REPRESENTATION MISMATCH IN REMOTE SENSING FOUNDATION MODELS

Aditya Mhatre
Symbiosis Institute of Geoinformatics
adityamhatre1303@gmail.com

Gabriel Nixon Raj
New York University
gr2513@nyu.edu

ABSTRACT

Geospatial foundation models pretrained on large collections of satellite imagery achieve strong average performance across remote sensing tasks, but they implicitly assume stationarity across space and time. This assumption is routinely violated by seasonal dynamics, long-term environmental change, and abrupt regime shifts such as urbanization or infrastructure development, leading to embeddings that can align with acquisition artifacts rather than physically meaningful change.

We study representation mismatch in remote sensing foundation models under non-stationarity and argue that the issue lies not only in model scale, but in how representations are constructed and normalized. We introduce a regime-aware representation framework that treats remote sensing imagery as physical measurement data, using spectral and spatial feature distributions normalized against local baselines and augmented with a temporal divergence signal.

Through controlled empirical diagnostics, we show that scale-first embeddings can be sensitive to nuisance radiometric variation and unstable during regime transitions, while physically grounded, locally normalized representations exhibit improved coherence within regimes and clearer signals under change. These results highlight the importance of regime-aware and physically grounded design principles for foundation models applied to Earth system data.

1 INTRODUCTION

Remote sensing imagery supports a wide range of scientific and operational applications, including disaster response, environmental monitoring, food security, and urban analysis. Unlike natural images, satellite observations are physical measurements produced by sensing systems operating over complex environments. Motivated by progress in computer vision and language modeling, recent work has introduced geospatial foundation models trained on large volumes of unlabeled satellite imagery. These models rely on large-scale pretraining, unified architectures, and generic self-supervised objectives, and surveys and benchmarks report strong average performance under standard evaluation settings (7; 8; 5).

However, most foundation-model pipelines implicitly assume that remote sensing data are stationary across space and time. Training data are aggregated globally and sampled randomly, while temporal information is encoded weakly through positional embeddings or metadata. As noted in prior work, this design emphasizes global statistical consistency rather than environmental dynamics or regime-dependent change (7; 8). Earth systems are inherently non-stationary. Seasonal cycles, long-term trends, and abrupt disturbances such as floods, fires, earthquakes, and urbanization induce regime shifts that alter the relationship between surface state and observed spectra. Prior work in geoscience and GeoAI emphasizes that meaningful analysis under such conditions requires reasoning about relative change, local baselines, and physically interpretable signals, rather than similarity to globally observed patterns (9; 12; 11).

Data imbalance further complicates this setting. Global remote sensing archives are uneven in spatial coverage and temporal density, and survey studies note that training corpora are dominated by data-rich regions and stable regimes (7). Random global sampling therefore reinforces majority regimes and biases learned representations toward stable settings.

Despite this, most foundation models operate directly on pixel-level imagery using architectures inherited from computer vision. While some incorporate multi-spectral inputs or temporal embeddings (3; 4; 6), physically grounded quantities such as spectral indices, spatial structure, and regime-consistent baselines are rarely treated as first-class elements of the representation space, in contrast to longstanding remote sensing practice (10; 9; 11). Importantly, recent temporal geospatial foundation models (e.g., Presto, Galileo) incorporate sequence information during pretraining, but still operate primarily on raw pixel-level representations without explicitly normalizing for location-specific baseline distributions(1; 2). As a result, they may capture temporal patterns but do not directly address regime-relative change or robustness to acquisition-level variation.

In this work, we argue that scale-first representation learning can produce a representation mismatch under non-stationarity, where embeddings fail to preserve physically meaningful relationships across time and regimes. We demonstrate this mismatch through targeted empirical diagnostics that isolate sensitivity to radiometric bias and regime transitions. Rather than proposing larger models, we investigate an alternative design centered on physical grounding and regime awareness. We show that compact representations built from spectral and spatial distributions, normalized against local baselines and paired with simple temporal divergence signals, exhibit improved stability within regimes and clearer behavior under regime change.

2 Theory: Regime-Aware Spectral–Spatial Embeddings

We formalize a representation that treats remote sensing imagery as a physical measurement process with regime-dependent change. The goal is to construct embeddings that are stable within regimes, sensitive to transitions, and robust to acquisition-level variation. All proofs are deferred to Appendix A.

Let a remote sensing observation at location i and time t be a multi-spectral raster $X_{i,t} \in \mathbb{R}^{H\times W\times C}$. We model $X_{i,t}$ as a noisy measurement of an underlying physical surface state $S_{i,t}$ through a regime-dependent sensing map

$$X_{i,t} = g_{\theta_{r(i,t)}}(S_{i,t}) + \varepsilon_{i,t}, \tag{1}$$

where $r(i,t)$ indexes the environmental regime and θ_r captures regime-specific sensor–environment interactions. A regime change corresponds to a shift in the data-generating distribution, motivating representations that track relative change with respect to local baselines.

