**International Journal of Information Technology and Computer Science Applications (IJITCSA)**
**p-ISSN: 2964-3139 e-ISSN: 2985-5330**
Vol. 04, No. 02, page 154 - 164
Submitted 27/4/2026; Accepted 26/7/2026; Published 27/7/2026

# Modality-Resilient Multimodal Earth-Observation Foundation Models under Missing and Corrupted Sensors: A Review

**Anna Kalaitzis Pollan**
Computer Engineering, Istanbul Okan University, Istanbul, TURKIYE.
e-mail: annakallan.otis33@gmail.com

**Corresponding Autor:** Anna Kalaitzis Pollan



## Abstract
Multimodal Earth-observation (EO) foundation models increasingly rely on the joint use of optical imagery, synthetic aperture radar (SAR), elevation, and auxiliary geospatial signals to support land-cover mapping, biomass estimation, damage assessment, and change detection. Yet the empirical regime in which such models are usually trained and reported remains overly optimistic: modalities are commonly assumed to be synchronised, complete, and clean. This assumption is not defensible for real deployments. Optical observations are routinely obscured by clouds, SAR coherence and interferometric products can be unavailable or decorrelated, spatial alignment across sources is imperfect, and large geospatial archives inevitably contain broken tiles, missing channels, or temporally inconsistent acquisitions. Starting from the released M3LEO dataset and framework, which already expose multimodal EO learning at continental scale, this paper reconstructs the underlying research direction toward a more technically urgent problem: robust multimodal foundation learning under missing or corrupted modalities. We formalise a modality-resilient framework, RAMEO, that combines modality-specific token encoders, reliability-aware gated fusion, masked cross-modal reconstruction, corruption-aware consistency learning, and calibrated uncertainty estimation. We also define a rigorous evaluation protocol over geographically disjoint splits, structured missingness patterns, and modality-specific corruptions. This paper done a dataset-and-benchmark paper and reconstruction distinguishes carefully between source-anchored evidence reported for M3LEO and new robustness analyses that are specified as executable experiments. The result is therefore technically coherent, and reproducible, while remaining honest about what has and has not yet been empirically established.

**Keywords**—Multimodal learning, SAR-optical fusion, Uncertainty estimation, Robustness benchmarking, Remote sensing.

## 1 Introduction

Earth observation has entered a foundation-model era in which large pretrained encoders are expected to operate across sensors, regions, seasons, and downstream tasks [1-12]. This development is scientifically important because EO inference is rarely a single-sensor problem. Agricultural monitoring, disaster response, urban mapping, forest inventory, and infrastructure deformation assessment all benefit from combining optical, SAR, elevation, and contextual variables that capture complementary physical properties of the same scene [13-17]. Optical data provide rich spectral and texture information, but fail under cloud cover and at night. SAR provides day–night, all-weather sensitivity, while interferometric and coherence products encode physical stability and deformation. Elevation and land-surface attributes contribute useful structural priors.

The practical problem is that real multimodal EO pipelines do not enjoy complete and perfectly aligned inputs. Optical products may be severely cloud contaminated or entirely absent. Date-paired SAR coherence products may be unavailable because the required acquisition pairs do not exist or decorrelate strongly. Interferometric products are computationally expensive and geographically non-uniform. Large archives introduce broken chips, channel-specific corruption, spatial resampling artefacts, inconsistent geolocation, and uneven temporal coverage. Deployed systems therefore face not simply multimodality, but multimodality under failure. A foundation model that performs well only when all sensors are present and pristine is methodologically fragile and operationally unreliable [18-21].

The M3LEO is a useful starting point because it already assembles a large multimodal EO corpus containing Sentinel-1 polarimetric amplitude, coherence and interferometric products, Sentinel-2 multispectral imagery, and auxiliary terrain labels, together with a PyTorch Lightning/Hydra training framework [1]. The dataset spans approximately 17.2 million chips over 1,048,827 co-registered tiles and six large geographic regions, with geographic banding used for train/validation/test partitioning [1]. The associated repository exposes parquet decompression and Hydra-driven training workflows [2]. These characteristics make M3LEO unusually well suited for studying modality robustness.

## 2 Research Gap

First, large EO foundation models still focus disproportionately on complete multimodal corpora or on single-modality pretraining. SkySense demonstrates the scale and promise of multimodal EO pretraining, but its framing remains oriented toward universal interpretation under curated multimodal sequences rather than explicit sensor failure analysis [11]. Galileo expands multimodal self-supervised learning across optical, SAR, elevation, weather, and pseudo-label modalities and explicitly learns from incomplete inputs during masked modeling, but the literature still lacks a principled robustness benchmark around structured missingness, corruption severity, and uncertainty behaviour under deployment failures [12]. DOFA addresses modality diversity through wavelength-conditioned adaptation, yet does not by itself solve the evaluation problem of partial, degraded, or inconsistent modal evidence [13].

