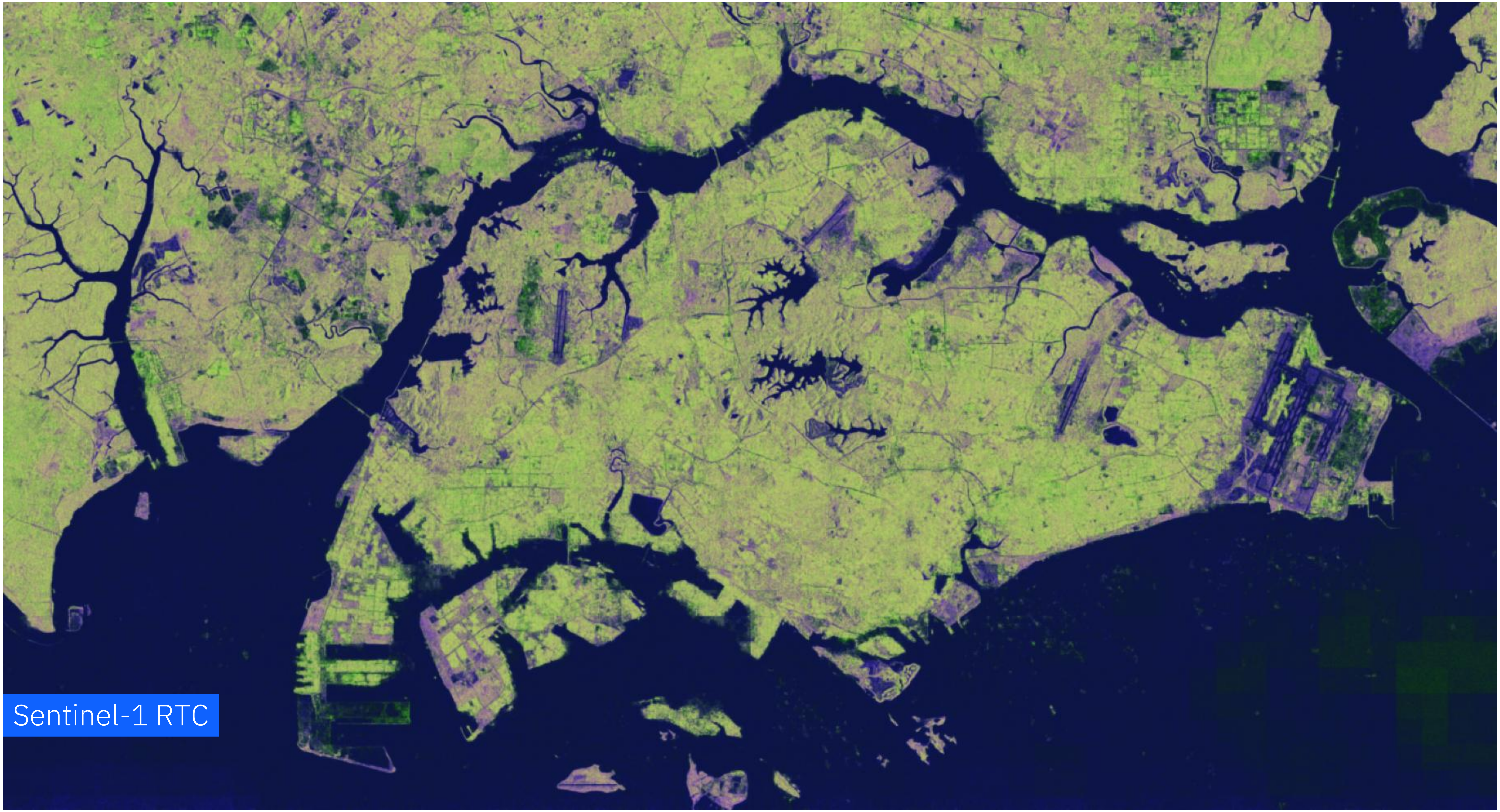
- ✓ **Mission**: Build foundation models for Earth Observation using self- and unsupervised multimodal learning
- ✓ **4M4EO Framework**: Fuse optical, SAR, hyperspectral, metadata & text into one masked modeling pipeline
- ✓ **Exascale Training**: Leverage JUPITER supercomputer for large-scale model training
- ✓ **Open Ecosystem:** TerraTorch for end-to-end data prep, training & evaluation
- ✓ **Applications:** Addressing regression, classification, and segmentation problems, involving text captioning in some of the applications

- UC1: Weather & Climate Disaster Analysis
- UC2: Detection of Methane Leaks
- UC3: Observation of Changes in Forest Above-Ground Biomass
- UC4: Estimation of Soil Properties
- UC5: Detection of Semantic Land Cover Changes
- UC6: Monitoring Expansion of Mining Fields into Farmlands

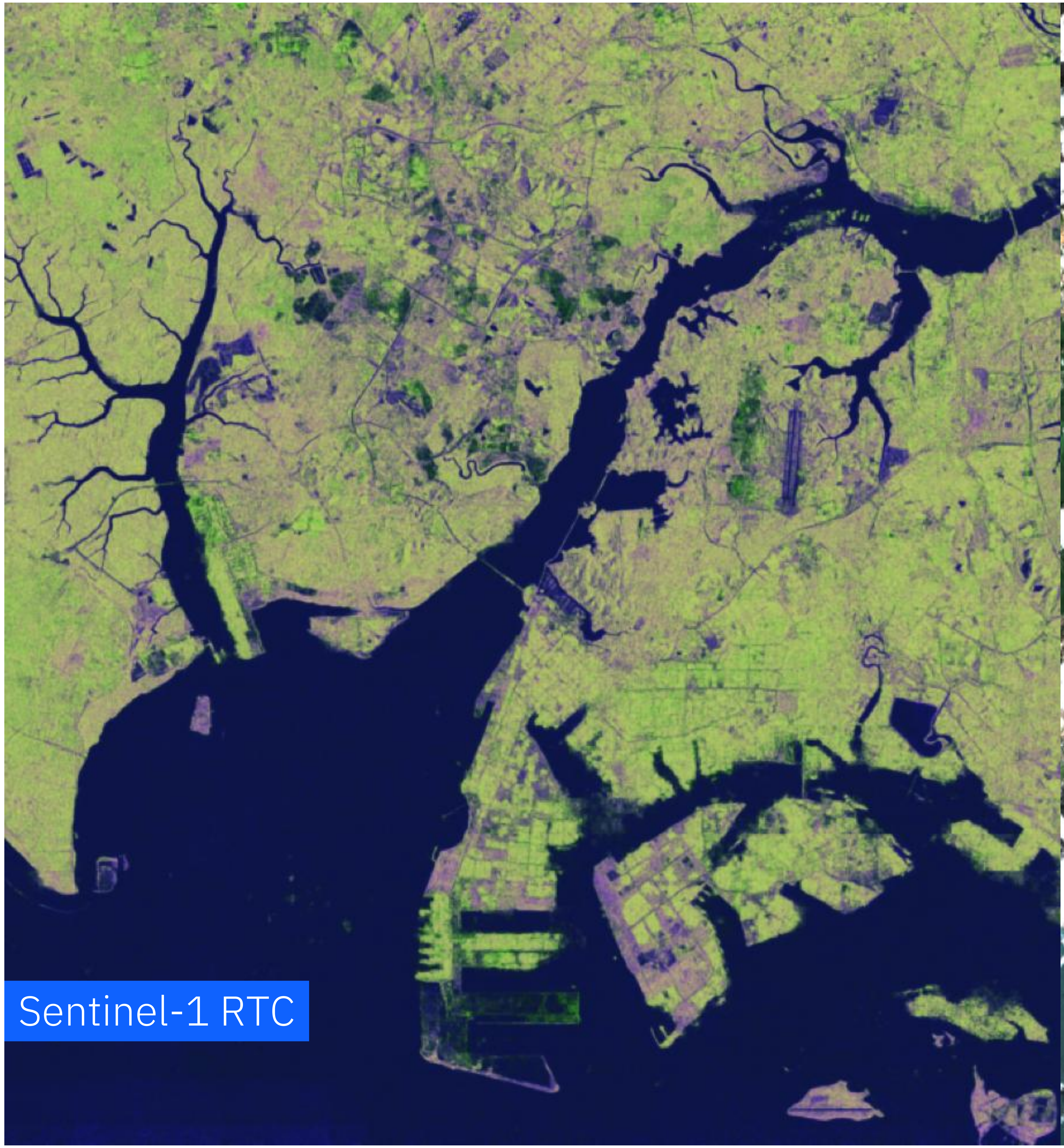
TerraMind – our first foundation model with cross-modal understanding

