
ATLAS: Auditing the Geometry of Geophysical Knowledge in EO Foundation Models

Stanislaw Kiedrzyński^{*1} Michal Nowicki^{*1} Mikolaj Wolinski^{*1} Bartlomiej Sobieski^{1,2}
Hubert Baniecki^{1,2} Przemyslaw Biecek^{1,2} Agata M. Wijata^{2,3}

¹Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

²Centre for Credible Artificial Intelligence (CCAI), Warsaw University of Technology, Warsaw, Poland

³Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland

Abstract Earth-observation (EO) foundation models increasingly underpin downstream tasks, yet they are evaluated primarily through task-level performance and remain largely black boxes. As a result, we still have limited understanding of what representations they encode, how their behaviour can be explained, and whether their internal knowledge can be reliably steered. We introduce *ATLAS*: Affine chArTs of LAtent Space, an architecture-agnostic probe that learns an invertible affine map, consisting of an orthogonal rotation, a translation, and a sparsity-regularized diagonal scaling, to project hidden states onto a two-dimensional, visually interpretable feature plane. Demonstrated on TerraMind, ATLAS isolates planes for longitude, latitude, altitude, and year across encoder depth and reveals three consistent regularities: geophysical knowledge organizes hierarchically with depth; periodic features become circular in deep layers; and latent-traversal cost tracks physical geography. Editing only the altitude plane and inverting raises a generated height map by about 500 m while preserving terrain shape—a causal effect confirmed by Shapley-value attribution. Feature planes thus provide a machine-readable, FAIR (Findable, Accessible, Interoperable, and Reusable)-aligned account of what a model encodes, how strongly it encodes it, and how it can be steered, information that current EO model cards and benchmarks do not capture.

1 Motivation

Self- and unsupervised pre-training has produced a generation of Geospatial Foundation Models (GFM)—SatMAE (Cong et al., 2022), Scale-MAE (Reed et al., 2023), Prithvi (Jakubik et al., 2023; Szwarcman et al., 2026), and the multimodal any-to-any model TerraMind (Jakubik et al., 2025)—that now underpin high-stakes EO tasks such as flood mapping and wildfire-scar segmentation. As these models become the shared substrate for downstream benchmarks, ensuring trustworthy and reproducible evaluation and providing FAIR, machine-readable model descriptions have become central challenges for the EO community. However, current practice judges a GFM almost entirely by task accuracy and otherwise treats it as a black box (Liu et al., 2025); a model card records the training data and the score, but says nothing about what geophysical structure the model has internalized (Zhu et al., 2026). For Large Language Models (LLMs) this gap has been closed from the inside: probing and representation-geometry studies show real-world attributes emerging as structured directions and manifolds—space and time as linear features (Gurnee and Tegmark, 2024), and periodic quantities such as days and months forming *circular* features (Engels et al., 2024). Comparable analyses are almost absent for specialized GFMs, making a principled characterization of a GFM’s internal knowledge impossible. Explainable Artificial Intelligence (XAI) for EO, meanwhile, has centered on per-prediction saliency and attribution maps (Selvaraju

^{*}These authors contributed equally.

GeoEx: extracting interpretable feature planes from a geospatial foundation model
an interpretable 2-D feature plane — not a single neuron — is the atomic unit of analysis

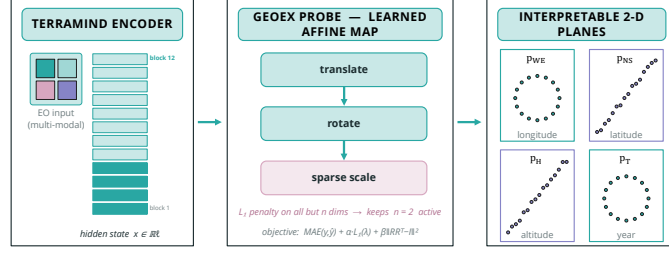


Figure 1: ATLAS as a probe. A hidden state $x \in \mathbb{R}^\ell$ is mapped by a learned affine transform—rotation R , translation t , and sparse diagonal scale $\text{Diag}(\lambda)$ retaining $n=2$ directions—onto a two-dimensional feature plane (p_{WE} , p_{NS} , p_H , p_T).

et al., 2017; Kakogeorgiou and Karantzas, 2021; Höhl et al., 2024), not on the structure a model internalizes—the data-to-model-centric shift urged by Model Science (Biecek and Samek, 2025; Biecek et al., 2026), whose verify—explain-control pillars ATLAS instantiates for GFM.