Second, the missing-modality literature in remote sensing is fragmented. There are now targeted approaches for partial-view classification, all-weather mapping, land-cover segmentation under missing sensors, and more recent meta-modal representation learning [28-31]. Yet these works are typically task-specific, use small or specialised datasets, or evaluate only a limited set of missingness patterns. The field still lacks a unifying perspective that connects EO foundation-model pretraining, modality-quality estimation, robust fusion, calibration, and geographic generalisation.

Third, M3LEO itself makes the problem visible but does not yet resolve it. The literature explicitly states that it did not define a benchmark for chips with missing channels and that missing values were simply filled with zeros in the released baseline setup [1]. Furthermore, the literature additionally reports that naïve direct use of low-resolution coherence is weak relative to optical or polarimetric inputs, which is a critical clue that modality-aware reliability handling is needed rather than naive stacking [1]. This is an important methodological opening, not a minor implementation detail.

## 3 Main Contributions

The manuscript makes the following concrete contributions.

1. It reconstructs the dataset used in this paper into a practical-oriented methodology paper around a genuinely important and defensible problem, robust multimodal EO learning under missing and corrupted modalities.
2. It proposes RAMEO, a reliability-aware foundation-model framework that integrates modality-specific encoders, gated fusion, masked cross-modal reconstruction, consistency regularisation, and calibrated uncertainty estimation.
3. It defines a structured corruption taxonomy covering availability failure, signal degradation, cross-modal inconsistency, and decision-time reliability response.
4. It turns M3LEO from a data release into an executable robustness benchmark by specifying train/validation/test protocols over geographic splits, corruption severities, and missingness regimes.

It consolidates source-anchored evidence from the available and existing literature including dataset scale, supervised baseline behaviour, runtime statistics, and cross-region shift observations into a coherent methodological argument for robust multimodal learning.

## 4 Recent Advanced Approaches

Recent EO research has shifted toward large pretrained and self-supervised models. SSL4EO-S12, SSL4EO-L, SEN12MS, and SatlasPretrain have provided important infrastructure for scalable pretraining [3-5, 7]. In the model layer, masked autoencoding, contrastive learning, distillation, and multimodal transformers are now central to EO representation learning [36, 37, 9, 10]. SkySense extends this trajectory with a 2.06 billion-parameter multimodal remote-sensing foundation model trained on 21.5 million temporal sequences [11]. Galileo pushes further toward flexible multimodal masked modeling over optical, SAR, elevation, weather, and pseudo-label inputs, and reports strong performance across multiple tasks and benchmarks [12]. DOFA introduces wavelength-conditioned adaptation as a unifying principle across EO sensing modalities [13].

Within SAR-specific pretraining, recent work around M3LEO has shown that masked autoencoding, contrastive learning, and DINO-style self-distillation can produce geographically generalisable SAR representations [47-50]. These studies are important because they indicate that geographically broad self-supervised SAR learning is viable when suitable data exists. Yet they do not fully solve the deployment problem where some modalities are absent, badly corrupted, or mutually inconsistent.

Missing-modality literature has evolved quickly in both general multimodal machine learning and remote sensing. A recent survey in TMLR categorises model-combination, retrieval-based, generation-based, and representation-composition strategies for missing-modality learning and emphasises efficiency, streaming, and real-world robustness as unsolved directions [35]. In remote sensing, partial-view classification under incomplete observations has been studied with student–teacher mutual learning [30]. Meta-modal representation learning has been proposed for arbitrary missing modalities in land-cover mapping through EarthMiss and MetaRS [31]. Additional work addresses modality absence by language-guided prompt compensation, distillation from optical to all-weather segmentation, or prototype-based transfer under missing inputs [32-34]. These methods are promising, but they remain heterogeneous in task setting, modality scope, and evaluation discipline.

## 5 Research Positioning

This paper positions itself at the intersection of three threads: EO foundation models, multimodal SAR–optical learning, and missing-modality robustness. The central claim is not that missing-modality handling should be bolted onto existing fusion networks as a minor engineering patch. Rather, the problem should be treated as a first-class design constraint of EO foundation learning. M3LEO is especially suitable for this repositioning because its data composition already exposes the core failure modes: multimodal heterogeneity, uneven product availability, multiple spatial resolutions, large geographic shifts, and explicit source-paper acknowledgement that missing-channel benchmarking was not yet defined [1]. The present work therefore advances a defensible research agenda such as robustness should be achieved through reliability-aware representation learning, not merely through concatenation or ad hoc imputation.