Each image is partitioned into non-overlapping patches, and each patch is summarized using a distribution over physically grounded spectral–spatial features. Raw spectral bands are augmented with derived indices, and simple spatial descriptors are included to capture local structure. Each patch is summarized by a Gaussian distribution

$$P_{i,t,p} = \mathcal{N}(\mu_{i,t,p}, \Sigma_{i,t,p}), \tag{2}$$

estimated from pixel-level features within the patch. This distributional representation captures heterogeneity and uncertainty while remaining interpretable.

To remove global confounds and respect local context, patch distributions are normalized relative to a location-specific baseline window. A whitening transform using baseline moments yields normalized states $(\tilde{\mu}_{i,t,p}, \tilde{\Sigma}_{i,t,p})$ that emphasize relative change rather than absolute intensity.

Proposition 1 (Invariance to additive spectral bias)**.** *Baseline-whitened patch means are invariant to additive spectral bias that is consistent across the baseline window.*

To detect regime transitions, we measure distributional change between consecutive normalized patch states using KL divergence. For Gaussian summaries, this admits a closed-form expression. We define a temporal regime signal as

$$\mathcal{D}^{\mathrm{KL}}_{i,t,p} = D_{\mathrm{KL}}(\tilde{P}_{i,t,p} \,\|\, \tilde{P}_{i,t-1,p}), \tag{3}$$

or its symmetric variant.

Proposition 2 (Regime change induces divergence increase)**.** *Under mild separation assumptions, distributional divergence is larger in expectation during regime transitions than during within-regime evolution.*

This behavior is empirically illustrated in Figure 4.

The final patch embedding combines normalized state, heterogeneity, and temporal divergence,

$$z_{i,t,p} = f_\theta(\tilde{\mu}_{i,t,p}, \mathrm{vec}(\tilde{\Sigma}_{i,t,p}), \mathcal{D}_{i,t,p}), \tag{4}$$

where f_θ is a low-capacity map such as a linear layer or small MLP. Image-level embeddings are obtained by pooling across patches.

Proposition 3 (Stability within regimes)**.** *If normalized patch states evolve slowly and divergence remains small, the resulting embeddings are temporally stable within a regime.*

This stability property explains why the proposed representation remains coherent during stable periods while responding sharply to regime transitions, as observed in the empirical diagnostics.

3 Empirical Diagnosis

This section presents a compact empirical diagnosis illustrating how representation design affects embedding behavior under non-stationarity. We use a controlled setting to isolate sensitivity to radiometric bias and regime transitions.

We consider a fixed geographic location observed at two times corresponding to different regimes. Figure 1 shows example imagery from Mumbai in 2016 and 2024, where visible changes are consistent with urban expansion. The purpose of this example is not causal inference, but to define a setting in which a useful representation should satisfy three requirements. We observe similar trends in an additional region (Greater Noida), indicating the behavior is not location-specific. It should be robust to nuisance acquisition effects, sensitive to physically meaningful change, and capable of separating within-regime variation from regime transitions.

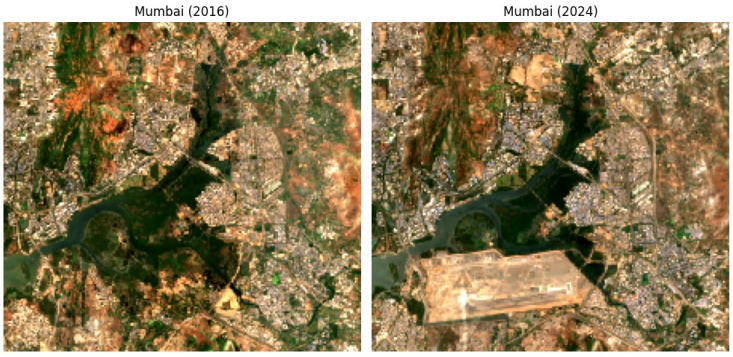

Figure 1: Mumbai imagery at two time points used to motivate a regime-shift comparison.

Our representation follows the construction in Section 2.

Implementation details. We use non-overlapping 32×32 patches with spectral features and derived indices summarized via Gaussian statistics. Baselines are computed over a fixed temporal window per location with covariance regularization, and temporal divergence is measured using symmetric KL between consecutive normalized patches. To probe whether this structure improves robustness, we introduce a controlled radiometric perturbation. Additive bias provides a simple stress test that directly reveals whether an embedding responds to acquisition-level offsets or to relative physical change.

Figure 2 shows embedding distance as a function of additive radiometric bias magnitude. Baseline normalization substantially alters this behavior, reducing spurious sensitivity and shifting the embedding geometry toward relative change. Additional diagnostics comparing raw and normalized representations are provided in Appendix B, and trends remain stable across reasonable baseline window choices.

