HELMHOLTZ ZENTRUM
DRESDEN ROSSENDORF

HELMHOLTZ INSTITUTE FREIBERG
FOR RESOURCE TECHNOLOGY

Towards a Foundation Model for Global Terrestrial 3D Above and Below Ground Carbon Stock Mapping (3D-ABC)

Aldino Rizaldy

Team: Weikang Yu, Gunjan Joshi, Peter Steinbach, Pedram Ghamisi

Introduction

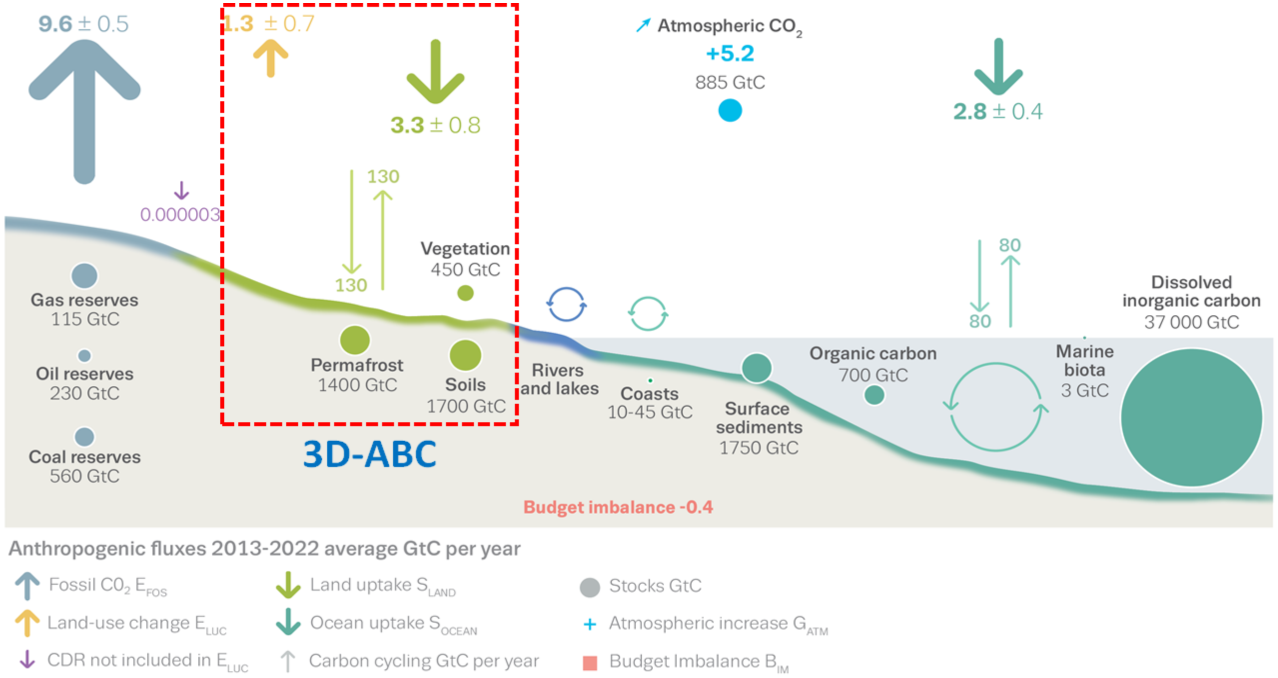
HELMHOLTZ Foundation Model Initiative

1 2 HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF 3 JÜLICH Forschungszentrum 4 GFZ 5 UFZ 6 DLR

We combine multimodal Earth Observation data into a Unified Foundation Model

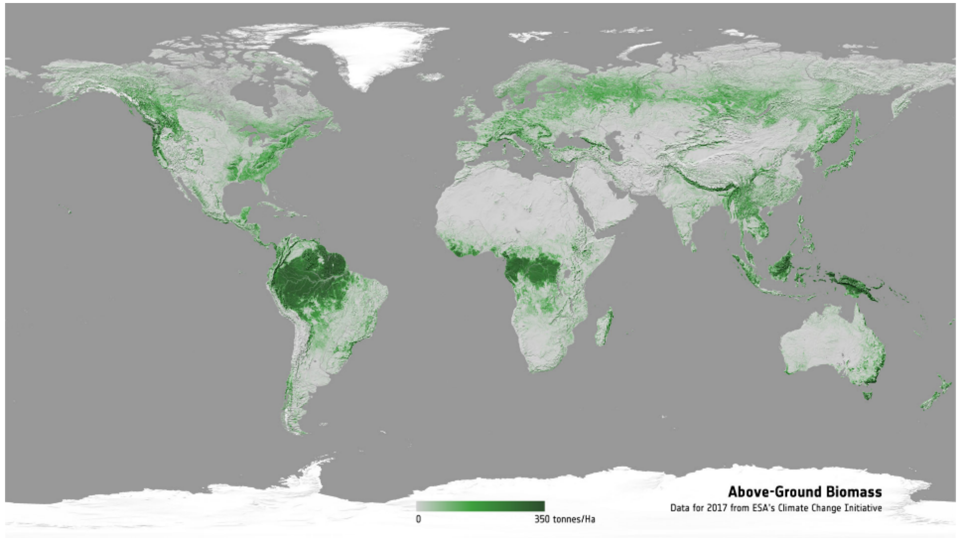
Global Carbon Stocks: A Dynamic Challenge

- Quantifying the global carbon budget is scientifically, economically, and politically important
- A large portion of the global carbon budget is linked to land use change, vegetation change, and soil processes
- Lack of globally comprehensive, spatially and temporally continuous, and consistent assessment of terrestrial C stocks

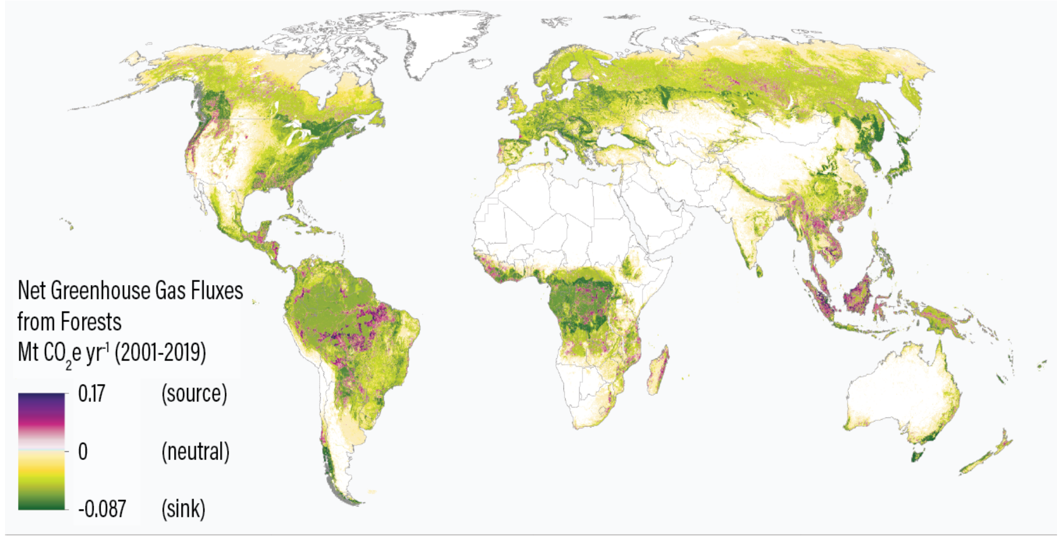


Mapping Global Terrestrial Carbon Stocks

Examples of current approaches



Global aboveground biomass for 2010, 2017-2020 at 100m spatial resolution. http://cci.esa.int/biomass (Santoro et al., 2021, ESSD)



Open data: https://gfw.global/315sv5h (Harris et al., 2021, Nature Climate Change)

Challenges

- Various approaches exist, but they either are snapshots, or lack high temporal, spatial, or thematic resolution
- Remote sensing projects and task-specific models require extensive labeled datasets and repeatedly significant computational resources

AI Foundation Models (FMs) in Earth Observation

- Large-scale, pre-trained AI models can perform a wide array of downstream tasks with unprecedented accuracy and efficiency

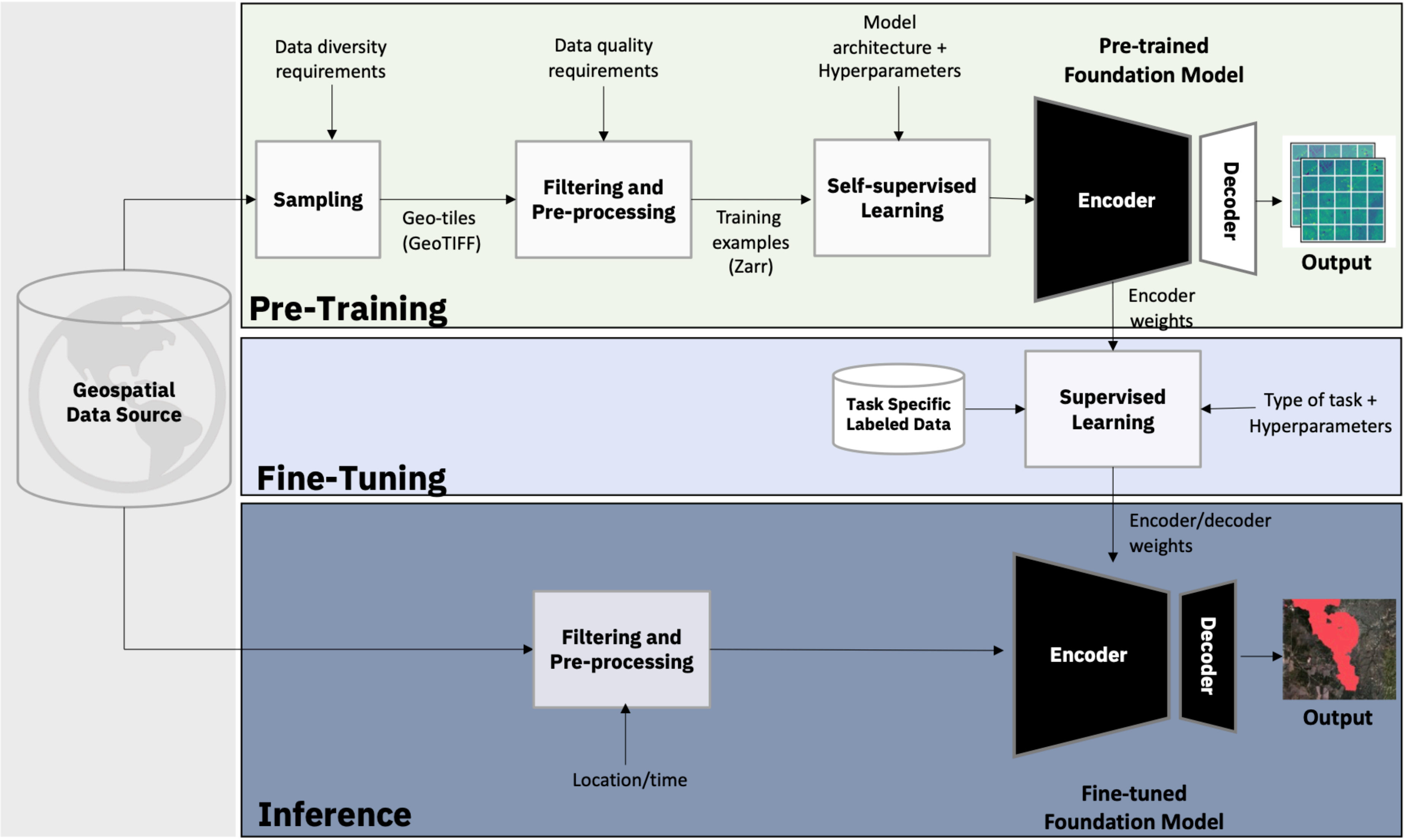
- Leverage extensive datasets and advanced architectures, enabling capturing complex patterns and features, can be fine-tuned for specific applications with minimal additional training

