

Operationalizing a Geospatial Foundation Model: Principles, Framework, Readiness Profile, Failure Modes, and Future Directions

Rajat Shinde^{1,*}, Venkatesh Kolluru¹, Abdelhak Marouane¹, Caden Helbling¹; Deepak Shah¹, Othneil Drew², Slesa Adhikari¹, Iksha Gurung¹, Manil Maskey³, Rahul Ramachandran³

¹Earth System Science Center, University of Alabama in Huntsville, Huntsville, AL 35805, USA

²Space and Earth Science Data Analysis, 7515 Mission Drive, Lanham, MD 20706

³NASA Marshall Space Flight Center, Huntsville, AL, 35808 USA

*Corresponding author

Rajat Shinde

Earth System Science Center, University of Alabama in Huntsville, Huntsville, AL 35899, USA

Email: rajat.shinde@uah.edu

Abstract

Geospatial foundation models (GeoFMs) are entering operational Earth observation (EO) pipelines for disaster response, agricultural monitoring, and environmental accounting. Although published performance on canonical benchmarks has improved, the conditions under which a deployed pipeline behaves predictably at scale remain poorly characterized. We argue that the operational behavior of a deployed GeoFM is predominantly shaped by the system surrounding the model. The system components consist of preprocessing protocols, reference products, post-processing steps that assemble chunked inferences into final geospatial products, serving infrastructure, and governance interfaces, and only secondarily by the foundation model in isolation. Yet, no existing evaluation practice systematically examines this scaffold. This article addresses that gap through three methodological contributions and a released implementation. First, we propose a unique systems framework that decomposes a deployed GeoFM pipeline into three pillars (data, model, deployment) and a governance layer and identifies eight axes (P1-P8) along which operational readiness can be characterized. Second, we introduce an Operational Readiness profile that converts the framework into a measurable vector with both structural and magnitude components per axis, computable for any deployed pipeline. Third, we apply the profile to deployed pipelines built on the NASA-IBM Prithvi-EO 2.0 GeoFM across three downstream tasks: burn scar detection, flood inundation mapping, and crop classification. The evaluation spans 17 operational case events and approximately 250 controlled red-team inferences validated against independent reference products per task family. The measurements identify failure modes that recur at the interfaces between pillars and are usually not surfaced by model-level evaluation. Two findings illustrate the stakes. First, the conventional HLS quality mask removes up to 17% of the burn target itself, misclassified as water, cutting per-event recall by up to 0.27 with no signal to the operator, while the identical layer serves as a validation reference in the flood pipeline. Second, evaluating the flood predictions against optical versus radar references yields statistically distinct cloud sensitivities (non-overlapping 95% confidence intervals). While Prithvi-EO 2.0 serves as our testbed, the profile itself is model-agnostic and extends to any deployed GeoFM pipeline, providing a transferable basis for assessing operational trustworthiness as these models enter production use.

The scripts, together with the profiling procedure packaged as a reusable, model-agnostic skill, are available on GitHub at <https://github.com/NASA-IMPACT/r2o-paper>. The operational platform underlying the use case is accessible on request at <https://explore.fm.odsii.io/>.

Index Terms - Foundation models, operational deployment, system evaluation, preprocessing protocols, AI governance.

1. Introduction

Earth observation (EO) has seen the rapid emergence of geospatial foundation models (GeoFMs) trained on petabyte-scale multi-sensor archives in recent years [1]-[5]. Several GeoFMs, released as open checkpoints fine-tunable for downstream EO, are consolidated in recent surveys [6]-[8]. Reported performance on canonical benchmarks [9]-[11] has shown improvement across burn-scar detection, flood mapping, land-cover classification, and crop-type identification. In parallel, the community has also begun to address the responsible development of these models [12]-[14] and the conditions under which their training data should be considered ready for use [15]-[17]. What remains underdeveloped is a comparable account of the *deployed pipeline* in which such a model operates, including (i) the data services that feed it, (ii) the preprocessing that binds inference inputs to the training distribution, (iii) the reference products against which its outputs are evaluated, (iv) the reconstruction logic that assembles geospatial outputs from chunked predictions, (v) the deployment infrastructure that serves users at scale, and (vi) the governance interfaces that determine what those users see.

The interaction of these components shapes the operational behavior of a deployed GeoFM. A GeoFM that excels on benchmarks can behave unpredictably once deployed in an operational pipeline. Additionally, accuracy is not the only currency of deployment: a real-time artificial intelligence (AI)-powered flood-mapping service that must return predictions within minutes may reasonably trade accuracy for latency and availability, requirements that a benchmark score neither captures nor ranks. Moreover, the failures can be specific, such as (a) the pipeline can silently violate the preprocessing protocol under which the model was trained; (b) the chosen reference product can shape apparent performance more than the prediction itself; (c) spectral or temporal modalities the model depends on can be absent or corrupted without an explicit warning being surfaced to the operator; (d) reconstruction or post-processing steps can introduce artifacts that are invisible at the tile-level metric; and (e) throughput, memory, and latency characteristics that determine whether the pipeline can serve users at scale follow non-monotonic curves that bear no relation to training time scaling laws [18], [19]. These are properties of the *system* in which the model is embedded rather than of any particular model release, and they recur across model families. A small but growing literature engages with this gap [20]-[23]. However, a general framework and standardized metrics for characterizing these failure modes across deployed GeoFM pipelines are still lacking.

The proposed work addresses this gap directly, building on several adjacent literatures on GeoFMs and EO to construct a framework in which the operational readiness of a deployed pipeline, and the failure modes that undermine it, become measurable. We organize the review around five threads: (i) the rapid development of GeoFMs and their open release; (ii) benchmarking and evaluation protocols proposed for these models; (iii) operational deployment of GeoFMs; (iv) production machine learning systems and data readiness; and (v) responsible and trustworthy AI for EO. This section closes by positioning the contribution.

i. GeoFMs for EO: GeoFMs have emerged over recent years as a distinct class of EO models, pretrained on multi-petabyte, multi-sensor archives under self-supervised objectives and released as open checkpoints that can be fine-tuned for downstream EO tasks. Major model families include the Prithvi series jointly developed by NASA and IBM [1], [2], SpectralGPT for spectral remote sensing [3],

RingMo for masked image modeling [4], RemoteCLIP for vision-language alignment in remote sensing [5], and a growing roster that includes SatMAE [24], SkySense [25], Clay [26], DOFA [27], TerraMind [28], and SatlasPretrain [29]. Xiao et al. [6] reviewed the field of remote sensing (RS) foundation models across visual, vision language model (VLM) and large language model (LLM) variants; Lu et al. [7] focused specifically on vision foundation models in RS; Vivone et al. [8] covered deep-learning-based image-fusion methods and protocols; and Wang et al. [30] released SSL4EO-S12, a foundational dataset and benchmark for self-supervised learning in EO. Earlier reviews from [31] and [32] established deep-learning and self-supervised-learning baselines, and Mai et al. [12] surveyed the opportunities and challenges of foundation models for GeoAI more broadly. Both the release artifacts and the surveys that consolidate them document the model checkpoint rather than the deployed system.

Table 1. Operational readiness dimensions as reported by the model cards, repositories, and papers of seven openly released GeoFMs (information as reported at time of access, Jul. 2026). "Yes/Partial/No" records whether the dimension is *tested and disclosed* in the released documentation, not whether the underlying model would pass the test. Footnotes give the evidence for non-obvious cells.

Model	Benchmark-tested	OOD / cross-region tested & reported	Missing-band robustness reported	Preprocessing protocol documented	Serving / scaling reported	Degraded-mode signaling	Per-run trust manifest	Weights open	Training data open
Prithvi-EO-2.0 [2]	Yes (GEO-Bench)	Partial	No	Yes	No	No	No	Yes (Apache-2.0)	Yes (HLS)
SatMAE [24]	Yes	No	No	Partial	No	No	No	Yes (research)	Yes (fMoW-Sentinel)
SkySense [25]	Yes (16 datasets)	Partial	No	No	No	No	No	Partial	No
Clay v1.5 [26]	Partial	Partial	No	Yes	No	No	No	Yes (Apache-2.0)	Partial
DOFA [27]	Yes	Partial	Partial	Partial	No	No	No	Yes (public)	Partial
TerraMind 1.0 [28]	Yes (PANGAEA)	Partial	Partial	Yes	No	No	No	Yes (Apache-2.0)	Yes (TerraMesh)
SatlasPretrain [29]	Yes	Partial	No	Partial	Partial	No	No	Yes (ODC-BY)	Yes (ODC-BY)

Seven representative GeoFMs were therefore selected because their public model cards, repositories, and papers permit cell-by-cell verification (Table 1) across the operational readiness dimensions developed in

this article [2], [24]-[29]. Table 1 quantifies the resulting asymmetry: the model-side dimensions (benchmark performance, weight availability, training-data openness) are broadly disclosed, whereas the deployment-side dimensions (serving and scaling behavior, degraded-mode signaling, per-run trust manifests) are tested and disclosed by none of the seven models, with a single partial entry. The framework we develop takes precisely these undisclosed dimensions as its object: each corresponds to an axis of the Operational Readiness profile, along which the missing evidence can be measured and reported.

ii. Benchmarking and evaluating geospatial foundation models: In parallel with model development, the community has also developed benchmarking suites for systematic evaluation. GEO-Bench [9] introduced twelve classification and segmentation tasks curated for Earth-monitoring relevance and well-suited to model comparison. PANGAEA [10] expanded the scope to a global, inclusive benchmark for GeoFMs, and PhilEO Bench [33] introduced n-shot and convergence-rate evaluation. REOBench [34] benchmarked the robustness of EO foundation models against image-level corruptions. The most recent contribution, GEO-Bench-2 [11], rethinks evaluation by introducing capability groups that rank models on shared characteristics (resolution, bands, temporality) and explicitly notes that no single model dominates across all tasks. This benchmark-driven evaluation culture is not unique to EO. In the adjacent weather and climate domain, WeatherBench 2 [35] standardizes the scoring of data-driven forecast models, and WxC-Bench [36] curates multi-modal, ML-ready datasets for downstream atmospheric tasks. These suites, like their EO counterparts, evaluate models under curated data distributions rather than in deployed pipelines.

These benchmarks provide controlled conditions for relative model comparison. However, the operational regime, in which the model interacts with a preprocessing layer, an inference and post-processing layer, a serving infrastructure, and an end-user consumer, is structurally distinct from what the benchmarks measure, and the failure modes that emerge at those interfaces are not what the benchmarks expose. Previous research also shows that the community has begun to acknowledge this gap explicitly [11]. Our framework complements these benchmarks: while standard benchmarks compare models under controlled input distributions, our framework profiles the operational context in which a chosen model is deployed. We therefore argue that a deployed pipeline can, and should, report both.

iii. Operational deployment of GeoFMs in EO: A small but growing literature collection engages directly with the question of how GeoFMs behave once placed into operational mapping systems. The WorldCereal team [20] reports a structured protocol for integrating GeoFMs into operational crop-mapping applications and explicitly documents that the integration of GeoFMs in operational mapping applications remains rare. They also present their three-step protocol that bridges controlled benchmark evaluation and the practical challenges of real-world deployment. SHRUG-FM [21] proposes reliability-aware foundation models for EO and connects model behavior to operational reliability indicators. Hsu, Li, and Wang [22] evaluate the domain adaptability of Prithvi specifically, surfacing the sensitivity of fine-tuned performance to deployment-time domain shift. A recent on-orbit demonstration [23] places a GeoFM onto a satellite payload, exposing additional operational constraints absent from ground-based deployment. Each of these cases examines a single deployment, model, or integration protocol. Their findings, though instructive, remain bound to the systems that produced them. Our contribution is complementary: a general framework and metrics for characterizing and exercising any

deployed GeoFM pipeline, demonstrated here in a single worked example as the first instance of a procedure intended to be repeated across systems.

iv. Production machine learning systems and data readiness: Broader literature originating in the engineering of large-scale machine learning (ML) services [18], [37] examines why models that perform well in development fail to serve reliably at scale, documenting hidden technical debt [18], deployment-stage failure modes [37], monitoring and degradation over the deployment lifetime [38], and the structural separation between research-stage and production-stage ML concerns [39]. Closest to our approach, Kolluru et al. [40] provide an end-to-end failure-mode study for assessing a deployed geospatial foundation model for flood-inundation mapping across nineteen out-of-distribution events and catalogue twenty-three distinct failure modes. The authors further find that pipeline engineering, rather than model capacity, dominates initial error. Their dual-reference validation further shows that part of the apparent model error reflects definitional inconsistency between reference products rather than detection failure.

A parallel body of work formalizes data readiness as a set of criteria (completeness, documentation, provenance, fitness-for-purpose, and machine-actionable metadata) [16], [17], which the EO community has codified for its own archives in an open data checklist [18]. Moreover, existing work on organizational AI readiness [41] identifies the technology-readiness factors that gate adoption in operational settings. Together these literatures supply the foundation upon which the present framework builds: the production-ML studies contribute the system-level perspective, and the data-readiness work contributes the inputs-side criteria. None of them, however, engages the operational interfaces specific to an EO pipeline (the preprocessing protocol, the reference products, and the reconstruction and serving layers), and it is at these interfaces that the measurements documented in Section 3 locate the consequential failure modes. The framework is organized around these interfaces, and the worked example extends the single-task study to a cross-task, per-axis assessment.

v. Responsible and trustworthy AI for Earth observation: Recent work has called for responsible development of GeoFMs and articulated taxonomies of risk for EO-based AI. Ghamisi et al. [13] proposed achievable and realistic paths for responsible AI in Earth observation, emphasizing stakeholder collaboration, domain involvement, and the regulatory context (NIST AI Risk Management Framework [42]; EU AI Act [43]; ISO/IEC 42001). Additionally, Zhu et al. [14] articulated the foundations of Earth and climate foundation models with explicit attention to trust and generalization. These contributions establish the need for operational accountability without specifying how such accountability would be measured or emitted by a deployed pipeline. Existing risk taxonomies remain categorical as they do not yield an instrument with which a specific deployed system can be characterized. The present article develops that instrument: the Operational Readiness profile, under which accountability becomes a measurable property of the deployed pipeline itself rather than an obligation asserted about it.

Overall, this article sits at the intersection of these five threads and argues that operational GeoFM deployment will require an evaluation dimension currently absent from the literature: a *measurable, pipeline-level account of operational readiness*. Consequently, we designed a system that is generalizable across GeoFM models, and this paper demonstrates it on one such model – NASA-IBM Prithvi-EO 2.0

[2] - as a first use case. Our core contributions are fourfold (framework, instrument, use case and implementation); for each, we state what it adds beyond the adjacent literatures reviewed above.