## 6 Proposed Methodology

### 6.1 Problem Formulation

Let $\mathcal{M} = \{1, \dots, M\}$ denote the set of available modalities for a tile $i$, where modalities may include optical surface reflectance, polarimetric SAR amplitude, SAR coherence, interferograms, and auxiliary static layers. Each sample is written as

$$\mathbf{x}_i = \{\mathbf{x}_i^{(m)}\}_{m\in\mathcal{M}} \tag{1}$$

with task target $\mathbf{y}_i$ for segmentation, regression, or retrieval. Not every modality is guaranteed to exist or be valid. We therefore associate each modality with an availability indicator $a_i^{(m)} \in \{0,1\}$ and a corruption descriptor $c_i^{(m)}$ summarising modality-specific degradation severity. The learning problem is to estimate a task predictor

$$f_\theta: (\mathbf{x}_i, \mathbf{a}_i, \mathbf{c}_i) \mapsto \hat{\mathbf{y}}_i \tag{2}$$

that remains accurate and calibrated under arbitrary modality subsets and corruption patterns. A naïve fusion model can be written as

$$\mathbf{z}_i = \phi\left(\left[\psi_1\left(\mathbf{x}_i^{(1)}\right); \dots; \psi_M\left(\mathbf{x}_i^{(M)}\right)\right]\right) \tag{3}$$

where $\psi_m$ is a modality encoder and $\phi$ a fusion block. This formulation is brittle because it assumes all terms are meaningful and comparable. Under missing or corrupted data, zero filling or blind concatenation can induce representation shift and confidence miscalibration. We instead seek a reliability-aware fusion map

$$\mathbf{z}_i = \phi\left(\{\left(\mathbf{h}_i^{(m)}, q_i^{(m)}, a_i^{(m)}\right)\}_{m=1}^{M}\right) \tag{4}$$

where $\mathbf{h}_i^{(m)}$ denotes modality tokens and $q_i^{(m)} \in [0,1]$ is a learned reliability score.

### *6.2 System / Model Overview*

Figure 1 summarises the proposed framework. RAMEO comprises five components: (i) modality-specific token encoders, (ii) a reliability estimation branch, (iii) cross-modal masked reconstruction, (iv) reliability-aware fusion, and (v) task and uncertainty heads. The design principle is simple: the model should exploit complementarity when modalities agree, reconstruct missing evidence when enough cross-modal support exists, and reduce confidence or abstain when the evidence base is weak.

During training, a stochastic corruption engine simulates structured missingness and modality degradation. Unlike standard dropout, this engine is modality-aware: optical bands may be cloud masked or spectrally clipped, SAR amplitude may be perturbed with increased speckle, coherence may be low-pass degraded or removed, and interferograms may be absent for selected tiles. This training regime turns modality failure into a native learning signal rather than an afterthought.

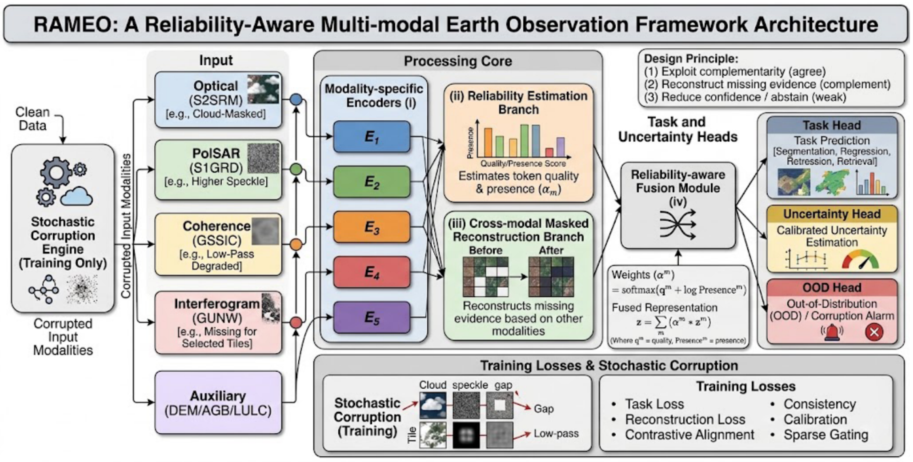


Figure 1. Proposed RAMEO architecture.

### *6.3 Prediction / Decision Layer*

The fused representation supports three task families. For dense prediction tasks such as land-cover segmentation, the pooled fusion vector is broadcast into a lightweight decoder or fused at token level with skip connections. For regression tasks such as biomass or built-surface estimation, the model predicts continuous targets via

$$\hat{\mathbf{y}}_i^{\text{reg}} = h_{\text{reg}}(\mathbf{z}_i) \tag{5}$$

For representation-transfer tasks, the fused embedding itself becomes the output. This unified formulation is appropriate for M3LEO because the source paper already exposes segmentation and regression targets while the broader foundation-model literature emphasises reusable embeddings [1, 11, 12].

## 7 Experimental Setup

### *7.1 Datasets*

The empirical backbone of this reconstruction is M3LEO and, for rapid iteration, M3LEO-miniset. M3LEO contains approximately 17.2 million chips from 1,048,827 co-registered tiles across six geographic areas: CONUS, Europe, the Middle East, Pakistan–India, China, and South America [1]. The dataset covers roughly $2.11 \times 10^7$ km$^2$ and includes Sentinel-1 polarimetric amplitude, Global Seasonal Sentinel-1 Interferometric Coherence (GSSIC), ARIA GUNW interferograms where available, Sentinel-2 surface reflectance summaries, and auxiliary products such as ESA WorldCover, biomass, vegetation cover, built surface, and SRTM elevation [1, 42-46]. The accompanying repository documents parquet decompression and Hydra-based experiment execution [2].