- Particularly valuable for the diverse and complex nature of multi-modal remote sensing data (optical, radar, lidar, multi-spectral, multi-resolution, multi-temporal, 2D, 3D)

- Strong performance in handling diverse remote sensing tasks such as classification, segmentation, object detection, regression

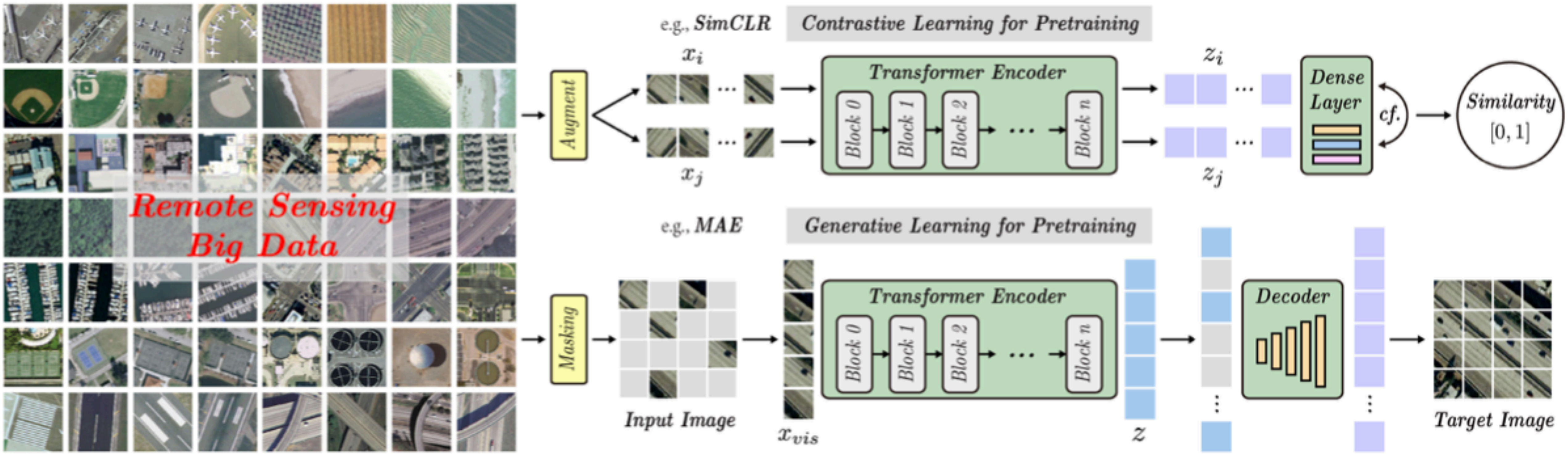
Lu et al. "AI foundation models in remote sensing: A survey." arXiv preprint arXiv:2408.03464 (2024).

Prithvi: Foundation Models for Generalist Geospatial Artificial Intelligence



https://arxiv.org/abs/2310.18660

SpectralGPT: Spectral Remote Sensing Foundation Model



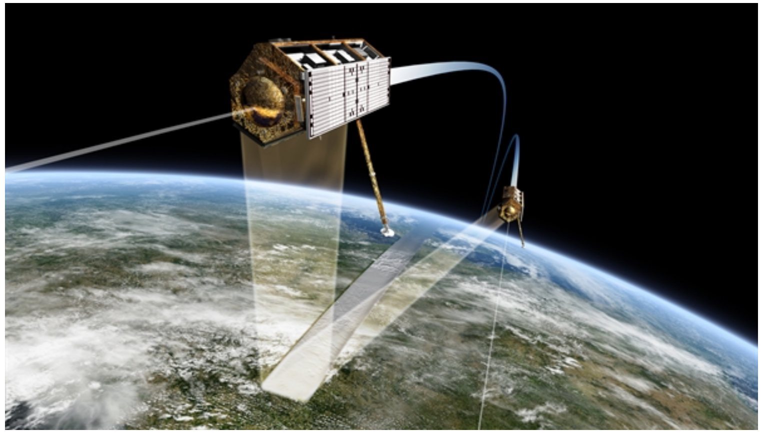
https://arxiv.org/abs/2311.07113

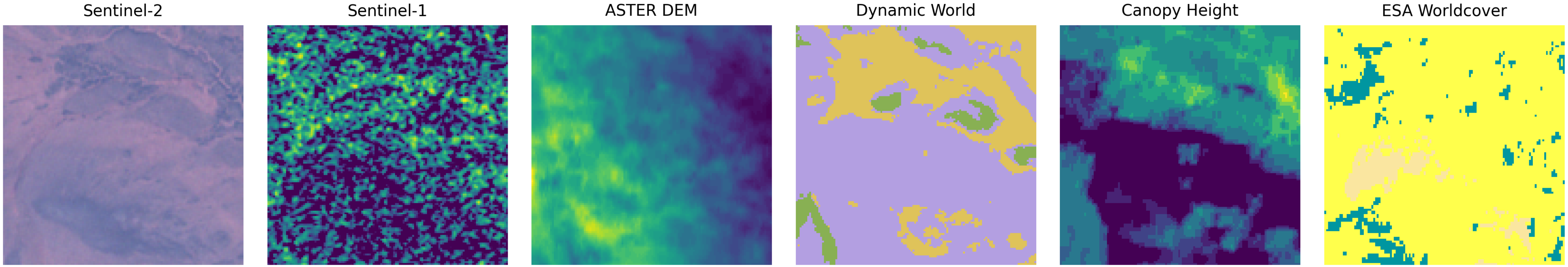
Limitations of Current FMs in Earth Observation for Carbon Monitoring

- Single-Modality or Weak Multimodal Integration
- Resolution Gaps
- Temporal Limitations
- Pretraining Objectives not Aligned with Carbon Estimation
  - Most are trained for semantic segmentation of land cover classification

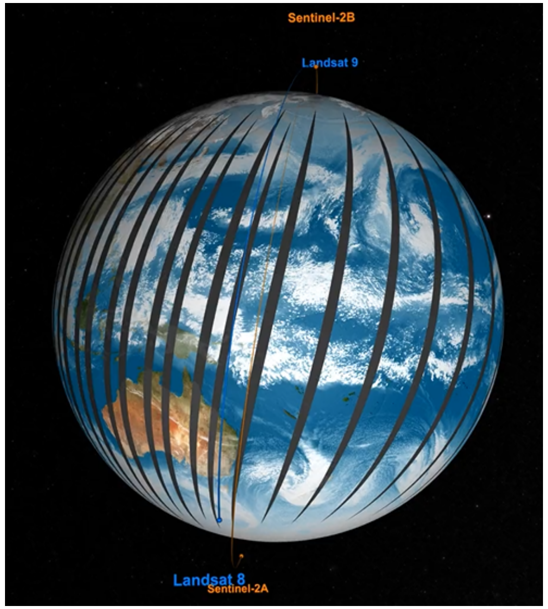
What is Earth Observation data?

- Definition: Remote sensing data collected from **satellites**, aircraft, or drones
- Types of Data:
  - Optical (RGB, infrared)
  - Radar (e.g. SAR, InSAR)
  - LiDAR (laser-based, 3D)
  - Climate & Environmental (e.g., temperature, rainfall from reanalysis)
- Global, Repeated, Non-Invasive
- Used to monitor land, oceans, atmosphere, and human activity
- Enables tracking of changes over time — vegetation, land use, water, climate

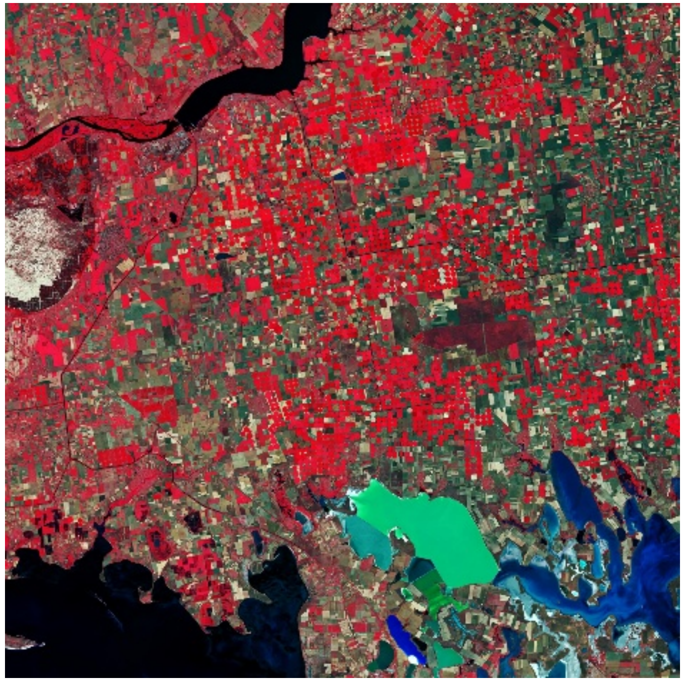




Input Datasets: HLS multispectral imagery



- Harmonized Landsat Sentinel
- Combined satellite data from NASA Landsat and ESA Sentinel-2
- 15 imaging bands: includes natural color, infrared, and thermal
- Global coverage, 2015–2024
- 30m resolution; 2–3 day revisit time
- Tiles: 3660×3660 pixels (~110×110 km^2)
- Ideal for monitoring vegetation health and seasonal land changes over time

