Geographical and Geospatial Sciences Committee (GGSC) Spring 2025 Meeting on June 11th:
Evolving Geodigital Data: Opportunities, Challenges & Disruptions

Evolving Role of AI in Earth Science: *Foundation Models and Accelerated Knowledge Discovery*

Rahul Ramachandran*, PhD
Senior Research Scientist, NASA MSFC

SMD by the Numbers

10 M
FY23 Unique users

140 PB
FY23 Total volume

50 PB/year
Ingest rate

>500 PB
Projected FY29 archive volume

1.5 B
Total files

NASA needs to adopt more efficient Data Science practices.

AI/ML 5+1 Strategy

The OCSDO will provide the AI expertise, system engineering, infrastructure, and training to facilitate the development of…

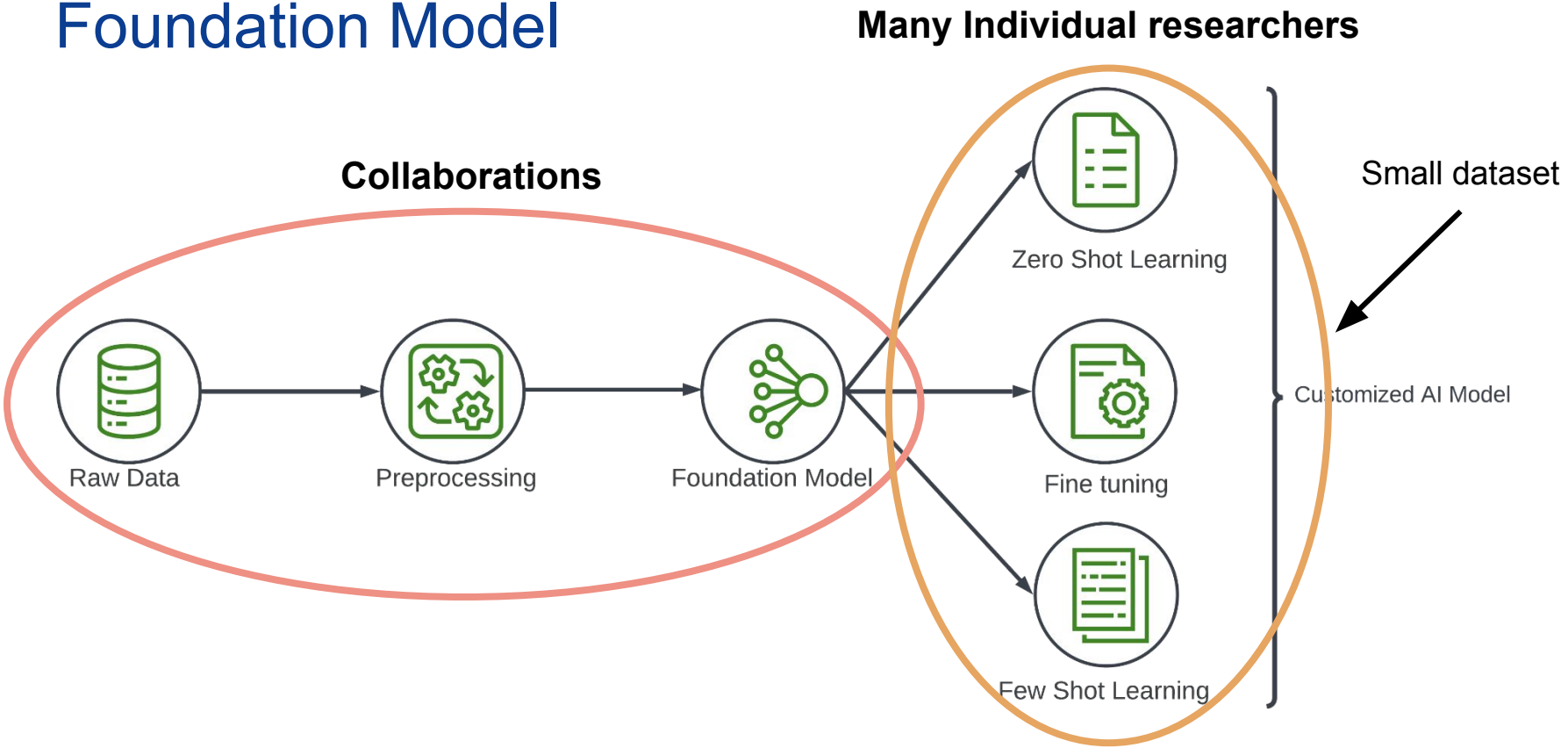
5 **Foundation models**
one for **EACH** division

