Urban-Aware Evaluation for Geospatial Foundation Models in Landslide Mapping

Zahra Irandegani, Reza Safarzadeh, and Xin Wang(✉)

University of Calgary, Calgary, AB, Canada
{zahra.irandegani,reza.safarzadeh,xcwang}@ucalgary.ca

Abstract. Landslides cause severe damage to infrastructure and human settlements, making reliable monitoring from Earth Observation imagery essential for disaster response. However, existing landslide mapping systems are often trained using future observations unavailable at prediction time, evaluated without considering infrastructure exposure, and provide limited support for risk assessment. We propose UrbanRisk-GFM, a framework that combines causal adaptation of the Geospatial Foundation Model (GFM) Prithvi-EO-2.0, urban-aware evaluation, and a proof-of-concept risk intelligence pipeline. The framework introduces Urban-F1 and Critical Miss Rate (CMR) to assess performance near roads and buildings and uses structured geospatial evidence to generate concise risk reports. Experiments on the Sen12Landslides benchmark show that causal Prithvi outperforms standard fine-tuned Prithvi, U-TAE, TSViT, and 3D-UNet, achieving the highest Urban-F1 (0.641) while reducing CMR from 0.497 to 0.342. Additional experiments demonstrate that the same causal model remains effective under partial temporal availability without retraining. These results highlight the potential of GFMs for urban-aware landslide assessment and risk intelligence.

Keywords: Geospatial Foundation Models · Landslide Monitoring · Urban-Aware Evaluation · Landslide Risk Assessment · Risk Reporting

1 Introduction

Landslides are among the most widespread and destructive natural hazards, affecting approximately 4.8 million people and causing over 18,000 deaths between 1998 and 2017 [1]. In practice, the greatest risks arise when landslides threaten roads, buildings, and other critical infrastructure, as delayed response can amplify societal impact. Earth Observation (EO) imagery is essential for landslide assessment and disaster response because its wide spatial coverage and frequent revisit cycles enable continuous monitoring of evolving slope conditions [2]. Geospatial Foundation Models (GFMs), such as Prithvi [3], are pre-trained on large-scale multitemporal EO data to learn transferable spatiotemporal representations that can be adapted to diverse downstream geospatial tasks [4], providing a promising foundation for multitemporal landslide mapping. However, strong benchmark performance does not guarantee reliable identification of landslides threatening nearby infrastructure during real-world disaster response [5].

Despite these advances, several challenges limit the real-world use of multitemporal landslide monitoring systems. First, multitemporal models are often trained with future observations unavailable during deployment, introducing temporal leakage and overly optimistic evaluation. Second, standard metrics such as F1-score (F1) and IoU weight all errors equally [6,7], despite the greater consequences of misses near critical infrastructure. Third, raw segmentation masks provide limited guidance because they do not prioritize which landslides threaten nearby roads and buildings. Collectively, these limitations reveal a gap between benchmark-oriented landslide segmentation and the information needed for effective disaster response. To address these challenges, we propose **UrbanRisk-GFM**. The main contributions of this work are:

- We introduce causal temporal attention masking into Prithvi for multitemporal landslide mapping, preventing future-frame leakage.
- We introduce Urban-F1 and CMR, two urban-aware evaluation metrics that incorporate proximity to roads and buildings into landslide segmentation assessment.
- We demonstrate a proof-of-concept risk intelligence pipeline that converts segmentation outputs into risk summaries.

2 Related Work

Landslide inventory mapping from EO imagery has progressed from single-date CNN segmentation toward multitemporal deep learning and GFMs. U-Net [8] established the encoder–decoder framework that underpins most landslide segmentation pipelines, but single-date models frequently confuse landslides with spectrally similar persistent features such as bare soil or exposed riverbeds [7]. U-TAE [9] and TSViT [10] addressed this through temporal attention and transformer-based multitemporal segmentation respectively, but both depend on supervised data from specific regions and generalize poorly across geographic domains [6,11]. GFMs address this through self-supervised pretraining on large-scale EO archives [4], learning transferable representations without task-specific labels. Among them, Prithvi [3] is particularly relevant because it is pretrained on multitemporal Sentinel-2 sequences, matching the temporal structure of landslide mapping tasks. Prior Prithvi-based landslide studies relied on single-timestep inputs from Landslide4Sense [5], which provides single-date imagery. In contrast, UrbanRisk-GFM fine-tunes Prithvi on temporally ordered multitemporal sequences for landslide segmentation.

