From Pretraining to Practice: Highlights from the 2026 ESA-NASA International Workshop on AI Foundation Models for Earth Observation

Hamed Alemohammad[1], Mina Burns[2], Ruben Cartuyvels[3], Nidhi Jha[4], Robert E. Kennedy[2], Nicolas Longepe[3], Valerio Marsocci[3], Rahul Ramachandran[5], Sujit Roy[4], Campbell Watson[6], Isabelle Wittmann[7]

1. Clark University - Center for Geospatial Analytics
2. Oregon State University - Environmental Monitoring, Analysis and Process Recognition Lab
3. European Space Agency - ϕ-Lab
4. University of Alabama in Huntsville - Earth Systems Science Center
5. National Aeronautics and Space Administration - Marshall Space Flight Center
6. IBM Thomas J. Watson Research Center - Yorktown Heights
7. IBM Research Europe - Zurich

The authors are listed in alphabetical order.

Abstract

The second ESA-NASA International Workshop on AI Foundation Models for Earth Observation brought together the Earth observation and AI communities from academic, commercial, and governmental organizations to assess the rapidly evolving role of foundation models in geospatial science applications. Hosted by NASA Marshall Space Flight Center in May of 2026, the workshop built on the inaugural meeting by shifting focus from early alignment and prototyping toward how geospatial foundation models should be pretrained, evaluated, adapted, trusted, deployed, and governed.

Across the workshop, a clear theme emerged: the community has moved beyond questioning whether large models can be pretrained on satellite data. Instead, discussions centered on what pretraining should teach, what information current models fail to capture, how model outputs can be trusted, and when foundation model approaches provide practical value over conventional methods. During sessions, attendees repeatedly emphasized that no single pretraining recipe, model architecture, benchmark, or deployment strategy fits all downstream tasks, making data quality, curation, uncertainty, evaluation design, operational constraints, and user-centered validation just as critical as scale.

The workshop also highlighted a transition from standalone foundation models toward broader geospatial AI systems. Open-source tools, scalable serving approaches, embedding products, onboard inference, operational decision-support applications, and agentic AI workflows are beginning to connect these models with data archives, benchmarks, scientific literature, code, and human expertise. This report summarizes pivotal discussions from the workshop and outlines future directions for benchmarking, embeddings, adaptation, operationalization, trust, agentic AI, and the unique role of public agencies in supporting an open, reliable geospatial AI ecosystem.

Index Terms*: Earth observation, foundation models, geospatial artificial intelligence, benchmarking, embeddings, agentic AI, operational deployment, Earth system science.*

I. INTRODUCTION

In May 2026, the European Space Agency (ESA) and the National Aeronautics and Space Administration (NASA) convened the second International Workshop on AI Foundation Models for Earth Observation. This year's workshop was hosted by NASA's Marshall Space Flight Center in Huntsville, Alabama. The workshop built on the first ESA–NASA workshop [1], which brought the artificial intelligence, Earth observation, and Earth science communities together around the development of open, trustworthy, and scalable geospatial foundation models. The second workshop retained this overarching ambition while reflecting a more mature and demanding phase of the field. Rather than asking only whether foundation models could be trained for Earth observation, participants examined how such models should be designed, evaluated, adapted, deployed, and used in scientific and operational settings. The 2026 event brought together 263 in-person participants and 464 virtual registrants, including agency leaders, researchers, tool developers, operational users, commercial practitioners, early-career researchers, and domain scientists. This broad participation created opportunities to assess technical advances against scientific requirements, institutional constraints, and operational needs. Organic interactions, particularly through breakout discussions, were intended to generate new research ideas, collaborations, and shared community priorities.

In his opening remarks, Dr. Rahul Ramachandran outlined the workshop's central objectives: exchanging advances in foundation-model science and applications for Earth observation; showcasing emerging artificial-intelligence capabilities; identifying research gaps and priorities; encouraging the shared use of datasets, benchmarks, and models; and engaging commercial organizations and early-career researchers. He also challenged participants to move beyond a narrow focus on computer vision and optical imagery, adopt a systems-level view of Earth observation, and establish a dedicated "AI Agents for Science" community. The program was organized around four interconnected themes. The first concerned the advancement and adaptation of foundation models for multimodal, multiresolution, multitemporal, and vision-language Earth observation data. The second examined emerging applications, including natural-language interfaces and agentic artificial intelligence. The third focused on science-based benchmarking, model evaluation, uncertainty, and the exploration of learned embedding spaces. The fourth addressed responsible artificial intelligence, operational deployment, and commercial use cases. Together, these themes emphasized that model architecture, evaluation, application, and governance must be considered as components of a shared Earth observation research agenda.

Dr. Nicolas Longepe summarized the inaugural workshop. That event helped align the artificial-intelligence and Earth observation communities around geospatial foundation models and established the long-term goal of moving from research prototypes toward operational, open, trustworthy, and scalable systems. The 2026 workshop extended this agenda through a more focused examination of reproducibility, benchmarking, hardware compatibility, natural-language interaction, compression, embeddings, model limitations, and multi-agent workflows.

Several scientific and technical directions emerged from the first workshop. Future models will need to integrate multiple data modalities, spatial resolutions, and temporal scales while

combining Earth observation and weather data to better represent biological, phenological, atmospheric, and other Earth-system processes. Participants also highlighted the growing convergence of physics-based and data-driven modeling, including approaches that incorporate physical constraints or hybrid techniques in areas such as radiative transfer and hydrology. Other themes included efficient adaptation strategies, such as head-only fine-tuning and knowledge distillation, as well as deployment frameworks for edge and onboard computing. The discussions also exposed unresolved questions about the value and limitations of geospatial foundation models. Geospatial foundation models do not provide clear advantages for every application, particularly when their performance is comparable to that of a task-specific architecture such as a custom U-Net. Participants therefore emphasized the need for benchmarking frameworks that are accessible, reproducible, scientifically meaningful, and actionable. Additional questions concerned what foundation models learn, how their representations should be interpreted, how confidence and uncertainty should be quantified, and how model limitations should be communicated to users. AI-driven compression and embedding pipelines offer promising mechanisms for scaling Earth observation analysis, but they introduce further scientific and operational concerns. These include determining acceptable levels of compression before critical semantic or physical information is lost, integrating physical models into differentiable processing pipelines, and ensuring that learned embeddings generalize beyond benchmark datasets to real applications. At the same time, natural-language interfaces may make complex Earth observation resources more accessible to non-experts, while multi-agent frameworks could coordinate tasks such as data discovery, ingestion, model training, analysis, and interpretation.

The findings of the inaugural workshop directly informed the design of the 2026 event, whose sessions revisited those earlier priorities while addressing the new scientific and operational challenges that had emerged as the field matured.

This paper follows the style and purpose of the 2025 ESA–NASA workshop report [1]. It is neither an exhaustive review of the geospatial foundation-model literature nor a proceedings volume that summarizes every presentation equally. Instead, it synthesizes the principal session-level outcomes and cross-cutting directions that emerged during the workshop. Its purpose is to document where the community appears to be heading, identify the scientific and operational tensions that remain unresolved, and clarify the shared infrastructure needed to move from promising models toward reliable Earth observation systems.

II. AGENCY PERSPECTIVES: FROM STRATEGY TO COORDINATION

Moderated by Dr. Rahul Ramachandran of NASA, the Agency Perspectives session examined how public institutions can respond to an AI landscape evolving faster than many organizational strategies and operational systems. Representatives from NASA, ESA, NOAA, and Oak Ridge

National Laboratory outlined their institutional priorities, the roles agencies can play alongside industry and academia, and opportunities for sustained cross-agency collaboration.

Matthew McClure presented AI as a fundamental capability for NASA science, with the potential to accelerate discovery, improve the analysis of large and complex datasets, and support systems that can adapt and operate more independently in demanding environments. His presentation connected AI with advanced computing, visualization, scientific instruments, and mission infrastructure, framing NASA's "5+1 AI for Science Strategy" as part of a coordinated effort to integrate AI across scientific workflows rather than pursue isolated applications.