…and

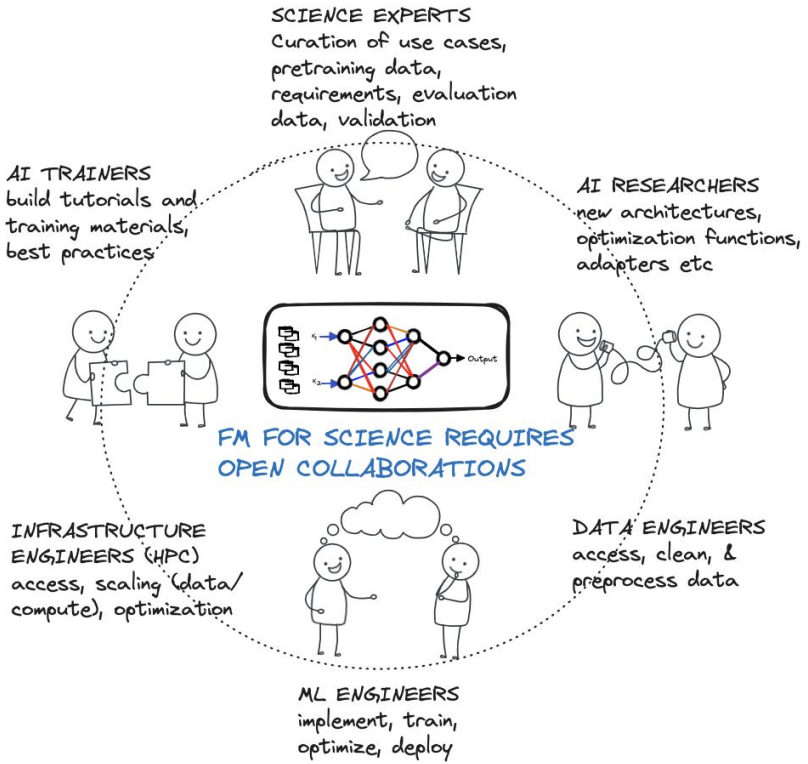
1 **Large language model**
for **ALL** divisions

AI Foundation Models: Tools

Foundation Model



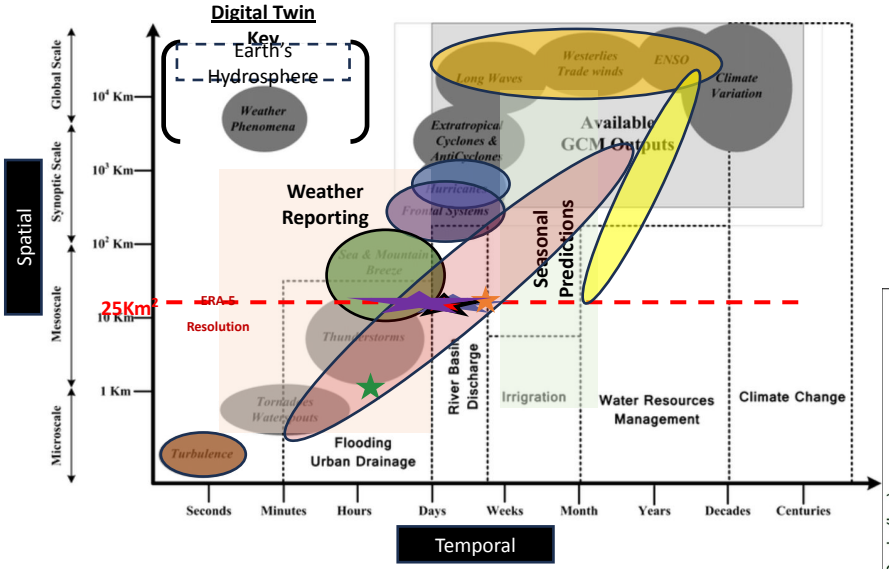
Building These FM Requires OPEN Collaborations

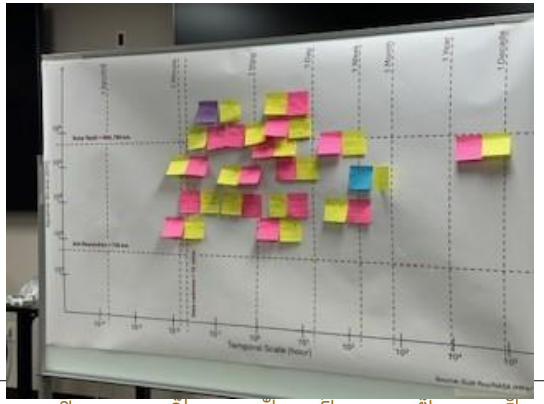


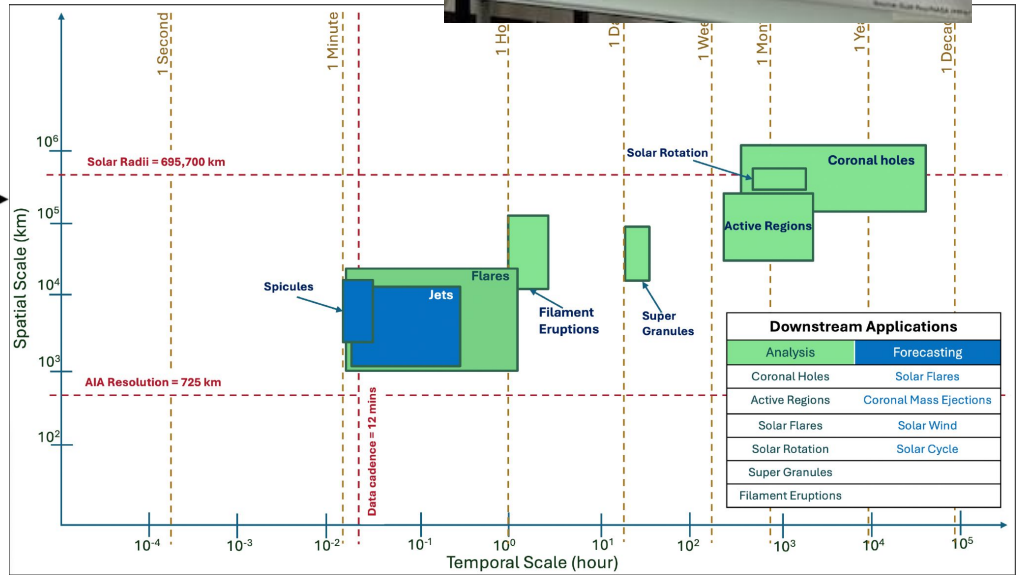
- Diverse skill sets are required to build a useful AI FM (beyond just writing a paper)
- Compute resources
- Access to data