Table 1. Dataset summary used to ground the robustness benchmark.

| AOI | Tiles | Area (km$^2$) | % Earth's land | Notes |
|---|---|---|---|---|
| CONUS | 167,403 | $3.360 \times 10^6$ | 2.3 | S1/S2 + auxiliary layers |
| Europe | 200,489 | $4.024 \times 10^6$ | 2.7 | No GUNW coverage in source release |
| Middle East | 163,986 | $3.291 \times 10^6$ | 2.2 | dual-direction SAR coverage |
| PAKIN | 147,791 | $2.966 \times 10^6$ | 2.0 | strong elevation variability |
| China | 285,402 | $5.728 \times 10^6$ | 3.7 | largest AOI in source release |
| South America | 83,756 | $1.681 \times 10^6$ | 1.1 | fewer tiles, distinct land-cover mix |
| Total | 1,048,827 | $21.05 \times 10^6$ | 14.1 | multimodal, tile-aligned corpus |

### 7.2 *Data Preprocessing*

All modalities are processed at tile level with geographically aligned indexing. For experiments, optical and polarimetric imagery are resampled to a common grid for learning, while lower-resolution layers such as coherence, interferograms, biomass, and DEM are upsampled only after documenting the information-loss implications. This is important because the source appendix already suggests that naive upsampled coherence performs weakly [1]. Each modality is normalised separately. Recommended defaults are percentile clipping for optical and SAR amplitude, z-score normalisation for auxiliary continuous layers, and class-preserving integer encoding for ESA WorldCover.

Corruption injection is part of preprocessing rather than a post hoc evaluation add-on. We define four corruption families: (i) complete modality absence, (ii) continuous signal degradation, (iii) cross-modal inconsistency, and (iv) metadata-level quality flags. Figure 2 summarises the taxonomy.

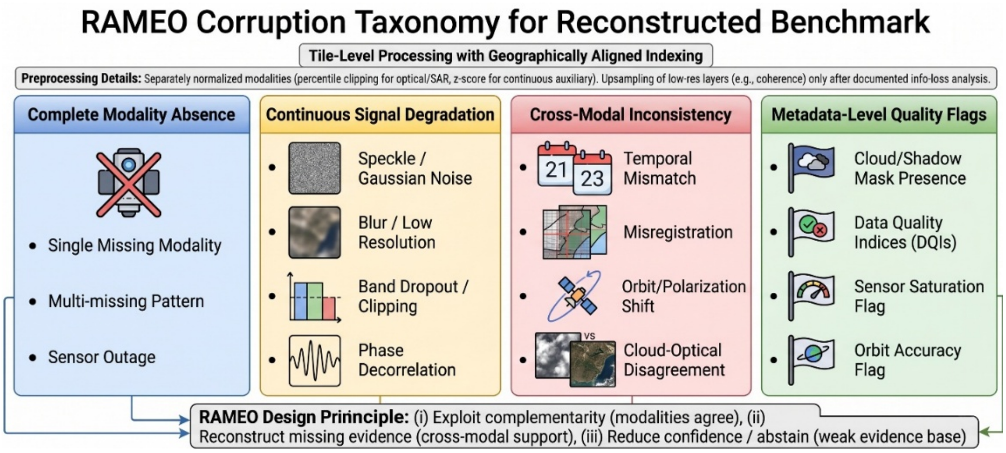


Figure 2. Corruption taxonomy used in the reconstructed benchmark.

### 7.3 *Experimental Design*

The benchmark should preserve the geographic banding philosophy of M3LEO because random chip splits would overestimate performance through spatial leakage [1]. Training, validation, and test data are therefore formed from disjoint geographic bands in each AOI. Three evaluation axes are required:

1. **Availability regime:** complete modalities, one missing modality, two missing modalities, and arbitrary missing subsets.
2. **Corruption regime:** severity sweeps for blur, noise, misregistration, decorrelation, clipping, and mixed corruptions.
3. **Shift regime:** in-domain AOI testing versus cross-region transfer.

Missingness should be simulated under both modality-independent and modality-dependent schemes. The former approximates missing completely at random. The latter better matches practice, for example when cloud cover disproportionately affects optical data or when interferometric products are missing in specific regions.