Framework. We develop a systems framework that decomposes the operational deployment of a GeoFM into three pillars (data, model, and deployment), along with a governance layer that ingests their emitted evidence. Within this framework, we articulate eight axes (P1-P8) along which the operational readiness of a deployed pipeline can be characterized. Each axis is anchored to a class of failure modes observable in deployed pipelines and measurable in principle. To our knowledge, this is the first framework in the GeoFM literature that takes the deployed pipeline, rather than the model checkpoint, as the unit of evaluation: existing benchmarks compare models under controlled input distributions [9]-[11], [35], [36], and existing Responsible-AI (RAI) taxonomies name risks categorically. Neither yields a decomposition of the operational system in which a chosen model runs.

Instrument. We convert the framework into a measurable artifact: the Operational Readiness profile, a vector of per-axis measurements computable for any deployed GeoFM pipeline and reportable alongside conventional accuracy metrics. Each axis measurement carries a structural component (whether the pipeline emits the relevant evidence at all) and a magnitude component (the observed severity of the corresponding failure mode where applicable). The profile turns the qualitative obligations of emerging AI-governance regimes [42], [43] into concrete claims that can be checked against the artifacts a pipeline emits in normal operation. The instrument's novelty is that it is computable and comparable: no existing benchmark, model card, or risk taxonomy yields a per-pipeline measurement that independent groups can compute and compare axis by axis. Section 2.2-D specifies the step-by-step procedure by which any deployed pipeline, regardless of the model, downstream task, or serving stack, can be profiled, and the appendix makes the measurement reproducible without proprietary instrumentation.

Use case. We illustrate the instrument by applying it to deployed pipelines built on Prithvi-EO 2.0 [2] across three downstream tasks: burn-scar detection, flood inundation mapping, and crop classification. The example spans 17 operational case events and approximately 250 controlled red-teaming inferences validated against independent reference products per task family. To the best of our knowledge, it is the first cross-task, multi-reference operational readiness assessment of a deployed GeoFM pipeline reported at this granularity. The measurements identify failure modes that recur at the interfaces between pillars rather than within any single pillar and quantify their magnitudes in a form that model-only evaluation does not produce.

Implementation. We release the framework as an executable readiness skill, so the profile is a tool, and not just a proposal. It packages the five-step procedure of Section 2.2-D, a per-axis scoring rubric with worked falsification anchors, a model-agnostic harness for the P2-P6 and P8 magnitudes, and a JSON reporting schema. It assumes nothing about the underlying pipeline: a user supplies a prediction callable and reference loaders through a documented adapter, and the resulting profile is directly comparable, axis by axis, with any other produced by the same procedure. Groups without perturbation capacity or budget can still report a structural profile from documentation review alone. By releasing the rubric with the code, we make the profile genuinely computable and comparable: shared anchors guide independent groups to the same vector and seed a community evidence base across GeoFMs.

We do not claim that the resulting evidence is representative of all GeoFMs or all downstream tasks. The profile is the instrument; the use case is the first instance of its application. We define that the five literature threads reviewed above situate the contribution. The proposed framework adopts the systems perspective of production ML but characterizes failure modes specific to deployed GeoFMs rather than to ML services in general. The profile then converts the qualitative claims of the responsible GeoAI literature into quantitative, per-pipeline evidence that can be computed, reported, and compared across systems. The worked example, finally, stands alongside the controlled benchmarks and individual operational case studies rather than in place of them.

2. Methods

This section presents the two methodological components of the study. Section 2.1 introduces a systems framework that decomposes a deployed GeoFM pipeline into three pillars (data, model, and deployment) and a governance layer and articulates the eight axes (P1-P8) along which its operational readiness is defined. Section 2.2 converts that framework into the Operational Readiness profile, a computable per-axis instrument, and specifies the step-by-step procedure by which any deployed pipeline can be scored. Together they define what is measured and how the measurement is produced, while the use case in Section 3 applies them to real deployments.

2.1. A Systems Framework for Operational Geospatial AI

We decompose the operational deployment of a GeoFM into three pillars and a governance layer (Fig. 1). The *data pillar* comprises sensor access, harmonization, preprocessing, and quality control. The *model pillar* comprises foundation model invocation and the adapter or fine-tuning heads attached to it for downstream tasks. The *deployment pillar* comprises reconstruction, post-processing, serving infrastructure, and the operational scaling characteristics that determine whether the pipeline is viable at the throughput required. The *governance layer* consumes evidence emitted by the three pillars and exposes it to auditors, operators, and end users.

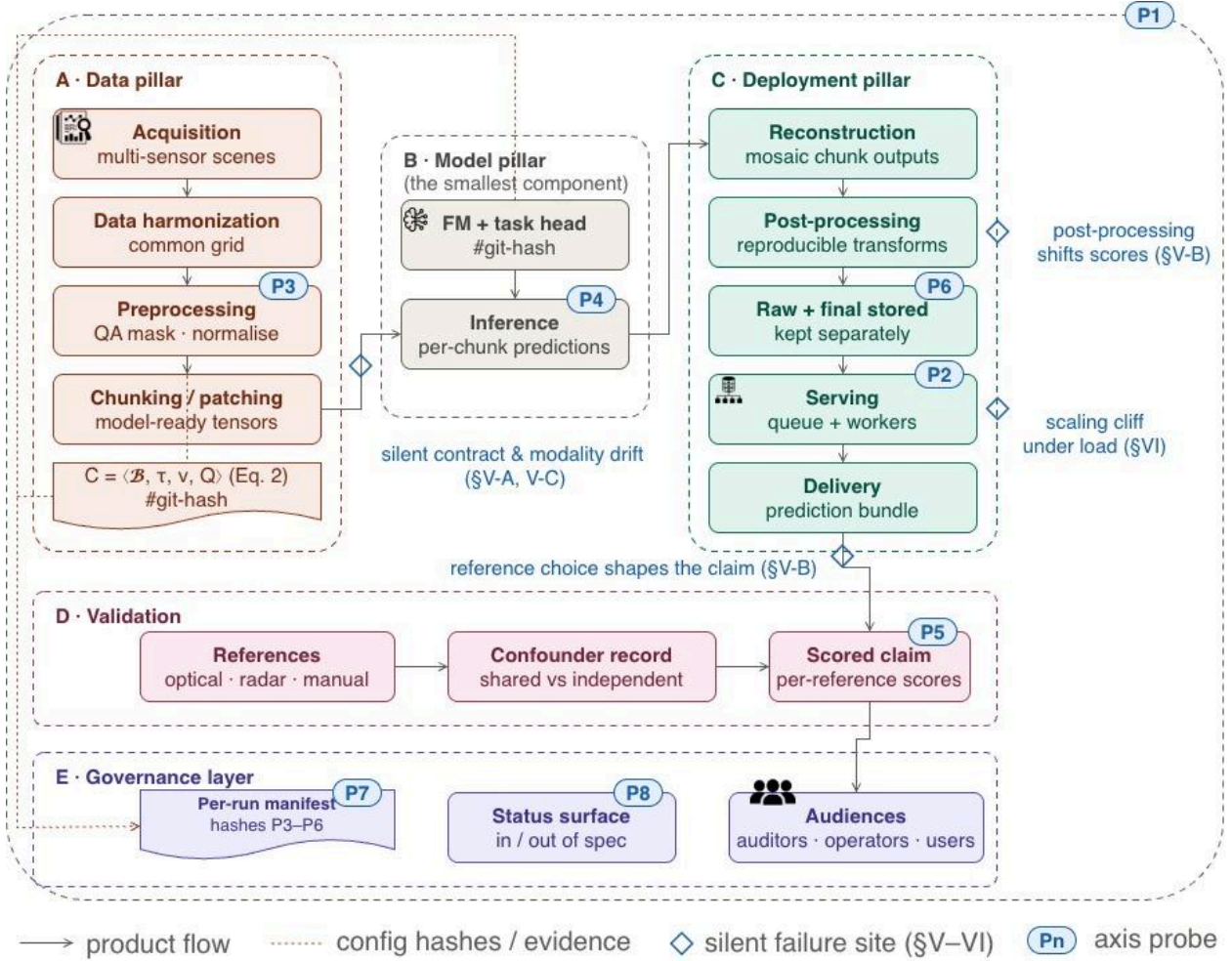


Fig. 1. Operational geospatial foundation models: the deployed pipeline, not the model checkpoint, is the unit of evaluation, and its operational readiness is measurable. Anatomy of an operational GeoFM pipeline, decomposed into the data (A), model (B), and deployment (C) pillars, the validation context (D), and the governance layer (E). The foundation model, deliberately drawn as the smallest component, is embedded in a system with preprocessing protocols (Eq. 2), reconstruction and post-processing rules, serving infrastructure, reference products, and governance interfaces that jointly determine operational behavior. Pills mark the eight axes of the Operational Readiness profile (P1–P8) at the components they characterize; diamonds mark the interfaces at which the worked example of Sections 3.1–3.2 observed silent failure modes; dashed leaders trace the git-hashed configuration artifacts into the per-run trust manifest (P7).

We characterize the operational readiness of a deployed pipeline along eight axes (P1–P8). Each axis corresponds to a specific pipeline failure mode, paired with a metric that can, in principle, quantify its magnitude. We adopt the layered design-tradeoff perspective of the production-ML systems literature [40]: each axis names a tradeoff space rather than a categorical best practice. Every deployed pipeline occupies a position in each of these spaces. The profile measurement makes that position explicit. We order the axes by their role in the framework: P1 states the operational-trust framing; P2 names the cross-cutting claim that distinguishes the operational regime from the training regime; P3–P6 are the four

pillar-internal axes (data, model, validation, deployment); and P7-P8 are the governance axes. Each axis is paired in Section 3.1 with the empirical observation that anchors it; these falsification anchors are summarized, alongside the measured profile, in Section 3.1D.

P1. Operational trust as an observable system property: The behavior of a deployed GeoFM pipeline is determined jointly by its data, model, and deployment components. The evidence required to establish operational trust is therefore emitted at the level of the pipeline rather than at the level of any single component. This axis is a framing claim rather than an empirical one; its function in the profile is to make explicit that the other seven axes are its measurable constituents.

P2. Operational scaling is distinct from training scaling: Published scaling laws relate compute, data, and parameters to held-out accuracy [19] which are properties of the training regime. However, at deployment, the operational object is different. We characterize it as a multi-objective vector

$$\mathbf{O}(\theta) = (T(\theta), M(\theta), L_{(p99)}(\theta), R_{(c)}(\theta)) \quad (1)$$

where T is throughput, M is peak memory, $L_{(p99)}$ is the tail latency, $R_{(c)}$ is condition-dependent reliability over an operational distribution of acquisition conditions c , and θ parameterizes the deployment configuration. The components of $\mathbf{O}$ are typically non-monotonic in θ , so deployment-readiness inferences cannot be drawn from training-scale evidence alone. This axis cuts across the three pillars and the governance layer. It is the axis along which the framework departs most explicitly from prior model-centric assessments of foundation model readiness. We note in Section 3.2 that its constituents are not unique to GeoFMs. The framework's contribution here is to make them a required component of the operational profile rather than to identify them as EO-specific.

P3. Preprocessing as a formal protocol: A foundation model is trained under a specific input preparation regime, comprising a band selection, a normalization transform, a *nodata* convention, and a quality-mask configuration. Inference under any other regime constitutes a *preprocessing-protocol violation*. We define a preprocessing protocol as a tuple

$$C = \langle \mathcal{B}, \tau, v, Q \rangle \quad (2)$$

where $\mathcal{B}$ is the set of input bands, τ the normalization transform, v the *nodata* convention, and Q the quality-mask configuration applied to the input prior to inference. For a model trained under protocol C_{train} and deployed under protocol C_{deploy} , the protocol violation is a distance $d(C_{\text{train}}, C_{\text{deploy}})$ that admits per-component decomposition. We choose these four components because they exhaust the transformations the pipeline can apply to the input tensor without invoking the model. A pipeline that does not record either protocol emits no evidence of violation. The silent case is the operationally consequential one.

P4. Per-task, per-temporal-slot modality disclosure: The dependence of a downstream task on individual spectral bands or temporal acquisitions is task-conditional. A foundation model can be tolerant of a band corruption for one downstream task and collapse under the same corruption for another. The unit at which modality dependence should be disclosed is therefore the (model, task, temporal-slot) triple, not the foundation model in isolation. We denote the modality-dependence function

$$\Delta(t,b,s) = \mathbb{E}[\mathcal{L}(y, \hat{y}(t)) - \mathcal{L}(y, \hat{y}(b,s \rightarrow 0)(t))] \quad (3)$$

where t is the downstream task, b is the band, s is the temporal slot (where applicable), $\hat{y}(t)$ is the prediction under nominal input, and $\hat{y}(b,s \rightarrow 0)(t)$ is the prediction under the corresponding band-slot zero corruption.

P5. Reference-target semantic alignment: The apparent performance of a GeoFM pipeline is shaped by the choice of reference product against which its predictions are scored. Reference products affected by the same conditions as the prediction conflate degradation of the reference with degradation of the model. An optically derived water mask used to validate an optical surface-water prediction, for example, is obscured by the same cloud cover that corrupts the prediction; the measured cloud dependence then reflects both and cannot be attributed to the model alone. An honest performance claim requires a reference product that is at least partially independent of the conditions that degrade the prediction, and a deployed pipeline should record which reference was used and the assumptions behind its semantic alignment with the model target.

P6. Observable spatial reconstruction: Chunked inference followed by post-processing steps such as mosaicking, terrain-shadow removal, aerosol overcorrection, or vegetation filtering can alter the prediction surface in ways invisible at the tile-level metric. A deployed pipeline should record both the raw inference output and the post-processed result, separately, with the corrections expressed as reproducible transforms.

P7. Pipeline-emitted trust evidence: Typically, audit and compliance evidence assembled after the fact has two weaknesses: it cannot inform the decision it is meant to legitimize, because it did not exist when the decision was made, and it depends on reconstructing the pipeline's configuration from whatever logs happen to survive. A pipeline that emits its trust evidence as part of normal operation avoids both: the evidence is complete by construction, and it reaches auditors, operators, and end users at the point of decision rather than in hindsight.

P8. Degraded operating modes as first-class outputs: Cloud cover, sensor anomalies, smoke plumes, and acquisitions outside the operational specification produce degraded predictions. By operational specification we mean a documented envelope of acquisition conditions under which the deployed pipeline's outputs are supported. Users are nonetheless entitled to know when outputs are degraded, and a deployed pipeline should signal such regimes explicitly rather than allowing them to propagate silently. We treat the operational specification itself as a documented artifact that the deployment operator publishes and updates.

The eight axes are not independent. P3 (preprocessing protocol) and P4 (modality disclosure) jointly constrain the data-model interface; P6 (reconstruction observability) and P7 (trust evidence) jointly constrain the model-deployment and deployment-governance interfaces; and P2 (operational scaling) cuts across all three pillars. We treat them as axes of a profile vector rather than as independent obligations, and the instrument introduced in Section 2.2 makes that vector measurable.

2.2. The Operational Readiness Profile

The framework described in Section 2.1 is qualitative. To make it operationally useful, we introduce the Operational Readiness profile, a vector of per-axis measurements computable for any deployed GeoFM pipeline. The profile is intended to be reported alongside conventional accuracy metrics whenever such a pipeline is published, deployed, or evaluated.

