WHAT BREAKS IN PRACTICE? A FAILURE TAXONOMY FOR FOUNDATION MODELS IN REMOTE SENSING

Mahule Roy
University of Oxford & Harvard Medical School
mroy25@bwh.harvard.edu

Subhas Roy
TATA Consumer Products Limited

ABSTRACT

Foundation models (FMs) pre-trained on large-scale satellite imagery promise transformative advances in remote sensing (RS). However, their real-world deployment often exposes failure patterns invisible in standard benchmarks. We introduce a failure taxonomy for RS FMs, validated through experiments on three models (SatMAE, Prithvi, SeCo) across five stress tests in Brazilian biomes. Results show performance drops of 12.4-34.2% under operational conditions, with cloud occlusion causing the most severe degradation (26.4% average drop). Confidence miscalibration increases significantly (ECE up to 0.28), highlighting overconfidence during failures. Our taxonomy provides a practical diagnostic framework for deployment testing, bridging the gap between research and reliable application.

1 INTRODUCTION

Foundation models for remote sensing show remarkable benchmark performance (1) but frequently fail in operational deployment (2). While most research focuses on aggregate accuracy metrics, there is limited understanding of real-world failure modes and the conditions under which these models underperform. **Brazil provides an ideal natural laboratory** for studying FM failures due to its extreme environmental diversity—including the Amazon rainforest, Cerrado savanna, and Pantanal wetlands—persistent cloud cover, pronounced seasonal variability, and rapid urbanization. These factors pose representative challenges for optical remote sensing systems and reflect failure modes likely to occur globally across tropical, temperate, and urban regions. To investigate these issues, we select three representative foundation models—SatMAE (transformer-based), Prithvi (CNN-temporal), and SeCo (seasonal contrastive)—covering major architectural paradigms and learning strategies. Our work addresses a critical gap by introducing and experimentally validating a *failure taxonomy*, defined as a structured mapping between (i) root causes, (ii) observable error patterns, and (iii) operational impacts. This taxonomy enables systematic evaluation of robustness, highlights vulnerabilities such as cloud occlusion and spatial generalization, and provides actionable insights for deployment. By bridging the gap between benchmark performance and operational reliability, our study offers a framework for stress-tested, uncertainty-aware remote sensing model evaluation.

2 EXPERIMENTAL METHODOLOGY

We evaluate three public FMs representing distinct architectural approaches: SatMAE (Vision Transformer) (3) learns spatial-spectral features via masked autoencoding, Prithvi (CNN-temporal) (4) focuses on temporal consistency, and SeCo (seasonal contrastive learning) (5) leverages seasonal invariance. Experiments ran on NVIDIA A100 GPU (40GB) with batch sizes optimized per architecture: 16 (SatMAE), 8 (Prithvi), 32 (SeCo). Random Forest baseline used 100 trees with spectral indices (NDVI, NDWI, NDBI) as features (6). Five targeted datasets isolate failure modes: (1)

Sensor shift: 100 Sentinel-2/Landsat-8 pairs from Amazon (2023); (2) **Cloud occlusion**: 150 Amazon images with 0-70% natural cloud cover (2022-23); (3) **Temporal drift**: 200 Cerrado images across dry/wet seasons; (4) **Spatial generalization**: 150 Pantanal wetland images excluded from pre-training; (5) **Label ambiguity**: 100 São Paulo urban images with mixed pixels. Sample sizes provide 80% power to detect effects $\geq$0.5 (α=0.05) (7). Preprocessing includes atmospheric correction (8), normalization, and cloud masking (s2cloudless, threshold=0.5) (9). Ground truth uses an 8-class system with Cohen's Kappa = 0.87. Statistical analysis employs paired t-tests with Bonferroni correction (15 comparisons, α=0.0033), 95% confidence intervals via bootstrapping (1000 resamples), and 5-fold cross-validation (12). We note that our stress test datasets are relatively small (100-200 images per scenario), while sufficient to illustrate key failure modes, reported effect sizes and accuracy drops are indicative and may not generalize across all regions, sensors, or biomes.