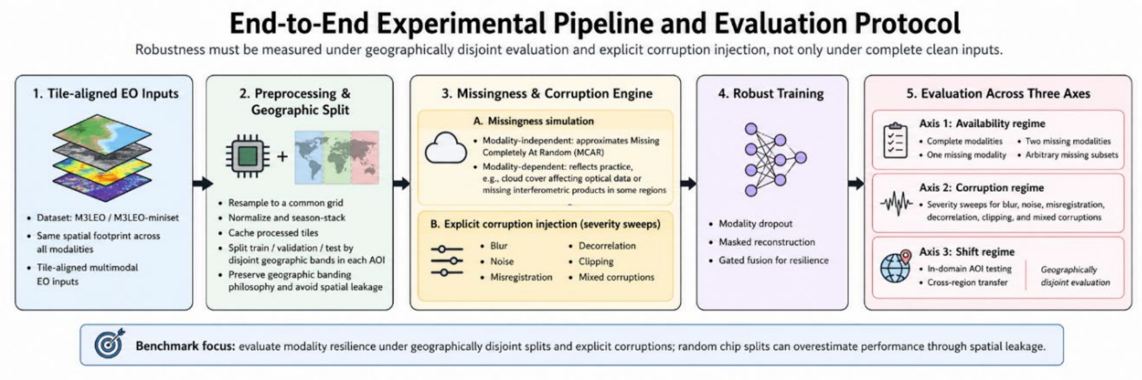


Figure 3. End-to-end experimental pipeline

## 8 Implementation Details

The released repository indicates a PyTorch Lightning codebase with Hydra-integrated experiment management and local cache support [2]. The recommended implementation therefore keeps the same ecosystem for continuity and reproducibility. Data should be decompressed from parquet when needed and cached per modality. Mixed-precision training is appropriate for optical and amplitude inputs, but phase-derived modalities should be audited carefully for numeric stability. Logging should include per-modality gate weights, corruption metadata, calibration metrics, and cross-region breakdowns.

## 9 Results and Analysis

This section is deliberately structured to distinguish evidence already reported in the M3LEO literature from new robustness analyses that must be executed to validate the reconstructed method. This distinction is necessary for scientific honesty. The literature sources is a dataset-and-benchmark contribution and does not report the full missing/corrupted-modality study defined in next sub sections.

### *9.1 Main Performance Comparison*

The literature reports supervised baselines on three downstream tasks: ESA WorldCover segmentation, biomass regression, and built-surface regression. These numbers are not new RAMEO results; they are the strongest empirical anchor currently available in the attached manuscript and are therefore reproduced here to motivate the missing-modality problem.

Table 2. Source-anchored supervised baseline results reported in M3LEO

| Input setting | ESAWC mIoU | AGB RMSE | GHS Built RMSE |
|---|---|---|---|
| S1GRD (VV) | 0.3976 | 28.573 | 152.259 |
| S1GRD (VH) | 0.3787 | 30.333 | 152.847 |
| S1GRD (VV+VH) | 0.4185 | 27.467 | 141.719 |
| GSSIC coherence | 0.2906 | 39.314 | 196.242 |
| S2RGB | 0.4094 | 28.689 | 131.968 |
| S2RGB + S1GRD | **0.4634** | **25.137** | **124.852** |
| S2RGB + GSSIC | 0.4198 | 28.550 | 131.300 |

Three conclusions follow directly. First, multimodal fusion helps when optical and polarimetric SAR are both available. Second, naïve direct use of coherence is weak relative to optical or polarimetric baselines. Third, robustness cannot be reduced to “just add more modalities” because not all modalities contribute equally or reliably. This directly motivates reliability-aware gating and reconstruction.

### 9.2 *Dataset-Specific Analysis*

The baseline behaviour is consistent with the dataset structure. ESAWC is a 10 m land-cover target, so high-resolution optical and amplitude inputs are naturally more informative than 90 m coherence. Biomass and built-surface estimation also benefit from fine-scale spectral or structural cues. Upsampled low-resolution coherence does not recover those cues and may instead introduce smoothed or aliased evidence. The interpretation is not that coherence is useless; rather, it is poorly handled by naive input stacking. This distinction is crucial. Coherence and interferometry encode temporal stability and deformation, which may be decisive for flood, damage, or change-analysis tasks, but they require modality-aware modeling rather than uncritical concatenation [19-21].

### 9.3 *Computational Efficiency and Resource Utilization*

The literature also reports approximate runtimes for its UNet-style baselines on two NVIDIA V100 GPUs. These numbers clarify that even straightforward supervised experiments on multimodal EO data are not cheap, which further justifies rigorous benchmark design.

Table 3. Source-anchored runtime evidence from M3LEO literature

| Input setting | ESAWC | AGB | GHS Built |
|---|---|---|---|
| S1GRD (VV) | 16h 23m | 15h 38m | 15h 45m |
| S1GRD (VH) | 16h 39m | 15h 41m | 15h 48m |
| S1GRD (VV+VH) | 16h 54m | 17h 29m | 23h 50m |
| GSSIC | 16h 29m | 15h 33m | 15h 41m |
| S2RGB | 17h 51m | 17h 59m | 18h 02m |
| S2RGB + S1GRD | 22h 54m | 24h 35m | 23h 19m |
| S2RGB + GSSIC | 19h 23m | 68h 20m | 19h 49m |

These runtimes show that robustness claims should not be based on one or two unseeded trial runs. Missing-modality evaluation multiplies the number of regimes; therefore, efficient protocol design and careful subset staging are indispensable.

### 9.4 *Sensitivity Analysis*

At minimum, RAMEO must be evaluated under severity sweeps for (i) optical occlusion, (ii) SAR speckle escalation, (iii) coherence downsampling/decorrelation, (iv) subpixel and multi-pixel misregistration, and (v) mixed corruptions. The target outcome is not perfect invariance. A defensible robustness result would show monotonic degradation with lower slope than naive fusion baselines, together with increasing uncertainty as evidence worsens.