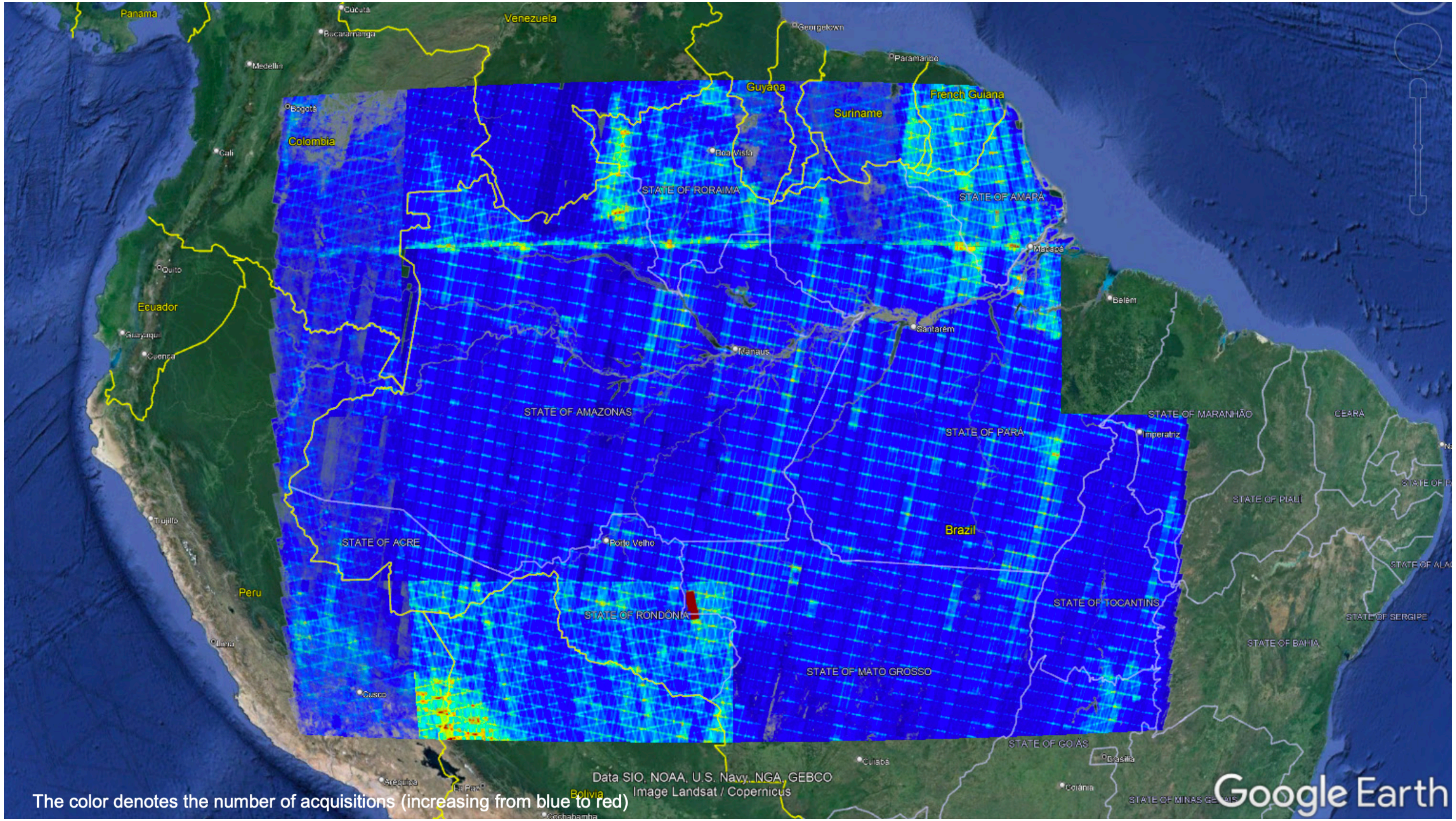


Input Datasets: TanDEM-X Interferometric Coherence

- Radar-based elevation and surface texture
- 20m resolution InSAR data for Amazon region (2018–2023)
- Penetrates clouds — works in tropical regions
- Measures surface roughness, forest structure, canopy gaps
- Complements optical and LiDAR for biomass modeling

Processed around 20k TanDEM-X InSAR acquisitions in 20m resolution between 2018 and 2023 for Amazon test region

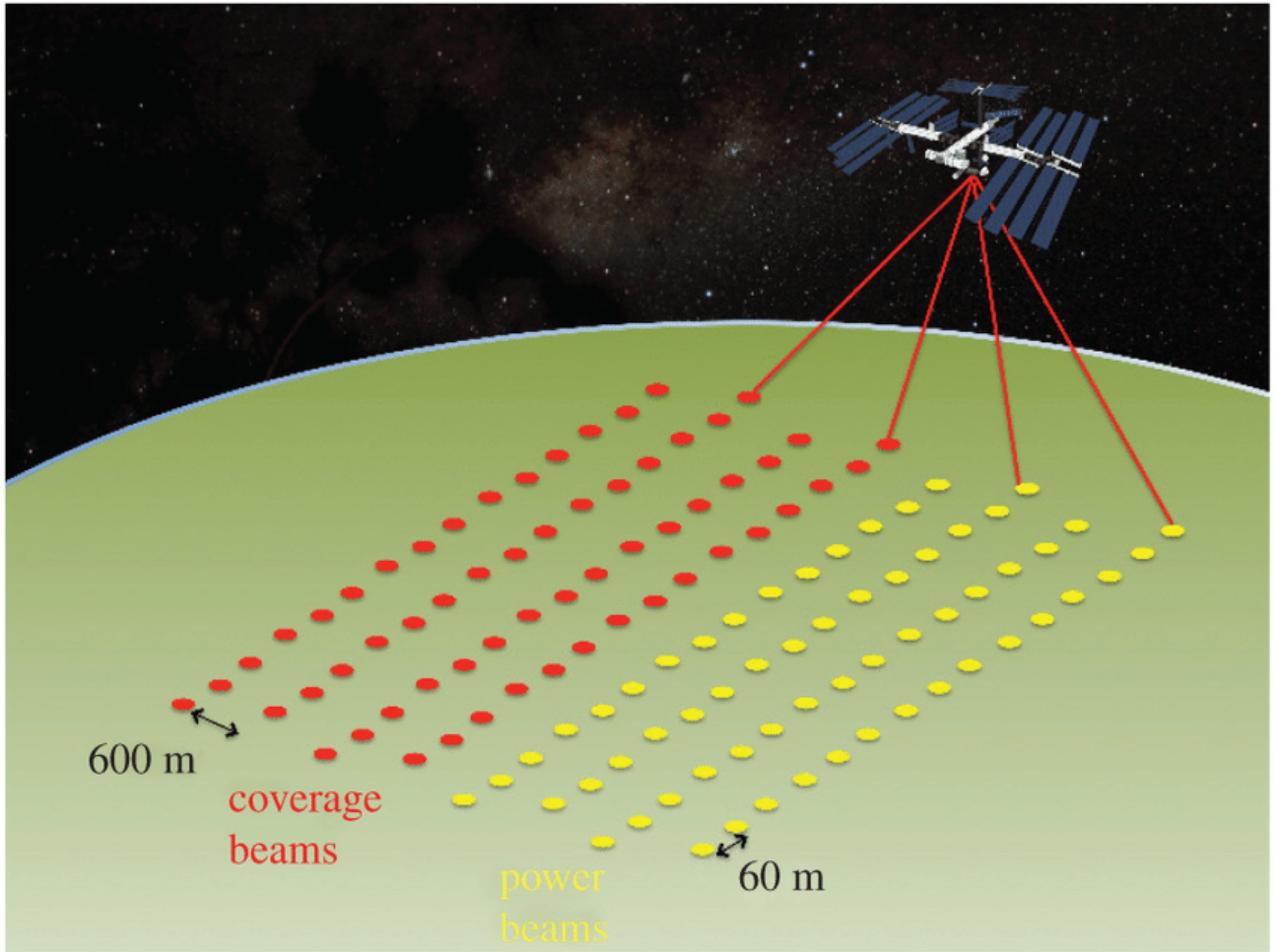
Input Datasets: TanDEM-X Interferometric Coherence