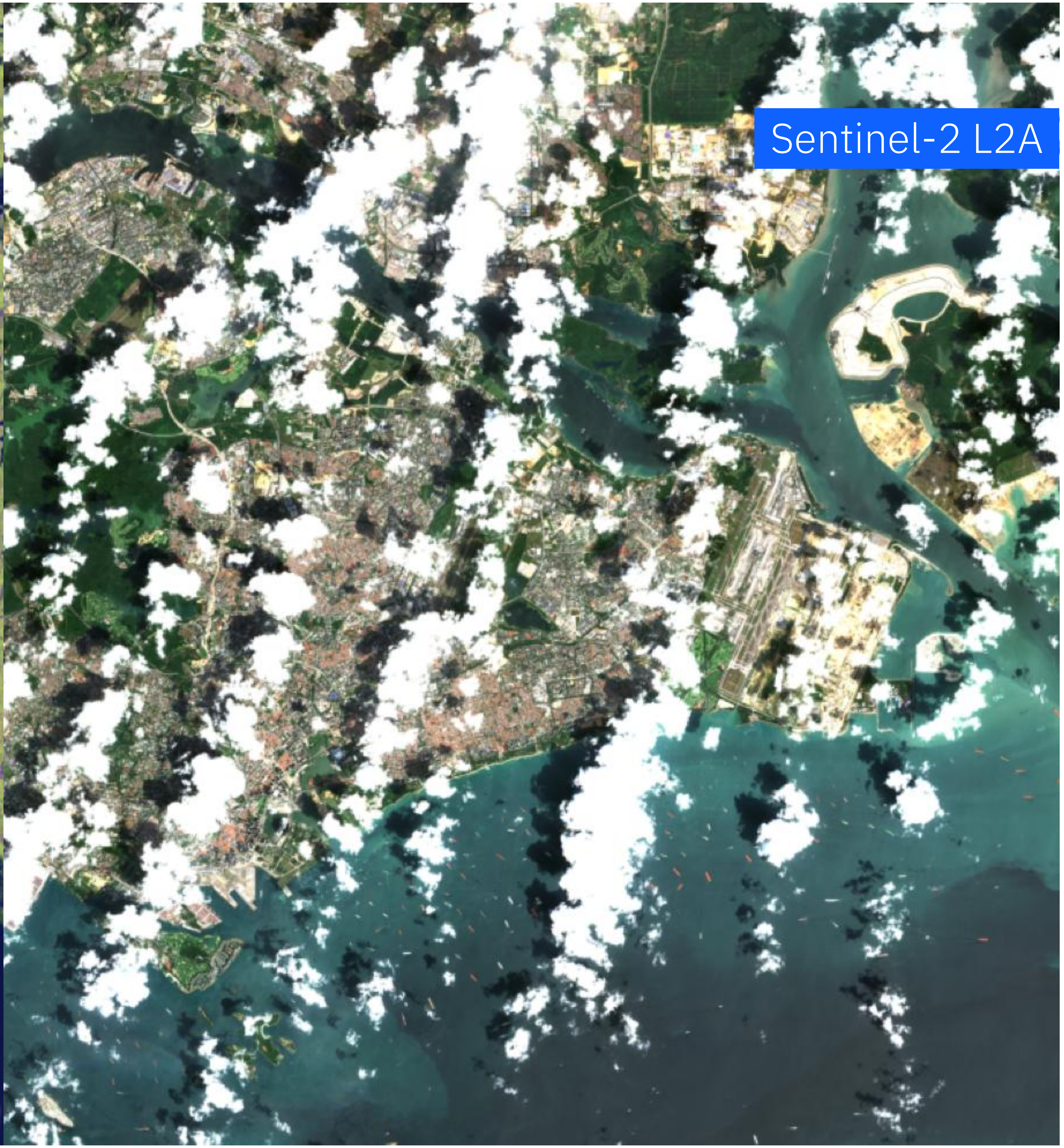


Sentinel-2 L2A

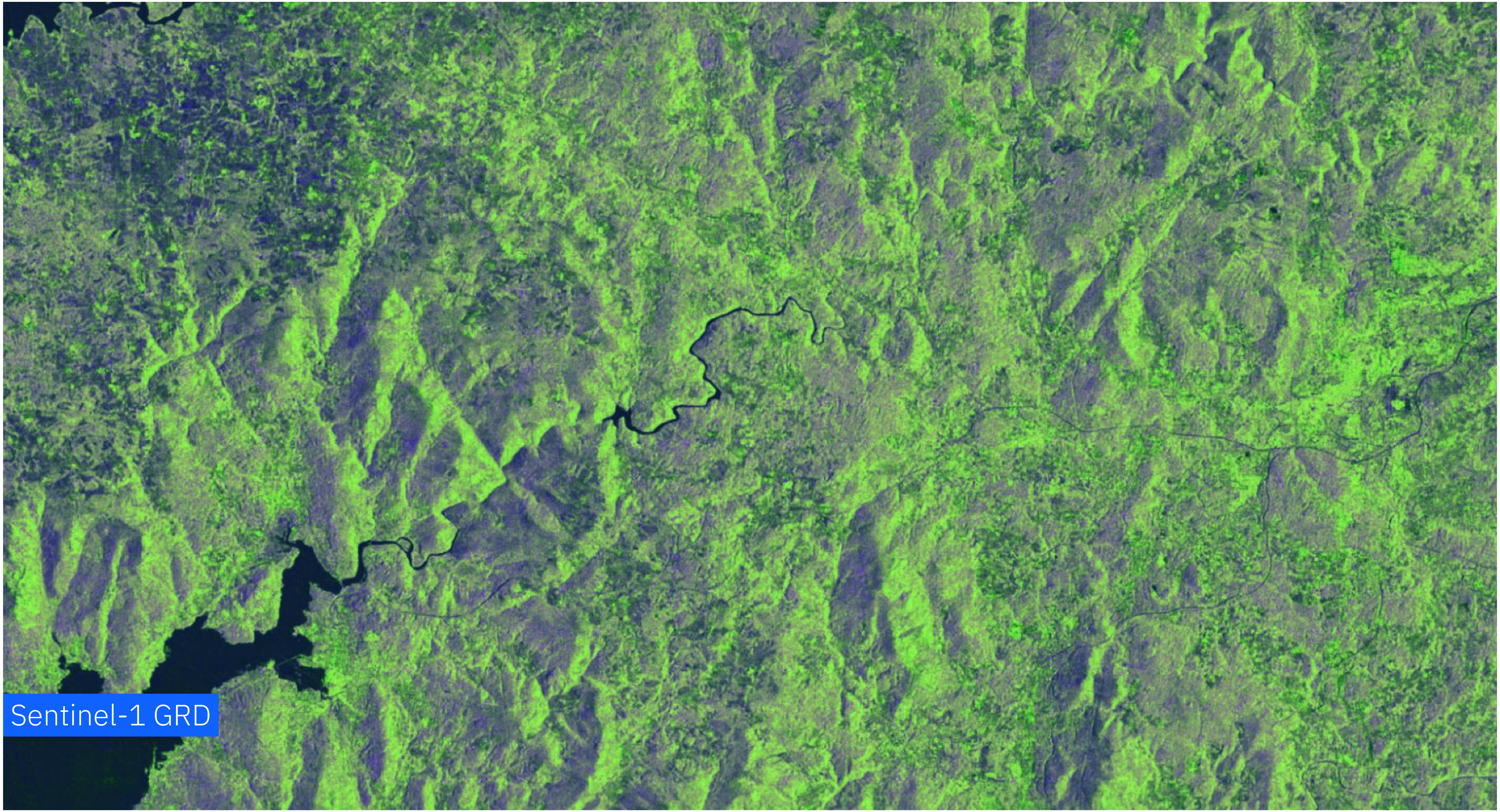


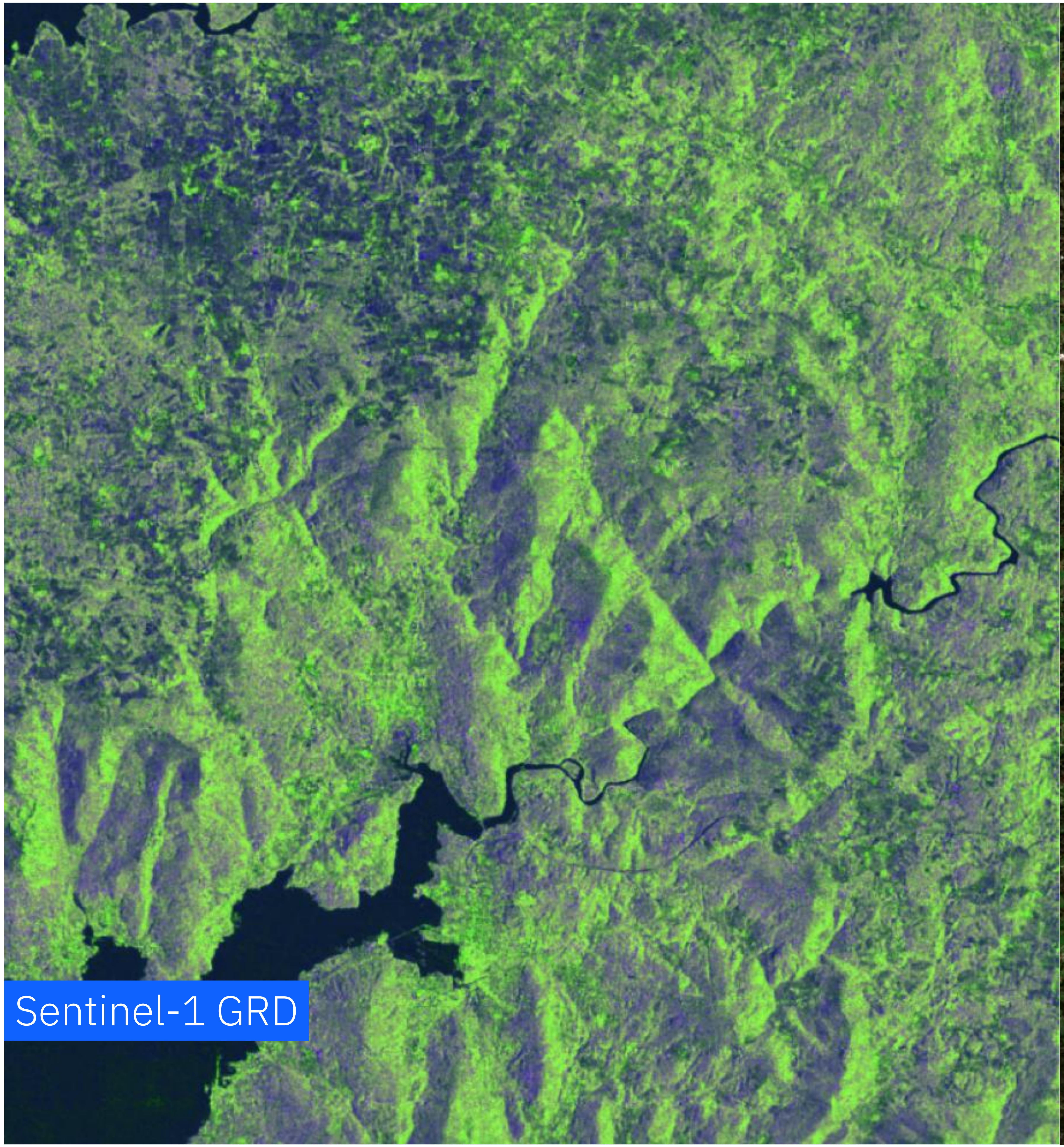
Sentinel-1 RTC

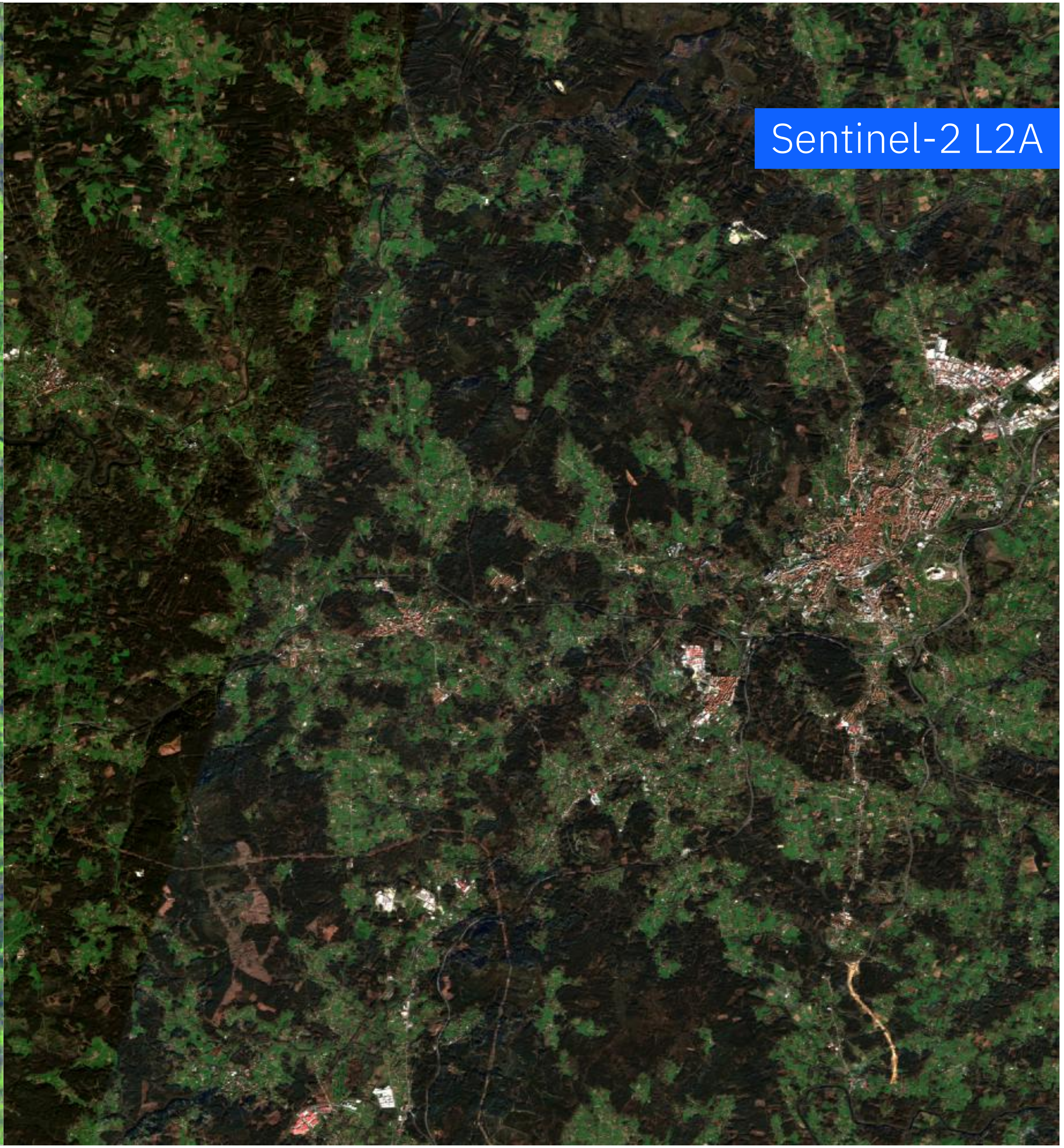

Sentinel-1 RTC


Sentinel-2 L2A

Sentinel-2 L2A


Sentinel-1 GRD


Sentinel-1 GRD

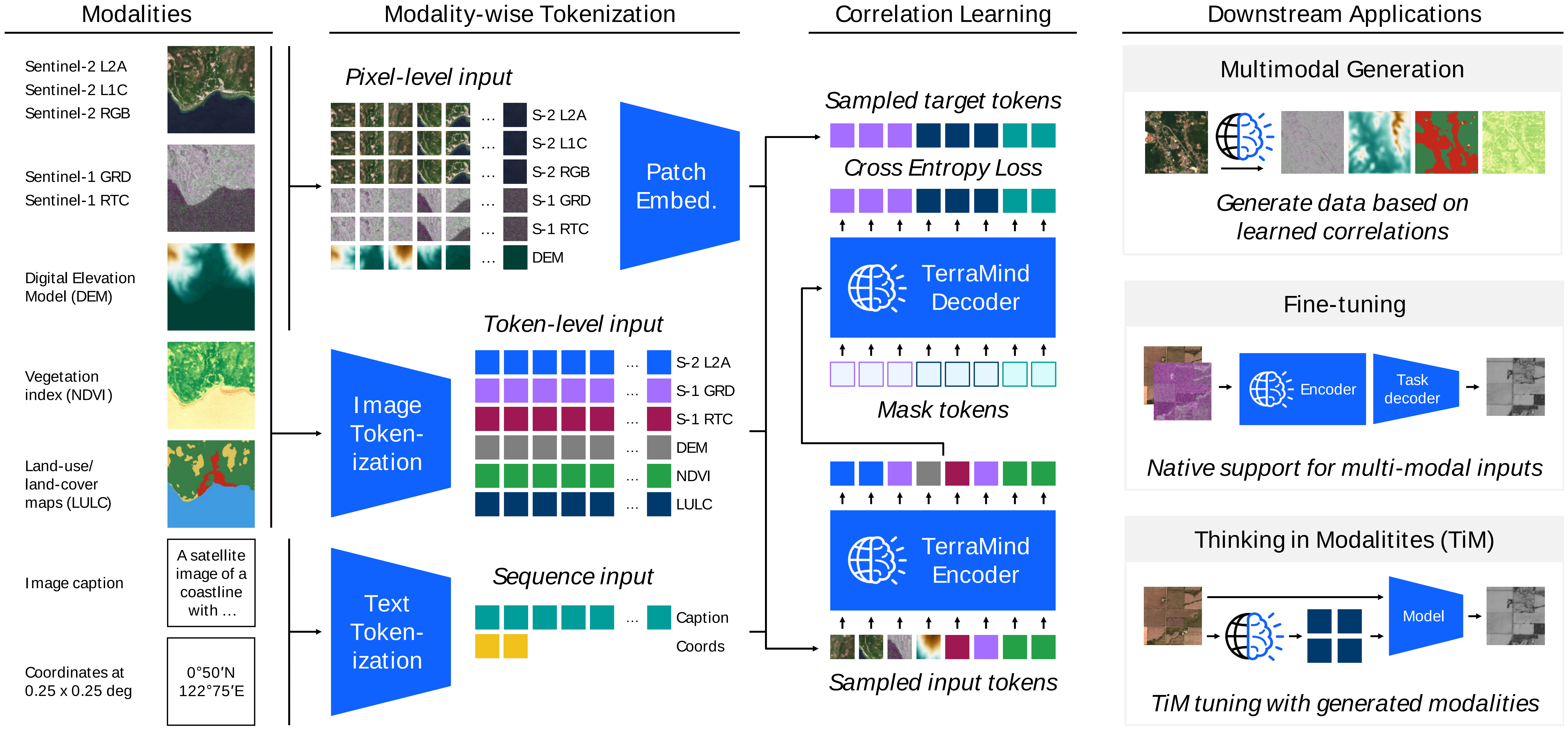

Sentinel-2 L2A

TerraMind

TerraMind represents the first any-to-any generative, and large-scale **multimodal** model for Earth observation pre-trained on **500 billion tokens** from global geospatial data.

The model digests inputs at **pixel-level, token-level, and as sequences**, simultaneously.

TerraMind outperforms other deep learning models for Earth observation in downstream applications and unlocks **any-to-any generation** and **Thinking-in-Modalities (TiM)** finetuning and inference.



Modalities

Sentinel-2 L2A
Sentinel-2 L1C
Sentinel-2 RGB

Sentinel-1 GRD
Sentinel-1 RTC

Digital Elevation Model (DEM)

Vegetation index (NDVI)