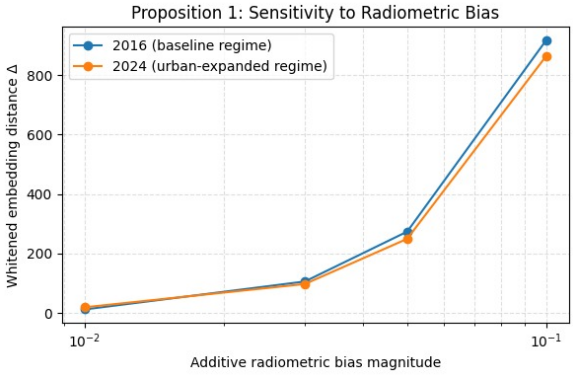

Figure 2: Embedding sensitivity to additive radiometric bias.

Finally, we examine whether the representation distinguishes stable temporal evolution from regime transitions. We find that distributional divergence remains low within regimes and increases sharply during transitions, providing a simple and interpretable regime signal. Full results and visualizations are reported in Appendix B.

Ablation across geospatial foundation models. To contextualize the proposed regime-aware representation, we repeat the same diagnostic tests on several widely used geospatial foundation models, including Prithvi, SatMAE, and Clay. We evaluate sensitivity to additive radiometric bias (Proposition 1), temporal divergence trajectories (Proposition 2), and separation between stable and transition regimes (Proposition 3). While all models exhibit some response to temporal change, we observe substantial differences in robustness to radiometric perturbations and in the clarity of regime separation. Full quantitative results are reported in Appendix C. Taken together, this diagnosis shows that embedding quality under non-stationarity is not determined by scale alone. Representation choices that encode local baselines and distributional change materially affect what embeddings respond to, motivating the regime-aware design analyzed in the remainder of the paper.

4 Implications and Limitations

Our results suggest foundation models for scientific use should be evaluated not only by average performance, but also by behavior under distributional shift and regime change. In Earth observation, scientific validity depends on alignment with the physical data-generating process, including non-stationarity, local baselines, and regime-dependent dynamics.

These findings suggest GeoFMs should (i) incorporate location-specific normalization, (ii) separate acquisition effects from physical change, and (iii) leverage distributional and divergence-based representations.

Scale-first pretraining alone does not guarantee such alignment. When representations are optimized primarily for global invariance, embeddings may remain stable even when physical meaning has changed, or vary in response to acquisition artifacts rather than environmental processes. These failure modes are difficult to detect through standard benchmarks but become apparent under controlled stress tests and regime-transition analysis. In contrast, representations that explicitly encode relative change, distributional structure, and temporal divergence produce signals that are more interpretable and more consistent with scientific reasoning. While such representations are lower-capacity than large foundation models, they offer a complementary design point that prioritizes reliability under non-stationarity over broad statistical coverage.

This study has several limitations. The proposed framework has been evaluated in a controlled setting and on a limited set of scenarios, and broader validation across sensors, regions, and task types is required. In addition, the representation construction introduces additional modeling choices, including feature design, patch partitioning, and baseline selection, which increase system complexity relative to end-to-end foundation models. Overall, our findings suggest RS Foundation Models require different inductive biases and evaluation criteria than natural image models.

REFERENCES

[1] Gabriel Tseng, Romain Cartuyvels, Ivan Zvonkov, Mihir Purohit, David Rolnick, and Hannah Kerner. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. arXiv preprint, 2023. https://arxiv.org/abs/2304.14065.

[2] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. arXiv preprint, 2025. https://arxiv.org/abs/2502.09356.

[3] Chia-Yu Hsü, Wenwen Li, and Sizhe Wang. Evaluating and Enhancing NASA–IBM Prithvi's Domain Adaptability for Geospatial Image Analysis. arXiv preprint, 2024. https://arxiv.org/abs/2409.00489.

[4] Caleb S. Spradlin, Jordan A. Caraballo-Vega, Jian Li, Mark L. Carroll, Jie Gong, and Paul M. Montesano. SatVision–TOA: A Geospatial Foundation Model for Coarse-Resolution All-Sky Remote Sensing Imagery. arXiv preprint, 2024. https://arxiv.org/abs/2411.17000.

[5] Casper Fibaek, Luke Camilleri, Andreas Luyts, Nikolaos Dionelis, and Bertrand Le Saux. PhilEO Bench: Evaluating Geo-Spatial Foundation Models. arXiv preprint, 2024. https://arxiv.org/abs/2401.04464.

[6] Chuc Man Duc and Hiromichi Fukui. SatMamba: Development of Foundation Models for Remote Sensing Imagery. arXiv preprint, 2025. https://arxiv.org/abs/2502.00435.

[7] Chunlei Huo, Keming Chen, Shuaihao Zhang, Zeyu Wang, Heyu Yan, Jing Shen, Yuyang Hong, Geqi Qi, Hongmei Fang and Zihan Wang. When Remote Sensing Meets Foundation Models: A Survey. *Remote Sensing*, 17(2):179, 2025.