F1 and IoU—the standard measures across major landslide benchmarks including Landslide4Sense [6] and Sen12Landslides [7]—weight every pixel equally regardless of where it falls. Broader remote sensing evaluation has introduced object-level, boundary, and class-weighted variants [11], and some studies report stratified performance by land-cover type. None of these account for the operational consequences of where errors occur: a missed detection adjacent to a road or building carries the same penalty as one in open terrain. To our knowledge, no prior EO segmentation work accounts for infrastructure proximity or explicitly measures infrastructure-adjacent misses.

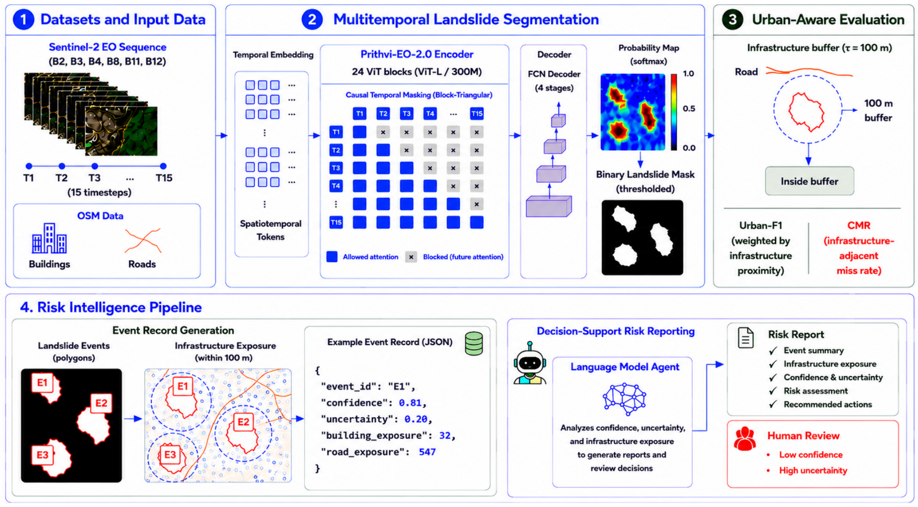

Fig. 1: Overview of the UrbanRisk-GFM framework.

Recent work has explored integrating EO encoders with large language models for grounded remote sensing analysis. GeoChat [12] and TEOChat [13] support scene understanding and temporal reasoning over EO imagery, demonstrating the potential of language-based interfaces for remote sensing. However, neither system addresses urban-aware prioritization or uncertainty-guided reporting under disaster-response constraints.

3 Methodology

UrbanRisk-GFM combines GFMs, urban-aware evaluation, and urban risk analysis for landslide assessment, as illustrated in Figure 1.

3.1 Multitemporal landslide Segmentation

Multitemporal landslide mapping requires modelling temporal changes across EO imagery. We target operational post-event monitoring, in which satellite acquisitions arrive sequentially and predictions may be required before the full sequence is observed. Under this formulation, attending to future frames ($t' > t$) at prediction time t introduces a train–test mismatch; the causal mask eliminates this by restricting each timestep to past and present observations. We compare standard bidirectional fine-tuning of Prithvi with this causally masked variant.

Prithvi-EO-2.0 [3] is pretrained using bidirectional temporal attention and adapted to a T=15 Sentinel-2 sequence through fine-tuning. The causal mask is defined as $M \in \{0, 1\}^{T \times T}$ with $M_{t,t'} = \mathbb{1}[t' \leq t]$. Applied to encoder self-attention logits before softmax, the mask introduces no additional trainable parameters and leaves the backbone unchanged. This prevents future-frame leakage while preserving temporal change information.

3.2 Urban-Aware Evaluation

Standard segmentation metrics such as F1-score weight all errors equally, regardless of their proximity to roads or buildings, so a model with strong aggregate performance can still miss landslides near critical infrastructure. We therefore introduce **Urban-F1** and **Critical Miss Rate (CMR)**, which weight errors by proximity to OSM-derived infrastructure (buildings and roads). The exposure weight $w(i, j)$ is defined from the Euclidean distance $d(i, j)$ to the nearest infrastructure pixel:

$$w(i,j) = 1 + \alpha \cdot \exp\left(-\frac{d(i,j)}{\sigma}\right) \tag{1}$$

where α sets the maximum emphasis and σ the spatial decay rate. We fix α=2.0 and σ=100 m a priori; across a sweep of 27 (α, σ, τ) combinations on a fixed reference seed, Urban-F1 and CMR each varied by less than 0.005. Exponential decay gives a smooth exposure gradient, assigning weight $1+\alpha$ to pixels adjacent to infrastructure ($d \rightarrow 0$) and approaching the unweighted baseline (w=1) far away, avoiding the discontinuity of a hard cutoff. The weighted confusion-matrix terms are per-pixel sums over the weight map:

$$\mathrm{TP}_w = \sum_{(i,j)} w(i,j)\,[\hat{y}_{ij}{=}1][y_{ij}{=}1], \quad \mathrm{FP}_w, \mathrm{FN}_w \text{ analogous} \tag{2}$$

where $\hat{y}_{ij}$ and y_{ij} are the predicted and ground-truth labels. Urban-F1 and CMR follow as:

$$\text{Urban-F1} = \frac{2\,\mathrm{TP}_w}{2\,\mathrm{TP}_w + \mathrm{FP}_w + \mathrm{FN}_w} \qquad \mathrm{CMR} = \frac{\mathrm{FN}_\tau}{\mathrm{TP}_\tau + \mathrm{FN}_\tau} \tag{3}$$

where TP_τ and FN_τ count ground-truth landslide pixels within τ=100 m of infrastructure. The two metrics are complementary by design: Urban-F1 applies continuous weighting as a general-purpose aggregate, while CMR uses a hard buffer to answer one operational question, namely what fraction of infrastructure-adjacent landslides were missed, independent of the weighting scheme.

3.3 Risk Intelligence Pipeline

Although Urban-F1 and CMR quantify performance near infrastructure, they do not provide information about the potential impact of detected landslides. Therefore, predicted landslide maps are combined with OSM roads and buildings to characterize infrastructure exposure. To estimate the reliability of each event, uncertainty is computed using Monte Carlo Dropout [14] (N = 20). Landslide events are then summarized as structured JSON records containing confidence, uncertainty, and infrastructure exposure information, which serve as the basis for risk assessment and report generation.

Decision-support risk reporting. Risk assessment depends on the interaction of confidence, uncertainty, infrastructure exposure, and building-density context, which is difficult to capture with a small set of fixed rules. Structured event records are therefore analysed by a tool-calling agent (Claude Sonnet [15], Anthropic Messages API) in a multi-turn loop of at most six iterations. The agent can retrieve infrastructure-proximity statistics, inspect pixel-level uncertainty,

Table 1: Urban-aware multitemporal landslide segmentation results on Sen12. Stratified columns report CMR. ‡ Lower CMR is achieved through higher urban recall (0.676 vs. 0.656) at the cost of lower urban precision (0.600 vs. 0.633) and substantially more infrastructure-adjacent false positives (238,555 vs. 94,519).

Model	F1	Urban-F1	CMR↓	Remote	Peri-Urban	Urban
Prithvi-C	**0.638**(±0.011)	**0.641**(±0.010)	0.342(±0.015)	**0.353**(±0.031)	0.389(±0.026)	0.347(±0.038)
Prithvi-S	0.452	0.448	0.497	0.526	0.507	0.560
U-TAE	0.604	0.614	0.366	0.419	0.382	0.411
3D-UNet	0.580	0.595	**0.324**‡	0.367	**0.298**	**0.370**
TSViT	0.523	0.529	0.374	0.434	0.372	0.431

estimate exposure from building counts, escalate an event for human review, and generate the final report. Escalation is an agent decision rather than a fixed gate: the agent is given heuristics such as low confidence with high uncertainty, or a high fraction of high-uncertainty pixels, but in practice escalates by integrating several corroborating signals, chiefly an elevated high-uncertainty pixel fraction together with a low data-quality flag. Report faithfulness is assessed with an LLM-as-judge protocol [16], extracting claims from each report and verifying them against the source event JSON.

4 Experiments

Experiments are conducted on the Sen12Landslides (Sen12) benchmark [7]. Evaluation uses the test split containing 2,043 multitemporal Sentinel-2 tiles spanning 15 geographic regions. We compare causal Prithvi against standard Prithvi, U-TAE [9], 3D-UNet [17], and TSViT [10]. Both Prithvi variants are adapted from the same pretrained Prithvi-EO-2.0 checkpoint and differ only in their temporal attention formulation. Code and an interactive dashboard are publicly available.[1]
Implementation details. Experiments are implemented in PyTorch. Sen12 is divided into 70/15/15% train/validation/test splits at the tile level (split seed=42), pooling tiles across all 15 geographic regions. Baseline models are trained from scratch using Adam (10^{-4}), while Prithvi-EO-2.0-300M is fine-tuned using AdamW (backbone 5×10^{-6}, decoder 10^{-5}) and cosine annealing. All models use weighted cross-entropy loss and early stopping based on validation F1. Baselines use seed= 42, while causal Prithvi is additionally evaluated across five seeds (42, 123, 999, 2025, 7777) with the data split held fixed.

