

AI-Enabled Remote Sensing for Multi-Hazard Disaster Risk Management: Advances, Challenges, and Future Research Directions

Mani Bhushan*

Government Engineering College, Khagaria, 848203, Bihar

*Corresponding author email ID: mani.tuntun@gmail.com

HIGHLIGHTS

- Integrates AI and multi-source remote sensing for multi-hazard disaster risk management.
- Identifies key gaps in transferability, uncertainty, interpretability, and operational reliability.
- Proposes a trustworthy disaster GeoAI framework for decision-oriented disaster intelligence

ARTICLE INFO

Article History:

Received: 3 August 2026

Accepted: 25 September 2026

Published: 01 October 2026

Keywords:

Artificial Intelligence

GeoAI

Disaster risk reduction

Multi-Hazard

Remote sensing

ABSTRACT

Earth observation (EO) and artificial intelligence (AI) are increasingly integrated to support disaster risk management through rapid hazard detection, impact assessment, monitoring and decision support. Advances in machine learning, deep learning, transformers, multimodal learning and geospatial foundation models have substantially improved the analysis of satellite, synthetic aperture radar (SAR), thermal, LiDAR, UAV and ancillary geospatial data. However, high predictive accuracy does not necessarily translate into operational reliability, particularly when models are evaluated using geographically narrow datasets, random data partitions or individual disaster events. This critical review synthesizes AI-enabled remote sensing for floods, droughts, wildfires, landslides, earthquakes, tropical cyclones and compound hazards across the disaster-management cycle. Particular attention is given to data representativeness, spatial and temporal leakage, domain shift, sensor interoperability, class imbalance, interpretability, physical consistency, uncertainty calibration and computational latency. The review argues that disaster Geospatial Artificial Intelligence (GeoAI) should be evaluated not only through conventional predictive metrics but also through independent-event validation, geographic transferability, uncertainty calibration, timeliness and decision relevance. Future research should prioritize multimodal and process-aware learning, self-supervised pretraining, rigorously evaluated foundation models, open multi-hazard benchmarks and human-in-the-loop systems. The transition from accurate hazard mapping to transferable, uncertainty-aware and operationally credible disaster intelligence is identified as a key priority for the next generation of AI-enabled disaster risk management.

1. INTRODUCTION

Disasters arise from interactions among hazardous processes, exposure, vulnerability and capacity; they are therefore not reducible to the physical footprint of a hazard. This distinction is

central to the use of remote sensing in disaster risk management. Satellite and airborne observations can characterize inundation, burn scars, surface deformation, vegetation stress, terrain instability and damage to visible assets with remarkable spatial consistency, but they observe social vulnerability,

<https://doi.org/10.66132/ngce20260304>

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

NG Civil Engineering, 2(3), 2026

institutional capacity and many dimensions of loss only indirectly. A rigorous review of AI-enabled remote sensing must consequently distinguish between hazard mapping, impact assessment and the broader problem of disaster risk. The Sendai Framework for Disaster Risk Reduction 2015–2030 places understanding risk, strengthening governance, investing in resilience, and improving preparedness and recovery at the centre of disaster-risk reduction (UNDRR, 2015). These priorities create a clear role for EO, but also establish a higher standard than producing accurate thematic maps.

The EO landscape has changed rapidly. Long-running Landsat observations are now complemented by Sentinel optical and SAR missions, high-frequency geostationary platforms, thermal sensors, commercial very-high-resolution imagery, satellite soil-moisture and gravimetry products, airborne and terrestrial LiDAR, and rapidly deployable unmanned aerial vehicles (UAVs). These observations are increasingly combined with digital elevation models (DEMs), meteorological reanalyses, numerical forecasts, river gauges, seismic observations, infrastructure layers, volunteered geographic information and socioeconomic data. Operational services demonstrate that such integration is no longer hypothetical. The Copernicus Emergency Management Service, for example, supports preparedness, rapid mapping and recovery using satellite imagery and other geospatial data, with rapid products intended for delivery within hours or days of activation (Copernicus Emergency Management Service, 2026).

The analytical challenge has therefore shifted. Earlier remote-sensing workflows relied heavily on thresholding, spectral indices, change detection, object-based image analysis and conventional supervised classifiers. Machine-learning methods such as random forests (RF), support vector machines (SVM) and gradient boosting improved nonlinear classification and susceptibility modelling. Deep learning subsequently enabled hierarchical feature extraction directly from imagery, accelerating semantic segmentation, object detection and change detection across remote-sensing applications (Ma et al., 2019). Disaster research adopted convolutional neural networks (CNNs), U-Net-type encoder-decoders, recurrent neural networks (RNNs), long short-term memory (LSTM) networks and gated recurrent units (GRUs). More recently, attention mechanisms, vision transformers, self-supervised learning, multimodal fusion and geospatial foundation models have broadened the methodological landscape.

This rapid progression has created an important imbalance between model sophistication and evaluation rigor. A model can achieve excellent performance on a randomly held-out subset of image patches yet fail when deployed on a new river basin, a different building typology, another SAR incidence angle, a different wildfire regime or an event outside the training climatology. Spatial autocorrelation can make nearby training and test samples deceptively similar; event-specific preprocessing can leak information; class imbalance can make overall accuracy meaningless; and a deterministic probability map can appear authoritative even when its uncertainty is poorly calibrated. These issues are not secondary statistical details. In disaster response, false negatives can hide affected settlements or blocked roads, false positives can divert scarce resources, and late products can be operationally irrelevant even when they are technically accurate.

Recent hazard-specific reviews confirm both the progress and fragmentation of the field. Deep-learning reviews have documented increasingly sophisticated wildfire detection, mapping and prediction systems (Ghali & Akhloufi, 2023a, 2023b); systematic reviews of landslide research show rapid uptake of AI for detection, susceptibility assessment and monitoring while repeatedly identifying data quality, overfitting and generalization as unresolved problems (Akosah et al., 2024; Li et al., 2024). Flood research has similarly expanded from threshold-based water extraction to semantic segmentation, forecasting and hybrid approaches (Li et al., 2024), while earthquake studies increasingly use pre/post-event imagery and deep models for building-damage assessment (Dong & Shan, 2013; Remote Sensing, 2023). The literature is therefore rich in hazard-specific advances but weaker in cross-hazard synthesis of what makes an AI system operationally trustworthy.

This review addresses that gap. Its central thesis is that the next scientific transition is not from one neural architecture to another, but from benchmark-oriented hazard mapping to trustworthy multi-hazard disaster intelligence. The paper has five objectives: (i) to synthesize EO sensors and AI architectures across major disaster classes; (ii) to map these capabilities to preparedness, early warning, response and recovery; (iii) to identify recurrent methodological weaknesses that undermine operational claims; (iv) to evaluate emerging opportunities in multimodal learning, physics/process-aware AI and geospatial foundation models; and (v) to propose a validation framework in which predictive skill is necessary but insufficient for operational adoption. The resulting perspective is intended for remote-sensing scientists, GeoAI

researchers, hydrologists, geoscientists, disaster-risk specialists and agencies responsible for emergency mapping and decision support.