We close this gap with three contributions. (i) *ATLAS*, a probe that extracts two-dimensional feature planes from a GFM latent space (Section 2); the plane, not a single neuron, is our atomic unit of analysis (Fig. 1). (ii) Demonstrated on TerraMind, ATLAS yields three auditable regularities—hierarchy with depth, emergent circularity, and latent-traversal costs that respect physical barriers (Section 3); whether these hold across GFMs is the open question we foreground for follow-up. (iii) Exploiting invertibility, we turn reading into steering with a causal intervention on altitude (Section 3), and argue that feature planes provide a structured representation-level description of what EO foundation models encode, information that current model cards and benchmarks do not capture (Section 4).

2 The ATLAS Probe

ATLAS operates on frozen latent activations and can be applied without modifying the underlying GFM. Standard probing trains an auxiliary predictor to recover geophysical targets, but probe capacity is a dilemma (Belinkov, 2022): an expressive Multilayer Perceptron (MLP) can recover information beyond the representation used by the model, while a linear probe misses structured geometry such as multi-dimensional or circular topologies. *ATLAS* strikes a compromise by restricting the probe to invertible affine transformations of the latent space. Rather than constructing new features, it can only reorient, translate, and selectively scale directions already present in the representation, providing a more faithful view of the information encoded by the model. Given a latent representation $\mathbf{x}$, *ATLAS* learns an orthogonal rotation R , a translation $\mathbf{t}$, and a diagonal scale matrix $\text{Diag}(\lambda)$ to obtain

$$\mathbf{z} = \text{Diag}(\lambda) R (\mathbf{x} - \mathbf{t}), \quad (1)$$

from which a prediction $\hat{y}$ is produced. The parameters are trained with

$$\mathcal{L}(y, \hat{y}) = \text{MAE}(y, \hat{y}) + \alpha \mathcal{L}_{1,n}(\lambda) + \beta \|RR^\top - I\|_2^2, \quad (2)$$

where MAE denotes mean absolute error, $\mathcal{L}_{1,n}$ applies an L_1 penalty to all but the n largest entries of λ , and $\|RR^\top - I\|_2^2$ enforces orthogonality. With $n=2$, the tuple $(R, \mathbf{t}, \lambda)$ defines a projection onto a 2-D plane that can be visually inspected (Fig. 1). We extract planes for longitude (p_{WE}), latitude (p_{NS}), altitude (p_H), and year (p_T) across depth; longitude and time of year use sine/cosine

Geophysical knowledge is organised hierarchically — and becomes circular with depth

schematic — illustrates the reported trend; axes are qualitative, not measured values

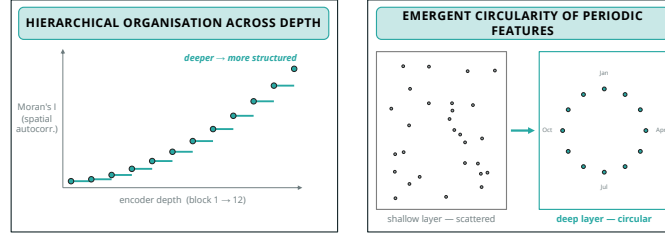


Figure 2: Knowledge is organized hierarchically and becomes circular with depth. Left: Moran’s I increases with depth. Right: a periodic feature (year, longitude) evolves from a scattered representation to a circular one. Panels illustrate observed trends.