Difficult for a single team/organization to do this well for science

Requirements Driven by Science Use Cases







1. What use cases do we plan to address?
2. What use cases do we plan not to address?
3. What (physical) processes do we expect to capture?

Image Source: Johannes Schmude, IBM Research

Prithvi-HLS (EO)-2.0: Versatile Multi-Temporal Foundation Model for Earth Observation

- Improved version based on feedback from users
- Built on masked autoencoders (MAE) with vision transformer (ViT) backbone
- Encoded temporal and location metadata as weighted sums for better context
- Pretrained using NASA's Harmonized Landsat Sentinel-2 (HLS) dataset (2015–2024)
- Updated sampling capturing global land-use diversity with 4-timestamp sequences (1–6 months apart)
- Trained on JUWELS HPC system with distributed PyTorch for scalability
- V2 300M trained on 80 GPUs (11 days); V2 600M on 240 GPUs (10 days)
- Performance validated on GEO-Bench benchmarking suite

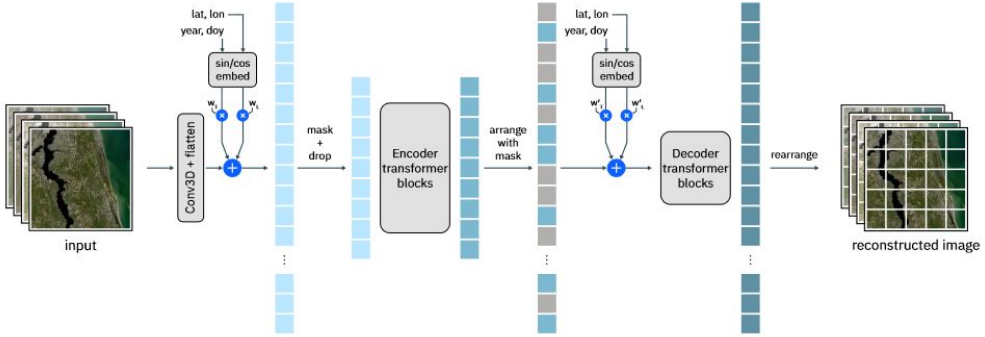


Fig. 4: Prithvi architecture and general pretraining framework.

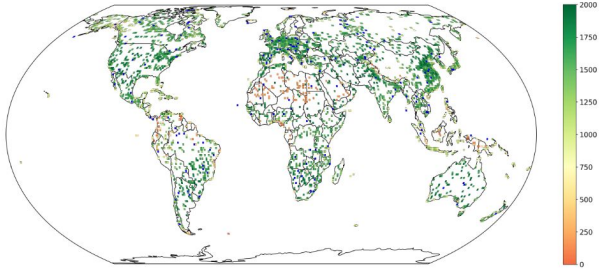


Fig. 3: Global HLS dataset distribution visualized on a tile-level. The number of training samples are color-coded in orange to green while validation tiles are visualized in blue.

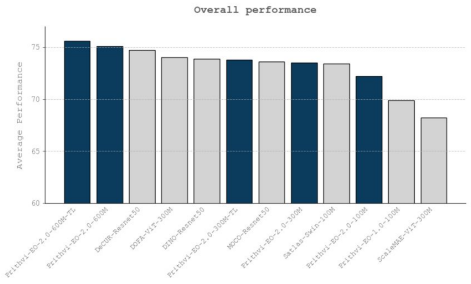
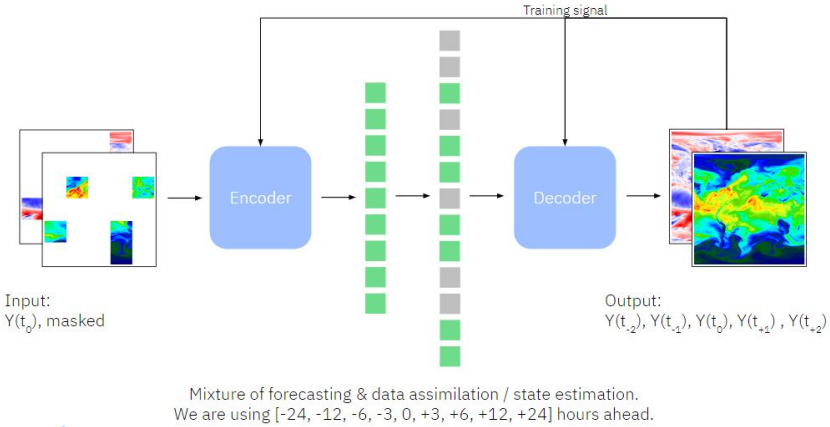


Fig. 1: Overall average performance across all GEO-Bench datasets.