2. REVIEW SCOPE AND EVIDENCE SYNTHESIS

This article is a critical integrative review rather than a formal meta-analysis. The literature synthesis emphasizes peer-reviewed studies and authoritative operational sources relevant to AI, EO and disaster management, with particular attention to developments from 2019 onward and to recent reviews that consolidate hazard-specific evidence. Foundational datasets and older methodological papers are retained where they continue to shape current practice. The scope includes floods, droughts, wildfires, landslides, earthquakes, tropical cyclones and storms, and compound or cascading hazards. The AI scope spans conventional machine learning, deep learning, temporal models, transformers, self-supervised learning, multimodal learning and geospatial foundation models.

A deliberate analytical choice is to organize the review around the disaster-management problem rather than a chronology of algorithms. For each hazard, three questions are asked. First, what physical or impact signal is actually observable from EO? Second, what role does AI play—classification, segmentation, forecasting, representation learning, data fusion or decision translation? Third, what evidence exists that the resulting model remains reliable outside the conditions in which it was developed? This structure avoids a common weakness of technology-led reviews in which increasingly complex architectures are listed without evaluating whether they improve operational transferability.

The review also uses a layered definition of disaster intelligence. The first layer is hazard observation, such as floodwater, active fire or ground deformation. The second is exposure, including buildings, roads, cropland and critical facilities. The third is vulnerability and impact, which may require structural, demographic, economic and institutional information beyond imagery. The fourth is decision context: lead time, acceptable false-negative rate, spatial unit of action, resource constraints and confidence requirements. AI models that operate only on the first layer should not be described as complete disaster-risk models unless the remaining dimensions are explicitly incorporated.

3. REMOTE-SENSING OBSERVATION SYSTEMS FOR DISASTER MANAGEMENT

3.1 Optical multispectral and very-high-resolution imagery

Optical remote sensing remains the most intuitive source of disaster information because many hazards and impacts produce strong changes in reflectance, texture and land-cover structure. Landsat and Sentinel-2 provide open multispectral archives suitable for long-term monitoring of vegetation stress, surface water, burn severity, sediment redistribution and land-cover change. Very-high-resolution commercial satellites add building-scale detail for post-earthquake or storm damage assessment. In machine-learning workflows, optical bands are frequently augmented with spectral indices such as NDVI, NBR, NDWI or variants designed for water, vegetation and burned-area discrimination.

The principal limitation is not algorithmic but observational: cloud, smoke, haze and illumination can remove or distort information exactly when rapid mapping is most needed. Tropical cyclones and floods are often associated with persistent cloud, while wildfire smoke can obscure the surface. Optical change detection also depends on radiometric comparability, seasonality and phenological state. A deep model can learn these nuisance differences as easily as it learns disaster signatures. Accordingly, pre/post-event pairing should control for season and viewing geometry where possible, and model validation should test whether performance persists under realistic atmospheric and phenological variability.

3.2 Synthetic aperture radar (SAR)

SAR is indispensable for all-weather, day-and-night disaster observation. Microwave backscatter responds to surface roughness, moisture, geometry and structure, enabling flood delineation under cloud, surface-deformation analysis through interferometry, and characterization of landslides, soil moisture and damaged built environments. Sentinel-1 has been especially influential because of its open archive and global coverage. Sen1Floods11, for example, provides georeferenced Sentinel-1 data across 11 flood events and broad ecological settings, explicitly supporting research on generalizable flood mapping (Bonafilia et al., 2020).

SAR also illustrates why AI cannot be detached from sensor physics. Smooth open water often produces low backscatter, but wind-roughened water, emergent vegetation, double-bounce scattering in urban areas, layover and radar shadow complicate this relationship. Models trained on one incidence-angle distribution or landscape may therefore learn context-specific signatures. Multi-temporal coherence and pre-event references can improve discrimination, but they introduce dependence on archive availability and registration quality. Strong SAR disaster studies

should report polarization, orbit, incidence-angle treatment, speckle filtering, terrain correction and temporal baselines rather than presenting SAR chips as generic images.

3.3 Thermal, hyperspectral, LiDAR and UAV observations

Thermal observations are central to active-fire and volcanic monitoring and can also characterize urban heat and post-fire surface conditions. Hyperspectral sensing offers richer material and vegetation discrimination but remains constrained by data volume, atmospheric sensitivity and comparatively limited operational coverage. LiDAR provides three-dimensional information on terrain, canopy and built structures, supporting flood modelling, landslide morphology and structural assessment. These sensors are therefore highly complementary to optical and SAR systems, but their availability and revisit characteristics vary substantially.

UAVs occupy a different operational niche. They can deliver centimetre-scale imagery and structure-from-motion products over priority zones after earthquakes, landslides, floods and storms. Their flexibility is valuable where satellite revisit or resolution is insufficient, but coverage, battery life, airspace restrictions, weather, communications and processing latency constrain scale. AI models developed on UAV imagery also face a severe domain problem: camera type, altitude, viewing angle and acquisition protocol can differ between deployments. The strongest use case is therefore targeted, human-supervised mapping integrated with wider-area satellite products rather than assuming that UAV imagery alone provides scalable regional intelligence.

3.4 Multi-source fusion as the emerging default

The practical conclusion across sensor types is that no single observation system is universally adequate. Flood response benefits from SAR for cloud penetration, optical data for contextual interpretation, DEMs for topographic plausibility and exposure layers for impact assessment. Drought requires temporal integration of precipitation, vegetation, evapotranspiration, land-surface temperature, soil moisture and sometimes terrestrial water storage. Wildfire risk combines fuel, weather, topography and human influence, whereas earthquake impact assessment combines pre/post imagery with building footprints and infrastructure networks (Table 1). The research frontier is consequently shifting from “which sensor is best?” toward “how should heterogeneous modalities be fused without hiding differences in

resolution, uncertainty, temporal support and physical meaning?”

4. EVOLUTION OF AI METHODS: FROM CLASSIFICATION TO TRANSFERABLE REPRESENTATIONS

4.1 Conventional machine learning remains a relevant baseline

The rise of deep learning has sometimes encouraged an unproductive assumption that conventional machine learning is obsolete. In disaster research, that is not defensible. RF, SVM and gradient-boosting methods remain strong baselines when predictors are tabular, sample sizes are moderate and physically interpretable covariates dominate the problem. Landslide susceptibility, for example, is often represented through slope, curvature, lithology, land cover, rainfall and distance variables; drought studies frequently combine climatic and remotely sensed indices. In such settings, tree ensembles can be competitive, computationally efficient and easier to diagnose than large neural networks.

The correct comparison is therefore not “traditional versus modern” but whether additional model complexity produces reproducible out-of-domain gains. Studies should compare against tuned non-neural baselines under identical spatial or temporal partitions. Without that control, apparent improvements can reflect preprocessing, sample design or data leakage rather than architecture. This point becomes particularly important in small disaster datasets, where high-capacity models can memorize event-specific context.

4.2 CNNs and encoder-decoder segmentation

CNNs transformed disaster-image analysis by replacing manually engineered texture and spectral descriptors with learned hierarchical features. Encoder-decoder architectures such as U-Net are especially well suited to dense prediction because they combine semantic context with spatial localization. Flood extent, burn scars, landslide polygons and damaged structures can therefore be formulated as segmentation tasks. The wider remote-sensing literature demonstrates the breadth of CNN use across scene classification, object detection and segmentation (Ma et al., 2019). The limitation is that convolutional inductive bias does not solve dataset shift: a network may learn vegetation, roof material, illumination or regional morphology as proxies for damage. Strong performance within a dataset is consequently not proof of causal or physical understanding.