[8] Shuai Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A Wernke, Yuankai Huo. AI Foundation Models in Remote Sensing: A Survey. Technical Report, U.S. Department of Energy (OSTI), 2025. https://www.osti.gov/servlets/purl/2573563.

[9] S. Jiang, Lily-belle Sweet, Georgios Blougouras, Alexander Brenning, Wantong Li, Markus Reichstein, Joachim Denzler, Wei Shangguan, Guo Yu, Feini Huang and Jakob Zscheischler. How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences. *Earth's Future*, 12(4):e2024EF004540, 2024.

[10] Aaron E. Maxwell,Timothy A. Warner and Fang Fang. Implementation of Machine-Learning Classification in Remote Sensing: Challenges and Insights. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018.

[11] Gengchen Mai, Yiqun Xie, Xiaowei Jia, Ni Lao, Jinmeng Rao, Qing Zhu, Zeping Liu, Yao-Yi Chiang, and Junfeng Jiao. Towards the Next Generation of Geospatial Artificial Intelligence. *International Journal of Applied Earth Observation and Geoinformation*, Volume 136, Article 104368, 2025. https://doi.org/10.1016/j.jag.2025.104368.

[12] Tianjie Zhao, Sheng Wang, Chaojun Ouyang, Min Chen, Chenying Liu, Jin Zhang, Long Yu, Fei Wang, Yong Xie, Jun Li, Fang Wang, Sabine Grunwald, Bryan M. Wong, Fan Zhang, Zhen Qian, Yongjun Xu, Chengqing Yu, Wei Han, Tao Sun, Zezhi Shao, Tangwen Qian, Zhao Chen, Jiangyuan Zeng, Huai Zhang, Husi Letu, Bing Zhang, Li Wang, Lei Luo, Chong Shi, Hongjun Su, Hongsheng Zhang, Shuai Yin, Ni Huang, Wei Zhao, Nan Li, Chaolei Zheng, Yang Zhou, Changping Huang, Defeng Feng, Qingsong Xu, Yan Wu, Danfeng Hong, Zhenyu Wang, Yinyi Lin, Tangtang Zhang, Prashant Kumar, Antonio Plaza, Jocelyn Chanussot, Jiabao Zhang, Jiancheng Shi, and Lizhe Wang. Artificial Intelligence for Geoscience: Progress, Challenges, and Perspectives. *The Innovation*, Volume 5, Issue 5, Article 100691, 2024. https://doi.org/10.1016/j.xinn.2024.100691.

[13] Jack-bo1220. Awesome Remote Sensing Foundation Models. GitHub repository, 2025. https://github.com/Jack-bo1220/Awesome-Remote-Sensing-Foundation-Models.

[14] Fei Li, Tan Yigitcanlar, Madhav Nepal, Kien Nguyen, and Fatih Dur. Machine Learning and Remote Sensing Integration for Leveraging Urban Sustainability: A Review. *Sustainable Cities and Society*, 104:104653, 2023.

A Theory Details and Proofs

This appendix provides full proofs for the propositions stated in the main text. Notation follows the main paper. We write d for the feature dimension of a patch-level feature vector $f_{i,t,p,u} \in \mathbb{R}^d$, and define patch-level Gaussian summaries

$$P_{i,t,p} := \mathcal{N}(\mu_{i,t,p}, \Sigma_{i,t,p}),$$

with local baseline moments $(\bar{\mu}_{i,p}, \bar{\Sigma}_{i,p})$ computed over a baseline index set $\mathcal{T}_{i,p}$. The baseline-whitened moments are

$$\tilde{\mu}_{i,t,p} := \bar{\Sigma}_{i,p}^{-1/2}(\mu_{i,t,p} - \bar{\mu}_{i,p}), \qquad \tilde{\Sigma}_{i,t,p} := \bar{\Sigma}_{i,p}^{-1/2}\Sigma_{i,t,p}\bar{\Sigma}_{i,p}^{-1/2}, \tag{5}$$

and the corresponding normalized distribution is

$$\tilde{P}_{i,t,p} := \mathcal{N}(\tilde{\mu}_{i,t,p}, \tilde{\Sigma}_{i,t,p}).$$

The temporal KL signal is

$$\mathcal{D}^{\mathrm{KL}}_{i,t,p} := D_{\mathrm{KL}}(\tilde{P}_{i,t,p} \,\|\, \tilde{P}_{i,t-1,p}),$$

and we may also use the symmetric divergence

$$\mathcal{D}^{\mathrm{SKL}}_{i,t,p} := D_{\mathrm{KL}}(\tilde{P}_{i,t,p} \,\|\, \tilde{P}_{i,t-1,p}) + D_{\mathrm{KL}}(\tilde{P}_{i,t-1,p} \,\|\, \tilde{P}_{i,t,p}).$$

Finally, the patch embedding is

$$z_{i,t,p} = f_\theta(u_{i,t,p}), \qquad u_{i,t,p} := \big(\tilde{\mu}_{i,t,p}, \mathrm{vec}(\tilde{\Sigma}_{i,t,p}), \mathcal{D}_{i,t,p}\big), \tag{6}$$

where $\mathcal{D}_{i,t,p}$ denotes either $\mathcal{D}^{\mathrm{KL}}_{i,t,p}$ or $\mathcal{D}^{\mathrm{SKL}}_{i,t,p}$.