For a deployed pipeline π , the profile is a vector

$$\mathbf{\Pi}(\pi) = (\pi_1, \pi_2, \dots, \pi_8), \quad \pi_i \in \mathcal{S}_i \times \mathcal{M}_i \quad (4)$$

where each axis measurement comprises a *structural* component $\mathcal{S}_i$ (a binary or low-cardinality categorical indicating whether the pipeline emits the relevant evidence at all) and a *magnitude* component $\mathcal{M}_i$ (a continuous measurement of the failure-mode severity, where applicable). For the preprocessing-protocol axis P3, for instance, $\mathcal{S}_3$ records whether the deployed pipeline emits a hashed representation of its preprocessing protocol, and $\mathcal{M}_3$ records the measured performance change under a controlled protocol perturbation. A pipeline that scores well on $\mathcal{S}_3$ but poorly on $\mathcal{M}_3$ has documented its protocol honestly and made the consequences of its choice visible; a pipeline that scores poorly on $\mathcal{S}_3$ has accepted silent protocol-drift risk regardless of what its accuracy numbers show. Table 2 extends this construction to all eight axes, stating for each the structural criterion, the magnitude quantity, and the falsification anchor that grounds the score.

Table 2. Per-axis measurement protocol. Each row specifies what the axis measures, how the structural and magnitude components are computed and which finding from the worked example anchors the axis. Structural components take the values yes / partial / no, determined by inspection of the artifacts the pipeline emits in normal operation: *yes* indicates the evidence is emitted and complete, *partial* that it is emitted in reduced or implicit form, and *no* that it is absent.

Axis	Measurement Dimension	Structural component $\mathcal{S}_i$	Magnitude component $\mathcal{M}_i$	Anchor in worked example
P1	System-level trust framing	Yes / No: pipeline scope is documented as system-level	<i>Qualitative; not a magnitude</i>	§2.1–2.2 framing
P2	Operational scaling	Yes / No: pipeline reports $\mathbf{O}(\theta)$ across an operational envelope	Δ error rate over the (throughput, memory, latency, condition) grid	§3.2
P3	Preprocessing-protocol fidelity	Yes / Partial / No: pipeline emits hashed $C = \langle \mathcal{B}, \tau, v, Q \rangle$ per run	Max Δ accuracy under protocol perturbation sweep	§3.1-A
P4	Per-task, per-slot modality dependence	Yes / No: pipeline records (band, slot) dependence per task	$\Delta(t, b, s)$ per Eq. (3)	§3.1
P5	Reference-target alignment	Yes / No: reference product and its documented confounders are recorded	95% CI gap in dose-response Pearson r between references	§3.1-B

			sharing and not sharing the confounder	
P6	Observable spatial reconstruction	Yes / No: raw inference and post-processed result are stored separately	Δ metric between raw and post-processed outputs under a documented reconstruction rule	§3.1-B
P7	Pipeline-emitted trust evidence	Yes / Partial / No: per-run manifest cites Git hashes for preprocessing, model, post-processing	<i>Structural only</i>	§3.2-B
P8	Degraded-mode signaling	Yes / Partial / No: consumer-visible signal indicates in-specification operation	Δ accuracy at documented degradation threshold vs baseline	§3.1-B, §3.2-C

Applying the protocol to the three worked-example pipelines yields the profiles reported in Section 3.1 (Fig. 8). For the profile to function as a shared instrument rather than a report specific to one system, it is constructed to satisfy the following three properties.

A. Computability. Each axis measurement is operationally defined in terms of artifacts the pipeline either emits or can be made to emit. The structural components are checked from configuration and documentation, whereas the magnitude components are obtained from controlled perturbation experiments of the type reported in our worked example. The per-axis measurement protocol is specified in Table 2 and in the appendix.

B. Comparability. The profile is reported as a fixed-length vector in a shared schema. Two deployed pipelines, built on different foundation models and serving different downstream tasks, can be compared on each axis individually. The profile does not impose a single scalar readiness score; instead, it exposes the tradeoffs. Where axis measurements are rendered numerically for visualization (as in the radar of Fig. 8), the scores are ordinal summaries of the underlying structural and magnitude components, not values on a common scale across axes. The falsification anchors in Fig. 11b ground each score in a specific observation from Section 3.1.

C. Governance interpretability. The structural components of the profile map to obligations articulated in current AI-governance regimes (NIST AI Risk Management Framework, EU AI Act, ISO/IEC 42001 [42], [43]). An Extended Data table gives the correspondence for the worked example, pairing each structural component with the obligation it evidences. The table is a proposed alignment rather than an endorsement, and the framework commits to no single regime. The profile is instead constructed so that, whichever regime applies, the required evidence is drawn from artifacts the pipeline already emits.

2.2.1. Applying the profile to a new pipeline.

The profile is designed so that the operator of any deployed GeoFM pipeline, including one whose foundation model, downstream task, and serving stack differ entirely from those of our worked example, can compute it by following a fixed, self-contained five-step procedure.

Step 1: Document the deployed preprocessing protocol $C = \langle \mathcal{B}, \tau, v, Q \rangle$ of Eq. (2) and, where available, the training-time protocol released with the model.

Step 2: Determine the eight structural components $\mathcal{S}_1\text{-}\mathcal{S}_8$ by inspection of the pipeline's emitted artifacts, using the Yes / Partial / No criteria of Table 2. This step requires no experiments.

Step 3: Assemble a small set of operationally meaningful case events for each downstream task, with at least one independent reference product per task and, where possible, one reference at least partially independent of the conditions that degrade the prediction.

Step 4: Run the perturbation experiments of Supplementary Material B with the reference implementations released with this article: namely the protocol sweep (P3), band and band-slot corruption (P4), multi-reference scoring (P5), reconstruction-scenario scoring (P6), the load sweep (P2), and degradation-threshold scoring (P8). Record the resulting magnitude components $\mathcal{M}_i$.

Step 5: Report the populated vector $\mathbf{\Pi}(\pi)$ in the shared JSON schema released with this article (Supplementary Material D) alongside the pipeline's conventional accuracy metrics. A pipeline that completes only Steps 1-2 still produces a publishable structural profile. The magnitude components can be added incrementally as perturbation capacity permits. Section 3.1 executes this procedure end-to-end on three pipelines and serves as a template for new applications.

3. Results

This section reports the results of applying the framework and the operational readiness profile to real deployments. Section 3.1 presents the use case in which pipelines built on Prithvi-EO 2.0 are profiled across three downstream tasks (burn-scar detection, flood inundation mapping, and crop classification), spanning 17 operational events and approximately 250 controlled inferences, and reports the failure modes and magnitudes measured at each axis. Section 3.2 reports the deployment infrastructure through which these pipelines are served, showing how the framework's requirements for observability, versioning, and degraded-mode signaling are realized in an operational system and how it behaves under load.

3.1. Use case: Deploying Prithvi-EO 2.0 across Three Downstream Tasks

We applied the Operational Readiness profile to deployed pipelines built on the Prithvi-EO 2.0 foundation model [2] across three downstream tasks: burn-scar detection, flood inundation mapping, and crop classification. Seventeen operational case events (Table 3) were processed through the production inference path of the corresponding task and subjected to a set of controlled input perturbations. Throughout, we tie each finding to the profile axes it anchors (Section 2.1). Sections 3.1-A to 3.1-C report the three pipelines in the order listed above, and Section 3.1-D synthesizes the findings across tasks.

Validation used one to four independent reference products, depending on the task. Burn-scar events were validated against official fire-perimeter polygons published by the responsible agencies (CalFire/NIFC, NPS/InciWeb, and Victoria Emergency/EFFIS; Table 3). Flood events were validated against up to four products spanning optical and radar provenance (Section 3.1-B), selected so that at least one reference does not share the cloud confounder that degrades the optical prediction. Crop events were validated against the USDA Cropland Data Layer, supplemented by the AAFC Annual Crop Inventory for the cross-border location. Per-event assignments are given in Table 3, and the documented confounders of each product are cataloged in Appendix Table B1.

A. Burn-scar Detection

Five wildfire events spanning Texas, Arizona, California, and Australia were processed through the Prithvi-EO 2.0 burn-scar production pipeline (300M-parameter variant, fine-tuned on HLSBurnScars [44], six-channel Harmonized Landsat and Sentinel-2 (HLS) input). Each event was perturbed under two protocols: a **quality-mask configuration sweep**, in which the *Fmask* cloud, shadow, snow/ice, and water bits were applied to the input individually and in combination across eight configurations prior to inference; and **spectral-band dropout**, in which one spectral band at a time was zeroed prior to normalization (six bands). The nominal baseline is the production-default configuration, in which all four *Fmask* bits are applied, as is conventional for HLS land products; the QA-relaxed configuration applies to none of them. Pixels without valid data (tile *nodata*, chunk padding) carry a neutral fill value (10^{-4}) in all configurations. Reference labels were rasterized from the fire-perimeter polygons published by the agency responsible for each event (Table 3).

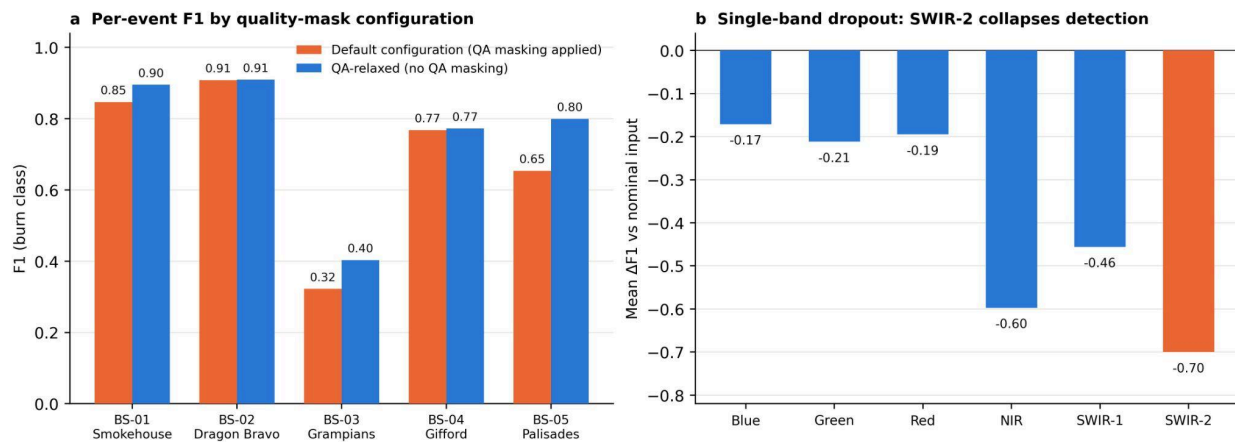


Fig. 2. Burn-scar pipeline: per-event accuracy and spectral dependence. (a) Per-event F1 of the burn class under the production-default configuration in which the four *Fmask* quality bits (cloud, shadow, snow/ice, water) are applied to the model input, as is conventional for HLS land products, and under the QA-relaxed configuration, in which none is applied. Relaxing the quality mask raises F1 by 0.15 at the coastal Palisades event (BS-05), where the *Fmask* water heuristic flags 17% of the burn scar and by 0.08 at the Grampians event (BS-03); the three inland events change by 0.05 or less. (b) Mean $\Delta F1$ under single-band dropout: SWIR-2 removal collapses detection to $F1 = 0.00$ at every event, while the other bands cost between 0.17 and 0.60.

Table 3. Case events used in the worked example. Seventeen operational case events across three downstream tasks, spanning five continents. Centroid coordinates are approximate representative points for each AOI; per-event bounding boxes are provided in the Extended Data. (BS- Burn Scar, FL- Flood and CR - Crop tasks). The dates listed reflect each task’s acquisition design. Burn-scar events use a single post-event acquisition. Flood events use a matched pair of acquisitions, one clear and one cloud-affected, which supports the cloud-substitution protocol of Section V-B. Crop events use three acquisitions spanning the growing season (T1-T3) plus one additional cloud-affected acquisition that substitutes for the mid-season slot in the cloudy-T2 protocol of Section V-C. Centroid coordinates are approximate representative points for each AOI.

ID	Event / site	Country	Centroid (Lat, Lon)	Acquisition date(s)	Reference product(s)	Significance
BS-01	Smokehouse Creek Fire, Texas Panhandle	USA	35.81°N, −100.73°W	2024-03-02	CalFire / NIFC	Largest US wildfire on record; F1 = 0.85 default, 0.90 under QA-relaxed
BS-02	Dragon Bravo Fire, Grand Canyon N. Rim	USA	36.41°N, −112.02°W	2025-08-09	NPS / InciWeb	Recovers to model-level ceiling under QA-relaxed
BS-03	Grampians Fires, Victoria	Australia	−37.34°S, 142.47°E	2025-02-11	VicEmergency EFFIS/GWIS	Residual gap: $F1 \leq 0.40$ under every QA configuration; different biome
BS-04	Gifford Fire, San Luis Obispo Co., CA	USA	35.13°N, −120.18°W	2025-08-18	CalFire / NIFC	F1 = 0.77; insensitive to the QA-masking choice
BS-05	Palisades Fire, Los Angeles Co., CA	USA	34.08°N, −118.59°W	2025-01-14	CalFire / NIFC	Coastal event; relaxing the QA mask raises recall from 0.55 to 0.81 (F1 0.65 → 0.80)
FL-01	Rio Grande do Sul floods, Porto Alegre	Brazil	−29.98°S, −52.31°W	2024-05-06 / 2024-05-07 (cloudy)	Fmask, OPERA DSWx, GFM, CEMS	23 chunks; matched-collection cloud pair
FL-02	Sacramento Valley floods	USA	38.89°N, −121.77°W	2023-01-22 / 2023-01-19 (cloudy)	Fmask, OPERA DSWx, GFM	45 chunks; largest per-event chunk count
FL-03	Thessaly Plain floods (Storm Daniel)	Greece	39.53°N, 22.30°E	2023-09-10 / 2023-09-08 (cloudy)	Fmask, OPERA DSWx, GFM, CEMS	18 chunks; cross-collection cloud pair