TanDEM-X InSAR coherence over the Amazon forest

Input Datasets: GEDI Lidar Data

ECOSYSTEM LIDAR

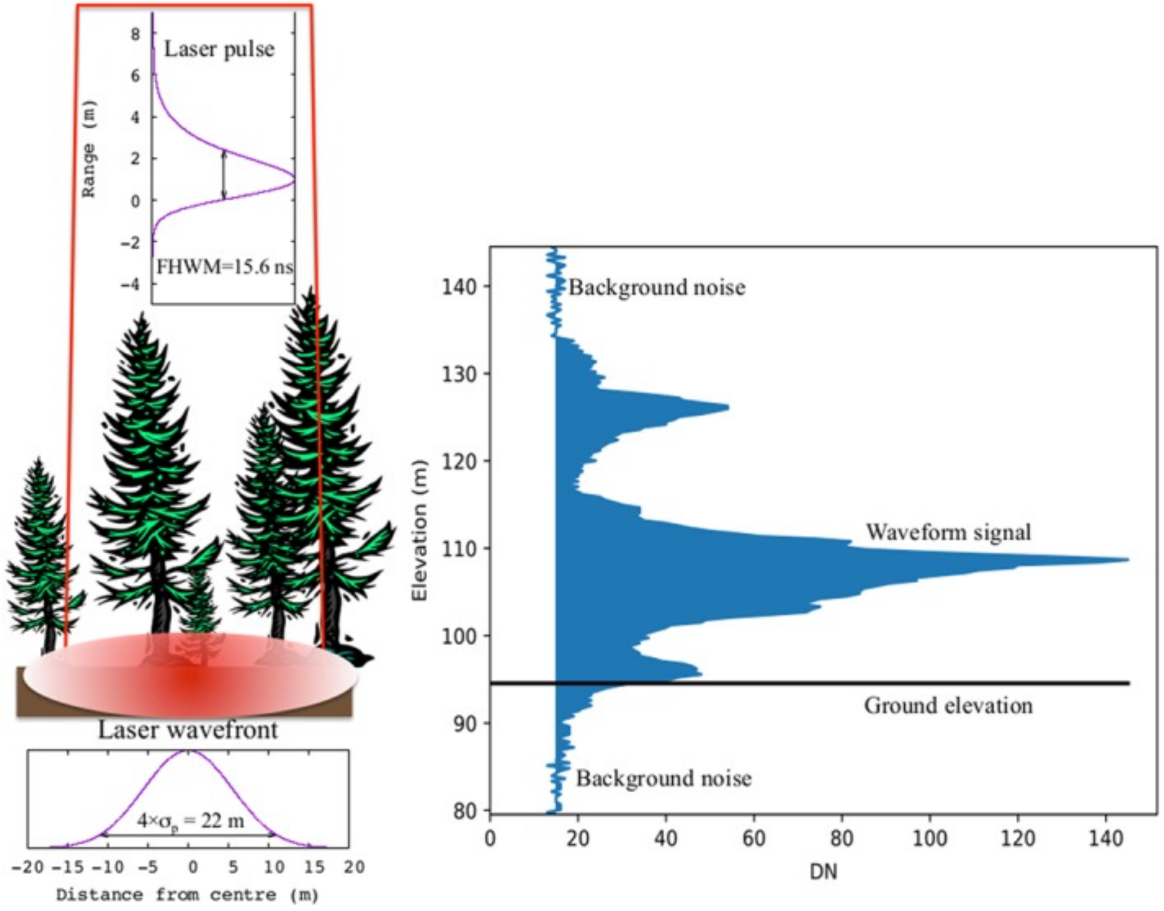


- Laser-based 3D measurements of Earth's forests, launched on the International Space Station (ISS)
- Captures full vertical structure of vegetation: canopy height, layering, and sub-canopy gaps
- Provides high-quality 3D profiles at 25-meter footprints, spaced ~60m apart
- Especially valuable for estimating forest biomass and carbon storage

Hancock, Steven & Mcgrath, Ciara & Lowe, Christopher & Davenport, Ian & Woodhouse, Iain. (2021). Requirements for a global lidar system: spaceborne lidar with wall-to-wall coverage. Royal Society Open Science. 8. 10.1098/rsos.211166.

HiF

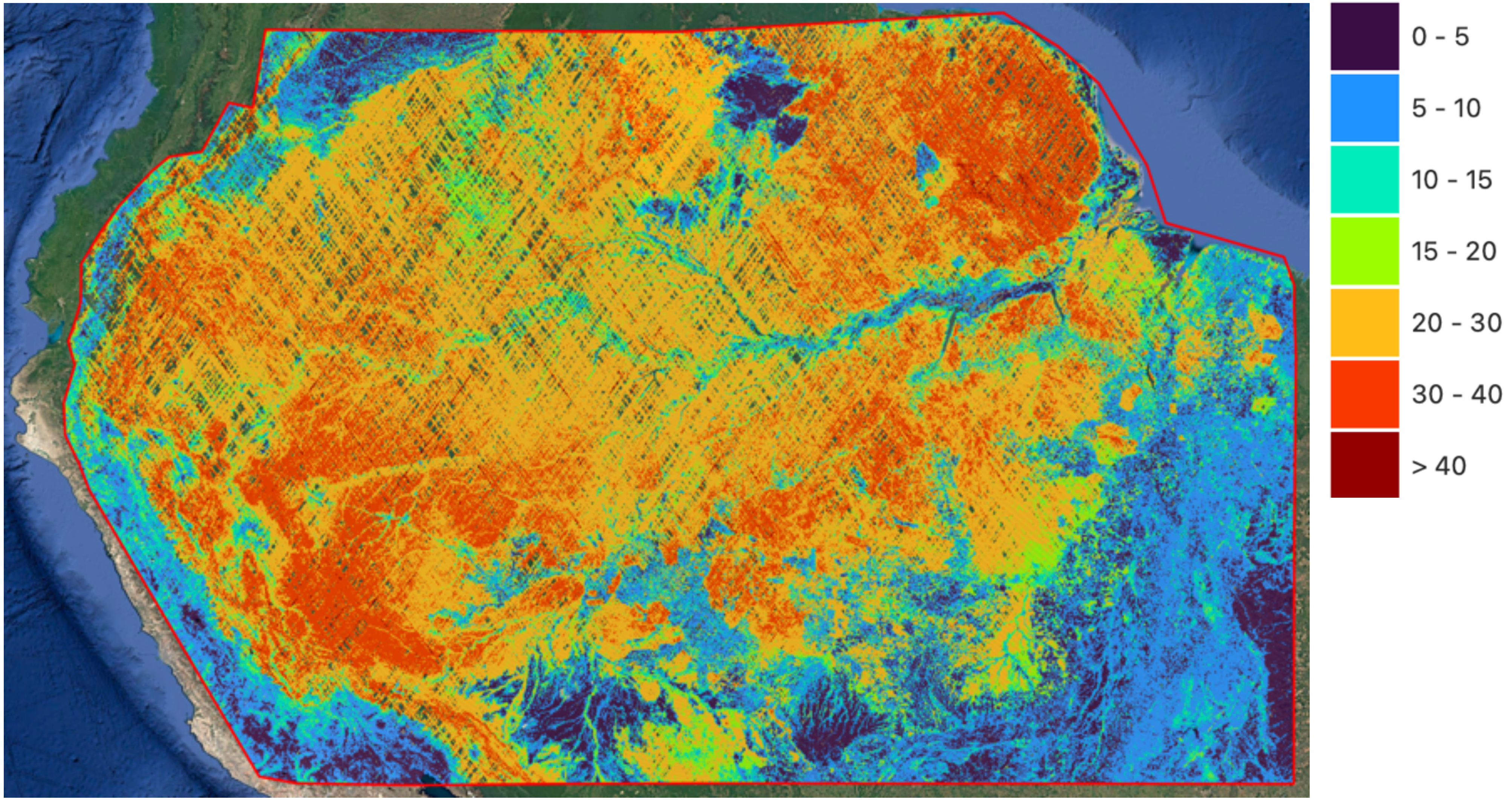
Input Datasets: GEDI Lidar Data



- Full-waveform LIDAR: Captures complete vertical energy return profile → Not just height, but how vegetation is distributed from ground to canopy
- Canopy cover, vertical distribution
- Critical for biomass and carbon stock estimation at tree-level precision

https://doi.org/10.1029/2018ea000506

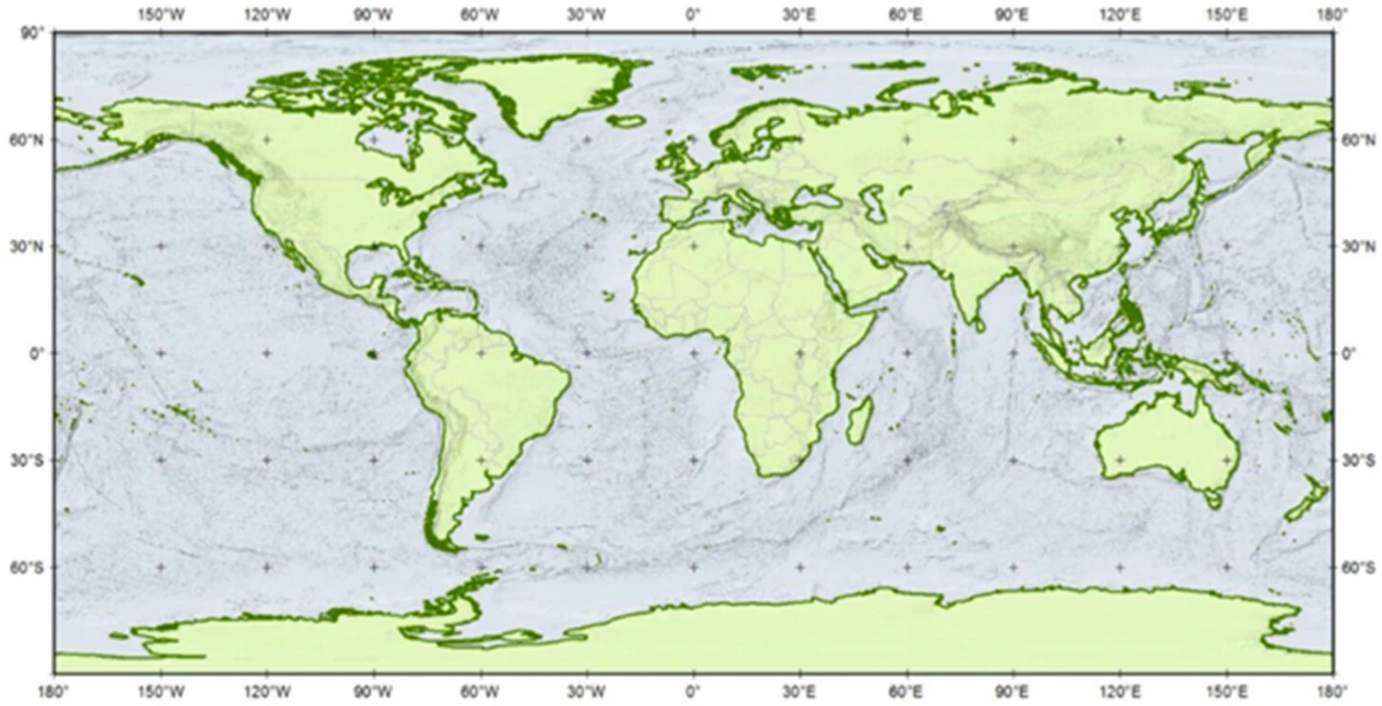
Input Datasets: GEDI Lidar Data



Canopy Height 2019-2023 (206,063,969 footprints)