Released Dec, 2024

Prithvi-WxC: Foundation Model for Weather and Climate

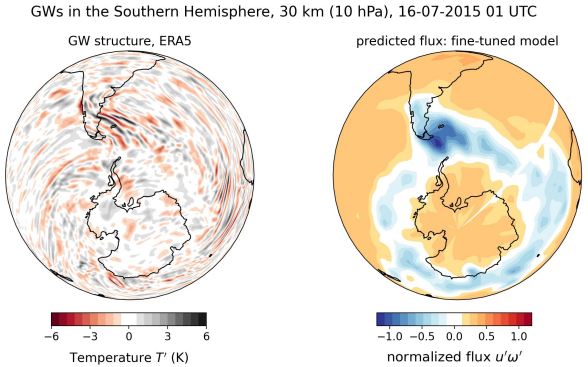
- Prithvi WxC is a 2.3 billion parameter model trained on 160 variables from MERRA-2 data.
- Jointly developed by NASA, IBM Research, ORNL, and several universities
- Pretrained for both forecasting and masked reconstruction tasks
- Capable of reconstructing atmospheric states from partial data and forecasting future states



Developed to address applications that aren't focused solely on forecasting (parameterization, downscaling, etc)

Released Sept, 2024

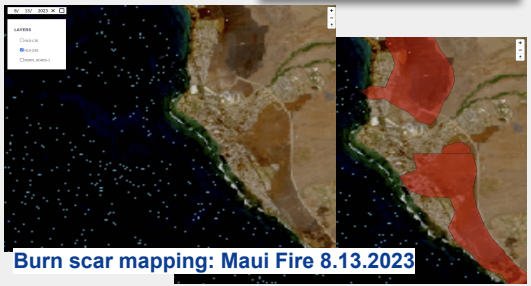
- Weather & climate Forecasting
- Downscaling
- Model output statistics/bias correction
- Multi-model ensembling
- Parameterization
- Data Assimilation
- Detection
- Tracking
- Impact Modeling



Applications

- Burn scar detection
- Flood water delineation
- Crop classification
- Cloud gap filling
- LULC (HK region)
- Eddy covariance
- Insect damage estimation
- Locust breeding ground prediction
- *Semantic segmentation of mangrove forest*

Model fine-tuned to segment the extent of floods on Sentinel-2



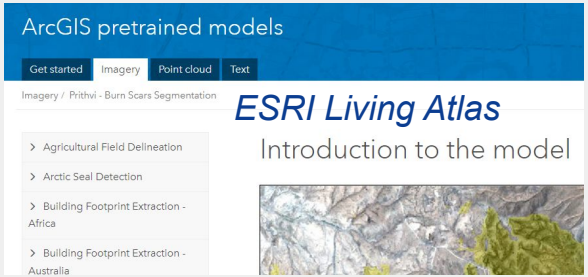
Burn scar mapping: Maui Fire 8.13.2023

Reuse

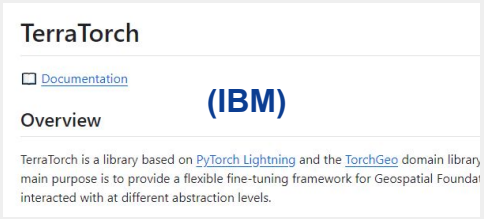
Locust breeding ground prediction
Source: Arnu Pretorius, Yusuf Ibrahim InstaDeep

200K+ Downloads (FMs and Fine Tuned)