4.3 Temporal learning

RNNs, LSTM and GRU architectures introduced explicit temporal representation into drought monitoring, hydrological forecasting, wildfire-risk modelling and displacement prediction. Their main value is not simply that they process sequences, but that they can integrate antecedent conditions that static imagery misses. Drought severity reflects accumulated water deficits; wildfire danger depends on evolving fuel moisture and weather; slope deformation may accelerate before failure. Yet sequence models inherit problems of nonstationarity and temporal leakage. Normalization must be fitted on training periods only, future information must not enter engineered features, and evaluation should use forward or blocked temporal validation rather than random shuffling.

4.4 Transformers and attention

Transformers offer long-range contextual modelling and flexible multimodal fusion, which are attractive for disaster scenes where spatial relationships extend beyond local convolutional receptive fields. They can link flooded road segments to surrounding inundation, integrate long temporal sequences or fuse modalities through cross-attention. However, transformer superiority should not be assumed. Training cost, memory demand and data requirements can be substantial, while gains may shrink when comparisons control for pretraining and parameter count. For emergency applications, inference latency and hardware availability are part of model quality, not implementation details.

4.5 Self-supervised learning and geospatial foundation models

Self-supervised learning addresses one of EO's defining asymmetries: unlabelled imagery is abundant, but expert disaster labels are scarce. Masked-image modelling and contrastive approaches can pretrain representations on large archives before fine-tuning on small event-specific datasets. Remote-sensing foundation models extend this logic by learning general representations across large, heterogeneous EO corpora. RingMo demonstrated masked-image modelling at remote-sensing scale, while the NASA-IBM Prithvi line has emphasized multitemporal EO representation. Recent evaluation of Prithvi shows both the promise and the need for explicit domain adaptation rather than assuming universal transfer (Zhang et al., 2024).

Prithvi-EO-2.0 represents a further step toward global, multitemporal pretraining, using millions of Harmonized Landsat-Sentinel samples and incorporating temporal and location embeddings (Szwarcman et al., 2025). Such models are highly

relevant to disaster management because the defining operational problem is often rapid adaptation to an unseen event with few labels. The correct benchmark is therefore not only average performance across standard datasets. Disaster-specific evaluation should measure few-shot adaptation, geographic holdout, cross-sensor robustness, calibration under domain shift and computational requirements during rapid response. A foundation model that is marginally better on an in-domain benchmark but poorly calibrated after transfer may be less useful than a smaller specialized model.

5. HAZARD-SPECIFIC EVIDENCE AND CRITICAL SYNTHESIS

5.1 Floods: the most mature test bed for rapid disaster GeoAI

Flood mapping is arguably the clearest demonstration of both the strengths and limitations of AI-enabled EO. SAR provides cloud-independent observations, while optical imagery and DEMs add spectral and topographic context. Early operational methods used thresholding and change detection; current work increasingly uses semantic segmentation and multimodal learning. Sen1Floods11 was an important step because it assembled Sentinel-1 flood data across 11 events, six continents and broad ecological settings rather than focusing on a single local event (Bonafilia et al., 2020). SpaceNet 8 pushed the task closer to emergency decision-making by combining building footprints, road networks and flood attribution, recognizing that responders need to know not merely where water is located but which assets and routes are affected (Hänsch et al., 2022).

The central unresolved problem is generalization. Water signatures vary with wind, vegetation, urban morphology, soil moisture and radar geometry. A segmentation model can therefore produce plausible but wrong boundaries when transferred to a different floodplain. Topographic plausibility provides one partial safeguard: isolated flood pixels on steep slopes or disconnected from plausible flow paths should trigger review. Multi-temporal references can distinguish permanent water from event inundation, but acquisition timing remains decisive. A technically excellent model cannot map peak flooding if the satellite observes the area after recession.

Flood studies also expose the difference between mapping and forecasting. A post-event SAR segmentation model detects inundation that has already occurred; it is not an early-warning model. Forecasting requires meteorological and hydrological information, temporal lead time and usually process or time-series modelling. Reviews of deep learning in

flood research show expansion across simulation, monitoring and forecasting, but these tasks should not be collapsed into a single claim of “flood prediction” (Li et al., 2024). For operational evaluation, flood AI should report event-wise transfer, acquisition-to-product latency, false-negative rates in exposed areas, and uncertainty alongside conventional IoU or F1-score.

5.2 Drought: a spatiotemporal and definitional challenge

Drought differs fundamentally from sudden-onset hazards because it develops through persistent anomalies and can be defined meteorologically, agriculturally, hydrologically or ecologically. EO contributes vegetation condition, land-surface temperature, evapotranspiration, soil moisture and terrestrial-water-storage information, while precipitation and reanalysis products supply atmospheric forcing. AI is valuable because these variables interact nonlinearly across different temporal lags and spatial supports.

The main methodological risk is target ambiguity. A model trained to reproduce a vegetation index anomaly is not necessarily detecting hydrological drought; a model trained against SPI may largely reproduce precipitation history; and a model trained on reservoir or streamflow deficits addresses another component of the drought system. Recent systematic reviews emphasize multimodal data fusion and hybrid spatiotemporal architectures, but also reveal geographic concentration and inconsistent model reporting (August, 2026). A strong drought-AI study must therefore define the drought process, target variable and lead time explicitly and avoid treating all drought indicators as interchangeable.

Transferability is particularly difficult under climate nonstationarity. Models fitted to historical relationships between rainfall, vegetation and temperature may encounter combinations outside the training distribution. This makes uncertainty and process consistency essential. Future drought GeoAI should integrate water-balance constraints, phenological context and explicit out-of-time evaluation, with attention to whether the model remains calibrated during unprecedented extremes.

5.3 Wildfire: rare-event detection under rapidly changing conditions

Wildfire applications span active-fire detection, smoke and flame recognition, burned-area mapping, burn-severity assessment, susceptibility and spread prediction. Thermal observations from MODIS and VIIRS provide frequent active-fire

information, while Sentinel and Landsat support higher-resolution burned-area and severity analysis. UAV and ground imagery can provide detailed local surveillance. Reviews by Ghali and Akhloufi (2023a, 2023b) document rapid expansion of CNN-based classification, object detection and segmentation and increasing integration of satellite data into fire prediction workflows.

The key statistical problem is extreme class imbalance. Most pixels and most times contain no active fire, yet the rare positive class carries the operational value. Overall accuracy can therefore be almost meaningless. Precision-recall analysis, event-level detection rate and false-alarm burden are more informative. A second issue is nonstationarity in fuels, weather and fire regimes. Models trained on historical fires can fail during extreme combinations of heat, drought and wind. Fire-spread models additionally require process awareness because direction and rate depend on wind, slope, fuel continuity and moisture.

AI can add substantial value when it fuses these drivers, but purely image-based fire models should not be interpreted as complete risk systems. Susceptibility before ignition, active-fire detection after ignition, spread prediction and post-fire severity are distinct tasks with different observations, labels and lead times. Conflating them obscures what the model can actually support operationally.