Emily Sylak-Glassman described NASA Earth Science Division's GeoAI portfolio as an end-to-end effort spanning missions, research, data systems, modeling, technology, and decision support. NASA is working to make its datasets AI-ready, improve data discovery and access, explore onboard processing, and determine where the agency should lead, support, or leverage advances from industry, academia, and nonprofit organizations. Within Earth Science Data Systems, the Earth Data Intelligence strategy emphasizes AI-assisted production pipelines, scalable infrastructure, intelligent data access, reproducible workflows, governance, and standards. Earth Action's strategy places decision-makers and applied users at the center of foundation-model development. Its activities include supporting applications based on the Prithvi Earth Observation and Weather and Climate models, incorporating disaster-science requirements into model design, and creating pathways for applied researchers and users to shape benchmarks. The program is also testing whether foundation models make applications faster, less expensive, equally accurate, or easier to scale than established approaches, while examining whether language models communicate uncertainty well enough for decision support.

Giuseppe Borghi situated Earth observation AI within ESA's agency-wide AI Strategy Plan, which spans mission design, operations, autonomous systems, infrastructure, data exploitation, and scientific applications. ESA Φ-lab supports this strategy by moving transformative ideas from early research toward application and commercialization, linking technology readiness with business readiness and European industrial competitiveness. The Φ-lab portfolio includes foundation models, generative and agentic AI, digital assistants and twins, explainable AI, physics-informed machine learning, onboard processing, and advanced computing. These technologies support ESA's broader transition from Earth observation data toward actionable information, with partnerships across member states, public institutions, industry, and international organizations serving as a central mechanism for implementation.

Rob Redmon characterized NOAA's approach as moving AI "at the speed of trust" from science into operations. NOAA evaluates AI according to whether it improves the accuracy, timeliness, efficiency, and effectiveness of its weather, climate, ocean, coastal, and resource-management missions. The NOAA Center for AI supports this work through governance, AI-ready data and metadata standards, training, communities of practice, partnerships, and reusable workflow resources. NOAA's applications illustrate the breadth of this operational focus, including experimental ensemble forecasting, AI-driven global weather models, lightning prediction, harmful algal bloom forecasting, automated fish detection, tropical-cyclone datasets, sonar analysis, and real-time fire monitoring. These examples demonstrate that operational AI requires more than an

accurate model: it also depends on dependable data pipelines, realistic testing, workflow integration, sustained maintenance, and clear communication of uncertainty and limitations.

Forrest Hoffman presented the Department of Energy's Genesis Mission as a national effort to connect AI, high-performance computing, scientific instruments, secure data, simulation tools, and national-laboratory expertise. The mission seeks to identify scientific and technological challenges in which AI can demonstrate a meaningful advantage, while the Genesis Mission Consortium brings together laboratories, universities, industry, nonprofits, government partners, and other organizations to contribute computing resources, technical expertise, and related capabilities. The American Science Cloud supports this vision by providing a common platform for AI-ready datasets, model services, large-scale training, simulations, scientific instruments, secure computing, and cross-laboratory workflows. This approach highlights the role of national laboratories as infrastructure providers, scientific integrators, and testing environments in which commercial technologies can be evaluated against public research objectives.

Across the presentations, agencies were not positioned as direct competitors with industry in model scale or development speed. Their distinct strengths lie in stewarding long-term scientific datasets, defining mission requirements, establishing standards, supporting reproducibility, validating systems in consequential settings, and maintaining infrastructure for sustained public use. Depending on the context, agencies may act as partners, customers, funders, conveners, validators, infrastructure providers, and guardians of public value. The presentations also emphasized that moving AI from research into operations is an institutional as well as a technical challenge. Agencies must balance rapid innovation with governance, security, scientific credibility, workforce capacity, and continuity of service. Over the next five to ten years, AI is likely to become embedded throughout the Earth observation lifecycle, from mission design and onboard processing to data discovery, forecasting, analysis, and decision support. Realizing that future will require complementary roles for agencies, industry, laboratories, and the scientific community, together with continued human oversight and clear accountability.

III. TECHNICAL KEYNOTES: REPRESENTATION, OPENNESS, AND COMPUTING CONVERGENCE

The technical keynotes situated Earth observation (EO) foundation models (FMs) within a broader paradigm shift in scientific computing and representation learning. Juan Bernabe Moreno (IBM Research) described the evolution from task-specific models toward foundation model platforms and, increasingly, agentic systems. In traditional workflows, separate models are typically developed for each individual use case. Foundation models disrupt this pattern by leveraging broad pretraining to create reusable representations that can later be fine-tuned, adapted, or

connected to downstream tools. Moreno also positioned AI, high-performance computing, and quantum computing as converging technologies for scientific discovery, arguing that these paradigms increasingly complement one another in scientific workflows.

Mikolaj Czerkawski (Asterisk Labs) emphasized the importance of open Earth intelligence and open-source geospatial AI, while highlighting the tension between quality and coverage in foundation model development. He discussed BetaEarth [2], an open-weight, open-embedding model used to explore whether a foundation model can be emulated exclusively from its embeddings. Additionally, Major TOM [3] and its extension revealed current systematic biases, noting that most models exhibit significantly larger errors for the Southern Hemisphere alongside longitudinal biases. A critical observation from this work was that generative foundation models tend to collapse toward average values, failing to capture the full distribution of possible outputs. Czerkawski concluded that "datasets are foundational more often than AI models" and introduced upcoming projects aimed at traceable, trustable, and scalable AI compression of Earth data.

Pierre Gentine (Columbia University) connected foundation models to physical representation learning. His keynote emphasized that useful latent spaces must separate meaningful physical and semantic modes rather than merely reconstructing pixels. The discussion traced the evolution of Earth observation foundation model architectures through recent iterations, including Prithvi-EO-2.0 (featuring 600M parameters, 4.2M training files, and 6 spectral bands) [4] and AlphaEarth Foundations, which fuses optical, radar, LiDAR, climate, and text into 64-bit embeddings [5]. Gentine highlighted the distinctive challenges of Earth observation: sensor heterogeneity, temporal dynamics, geographic distribution shifts, data-label scarcity, and scale mismatch. These complexities explain why a straightforward transfer of natural language processing or natural image computer vision remains insufficient for Earth system science.

Together, the keynotes framed a central tension for the workshop: while foundation models promise reusable, scalable, and adaptable AI-native representations, Earth observation requires those representations to remain physically meaningful, open for community use, efficient for operational deployment, and transparent for scientific trust.

IV. GOOD PRACTICES FOR GEOFM DEVELOPMENT AND DEPLOYMENT

The session on good practices demonstrated that the community is actively building the software and infrastructure stack needed to transition from research prototypes to repeatable workflows. TerraKit [6] was introduced as an open-source Python library for multimodal and multiresolution geospatial dataset curation, enabling users to retrieve data from multiple sources and prepare it for training, fine-tuning, and inference. Similarly, TerraTorch [7] was presented as a standardized geospatial fine-tuning toolkit built on PyTorch, PyTorch Lightning [8], and TorchGeo [9] that supports model factories, custom model integration, and generic embedding generation.

Several presentations addressed the operational gap between successful model training and scalable inference. Work featuring vLLM [10], Kubernetes [11], and KServe [12] emphasized that

production success metrics must extend beyond accuracy to include latency, throughput, request handling, and deployment scalability. GEOStudio [13] extended this workflow into an integrated exploration and orchestration environment, seamlessly connecting dataset curation, fine-tuning, inference, infrastructure, graphical interfaces, and LLM-agent workflows.

Practical application examples highlighted the real-world value of these tools. For instance, the Hawaii Cropland Data Layers project explored how satellite embeddings can support operational crop mapping. Additionally, a discussion on onboard autonomy illustrated how geospatial foundation models may eventually support processing directly on satellites, shifting inference and decision logic closer to the sensor. These examples suggest that best practices now span the entire lifecycle: data access, curation, fine-tuning, benchmarking, inference, serving, deployment, interfaces, agents, and edge or onboard autonomy.