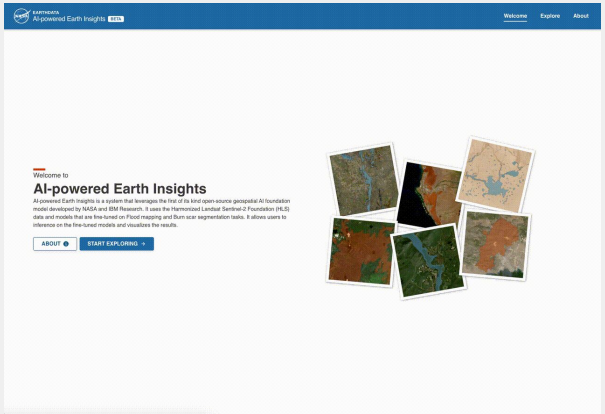
Industry Adoption



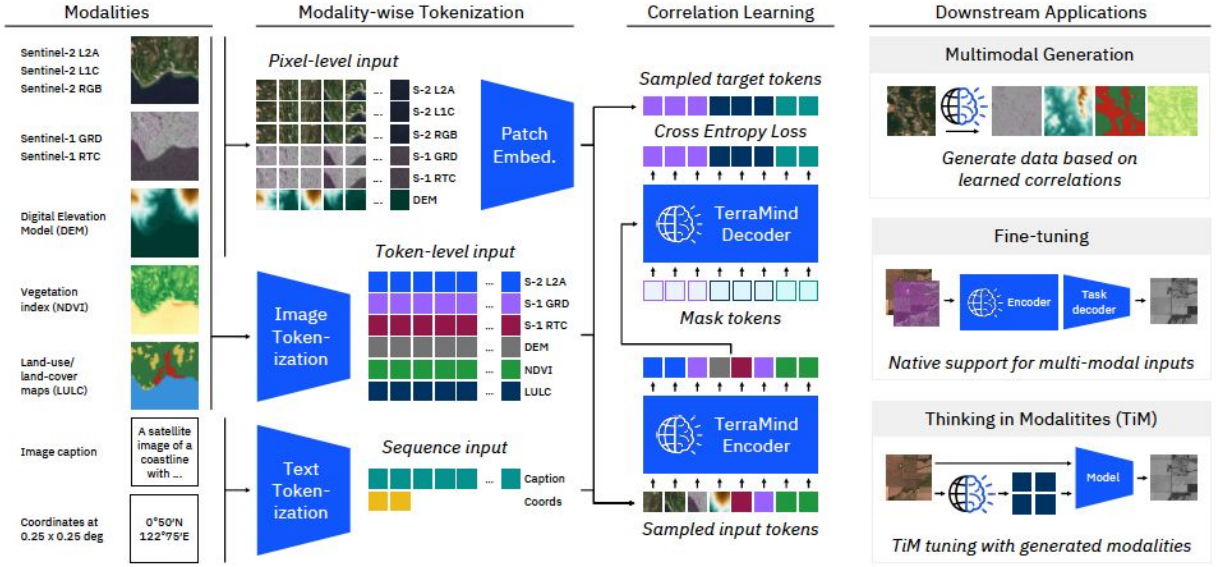
Community Developed Library



NASA Data Systems Integration



Terra Mind - Multi Modal Model



- Multimodal foundation model for Earth observation (EO)
- Collaboration between IBM Research – Europe, ETH Zurich, Jülich Supercomputing Center, ESA Phi Labs, and NASA
- Modalities – Sentinel-1, Sentinel-2, LULC, NDVI, DEM, geographic coordinates, and natural language captions
- Supports any-to-any generation of modalities, such as text-to-image generation, land-use segmentation, geolocation prediction, vegetation assessments, and caption generation
- Combining multiple modalities typically provides better performance than using any modality alone

Johannes Jakubik et al. 2025- TerraMind: Large-Scale Generative Multimodality for Earth Observation

Investigating WxC FM Through the Lens of Fundamental Scientific Principles

Objective: Evaluate Prithvi WxC model performance in the context of adherence to fundamental physical constraints, consistency to QG in midlatitudes, and consistency of flow evolution with respect to physical processes

Science Experiments / Tests

1. Conservation of mass
2. Geostrophic balance
3. Hypsometric equation
4. Thermal wind
5. Radiative parameterization
6. Convective parameterization

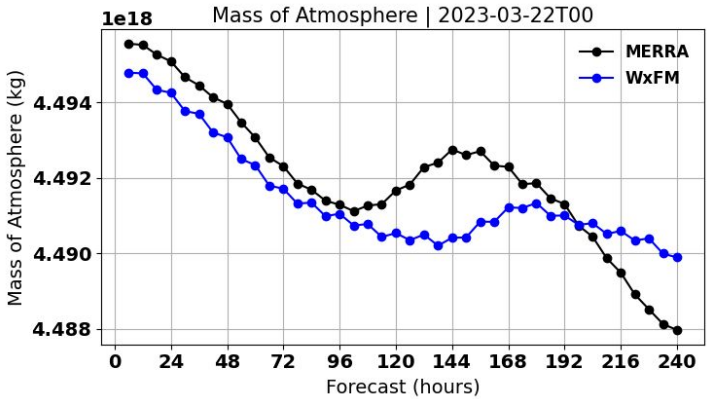


Fig. Mass of the atmosphere from Prithvi WxC model, comparison with mass calculated from MERRA for 10-days forecast.