5.4 Landslides: strong AI uptake, persistent inventory bias

Landslide research combines image-based detection, susceptibility assessment and deformation or displacement monitoring. Optical and very-high-resolution imagery can delineate fresh scars; SAR and InSAR provide deformation information; DEMs and LiDAR characterize terrain; and susceptibility models integrate slope, geology, rainfall, land cover and anthropogenic factors. Recent systematic reviews confirm widespread adoption of CNNs and other AI methods while identifying data standardization, robustness, overfitting and generalization as continuing limitations (Akosah et al., 2024; Li et al., 2024).

Inventory quality is the central constraint. Landslide databases are rarely complete, and mapping quality varies with image availability, event magnitude, terrain and analyst effort. “Non-landslide” samples may include unmapped failures, creating label noise. Random sampling can also place neighbouring slope units into both training and testing sets, inflating apparent performance because environmental conditions are spatially autocorrelated. Event-based or spatial-block validation is therefore essential for credible susceptibility claims.

Interpretability requires caution. Feature-importance rankings may identify slope or rainfall as influential, but they do not establish process causality. Correlated terrain variables can redistribute importance, while land-cover variables may proxy for elevation or human access. The stronger approach combines explainability with sensitivity tests, known geomorphic relationships and transfer to independent terrain and geological settings.

5.5 Earthquakes: rapid damage screening, not automatic structural diagnosis

Earthquake EO serves two major roles: measurement of surface deformation and rapid assessment of visible damage. InSAR provides spatially continuous deformation information, while high-resolution optical, SAR, LiDAR and UAV imagery support mapping of collapsed structures, debris and access disruption. Deep-learning reviews of earthquake disaster assessment describe growing use of object detection, semantic segmentation and change-detection architectures (Remote Sensing, 2023). Earlier remote-sensing assessments already emphasized that identifying the spatial extent of damage and assigning engineering damage levels are different problems (Dong & Shan, 2013). That distinction remains crucial in the AI era.

Roof appearance is an imperfect proxy for structural safety. A building can suffer severe internal or lateral-system damage without an obvious roof signature, while shadows, viewing geometry and pre/post registration differences can resemble damage. Consequently, satellite AI should generally be framed as rapid screening or prioritization unless validated against engineering inspection. Models must also transfer across building materials, roof types, urban density and construction practices, which vary strongly between regions.

Pre/post-event datasets have enabled major advances, but they can also encourage shortcut learning. If pre- and post-images come from systematically different sensors or acquisition conditions, a network may learn those differences. Robust evaluation therefore requires careful registration, balanced geographic partitions and testing across events rather than within a single disaster.

5.6 Cyclones, storms and compound impacts

Tropical cyclones illustrate why multi-hazard framing is necessary. The storm itself is an atmospheric system, but losses emerge through wind, storm surge, extreme rainfall, river flooding, coastal erosion, landslides and infrastructure disruption. Geostationary satellites and scatterometers support

storm monitoring, while SAR and optical imagery contribute inundation and damage assessment. A single “cyclone damage” model can therefore hide several physically distinct processes.

The appropriate AI architecture depends on the decision problem. Track or intensity estimation is a spatiotemporal forecasting task; storm-surge mapping requires coastal topography and hydrodynamics; post-event building damage is a change-detection problem; road accessibility is a network and segmentation problem. Multi-hazard systems should preserve these distinctions while integrating their outputs at the impact and decision layers.

6. FROM SINGLE HAZARDS TO COMPOUND AND CASCADING DISASTER INTELLIGENCE

Real disasters frequently involve sequences rather than isolated hazards: cyclone-surge-flood chains, earthquake-landslide-river-blockage cascades, drought-vegetation-stress-wildfire interactions, and wildfire-erosion-debris-flow sequences. The analytical difficulty is that each process operates at different spatial and temporal scales and may require different observations. A multi-hazard model is therefore not simply a multiclass image classifier.

A more defensible architecture is modular. Hazard-specific components estimate physically distinct states; an integration layer represents dependencies and exposed systems; and a decision layer translates joint conditions into consequences. Graph neural networks and knowledge graphs are promising because they can represent relationships among roads, rivers, buildings, utilities and hazards. Multimodal transformers can fuse heterogeneous observations, while probabilistic graphical or ensemble approaches can propagate uncertainty across modules. However, benchmark data for cascading hazards remain scarce, and causal relationships should not be inferred solely from co-occurrence.

The strongest near-term research opportunity is to move from hazard footprints to impact pathways. SpaceNet 8 exemplifies this shift by linking flood state to buildings and roads (Hänsch et al., 2022). Similar approaches are needed for wildfire evacuation routes, landslide blockage of transport corridors, cyclone damage to power networks and drought impacts on agricultural production. Such products are closer to the decisions agencies actually make (Figure1).