FL-04	Marj, Cyrenaica (Storm Daniel)	Libya	32.53°N, 20.92°E	2023-09-15 / 2023-09-17 (cloudy)	Fmask, OPERA DSWx, GFM	1 chunk; small-AOI dynamic-range compression
FL-05	Lower Silesia floods (Storm Boris)	Poland	51.03°N, 17.11°E	2024-09-18 / 2024-09-23 (cloudy)	Fmask, OPERA DSWx, GFM, CEMS	5 chunks; 99% mean cloud on cloudy substitute; catastrophic F1 collapse
FL-06	Valencia floods (DANA)	Spain	39.47°N, −0.38°W	2024-10-30 / 2024-10-31 (cloudy)	Fmask, OPERA DSWx, GFM, CEMS	2 chunks; small-AOI case
FL-07	Sardoba Dam collapse	Uzbekistan	40.42°N, 68.42°E	2020-05-04 / 2020-05-08 (cloudy)	Fmask, GFM	9 chunks; dam-break regime; largest NIR sensitivity in the flood set
CR-01	International Falls (US–Canada border)	USA / Canada	48.60°N, −93.41°W	2024-04-14 / 08-01 / 11-07 (+ 08-12 cloudy T2)	USDA CDL, AAFC ACI	Cross-border AOI; AAFC + CDL dual reference
CR-02	Prineville, Oregon	USA	44.30°N, −120.83°W	2024-04-23 / 08-06 / 11-07 (+ 08-09 cloudy T2)	USDA CDL	Rangeland cover; worst cell in NIR-narrow rather than SWIR-2
CR-03	Mitchell, South Dakota	USA	43.71°N, −98.02°W	2025-03-27 / 06-30 / 09-28 (+ 07-05 cloudy T2)	USDA CDL	Corn/soybean-dominant; −48 pp OA under cloudy-T2 substitution
CR-04	Urbana, Illinois	USA	40.11°N, −88.20°W	2024-03-30 / 07-13 / 10-06 (+ 07-18 cloudy T2)	USDA CDL	Most fragile site; Red × T2 removes 63.8 pp OA
CR-05	Kearney, Nebraska	USA	40.70°N, −99.08°W	2025-02-26 / 05-31 / 08-21 (+ 06-12 cloudy T2)	USDA CDL	Most robust site; worst band-corruption only −47 pp

*Note: **CAL FIRE**: California Department of Forestry and Fire Protection incident perimeters; **NIFC**: National Interagency Fire Center interagency perimeter archive; **NPS/InciWeb**: National Park Service perimeters published via the interagency InciWeb incident information system; **VicEmergency/EFFIS**: Victoria State Government emergency service and the European Forest Fire Information System; **Fmask**: the HLS Fmask quality layer (water bit) accompanying the input imagery, which is also the pipeline's cloud-screening layer (Section V-B); **OPERA DSWx**: the Dynamic Surface Water eXtent product (DSWx-HLS) of the JPL OPERA project; **GFM**: the Global Flood Monitoring component of the Copernicus Emergency Management Service, derived from Sentinel-1 SAR; **CEMS**: Copernicus Emergency Management Service rapid-mapping flood delineations; **USDA CDL**: United States Department of Agriculture, National Agricultural Statistics Service Cropland Data Layer; **AAFC ACI**: Agriculture and Agri-Food Canada Annual Crop Inventory.*

Operational HLS pipelines conventionally apply the *Fmask* quality bits as an input mask: cloud, shadow, snow and water pixels are unreliable for most land applications, and screening them protects downstream products. Whether that convention helps, however, depends on the target class. Under the production-default configuration, in which all four bits are applied, the cross-event mean F1 was 0.70, with per-event values from 0.32 (Grampians) to 0.91 (Dragon Bravo) (Fig. 2a; Table 4). Relaxing the quality mask entirely raised the cross-event mean to 0.76.

The gain concentrates where the *Fmask* water heuristic intersects the burn scar itself. *Fmask* detects water through a near-infrared darkness test, and freshly burned vegetation is also NIR-dark. At the coastal Palisades event (BS-05), 17% of the ground-truth burn area carries the water flag, so masking removes the pixels the model needs, and relaxing the mask recovers 0.15 F1 ($0.65 \rightarrow 0.80$), almost entirely through recall ($0.55 \rightarrow 0.81$). The Grampians event gains 0.08 for the same reason; the three remaining events change by 0.05 or less. Fig. A1 shows the per-event detections under both configurations, where the gain is visible at Palisades and negligible inland. Duan et al. [45] documented this behavior of the QA water and shadow flags over recently burned vegetation and excluded them from their burned-area mapping. Heritage operational products instead exclude water at preprocessing from static land/water layers rather than scene-level classification [46]. The flood pipeline illustrates the other side of the same dependence: there the *Fmask* water bit is semantically aligned with the target and serves as a validation reference (Table A2), while for burn-scar mapping the identical layer flags the target for removal. The value of a quality layer is a property of the (layer, task) pair, not of the layer alone. Four of the five acquisitions in this battery are largely cloud-free (the Grampians scene is the exception; see below), so the protective role of input masking is barely exercised here. On cloud-contaminated scenes, input masking may remain the appropriate choice, which is precisely why the masking policy must be recorded and evaluated per task rather than inherited as a fixed default. Neither configuration signals this difference to the operator: the pipeline surfaces a complete-looking burn mask either way. This is the failure mode P3 characterizes: the deployed masking choice changes the product, and nothing at the operator interface reveals the change.

Table 4. Burn-scar results (5 events). Per-event F1 of the burn class under the production-default configuration, in which the four *Fmask* quality bits (cloud, shadow, snow/ice, water) are applied to the model input, and under the QA-relaxed configuration, in which no quality bit is applied; $\Delta F1$ is the difference. Accuracy, precision, recall and IoU of the burn class are reported for both configurations. Metrics are averaged over tiles containing at least 1,000 ground-truth burn pixels. SWIR-2 dropout, the single-band corruption result of the P4 battery, reduces F1 to 0.00 at every event under both configurations.

ID	Site	Default F1	QA-relaxed F1	$\Delta F1$	Configuration	Accuracy	Precision	Recall	F1	IoU
BS-01	Smokehouse Creek, TX	0.85	0.90	+0.05	Default	0.914	0.940	0.774	0.847	0.737
					QA-relaxed	0.947	0.947	0.849	0.895	0.813

BS-02	Dragon Bravo, AZ	0.91	0.91	+0.00	Default	0.947	0.957	0.866	0.908	0.833
					QA-relaxed	0.947	0.953	0.872	0.909	0.835
BS-03	Grampians, AU	0.32	0.40	+0.08	Default	0.869	0.314	0.330	0.322	0.312
					QA-relaxed	0.881	0.605	0.378	0.403	0.357
BS-04	Gifford, CA	0.77	0.77	+0.00	Default	0.865	0.667	0.905	0.768	0.623
					QA-relaxed	0.866	0.668	0.916	0.772	0.629
BS-05	Palisades, CA	0.65	0.80	+0.15	Default	0.907	0.816	0.545	0.654	0.486
					QA-relaxed	0.934	0.786	0.812	0.799	0.665
-	Mean (all 5)	0.70	0.76	+0.06	Default	0.900	0.739	0.684	0.700	0.598
					QA-relaxed	0.915	0.792	0.765	0.756	0.660

Spectral-band dropout produced a complementary failure pattern. Zeroing the SWIR-2 band reduced F1 to 0.00 in every one of the five events tested (Fig. 2b; Table 4). The mean $\Delta F1$ for the other bands was -0.60 for NIR, -0.46 for SWIR-1, -0.21 for Green, -0.19 for Red, and -0.17 for Blue. Only SWIR-2 produces uniform collapse; it is therefore the load-bearing band for burn-scar detection in this model. The asymmetry against the other downstream tasks (below) makes this a task-specific rather than model-general finding; it is the canonical anchor for P4.

The default configuration that produced F1 of 0.77 or higher for four of the five events did not do so for the Grampians event. For this event, F1 remained at or below 0.40 under every quality-mask configuration tested (Fig. 2a, BS-03 Grampians, Australia). The event differs from the others in the biome, with eucalyptus forest cover under a Southern Hemisphere summer, and the acquisition carries substantial cloud cover. The residual gap indicates a training-distribution limit that no upstream intervention recovers. Ruling out data-side explanations, such as a different QA representation in eucalyptus scenes, requires further work; the finding as reported here is that no available preprocessing intervention closes the gap.

B. Flood Inundation Mapping

Seven flood events spanning five continents were processed through the flood-detection production pipeline. The model consumes a six-channel surface-reflectance input and produces a binary water mask. Each event was perturbed under two protocols: **spectral-band corruption**, in which one band at a time was replaced by its training-set mean prior to inference; and **cloud-paired substitution**, in which a

temporally proximate cloud-contaminated acquisition was substituted for the nominal clean acquisition. Validation used up to four reference products: the *Fmask* water bit (optical), the OPERA Dynamic Surface Water Extent product (optical), the Global Flood Monitor based on Sentinel-1 synthetic aperture radar (SAR), and the Copernicus Emergency Management Service (CEMS) rapid-mapping product where available.

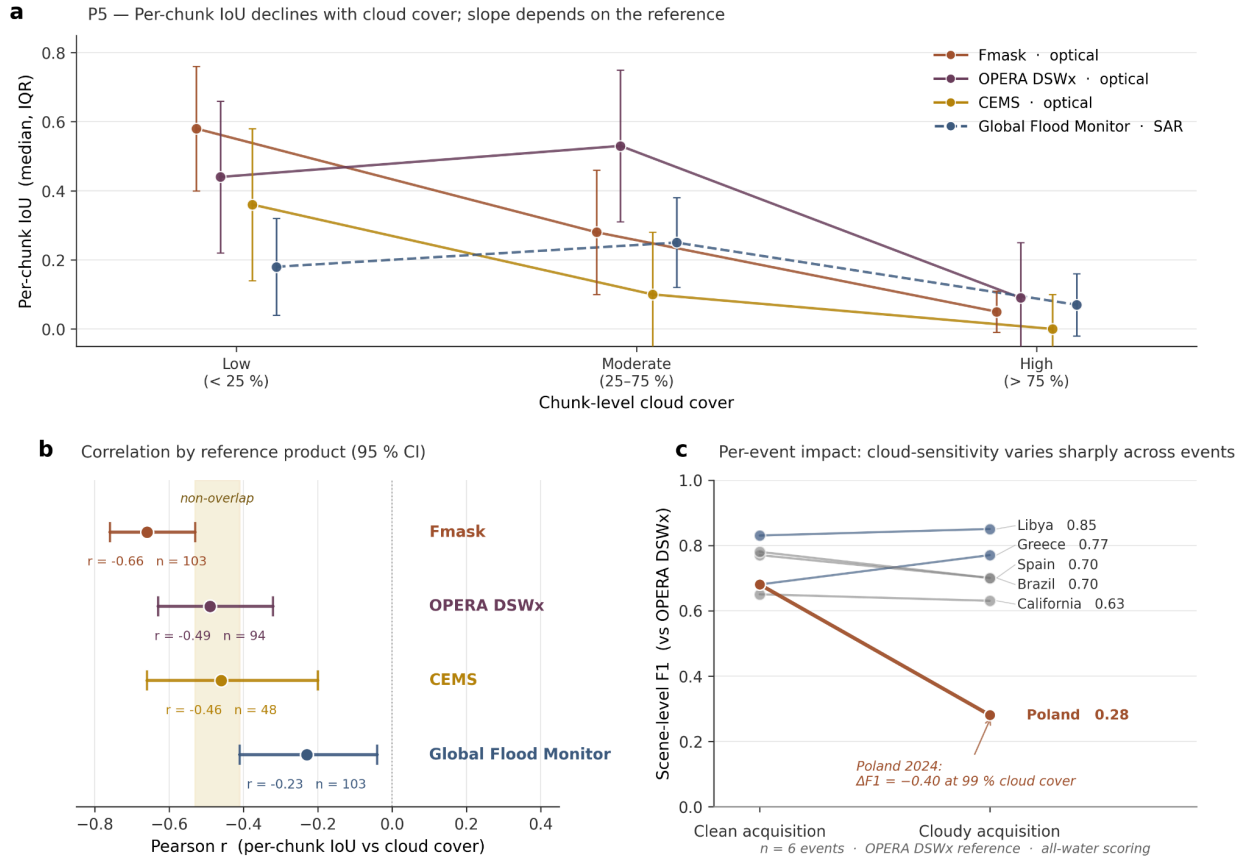


Fig. 3. P5: Reference-target alignment shapes apparent flood-pipeline cloud sensitivity. (a) Per-chunk IoU declines with chunk-level cloud cover for all four reference products, but the slope is reference-conditional. Markers show median IoU within Low (<25%), Moderate (25–75%), and High (>75%) cloud-cover bins; vertical bars show inter-quartile range. The SAR-based Global Flood Monitor (dashed) is approximately flat, reflecting its conceptual independence from the optical cloud confounder; the three optical references decline steeply. (b) Per-reference Pearson correlation of per-chunk IoU with cloud cover, with 95% confidence intervals (*Fmask* $n = 103$; OPERA DSWx $n = 94$; GFM $n = 103$; CEMS $n = 48$). The *Fmask* CI $[-0.76, -0.53]$ does not overlap the Global Flood Monitor CI $[-0.41, -0.04]$ (shaded amber band): the same prediction, evaluated against the two references, produces structurally different accounts of its cloud sensitivity. (c) Per-event clean $\rightarrow$ cloudy F1 trajectories against OPERA DSWx ($n = 6$ events, all-water scoring). Poland 2024 ($\Delta F1 = -0.40$ at 99% mean cloud cover) collapses; Libya and Greece (slate blue) improve modestly, indicating that the reference product itself contains cloud-induced gaps.

The dominant spectral dependence was on the Red band, with a per-event mean $\Delta F1$ of -0.28 (Fig. 5; Appendix Table A2). The dependence was largest at the events characterized by vegetation-water boundary discrimination, such as agricultural canals of FL-07, where the corresponding Near-Infrared dependence also concentrated. The band-vulnerability ranking was preserved across all four reference products and both scoring scenarios (Fig. 6d). The FL-06 Spain event was anomalous in that its small area of interest and near-saturated nominal F1 compressed the dynamic range of its perturbation responses; excluding it shifted the mean Red $\Delta F1$ to -0.31 and did not change the band ranking.

Substituting a real cloud-contaminated acquisition degraded F1 in six of the seven events (Fig. 3c). Among these, the most severely affected event at 99.1% mean *Fmask* cloud cover declined from a nominal F1 of 0.68 to 0.28 ($\Delta F1 = -0.40$; OPERA DSWx reference, all-water scoring; Fig. 3c). Pooling 103 per-chunk measurements across the seven events revealed a monotonic decline of intersection-over-union with chunk-level cloud cover that was strongly reference-dependent (Fig. 3a). Against the optical *Fmask* reference, the per-chunk dose-response correlation was Pearson $r = -0.66$ with a 95% confidence interval $[-0.76, -0.53]$ over $n = 103$ chunks. Against the SAR-based reference, $r = -0.23$ with a 95% confidence interval $[-0.41, -0.04]$ over $n = 103$ chunks (Fig. 3b). The confidence intervals do not overlap. The dose-response slope is structurally shallower against the radar reference because the radar product is partially independent of the optical confounder (cloud) that corrupts the prediction. This is the operational signature characterized by P5: the same prediction, evaluated against two references, produces two qualitatively different accounts of its cloud sensitivity. The choice of reference shapes the apparent claim. Restricting validation to references that share the dominant confounder with the prediction therefore misattributes reference degradation to the model and misstates the pipeline’s operational cloud sensitivity.