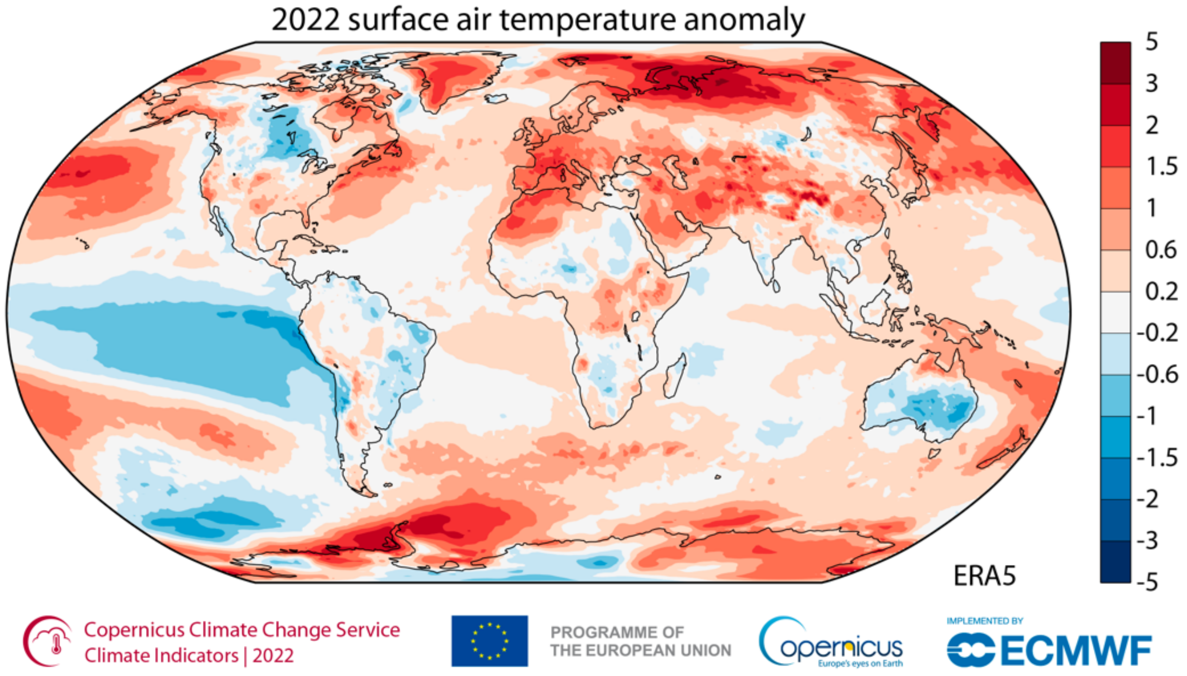
Input Datasets: GLO-30

- High-resolution digital elevation model (DEM) at 30-meter resolution
- Covers the entire globe
- Derived from optical stereo satellite data (from Copernicus program)
- Provides ground elevation: valleys, hills, mountain ranges
- Essential for understanding topographic variation, water flow, and vegetation patterns



Input Datasets: ERA5

- Global dataset of hourly to yearly climate variables (temperature, rainfall, wind, etc.)
- Produced by ECMWF, covering 1979 to present
- Resolution: ~30 km spatial, hourly temporal
- Combines satellite observations and weather models
- Provides consistent, gap-free climate data over space and time
- Helps understand how climate influences vegetation, soil, and carbon fluxes

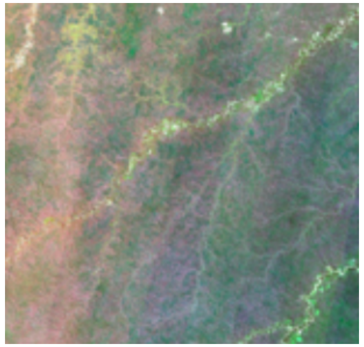


C3S/ECMWF

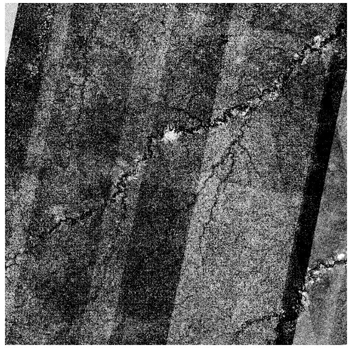
Capturing Our Complex Planet through Multimodal Earth Observation

The world

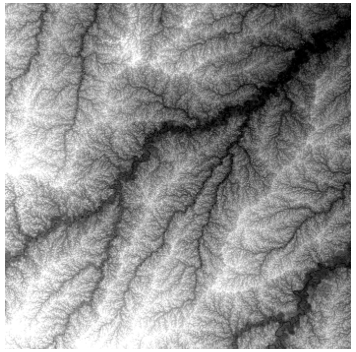
Multispectral
(RGB/ Infrared/ Thermal)



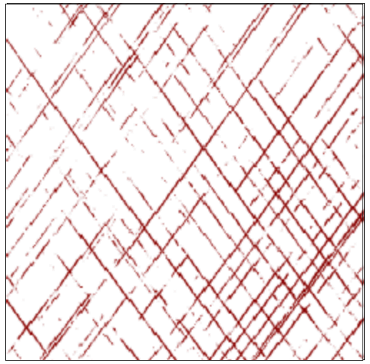
Radar



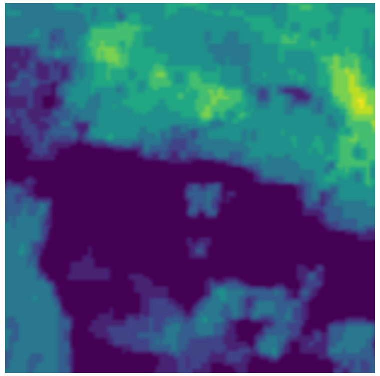
Digital Elevation Maps
(DEM)



Space/Air borne Lidar



Canopy Height



Climate Reanalysis data

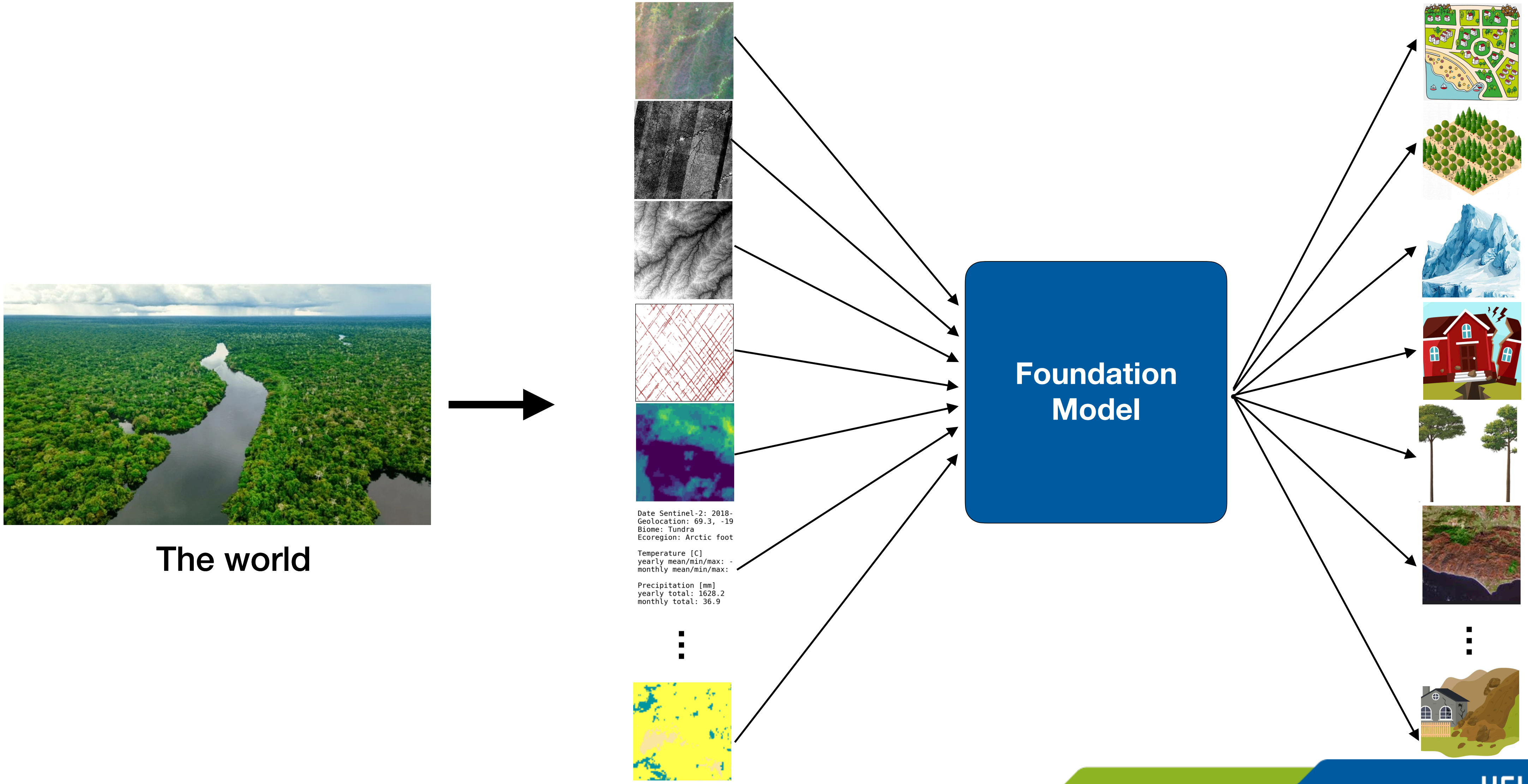
```
Date Sentinel-2: 2018-
Geolocation: 69.3, -19
Biome: Tundra
Ecoregion: Arctic foot

Temperature [C]
yearly mean/min/max: -
monthly mean/min/max:

Precipitation [mm]
yearly total: 1628.2
monthly total: 36.9
```

World cover

more

...