AI-enabled remote sensing across the disaster-risk-management cycle

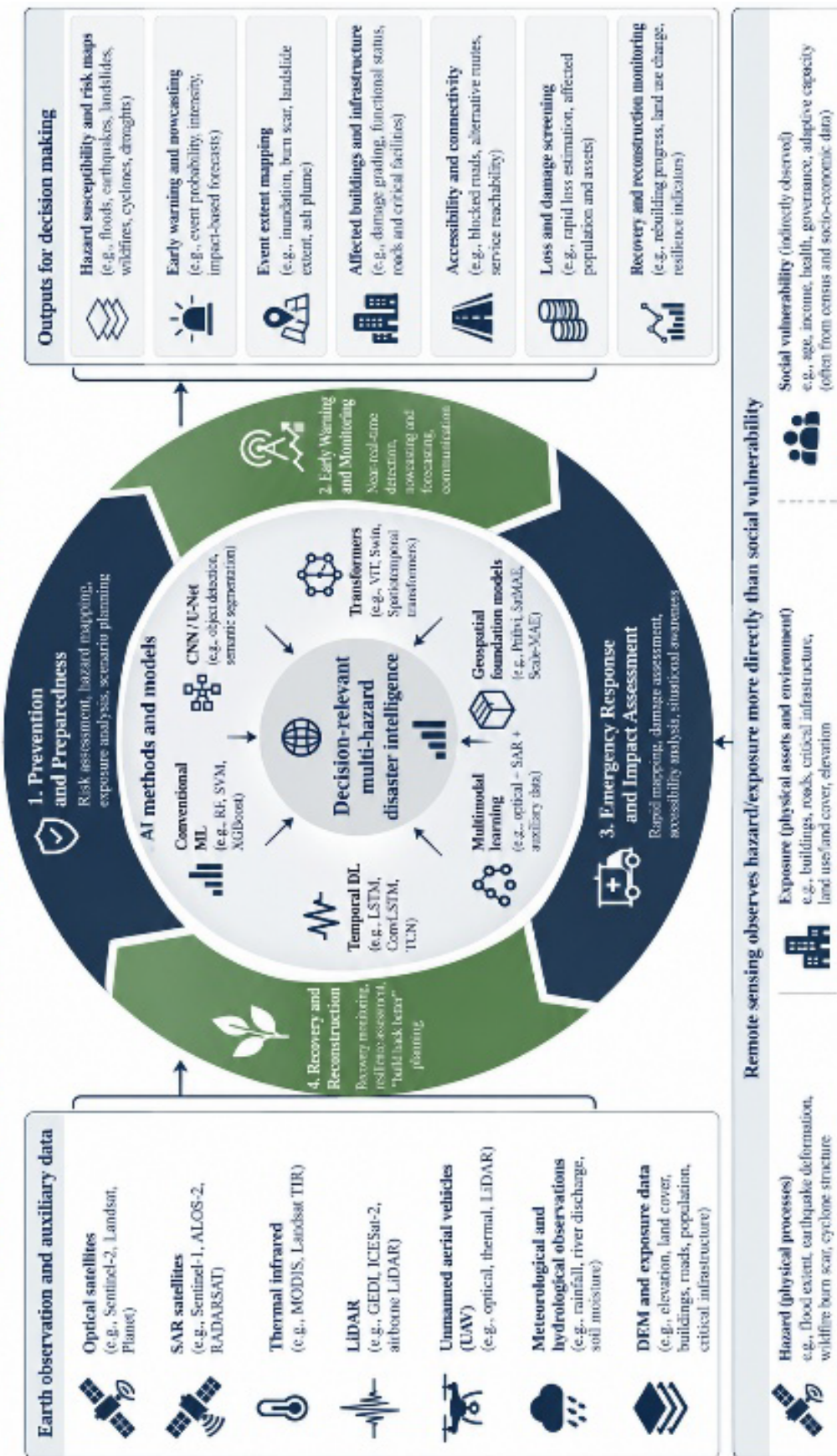


Figure 1. Integration of heterogeneous Earth-observation data and AI across the disaster-risk-management cycle.

7 WHY HIGH ACCURACY IS NOT OPERATIONAL RELIABILITY

7.1 Spatial, temporal and event leakage

One of the most consequential weaknesses in disaster GeoAI is evaluation design. Randomly splitting pixels or small patches assumes approximate independence, yet EO data are spatially autocorrelated. Adjacent patches share land cover, terrain, atmospheric conditions, sensor geometry and often the same disaster signature. When they are distributed across training and testing sets, the test set no longer represents an independent deployment environment. The resulting score may measure interpolation within an event rather than generalization to a new event.

Temporal leakage produces an analogous problem. Normalization, feature engineering or decomposition performed on the full-time series can expose future information to training. Random shuffling of drought or hydrological sequences can mix climatic regimes. Pre-event and post-event images can also encode acquisition-specific cues. The remedy is problem-dependent but conceptually simple: partitions should follow the intended deployment boundary. If the model will be used on a future event, the test event must be unseen. If it will be transferred to a new region, geographic blocks or regions must be held out.

7.2 Domain shift and transferability

Domain shift arises when the joint distribution of inputs and targets differs between development and deployment. In disaster EO, shifts occur across sensors, seasons, landscapes, building typologies, hazard magnitude, climate regimes and preprocessing pipelines. Cross-region performance should therefore be treated as a primary endpoint rather than an optional robustness test. Domain adaptation can help, but adaptation itself requires realistic assumptions about the amount of target-domain data available after an event. Few-shot and zero-shot claims are particularly relevant to foundation models and should be evaluated using strict event and geographic holdouts.

7.3 Class imbalance and metric selection

Disaster classes are often rare. A burned or damaged area may occupy only a small fraction of a scene; collapsed buildings may be far fewer than intact buildings. Overall accuracy can therefore remain high even when the model misses most operationally important cases. Precision, recall, F1-score, IoU, Dice coefficient and precision-recall area under the curve should be selected according to the task. For response

applications, class-specific recall and false-negative burden can be more consequential than a single aggregate score. Metrics should also be reported per event, not only pooled across all pixels, because large events can dominate global statistics.

7.4 Uncertainty and calibration

Most disaster maps are displayed as categorical products even when the underlying model is uncertain. This presentation can create false precision. Uncertainty arises from sensor noise, ambiguous labels, model parameters, environmental variability, domain shift and missing modalities. Bayesian approximations, deep ensembles, test-time augmentation, evidential approaches and conformal prediction provide different mechanisms for uncertainty estimation, but uncertainty quality must itself be evaluated.

Calibration asks whether predicted confidence corresponds to empirical correctness. A model that labels a pixel with 0.9 confidence should be correct at approximately that rate under the relevant deployment distribution. Domain shift often destroys calibration even when ranking performance remains acceptable. Reliability diagrams, expected calibration error, Brier score and coverage of prediction sets or intervals can therefore complement discrimination metrics. In operational mapping, uncertainty can be used to prioritize human review rather than merely appended as a supplementary map.

7.5 Interpretability and physical consistency

Explainable AI is useful but should not be confused with physical validity. Saliency maps, SHAP values or attention weights can identify model sensitivities, yet they do not prove that the learned relationship is causal or physically credible. Disaster models should be tested against process constraints where appropriate: flood outputs against terrain and connectivity; landslide susceptibility against geomorphic controls; wildfire spread against wind, fuel and slope; drought outputs against water-balance logic. Interpretability is strongest when post-hoc explanation, sensitivity analysis and process diagnostics agree (Table 2).

7.6 Latency, compute and reproducibility

Operational performance includes time. The relevant latency is not neural-network inference alone but acquisition → downlink → preprocessing → model execution → quality control → product generation → delivery. Copernicus rapid mapping explicitly targets emergency information within hours

or days, illustrating that product timeliness is a service-level property (Copernicus Emergency Management Service, 2026). Large foundation models may improve representation quality but impose compute and memory costs. Disaster studies should therefore report hardware, inference time, preprocessing time and data dependencies.

Reproducibility additionally requires code, model configuration, data provenance, partition definitions and versioned preprocessing; otherwise, high reported scores cannot be independently assessed (Figure 2).

Table 2. Trustworthiness dimensions that should accompany conventional predictive metrics.

Dimension	Question	Minimum evidence	Common practice	weak
Predictive skill	Does the model identify the target accurately?	Task-appropriate class metrics; per-event results	Overall accuracy only	
Temporal generalization	Does it work on a future/unseen event?	Event-wise or forward holdout	Random shuffle	temporal
Spatial transfer	Does it work in a new region?	Geographic/block holdout	Adjacent patches in train and test	
Sensor robustness	Does it survive sensor/acquisition changes?	Cross-sensor or acquisition-condition test	Single sensor/preprocessing chain	
Physical/contextual consistency	Are outputs plausible given process constraints?	Terrain/process/bounds diagnostics	Accuracy interpreted as physical validity	
Uncertainty calibration	Does confidence reflect error?	Calibration/coverage metrics; uncertainty map	Uncalibrated softmax probabilities	
Interpretability	Can important sensitivities be interrogated?	XAI plus sensitivity/process checks	One saliency figure without validation	
Latency	Is information available in time for the decision?	End-to-end processing time	Inference speed only	
Reproducibility	Can results be independently reproduced?	Data provenance, code/configuration, exact splits	Incomplete preprocessing/model details	

8. EMERGING FRONTIERS

8.1 Multimodal learning

Multimodal learning is likely to deliver larger operational gains than replacing one image backbone with another. Each modality has a different error structure: optical imagery is spectrally rich but cloud-sensitive; SAR is weather-independent but geometrically complex; DEMs provide static terrain constraints; meteorological and hydrological data provide forcing and temporal context; infrastructure and socioeconomic layers translate hazard into consequences. Effective fusion should preserve these differences rather than simply concatenate resampled rasters. Cross-attention, modality-specific encoders and late-fusion ensembles provide mechanisms to integrate evidence while allowing missing modalities.