Land-use/ land-cover maps (LULC)

Image caption: A satellite image of a coastline with …

Coordinates at 0.25 x 0.25 deg: 0°50′N 122°75′E

Modality-wise Tokenization

Pixel-level input: S-2 L2A, S-2 L1C, S-2 RGB, S-1 GRD, S-1 RTC, DEM → Patch Embed.

Token-level input: Image Token-ization → S-2 L2A, S-1 GRD, S-1 RTC, DEM, NDVI, LULC

Sequence input: Text Token-ization → Caption, Coords

Correlation Learning

Sampled target tokens

Cross Entropy Loss

TerraMind Decoder

Mask tokens

TerraMind Encoder

Sampled input tokens

Downstream Applications

Multimodal Generation

Generate data based on learned correlations

Fine-tuning

Encoder → Task decoder

Native support for multi-modal inputs

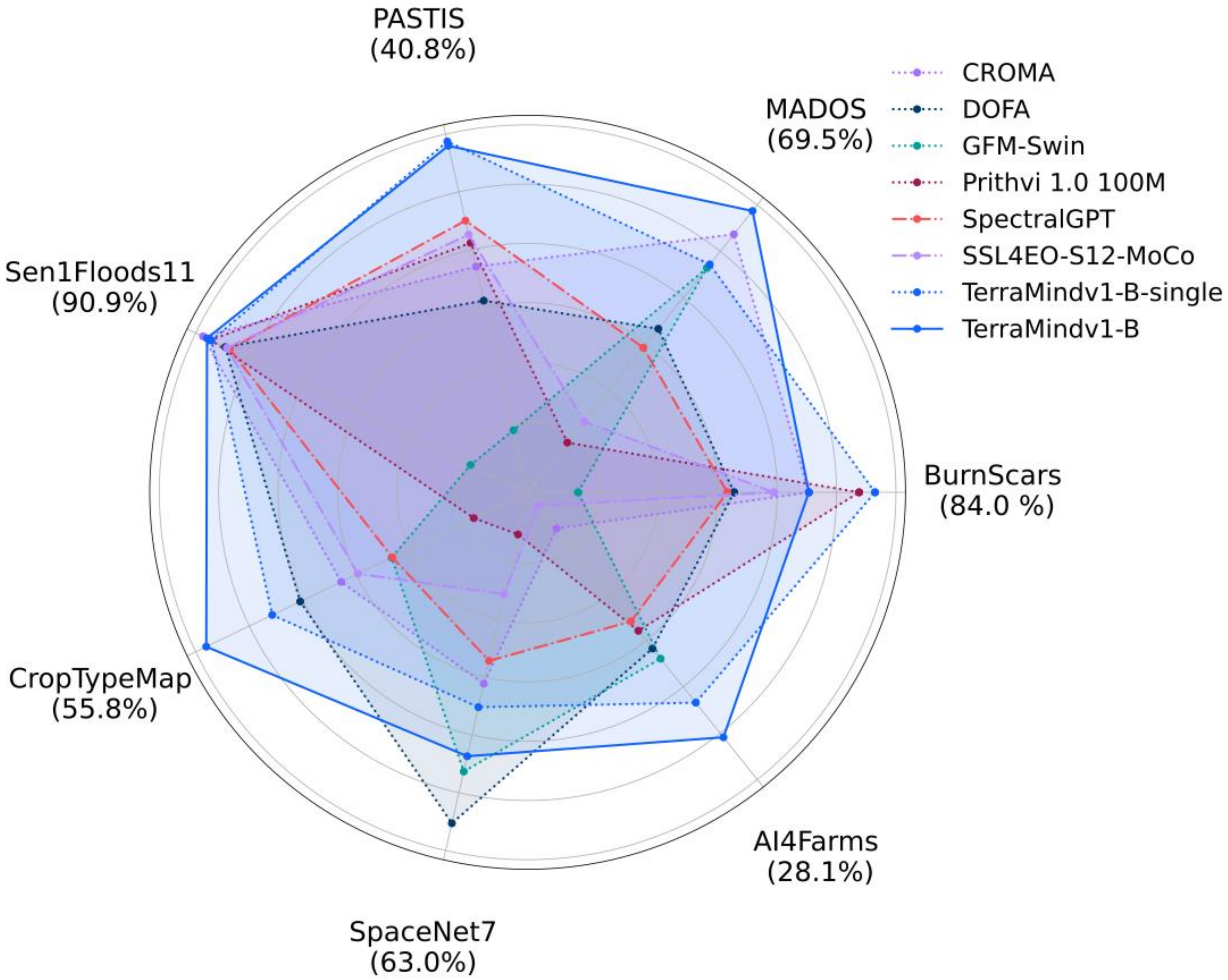
Thinking in Modalitites (TiM)

Model

TiM tuning with generated modalities

Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., ... & Longépé, N. (2025). TerraMind: Large-Scale Generative Multimodality for Earth Observation. *arXiv preprint arXiv:2504.11171*.