Capturing Our Complex Planet through Multimodal Earth Observation



Capturing Our Complex Planet through Multimodal Earth Observation

Goal: Develop a large-scale, pre-trained foundation model tailored for multi-modal remote sensing data to enable accurate and efficient performance across diverse geospatial tasks.

Multi-modal integration: Handle optical, radar, LiDAR, multi-spectral, multi-resolution, multi-temporal, 2D, and 3D data

Support **multiple spatial resolutions** (e.g., 20m, 25m, 30m, 9km) and **multi-temporal data**

Integrate **rich metadata** (e.g., location, time, climate) and handle **sparse modalities** like LiDAR

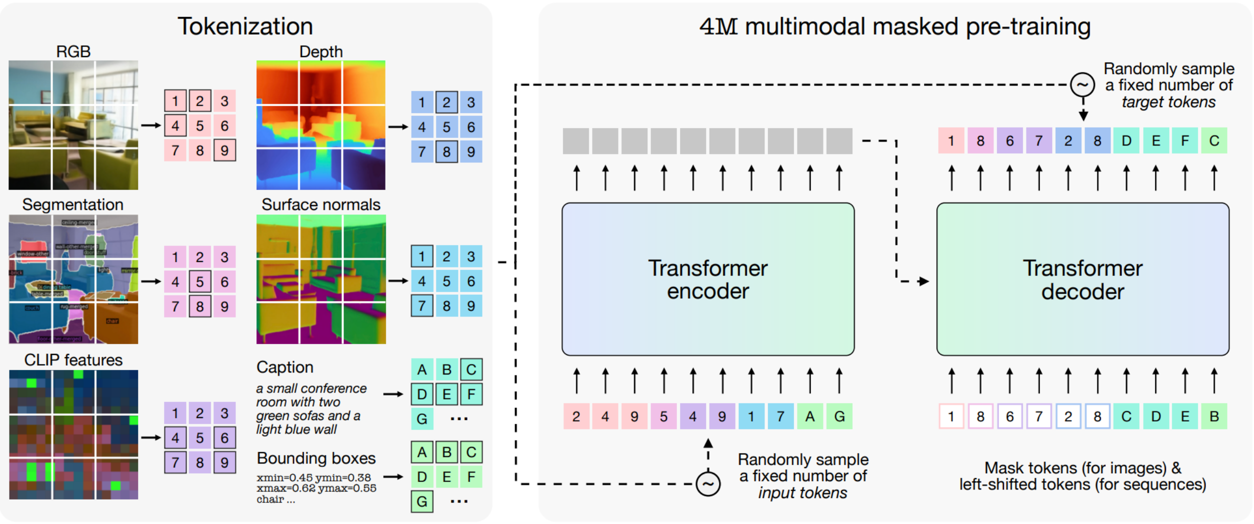
Efficient fine-tuning: Adapt to specific applications with minimal additional training

Versatile task support: Achieve strong performance in classification, segmentation, object

3D-ABC FM Backbone

Evaluated various models types

- ViT (e.g., Prithvi), SWINv2 transformer, DOFA
- New paradigm: Intro of Tokenization in Multimodal Processing (e.g., SOTA)
- Decided on **4M: Massively Multimodal Masked Modeling** as 3D-ABC backbone



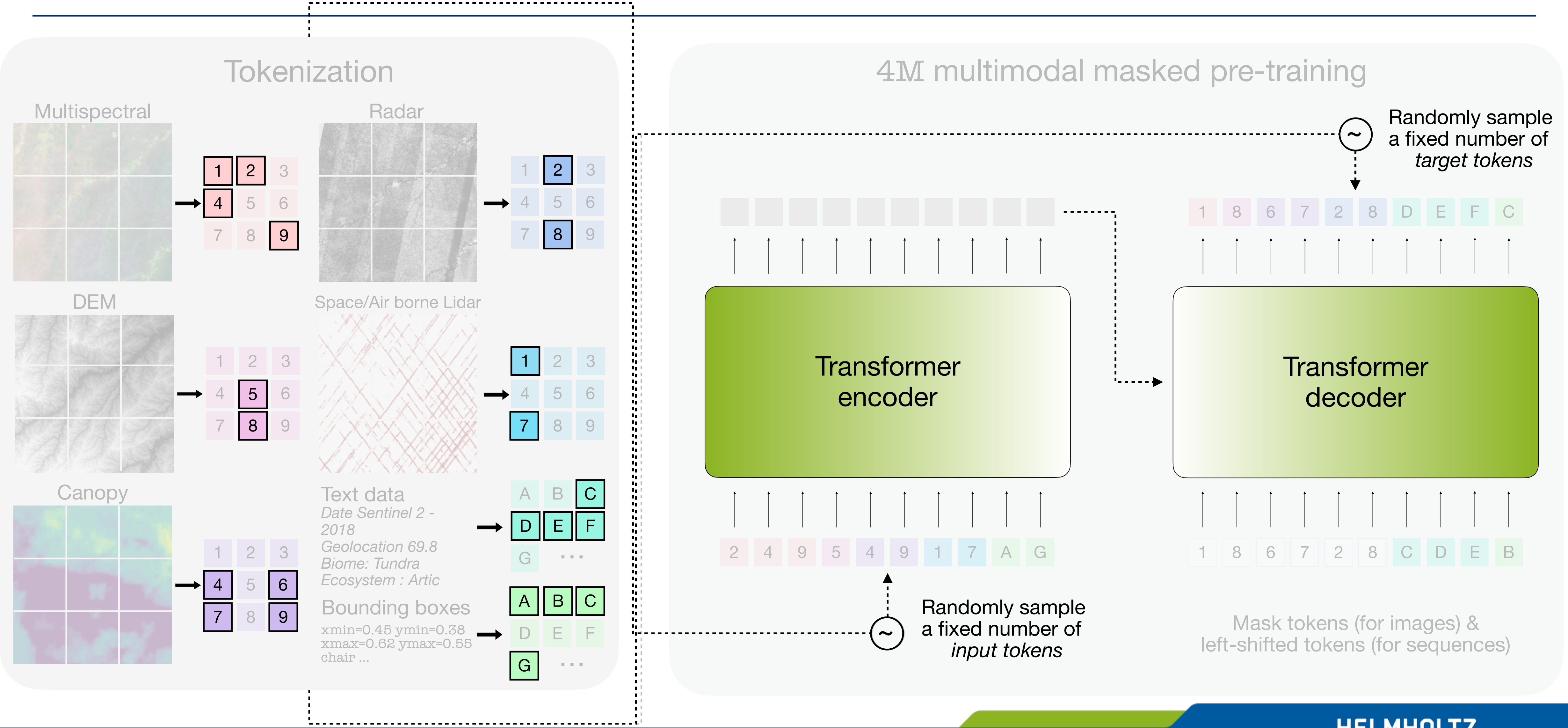
Mizrahi et al. (2023): 4M: Massively multimodal masked modeling (Advances in Neural Information Processing Systems)

4M Foundation Model:

- Tokenization can be a solution to unify data from all modalities in a similar latent space
- FM serves to “Generate” input tokens into output tokens
- Self-supervised learning directly masks tokens rather than raw data, enabling token interaction across the modalities
- Model designed originally for vision language tasks, but we can build tokenizers for geospatial data/metadata/…
- Tokens from different modalities can interact with each other during pretrain
- Input tokens and output tokens are both randomly selected, maximizing generalization capacity

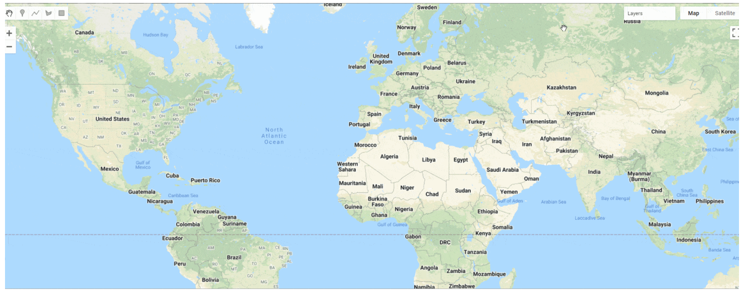
4M: Massively Multimodal Masked Modeling

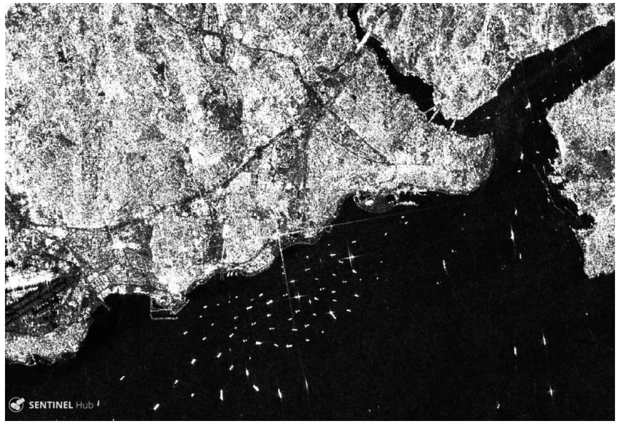
Unifying modalities through tokenization
Scalable cross-modal learning through masking

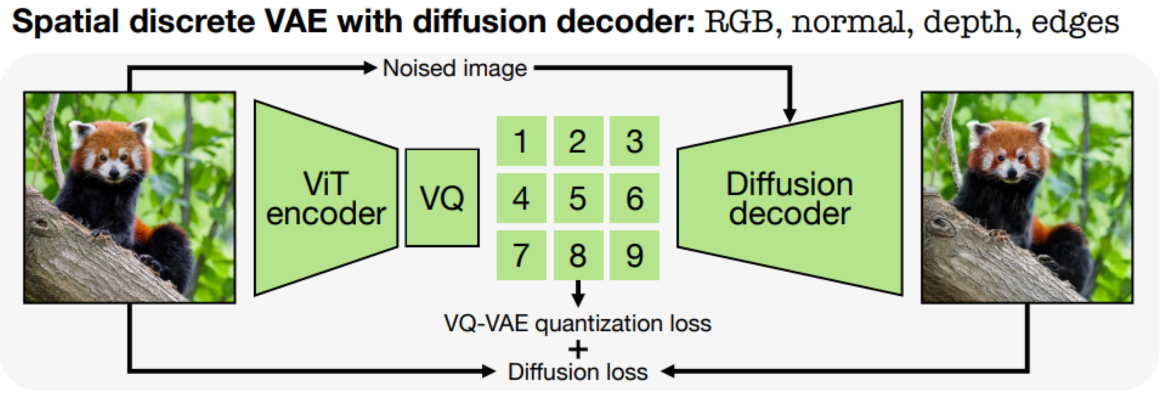


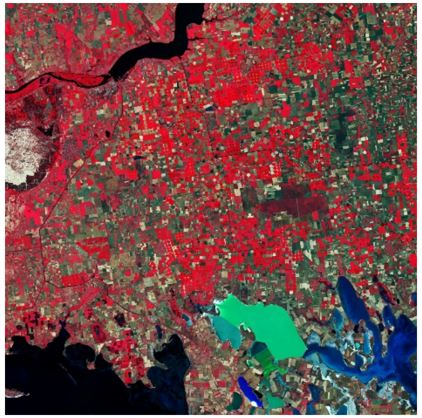
Modified from David Mizrahi et al., *4M: Massively Multimodal Masked Modeling*, NeurIPS 2023.