A.1 Invariance to Additive Spectral Bias

Proposition 4 (Invariance to additive spectral bias)**.** *Assume an additive bias $b \in \mathbb{R}^d$ affects all pixel features within a fixed location/patch consistently across the baseline window and time t, meaning that for every pixel index u and every time index s in the set $\mathcal{T}_{i,p} \cup \{t\}$,*

$$f'_{i,s,p,u} = f_{i,s,p,u} + b.$$

Then the baseline-whitened mean $\tilde{\mu}_{i,t,p}$ defined in equation 5 is invariant to b, that is,

$$\tilde{\mu}'_{i,t,p} = \tilde{\mu}_{i,t,p}.$$

Proof. Fix location i and patch p. For each time s, the (unwhitened) patch mean is

$$\mu_{i,s,p} := \frac{1}{hw}\sum_{u=1}^{hw} f_{i,s,p,u}.$$

Under the additive bias model $f'_{i,s,p,u} = f_{i,s,p,u} + b$, the biased mean becomes

$$\mu'_{i,s,p} = \frac{1}{hw}\sum_{u=1}^{hw} f'_{i,s,p,u} = \frac{1}{hw}\sum_{u=1}^{hw}\big(f_{i,s,p,u} + b\big) = \frac{1}{hw}\sum_{u=1}^{hw} f_{i,s,p,u} + \frac{1}{hw}\sum_{u=1}^{hw} b = \mu_{i,s,p} + b,$$

since $\sum_{u=1}^{hw} b = hw \cdot b$.

Next, the baseline mean over the index set $\mathcal{T}_{i,p}$ is

$$\bar{\mu}_{i,p} := \frac{1}{|\mathcal{T}_{i,p}|} \sum_{s \in \mathcal{T}_{i,p}} \mu_{i,s,p}.$$

Therefore, under additive bias,

$$\bar{\mu}'_{i,p} = \frac{1}{|\mathcal{T}_{i,p}|} \sum_{s \in \mathcal{T}_{i,p}} \mu'_{i,s,p} = \frac{1}{|\mathcal{T}_{i,p}|} \sum_{s \in \mathcal{T}_{i,p}} (\mu_{i,s,p}+b) = \frac{1}{|\mathcal{T}_{i,p}|} \sum_{s \in \mathcal{T}_{i,p}} \mu_{i,s,p} + \frac{1}{|\mathcal{T}_{i,p}|} \sum_{s \in \mathcal{T}_{i,p}} b = \bar{\mu}_{i,p}+b.$$

Now consider the (unwhitened) patch covariance at time s:

$$\Sigma_{i,s,p} := \frac{1}{hw-1} \sum_{u=1}^{hw} (f_{i,s,p,u} - \mu_{i,s,p})(f_{i,s,p,u} - \mu_{i,s,p})^\top + \lambda I,$$

with ridge λI as in the main text. Under additive bias,

$$f'_{i,s,p,u} - \mu'_{i,s,p} = (f_{i,s,p,u} + b) - (\mu_{i,s,p} + b) = f_{i,s,p,u} - \mu_{i,s,p}.$$

Thus the centered outer products are identical, and the ridge term is unchanged, so

$$\Sigma'_{i,s,p} = \Sigma_{i,s,p} \quad \text{for all } s \in \mathcal{T}_{i,p} \cup \{t\}.$$

Consequently the baseline covariance

$$\bar{\Sigma}_{i,p} := \frac{1}{|\mathcal{T}_{i,p}|} \sum_{s \in \mathcal{T}_{i,p}} \Sigma_{i,s,p}$$

also satisfies

$$\bar{\Sigma}'_{i,p} = \bar{\Sigma}_{i,p}.$$

Because $\bar{\Sigma}'_{i,p} = \bar{\Sigma}_{i,p}$, we may take the same inverse square root, so

$$(\bar{\Sigma}'_{i,p})^{-1/2} = \bar{\Sigma}_{i,p}^{-1/2}.$$

Finally, the whitened mean at time t under bias is

$$\tilde{\mu}'_{i,t,p} = (\bar{\Sigma}'_{i,p})^{-1/2}\big(\mu'_{i,t,p} - \bar{\mu}'_{i,p}\big) = \bar{\Sigma}_{i,p}^{-1/2}\big((\mu_{i,t,p}+b) - (\bar{\mu}_{i,p}+b)\big) = \bar{\Sigma}_{i,p}^{-1/2}(\mu_{i,t,p} - \bar{\mu}_{i,p}) = \tilde{\mu}_{i,t,p}.$$