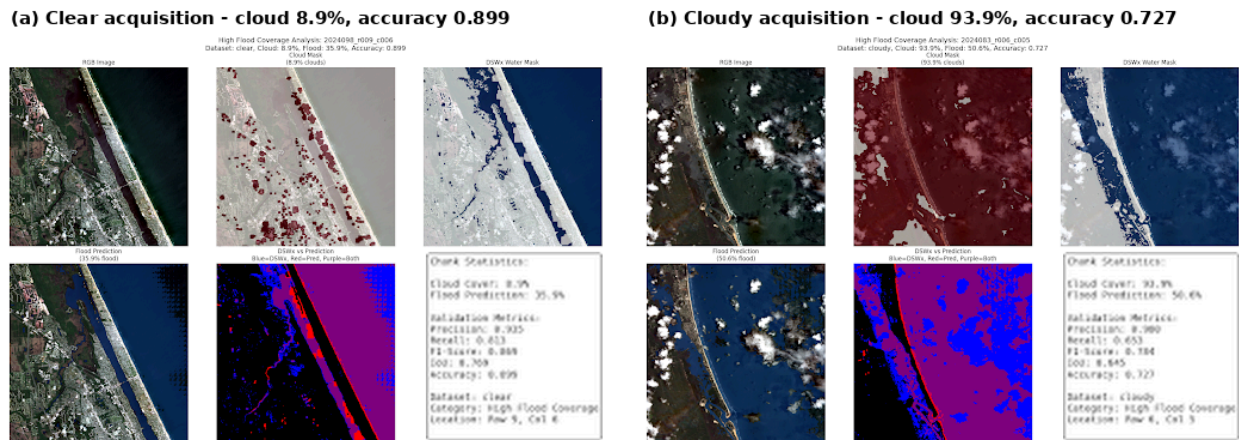


Fig. 4. Qualitative prediction-reference comparison for a matched clear/cloudy chunk pair (left-panel: clear, 8.9% cloud, accuracy 0.899; right-panel: cloudy, 93.9% cloud, accuracy 0.727; [all-water] scoring). Panels per pane: HLS RGB composite, *Fmask* cloud mask, OPERA DSWx water mask, model flood prediction, and DSWx-vs-prediction agreement (blue = reference only, red = prediction only, purple = both). Under heavy clouds the loss is recall-driven (recall 0.813 $\rightarrow$ 0.653 at near-constant precision), and disagreement concentrates where the optical reference is itself degraded.

Figure 4 shows a matched clear/cloudy chunk pair scored against the OPERA DSWx reference; the disagreement concentrated in cloud-affected regions of the agreement map is the spatial form of the dose–response quantified in Fig. 3(a)–(b).

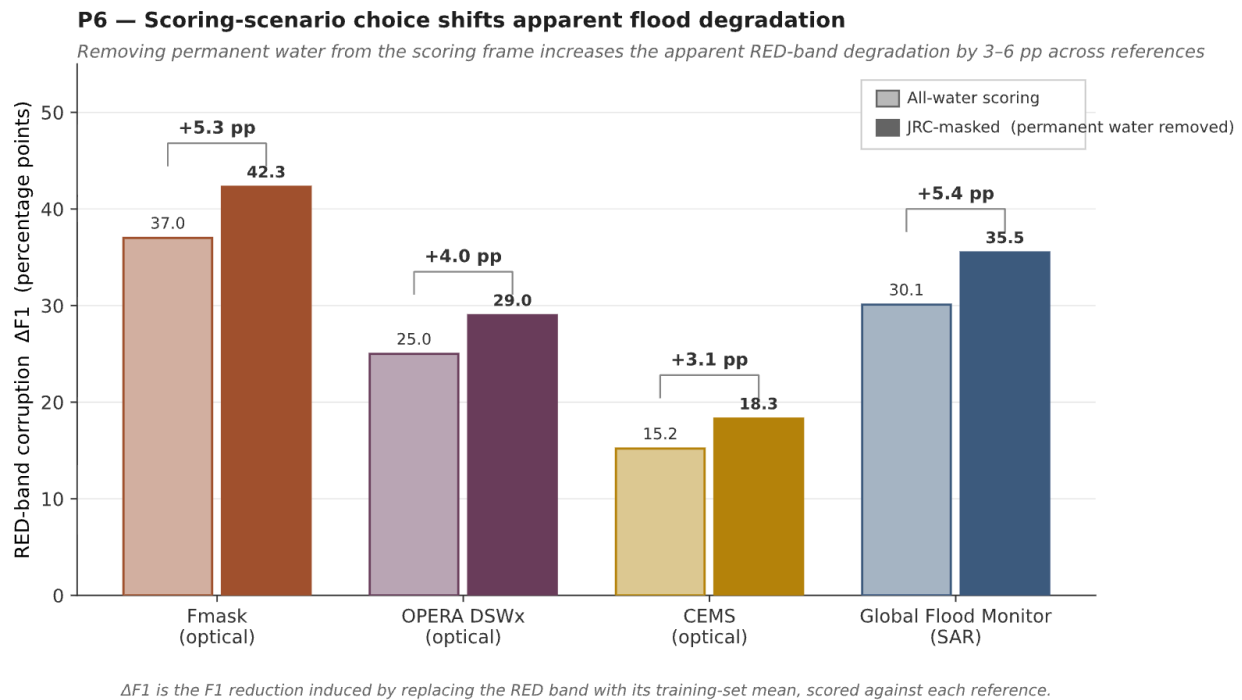


Fig. 5. P6: Scoring-scenario choice shifts apparent flood degradation. Red-band corruption $\Delta F1$ (the F1 reduction induced by replacing the Red band with its training-set mean) scored against each of four reference products, under two scoring scenarios. Light-fill bars: all-water scoring (every water pixel in the reference contributes to the score). Dark-fill bars: JRC-masked scoring (pixels labeled as permanent water by the JRC Global Surface Water product are removed from the score). The JRC-masked scenario produces a larger apparent $\Delta F1$ in every case, by 3.1 to 5.4 percentage points (brackets). The discrimination of operationally relevant transient water is therefore more fragile than the unmasked metric suggests; the choice of scoring scenario is a post-processing decision (P6) that materially shifts the apparent claim.

Masking permanent water bodies in the reference (using either the JRC Global Surface Water product or the radar-derived persistent-water mask) preserved the band-vulnerability ranking but increased the apparent $\Delta F1$ magnitudes by 3.1 to 5.4 percentage points (Fig. 5). The shift quantifies how much of the nominal performance derives from reliably detecting static water bodies rather than from identifying transient inundation. The discrimination of operational interest is thus more fragile than the unmasked metric suggests. This is an instance of the failure mode P6 addresses. Appendix Table A2 reports the per-event flood results, including baseline F1 and $\Delta F1$ under Red-band and cloud-paired perturbations, together with the pooled per-reference Pearson correlations.

C. Crop Classification

Five crop-cover phenology cycles spanning four US states and one US–Canada border location were processed through the multi-temporal crop-classification pipeline. The model consumes a six-channel, three-timestep input and produces a per-pixel class label. Each event was perturbed under four protocols: **single-band corruption**, **single-timestep corruption**, **band-by-timestep cell corruption** (an eighteen-cell grid for each event), and **cloud-paired substitution** of the mid-season acquisition. Validation used the USDA Cropland Data Layer, supplemented by the AAFC Annual Crop Inventory for the cross-border location.

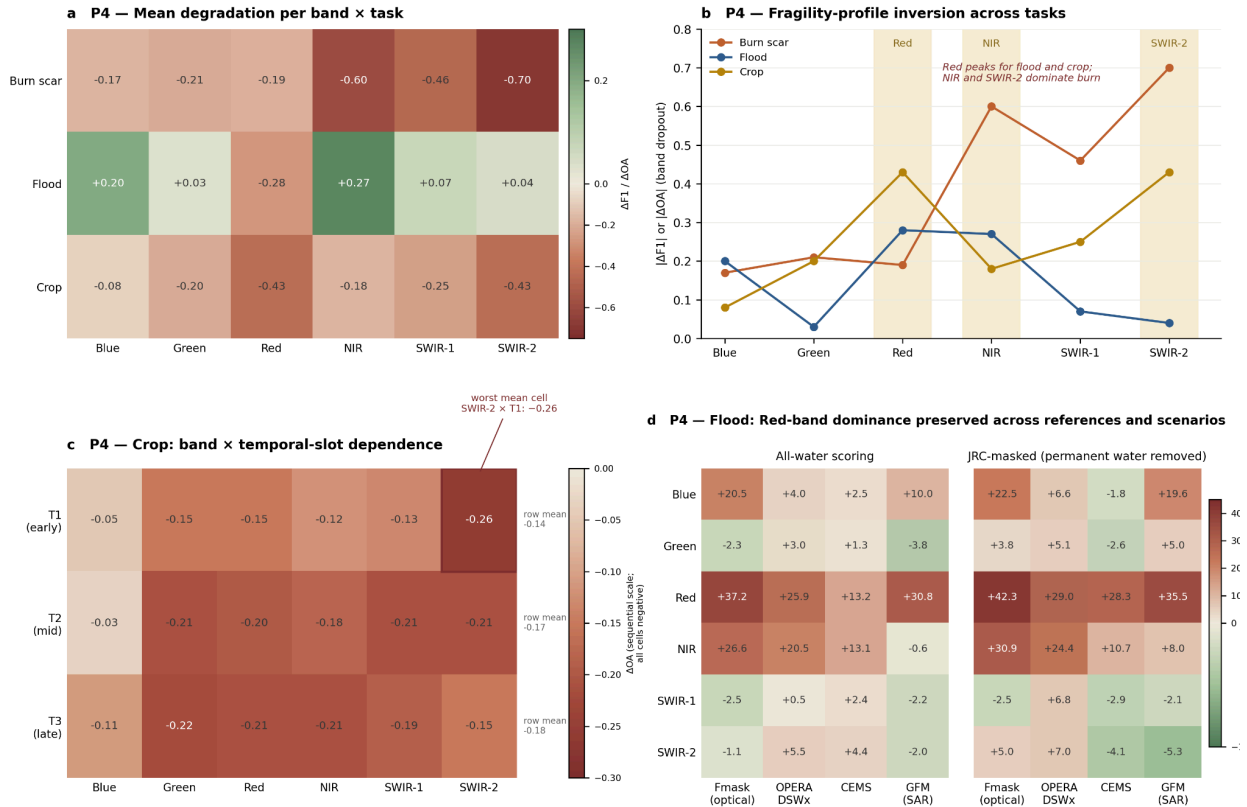


Fig. 6. P4: Modality dependence is task-conditional, reference-conditional, and scenario-conditional. (a) Cross-task heatmap of mean spectral-band degradation across burn-scar ($\Delta F1$), flood ($\Delta F1$), and crop (ΔOA) tasks under single-band corruption. Asymmetry is visible at a glance: SWIR-2 collapses burn and crop but is tolerated by flood. Red dominates flood and crop but is tolerated by burn. (b) Fragility profile per task; the crossing lines visualize the modality-dependence inversion, with the Red and SWIR-2 columns highlighted as the two inverting bands. (c) Crop band × temporal-slot dependence: the model is most sensitive at the earliest frame (T1, row mean -0.15), and the worst single (band, slot) cell is SWIR-2 × T1 ($\Delta OA = -0.26$, outlined). (d) Flood band × reference under two scoring scenarios (all-water; JRC-masked, with permanent water removed). The Red row (outlined) is dominant across all four references in both scenarios, confirming that the band-vulnerability ranking is a property of the model rather than the reference or the scoring choice. Per-row degradation magnitudes shift by 3-6 percentage points between the two scenarios; this scenario-shift information is shown directly in Fig. 5.

Table 5. Crop results (5 events). Per-event Δ OA under single-band and single-timestep corruption; worst (band, slot) cell per event; Δ OA under cloudy T2 substitution. All values are percentage points relative to the event's clean baseline OA against CDL.

ID	Site	Baseline OA	Worst band (Δ OA, pp)	T1 Δ OA (pp)	T2 Δ OA (pp)	T3 Δ OA (pp)	Worst cell (Δ OA, pp)	Cloudy T2 Δ OA (pp)
CR-01	International Falls, MN/ON	0.79	SWIR-2 (−71.8)	−68.2	−48.2	−35.2	SWIR-2 $\times$ T3 (−39.1)	−9.0
CR-02	Prineville, OR	0.79	NIR-narrow (−62.0)	−38.7	−11.8	−14.8	NIR-narrow $\times$ T3 (−32.3)	−2.9
CR-03	Mitchell, SD	0.77	Red (−49.9)	−19.8	−20.2	−13.3	SWIR-2 $\times$ T1 (−29.1)	−48.1
CR-04	Urbana, IL	0.77	Red (−67.9)	−37.6	−49.6	−0.8	Red $\times$ T2 (−63.8)	−36.2
CR-05	Kearney, NE	0.85	SWIR-2 (−47.1)	−21.5	−3.2	−32.0	SWIR-2 $\times$ T3 (−41.6)	−4.4
-	Mean	0.79	-	−37.2	−26.6	−19.2	-	−20.1

The two most damaging single-band corruptions were Red (mean Δ OA = −0.43) and SWIR-2 (mean Δ OA = −0.43) (Fig. 7a, crop row; Table 5). The per-event variance was large. At the most sensitive site, corrupting the red band alone reduced overall accuracy by 0.68, whereas at the most robust site only two of the six bands caused a substantial drop when corrupted. The crop band-vulnerability ranking (Red and SWIR-2 dominant) differs from the flood ranking (Red and Near-Infrared dominant) and the burn-scar ranking (SWIR-2 alone catastrophic), reinforcing the task-conditional character of the modality dependence quantified by Eq. (3). Figure 6b visualizes the inversion across the three tasks: SWIR-2 dominates burn and crop but is tolerated by flood; Red dominates flood and crop but is tolerated by burn.

A second observation from this task is empirically distinctive. The pipeline was most sensitive to corruption of the *earliest* temporal frame, with a cross-event mean Δ OA of −0.37 at the first frame, −0.27 at the second, and −0.19 at the third, a pattern that was held in three of five events. This contradicts the biological prior that the mid-season peak-biomass frame would carry the most discriminative information. The eighteen-cell band-by-timestep decomposition identified SWIR-2 at the first frame as the most damaging cell on average (cross-event mean Δ OA = −0.26), with SWIR-2 the leading band in the worst-cell ranking at each timestep (Fig. 6c). The implication for P4 is that, for multi-temporal pipelines, modality dependence must be disclosed at the (band, temporal slot) level, not at the band level alone.

The crop pipeline also surfaced a second preprocessing-protocol violation. When a temporal frame was sourced from the HLSS30 collection rather than the HLSS30 collection, the operational pipeline silently substituted HLSS30's broad Near-Infrared band for the narrow Near-Infrared band on which the model was trained. The substitution did not appear in any operational metadata. We report it as a second instance of the protocol-violation failure mode P3 characterizes.