Tokenization for Different Modalities







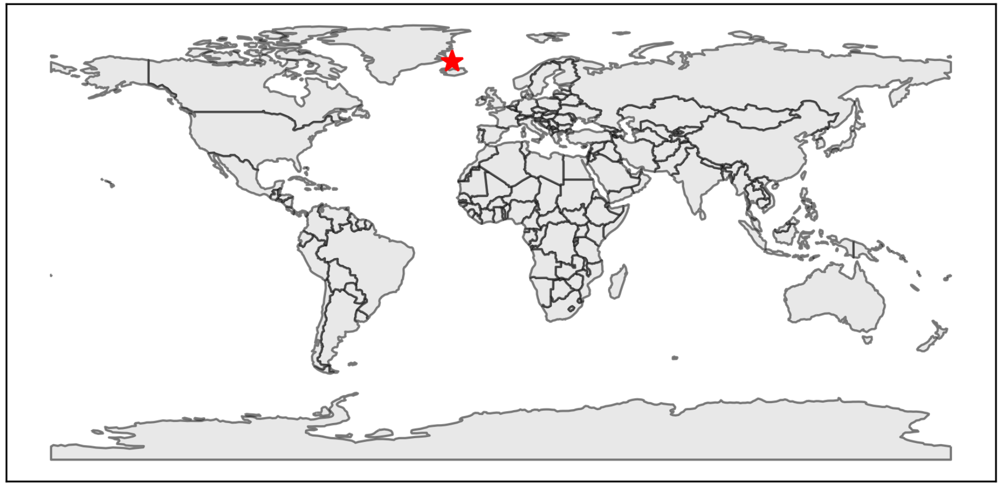


We use Diffusion VAE to encode the complex spatial data.
(4 Modalities)

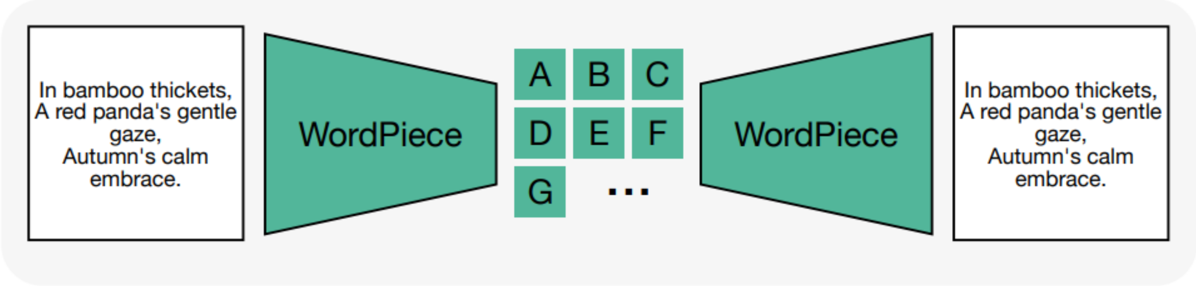
Tokenization for Different Modalities

- Input: Multiband images (e.g., 15-band HLS: RGB, NIR, SWIR, TIR, etc.)
- Encoder: Vision Transformer (ViT) processes the spatial + spectral context
- Latent Representation: Discrete tokens using Vector Quantization
- Decoder: Diffusion model reconstructs images from token sequence
- Training Objective:
  - Quantization loss (VQ-VAE): Forces latent tokens to match learned codebook entries
  - Diffusion loss: Trains model to denoise and recover full image from noisy latents
- Enables unified, spatially-aware tokens across multi-spectral data

Tokenization for Different Modalities



```
Date Sentinel-2: 2018-10-06
Geolocation: 69.3, -19.6
Biome: Tundra
Ecoregion: Arctic foothills tundra

Temperature [C]
yearly mean/min/max: -6.6/-42.0/24.4
monthly mean/min/max: -4.6/-28.2/11.9

Precipitation [mm]
yearly total: 1628.2
monthly total: 36.9
```



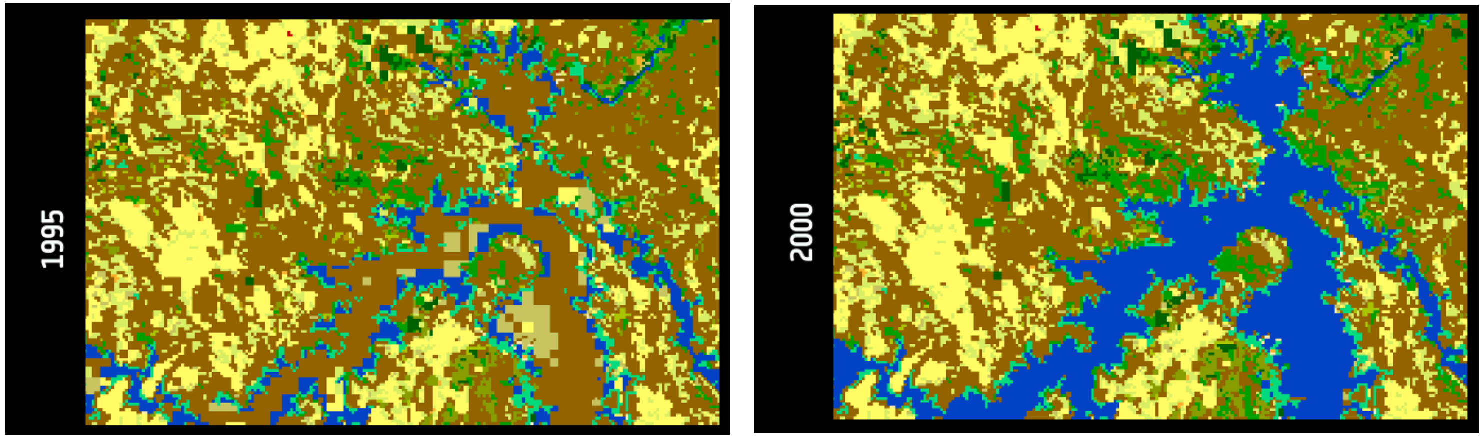
We use WordPiece Tokenizer to encode the ERA5 MetaData (1 Modality)

Tokenization for Different Modalities

- ERA5 metadata: Structured variables like temperature, precipitation, humidity, wind, time, location
- Preprocessing: Convert numerical and categorical metadata into token sequences
- WordPiece tokenizer:
- Originally used in NLP (e.g., BERT)
- Breaks input into sub-word units based on frequency
- Enables:
  - Handling of structured metadata as text-like input
  - Shared embedding space with other modalities (e.g., image tokens)
- Learned jointly with other tokenizers during pretraining
- Helps model understand climate context across space and time

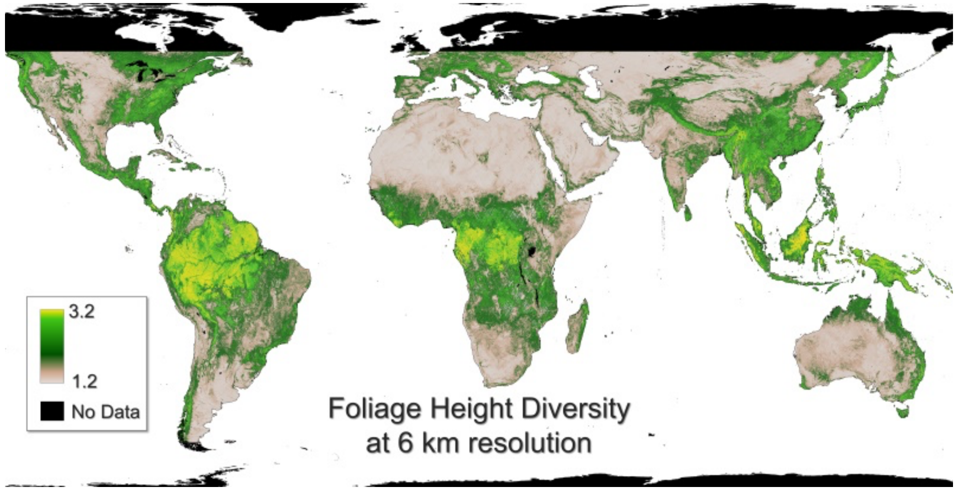
Downstream Task: Land Cover and Change

- Objective: Classify land cover types (e.g., forest, urban, agriculture, water)
- Change detection: Identify and quantify changes over time (e.g., deforestation, urban expansion)
- Input: Multitemporal satellite imagery (HLS), radar (TanDEM-X), climate (ERA5)
- Model: Finetune foundation model to detect subtle land surface changes
- Output:
  - Land cover maps for multiple years
  - Temporal trends and transition matrices
- Applications:
  - Environmental monitoring
  - Urban planning
  - Biodiversity & conservation
  - Emissions tracking (e.g., forest loss)