Evaluation on PANGAEA bench



TerraMind is evaluated on PANGAEA bench with a diverse set of modalties and downstream tasks – with a frozen encoder.

It outperforms all other evaluated geospatial foundation models and even fully fine-tuned UNet and ViT models.

TerraMind benefits from multi-modal inputs and the new Thinking-in-Modalities approach for improved performance results.

Marsocci, V., Jia, Y., Bellier, G. L., Kerekes, D., Zeng, L., Hafner, S., ... & Nascetti, A. (2024). PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*.

MS CLIP – Zero-shot applications via contrastive learning

Vision Language Models enable interactive applications based on natural language.

CLIP is the most promindent model with zero-shot classification and text-to-image retrieval capabilities.

Show me images of “agricultural fields next to a canal”

Check if these images include a “solar farm”

Retrieved images based on image-text similarity

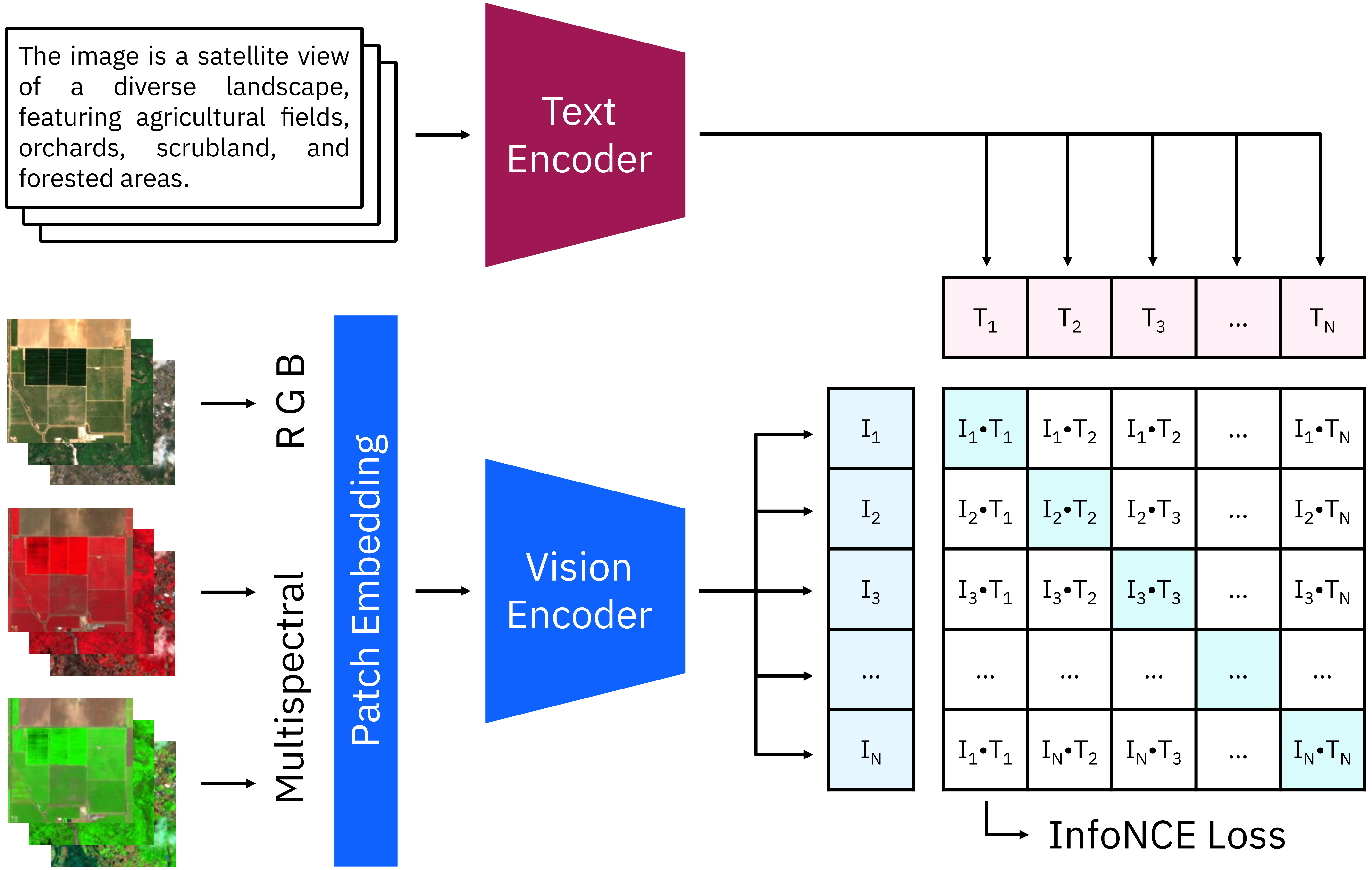
Classification with *solar farm* vs. *others*

MS CLIP

CLIP[1] is trained with Contrastive Learning on 400M image-text pairs.

But the model does not generalize well on domain specific tasks.

Continous pre-training with additional channels for EO domain adoption.

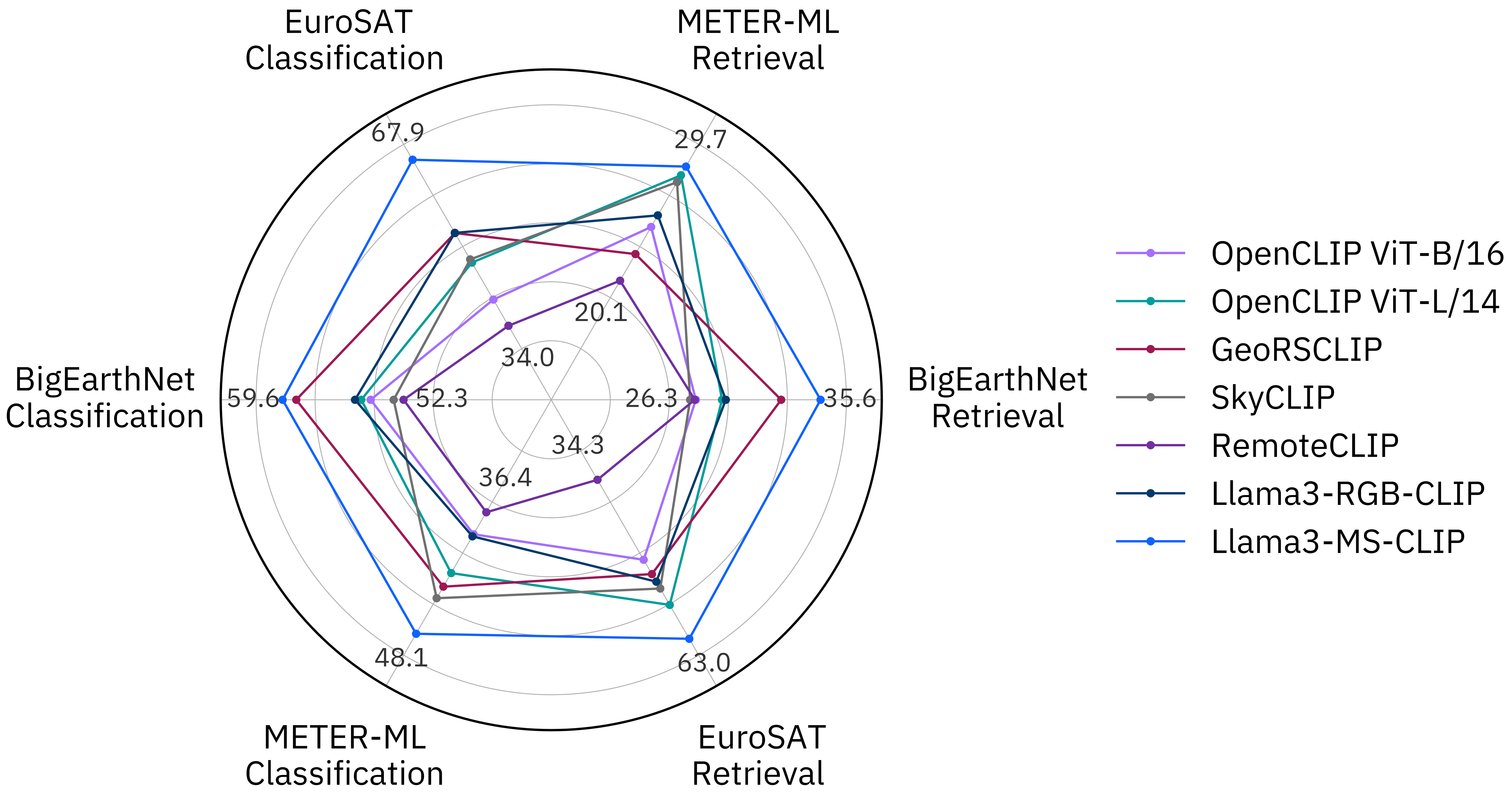


The image is a satellite view of a diverse landscape, featuring agricultural fields, orchards, scrubland, and forested areas.

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_2$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
...	...	...	...	...	...
I_N	$I_1 \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

MS CLIP – Zero-shot evaluation



Marimo, C. T., Blumenstiel, B., Nitsche, M., Jakubik, J., & Brunschwiler, T. (2025). Beyond the Visible: Multispectral Vision-Language Learning for Earth Observation. *arXiv preprint arXiv:2503.15969*.

MS CLIP outperforms all baselines and EO-specific VLMs.