D. Cross-task synthesis

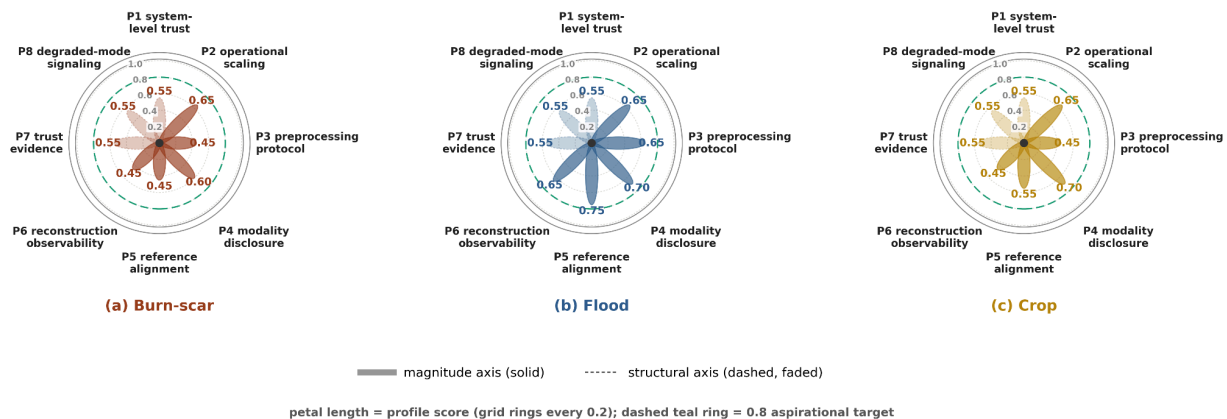


Fig. 7. Per-task Operational Readiness profiles for the three worked-example pipelines. Each panel shows one pipeline - (a) burn-scar, (b) flood, (c) crop, with petal length encoding the profile score on each of the eight axes (grid rings every 0.2; dashed teal ring at the 0.8 aspirational target). Solid petals are magnitude axes (P2-P6, P8), anchored in the measurements of Sections 3.1-3.2; faded dashed petals are structural axes (P1, P7), which record whether the pipeline emits the relevant evidence. The pipelines separate on the data- and validation-side axes - the flood pipeline's documented preprocessing protocol (P3 = 0.65) and four-reference validation (P5 = 0.75) against the burn-scar pipeline's silent quality-mask violation (P3 = 0.40), while sharing a uniform structural deficit on P1/P7/P8.

Three findings recur across the tasks. The per-task Operational Readiness profiles that summarize them are shown side by side in Fig. 7.

First, modality dependence is task-conditional and asymmetric (Fig. 6 a,b). The SWIR-2 band is load-bearing for burn-scar detection (its uniform collapse is shown in Fig. 2b) and for crop classification but is operationally tolerated by flood detection. A plausible physical explanation is that SWIR-2 reflectance is directly sensitive to vegetation-moisture loss (the same sensitivity the normalized burn ratio exploits), whereas open-water discrimination is carried chiefly by red and near-infrared contrast, leaving flood detection with little dependence on SWIR-2. The asymmetry suggests that the spectral physics of each downstream task, not the model architecture alone, determines which perturbations are operationally consequential, and that readiness must therefore be assessed per (model, task) pair. The Red band shows the mirror pattern, dominant for flood and crop but tolerated by burn. The implication for P4 is that modality dependence must be disclosed at the (model, task, temporal slot) level rather than at the model level alone; model cards that report band sensitivity for the model as a whole understate the operational risk.

Second, the deployed preprocessing protocol is the single most consequential intervention point we identified. The burn-scar quality-mask configuration accounts for a per-event $\Delta F1$ of up to 0.15, driven almost entirely by recall (up to 0.27), under full QA relaxation (Section 3.1-A; Fig. 2a; Table 4), and the crop pipeline harbors a second, undocumented, protocol violation (Section 3.1-C). Spectral-band dropout produced larger magnitudes (Fig. 2b), but band availability is a property of the acquisition; the masking policy is a choice the operator makes, can disclose, and can change. This identifies the data-model interface, specifically the documentation and emission of the preprocessing protocol under P3, as the highest-leverage controllable axis of the profile in our use-case.

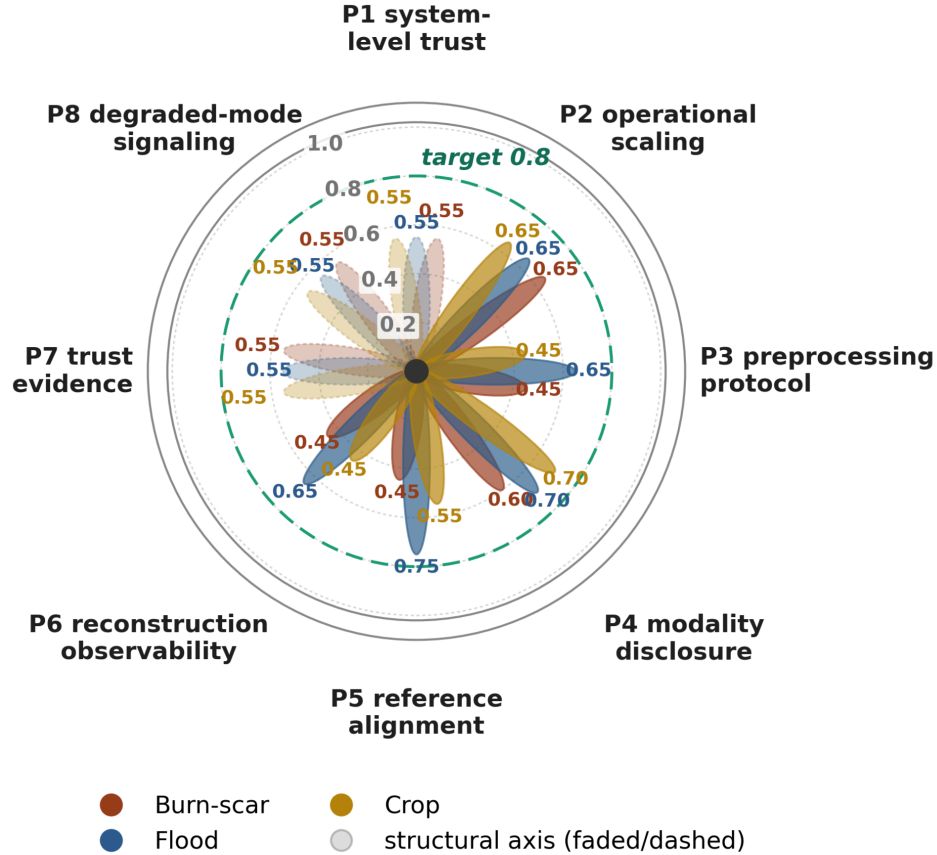


Fig. 8. The Operational Readiness profile ($\Pi(\pi)$ of Eq. (4)), measured for the three worked-example pipelines built on Prithvi-EO 2.0 (burn-scar, flood, crop; Sections 3.1-3.2). Higher scores indicate more observable operational evidence; the dashed teal ring at 0.8 marks an aspirational target that no pipeline reaches on any axis. The pipelines separate on the data- and validation-side axes: the flood pipeline scores highest on P3 (its preprocessing protocol is documented) and on P5 (its reference dependence is characterized against four independent products), while the burn-scar pipeline scores lowest on both. The three structural axes (P1, P7, P8) sit at a uniform mid-ordinal value for all pipelines: the system-level framing is documented, but none of the three pipelines emits the corresponding structural evidence today. Exact per-cell values and the falsification anchor behind each score are given in Fig.9.

Third, real cloud-contaminated acquisitions provide stronger and more operationally informative evidence than synthetic mean-replacement perturbations. The flood per-chunk dose-response (Section 3.1-B; Fig. 3 a, b) was reference-dependent in a way the perturbation experiments alone would not have surfaced, and the crop mid-frame substitution (Section 3.1-C) degraded accuracy in every event tested. Pixel-level cloud filtering on the reference side was insufficient to recover the loss in either pipeline. The operational mitigation indicated by both is a chunk-level cloud-cover gate. This is the empirical specification of the failure mode P8 addresses: the deployed pipeline should refuse inference under conditions outside the operational specification, rather than producing predictions with silent degradation at the user interface.

Figure 8 reports the resulting Operational Readiness profile for the three pipelines on a single set of axes; the per-task dahliagrams of Fig. 7 decompose the same profiles pipeline by pipeline. The three findings populate three of the eight axes with quantitative magnitudes, and the remaining axes are addressed by the deployment infrastructure described in Section 3.2.

3.2. Deployment Infrastructure for Operational GeoFM Pipelines

The same foundation model, deployed across two environments differing only in hardware capacity, produced qualitatively different operational regimes. The research-grade environment sustained acceptable performance for up to 75 concurrent users across all spatial-query scales tested. The production-grade environment exhibited a performance discontinuity between 20 and 25 concurrent users: error rates rose from approximately 5% to 36% at the smallest spatial scale, exceeded 80% at larger scales by 50 users, and tail latency (p_{95}) grew exponentially where the research-grade environment grew approximately linearly (Fig. 9a,b,d). Throughput plateaued at 8-9 requests per second on the production-grade environment while continuing to scale to 10-14 requests per second on the research-grade environment under matched load (Fig. 9b, trajectory arrows). The model is identical in both environments, but the operational viability is not. The choice of serving-layer language and framework produced a measurable throughput difference on matched hardware (Fig. 9c). Both observations reinforce the central claim: operational scaling depends on the deployment software stack and the deployment hardware in ways that conventional model-card evaluation neither predicts nor surfaces.

We note that the components of $\mathbf{O}(\theta)$ (throughput, memory, tail latency, and condition-dependent reliability) are properties of any ML-serving system and are not unique to GeoFMs. What the profile contributes is not the identification of these components but the requirement that they be reported alongside the accuracy metrics on which model cards currently focus, and the observation that in the GeoFM, setting the accuracy metric alone conceals the operational cliff that Fig. 9 exposes.

These observations are the empirical anchor of P2 and of Eq. (1). The remainder of this section identifies the architectural primitives of the deployment pipeline that determine where on the operational scaling vector $\mathbf{O}(\theta)$ a given deployment sits and how each primitive enables (or, when absent, defeats) the corresponding axis of the Operational Readiness profile.

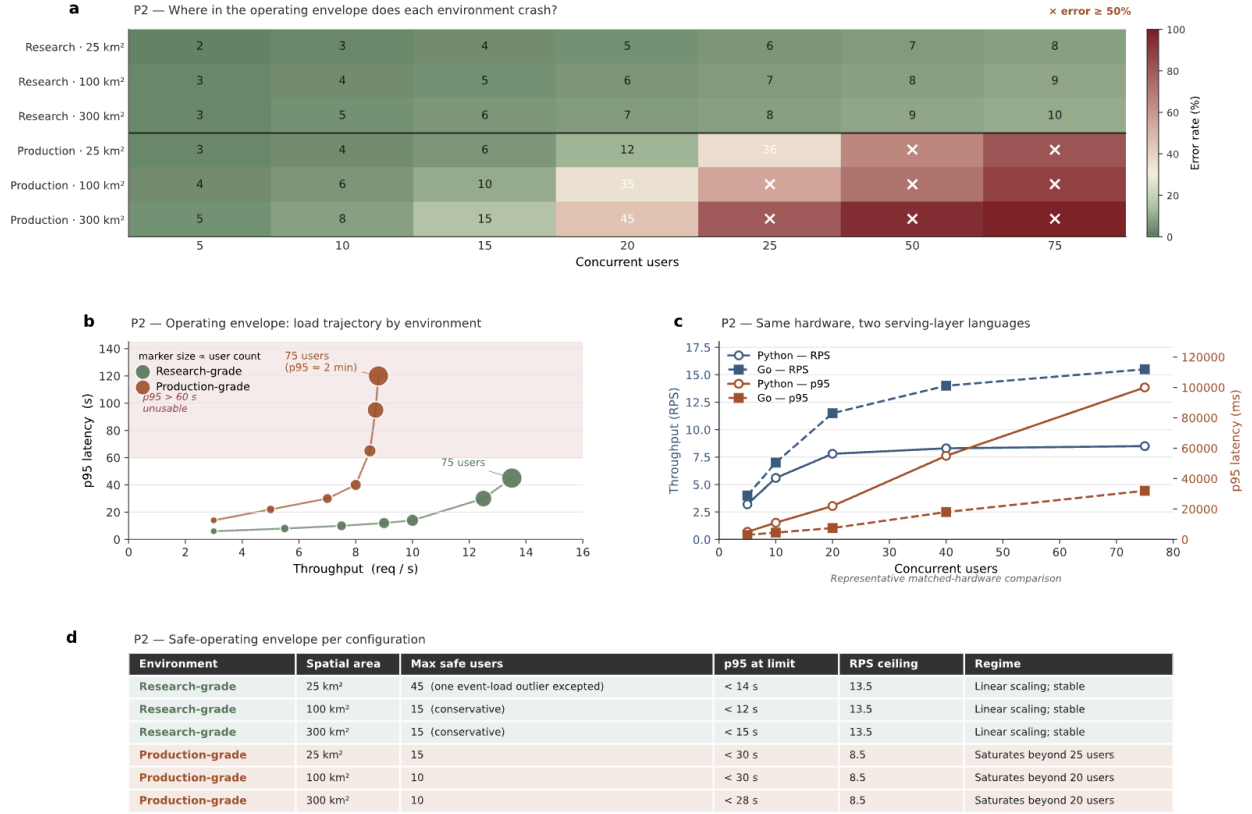


Fig. 9. Operationalizing foundation models: the measured operational scaling (a–d). (a) Error-rate heatmap over the (environment $\times$ spatial-area $\times$ concurrent-users) operating envelope. Cells marked $\times$ indicate error rate $\geq 50\%$ (effectively unusable). Production-grade configurations cross the 50% threshold at 25 concurrent users for the largest spatial scale and across all scales by 50 users; research-grade configurations remain below 10% throughout the tested range. (b) Operating envelope: per-environment trajectory through (throughput, p_{95} -latency) space as user count increases (marker size $\propto$ concurrent users). The production trajectory crosses the unusable-latency band ($p_{95} > 60$ s, shaded) before reaching its throughput plateau at approximately 8.5 RPS, while the research trajectory continues to scale toward 13.5 requests per second (RPS) at 75 users. (c) Same hardware, two serving-layer languages: Go (dashed, filled markers) achieves approximately $2\times$ the throughput of Python (solid, open markers) at peak load and substantially lower tail latency; the deployment software stack is therefore a non-trivial component of the operational scaling vector $\mathbf{O}(\theta)$ in Eq. (1). (d) Safe-operating envelope per configuration; the rightmost column summarizes the regime.

A. Asynchronous Queuing Mechanism

Geospatial inference is a long-tail workload. A single request may span hundreds of chunks, each taking anywhere from under a second to many seconds to process, and reconstructing the final output adds further latency that grows with the area of interest. Synchronous request-response semantics do not accommodate this distribution, and constrained GPU capacity compounds it: inference workers contend for a small pool of accelerators, so requests must be admitted at the rate the hardware sustains rather than the rate at which they arrive. Operational GeoFM pipelines must instead decouple request reception from

inference execution: a request enters a queue, a worker pool draws from the queue and executes inference, and the requester retrieves the result once available.