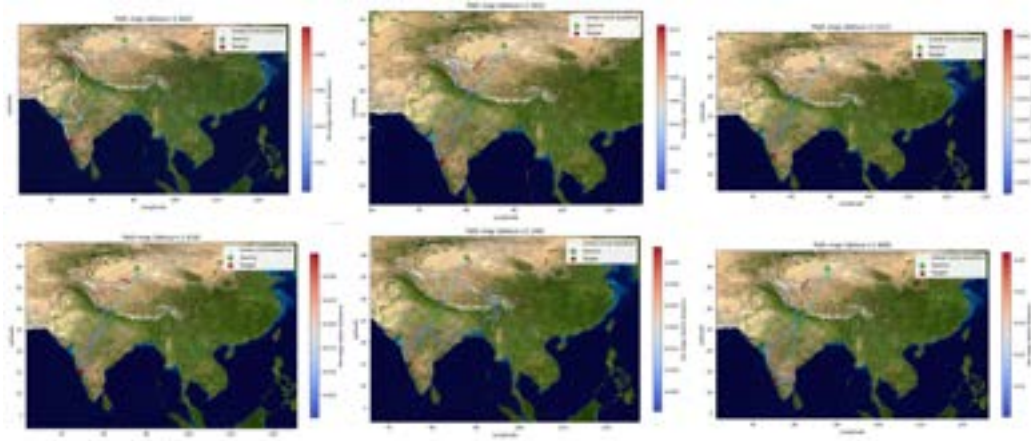


Figure 3: Shortest paths in TerraMind’s deepest latent space over South and Central Asia for six source–target pairs (green/red stars; dashed cyan: great-circle baseline). Edge colors denote latent cosine distance (blue low, red high): costs spike across the Himalayas and Tibetan Plateau, indicating that latent geometry reflects physical geography rather than coordinates.

encodings, and p_T is fit on Northern-Hemisphere data to avoid conflating opposite seasons. Because $\text{Diag}(\lambda)$, R , and t are invertible, a projection can be edited and mapped back to the full latent space, a property exploited in Section 3.

3 What TerraMind Encodes—and How to Steer It

Applying ATLAS to TerraMind reveals three regularities, summarized in Table 1; Fig. 2 schematically illustrates the observed trends across encoder depth. *Hierarchical organization with depth*: spatial autocorrelation of the planes, measured by Moran’s I (Moran, 1950), rises with encoder depth—the model does not store a static map but iteratively *organizes* its knowledge of Earth. *Emergent circularity*: periodic features collapse from a scattered shallow representation into a *circular* deep-layer one, quantified by a circular-correlation coefficient (Jammalamadaka and Sengupta, 2001); this mirrors the circular features found for days and months in LLMs (Engels et al., 2024). *Physical barriers*: on a hybrid geo-latent graph (geographic gating plus a cosine-distance threshold), paths are cheap within homogeneous regions but incur a sharp latent-distance spike where the route crosses a barrier such as the Himalayas (Fig. 3)—the geometry encodes physical geography, not merely coordinates.

From reading to steering. Reading structure off a probe leaves open whether the model *uses* it. We tested this on the altitude plane. Because the pre-projection operations are invertible, we compute

Table 1: ATLAS turns TerraMind’s latent space into an auditable set of feature planes—each a candidate field for a machine-readable EO model card (Section 4).

Plane	Feature	Encoding	Deep-layer geometry
p_{WE}	longitude	sin/cos	circular
p_{NS}	latitude	linear	linear / clustered
p_H	altitude	linear	linear, steerable
p_T	year	sin/cos (N. hem.)	circular

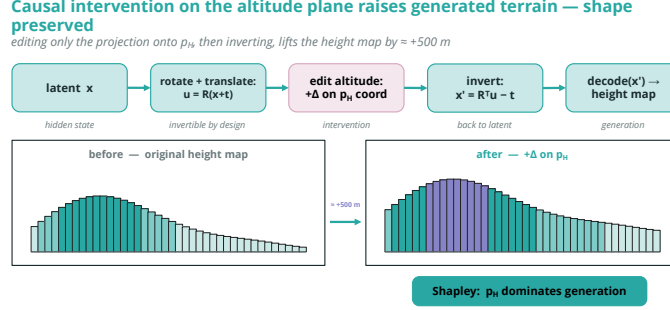


Figure 4: Causal intervention on the altitude plane. Editing only p_H raises the decoded height map by ≈ 500 m while preserving terrain shape; Shapley attribution confirms its role in generation.

the projection onto p_H , add an offset to its active coordinate, and map the edited representation back to the latent space before decoding (Fig. 4). The generated topographic height map increases by roughly 500 m while the shape of the terrain is preserved, and a Shapley-value analysis (Lundberg and Lee, 2017) confirms that p_H drives the generation of the height map; as attribution methods can be fragile, we treat this as a validation check (Adebayo et al., 2018), not as a faith statement. The intervention demonstrates controllability; it demonstrates *controllability*—assessable only from inside the model, and exactly what a trustworthy evaluation protocol for generative GFM should report.