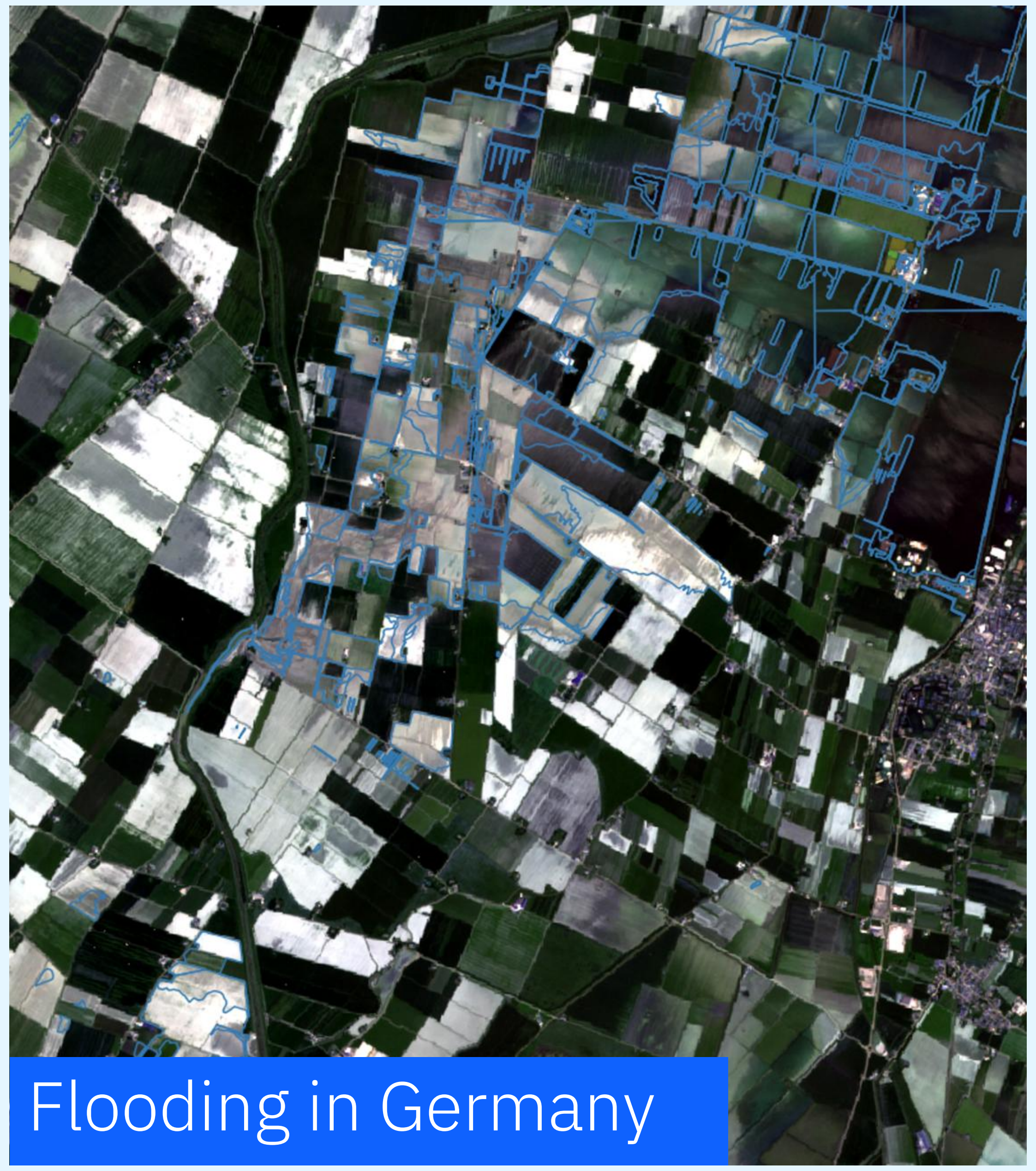
MS CLIP improves classification accuracy by +6.77% on average and retrieval by +4.63% mAP compared to the second-best model.

Benefit of multi-spectral data as our RGB-CLIP only performs on pair with other EO VLMs.

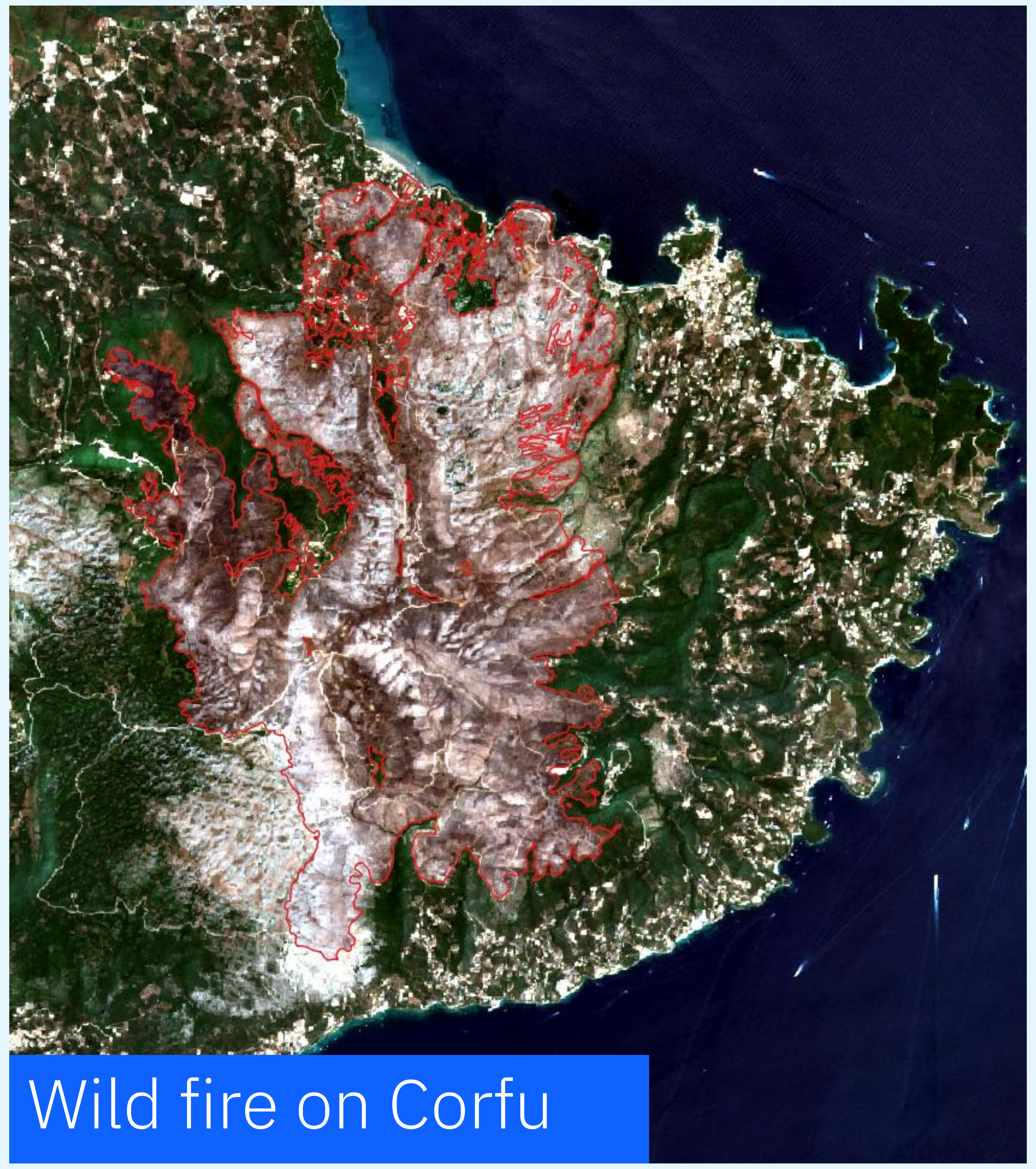
Downstream Applications

Climate impact analysis

Building a large-scale, multi-modal, multi-temporal dataset for predicting various disaster types. Release in Q2 – stay tuned!