4.1 Urban-aware Multitemporal Landslide Segmentation

Urban-aware evaluation. We assess whether urban-aware metrics provide additional insight beyond standard F1 when evaluating landslide predictions near roads and buildings. Table 1 shows that causal Prithvi (Prithvi-C) achieves the highest F1 (0.638) and Urban-F1 (0.641), outperforming both standard Prithvi (Prithvi-S) and all multitemporal baselines. Relative to standard Prithvi, causal

[1] `https://github.com/ZahraIrandegani/fm4urban-risk-intelligence`

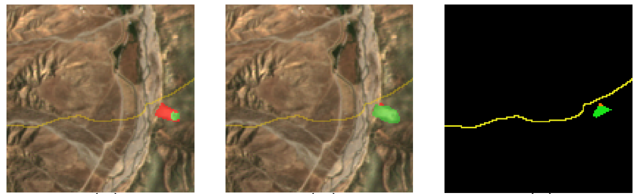

Fig. 2: Urban-critical failure case from Sen12. (a) standard Prithvi misses infrastructure-adjacent landslide pixels; (b) causal modeling recovers most of them. Panel (c) highlights false negatives recovered by causal attention (green).

attention improves Urban-F1 from 0.448 to 0.641 while reducing CMR from 0.497 to 0.342, indicating that the benefits of causal training extend to infrastructure-adjacent regions rather than being limited to aggregate segmentation performance.

Figure 2 illustrates a representative infrastructure-adjacent failure case where causal Prithvi recovers landslide regions missed by standard Prithvi near exposed infrastructure. This example is consistent with the quantitative results in Table 1, suggesting that causal attention can improve performance near infrastructure by enforcing consistency with the temporal information available at prediction time.

Although 3D-UNet achieves the lowest CMR, this advantage is obtained through higher recall at the cost of substantially more infrastructure-adjacent false positives (Table 1, note ‡). Urban-F1 therefore serves as the primary metric because it balances both false negatives and false positives near infrastructure.

Causal masking and training stability. We verify that the observed improvements arise from causal masking rather than differences in model capacity or temporal context. At fixed T=15, causal Prithvi improves Urban-F1 from 0.448 to 0.641 and reduces CMR from 0.497 to 0.342, while also increasing F1 from 0.452 to 0.638. The bidirectional variant's lower score does not reflect under-training, since both variants share the same backbone, training budget, and early-stopping protocol; bidirectional fine-tuning at T=15 creates a train–test mismatch by exposing the model to future frames during training that are unavailable at prediction time, a mismatch the causal mask eliminates. Five-seed evaluation yields 0.638±0.011 F1 and 0.641±0.010 Urban-F1 (Table 1), indicating that the gains ($\mathit{\Delta}$Urban-F1 $\approx +0.19$) are consistent across random initialisations.

Performance under partial temporal availability. In operational monitoring scenarios, only a subset of the temporal sequence may be available when predictions are required. We therefore assess whether causal Prithvi remains effective when only a prefix $T^* < 15$ of the sequence is provided at inference time. Because causal attention restricts each frame to past observations, the same checkpoint can be evaluated on any prefix $T^* \in \{1, 3, 6, 9, 12, 15\}$ without retraining. Figure 3 reports F1, Urban-F1, and CMR as a function of T^*.

Performance improves monotonically as additional observations become available. Notably, at $T^* = 12$, causal Prithvi already surpasses the full-sequence non-causal baseline (Urban-F1: 0.629 vs. 0.448), despite using fewer temporal

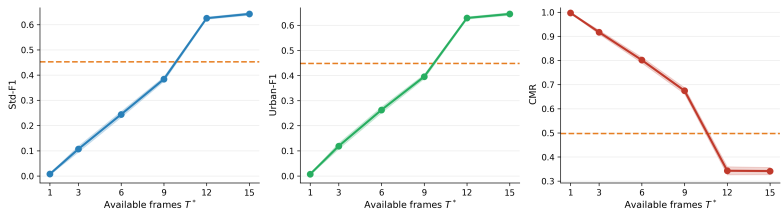

Fig. 3: F1, Urban-F1 and CMR as a function of available frames T^* for causal Prithvi (averaged over seeds). Dashed line: standard Prithvi T=15 reference.

observations. The largest improvement occurs between $T^* = 9$ and $T^* = 12$ (Urban-F1: 0.396 → 0.629; CMR: 0.675 → 0.343), suggesting that the model can exploit newly available temporal evidence without requiring retraining. These results indicate that causal training enables robust performance under incomplete temporal observations while retaining the ability to benefit from longer sequences when they become available.