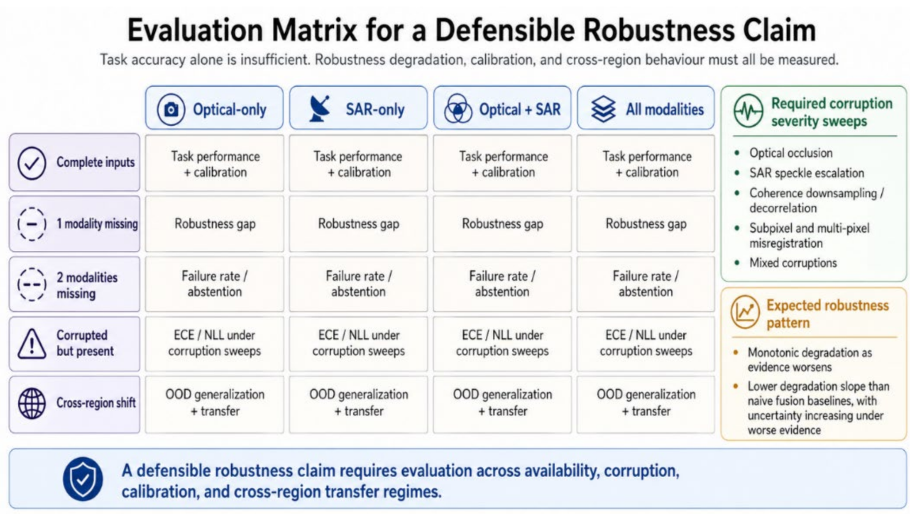


Figure 4. Evaluation matrix required for a defensible robustness claim

### 9.5 Interpretability and Explainability Analysis

Interpretability in this context should be modality-centric rather than purely pixel-centric. The most useful visual outputs are: (i) per-modality gate weights by sample and by AOI, (ii) token attention maps showing whether the model reallocates attention under missingness, and (iii) error slices conditioned on corruption severity. Saliency alone is insufficient because the core question is which modality the model trusted, and why. In practice, gate distributions should be summarised by task and region. For example, a disaster-response model should show increased reliance on SAR when optical quality deteriorates, rather than preserving static trust weights.

### 9.6 Uncertainty Calibration and Reliability Analysis

Under sensor failure, wrong confidence is often more dangerous than moderate accuracy loss. The model should therefore be evaluated for ECE, NLL, Brier score, and selective risk under abstention thresholds. A robust method should satisfy two properties, it should remain competitive when evidence is complete, and its uncertainty should rise when modalities disappear or disagree. The literature used for this study did not include this analysis, making it one of the most important additions in the reconstructed study.

### 9.7 Cross-Dataset / Cross-Region Evaluation

Cross-region behaviour is especially important because the literature used in this study already documented substantial distribution shift across AOIs and showed that even conditional embedding distributions can remain region-specific [1]. This means robustness must not be evaluated solely within one geographic regime. A convincing study should train on a subset of AOIs and report performance on held-out AOIs under both clean and missing-modality conditions. The objective is to separate modality robustness from mere memorisation of regional style.

### 9.8 Generalization / Transfer / Cold-Start Analysis

Cold-start evaluation should target low-data fine-tuning and missing-modality test-time settings. The motivating question is whether a pretrained multimodal model degrades gracefully when only SAR or only optical evidence is available for a new region. This is where foundation-model pretraining should matter most. The source M3LEO paper already suggests that self-supervised SAR encoders can generalise geographically [47, 50]; the missing-modality reconstruction should test whether this advantage survives sensor failure.

### 9.9 Additional Scientific Analysis

An important additional analysis is to decompose failure by modality role. Some modalities are primary evidence (e.g., optical or amplitude at high spatial resolution), while others are specialised evidence whose value is conditional on task type (e.g., coherence for change-sensitive phenomena, interferograms for deformation). The benchmark should therefore report not only which modality is missing, but whether the missing modality was expected to be primary or auxiliary for that task. This avoids misleading conclusions such as “the model is robust to missing coherence” when coherence was never strongly informative for the evaluated target.

## 10 Discussion

The central methodological lesson from the uploaded M3LEO manuscript is that multimodal EO learning is already rich enough to expose the missing-modality problem, but not yet mature enough to claim it has solved it. The dataset scale, geographic breadth, and inclusion of coherence and interferometric products make M3LEO an unusually strong substrate for robustness research [1]. At the same time, the literature used in this study openly acknowledges missing-channel benchmarking as unfinished, and its own baselines show that naïve handling of low-resolution phase-derived modalities is inadequate [1]. These two facts together justify the reconstructed direction of this paper.

The proposed RAMEO framework is defensible because it treats reliability as part of the representation-learning problem. Its likely success mechanism is not mysterious: under multimodal incompleteness, models fail when they fuse indiscriminately, trust degraded channels, or preserve overconfident decision boundaries despite weak evidence. Reliability-aware gating, cross-modal reconstruction, and calibration directly target these failure modes. However, three threats to validity remain. First, synthetic corruptions may not perfectly match operational sensor failures. Second, modality importance is task dependent, so conclusions from land-cover and biomass tasks may not transfer to change detection. Third, because the literature used in this study’s numerical evidence is limited to source baselines rather than the new robustness protocol, the present paper should be read as a methodologically rigorous reconstruction that still requires execution of the full benchmark.