Missing-data robustness is particularly important during disasters because the ideal sensor combination is rarely available at the required time.

8.2 Physics- and process-aware AI

Process-aware AI seeks to reduce physically implausible predictions and improve extrapolation by embedding domain knowledge in features, architectures, loss functions or hybrid models. Flood models can incorporate topographic connectivity, water balance or hydrodynamic outputs; drought models can incorporate water and energy constraints; wildfire models can use fuel, moisture, wind and terrain relationships; landslide models can use slope-stability concepts and rainfall thresholds. The goal is not to label a network “physics-informed” because a physical variable is included. Evidence should show

that the constraint changes behaviour in scientifically meaningful ways and improves performance or

reliability under conditions outside the training distribution.

Evolution of AI-enabled remote sensing for disaster management

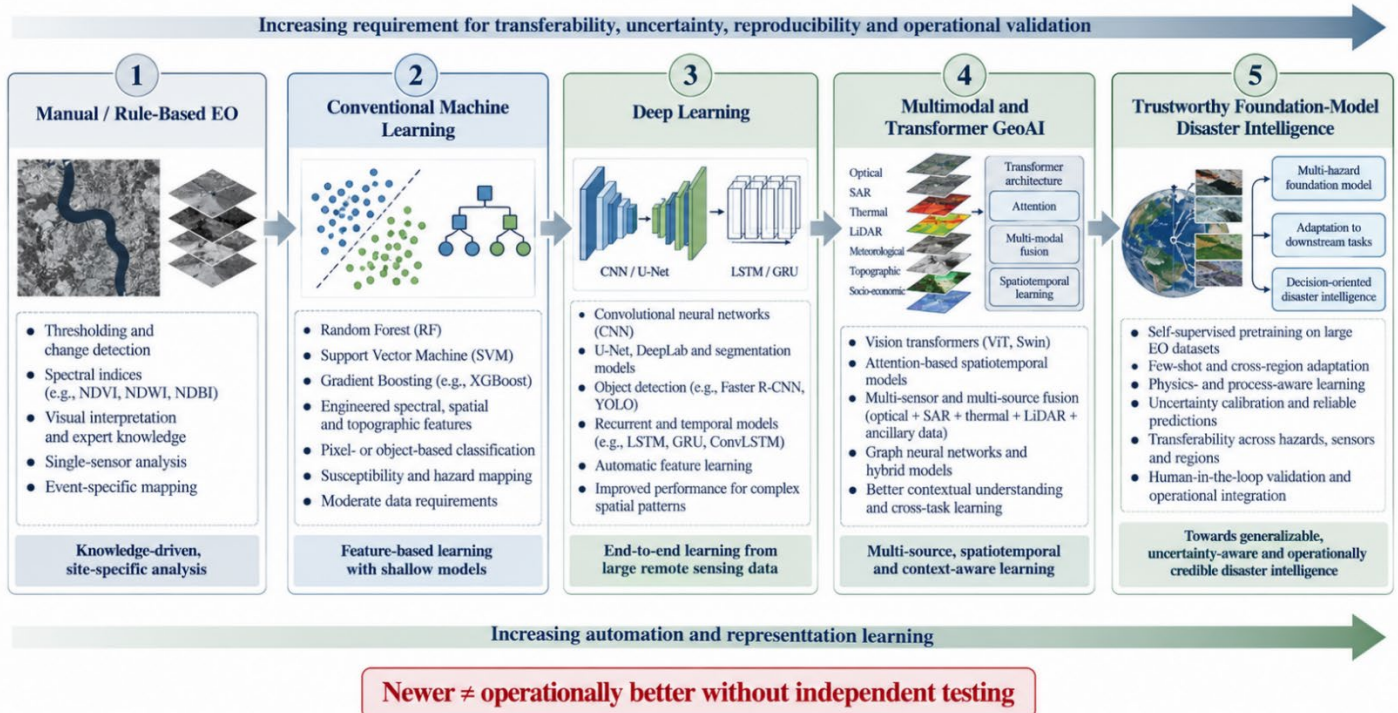


Figure 2. Conceptual figure prompt: Evolution from hazard mapping to trustworthy disaster GeoAI

8.3 Foundation models: opportunity with a demanding validation burden

Geospatial foundation models can reduce dependence on task-specific labels by pretraining on massive EO archives. RingMo demonstrated remote-sensing masked-image modelling, and Prithvi has provided an open, multitemporal EO foundation-model pathway. Independent evaluation of Prithvi emphasizes domain adaptability as a central issue rather than an automatic property (Zhang et al., 2024). Prithvi-EO-2.0 expands pretraining scale and global diversity and reports gains across heterogeneous EO benchmarks (Szwarcman et al., 2025).

For disasters, the most compelling use case is rapid adaptation. A new earthquake, flood or landslide event often offers few verified labels during the critical response window. A useful foundation model should therefore provide strong representations with minimal fine-tuning and should remain calibrated under geographic and sensor shift. Disaster benchmarks should compare frozen features, linear probing, full fine-tuning and parameter-efficient adaptation under controlled label budgets. They should also include compute, latency and energy requirements. Without these tests, “foundation model” risks becoming a scale descriptor rather than evidence of operational generality.

8.4 Human-in-the-loop and active learning

Emergency mapping is a high-consequence environment in which fully autonomous output may be inappropriate. Human analysts can recognize acquisition artefacts, contextual inconsistencies and rare failure modes that a model has never encountered. Uncertainty-aware active learning can focus analyst attention on ambiguous areas and then use verified corrections to improve the model. This creates a more realistic operational loop: automated screening → uncertainty triage → expert validation → product release → feedback. The value of AI is then measured partly by reduction in analyst workload and turnaround time rather than by replacement of human expertise.

8.5 Edge, cloud and distributed processing

Disaster response often occurs under bandwidth and infrastructure constraints. Cloud platforms facilitate large-scale EO processing, while edge inference on UAVs or local devices can reduce communication dependence. Model compression, quantization and efficient architectures therefore deserve more attention. The optimal system may use a large pretrained model centrally to produce transferable features and smaller task-specific models for rapid field deployment. Research should report

not only model accuracy but also memory footprint, throughput and operational dependencies.

9. PROPOSED FRAMEWORK FOR TRUSTWORTHY MULTI-HAZARD DISASTER GEOAI

A credible operational architecture can be represented as seven linked stages. Stage 1 is multi-source observation: optical, SAR, thermal, LiDAR, UAV, meteorological, hydrological, terrain, exposure and where appropriate socioeconomic information. Stage 2 is harmonization: georegistration, radiometric treatment, temporal alignment, resolution management, quality control and provenance. Stage 3 is representation learning using conventional features, CNNs, transformers, self-supervised encoders or foundation models. Stage 4 integrates modalities and process knowledge. Stage 5 produces task-specific inference such as detection, segmentation, forecasting, damage screening or risk estimation. Stage 6 is the critical gateway: trustworthiness validation. The output must be tested for predictive skill, independent-event performance, geographic transfer, sensor robustness, physical or contextual

consistency, interpretability, uncertainty calibration and reproducibility. Failure on one dimension does not always invalidate the model, but it limits the claim that can responsibly be made. For example, a model with excellent in-event segmentation but poor geographic transfer may still be useful after local retraining, but it should not be presented as globally generalizable.