Ultimately, the main takeaway was that sound methodology is no longer confined to model architecture alone. It now encompasses reproducible workflows, interoperable software, standardized interfaces, scalable serving, user-facing systems, and deployment-aware design. This robust infrastructure will be necessary to sustain the next generation of scalable geospatial AI.

V. BENCHMARKING AND EVALUATION: FROM LEADERBOARDS TO CAPABILITY

Benchmarking emerged as a central focus of the workshop. The first benchmarking session emphasized that no single geospatial foundation model reigns supreme; instead, evaluation should help users understand when and why a specific model should be applied, rather than merely ranking models by aggregate scores. To address this, GEO-Bench-2 [14] was presented as a capability-oriented framework designed to evaluate models across different dataset types, resolutions, temporal structures, and downstream tasks. This framing reflects a broader shift from simple performance measurement to comprehensive capability assessment [15].

Other presentations expanded the scope of evaluation far beyond standard accuracy metrics. Holistic benchmarking now considers representation robustness, interpretability, perturbation stability, embedding-space comparisons, resource requirements, power consumption, and carbon impacts. Specifically, GELOS [39] focused on an embedding-centered evaluation approach. Furthermore, studies on dynamic irrigation, operational-trust taxonomies, coastal and marine benchmarks, and adversarial robustness demonstrated that evaluation is becoming both broader and more use-case specific. The consensus was not that every benchmark must include every dimension, but rather that the community requires evaluation approaches that reflect the real-world constraints under which models operate.

The second benchmarking session deepened this discussion through comparative and domain-specific case studies. A comparison of TerraMind [16] and THOR [17] revealed that architectural and experimental design choices – such as patch size, decoder design, data regime, modality fusion, compute budget, and frozen versus fine-tuned backbones – can swing performance by

5% to 20%. Consequently, these choices often outweigh the importance of model identity and pretraining strategy alone. Notably, increased compute did not always yield better results. Adversarial robustness testing further revealed that high accuracy does not guarantee robustness, and scale does not imply security; for example, a standard ResNet-50 demonstrated greater average robustness in segmentation tasks than larger, transformer-based foundation models. Independent benchmarking of 11 foundation models across six LULC datasets revealed that a CNN baseline (SSL4EO-ResNet50 [18]) remains competitive with many newer architectures, prompting the observation that researchers may be prematurely overlooking CNNs.

The collaborative benchmarking discussion reinforced these concerns from a practitioner perspective, with participants questioning whether existing datasets, protocols, and leaderboards are sufficient for real-world evaluation. They argued that benchmarks should inform users what a model is doing, when it is likely to work, which constraints matter, and how results connect to operational decisions. Additionally, a tournament-style vote across eight benchmarking innovation categories produced a decisive result: time-series benchmarks were identified as the most critically needed capability, winning unanimously in the semifinal and narrowly defeating thematic benchmarks in the final. The community also called for end-user-defined benchmarks under realistic conditions rather than the curated settings that dominate current leaderboards. From a research and development standpoint, participants emphasized the need for community-wide shared practices and tools to enforce consistent, fair, and transparent evaluations.

Overall, benchmarking is shifting from leaderboard competition to a scientific and operational design problem. The most useful benchmarks will help practitioners select models for specific scenarios, understand failure modes, evaluate robustness, and connect model behavior to real-world decisions. The next phase of benchmarking must therefore ask not only which model scores highest, but which model works for which task, under what data and compute constraints, with what uncertainty, and for whose ultimate decision.

VI. EMBEDDINGS AS INFRASTRUCTURE FOR DISCOVERY AND DECISION WORKFLOWS

Embeddings were discussed throughout the workshop not only as internal model representations but also as potential geospatial infrastructure. This collaborative discussion raised foundational questions about how embeddings are located on Earth, how they operate across spatial and temporal resolutions, how they should be stored and discovered, and whether precomputed embeddings require a dedicated benchmark category. Participants emphasized that embeddings can lower barriers for new geospatial data users, while questioning whether they solve real-world problems on their own or require additional interpretation, validation, and iteration.

Practical concerns for embedding ecosystems were deemed just as important as scientific questions. The discussion highlighted community infrastructure, potential cloud-native storage and metadata practices, and the need to distinguish broad foundation models from domain-

specific ones. Convergence speed, data efficiency, usability for non-specialists, and cross-version compatibility were identified as critical evaluation dimensions. A particularly vital issue was backward compatibility: if embeddings become outdated whenever a model changes, they risk becoming single-use artifacts rather than durable data products.

The embeddings session demonstrated how these questions are transitioning into real-world applications. Context-aware, multimodal spatiotemporal embeddings were presented as a way to provide a unified latent representation per location and timestamp, allowing downstream models to select the precise spatial and temporal context needed for a task [19]. Queryable Earth showed how text-aligned embeddings enable natural-language searches over Earth observation imagery, with larger vision-language models achieving 86% captioning accuracy (rising to 94% with geolocation context). Finally, PiCSRL [21] combined physics-informed spectral indices with reinforcement learning for adaptive environmental sensing, achieving 98.4% accuracy in harmful algal bloom detection with representations that generalize better than raw-band models. Deployable learned compression demonstrated 124x to 315x compression ratios at PSNR above 35 dB on payload-grade ARM processors, directly addressing the LEO downlink bottleneck.

The report-outs summarized this shift directly: embeddings are becoming the deliverable. Difficult tasks such as change detection can sometimes be reframed as similarity lookups; non-experts can interact with Earth observation data without relying on expert and task-dependent data pipelines; and learned representations may become a bridge between raw imagery, databases, agents, and decision-support tools. However, this future requires embedding-specific benchmarks, storage standards, metadata conventions, versioning practices, compatibility strategies, and clearer success stories beyond proof-of-concept demonstrations.

VII. LATEST ADVANCES IN AI FOUNDATION MODELS

The sessions on latest advances demonstrated that the frontier of Earth observation foundation models is shifting from mere scale toward purposeful pretraining, uncertainty quantification, temporal reasoning, language interfaces, physical structure, and operational utility. Presentations highlighted a wide range of innovations, including knowledge-guided multimodal pretraining [22], TerraMind 2.0, THOR [17], TerraVerse, and spectral targeted masking [23]. Researchers also introduced stochastic generative modeling of Copernicus data [24-25], partial differential equation (PDE) foundation models for planetary atmospheres [26], reliability-aware foundation models [27], semantic change captioning, open-world change detection, zero-shot regional weather forecasting, geospatial vision-language models, and spatiotemporal cloud and convection prediction.

A recurring theme throughout these sessions was the fundamental question of what pretraining should teach. For example, knowledge-guided variable-step forecasting [22] utilized relationships among environmental modalities to define a pretraining objective, moving away from a sole reliance on masked reconstruction or next-token prediction. Similarly, SpecTM [23] introduced targeted spectral masking that prioritizes diagnostic wavelengths carrying physical information;

this approach achieved 1.8x to 2.2x accuracy improvements over physics-blind baselines using only 5% labeled data. Furthermore, PDE foundation models pretrained on partial differential equation dynamics enabled accurate Martian atmospheric forecasting using just four years of reanalysis data on a single GPU. Together, these presentations underscored a broader consensus in the field: EO foundation models must learn the underlying relationships among weather, land cover, spectral response, temporal dynamics, and physical processes, rather than focusing strictly on pixel-level reconstructions.

Other advances emphasized temporal awareness, compute adaptivity, multimodality, and efficient scaling. TerraMind 2.0 introduced temporal disjoint sampling, multi-resolution processing, and multi-codebook quantization, achieving 10x parameter efficiency gains on temporal tasks at constant accuracy. Its open-vocabulary semantic segmentation – trained on 9 million annotated samples from OpenStreetMap and Overture across 174 countries – outperformed SAM3 [41], DINOv3 [42], and GeoPixel [43] by over 115%. Meanwhile, THOR [17] illustrated a complementary approach through GSD-aware positional encoding and FlexiViT-style randomized patch sizes, performing strongly in low-labeled-data regimes. TerraVerse and COP-GEN further pointed toward generative and simulation-oriented applications for foundation models.