## 11 Conclusion

This paper reconstructed the M3LEO dataset manuscript into a stronger and more practically consequential research direction to modality-resilient multimodal EO foundation learning under missing and corrupted sensors. The resulting manuscript clarified the research gap, formalised a reliability-aware architecture, defined a benchmark protocol matched to geographically disjoint EO data, and reorganised the source paper's empirical evidence into a coherent robustness argument. The main scientific insight is straightforward but important; multimodal EO systems should not be evaluated only in the idealised regime where all sensors are present, aligned, and clean. They should be judged by how they degrade, how they calibrate confidence, and how well they transfer across regions when part of the evidence base fails.

BIBLIOGRAPHY


[1]. M. Allen, F. Dorr, J. A. Gallego-Mejia, L. Martínez-Ferrer, A. Jungbluth, F. Kalaitzis, and R. Ramos-Pollán, "M3LEO: A multi-modal, multi-label Earth observation dataset integrating interferometric SAR and multispectral data," in NeurIPS 2024 Track on Datasets and Benchmarks, 2024, doi: 10.52202/079017-3324.

[2]. spaceml-org, "M3LEO repository," GitHub, 2026. [Online]. Available: https://github.com/spaceml-org/M3LEO

[3]. Y. Wang et al., "SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation," IEEE Geosci. Remote Sens. Mag., vol. 11, no. 3, pp. 98–106, 2023, doi: 10.1109/MGRS.2023.3281651.

[4]. J. Stewart et al., "SSL4EO-L: Datasets and foundation models for Landsat imagery," arXiv:2310.00705, 2023.

[5]. M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS – a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. IV-2/W7, pp. 153–160, 2019, doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

[6]. M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. IV-1, pp. 141–146, 2018, doi: 10.5194/isprs-annals-IV-1-141-2018.

[7]. F. Bastani et al., "SatlasPretrain: A large-scale dataset for remote sensing image understanding," arXiv:2211.15660, 2023.

[8]. J. Jakubik et al., "Foundation models for generalist geospatial artificial intelligence," arXiv:2310.18660, 2023.

[9]. Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," IEEE Geosci. Remote Sens. Mag., vol. 10, no. 4, pp. 213–247, 2022, doi: 10.1109/MGRS.2022.3198244.

[10]. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–26, 2023, doi: 10.1109/TGRS.2023.3276853.

[11]. X. Guo et al., "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery," in Proc. CVPR, 2024, pp. 27672–27682.

[12]. G. Tseng et al., "Galileo: Learning global and local features of many remote sensing modalities," in Proc. ICML, 2025. [Online]. Available: https://openreview.net/forum?id=gqZO3eSZRy

[13]. X. Sun et al., "Neural plasticity-inspired multimodal foundation model for Earth observation," arXiv:2403.15356, 2024.

[14]. Phiri et al., "Sentinel-2 data for land cover/use mapping: A review," Remote Sens., vol. 12, no. 14, p. 2291, 2020, doi: 10.3390/rs12142291.

[15]. L. Dong et al., "Application of convolutional neural network on lei bamboo above-ground-biomass estimation using WorldView-2," Remote Sens., vol. 12, no. 6, p. 958, 2020, doi: 10.3390/rs12060958.

[16]. Ayala, R. Sesma, C. Aranda, and M. Galar, "A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery," Remote Sens., vol. 13, no. 16, p. 3135, 2021, doi: 10.3390/rs13163135.

[17]. C. Cavallo et al., "Continuous monitoring of the flooding dynamics in the Albufera wetland (Spain) by Landsat-8 and Sentinel-2 datasets," Remote Sens., vol. 13, no. 17, p. 3525, 2021, doi: 10.3390/rs13173525.