ESA CCI High Resolution Land Cover Change maps

Downstream Task: Vegetation Height

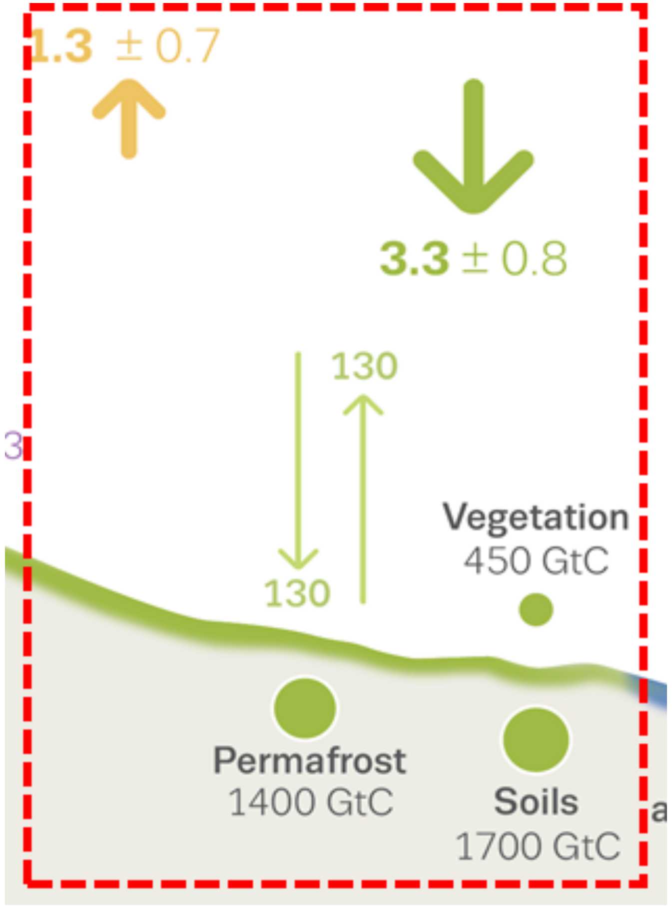


https://doi.org/10.3334/ORNLDAAC/2339

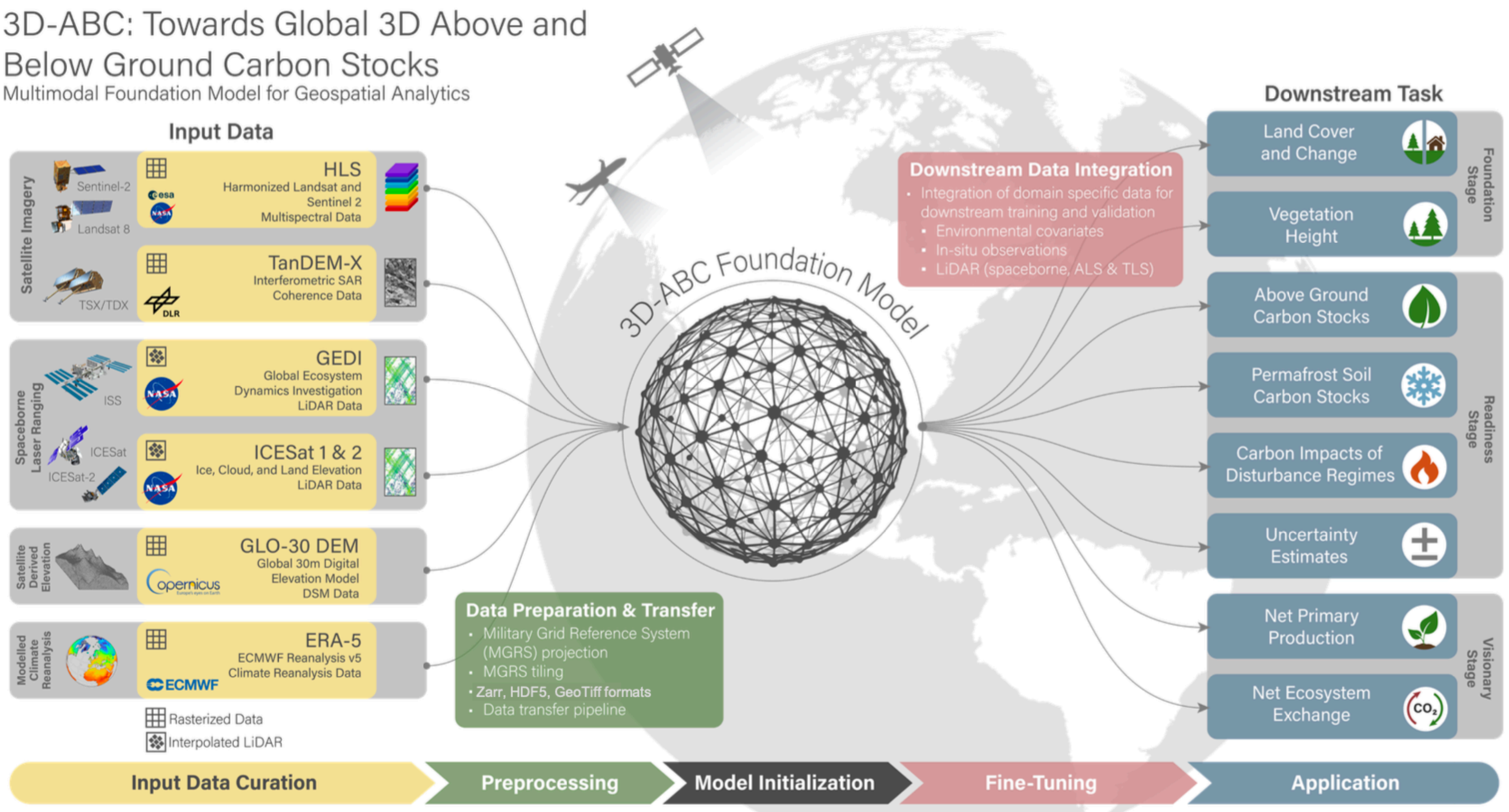
- Goal: Estimate vegetation height (especially tree height) across large areas
- Why it matters:
  - Critical for biomass and carbon stock estimation
  - Helps distinguish forests, shrubs, grasslands
- Inputs:
  - GEDI LiDAR (sparse but accurate vertical structure)
  - TanDEM-X radar (structure and surface roughness)
  - Multispectral imagery (HLS, texture + spectral cues)
- Model: Finetune foundation model to learn 3D canopy structure from sparse and dense inputs
- Output:
  - Wall-to-wall maps of vegetation height
  - Fine spatial resolution (e.g., 30m)
- Applications:
  - Forest monitoring and management
  - Ecosystem classification

Downstream Task: Above Ground Carbon Stock Estimation

- Goal: Estimate the amount of carbon stored in vegetation above the soil
- Why it matters:
  - Key to understanding the global carbon budget
  - Essential for climate models, REDD+, and land-based mitigation
- Inputs:
  - GEDI LiDAR (anchor for tree structure and height)
  - Vegetation height & land cover maps
  - Radar (TanDEM-X), optical (HLS), climate (ERA5)
- Model: Finetune the foundation model using field inventory plots and biomass reference maps
- Output:
  - Continuous maps of aboveground biomass (converted to carbon)
  - High-resolution, time-resolved estimates
- Applications:
  - Carbon accounting and emissions reporting
  - Forest conservation and climate finance
  - Land degradation assessment



3D-ABC FM, Input Data and Downstream Tasks



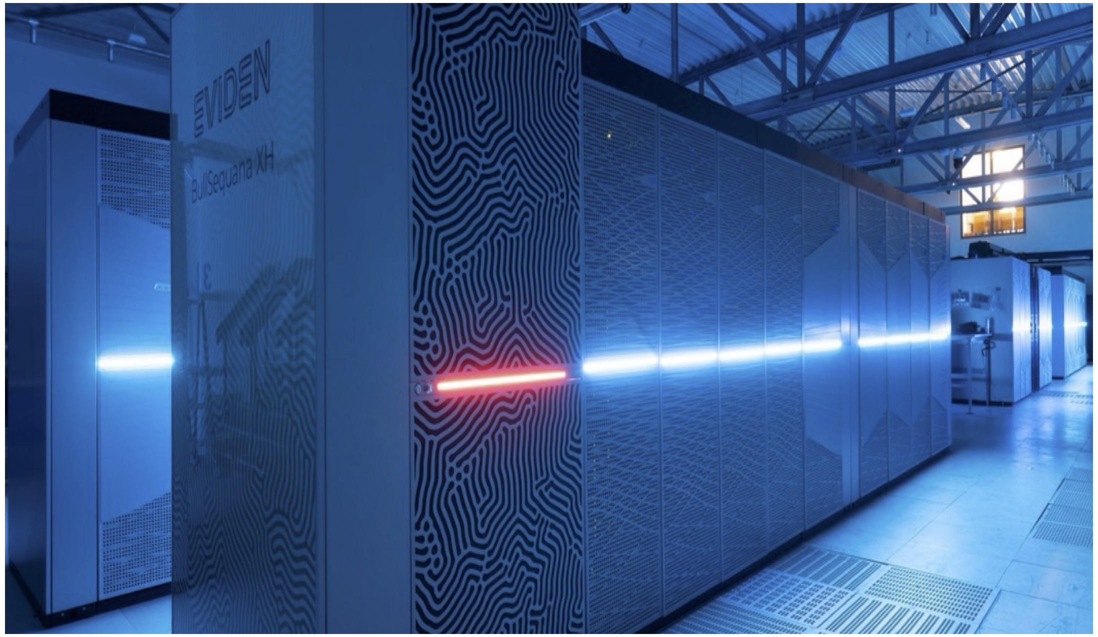
3D-ABC HPC Resources

JUPITER: The arrival of Exascale in Europe
https://www.fz-juelich.de/en/ias/jsc/jupiter

HPC status and plans

- Computing time now available on HPC systems at JSC JUWELS (>3,000 Nvidia A100 GPUs)

- JUPITER Exascale Development Instrument (JEDI): Benchmarks on different hardware accelerators: A100, H100, GH200, AMD MI250, IPU GC200, WestAI H100 and JEDI GH200



JEDI: The first module of the exascale supercomputer JUPITER

Summary

- **Status**

4M FM selected, input data for Amazon ready, in prep for Alaska-NW Canada region, data pipeline ready, finetuning data for first downstream tasks ready

- **Scope**

3D-ABC FM leverages huge remote sensing, field, and modeling datasets to capture carbon stock dynamics at enhanced spatial resolution

Finetuning towards multiple downstream tasks will deliver thematical insights with a broad range of applications and uses

- **Relevance**

Detailed quantification of global terrestrial carbon stocks and dynamics for science, society, and policy making