4.2 Risk Intelligence Pipeline

We assess whether structured geospatial evidence can support reliable risk reporting. The pipeline is evaluated as a proof of concept on a stratified sample of 76 events spanning 15 geographic regions.

Decision-support risk reporting. The pipeline detects 2,129 events on the Sen12 test set. Of the 76 sampled events, the agent escalates 23 (30.3%) for human review and generates reports for the remaining 53 (69.7%). Escalations are concentrated in high-uncertainty cases (21 of 23), with two further cases triggered by conflicting event-level and tile-level uncertainty estimates, reflecting the agent's integration of multiple signals rather than a single threshold. Faithfulness is evaluated on the 53 generated reports using an LLM-as-judge protocol [16], with claims verified against the source event JSON; escalated events are excluded. Reports achieve a faithfulness score of 0.986, with 0.13 unsupported claims and 0.11 contradictions per report on average.

5 Conclusion

UrbanRisk-GFM is a framework for multitemporal landslide assessment that combines causal foundation-model adaptation, urban-aware evaluation, and a risk-reporting pipeline. Experiments on Sen12 show that standard segmentation metrics alone are insufficient for assessing performance near roads and buildings, motivating the proposed Urban-F1 and CMR metrics. Compared with standard bidirectional fine-tuning, causal temporal masking improves F1 from 0.452 to 0.638 and Urban-F1 from 0.448 to 0.641 while reducing CMR from 0.497 to 0.342, and the same causal model remains effective on incomplete temporal sequences without retraining. Beyond segmentation, the risk intelligence pipeline shows how uncertainty estimates and OSM infrastructure context can be turned into structured, traceable risk reports. Future work will extend urban-aware assessment to additional datasets and broader environmental settings, including a prefix non-causal baseline to isolate the contribution of causal masking from sequence-length effects, and region-held-out evaluation to assess cross-region transfer.

Acknowledgments. Generative AI tools were used for proofreading and grammar correction only; all scientific content, experimental design, analysis, and conclusions are the authors' own.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. World Health Organization. Landslides. `https://www.who.int/health-topics/landslides#tab=tab_1`
2. Guzzetti, F., et al.: Landslide inventory maps: New tools for an old problem. Earth-Science Reviews **112**(1–2), 42–66 (2012).
3. Szwarcman, D., et al.: Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth observation applications. IEEE Trans. Geosci. Remote Sens. (2025).
4. Tuia, D.: Artificial intelligence to advance Earth observation: A review of models, recent trends, and pathways forward. IEEE Geoscience and Remote Sensing Magazine, pp. 2–25 (2025).
5. Li, W., et al.: Landslide hazard mapping with geospatial foundation models: Geographical generalizability, data scarcity, and band adaptability. arXiv preprint arXiv:2511.04474 (2025). `https://arxiv.org/abs/2511.04474`
6. Ghorbanzadeh, O., et al.: The outcome of the 2022 Landslide4Sense competition: Advanced landslide detection from multisource satellite imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–11 (2022).
7. Höhn, P., et al.: Sen12Landslides: A multitemporal benchmark dataset for landslide detection using Sentinel-2 imagery. Scientific Data 12, 1–12 (2025).
8. Ronneberger, O.,et al.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., et al. (eds.) MICCAI 2015, LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015).
9. Garnot, V.S.F., Landrieu, L.: Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Proc. ICCV 2021, pp. 4872–4881. IEEE (2021).
10. Tarasiou, M., et al.: ViTs for SITS: Vision transformers for satellite image time series. In: Proc. CVPR 2023, pp. 10654–10663. IEEE (2023).
11. Tuia, D., et al.: Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE Geoscience and Remote Sensing Magazine **4**(2), 41–57 (2016).
12. Kuckreja, K., et al.: GeoChat: Grounded Large Vision-Language Model for Remote Sensing. arXiv preprint arXiv:2311.15826 (2024).
13. Irvin, J., et al.: TEOChat: A large vision-language assistant for temporal Earth observation data. arXiv:2410.06234 (2024).
14. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proc. ICML 2016, JMLR W&CP, vol. 48, pp. 1050–1059. PMLR (2016)
15. Anthropic: Claude. `https://www.anthropic.com` (2024)
16. Zheng, L., et al.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023).
17. Çiçek, Ö., et al.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 424–432 (2016).