Flooding in Germany



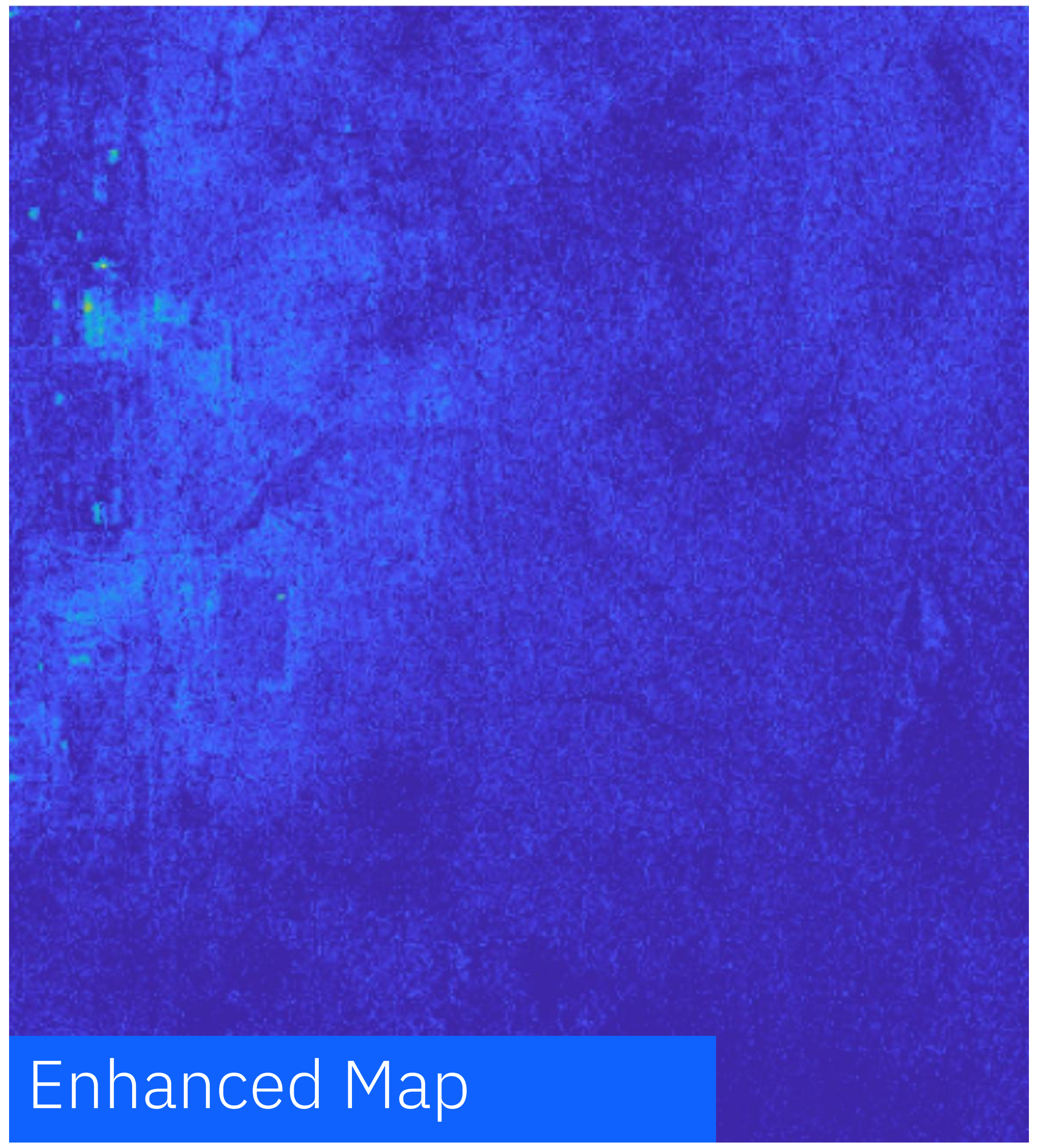
Wild fire on Corfu

FAST-EO / ESA Contract No. 4000143501/23/I-DT

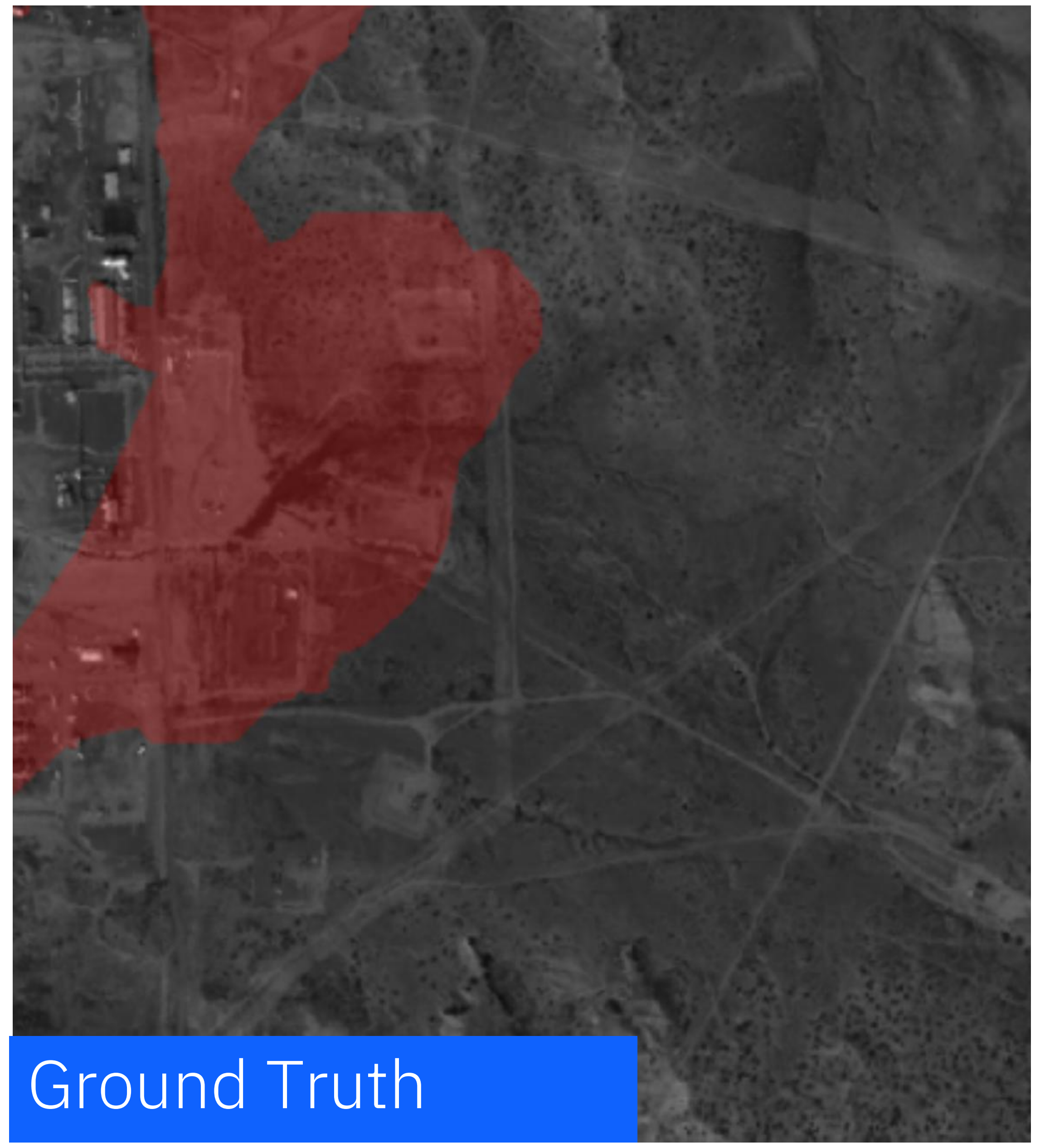
Detection of Methane Leaks

Fine-tuning TerraMind to detect methane leaks in airborne and satellite imagery, thereby surpassing the benchmarks.

Model	ACC	SPE	SEN	F-Score	MCC
FM – 12 bands	**0.841**	**0.823**	**0.869**	**0.853**	**0.676**
Baseline – 432 bands	*0.680*	*0.380*	*0.990*	*0.760*	*0.460*



Enhanced Map



Ground Truth

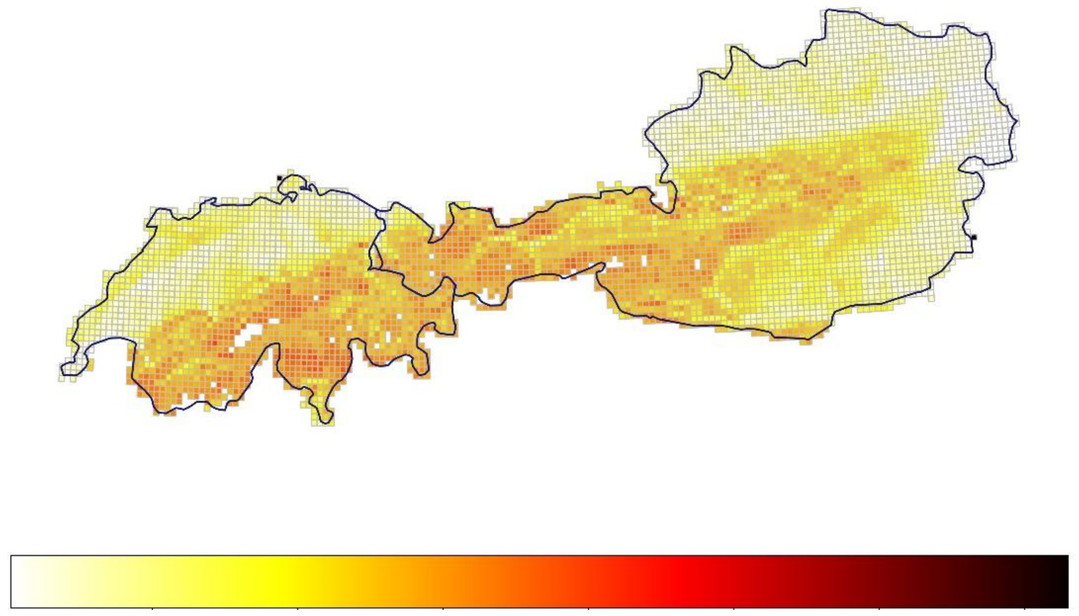
Downstream Applications

Forest Biomass Change Monitoring

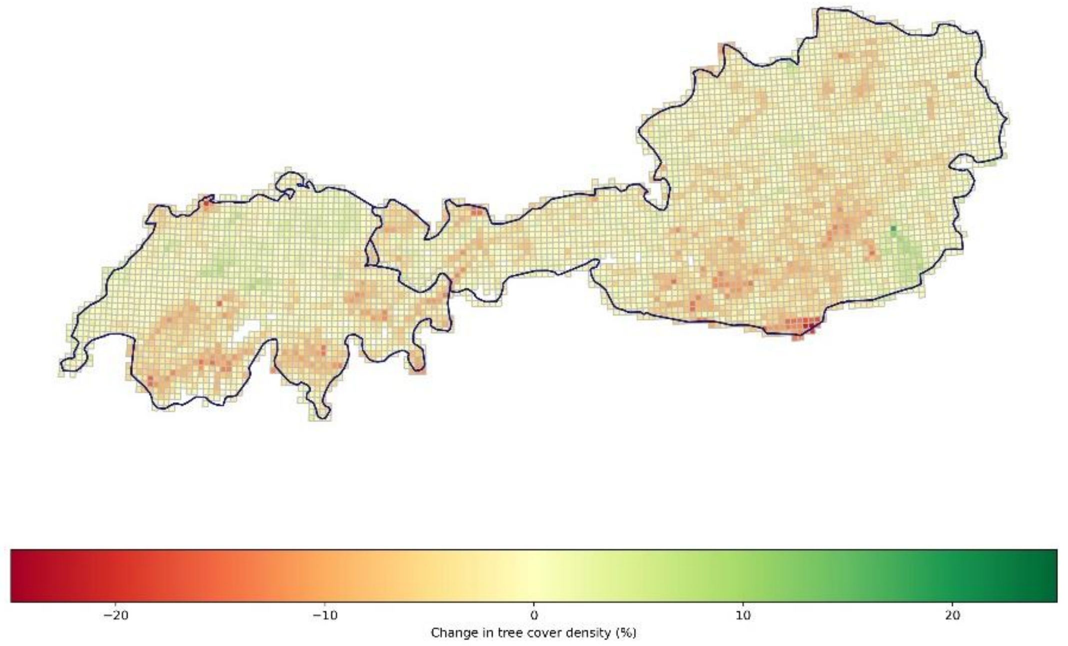
Investigating how terrain topography affects the detection of forest biomass changes and seeking improvements for mountainous regions using the TerraMind model.

.

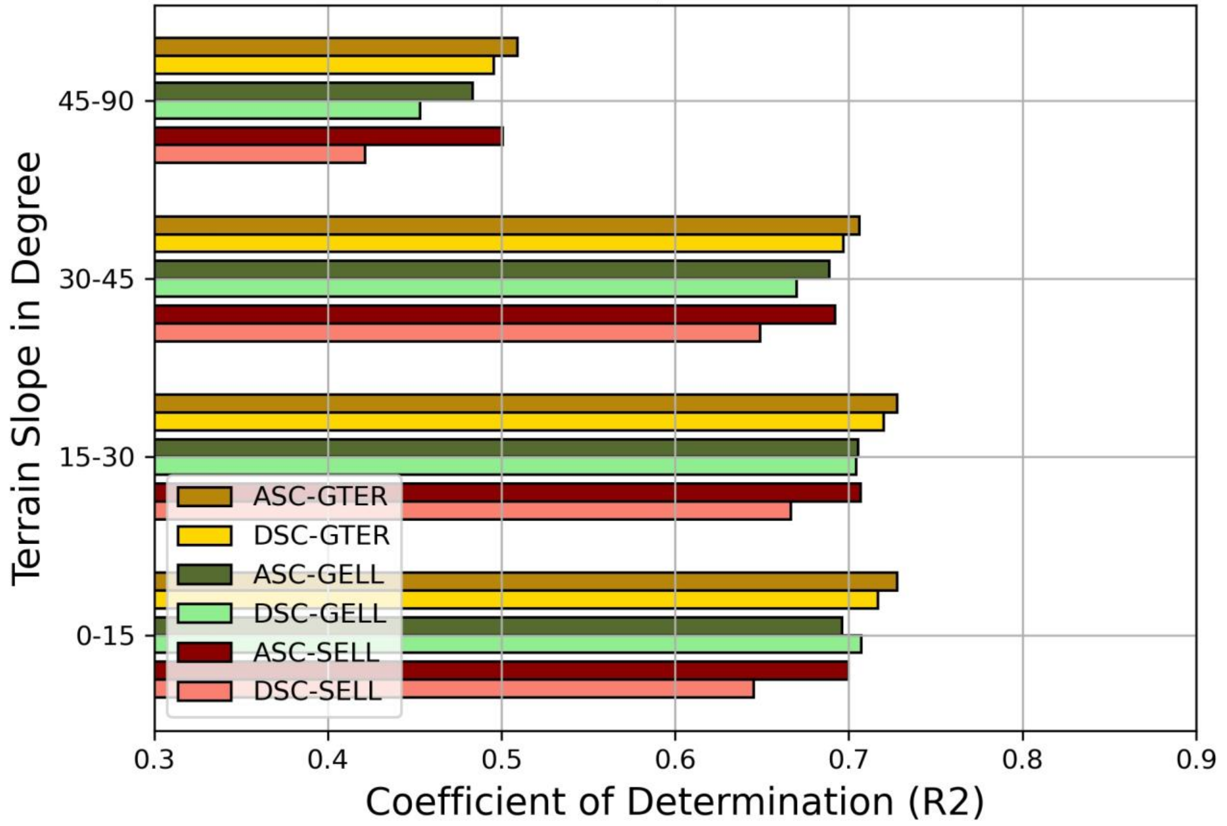
Terrain slope map in Austria and Switzerland