Evaluation paradigms themselves also came under scrutiny. COP-GEN [25], a stochastic generative model, challenged the field's reliance on single-reference metrics by evaluating whether model output distributions match true observation distributions. Specifically, COP-GEN covers 90% of the real observation manifold versus just 2.8% for deterministic approaches, and spans 63% of the real per-band reflectance range versus 18%. The argument that evaluating distributions, rather than single samples, is essential for generative EO models stood out as one of the session's most important methodological contributions. Finally, SHRUG-FM [27] addressed the common limitation that foundation models rarely provide calibrated confidence. It disentangles aleatoric and epistemic uncertainty by feeding input data flags, embedding distribution flags, and decoder uncertainty into interpretable decision trees for selective prediction.

The second session on latest advances placed trust and language at the center of the discussion. Research in semantic change captioning, open-world change detection, and geospatial vision-language modeling demonstrated a shift beyond simply asking where an image changed toward understanding what changed, how it changed, and how users can express EO questions in natural language. Additionally, zero-shot regional weather forecasting showed that pretraining on reanalysis data introduces realistic high-resolution structures in entirely held-out regions, reducing local data requirements by 95 to 100%. Together, these directions point toward a future in which users interact with imagery through language-mediated, model-backed analytic systems.

The major takeaway is that model scale is no longer the sole frontier. Instead, the community is increasingly investigating what models know, how they know it, what physical or semantic relationships they encode, how uncertainty is surfaced, and how users can effectively interrogate model outputs.

VIII. ADAPTING FOUNDATION MODELS TO GEOSPATIAL DATA AND SPECIFIC EO TASKS

The adaptation session emphasized that a model being “foundational” does not mean it is universally sufficient. Geospatial foundation models must frequently be adapted to the specific signals, resolutions, modalities, labeling regimes, and scientific constraints of Earth observation tasks. Presentations covered scaling studies [28] on MajorTOM and FastTOM using the PhilEO Bench [29], phase-aware temporal models for conflict damage mapping, contrastive learning for P-band PolSAR soil moisture retrieval [46], and methods for handling sparse, asynchronous in situ air pollution data. Additional topics included the integration of agricultural reference data from satellites and field photos, ICESat-2 applications for canopy height change, and regional-scale biomass monitoring.

Scaling and benchmarking work [28] demonstrated that while increasing model size and data scale can yield performance gains, optimal choices ultimately depend on the specific task, labeled-data availability, architecture, compute constraints, and data regime. CNN-based approaches often remain efficient for smaller datasets, whereas transformer-based models can exploit very large datasets but require greater resources. Furthermore, low-shot settings, pixel-wise regression, decoder choice, and task-specific data structure all influence which model architecture is most appropriate.

The phase-aware SAR presentation provided a stark example of the limitations of current foundation models. In conflict damage mapping, SAR coherence carries the strongest signal for building damage detection, yet no current foundation model integrates phase information. Quantitative results underscored this gap: amplitude-only approaches achieved a mere 5% precision, and amplitude-based foundation model z-scores reached 44%, whereas coherence-based methods achieved 61%. This disparity demonstrates that a modality entirely absent from FM pretraining can dramatically outperform the models themselves in rare-event detection. Consequently, these findings suggest that anomaly-based downstream workflows combining FMs with physics-derived features may be more appropriate for extreme events than end-to-end approaches.

Other presentations reinforced the need to combine foundation model representations with external data and task-specific information. For instance, integrating sparse in-situ NO2 measurements into FM embeddings improved the R-squared value of air pollution mapping from 0.71 to 0.84 and halved the mean prediction error. Similarly, AgroVision [44] combined satellite time series, geotagged field photographs from the EU's LUCAS survey (comprising 250,000 fully open-source samples across 40 crop types), and embeddings from five different FMs. This integration demonstrated that multi-modal fusion overcomes single-modality observational limitations and achieves highest accuracy on crop type classification. Furthermore, canopy-height change mapping using AlphaEarth satellite embeddings with ICESat-2 LiDAR data revealed

asymmetric forest dynamics across the southeastern United States, characterized by stronger growth in northern regions and greater losses along the Gulf Coast due to hurricane impacts.

Ultimately, successful adaptation depends on identifying what a model has learned, what it lacks, and what external information must be introduced for a given task. This makes model adaptation central to scientific utility rather than being a secondary step after pretraining.

IX. OPERATIONAL AND COMMERCIAL USE OF FOUNDATION MODELS IN EARTH OBSERVATION

The operational and commercial sessions demonstrated that foundation models are beginning to deliver tangible value in applied settings. However, they also underscored that operational success depends primarily on usability, trust, latency, spatial context, uncertainty, and decision relevance.

Practical testing of Prithvi-EO-2.0 on forest biomass, land use, and landscape change revealed how embeddings improve spatial context beyond traditional pixel-based methods. Notably, the embeddings approach achieved the highest accuracy in Land Use and Land Cover (LULC) classification. Prithvi also detected features missed by imperfect training datasets, prompting the observation that apparent "false positives" were often actually correct. Prediction of forest biomass also illustrated the challenges of mapping with chip-based paradigms, with chip boundary and positional artifacts emerging when wall-to-wall maps are built. As one discussion theme emphasized, maps matter: operational users require outputs they can readily interpret and trust.

Several commercial and operational deployments highlighted exactly where FMs are already creating value. For instance, OroraTech demonstrated wildfire detection from its eight satellite constellation using dense MWIR embeddings [30]. In this setup, fires naturally separate from background within frozen embeddings without requiring labels, achieving a 70% to 90% joint detection rate with VIIRS while detecting smaller fires much sooner. Similarly, MineSite, built on Clay embeddings [31], applies cosine anomaly change detection to monitor artisanal mining, border activity, and dam risks. The system assesses three years of activity in under 30 minutes, enabling non-expert decision-makers to leverage FM products effectively. In addition, EarthDaily Analytics presented an agricultural forecasting tool that predicts corn harvest dates with 7-day median error up to 60 days in advance, though accuracy degrades significantly at coarser resolutions and lower temporal frequencies. As the presenter explicitly noted, "datasets are foundational more often than AI models," reinforcing that commercial applications still require scientific rigor.

A transition from static classification toward dynamic, predictive modeling was evident. Hierarchical data assimilation for wind farm digital twins blended LiDAR, wind towers, satellite, and numerical weather data across three scales, proving less computationally expensive and

more accurate than monolithic approaches when local observations exist. Multi-hazard vulnerability modeling combined three foundation models (Prithvi [32], PrithviWxC [33], and Granite-LST [34]) to generate hourly land surface temperature fields. This approach identified 33,000 additional Indianapolis residents at risk for extreme heat that had been missed by static vulnerability indices. Overall, this trajectory represents a shift from EO-informed static indices to multi-model, calibrated dynamic systems.

The key conclusion was that operational adoption will depend less on whether a method is labeled a “foundation model” and more on whether it produces timely, trustworthy, and usable information for real-world decisions. While FMs are valuable for improving context, reducing labeling burdens, enabling faster monitoring, supporting higher spatial or temporal resolution, and integrating uncertainty into decision workflows, they must be validated against operational needs, rather than just research benchmarks.

X. AGENTIC AI FOR EARTH OBSERVATION AND SCIENTIFIC WORKFLOWS

Agentic AI emerged as one of the most forward-looking themes of the workshop. The oral session demonstrated that geospatial agents are rapidly evolving beyond simple demos and chatbots into sophisticated scientific systems. These systems integrate planners, specialized tools, sub-agents, generated code, retrieval-augmented generation (RAG), knowledge graphs, memory, structured states, and human evaluation [37].