Stage 7 is decision translation. Model probabilities and maps are converted into operational products at the spatial and temporal scale of action: warning zones, priority inspection areas, potentially isolated settlements, affected road segments, high-uncertainty regions requiring analyst review, or recovery indicators. Decision translation should include thresholds chosen according to consequences rather than generic 0.5 probability cutoffs. Feedback from analysts and field verification should return to the data and model stages, creating an adaptive rather than static system (Figure 3).

Stage 6 is the critical gateway: trustworthiness validation. The output must be tested for predictive skill, independent-event performance, geographic transfer, sensor robustness, physical or contextual consistency, interpretability, uncertainty calibration and reproducibility. Failure on one dimension does not always invalidate the model, but it limits the claim that can responsibly be made. For example, a model with excellent in-event segmentation but poor geographic transfer may still be useful after local retraining, but it should not be presented as globally generalizable.

Framework for trustworthy AI-enabled multi-hazard disaster intelligence

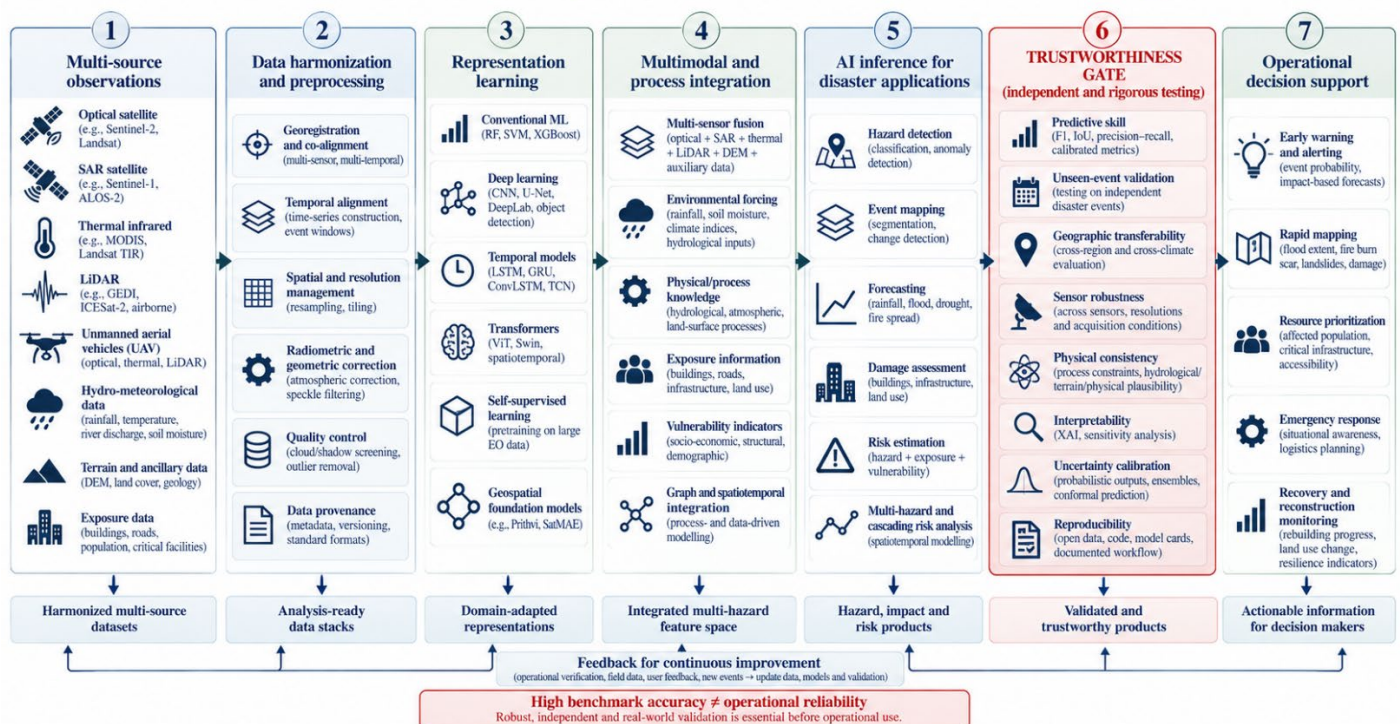


Figure 3. Conceptual figure prompt: Trustworthy multi-hazard disaster intelligence framework

10. RESEARCH PRIORITIES

10.1 Replace random splits with deployment-realistic validation

The benchmarks should require authors to justify how the train-validation-test design represents intended deployment. Event-wise, basin-wise, city-wise or geographic-block holdouts should become

routine. For temporal forecasting, forward validation should be the default. Random patch splits may still be useful for ablation studies, but they should not support claims of operational generalization.

10.2 Build open multi-hazard and multi-sensor benchmarks

The field needs datasets that connect hazard footprints with exposed assets, infrastructure functionality and verified impacts across multiple events and regions. Benchmark design should include

sensor metadata, uncertainty in labels and standardized geographic splits. Multi-hazard benchmarks should preserve process identity rather than reducing all impacts to a single damage class.

Table 3. Representative evidence base and its contribution to the review argument.

Study/source	Domain	Contribution	Critical implication
Ma et al. (2019)	Remote-sensing DL	Meta-analysis/review of >200 DL remote-sensing studies	Large training requirements and broad DL uptake establish baseline context
Bonafilia et al. (2020)	Flood	Sen1Floods11 multi-event Sentinel-1 dataset	Generalization requires geographically diverse event data
Hänsch et al. (2022)	Flood impacts	SpaceNet 8 buildings, roads and flood attribution	Decision relevance improves when hazard is linked to exposed assets
Ghali & Akhloufi (2023a,b)	Wildfire	Reviews of DL detection, mapping and prediction	Wildfire tasks differ in labels, lead time and class imbalance
Akosah et al. (2024)	Landslide	Systematic review of 186 studies	Data quality, model robustness and standardization remain limiting
Li et al. (2024)	Landslide	AI detection and susceptibility review	Overfitting and robustness remain central despite methodological maturity
Li et al. (2024)	Flood	Review of DL strategies in flood research	Flood AI spans distinct simulation, monitoring and forecasting tasks
Zhang et al. (2024)	Foundation models	Evaluation/adaptation of NASA-IBM Prithvi	Foundation-model transfer requires explicit domain adaptation
Szwarcman et al. (2025)	Foundation models	Prithvi-EO-2.0 multitemporal global pretraining	Scale improves versatility but disaster-specific transfer still requires testing
Copernicus EMS (2026)	Operations	Preparedness, rapid and recovery mapping	Timeliness and product delivery are operational performance dimensions
UNDRR (2015)	Policy	Sendai Framework	Disaster risk includes hazard, exposure, vulnerability and capacity

10.3 Make uncertainty a standard product

Future papers should report calibration and uncertainty whenever model outputs are intended to inform decisions. Benchmark leaderboards could include calibration error and selective performance, measuring how accuracy changes when uncertain predictions are deferred to human review. This would reward systems that know when they are likely to fail.