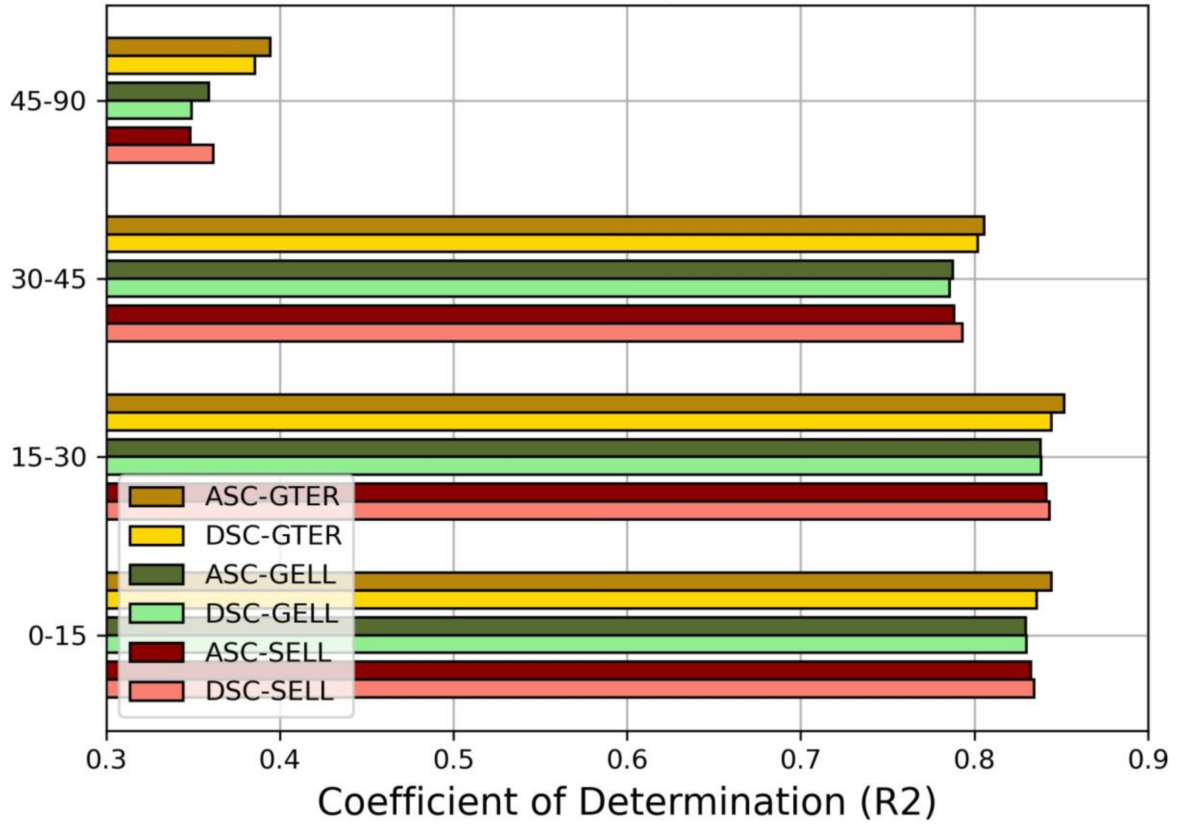
Tree cover density change map in Austria and Switzerland - between 2015-2018



R2 vs Terrain slope for Switzerland

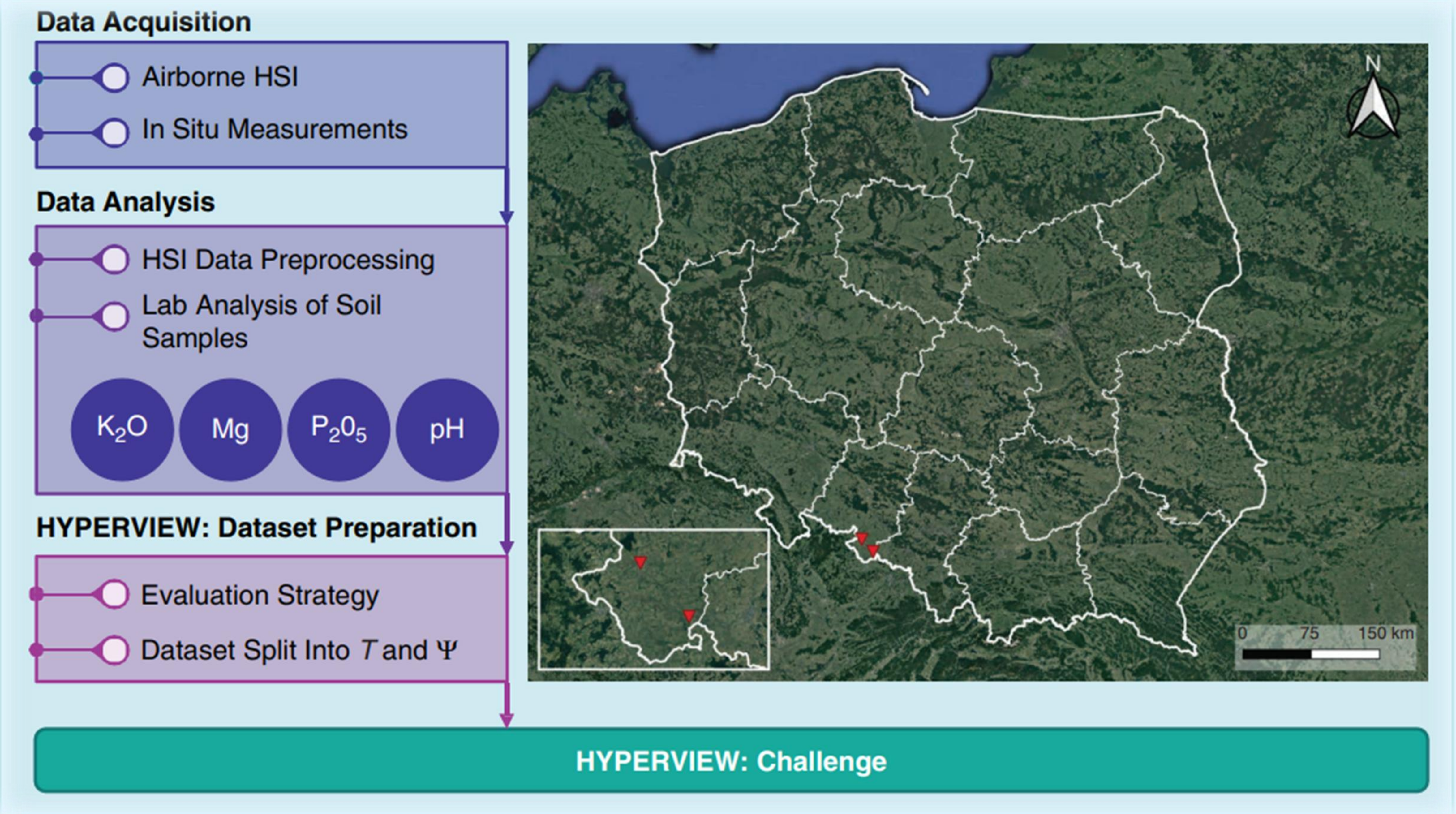


R2 vs Terrain slope for Austria



Estimation of Soil Properties

Fine-tuning the TerraMind model to regress soil properties, achieving a HYPERVIEW score of 0.7943 and securing 1th place in the benchmark.



Data Acquisition
- Airborne HSI
- In Situ Measurements

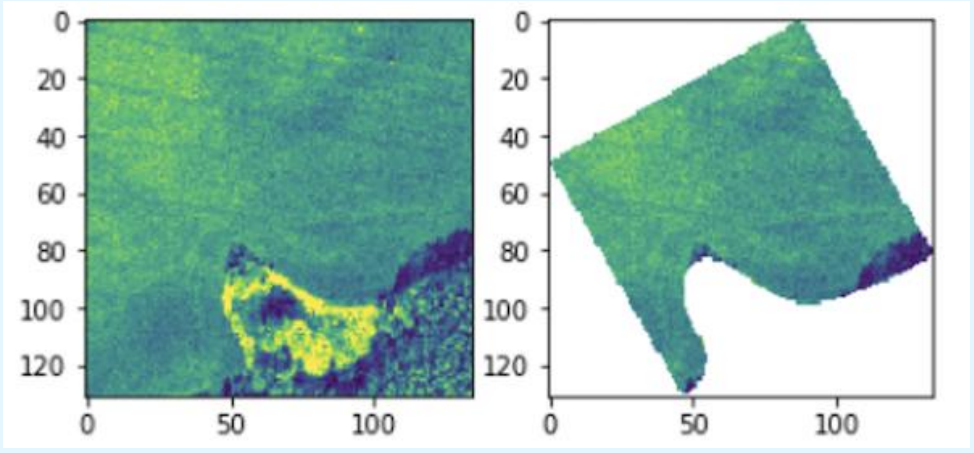
Data Analysis
- HSI Data Preprocessing
- Lab Analysis of Soil Samples

K_2O Mg P_2O_5 pH

HYPERVIEW: Dataset Preparation
- Evaluation Strategy
- Dataset Split Into T and Ψ

HYPERVIEW: Challenge

Model	HYPERVIEW Score
TerraMind	0.8374
TerraMind & KNN	0.7943
Prithivi	0.8483
ResNet50	0.8444
ViT	0.8958