Trust served as the dominant design theme across these developments. For instance, a knowledge-graph-backed pipeline designed for EO literature gap analysis constructs persistent graphs across academic papers. It successfully identifies support, contradiction, and refinement relationships with associated confidence scores, all while preserving strict provenance from the synthesized gaps back to specific sentences, sections, and papers. Similarly, SciELF, a fact-checking benchmark developed alongside five NASA domain experts specializing in climate, hurricanes, and astrophysics, demonstrated current technological limits. While verification systems achieve an F1 score of approximately 0.84 within their domain, their performance degrades significantly when applied to out-of-domain web sources. To counter this, a two-stage fact-correction pipeline successfully raised factual precision from 0.42 to above 0.90. Throughout these discussions, presenters emphasized that "human expertise and knowledge is definitely the most valuable thing you can get" for robust scientific benchmarking.

Application-level agents further highlighted this transition from experimental prototypes to operationally useful systems. Drawing from three years of production experience in building geo-LLM agents, Development Seed distilled a core design principle: "Don't let the LLM generate the analysis; use the tooling to write code that produces it." Their recommended architecture relies on a light orchestrator paired with smart deterministic tools, rich tool descriptions, narrow code-generation scope, data passed by reference rather than directly in-context, and established EO

infrastructure like STAC [35]. Additionally, the Hydrology Copilot [45], a cloud-native, multi-agent system for drought analysis built on NASA's NLDAS-3 dataset, completes complex, multi-decade climatology analyses in just two to three minutes. Crucially, it generates transparent, inspectable Python code instead of black-box outputs. As the presenter noted, "NASA data is free but expensive to use," and the speed of this workflow is transformative.

The EVE platform [36], developed under ESA Phi-Lab in collaboration with Pi School, Mistral, and Wiley, provided one of the most developed examples of an LLM-based assistant for EO. It features a 24B-parameter model fine-tuned on a 2.8-billion-token, domain-specific corpus, utilizing retrieval-augmented generation across approximately 270,000 open-access documents, chunk-level source attribution, and a dedicated hallucination detection module. Over 350 users from academia, space agencies, and government have piloted the system over a six-month period. In parallel, an agentic AI framework for ILAMB [38] benchmarking proposed connecting foundation models with process-based Earth System Model evaluation, demonstrating automated multi-variable, multi-model benchmarking that traditionally takes months to complete.

Following the oral session, an ad-hoc collaborative discussion session was scheduled with the theme of building a community for AI agents. Conversations in this session extended beyond technical concerns into governance, sharing, reproducibility, credit, and accountability. Participants discussed how scientists might contribute to agentic workflows through user groups, structured subject-matter-expert intake, reusable scripts, shared evaluations, and public examples. They also emphasized the need to cite agents in papers, preserve execution artifacts and tool traces, benchmark agents across various models and use cases, and clarify liability when an agent produces incorrect results.

Ultimately, agentic AI may become the connective tissue between foundation models and scientific use. However, the workshop made clear that widespread adoption will require community practices for transparent traces, reproducible workflows, versioned tools, proper citations, safety, cybersecurity, accessibility, and domain-expert co-design. Without these safeguards, agents risk becoming another opaque layer between users and Earth observation data.

XI. GENERAL REMARKS AND FUTURE DIRECTIONS

A. Purposeful Pretraining Beyond Model Scale

The workshop discussions suggest that the next phase of Earth observation foundation-model development should be guided less by model size alone and more by explicit decisions about what pretraining is intended to teach. Reconstruction, masking, and generative objectives remain useful, but they should be assessed according to whether they encode the temporal, spectral, spatial, semantic, and physical relationships required by downstream applications. For many Earth science tasks, reproducing plausible pixels is not equivalent to learning environmental

processes, rare-event signatures, or meaningful cross-modal relationships. Pretraining design should therefore begin with a stated capability hypothesis, such as learning seasonal dynamics, identifying diagnostically important spectral features, representing relationships among environmental variables, or transferring across sensors and spatial scales.

Greater transparency about pretraining data and objectives would also make model capabilities and limitations easier to interpret. Developers should document the sensors, modalities, geographic regions, time periods, spatial resolutions, environmental conditions, and physical variables represented in the training corpus, as well as important omissions. Evaluation should then test whether the resulting representations retain the intended information across geographic regions, seasons, resolutions, and downstream tasks. This would help distinguish broadly reusable representations from models whose apparent generality depends on overlap between pretraining data and evaluation conditions.

B. Task-Aware Model and Adaptation Selection

The workshop provided little support for the idea that a single model will be optimal across the full range of Earth-observation problems. Model suitability depends on the target phenomenon, sensor modality, spatial and temporal resolution, label regime, geographic transfer requirement, compute budget, latency constraint, and level of uncertainty characterization required by the user. Broad multimodal models may be valuable when tasks benefit from reusable context or limited labels, while specialized self-supervised models, conventional neural networks, physics-based methods, or hybrid workflows may remain more effective when the relevant signal is narrow, well understood, or absent from foundation-model pretraining. The phrase “foundation model” should therefore describe a development approach rather than imply universal superiority.

Adaptation studies should report not only whether performance improved, but also which information had to be added before the model became useful. Relevant additions may include in situ observations, field imagery, task-specific labels, physical indices, SAR phase or coherence, numerical-model outputs, or domain-specific temporal features. These additions reveal which capabilities were not present in the pretrained representation and whether adaptation is correcting a modest domain mismatch or compensating for a fundamental omission. Comparisons should also include competitive conventional and specialized baselines so that the value attributable to pretraining can be separated from the value of the added data, decoder, or workflow.

A practical selection framework could organize models along a small set of decision axes: required modality, spatial support, temporal structure, label availability, geographic transfer, computational resources, deployment environment, and tolerance for uncertain outputs. Reporting results along these axes would be more useful than presenting a single aggregate score because it would allow researchers and practitioners to identify models that are appropriate under their actual constraints. Such a framework could also accommodate mixtures of experts or agentic systems that select among broad models, specialized models, deterministic tools, and physical methods for different parts of a workflow.

C. Diagnostic Evaluation of Model Capabilities

Benchmarking should evaluate identifiable model capabilities under controlled, reproducible, and sufficiently realistic conditions. Priority areas include time-series reasoning, geographic transfer, thematic specialization, robustness to perturbation and missing data, uncertainty calibration, semantic correctness, spatial coherence, and performance under scarce-label conditions. The purpose of these benchmarks is to expose capability boundaries and support fair comparison, rather than to reproduce the full complexity of an operational workflow. The community should converge towards a set of shared benchmarks, tools and practices to facilitate fairness and reproducibility.

Evaluation should also reflect the conditions under which systems will operate. Relevant dimensions include inference latency, memory and energy requirements, label and data-access costs, update frequency, geographic coverage, hardware compatibility, and performance degradation under resolution or temporal-density changes. For user-facing products, evaluation may additionally need to measure whether outputs are spatially consistent, interpretable, timely, and suitable for the decision being supported. Including these constraints would make benchmark results more predictive of real-world usefulness and reduce the gap between research performance and operational adoption.

The workshop discussions further suggest that benchmark design should involve end users and independent evaluators earlier in the process. End users can define realistic decision contexts, acceptable failure rates, meaningful spatial and temporal units, and the costs associated with false positives, false negatives, delayed outputs, or uncertain predictions. Independent testing, hidden evaluation sets where appropriate, versioned protocols, and reproducible evaluation environments can reduce overfitting to familiar leaderboards. Benchmarks themselves should also be periodically assessed against downstream outcomes: if high benchmark performance does not predict usefulness in realistic workflows, the benchmark should be revised rather than treated as authoritative.

D. Embeddings as Managed Geospatial Infrastructure

The workshop showed that embeddings are becoming reusable products rather than temporary intermediate outputs. To function as durable infrastructure, embedding collections should be accompanied by metadata describing the source observations, model name and version, spatial footprint, timestamp or temporal aggregation, spatial resolution, modality, preprocessing, uncertainty information where available, and intended or validated uses. Users should also be able to determine whether an embedding represents a single acquisition, a temporal composite, a location-specific history, or a multimodal fusion product. Without this information, embedding stores may lower the technical barrier to access while increasing uncertainty about what the representation actually contains.