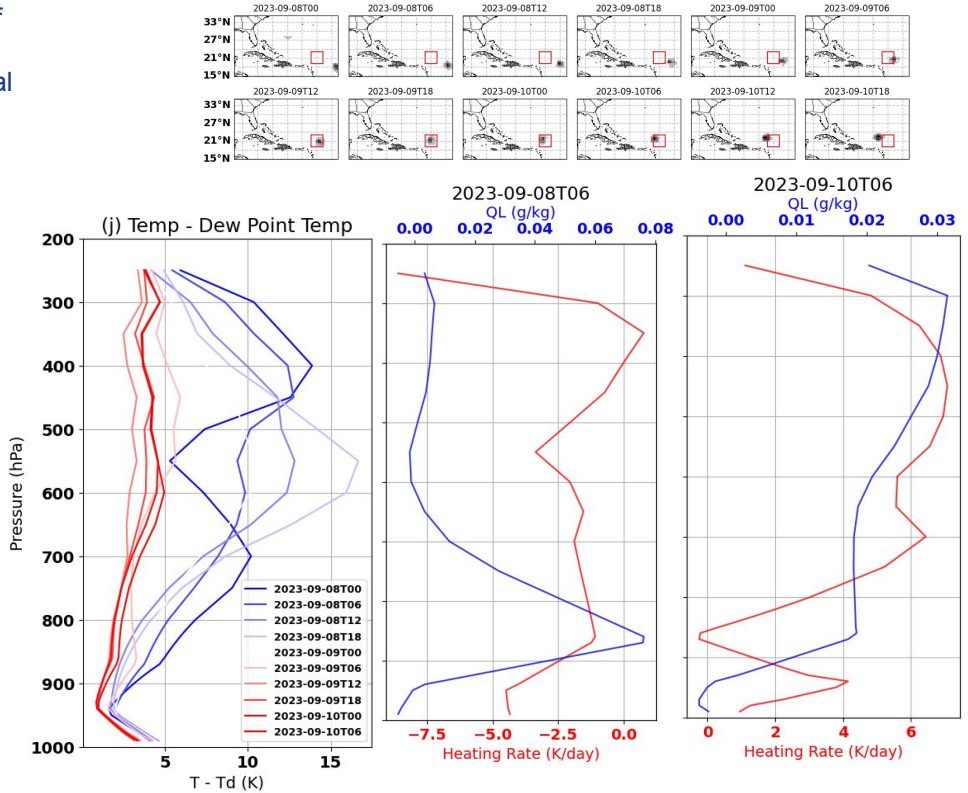


Fig. Vertical profiles showing the difference between temperature and dew point temperature across all forecast hours. The profiles in the second and third columns represent the heating rate and the cloud mass fraction that is water at 2023-09-08T06 and 2023-09-10T06.

A Primer for Assessing Foundation Models for Earth Observation

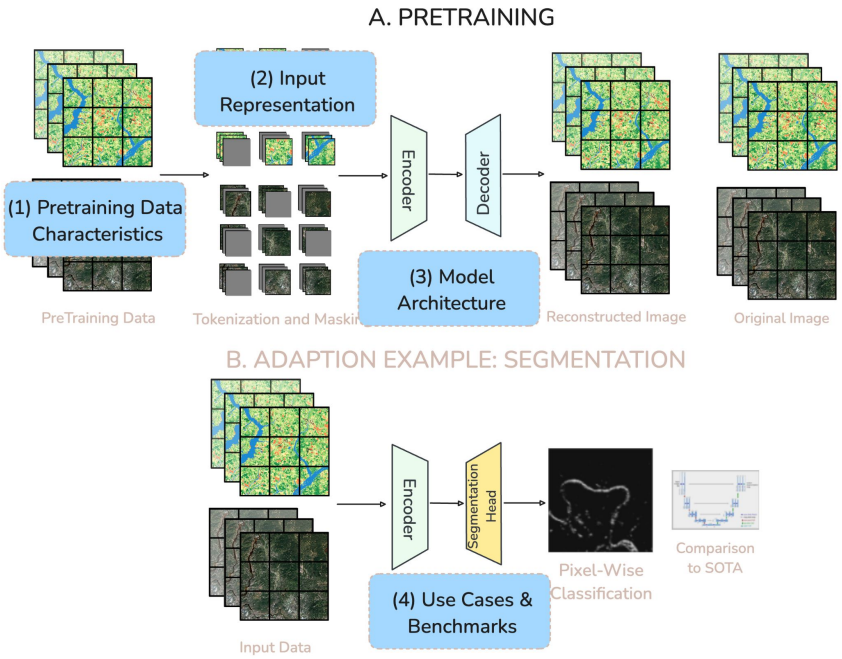
Ramachandran, R., Roy, S., Maskey, M., Szwarcman, D., & Fraccaro, P. (2025). A Primer for Assessing Foundation Models for Earth Observation. Zenodo. https://doi.org/10.5281/zenodo.15264752

A key **challenge** is selecting models suited for specific EO tasks.

- Model selection currently depends on benchmark-based evaluations of downstream tasks.
- EO tasks involve signals with complex spatial, spectral, and temporal patterns, meaning an FM cannot be judged by downstream performance metrics alone.

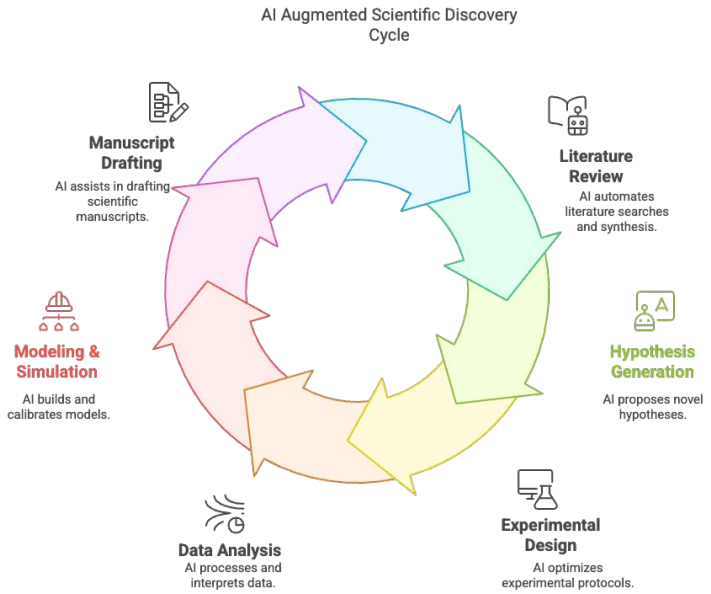
Primer introduces a structured assessment framework centered on four questions:

1. **Pretraining Data**: What are the characteristics of the data used to pretrain the FM, including preprocessing?
2. **Tokenization**: How was the pretrained data **modified and structured** before model input?
3. **Architecture**: What are the model's architectural and design choices?
4. **Evaluation**: Do the benchmarks test the model's ability to differentiate relevant and representative signals in the data?