4 Feature Planes as an Audit Layer, and Open Questions

An audit layer. ATLAS’s artifacts should outlive the run that produced them, so FAIR, machine-readable metadata is part of the method—linking it to trustworthy EO evaluation and to model and dataset cards (Mitchell et al., 2019; Gebru et al., 2021). We propose a *feature-plane audit* extending an EO model card with, for each geophysical attribute, the plane’s encoding, the depth at which it becomes organised (Moran’s I) or circular (circular correlation), and its intervention sensitivity (decoded change per unit edit). Table 1 is a minimal instance: machine-readable fields a benchmark repository can accumulate and an agentic AutoML pipeline can read and write.

Open questions. Three directions follow, each also a reproducibility requirement for the audit. *Baselines and statistics:* ATLAS must be compared against linear and MLP probes on identical data, with multi-seed confidence intervals on the metric curves and a quantified intervention (with a random-plane control), so “interpretable” and “steerable” become numbers. *Cross-model replication*—the key next step: repeating the pipeline on other GFM (Prithvi, SatMAE, Scale-MAE) tests whether hierarchy and circularity are properties of the model *class* or artifacts of TerraMind. *Search efficiency:* scoring these properties is costlier than scoring accuracy, inviting surrogate and multi-fidelity strategies in an AutoML loop. Our position: for EO foundation models, *what a model encodes and whether it can be steered* is measurable, and belongs inside trustworthy evaluation rather than beside it—turning a GFM’s latent geometry into an audit layer for model cards and benchmarks.

Acknowledgements. Work on this project is financially supported by the Foundation for Polish Science (FNP) grant ‘Centre for Credible AI’ No. FENG.02.01-IP.05-0058/24.

References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Neural Information Processing Systems*.
- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Biecek, P., Longo, L., Zhou, J., Fel, T., Holzinger, A., and Samek, W. (2026). The case for model science: Verify, explore, steer, refine.
- Biecek, P. and Samek, W. (2025). Model science: Getting serious about verification, explanation and control of AI systems. In *Proc. 24th European Conf. on Artificial Intelligence (ECAI 2025)*, pages 46–53. IOS Press.
- Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., and Ermon, S. (2022). SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. (2024). Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):86–92.
- Gurnee, W. and Tegmark, M. (2024). Language models represent space and time. In *International Conference on Learning Representations (ICLR)*.
- Höhl, A., Obadic, I., Fernández-Torres, M.-, Najjar, H., Oliveira, D. A. B., Akata, Z., Dengel, A., and Zhu, X. X. (2024). Opening the black box: A systematic review on explainable artificial intelligence in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 12(4):261–304.
- Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., and Ramachandran, R. (2023). Foundation models for generalist geospatial artificial intelligence.
- Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., Ramachandran, R., Fraccaro, P., Brunschweiler, T., Cavallaro, G., Bernabe-Moreno, J., and Longépé, N. (2025). TerraMind: Large-scale generative multimodality for Earth observation. *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Jammalamadaka, S. and Sengupta, A. (2001). *Topics in Circular Statistics*.
- Kakogeorgiou, I. and Karantzas, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103:102520.
- Liu, Q., Chen, Z., and Ma, J. (2025). Foundation models for exploration geophysics. *Geophysics*, 90(6):H29–H52.

- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.
- Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., and Darrell, T. (2023). Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4065–4076.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.
- Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, E., Blumenstiel, B., Ghosal, R., de Oliveira, P. H., de Sousa Almeida, J. L., Sedona, R., Kang, Y., Chakraborty, S., Wang, S., Gomes, C., Kumar, A., Gaur, V., Truong, M., Godwin, D., Khallaghi, S., Lee, H., Hsu, C.-Y., Asanjan, A. A., Mujeci, B., Shidham, D., Balogun, R. O., Kolluru, V., Keenan, T., Arevalo, P., Li, W., Alemohammad, H., Olofsson, P., Mayer, T., Hain, C., Kennedy, R., Zadrozny, B., Bell, D., Cavallaro, G., Watson, C., Maskey, M., Ramachandran, R., and Moreno, J. B. (2026). Prithvi-EO-2.0: A versatile multitemporal foundation model for Earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20.
- Zhu, X. X., Xiong, Z., Wang, Y., Stewart, A. J., Heidler, K., Wang, Y., Yuan, Z., Dujardin, T., Xu, Q., and Shi, Y. (2026). On the foundations of Earth foundation models. *Communications Earth & Environment*, 7(1):103.