10.4 Evaluate foundation models under scarce-label emergency conditions

Foundation models should be tested under realistic label budgets and rapid adaptation windows. The decisive experiment is not unlimited fine-tuning on a mature benchmark; it is whether the model can adapt to an unseen event using a small number of verified examples while preserving calibration and acceptable latency.

10.5 Integrate process knowledge without overstating physical validity

Physics-guided losses, hybrid simulators and process-informed features should be evaluated using explicit physical diagnostics. A model should not be described as physically consistent merely because it uses elevation, rainfall or another physical predictor. Constraints must be linked to observable improvements in plausibility or extrapolation.

10.6 Shift evaluation toward decision utility

Disaster AI research should increasingly report the consequences of errors at the decision unit. For a road network, this may be missed inaccessible segments; for building screening, missed severely affected structures; for flood mapping, false negatives within populated areas; for wildfire detection, time to detection and false-alarm burden. Such metrics align algorithm development with emergency-management objectives.

10.7 Address geographic and institutional inequity

High-resolution commercial imagery, cloud computing and expert annotation are unevenly available. Models trained primarily in data-rich regions can underperform where disaster vulnerability is greatest. Open sensors, efficient models, transparent benchmarks and regional partnerships are therefore not only reproducibility measures but components of equitable disaster-risk reduction.

10.8 Establish model governance and versioned operational monitoring

Operational AI systems change as sensors, preprocessing, models and data distributions evolve. Agencies need versioned models, documented thresholds, audit trails, post-event performance review and mechanisms for detecting drift. Disaster GeoAI should be treated as a maintained socio-technical system rather than a one-time research model.

11. CONCLUSIONS

AI-enabled remote sensing has progressed from rule-based interpretation and conventional classification to dense deep-learning segmentation, spatiotemporal modelling, multimodal fusion and geospatial foundation models. Across floods, droughts, wildfires, landslides, earthquakes and storms, these methods can accelerate mapping, identify complex spatial patterns and integrate data volumes that exceed manual analytical capacity. Operational EO services demonstrate that satellite-

derived information already supports preparedness, response and recovery.

The review nevertheless identifies a consistent methodological gap: benchmark accuracy is often stronger than evidence of transferability. Spatially random splits, event-specific datasets, sensor dependence, incomplete inventories, class imbalance and uncalibrated confidence can make models appear more reliable than they are. The problem becomes more severe as claims move from hazard detection to risk or decision support, because exposure, vulnerability, consequences and institutional context extend beyond image classification.

The next phase of disaster GeoAI should therefore be defined by trustworthiness rather than architecture novelty. Independent-event and geographic validation, sensor robustness, uncertainty calibration, process and contextual diagnostics, reproducibility, latency and human review should accompany conventional predictive metrics. Multimodal learning and foundation models are promising precisely because they may improve adaptation across events and sensors, but those claims must be demonstrated under scarce-label, out-of-domain conditions.

The long-term objective is not an autonomous algorithm that produces a visually convincing map. It is an evidence-based system that transforms heterogeneous observations into timely, transferable and uncertainty-aware information while making its limitations visible to decision-makers. Progress toward that objective will require closer collaboration among EO scientists, AI researchers, hazard-process specialists, emergency managers and affected communities. When evaluated at this system level, AI-enabled remote sensing can move beyond increasingly accurate hazard maps toward operationally credible multi-hazard disaster intelligence.

DECLARATIONS

Conflicts of Interest: The author declares no conflicts of interest.

Funding: No external funding was reported in the source thesis.

Author Contributions: The sole author was responsible for the conceptualization, methodology, data collection and curation, analysis, visualization, interpretation of results, writing of the original draft, and review and editing of the manuscript. The author has read and approved the final version of the manuscript.

AI Disclosure: The authors declare that artificial intelligence (AI) tools were used solely to assist in

language refinement, organization, and editing of the manuscript. AI was not used to generate original research data, conduct analyses, or draw scientific conclusions. All technical content, interpretations, and final decisions were developed and verified by the authors, who take full responsibility for the accuracy, integrity, and originality of the work.

REFERENCES

- Akosah, S., Gratchev, I., Kim, D.-H., & Ohn, S.-Y. (2024). Application of artificial intelligence and remote sensing for landslide detection and prediction: Systematic review. *Remote Sensing*, 16(16), 2947. <https://doi.org/10.3390/rs16162947>
- August, O. (2026). Remote sensing and machine learning approaches for hydrological drought detection: A PRISMA review. *Water*, 18(3), 369.
- Bonafilia, D., Tellman, B., Anderson, T., & Issenberg, E. (2020). Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 835–845. <https://doi.org/10.1109/CVPRW50498.2020.00113>
- Copernicus Emergency Management Service. (2026). CEMS On-Demand Mapping: About the service. European Commission Joint Research Centre.
- Dong, L., & Shan, J. (2013). A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, 84, 85–99.
- Ghali, R., & Akhloufi, M. A. (2023a). Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. *Remote Sensing*, 15(7), 1821. <https://doi.org/10.3390/rs15071821>
- Ghali, R., & Akhloufi, M. A. (2023b). Deep learning approaches for wildland fires using satellite remote sensing data: Detection, mapping, and prediction. *Fire*, 6(5), 192. <https://doi.org/10.3390/fire6050192>
- Hänsch, R., Arndt, J., Lunga, D., Gibb, M., Pedelose, T., Boedihardjo, A. P., Petrie, D., & Bacastow, T. M. (2022). SpaceNet 8 – The detection of flooded roads and buildings. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1471–1479. <https://doi.org/10.1109/CVPRW56347.2022.00153>
- Li, H., Zhu, M., Li, F., & Skitmore, M. (2024). Solving flood problems with deep learning technology: Research status, strategies, and future directions. *Sustainable Development*, 32(6), 7011–7035. <https://doi.org/10.1002/sd.3074>
- Li, Y., Fu, B., Yin, Y., Hu, X., Wang, W., Wang, W., Li, X., & Long, G. (2024). Review on the artificial intelligence-based methods in landslide detection and susceptibility assessment: Current progress and future directions. *Intelligent Geoengineering*, 1(1), 1–18. <https://doi.org/10.1016/j.ige.2024.10.003>
- Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166–177. <https://doi.org/10.1016/j.isprsjprs.2019.04.015>
- Saleh, A., Zulkifley, M. A., Harun, H. H., et al. (2024). Forest fire surveillance systems: A review of deep learning methods. *Heliyon*, 10(1), e23127. <https://doi.org/10.1016/j.heliyon.2023.e23127>
- Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J., & Fu, K. (2023). RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–22. <https://doi.org/10.1109/TGRS.2022.3194732>
- Szwarcman, D., et al. (2025). Prithvi-EO-2.0: A versatile multitemporal foundation model for Earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64, Article 4400120. <https://doi.org/10.1109/TGRS.2025.3642610>
- United Nations Office for Disaster Risk Reduction. (2015). Sendai framework for disaster risk reduction 2015–2030. United Nations.
- Zhang, C., et al. (2024). Geospatial foundation models for image analysis: Evaluating and enhancing NASA-IBM Prithvi's domain adaptability. *International Journal of Geographical Information Science*, 39(9), 2096–2125. <https://doi.org/10.1080/13658816.2024.2397441>