Versioning and compatibility are particularly important because embeddings may be costly to generate and may support many downstream products. Model updates should not silently invalidate existing indexes, labels, similarity thresholds, or trained downstream models. The community should explore versioned embedding specifications, compatibility tests, migration procedures, and reference datasets that quantify how representation spaces change across

model releases. Where backward compatibility cannot be preserved, providers should clearly communicate the consequences for downstream users and archive the information needed to reproduce earlier products.

Embedding-specific evaluation should examine transferability, temporal stability, geographic bias, data efficiency, semantic retrieval quality, sensitivity to resolution and preprocessing, and suitability for different classes of downstream tasks. Such evaluation should also identify cases in which embedding similarity is misleading or in which a representation suppresses rare but decision-relevant signals. Shared embedding infrastructure can make Earth-observation data more accessible, but it does not eliminate the need for interpretation, calibration, or task-specific validation.

E. Operational Validation of End-to-End Systems

Operational validation should begin after a model has been incorporated into a complete product or workflow. The relevant unit of analysis is therefore not the pretrained backbone alone, but the resulting map, forecast, alert, monitoring service, risk estimate, or decision-support system. Evaluation should determine whether that system remains reliable over repeated use, integrates with established practices, communicates uncertainty effectively, and improves outcomes relative to the method currently used.

Claims of practical benefit should be evaluated against current practice rather than only against other foundation models. Depending on the application, the relevant comparator may be a conventional neural network, a physics-based model, a manual analysis workflow, an existing operational product, or a hybrid system already trusted by users. Evaluation should ask whether the new approach improves coverage, timeliness, spatial or temporal detail, labeling efficiency, uncertainty communication, or the quality of a downstream decision. This comparison would also clarify where foundation models introduce unnecessary complexity without producing a corresponding user benefit.

Domain experts and intended users should participate in problem formulation, validation, interface design, and acceptance testing. Their involvement is especially important when model outputs require interpretation, when error costs are asymmetric, or when local conditions are poorly represented in training data. User-centered validation can also reveal whether uncertainty displays, maps, natural-language interfaces, or automated recommendations are being interpreted as intended. The workshop discussions suggest that operational trust is more likely to emerge from sustained co-design than from technical performance claims alone.

F. Traceable, Reproducible, and Accountable AI Systems

Trust in geospatial AI depends on the full workflow, not only on the performance of the trained model. Scientific results are shaped by the data and preprocessing used, the model and adaptation strategy selected, the evaluation protocol applied, the software environment in which the analysis runs, and the conditions under which the system is deployed. Users should therefore

be able to determine which data, model, and software versions produced an output; how the system was adapted and evaluated; and how uncertainty and known limitations were handled. Model cards and dataset documentation provide a useful foundation, but consequential workflows also require versioned dependencies, inspectable evaluation artifacts, and provenance records that make results auditable and reproducible.

Agentic systems introduce additional traceability requirements because they can retrieve information, generate code, select tools, transform data, and revise intermediate results. Scientific agent workflows should preserve retrieved sources, generated code, tool calls, execution logs, model and tool versions, structured intermediate states where feasible, and the artifacts used to produce final conclusions. These records would allow users to distinguish a reproducible analysis from an unsupported natural-language response and would support debugging when an agent selects an inappropriate dataset, tool, or computational procedure.

Agent evaluation should include factuality, tool-selection accuracy, code and execution correctness, reproducibility, failure detection, sensitivity to model or tool updates, and performance outside the agent's primary domain. Systems used for consequential scientific or operational tasks should also define when human review is required, who is authorized to approve outputs, and who remains accountable when the system fails.

G. Public Agencies as Coordinators and Stewards

Public agencies have a distinct role that does not depend on matching the scale of private-sector model development. ESA, NASA, and partner organizations can steward trusted archives, preserve long-duration records, support open standards, convene diverse communities, and provide independent evaluation infrastructure. They can also help ensure that model development remains connected to Earth science priorities, operational users, and public-interest applications rather than being driven only by benchmark visibility or commercial demand. This comparative advantage is particularly important where scientific validity, continuity, geographic inclusion, and public accountability are central requirements.

Specific shared infrastructure could include curated evaluation datasets, independent benchmark services, reproducible computing environments, embedding registries, metadata and versioning standards, model and dataset documentation templates, and repositories for traceable agent workflows. Agencies could also support sustained comparisons among foundation models, specialized models, conventional approaches, and hybrid systems without presuming that one model category should prevail. Coordinating these resources across agencies and international partners would reduce duplicated effort and make evaluation results easier to compare and reuse.

Agency-supported programs should prioritize use cases for which market incentives may be insufficient, including long-term environmental monitoring, low-resource regions, rare hazards, open scientific infrastructure, and applications requiring transparent validation. Procurement, funding, and partnership mechanisms can encourage open interfaces, documented limitations, reproducible evaluations, and continuity beyond short project cycles. Agencies can also use their convening authority to bring researchers, commercial developers, infrastructure providers,

domain experts, and operational users into the same validation and governance process.

H. A Coordinated Agenda for the Next Phase

Taken together, the workshop discussions point toward an ecosystem rather than a single technical solution. The immediate need is for interoperable evidence: clear accounts of what pretraining is intended to encode, diagnostic evaluations of what models can and cannot do, adaptation studies that reveal missing information, reusable and versioned representations, operational validation against real decisions, and traceable systems that preserve human oversight. Progress in one area cannot substitute for weaknesses in the others; a powerful model evaluated on an uninformative benchmark or embedded in an opaque workflow will remain difficult to trust and use.

The field should therefore resist defining progress primarily through the size, number, or novelty of foundation models. More consequential indicators include whether models capture task-relevant Earth-system relationships, whether users can select them under realistic constraints, whether outputs improve established workflows, whether results can be reproduced, and whether institutions can sustain the required data and evaluation infrastructure. The workshop's broader message is that the speed of technical development should be matched by the speed at which the community builds evidence, standards, and trust.

XII. CONCLUSION

The 2026 ESA–NASA International Workshop on AI Foundation Models for Earth Observation showed that the field has entered a more mature and demanding phase. The central question is no longer simply whether large models can be pretrained on Earth observation data, but what those models should learn, which signals and processes they fail to capture, how their outputs should be evaluated and trusted, and under what conditions they provide meaningful advantages over specialized or conventional approaches. This shift places greater emphasis on purposeful pretraining, data quality, task-aware adaptation, uncertainty, and evidence of scientific or operational value.

Three broader changes in emphasis emerged across the workshop. First, progress is shifting from scale alone toward representations that capture relevant temporal, spectral, spatial, semantic, and physical relationships. Second, evaluation is moving from aggregate leaderboards toward diagnostic assessments of capability, robustness, constraints, and decision relevance. Third, foundation models are increasingly being treated as components of larger geospatial AI systems that include curated data, embeddings, adaptation tools, scalable inference, operational interfaces, scientific software, agentic workflows, and human expertise.

Credible progress will require stronger connections across the full model and system lifecycle. Developers and users need clearer documentation of pretraining data and objectives, guidance for selecting models under realistic task and resource constraints, benchmarks that expose capability boundaries, reusable and versioned embedding infrastructure, and validation against maps, forecasts, alerts, and decisions. Agentic systems will require additional safeguards, including inspectable tool use, preserved provenance, reproducible execution artifacts, and explicit human accountability. Together, these requirements provide a basis for distinguishing technically impressive demonstrations from systems that are scientifically defensible and operationally useful.