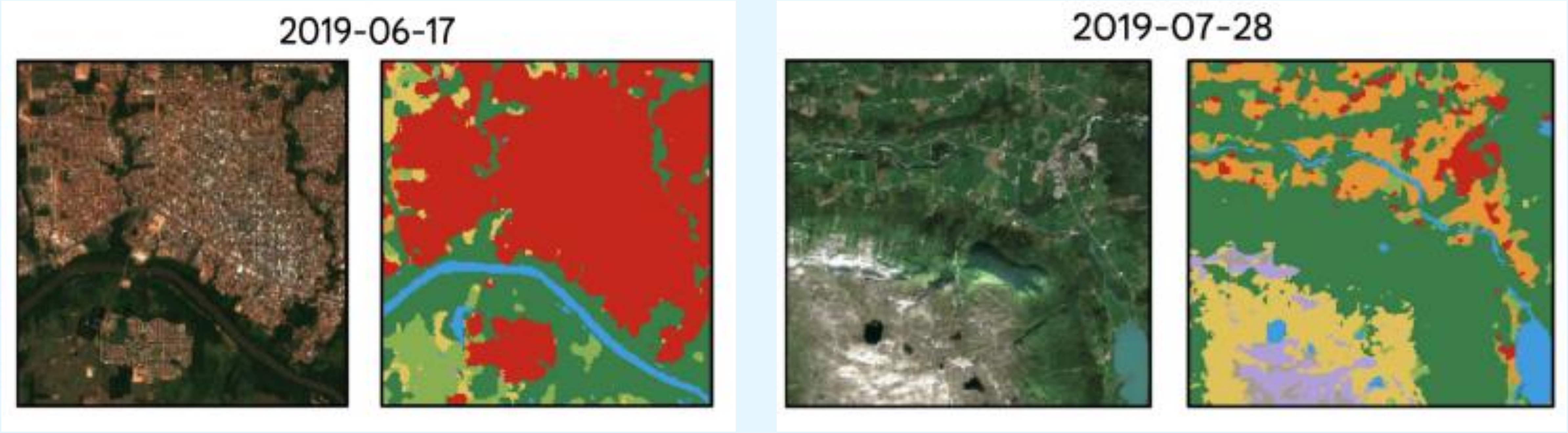


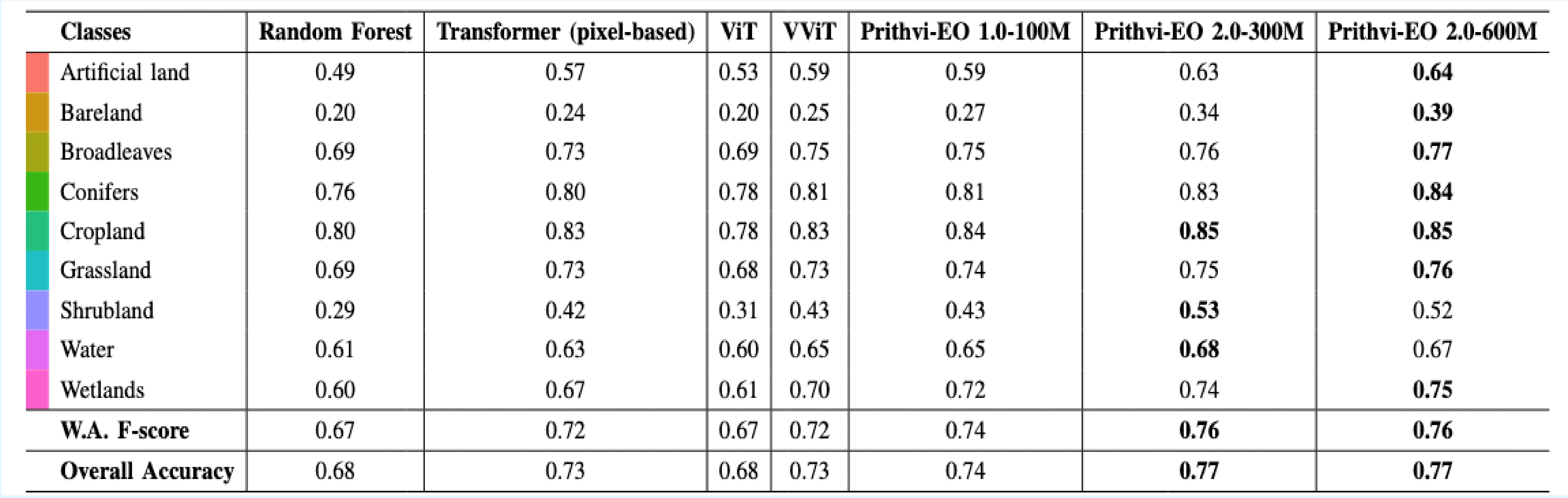
Downstream Applications

Detection of Semantic Land Cover Changes

Curating the 335,125-patch Sen4Map Sentinel-2 time-series dataset and fine-tuning Geo-FMs for semantic land-cover change detection—establishing robust benchmarks with Random Forest, pixel-based Transformers, ViTs, and Video ViTs.

.



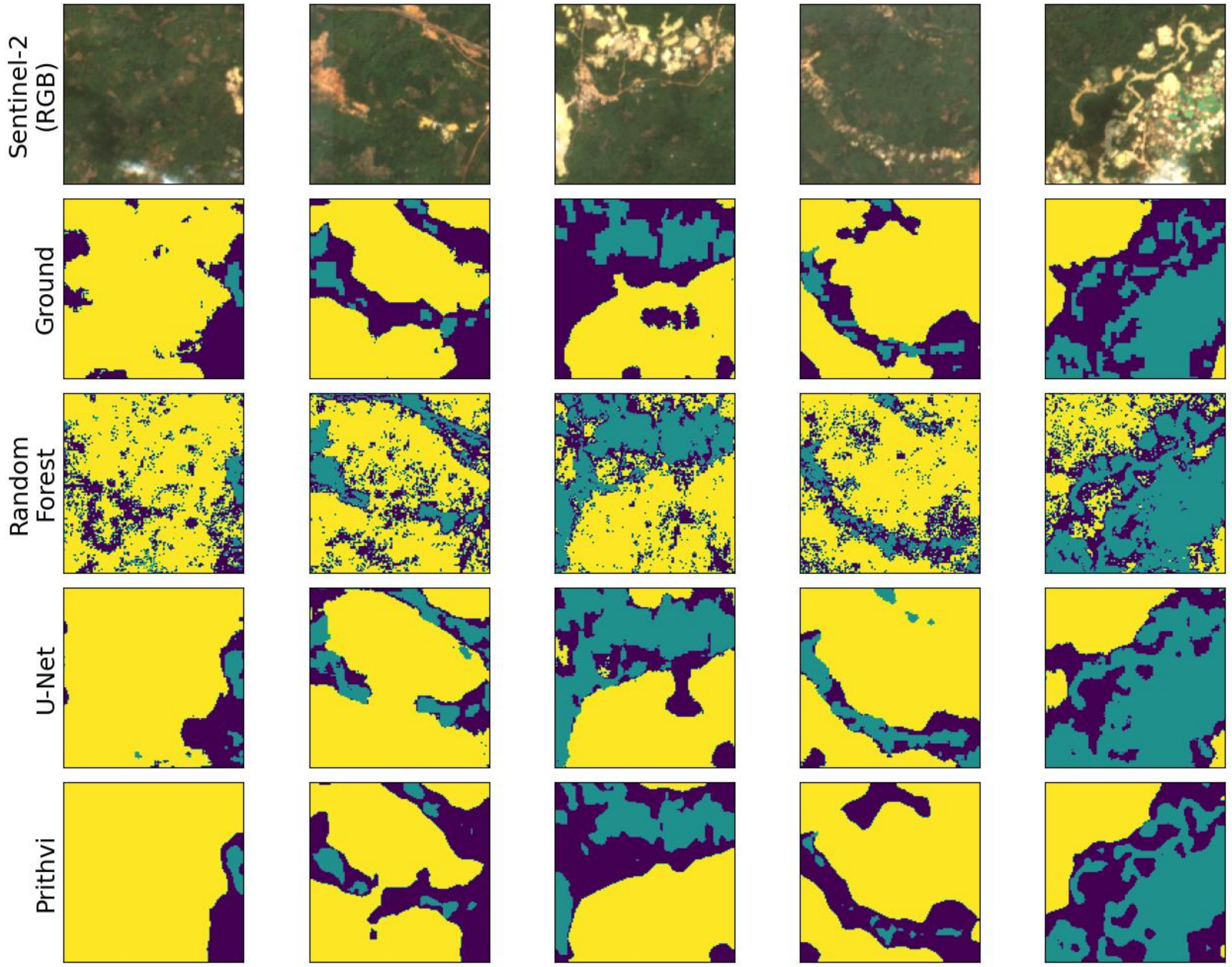


Classes	Random Forest	Transformer (pixel-based)	ViT	VViT	Prithvi-EO 1.0-100M	Prithvi-EO 2.0-300M	Prithvi-EO 2.0-600M
🟥 Artificial land	0.49	0.57	0.53	0.59	0.59	0.63	**0.64**
🟧 Bareland	0.20	0.24	0.20	0.25	0.27	0.34	**0.39**
🟨 Broadleaves	0.69	0.73	0.69	0.75	0.75	0.76	**0.77**
🟩 Conifers	0.76	0.80	0.78	0.81	0.81	0.83	**0.84**
🟩 Cropland	0.80	0.83	0.78	0.83	0.84	**0.85**	**0.85**
🟦 Grassland	0.69	0.73	0.68	0.73	0.74	0.75	**0.76**
🟪 Shrubland	0.29	0.42	0.31	0.43	0.43	**0.53**	0.52
🟪 Water	0.61	0.63	0.60	0.65	0.65	**0.68**	0.67
🟪 Wetlands	0.60	0.67	0.61	0.70	0.72	0.74	**0.75**
W.A. F-score	0.67	0.72	0.67	0.72	0.74	**0.76**	**0.76**
Overall Accuracy	0.68	0.73	0.68	0.73	0.74	**0.77**	**0.77**

FAST-EO / ESA Contract No. 4000143501/23/I-DT

Monitoring Expansion of Mining Fields into Farms

Artisanal gold-mining segmentation using Sentinel-1, Sentinel-2, and DEM data—achieving a mean IoU of 0.76 surpassing the benchmarks. Stay tuned for the improved results with TerraMind 1.0 soon!



S. Ofori-Ampofo, A. Zappacosta, R. S. Kuzu, P. Schauer, M. Willberg and X. X. Zhu, "SmallMinesDS: A Multi-Modal Dataset for Mapping Artisanal and Small-Scale Gold Mines," in *IEEE Geoscience and Remote Sensing Letters*, doi: 10.1109/LGRS.2025.3566356

Next Steps for FAST-EO

- ✓ Develop and Open-Source TerraMesh+
- ✓ Integrate Advanced SAR Capabilities
- ✓ Embed Trust and Governance Tools
- ✓ Optimize for Edge and Cloud Deployment
- ✓ Demonstrate End-to-End Operational Workflows

THANK YOU