Accelerated Knowledge Discovery: Collaborator

Accelerated Knowledge Discovery



- Change in the scientific discovery process where AI will fully integrate into the research workflow
  - Integration of **AI based on agentic systems** with contextual awareness and sophisticated decision-making
- This transformation envisions ***AI systems transitioning from static tools to dynamic collaborators***, enabling human-AI symbiotic scientific discovery.
- A central tenet of the AKD paradigm is AI's role as a "**thinking partner**," coupling human insight and intuition with AI's computational power and analytical capabilities.

Image Credit: Rahul Ramachandran

Closed Loop Scientific Workflows (CLSW)

CLSWs — automated, self-updating experimental cycles that operate within **clearly defined boundaries**

- Operate within a **well-defined scientific scope** that includes a machine-readable statement of the research question, hypotheses, success metrics, and acceptable data domains
- Utilize an declarative **inventory of approved tools and services**
- **Autonomous closed-loops** convert static workflows into adaptive workflows to **refine hypotheses and execute experiments**.
- **Human-in-the-loop** guardrails ensure scientific rigor and ethical compliance without sacrificing speed.

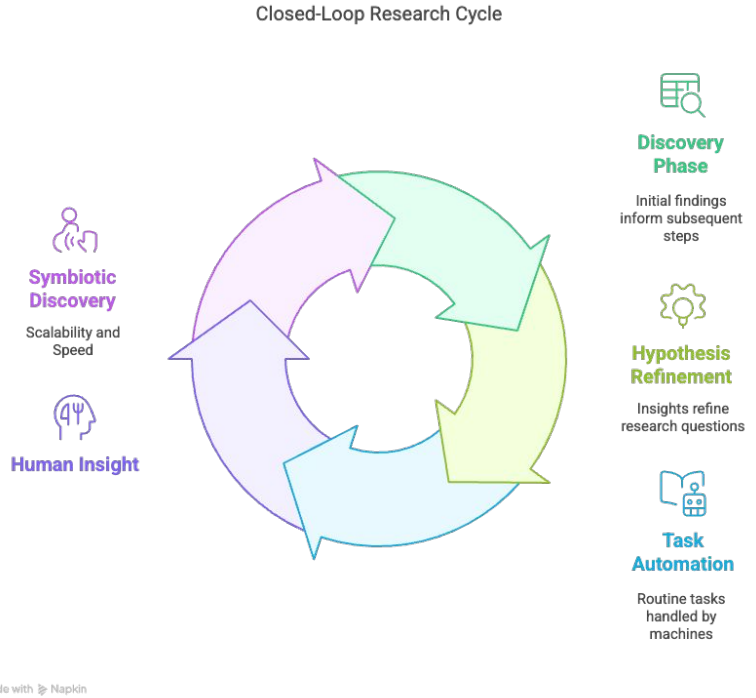
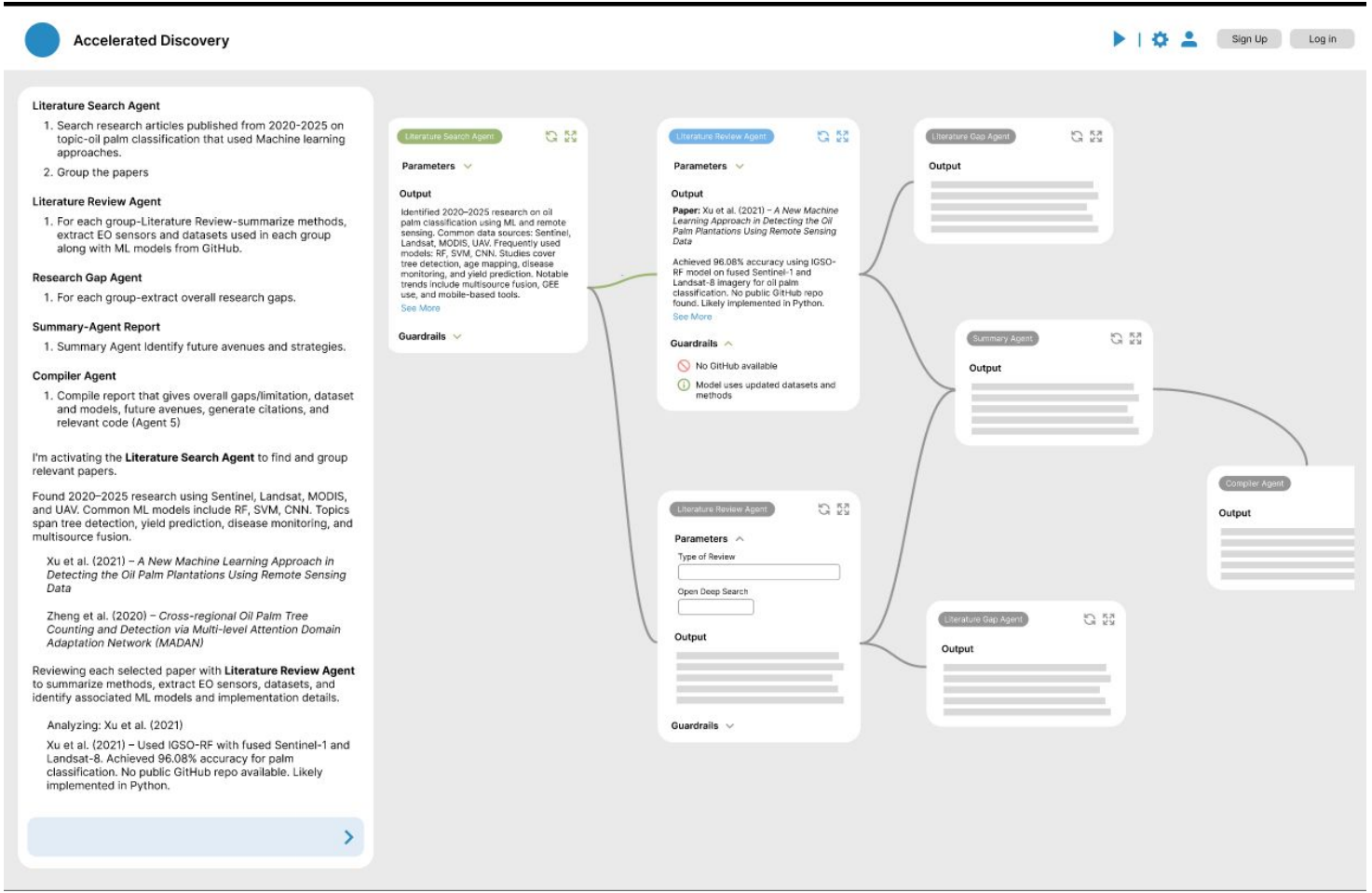
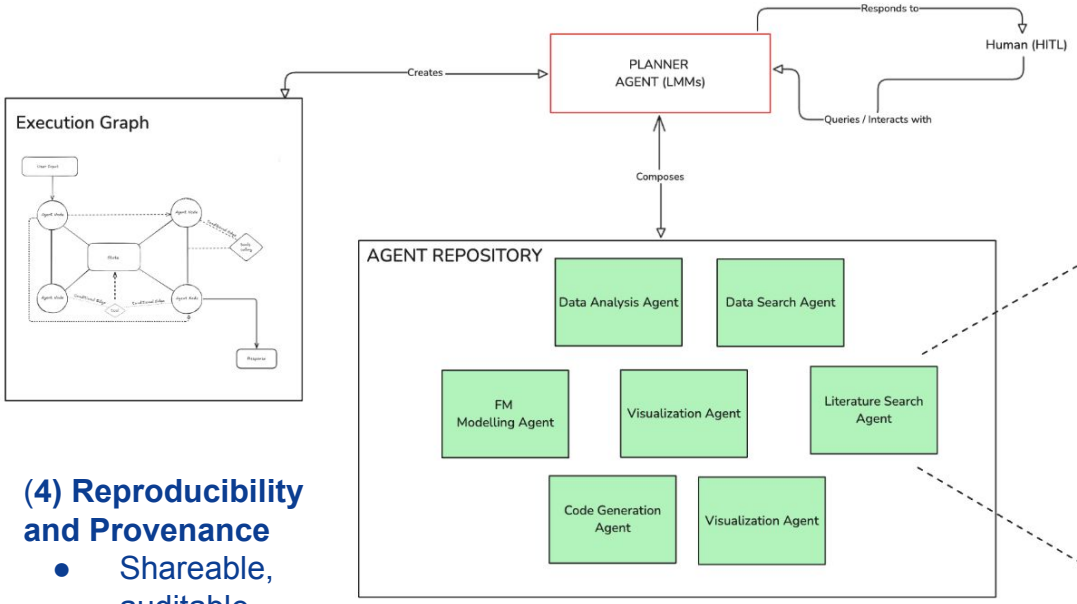