ESA, NASA, and partner agencies can play a catalytic role by stewarding trusted data, supporting open evaluation and reproducibility infrastructure, convening research and user communities, enabling independent validation, and sustaining public-interest applications that may not be adequately supported by short-term market incentives. If the first ESA–NASA workshop helped align the Earth observation and foundation-model communities, the second clarified the work required to turn that alignment into practice. Success should therefore be measured not by the number or size of new models, but by whether foundation-model-enabled systems become reliable, transparent, and useful instruments for Earth science, Earth action, and decision-making.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the members of the Organizing Committee for their leadership, guidance, and support in enabling this workshop, as well as the members of the Technical Committee for their contributions to the technical planning and execution of the workshop. The authors further acknowledge Amazon Web Services (AWS) for providing cloud computing credits that enabled the hands-on tutorial sessions.

Organizing Committee

Anca Anghelea; ESA Climate Office, United Kingdom

Katie Baynes; NASA Headquarters Earth Science Data Systems, United States

Giuseppe Borghi; ESA Φ-lab, Italy

Claudio Iacopino; ESA Φ-lab, Italy

Nicolas Longépé; ESA Φ-lab, Italy

Matt McClure; NASA Headquarters Office of the Chief Science Data Officer, United States

Kevin Murphy; NASA Headquarters Office of the Chief Information Officer, United States

Rahul Ramachandran; NASA Marshall Space Flight Center, United States

Emily Sylak-Glassman; NASA Headquarters Earth Science Division, United States

Programme Committee

Hamed Alemohammad - Clark University, USA

Valerio Marsocci - ESA Φ-lab, Italy

Sujit Roy - NASA IMPACT, University of Alabama in Huntsville, United States

Xiaoxiang Zhu - Technical University of Munich, Germany

Campbell Watson - IBM Research, USA

Technical Committee

Paulo Arevalo; Boston University, United States

Samantha Arundel; United States Geological Survey, United States

Thomas Brunschwiler; IBM Research Europe - Zurich, Switzerland

Amy Campbell; ESA, U.K

Ruben Cartuyvels; ESA Φ-lab, Italy

Jocelyn Chanussot; University Grenoble, France

Hoanen Chen; Colorado State University, United States

Begum Demir; Technische Universität Berlin, Germany

Nikolaos Dionelis; University of Alabama in Huntsville, United States

Paolo Fraccaro; IBM Research Europe - Zurich, Switzerland

William Jones; ESA Climate Office, United Kingdom

Robert Kennedy; Oregon State University, United States

Sebastien Lefèvre; Institute for Research in Computer Science and Random Systems, France

Wen Wen Li; Arizona State University, United States

Nicolas Longépé; ESA Φ-lab, Italy

Dalton Lunga; Oak Ridge National Laboratory, United States

Valerio Marsocci; ESA Φ-lab, Italy

Anirbit Mukherjee; University of Manchester, United Kingdom

Jakub Nalepa; KP Labs, Poland

Muthukumaran Ramasubramanian; University of Alabama in Huntsville, United States

Johannes Schmude; IBM Thomas J. Watson Research Center, United States

Rajat Shinde; University of Alabama in Huntsville, United States

Devis Tuia; École Polytechnique Fédérale de Lausanne, Switzerland

AWS Representative

Chris Stoner; Amazon Web Services, United States

REFERENCES

[1] Longépé, Nicolas, Hamed Alemohammad, Anca Anghelea, Thomas Brunschwiler, Gustau Camps-Valls, Gabriele Cavallaro, Jocelyn Chanussot, et al. 2025. "Earth Action in Transition: Highlights from the 2025 ESA–NASA International Workshop on AI Foundation Models for EO." *IEEE Geoscience and Remote Sensing Magazine* 13 (4): 457–62. https://doi.org/10.1109/MGRS.2025.3592035.

[2] Asterisk Labs. 2026. "BetaEarth: Emulator of AlphaEarth." Software repository. Accessed July 15, 2026. https://github.com/asterisk-labs/beta-earth.

[3] Francis, Alistair, and Mikolaj Czerkawski. 2024. "Major TOM: Expandable Datasets for Earth Observation." In *IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium*, 2935–40. https://doi.org/10.1109/IGARSS53475.2024.10640760.

[4] Szwarcman, Daniela, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, et al. 2024. "Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications." arXiv preprint arXiv:2412.02732. https://doi.org/10.48550/arXiv.2412.02732.

[5] Brown, Christopher F., Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, et al. 2025. "AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data." arXiv preprint arXiv:2507.22291. https://doi.org/10.48550/arXiv.2507.22291.

[6] Edwards, Blair, Rosie Lickorish, Benedikt Blumenstiel, Paolo Fraccaro, Fred Otieno, and Tong Luo. 2025. *TerraKit: An End-to-End Library for Preparation of AI-Ready Geospatial Datasets*. Computer software. IBM Research. https://research.ibm.com/publications/terrakit-an-end-to-end-library-for-preparation-of-ai-ready-geospatial-datasets

[7] Gomes, Carlos, Benedikt Blumenstiel, Joao Lucas de Sousa Almeida, Pedro Henrique de Oliveira, Paolo Fraccaro, Francesc Marti Escofet, Daniela Szwarcman, Naomi Simumba, Romeo Kienzler, and Bianca Zadrozny. 2025. "TerraTorch: The Geospatial Foundation Models Toolkit." arXiv preprint arXiv:2503.20563. https://doi.org/10.48550/arXiv.2503.20563.

[8] Lightning AI. n.d. "PyTorch Lightning Documentation." Software documentation. Accessed July 15, 2026. https://lightning.ai/docs/pytorch/stable/.

[9] Stewart, Adam J., Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. 2022. “TorchGeo: Deep Learning with Geospatial Data.” In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–12. https://doi.org/10.1145/3557915.3560953.

[10] Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. “Efficient Memory Management for Large Language Model Serving with PagedAttention.” In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–26. https://doi.org/10.1145/3600006.3613165.

[11] Kubernetes Authors. n.d. “Kubernetes Documentation.” Software documentation. Accessed July 15, 2026. https://kubernetes.io/docs/.

[12] KServe Community. n.d. “KServe Documentation.” Software documentation. Accessed July 15, 2026. https://kserve.github.io/website/docs/.

[13] TerraStackAI. 2026. “Geospatial Exploration and Orchestration Studio.” Software repository. Accessed July 15, 2026. https://github.com/terrastackai.

[14] Simumba, Naomi, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longépé, Xiao Xiang Zhu, Hannah Kerner, Juan Bernabe-Moreno, and Alexandre Lacoste. 2025. “GEO-Bench-2: From Performance to Capability, Rethinking Evaluation in Geospatial AI.” arXiv preprint arXiv:2511.15658. https://doi.org/10.48550/arXiv.2511.15658.

[15] Marsocci, Valerio, Yuru Jia, Georges Le Bellier, Dávid Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, et al. 2026. "PANGAEA: Assessing Geospatial Foundation Models Capabilities through a Global and Inclusive Benchmark." *IEEE Geoscience and Remote Sensing Magazine*. https://arxiv.org/abs/2412.04204

[16] Jakubik, Johannes, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, et al. 2025. “TerraMind: Large-Scale Generative Multimodality for Earth Observation.” arXiv preprint arXiv:2504.11171. https://doi.org/10.48550/arXiv.2504.11171.

[17] Forgaard, Theodor, Jarle H. Reksten, Anders U. Waldeland, Valerio Marsocci, Nicolas Longépé, Michael Kampffmeyer, and Arnt-Børre Salberg. 2026. “THOR: A Versatile Foundation Model for Earth Observation Climate and Society Applications.” arXiv preprint arXiv:2601.16011. https://doi.org/10.48550/arXiv.2601.16011.

[18] Wang, Yi, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. 2023. “SSL4EO-S12: A Large-Scale Multimodal, Multitemporal Dataset for Self-Supervised Learning in Earth Observation.” *IEEE Geoscience and Remote Sensing Magazine* 11 (3): 98–106.

[19] Peters, Julia, Karin Mora, Miguel D. Mahecha, Chaonan Ji, David Montero, Clemens Mosig, and Guido Kraemer. 2025. “Context-Aware Multimodal Representation Learning for Spatio-Temporally Explicit Environmental Modelling.” arXiv preprint arXiv:2511.11706. https://doi.org/10.48550/arXiv.2511.11706.