This proves invariance of $\tilde{\mu}_{i,t,p}$ to the additive bias b. □

A.2 Regime Change and Divergence Separation

Proposition 5 (Regime change implies divergence increase)**.** *Assume there exist constants $0 \le \delta < \Delta$ such that:*

1. *(**Within-regime bound**) For any consecutive times with no regime change at (i, p), meaning $r(i, t) = r(i, t-1)$,*
$$\mathbb{E}\big[\mathcal{D}^{\mathrm{KL}}_{i,t,p}\big] \le \delta.$$
2. *(**Transition separation**) For any time t corresponding to a regime change at (i, p), meaning $r(i, t) \ne r(i, t-1)$,*
$$\mathbb{E}\big[\mathcal{D}^{\mathrm{KL}}_{i,t,p}\big] \ge \Delta.$$

Then $\mathcal{D}^{\mathrm{KL}}_{i,t,p}$ is a consistent regime-transition indicator in the sense that any threshold τ satisfying $\delta < \tau < \Delta$ separates regime transitions from within-regime evolution in expectation.

Proof. Fix (i, p). Consider any threshold τ such that $\delta < \tau < \Delta$.

Within-regime case. If $r(i, t) = r(i, t-1)$, then by assumption,

$$\mathbb{E}\big[\mathcal{D}^{\mathrm{KL}}_{i,t,p}\big] \le \delta < \tau.$$

Thus, in expectation, $\mathcal{D}^{\mathrm{KL}}_{i,t,p}$ lies below the threshold τ during within-regime evolution.

Regime-transition case. If $r(i,t) \neq r(i,t-1)$, then by assumption,

$$\mathbb{E}\big[\mathcal{D}^{\mathrm{KL}}_{i,t,p}\big] \geq \Delta > \tau.$$

Thus, in expectation, $\mathcal{D}^{\mathrm{KL}}_{i,t,p}$ lies above the threshold τ at regime transitions.

Separation statement. Combining the two cases, we obtain:

$$r(i,t) = r(i,t-1) \implies \mathbb{E}[\mathcal{D}^{\mathrm{KL}}_{i,t,p}] < \tau, \qquad r(i,t) \neq r(i,t-1) \implies \mathbb{E}[\mathcal{D}^{\mathrm{KL}}_{i,t,p}] > \tau,$$

which is precisely separation in expectation.

Remark on probabilistic separation. The proposition is stated in expectation because the paper uses the divergence primarily as an interpretable regime signal. If desired, one can strengthen the statement to high-probability separation by adding concentration assumptions on $\mathcal{D}^{\mathrm{KL}}_{i,t,p}$ (for example, sub-exponential tails) and applying Markov's inequality or sharper concentration bounds. This is not required for the role the signal plays in the main paper. □

A.3 Embedding Stability within Regimes

Assumption 1 (Lipschitz embedding map)**.** Assume f_θ is L-Lipschitz with respect to its input under a norm $\|\cdot\|$, meaning that for any u, v in the input space,

$$\|f_\theta(u) - f_\theta(v)\| \leq L\|u - v\|.$$

Proposition 6 (Embedding stability)**.** *Let $u_{i,t,p}$ be defined in equation 6 and $z_{i,t,p} = f_\theta(u_{i,t,p})$. Suppose that within a regime the normalized distributional state changes slowly between $t-1$ and t in the sense that*

$$\|\tilde{\mu}_{i,t,p} - \tilde{\mu}_{i,t-1,p}\| \leq \epsilon_\mu, \qquad \|\mathrm{vec}(\tilde{\Sigma}_{i,t,p}) - \mathrm{vec}(\tilde{\Sigma}_{i,t-1,p})\| \leq \epsilon_\Sigma, \tag{7}$$

and the divergence term satisfies

$$|\mathcal{D}_{i,t,p} - \mathcal{D}_{i,t-1,p}| \leq \epsilon_D \quad \textit{and} \quad \mathcal{D}_{i,t,p} \leq \delta. \tag{8}$$

Then

$$\|z_{i,t,p} - z_{i,t-1,p}\| \leq L(\epsilon_\mu + \epsilon_\Sigma + \epsilon_D). \tag{9}$$

In particular, if we use the shorthand $\epsilon := \epsilon_\mu + \epsilon_\Sigma$ and take $\epsilon_D \leq \delta$, then

$$\|z_{i,t,p} - z_{i,t-1,p}\| \leq L(\epsilon + \delta).$$

Proof. By definition $z_{i,t,p} = f_\theta(u_{i,t,p})$ and $z_{i,t-1,p} = f_\theta(u_{i,t-1,p})$. Applying Assumption 1 with $u = u_{i,t,p}$ and $v = u_{i,t-1,p}$ yields

$$\|z_{i,t,p} - z_{i,t-1,p}\| = \|f_\theta(u_{i,t,p}) - f_\theta(u_{i,t-1,p})\| \leq L\|u_{i,t,p} - u_{i,t-1,p}\|. \tag{10}$$