This pattern is widespread in machine-learning serving [47] but has specific consequences for operational readiness. The queue depth becomes a load-signaling primitive: when depth exceeds the worker pool's drain rate, the deployment is in a degraded regime that should be signaled to downstream consumers. The worker pool itself becomes the unit of operational scaling, and its sizing is governed by the throughput-memory function characterized in $O(\theta)$ rather than by any model-level capacity argument. In our deployment, this pattern is implemented through a Go-based API layer, Temporal-managed orchestration and worker execution, pipeline-specific FastAPI inference services, and a relational database that persists request state, metadata and inference outputs. Comparable functionality can be provided by alternative queueing frameworks, and the framework makes no commitment to any particular implementation.

B. Version Control to Ensure Trust

Operational pipelines often evolve over their deployment lifetime: preprocessing protocols are revised, reference-product versions are updated, and post-processing rules are tuned. A deployed pipeline that cannot produce, on demand, the exact configuration under which a past prediction was generated cannot satisfy P7, because the evidence must then be reconstructed from surviving records rather than read from an emitted artifact. The reconstruction process admits silent drift between the configuration that produced the prediction and the configuration the records imply.

The deployment architecture must therefore make the pipeline configuration a *declarative, version-controlled artifact* - a state representable by a Git commit hash, applied to the cluster by a continuous-reconciliation process. The widely-used pattern [48] treats the desired cluster state as code stored in a Git repository and deploys a controller that reconciles the live cluster state to the Git state continuously. Under this pattern, the preprocessing protocol defined in Eq. (2) has a Git hash; the model checkpoint has a Git hash; the post-processing rule set has a Git hash; and the manifest emitted alongside every prediction cites these hashes as the verifiable record of the pipeline configuration that produced the output (Fig. 9, manifest path). Rollback to a prior pipeline state is the application of a prior Git commit, not the recovery of a forgotten configuration.

C. User-centric Interaction and Observability

P8 specifies that degraded operating modes should be first-class outputs of a deployed pipeline. The architectural primitive that delivers on this requirement is a *user-centric interface* - a public surface (typically a status page, but conceptually any consumer-visible signal) that reports whether the pipeline is currently within its operational specification, strengthening observability.

The signaling layer is not the same as internal monitoring. Internal monitoring serves the operator; the signaling layer serves the consumer of the pipeline's outputs. The two have different audiences and different update cadences, and the difference matters: an end user inspecting a flood-extent map produced by the pipeline during a queue-overload event is entitled to know that the prediction was generated under degraded operational conditions, regardless of whether the operator's internal dashboards reflect the same

fact. Our implementation uses a static, declaratively-managed status page that consumes the same monitoring signals available to the operator; the conceptual requirement itself is independent of the specific tooling.

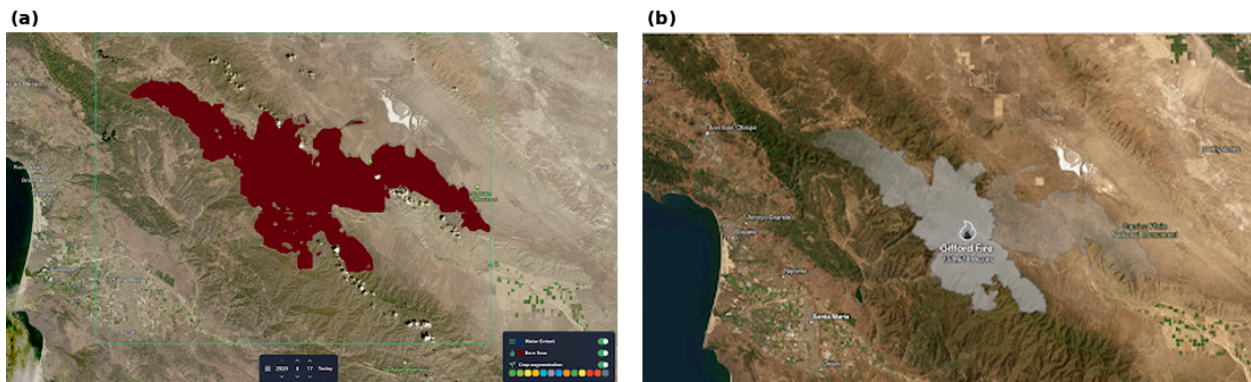


Fig. 10. User-centric interfaces for the Gifford Fire (California, August 2025; the fire underlying perturbation case event BS-04 (Table 3). (a) The operational platform's burn-scar inference for the 2025-08-17 acquisition with layer and acquisition-date controls. (b) The CalFire incident perimeter for the same fire. The platform's inference is visually consistent with the CalFire perimeter, illustrating routine-operation output alongside the reference an end user would consult.

Figure 10 shows the user-facing interface of the operational platform for a wildfire event processed in routine operation alongside the corresponding agency perimeter in the CalFire portal. The interface surfaces the prediction, layer selection, and acquisition date.

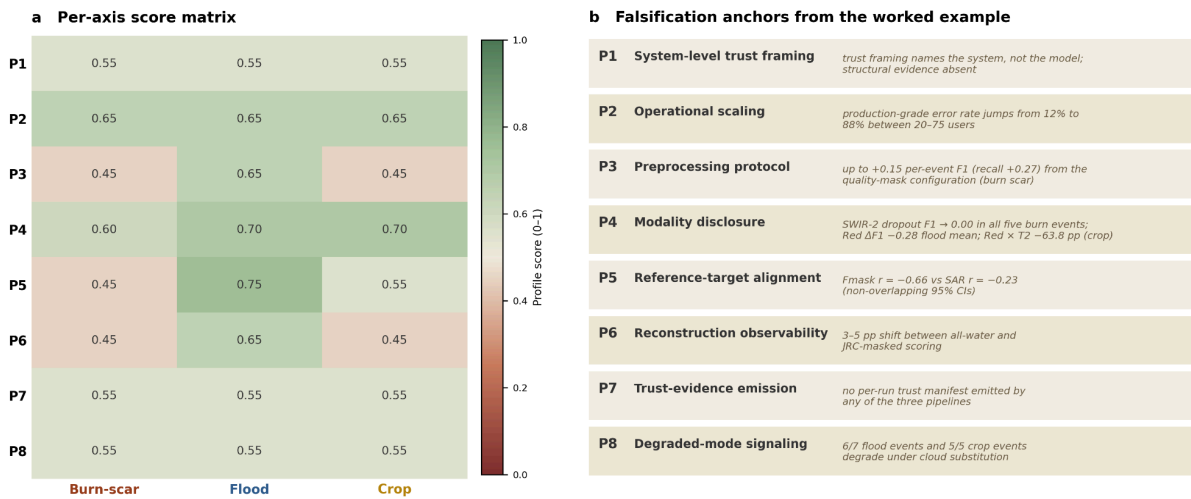


Fig. 11. The Operational Readiness profile in detail: per-axis scores and their falsification anchors. (a) Per-axis score matrix giving exact values for the three worked-example pipelines. The lowest cell is the burn-scar pipeline's P3 (0.40), reflecting the silent quality-mask protocol violation of Fig. 2; the highest is the flood pipeline's P5 (0.75), reflecting its four-reference validation with documented confounders (Fig. 3). The structural axes (P1, P7, P8) take a uniform mid-ordinal value (0.55) across all three pipelines: the

framing is documented, but the structural evidence the axes call for, a per-run trust manifest (P7) and a consumer-visible degraded-mode signal (P8), is emitted by none of the pipelines today. (b) Falsification anchors: each axis paired with the empirical observation from the worked example that justifies its score. The structural axes (P1, P7) are qualitative claims about what the pipeline emits; the magnitude axes (P2-P6, P8) are anchored in the measurements reported in Figs. 2-6 and 9 described in Sections 3.1 and 3.2. The numeric scores in Fig. 8 are ordinal summaries of the underlying structural determinations and magnitude measurements.

These four deployment primitives (queue-based serving, declarative configuration, hardware-conditioned scaling, and user-facing degraded-mode signaling) are the deployment-pillar contributions to the Operational Readiness profile. They are necessary, but not sufficient: a pipeline may instantiate all four and still fail along the data-pillar axes characterized in Section 3.1. The profile reports these failures distinctly precisely because the framework treats them as distinct. Figure 11 reports the completed Operational Readiness profile for the three worked-example pipelines, populated by the perturbation measurements of Section 3.1 on the data-, model-, and validation-side axes, and by the deployment evidence of this section on P2, P7, and P8.

4. Discussion

The use case supports one central claim: the operational behavior of a deployed GeoFM pipeline is shaped predominantly by system-level properties (preprocessing protocols, reference-target alignment, task-and-slot-specific modality dependence, observable reconstruction, infrastructure-bounded operational scaling, and degraded-mode signaling), and only secondarily by the foundation model in isolation. Across the three downstream tasks measured, the largest single magnitudes came from modality dropout, which is a property of the acquisition rather than of the deployment. Among interventions under the operator's control, none exceeded the preprocessing-protocol and reference-alignment effects, and no model architecture change would have removed the operational-scaling discontinuity of Section 3.2.

For the developers of geospatial foundation models, the implication is concrete. The assets released alongside a model checkpoint should include the preprocessing protocol under which the model was trained, expressed in a machine-actionable form sufficient for a downstream operator to detect protocol drift at inference time; per-task modality dependence at the (task, band, temporal-slot) level rather than at the model level alone; and the reference products against which downstream performance was evaluated, with explicit annotation of confounders shared between prediction and reference. Model cards in their current form require none of these disclosures and would not have surfaced the failures we observed. The eight axes of the profile name the dimensions along which such disclosure would need to extend.

For deployment operators and the governance actors who consume their outputs, the profile's value lies in answering a concise set of structural questions in a standardized format. By analyzing a pipeline's emitted artifacts, evaluators can assign a straightforward 'yes,' 'partial,' or 'no' rating to core capabilities: emitting a per-run trust manifest (P7), signaling degraded conditions (P8), separating raw inference from reconstruction (P6), and characterizing operational scaling (P2). The profile vector, in turn, translates the qualitative obligations articulated by the NIST AI Risk Management Framework, the EU AI Act, and

ISO/IEC 42001 into a small set of pipeline-level claims verifiable by inspecting the manifest that the pipeline emits in normal operation.

4.1. Limitations and Future Work

We flag four limitations explicitly. First, the measurements characterize a single foundation model across three downstream tasks. The instrument is general, but the evidence base is not, and the rank order of load-bearing axes may differ for other GeoFMs and other task families. Second, the HLSL30 broad-to-narrow near-infrared substitution reported in Section 3.1-C was discovered during the perturbation experiments rather than sought deliberately. It is reported as an instance of the P3 failure mode, not as an estimate of how prevalent such substitutions are across deployed pipelines. Third, the operational-scaling measurements in Section 3.2 reflect two specific hardware configurations and one workload distribution. The research-to-production discontinuity is sharp in our data but may scale, soften, or move under other configurations, and the throughput, memory, and latency properties that constitute $O(\theta)$ are not GeoFM-specific in origin, though they are acute for GeoFM workloads. Fourth, the community evidence base outlined below is infrastructure awaiting adoption. The profiling procedure, perturbation implementations, and reporting schema are released with this article, but the framework's utility as a comparative standard depends on independent profiles from other groups, which we have not yet demonstrated. The community evidence base is the natural mechanism for addressing all four limitations: profiles of further models and task families address the first, accumulated profiles convert the second and third from single instances into prevalence estimates, and every independently submitted profile is itself the demonstration the fourth calls for. Beyond adoption, two extensions are immediate: a second-model instance of the P4 band-corruption battery, which would test the generality claim directly, and an inter-rater exercise in which two groups independently score the same pipeline against the released rubric, which would test the comparability claim.

5. Conclusions

As geospatial foundation models move from benchmarks into operational service, the object of evaluation has to move with them, from the model checkpoint to the pipeline that deploys it. This article has argued that case and provided the instrument the argument requires. The Operational Readiness profile measures a deployed pipeline along eight axes spanning its data, model, deployment, and governance layers. Applied to three pipelines built on a single foundation model, the profile located the failures that determine operational behavior not in the model but in the system around it, each invisible to model-level evaluation. A preprocessing-protocol choice silently shifts per-event F1 by up to 0.15, driven by recall (up to 0.27); optical and radar references yield apparent cloud sensitivities with confidence intervals that do not overlap; and two deployment environments serving the same model sit on opposite sides of a discontinuous operational-scaling cliff. The five-step procedure for applying the profile to a new pipeline (Section 2.2-D), together with the released perturbation experiments, lets users compute this profile for their own pipeline, whatever the foundation model, downstream task, or serving stack. A single profile characterizes one deployment; aggregated across groups and models, such profiles would turn operational trustworthiness from a claim asserted about a model into concrete evidence read from the underlying pipeline.

Supplementary Material

This appendix specifies the per-axis measurement protocol of the Operational Readiness profile and the reference-product cross-walk that anchors the P5 axis. The protocol is designed so that an independent group with access to a deployed pipeline and its inference outputs can reproduce a profile measurement without proprietary instrumentation.

A. Axis structural components ($\mathcal{S}_i$). For each axis, the structural component is determined by inspection of artifacts the pipeline emits in normal operation.

P3 (preprocessing protocol): does the pipeline emit a hashed representation of $C = \langle \mathcal{B}, \tau, v, Q \rangle$ alongside each prediction?

P5 (reference alignment): does the pipeline record the reference product and the documented confounders of its semantic alignment?

P6 (reconstruction observability): does the pipeline record raw inference output and post-processed results separately?

P7 (trust evidence): does the pipeline emit a per-run manifest citing the Git hashes of preprocessing, model checkpoint, and post-processing rules?

P8 (degraded-mode signaling): does the pipeline expose a consumer-visible signal indicating whether it is operating within specification? Each structural component admits a yes / partial / no determination.

B. Axis magnitude components ($\mathcal{M}_i$). The magnitude components are obtained from controlled perturbation experiments. For P3, the protocol sweeps each component of C across operationally plausible alternatives and reports the maximum-magnitude Δ_{accuracy} across the sweep. For P4, the protocol replaces each band (and, for multi-temporal pipelines, each (band, slot) cell) by its training-set mean and reports the resulting $\Delta(t, b, s)$ per Eq. (3). For P5, the protocol scores the same predictions against at least two reference products with documented semantic alignment to the prediction, computes the per-chunk Pearson correlation of accuracy with the dominant confounder against each reference, and reports the maximum confidence-interval gap. For P6, the protocol scores the predictions under at least two reconstruction scenarios (for example, all-water vs permanent-water-masked) and reports the resulting Δ in the operational metric. For P2, the protocol measures throughput, peak memory, p_{99} latency, and condition-dependent reliability across an operational range of θ (concurrent users; spatial area; acquisition conditions).