[20] Wang, Zhecheng, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. 2024. “SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing.” *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (6): 5805–13. https://doi.org/10.1609/aaai.v38i6.28393.

[21] Azadani, Mitra Nasr, Syed Usama Imtiaz, and Nasrin Alamdari. 2026. “PiCSRL: Physics-Informed Contextual Spectral Reinforcement Learning.” arXiv preprint arXiv:2603.26816. https://doi.org/10.48550/arXiv.2603.26816.

[22] Ravirathinam, Praveen, Ajitesh Parthasarathy, Ankush Khandelwal, Rahul Ghosh, and Vipin Kumar. 2024. “Towards Knowledge Guided Pretraining Approaches for Multimodal Foundation Models: Applications in Remote Sensing.” arXiv preprint arXiv:2407.19660. https://doi.org/10.48550/arXiv.2407.19660.

[23] Imtiaz, Syed Usama, Mitra Nasr Azadani, and Nasrin Alamdari. 2026. “SpecTM: Spectral Targeted Masking for Trustworthy Foundation Models.” arXiv preprint arXiv:2603.22097. https://doi.org/10.48550/arXiv.2603.22097.

[24] Espinosa, Miguel, Eva Gmelich Meijling, Valerio Marsocci, Elliot J. Crowley, and Mikolaj Czerkawski. 2026. “COP-GEN: Latent Diffusion Transformer for Copernicus Earth Observation Data—Generation Stochastic by Design.” arXiv preprint arXiv:2603.03239. https://doi.org/10.48550/arXiv.2603.03239.

[25] Espinosa, Miguel, Valerio Marsocci, Yuru Jia, Elliot J. Crowley, and Mikolaj Czerkawski. 2025. “COP-GEN-Beta: Unified Generative Modelling of Copernicus Imagery Thumbnails.” In *Proceedings of the CVPR 2025 Workshop on Multi-Modal Reasoning for Earth Observation*.

[26] Schmude, Johannes, Sujit Roy, Liping Wang, Theodore van Kessel, Levente Klein, Marcus Freitag, Eloisa Bentivegna, et al. 2026. “PDE Foundation Models Are Skillful AI Weather Emulators for the Martian Atmosphere.” arXiv preprint arXiv:2602.15004. https://doi.org/10.48550/arXiv.2602.15004.

[27] Cohrs, Kai-Hendrik, Zuzanna Osika, Maria Gonzalez-Calabuig, Vishal Nedungadi, Ruben Cartuyvels, Steffen Knoblauch, Joppe Massant, Shruti Nath, Patrick Ebel, and Vasileios Sitokonstantinou. 2025. “SHRUG-FM: Reliability-Aware Foundation Models for Earth Observation.” arXiv preprint arXiv:2511.10370. https://doi.org/10.48550/arXiv.2511.10370.

[28] Dionelis, Nikolaos, Riccardo Musto, Jente Bosmans, Simone Sarti, Giancarlo Paoletti, Peter Naylor, Valerio Marsocci, Sébastien Lefèvre, Bertrand Le Saux, and Nicolas Longépé. 2025. “Scaling Laws for Geospatial Foundation Models: A Case Study on PhilEO Bench.” arXiv preprint arXiv:2506.14765. https://doi.org/10.48550/arXiv.2506.14765.

[29] Fibaek, Casper, Luke Camilleri, Andreas Luyts, Nikolaos Dionelis, and Bertrand Le Saux. 2024. “PhilEO Bench: Evaluating Geo-Spatial Foundation Models.” In *IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium*. https://doi.org/10.1109/IGARSS53475.2024.10642780.

[30] Rötzer, Matthias, Veronika Pörtge, Martin Ickerott, Jayendra Praveen Kumar Chorapalli, Dimitri Scheftelowitsch, Max Bereczky, Dmitry Rashkovetsky, Sai Manoj Appalla, and Julia Gottfriedsen. 2026. “On-Orbit Real-Time Wildfire Detection under On-Board Constraints.” arXiv preprint arXiv:2605.06273. https://doi.org/10.48550/arXiv.2605.06273.

[31] Clay Foundation. n.d. “Clay Foundation Model: An Open-Source AI Model for Earth.” Model documentation and software repository. Accessed July 15, 2026. https://clay-foundation.github.io/model/.

[32] Jakubik, Johannes, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, et al. 2023. “Foundation Models for Generalist Geospatial Artificial Intelligence.” arXiv preprint arXiv:2310.18660. https://doi.org/10.48550/arXiv.2310.18660.

[33] Schmude, Johannes, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, et al. 2024. “Prithvi WxC: Foundation Model for Weather and Climate.” arXiv preprint arXiv:2409.13598. https://doi.org/10.48550/arXiv.2409.13598.

[34] IBM Granite. n.d. “Granite Geospatial Land Surface Temperature.” Model card and documentation. Accessed July 15, 2026. https://huggingface.co/ibm-granite/granite-geospatial-land-surface-temperature.

[35] SpatioTemporal Asset Catalog Community. n.d. “The SpatioTemporal Asset Catalog Specification.” Technical specification. Accessed July 15, 2026. https://stacspec.org/.

[36] Atrio, Àlex R., Antonio Lopez, Jino Rohit, Yassine El Ouahidi, Marcello Politi, Vijayasri Iyer, Umar Jamil, Sébastien Bratières, and Nicolas Longépé. 2026. “EVE: A Domain-Specific LLM Framework for Earth Intelligence.” arXiv preprint arXiv:2604.13071. https://doi.org/10.48550/arXiv.2604.13071.

[37] Bell, Aaron, Amit Aides, Amr Helmy, Arbaaz Muslim, Aviad Barzilai, Aviv Slobodkin, Bolous Jaber, et al. 2025. “Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning.” arXiv preprint arXiv:2510.18318. https://doi.org/10.48550/arXiv.2510.18318.

[38] Collier, Nathan, Forrest M. Hoffman, David M. Lawrence, Gretchen Keppel-Aleks, Charles D. Koven, William J. Riley, Mingquan Mu, and James T. Randerson. 2018. “The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation.” *Journal of Advances in Modeling Earth Systems* 10 (11): 2731–54. https://doi.org/10.1029/2018MS001354.

[39] Godwin, Denys, Rufai Omowunmi Balogun, Sam Khallaghi, Yuting Yao, Sujit Roy, Rahul Ramachandran, et al. 2025. “GELOS: A Benchmark Dataset for Geospatial Exploration of Latent Observation Space.” Poster presented at the AGU Annual Meeting, December 16, 2025.

[40] Andres, Brian. 2026. “Building and Validating LLM-Generated Captions and Embeddings for Geospatial Imagery in Queryable Earth.” Element 84, March 27, 2026.

[41] Carion, Nicolas, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, et al. 2025. “SAM 3: Segment Anything with Concepts.” arXiv preprint arXiv:2511.16719. doi:10.48550/arXiv.2511.16719.

[42] Siméoni, Oriane, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, et al. 2025. “DINOv3.” arXiv preprint arXiv:2508.10104. doi:10.48550/arXiv.2508.10104.

[43] Shabbir, Akashah, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. 2025. “GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing.” In *Proceedings of the 42nd International Conference on Machine Learning*, 54095–54111. *Proceedings of Machine Learning Research* 267.

[44] University of Twente. 2025. “Translating Field Photos to Detailed Crop Descriptions: ESA Project Advances AI Technology to Agricultural Assessment.” May 20, 2025.

[45] Microsoft. 2025. “Microsoft and NASA Apply AI Agents to Key Hydrology Data, Deepening Our Understanding of Earth.” December 18, 2025.

[46] Khallaghi, Sam, Anke Fluhrer, Thomas Jagdhuber, Hamed Alemohammad. 2025. “A Contrastive Self-supervised Learning Model for Soil Moisture Retrieval from P-band PolSAR.” Oral presentation presented at the AGU Annual Meeting. December 16, 2025.