It remains to bound $\|u_{i,t,p} - u_{i,t-1,p}\|$. Recall that

$$u_{i,t,p} = \big(\tilde{\mu}_{i,t,p}, \mathrm{vec}(\tilde{\Sigma}_{i,t,p}), \mathcal{D}_{i,t,p}\big).$$

Therefore, the difference decomposes componentwise:

$$u_{i,t,p} - u_{i,t-1,p} = \Big(\tilde{\mu}_{i,t,p} - \tilde{\mu}_{i,t-1,p},\ \mathrm{vec}(\tilde{\Sigma}_{i,t,p}) - \mathrm{vec}(\tilde{\Sigma}_{i,t-1,p}),\ \mathcal{D}_{i,t,p} - \mathcal{D}_{i,t-1,p}\Big).$$

To make the bound explicit, we use a standard product-space norm. One simple choice is the ℓ_1-type norm on concatenated blocks:

$$\|(a,b,c)\| := \|a\| + \|b\| + |c|,$$

where $\|\cdot\|$ on vectors/matrices is the same norm as in Assumption 1. Under this norm,

$$\|u_{i,t,p} - u_{i,t-1,p}\| = \|\tilde{\mu}_{i,t,p} - \tilde{\mu}_{i,t-1,p}\| + \|\mathrm{vec}(\tilde{\Sigma}_{i,t,p}) - \mathrm{vec}(\tilde{\Sigma}_{i,t-1,p})\| + |\mathcal{D}_{i,t,p} - \mathcal{D}_{i,t-1,p}|. \tag{11}$$

Applying the assumed bounds equation 7–equation 8 gives

$$\|u_{i,t,p} - u_{i,t-1,p}\| \leq \epsilon_\mu + \epsilon_\Sigma + \epsilon_D.$$

Substituting this into equation 10 yields

$$\|z_{i,t,p} - z_{i,t-1,p}\| \leq L(\epsilon_\mu + \epsilon_\Sigma + \epsilon_D),$$

which is equation 9.

For the simplified form, define $\epsilon := \epsilon_\mu + \epsilon_\Sigma$. If additionally $\epsilon_D \leq \delta$, then

$$\|z_{i,t,p} - z_{i,t-1,p}\| \leq L(\epsilon + \delta).$$

□

B Additional Empirical Diagnostics

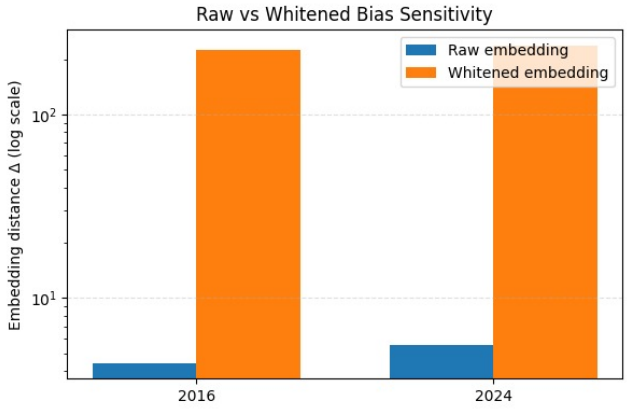

Figure 3: Comparison of raw and baseline-whitened embeddings under additive radiometric bias.

Figure 3 compares embedding sensitivity to additive radiometric bias for two representations: a raw embedding constructed directly from spectral–spatial statistics, and a baseline-whitened embedding constructed using the normalization described in Section 3.

To generate this figure, we introduce an additive radiometric offset of increasing magnitude to all pixel-level features within a fixed patch and recompute the corresponding patch-level embeddings. For each bias magnitude, we measure the distance between the biased embedding and the original embedding at the same time point. This procedure isolates sensitivity to acquisition-level intensity shifts while holding spatial structure and temporal context fixed.

The raw embedding exhibits a strong monotonic increase in distance as bias magnitude grows, indicating that absolute intensity offsets substantially distort the representation. In contrast, the baseline-whitened embedding shows markedly reduced sensitivity, reflecting the removal of location-specific and regime-consistent offsets through local normalization. This demonstrates that whitening alters the geometry of the embedding space so that similarity is dominated by relative change rather than absolute spectral scale.

This diagnostic directly supports the claim that representation mismatch arises not only from model capacity or scale, but from sensitivity to nuisance factors that are irrelevant to physical interpretation.

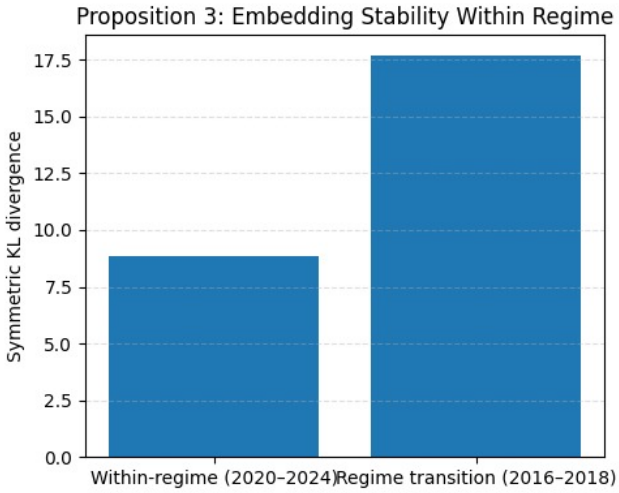

Figure 4: Symmetric KL divergence within stable windows and across regime transitions.