Image Credit: Accelerated Discovery Team



Multi-Agent Architecture: Trust is Foundational



(1) Alignment with Open Science Principles

- agentic workflows support open data, code and results

(2) Human Oversight and Scientific Accountability

- Scientist direct and verify
- Explainability/**groundedness**, **factuality & bias checks**

(3) Adaptable and Extensible

- Use part or full workflow
- Add new agents (and tools)

(4) Reproducibility and Provenance

- Shareable, auditable, repeatable
- Stateful and deterministic*

Link to Resources

https://science.nasa.gov/artificial-intelligence-science/

Artificial Intelligence Science Data Projects

Prithvi Geospatial Foundation Model

Pre-trained on Harmonized Landsat and Sentinel-2 (HLS) satellite data, the Prithvi Geospatial Foundation Model helps Earth science researchers conduct AI-powered studies such as flood mapping, fire-scar identification, and crop classification. The model is released openly on the Hugging Face platform.

Access the Model

Prithvi Weather and Climate Foundation Model

The Prithvi Weather-Climate Foundation Model enables Earth science projects like hurricane tracking, severe weather risk mapping, and renewable energy forecasting. It is pre-trained on Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA2) satellite data. The model is released openly on the Hugging Face platform.

Access the Model

INDUS Large Language Models

Named for the southern sky constellation, INDUS (stylized in all caps) is a suite of large language models tailored for astrophysics, biological and physical sciences, Earth science, heliophysics, and planetary science. The model is released openly on the Hugging Face platform.

Access the Model