3 Experimental Results

Baseline accuracies under optimal conditions were 86.2% (SatMAE), 84.1% (Prithvi), and 82.3% (SeCo). Table 1 reports performance degradation under stress tests with 95% confidence intervals. While fine-tuning may reduce some failures, many deployment scenarios rely on zero-shot or lightly adapted models, making these failure modes operationally relevant.

Table 1: Performance Degradation Under Stress Tests (Accuracy Drop %)

Stress Test	**SatMAE**	**Prithvi**	**SeCo**	**Average**
Sensor Shift	14.3 ± 1.8	12.4 ± 1.9	16.8 ± 1.7	14.5
Cloud Occlusion	26.4 ± 2.2	18.7 ± 2.4	34.2 ± 2.1	26.4
Temporal Drift	19.2 ± 2.0	10.8 ± 1.8	22.5 ± 1.9	17.5
Spatial Generalization	22.8 ± 2.1	18.3 ± 2.3	27.4 ± 2.0	22.8
Label Ambiguity	15.1 ± 1.6	8.4 ± 1.7	18.2 ± 1.5	13.9

Key findings: Cloud occlusion causes the largest degradation (26.4% average), followed by spatial generalization (22.8%). SeCo is most sensitive to clouds (34.2%), while Prithvi is temporally robust (10.8% drift). Inference times are 1.8s (SatMAE), 1.2s (Prithvi), 0.9s (SeCo) per 256$\times$256 patch with 10 spectral bands; larger or full-resolution images may increase computation. Paired t-tests and Cohen's d indicate effect magnitude, but small sample sizes (100–200 images) and stochastic predictions may overestimate significance. Performance drops and confidence intervals are illustrative, highlighting relative model sensitivity rather than absolute operational performance.

3.1 Confidence Calibration and Error Analysis

Confidence calibration degrades under stress, with ECE rising from 0.08–0.15 (baseline) to 0.24–0.28 under cloud occlusion and $\sim$0.19–0.21 for spatial generalization. While foundation models outperform Random Forest in clean settings (+10.4% accuracy), they degrade more under stress (26.4% vs 15.2%) and show worse calibration (ECE +0.13 vs +0.05). Errors follow consistent patterns across conditions (e.g., vegetation–water, wetland–water, urban–vegetation), with forest and wetland classes most affected (F1 drops 35–40%), and performance declines sharply when cloud cover exceeds 50% ($>$30% drop).

4 Conclusion

Remote sensing foundation models, while strong on benchmarks, exhibit substantial performance degradation under real-world conditions, with accuracy drops of 12.4–34.2% and ECE up to 0.28. We introduce a five-category failure taxonomy—**sensor shift, cloud occlusion, temporal drift, spatial generalization, and label ambiguity**—that systematically identifies key vulnerabilities, with cloud occlusion, spatial generalization, and overconfident predictions emerging as dominant risks. This framework enables more rigorous stress-tested evaluation and supports robust, uncertainty-aware deployment in operational settings.

REFERENCES

[1] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[2] Mania, H., Miller, J., Schmidt, L., Hardt, M., & Recht, B. (2019). Model similarity mitigates test set overuse. Advances in Neural Information Processing Systems, 32.

[3] Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., ... & Ermon, S. (2022). Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems, 35, 197-211.

[4] Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., ... & Ramachandran, R. (2023). Foundation models for generalist geospatial artificial intelligence. arXiv preprint arXiv:2310.18660.

[5] Manas, O., Lacoste, A., Giró-i-Nieto, X., Vazquez, D., & Rodriguez, P. (2021). Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9414-9423).

[6] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[7] Cohen, J. (2013). Statistical power analysis for the behavioral sciences. routledge.

[8] Richter, R., & Schläpfer, D. (2011). Atmospheric/topographic correction for airborne imagery. ATCOR-4 user guide, 565-02.

[9] Zupanc, A. (2017). Improving cloud detection with machine learning. Accessed: Oct, 10, 2019.

[10] Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American statistical association, 56(293), 52-64.

[11] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics: Methodology and distribution (pp. 569-593). New York, NY: Springer New York.

[12] Browne, M. W. (2000). Cross-validation methods. Journal of mathematical psychology, 44(1), 108-132.