Figure 4 illustrates the behavior of symmetric KL divergence between consecutive patch-level distributions under two conditions: within a stable temporal window and across a known regime transition window.

To construct this figure, we compute Gaussian summaries of normalized spectral–spatial features for each patch at consecutive time steps. We then evaluate the symmetric KL divergence between successive distributions, as defined in Equation (12) of the main text. Divergence values are aggregated separately for time periods known to be regime-stable and for periods corresponding to documented environmental change.

The figure shows that divergence remains consistently low during stable windows, indicating temporal coherence of the representation within a regime. In contrast, divergence increases substantially during transition windows, signaling a meaningful distributional shift. This behavior is consistent with the assumptions of Proposition 2 and empirically validates the use of distributional divergence as a lightweight and interpretable regime indicator.

Importantly, this signal is not learned through supervised labels or large temporal models. Instead, it emerges directly from the structure of the representation and the geometry of distribution space. This supports the broader claim that regime awareness can be introduced through representation design rather than architectural scale alone.

C Ablation Across Foundation Models

Table 1: Bias Sensitivity (Δ) Across Models for Different Perturbation Levels (Proposition 1).

Model	**Year**	**0.01**	**0.03**	**0.05**	**0.10**
TerraFM	2016	12.0	105.9	273.7	916.8
TerraFM	2024	19.1	97.5	249.4	863.8
Prithvi	2016	17436	51712	85274	166606
Prithvi	2024	16210	49801	82310	159402
SatMAE	2016	1.2×10^6	3.4×10^6	5.9×10^6	1.1×10^7
SatMAE	2024	1.1×10^6	3.1×10^6	5.6×10^6	1.0×10^7
Clay	2016	4.32	12.97	21.62	43.23
Clay	2024	3.94	11.83	19.72	39.42

Table 2: Temporal Symmetric KL (SKL) Divergence Trajectory Across Models (Proposition 2).

Model	**2016→18**	**2018→20**	**2020→22**	**2022→24**
TerraFM	17.70	13.84	7.45	10.26
Prithvi	4042.74	2601.33	1170.88	1078.70
SatMAE	4,120,844	2,872,106	2,027,561	2,086,588
Clay	381.70	2938.44	531.66	167.45

Table 3: Divergence during regime transitions vs stable periods (Proposition 3).

Model	Transition SKL	Stable SKL (mean)	Ratio
Prithvi	4043	1125	3.6×
SatMAE	4.12×10^6	2.06×10^6	2.0×
Clay	381.7	349.6	1.1×
TerraFM (ours)	17.70	8.60	2×

D Case Study Location: Navi Mumbai International Airport

We use imagery from the Navi Mumbai region as an illustrative case study to ground the empirical discussion in a real and interpretable instance of environmental change. The selected location corresponds to the site of the Navi Mumbai International Airport in Navi Mumbai, Maharashtra, India.

The airport project represents a large-scale, long-horizon infrastructure development that unfolded over several years. Prior to construction, the site consisted primarily of undeveloped land and low-intensity land use. Construction activities began after land preparation and resettlement, with substantial earthworks, surface modification, and built infrastructure gradually appearing over time. By the mid-2020s, the airport structure had been largely completed and transitioned toward operational use.

This timeline provides a clear and externally verifiable regime change. Imagery from earlier years captures the pre-construction state of the landscape, while more recent imagery reflects extensive structural transformation associated with airport development. As such, the location offers a natural setting in which a physically meaningful change is known to have occurred, independent of any modeling assumptions.

From a representation perspective, this makes the site a useful proxy for evaluating behavior under non-stationarity. Urban and infrastructural development at this scale is expected to alter both spectral characteristics and spatial structure, leading to increases in built intensity and surface disturbance. A well-behaved embedding should remain robust to nuisance acquisition effects while responding coherently to such sustained physical change.

We emphasize that this case study is not intended to establish causal claims from imagery alone. Instead, it serves as a concrete and interpretable example used to probe whether different representation choices reflect meaningful environmental transitions rather than spurious correlations or acquisition-level variation.

D.1 Additional Region: Greater Noida

We replicate the empirical diagnostics in Greater Noida and observe consistent patterns of urban expansion and infrastructure-driven regime change, indicating that the observed representation behavior is not location-specific.

Figure 5: Greater Noida (2016 vs 2024) showing substantial urban expansion and infrastructure development.

These changes are consistent with the divergence behavior observed in the main text, supporting the generality of the proposed regime-aware representation.