C. Worked-example specifics. The burn-scar perturbation experiments comprised one baseline, ten QA-bit configurations, and four spectral-band dropouts per event (five events). The flood experiments comprised one baseline, six single-band mean-replacement perturbations, and one cloud-paired substitution per event (seven events). The crop experiments comprised one baseline, six single-band perturbations, three single-timestep perturbations, 18 (band $\times$ slot) cell perturbations, and one cloud-paired substitution per event (five events). All scoring was at the chunk level (per 224×224 inference window) and aggregated to the event level by area-weighted means; reference-product semantic alignment was documented per event and per reference.

D. Reproducibility. The Prithvi-EO 2.0 checkpoint used is the publicly released variant of [2]. All the scripts for this paper are available on GitHub <https://github.com/NASA-IMPACT/r2o-paper>. The Operational Readiness profile is available as a skill at <https://github.com/NASA-IMPACT/r2o-paper/blob/development/skills/geofm-operational-readiness/SKI-LL.md>. The operational platform underlying the downstream use cases is accessible on request at <https://explore.fm.odsio.io/>.

Appendix A

A.1 Burn-scar detection maps.

Fig. A1 shows the detected burn-scar area for each of the five case events under the two quality-mask configurations of Table 4.

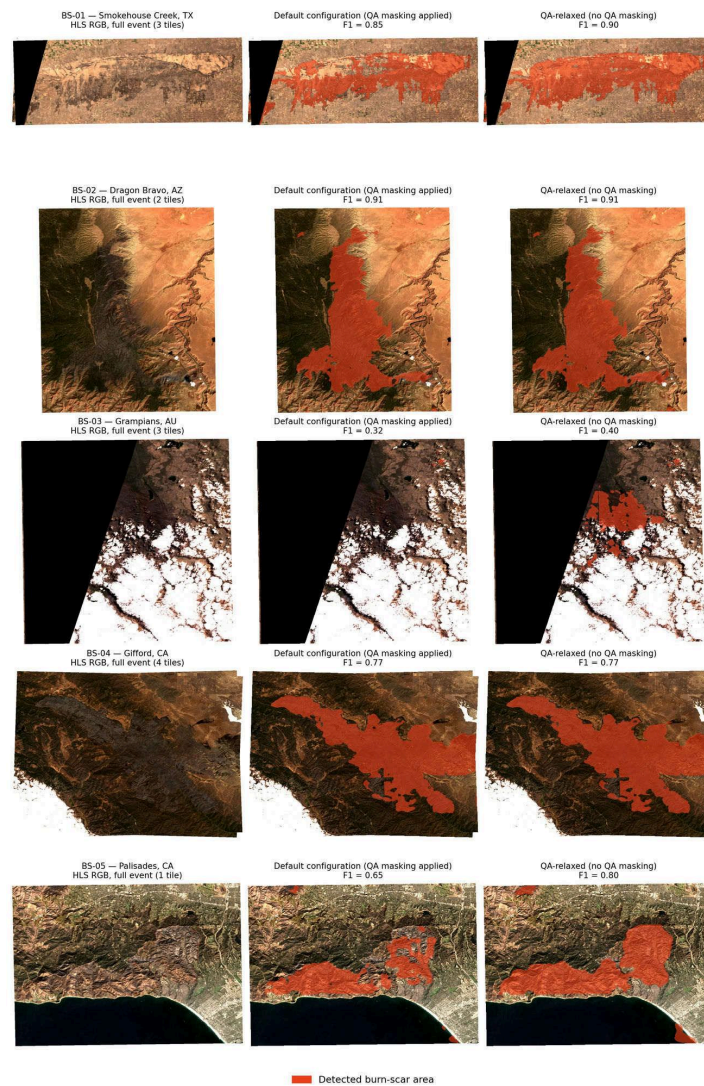


Fig. A1. Detected burn-scar area for the five case events. Each row shows one complete event, with all of its tiles mosaicked.

A.2 Per-event flood results

Table A2. Flood results (7 events × up to 4 reference products). Per-event baseline F1 under nominal input, mean ΔF1 under Red-band corruption, and ΔF1 under cloud-paired substitution, scored against the *Fmask* water-bit reference. The pooled per-reference Pearson correlations at the bottom of the table are computed on the per-chunk IoU dose-response with chunk-level cloud cover (*Fmask* $n = 103$; GFM $n = 103$; OPERA DSWx $n = 94$; CEMS $n = 48$). The *Fmask* and GFM 95% confidence intervals do not overlap, which is the empirical anchor for P5.

Event	Baseline F1 (Fmask)	Red ΔF1 (pp)	Cloudy ΔF1 (pp)
Brazil 2024 (FL-01)	0.82	−36.6	−7.3
California 2023 (FL-02)	0.82	−31.2	−3.3
Greece 2023 (FL-03)	0.65	−28.0	−4.9
Libya 2023 (FL-04)	0.48	−7.1	−9.8
Poland 2024 (FL-05)	0.82	−48.6	−44.8
Spain 2024 (FL-06)	0.08	−12.8	+6.5
Uzbekistan 2020 (FL-07)	0.87	−31.1	−14.5

Per-reference mean ΔF1 under Red-band corruption (pooled across events).

Reference	Baseline F1	Red ΔF1 (pp)	Pearson r (per-chunk, [95% CI])
<i>Fmask</i> water bit	0.65	−37.0	−0.66 [−0.76, −0.53]
OPERA DSWx-HLS	0.67	−25.0	−0.49 [−0.63, −0.32]
Global Flood Monitor (SAR)	0.60	−30.1	−0.23 [−0.41, −0.04]
CEMS rapid mapping	0.52	−15.2	−0.46 [−0.66, −0.20]

Appendix B

Table B1. Reference-product confounders (P5 anchor). Reference products used in the worked example, their sensing modality, documented confounders shared with optical GeoFM predictions, and the tasks in which each was used.

Reference product	Role	Modality	Documented confounder(s) shared with optical GeoFM prediction	Tasks used
-------------------	------	----------	---	------------

<i>Fmask</i> water bit	Reference (label)	Optical (Landsat/Sentinel-2)	Cloud, cloud shadow, aerosol, near-infrared darkness heuristic overlap with burn signatures	Burn scar (BS-01 through BS-05); Flood (all 7 events)
OPERA Dynamic Surface Water Extent (DSWx-HLS)	Reference (label)	Optical (HLS)	Cloud, cloud shadow, aerosol (shared upstream with <i>Fmask</i>)	Flood (6 of 7 events; excl. FL-07)
Global Flood Monitor (GFM)	Reference (label)	Synthetic-aperture radar (Sentinel-1)	Partially independent of optical cloud confounder; sensitive to wet-soil / vegetation-water interaction	Flood (all 7 events)
Copernicus Emergency Management Service (CEMS) rapid mapping	Reference (label)	Composite (optical + human interpretation)	Cloud (partial); interpreter bias	Flood (4 of 7 events)
JRC Global Surface Water	Mask (applied in scoring)	Optical historical composite	Permanent-water bias in scoring	Flood (scoring scenario in Fig. 4)

Acknowledgements

The authors thank [collaborators and reviewers - to be completed].

Use of AI-assisted tools. The authors disclose the use of large-language-model assistance during drafting and editing of this manuscript. AI-generated text was used for prose polishing and structural suggestions; all technical content, empirical claims, figure construction, and conclusions were authored, verified, and finalised by the human authors. No AI-generated text was incorporated without human review.

6. References

- [1] J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards *et al.*, "Foundation models for generalist geospatial artificial intelligence," arXiv:2310.18660, 2023.
- [2] D. Szwarcman, S. Roy, P. Fraccaro, P. E. Gislason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, S. Chakraborty, S. Wang, A. Kumar, M. Truong, D. Godwin, H. Lee, C.-Y. Hsu, A. Akbari Asanjan, B. Mujeci, T. Keenan, P. Arévalo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, and J. Bernabe Moreno, "Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth observation applications," arXiv:2412.02732, 2024.
- [3] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

- [4] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–22, 2023.
- [5] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "RemoteCLIP: A vision–language foundation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [6] A. Xiao, W. Xuan, J. Wang, J. Huang, D. Tao, S. Lu, and N. Yokoya, "Foundation models for remote sensing and Earth observation: A survey," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–29, 2025.
- [7] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. vanValkenburgh, S. A. Wernke, and Y. Huo, "Vision foundation models in remote sensing: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 3, pp. 190–215, 2025.
- [8] G. Vivone, L.-J. Deng, S. Deng, D. Hong, M. Jiang, C. Li, W. Li, H. Shen, X. Wu, J.-L. Xiao, J. Yao, M. Zhang, J. Chanussot, S. García, and A. Plaza, "Deep learning in remote sensing image fusion: Methods, protocols, data, and future perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 1, pp. 269–310, 2025.
- [9] A. Lacoste, N. Lehmann, P. Rodriguez, E. Sherwin, H. Kerner, B. Lütjens, J. Irvin, D. Dao, H. Alemohammad, A. Drouin *et al.*, "GEO-Bench: Toward foundation models for Earth monitoring," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 51080–51093.
- [10] V. Marsocci, Y. Jia, G. Le Bellier, D. Kerekes, L. Zeng, S. Hafner, S. Gerard, E. Brune, R. Yadav, A. Shibli *et al.*, "PANGAEA: A global and inclusive benchmark for geospatial foundation models," arXiv:2412.04204, 2024.
- [11] N. Simumba, N. Lehmann, P. Fraccaro, H. Alemohammad, G. De Mel, S. Khan, M. Maskey, N. Longepe, X. X. Zhu, H. Kerner, J. Bernabe-Moreno, and A. Lacoste, "GEO-Bench-2: From performance to capability, rethinking evaluation in geospatial AI," arXiv:2511.15658, 2025.
- [12] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu *et al.*, "On the opportunities and challenges of foundation models for GeoAI (vision paper)," *ACM Trans. Spatial Algorithms Syst.*, vol. 10, no. 2, pp. 1–46, 2024.
- [13] P. Ghamisi, W. Yu, A. Marinoni, C. M. Gevaert, C. Persello, S. Selvakumaran, M. Girotto, B. P. Horton, P. Rufin, P. Hostert *et al.*, "Responsible artificial intelligence for Earth observation: Achievable and realistic paths to serve the collective good," *IEEE Geosci. Remote Sens. Mag.*, 2025.
- [14] X. X. Zhu, Z. Xiong, Y. Wang, A. J. Stewart, K. Heidler, Y. Wang, Z. Yuan, T. Dujardin, Q. Xu, and Y. Shi, "On the foundations of Earth and climate foundation models," *Commun. Earth Environ.*, 2024 (preprint arXiv:2405.04285).

- [15] N. F. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, 2021, pp. 560–575.
- [16] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [17] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, 160018, 2016.
- [18] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 2503–2511.
- [19] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," arXiv:2001.08361, 2020.
- [20] WorldCereal Team *et al.*, "Deploying geospatial foundation models in the real world: Lessons from WorldCereal," arXiv:2508.00858, 2025.
- [21] Frontier Development Lab Team, "SHRUG-FM: Reliability-aware foundation models for Earth observation," arXiv:2511.10370, 2025.
- [22] C.-Y. Hsu, W. Li, and S. Wang, "Geospatial foundation models for image analysis: Evaluating and enhancing NASA-IBM Prithvi's domain adaptability," *Int. J. Geogr. Inf. Sci.*, pp. 1–30, 2024.
- [23] R. Del Prete *et al.*, "First on-orbit demonstration of a geospatial foundation model," arXiv:2512.01181, 2025.
- [24] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 197–211.
- [25] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27672–27683.
- [26] Clay Foundation, "Clay: An open-source foundation model of Earth," technical documentation, release v1.5, Radiant Earth, 2024.
- [27] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. Le Saux, G. Camps-Valls, and X. X. Zhu, "Neural plasticity-inspired multimodal foundation model for Earth observation," arXiv:2403.15356, 2024.

- [28] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp et al., "TerraMind: Large-scale generative multimodality for Earth observation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.
- [29] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "SatlasPretrain: A large-scale dataset for remote sensing image understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 16772–16782.
- [30] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "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, Sep. 2023.
- [31] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [32] 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, Dec. 2022.
- [33] C. Fibaek et al., "PhilEO Bench: Evaluating geo-spatial foundation models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2024.
- [34] X. Li, Y. Tao, S. Zhang, S. Liu, Z. Xiong, C. Luo, L. Liu, M. Pechenizkiy, X. X. Zhu, and T. Huang, "REOBench: Benchmarking robustness of Earth observation foundation models," arXiv:2505.16793, 2025.
- [35] S. Rasp et al., "WeatherBench 2: A benchmark for the next generation of data-driven global weather models," *J. Adv. Model. Earth Syst.*, vol. 16, e2023MS004019, 2024.
- [36] R. Shinde, K. Ankur, C. E. Phillips, A. Gupta, S. Pfreundschuh, S. Roy et al., "WxC-Bench: A novel dataset for weather and climate downstream tasks," *Sci. Data*, vol. 13, art. no. 596, 2026, doi: 10.1038/s41597-026-06839-7.
- [37] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: A survey of case studies," *ACM Comput. Surv.*, vol. 55, no. 6, pp. 1–29, 2022.
- [38] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ML test score: A rubric for ML production readiness and technical debt reduction," in *Proc. IEEE Int. Conf. Big Data*, 2017, pp. 1123–1132.
- [39] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey," *ACM SIGMOD Rec.*, vol. 47, no. 2, pp. 17–28, 2018.
- [40] V. Kolluru, R. Shinde, A. Marouane, C. Helbling, D. Shah, O. Drew, I. Gurung, M. Maskey, and R. Ramachandran, "Land cover and flood type govern the detection limits of satellite-based flood mapping across diverse global flood events," arXiv:2606.07780, 2026.

- [41] R. C. Holmström, "From AI to digital transformation: The AI readiness framework," *Bus. Horiz.*, vol. 65, no. 3, pp. 329–339, 2022.
- [42] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
- [43] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 (the AI Act)," *Off. J. Eur. Union*, Jun. 2024.
- [44] C. Phillips, S. Roy, K. Ankur, and R. Ramachandran, "HLS Foundation Burnscars dataset," Hugging Face, 2023, doi: 10.57967/hf/0956.
- [45] Q. Duan, R. Liu, J. Chen et al., "Burned area detection from a single satellite image using an adaptive thresholds algorithm," *Int. J. Digit. Earth*, vol. 17, no. 1, art. no. 2376275, 2024, doi: 10.1080/17538947.2024.2376275.
- [46] L. Giglio, L. Boschetti, D. P. Roy, M. L. Humber, and C. O. Justice, "The Collection 6 MODIS burned area mapping algorithm and product," *Remote Sens. Environ.*, vol. 217, pp. 72–85, 2018.
- [47] D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica, "Clipper: A low-latency online prediction serving system," in *Proc. USENIX NSDI*, 2017, pp. 613–627.
- [48] F. Beetz and S. Harrer, "GitOps: The evolution of DevOps?," *IEEE Softw.*, vol. 39, no. 4, pp. 70–75, 2022.