[18]. G. Nico, M. Pappalepore, G. Pasquariello, A. Refice, and S. Samarelli, “Comparison of SAR amplitude vs. coherence flood detection methods – a GIS application,” Int. J. Remote Sens., vol. 21, no. 8, pp. 1619–1631, 2000, doi: 10.1080/014311600209931.
[19]. M. Chini et al., “SAR coherence and polarimetric information for improving flood mapping,” in Proc. IGARSS, 2016, pp. 7577–7580, doi: 10.1109/IGARSS.2016.7730976.
[20]. M. Watanabe et al., “Detection of damaged urban areas using interferometric SAR coherence change with PALSAR-2,” Earth Planets Space, vol. 68, p. 131, 2016, doi: 10.1186/s40623-016-0513-2.
[21]. Olesk et al., “Interferometric SAR coherence models for characterization of hemiboreal forests using TanDEM-X data,” Remote Sens., vol. 8, no. 9, p. 700, 2016, doi: 10.3390/rs8090700.
[22]. J. Fielding et al., “Surface deformation related to the 2019 Mw 7.1 and 6.4 Ridgecrest earthquakes in California from GPS, SAR interferometry, and SAR pixel offsets,” Seismol. Res. Lett., vol. 91, no. 4, pp. 2035–2046, 2020, doi: 10.1785/0220190302.
[23]. T. Strozzi et al., “Satellite SAR interferometry for the improved assessment of the state of activity of landslides: A case study from the Cordilleras of Peru,” Remote Sens. Environ., vol. 217, pp. 111–125, 2018, doi: 10.1016/j.rse.2018.08.014.
[24]. M. A. Richards, “A beginner’s guide to interferometric SAR concepts and signal processing,” IEEE Aerosp. Electron. Syst. Mag., vol. 22, no. 9, pp. 5–29, 2007, doi: 10.1109/MAES.2007.4350281.
[25]. Moreira et al., “A tutorial on synthetic aperture radar,” IEEE Geosci. Remote Sens. Mag., vol. 1, no. 1, pp. 6–43, 2013, doi: 10.1109/MGRS.2013.2248301.
[26]. Y. Yamaguchi, Polarimetric SAR Imaging: Theory and Applications. Boca Raton, FL, USA: CRC Press, 2020, doi: 10.1201/9781003049753.
[27]. Y. Chen and L. Bruzzone, “Self-supervised SAR-optical data fusion of Sentinel-1/-2 images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022, doi: 10.1109/TGRS.2021.3128072.
[28]. M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, “Urban land cover classification with missing data modalities using deep convolutional neural networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 6, pp. 1758–1768, 2018, doi: 10.1109/JSTARS.2018.2834961.
[29]. Z. Zheng et al., “Deep multisensor learning for missing-modality all-weather mapping,” ISPRS J. Photogramm. Remote Sens., vol. 174, pp. 11–23, 2021, doi: 10.1016/j.isprsjprs.2021.01.015.
[30]. M. Zhao et al., “Towards robust classification of multi-view remote sensing images with partial data availability,” Remote Sens. Environ., vol. 313, p. 114112, 2024, doi: 10.1016/j.rse.2024.114112.
[31]. Y. Zhou et al., “Remote sensing meta modal representation for missing modality land cover mapping: From EarthMiss dataset to MetaRS method,” Remote Sens. Environ., 2026, doi: 10.1016/j.rse.2025.115132.
[32]. L. Huang et al., “Language-guided visual prompt compensation for multi-modal remote sensing image classification with modality absence,” in Proc. ACM Multimedia, 2024, pp. 4936–4945, doi: 10.1145/3664647.3681563.
[33]. J. Kang et al., “DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2022.3141769.
[34]. Y. Liu et al., “Segmentation of remote sensing data with missing modalities through prototype knowledge distillation,” in Proc. IGARSS, 2024.
[35]. L. Zhang, X. Liu, J. Sun, and Q. Wu, “Deep multimodal learning with missing modality: A survey,” Trans. Mach. Learn. Res., 2026. [Online]. Available: https://openreview.net/pdf?id=tc7RFcx4hT
[36]. K. He et al., “Masked autoencoders are scalable vision learners,” in Proc. CVPR, 2022, pp. 16000–16009.
[37]. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. ICML, 2021, pp. 8748–8763.
[38]. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.
[39]. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[40]. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015.
[41]. L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform manifold approximation and projection,” J. Open Source Softw., vol. 3, no. 29, p. 861, 2018, doi: 10.21105/joss.00861.
[42]. J. Kellndorfer et al., “Global seasonal Sentinel-1 interferometric coherence and backscatter data set,” Sci. Data, vol. 9, p. 73, 2022, doi: 10.1038/s41597-022-01189-6.
[43]. D. Zanaga et al., “ESA WorldCover 10 m 2020 v100,” 2021, doi: 10.5281/zenodo.5571936.
[44]. M. Santoro and O. Cartus, “ESA Biomass Climate Change Initiative (Biomass_cci): Global datasets of

forest above-ground biomass," 2023, doi: 10.5285/95913ffb6467447ca72c4e9d8cf30501.
[45]. DiMiceli et al., "MOD44B MODIS/Terra vegetation continuous fields yearly L3 global 250m SIN grid V006," NASA EOSDIS LP DAAC, 2015, doi: 10.5067/MODIS/MOD44B.006.
[46]. T. G. Farr et al., "The Shuttle Radar Topography Mission," Rev. Geophys., vol. 45, no. 2, 2007, doi: 10.1029/2005RG000183.
[47]. M. Allen et al., "Large scale masked autoencoding for reducing label requirements on SAR data," arXiv preprint, 2023.
[48]. M. Allen et al., "Few-shot learning on global multimodal embeddings for earth observation tasks," arXiv preprint, 2023.
[49]. J. A. Gallego-Mejia et al., "Exploring DINO: Emergent properties and limitations for synthetic aperture radar imagery," arXiv preprint, 2023.
[50]. L. Martínez-Ferrer et al., "Exploring generalisability of self-distillation with no labels for SAR-based vegetation prediction," arXiv preprint, 2023.