WILDFIREIA: CAN ENVIRONMENTAL DATA AT FIRE DISCOVERY TIME PREDICT WHETHER A WILDFIRE WILL ESCAPE INITIAL ATTACK?

Anonymous authors
Paper under double-blind review

ABSTRACT

Initial attack (IA) refers to the first suppression actions taken to contain or extinguish a wildfire following its detection. Existing IA failure prediction studies often use non-public response records or regional settings, so it remains unclear how well public data aligned to the reported discovery day can support IA failure prediction at national scale. We present WILDFIREIA, a novel U.S. national-scale benchmark for IA failure prediction under a discovery-day information contract using public environmental and contextual data. WILDFIREIA aligns 38,128 naturally caused FPA-FOD wildfire events with FIRMS/VIIRS thermal detections, gridMET weather and fire-danger variables, LANDFIRE vegetation, fuel, and topography, OpenStreetMap access features, and WorldPop population density. To prevent data leakage, the benchmark fixes the event unit, size-based label rule, chronological split, metrics, and forbidden-feature list, and excludes final fire size, containment timestamps, and D+1 and later satellite detections from model inputs. We evaluate 16 representative models across tabular, temporal, spatial, and spatiotemporal families under the same protocol. Results show that public discovery-day data provides useful but incomplete signal for IA failure prediction: XGBoost achieves the best AUPRC of 53.3%; FIRMS/VIIRS is the least redundant source; and fuel is the strongest static predictor when dynamic observations are unavailable. We release preprocessing outputs and model-ready caches to support reproducible research on early wildfire risk assessment: `https://github.com/WildfireIA-anonymous/WildfireIA`.

1 INTRODUCTION

The initial attack (IA) phase is the first suppression phase after a wildfire is discovered, when first-arriving firefighting resources attempt to halt or slow fire spread (Lee et al., 2013; Korkola et al., 2024). At this stage, fire management agencies face an event-level triage problem: they must quickly predict whether a reported wildfire can be contained by the initial response or is likely to escape early control (Korkola et al., 2024; Cardil et al., 2025; Xu et al., 2023; Katuwal et al., 2018). If this risk is underestimated, a wildfire that still appears manageable may be treated just as a routine event, causing agencies to miss the narrow window for early escalation and forcing a shift from rapid containment to a larger, longer, and more costly suppression effort with greater operational complexity and life-safety risk (Katuwal et al., 2018; Rashidi et al., 2018; Cardil et al., 2025; Arienti et al., 2006). These stakes are becoming increasingly severe in U.S. recently as fire-weather seasons lengthen, large-fire activity increases, fuels become drier, and ignition patterns expand the conditions under which damaging wildfires can occur (Westerling et al., 2006; Jolly et al., 2015; Dennison et al., 2014; Abatzoglou & Williams, 2016; Balch et al., 2017).

Previous IA failure prediction studies show that early suppression outcomes are associated with fire weather, fuels, terrain, access, location, and response-related factors (Arienti et al., 2006; Lee et al., 2013; Korkola et al., 2024; Cardil et al., 2025). However, these studies leave two immediate barriers to building reliable early-warning models. First, many rely on operational variables such as dispatch time, crew size, resource availability, aircraft use, and deployment logs (Lee et al., 2013; Katuwal et al., 2018; Rashidi et al., 2018; Korkola et al., 2024), which are important for explaining suppression outcomes but are often unavailable in public U.S. national databases or unknown at fire discovery

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

time. This makes their results difficult to reproduce and can mix the intrinsic early risk of a fire with information about the later suppression response. Second, existing evaluations are often regional or agency-specific, so their labels, response standards, source availability, and evaluation protocols are not directly comparable across studies (Korkola et al., 2024; Butler et al., 2025; Xu et al., 2023). As a result, agencies and researchers still lack clear evidence on how well public discovery-day information can predict IA failure, which sources add non-redundant signal, and which models are reliable under the same evaluation setting, limiting evidence-based early warning and resource prioritization. This uncertainty has practical consequences as wildfire activity continues to strain communities and response systems, with U.S. fires in 2026 already exceeding the 10-year average for the same year-to-date (YTD) period in both fire counts and burned area, and recent evidence showing that wildfire smoke is imposing growing public-health burdens (National Interagency Fire Center, 2026; Deng et al., 2026). Therefore, a U.S. national public benchmark is urgently needed to turn IA failure prediction into a reproducible, leakage-controlled task where data sources and model families are compared under the same rules.

To tackle the research gaps, we present WILDFIREIA, a novel U.S. national-scale benchmark for wildfire IA failure prediction from public environmental data under a fixed discovery-day information contract. To address the dependence on non-public operational records, WILDFIREIA builds on 38,128 naturally caused FPA-FOD wildfire events as a national event backbone and aligns them with public discovery-day or static signals from FIRMS/VIIRS, gridMET, LANDFIRE, OpenStreetMap, and WorldPop (Short, 2022; NASA, 2026; Abatzoglou, 2013; LANDFIRE, 2026; Haklay & Weber, 2008; Tatem, 2017). To address the lack of comparable evaluation protocols, WILDFIREIA fixes the event unit, size-based label rule, chronological split, forbidden-feature list, metrics, and model-ready representations, while excluding outcome-derived information such as final fire size, containment timestamps, and D+1 and later satellite detections from all model inputs (Short, 2022; Xu et al., 2023; Korkola et al., 2024; Bhardwaj, 2025; U.S. Geological Survey & USDA Forest Service, 2026). We focus on naturally caused wildfires because their early spread is more directly governed by ignition environment, fire weather, fuel, and terrain, whereas human-caused fire outcomes are more strongly shaped by human activity distributions and reporting patterns (Balch et al., 2017; Finney, 1998; 2006; Jolly et al., 2015). Using this benchmark, we evaluate 16 representative models that cover both standard classical baselines and recent state-of-the-art architectures across tabular, temporal, spatial, and spatiotemporal prediction families (Chen & Guestrin, 2016; Cho et al., 2014; Shi et al., 2015; Tang et al., 2023; Fare Garnot & Landrieu, 2021). The evaluation shows three key insights: (i) public discovery-day data contains meaningful but incomplete signal for IA failure prediction, with structured event-level models providing strong baselines; (ii) discovery-day FIRMS/VIIRS evidence is the least redundant input source, while fuel provides the strongest static fallback signal when dynamic observations are unavailable; and (iii) containment duration is only weakly explained by discovery-day public inputs. Overall, our contributions are summarized as follows:

- **The first U.S. national benchmark for wildfire IA failure prediction.** To the best of our knowledge, WILDFIREIA is the first to standardize this event-level task at fire discovery day using public environmental data, with fixed labels, splits, metrics, and leakage controls.
- **A leakage-controlled multi-source data infrastructure.** We align 38,128 natural FPA-FOD wildfire events with FIRMS/VIIRS, gridMET, LANDFIRE, OpenStreetMap, and WorldPop data, yielding tabular, temporal, spatial, and spatiotemporal model-ready inputs.
- **A broad baseline suite across prediction paradigms.** We evaluate 16 representative models covering classical baselines and recent architectures for tabular, temporal, spatial, and spatiotemporal learning, establishing a reproducible performance reference for future IA prediction research.
- **Benchmark-driven insights into early wildfire risk.** Our experiments show that public discovery-day data provides meaningful but incomplete signal for IA failure prediction, identify FIRMS/VIIRS as the least redundant discovery-day source and fuel as the strongest static fallback source, and show that containment duration is weakly explained by discovery-day inputs.

2 BENCHMARK DESIGN AND DATA CONSTRUCTION

This section first defines the event-level IA failure prediction task and research questions, then describes the public data sources used in WILDFIREIA, and finally presents the canonicalization pipeline that converts these sources into leakage-controlled, model-ready artifacts.

Table 1: Dataset-source inventory for WILDFIREIA. Each row corresponds to a downloaded public data product or raster layer. Area is reported in km^2 for the nominal source coverage.

Name	Source	Country	Area (km^2)	Task	Period	Spatial resolution	Update frequency
FPA-FOD	fire reports	USA	9,834,000	labels and metadata	1992–2020	event point	periodic release
FIRMS/VIIRS	fire signal	Worldwide	149,000,000	discovery-day thermal signal	2012–2026	375 m	near-real-time
gridMET	weather	USA	9,834,000	weather and fire danger	1979–2020	$\sim$4 km	daily
Landfire	fuel	USA	9,834,000	Fuel Disturbance, Vegetation, Canopy	2001–2024	30 m	versioned release
Landfire	vegetation	USA	9,834,000	Existing Vegetation	2001–2024	30 m	versioned release
Landfire	topography	USA	9,834,000	Elevation, Aspect, Slope Degrees	2001–2024	30 m	versioned release
OpenStreetMap	access	Worldwide	149,000,000	roads density and fire stations	2004–2026	vector	continuously updated
WorldPop	population	Worldwide	149,000,000	population density	2000–2026	$\sim$100 m	annual

2.1 EVENT-LEVEL TASK DESIGN AND RESEARCH QUESTIONS

This subsection defines the event unit, input rule, label construction, and research questions for WILDFIREIA. Let $i \in \{1, \ldots, n\}$ index wildfire events, and let t_i^d denote the reported discovery time of event i. Let D_i denote the calendar day containing t_i^d. Let a_i denote its final burned area, which is used only for label construction and never used as a model input. Let $\mathcal{C}$ denote the set of public source groups and $\mathcal{S} \subseteq \mathcal{C}$ denote any source subset. For event i, let $\boldsymbol{x}_i(\mathcal{S})$ denote the model-ready input constructed from sources in $\mathcal{S}$ using only public information aligned to D_i, while excluding satellite detection from $D_i + 1$ and later. The input $\boldsymbol{x}_i(\mathcal{S})$ may be tabular, temporal, spatial, or spatiotemporal depending on the model family. Following prior wildfire-management studies (Cardil et al., 2025; Lee et al., 2013; Korkola et al., 2024; Xu et al., 2023; Butler et al., 2025; Arienti et al., 2006), we define the initial attack below.

Definition 2.1 (Initial Attack)**.** For a wildfire event i, the *initial attack* phase is the first suppression phase beginning at t_i^d, during which first-arriving firefighting resources attempt to halt or contain fire spread. A wildfire *escapes initial attack* if this first response fails and the incident requires suppression beyond the initial response capacity (National Wildfire Coordinating Group, 2026).

Direct operational IA outcomes are not consistently available in public national fire records, and no universal final-area threshold is prescribed for IA escape (National Wildfire Coordinating Group, 2026; Short, 2022). Prior U.S. and international studies have therefore used final fire size to operationalize IA outcomes, with thresholds varying across studies and regions (Cardil et al., 2025; Lee et al., 2013; Rodrigues et al., 2019; Xu et al., 2023; Arienti et al., 2006). Following this size-based practice, we define fires no larger than 10 ha as successful IA ($y_i = 0$), fires at least 50 ha as IA failure ($y_i = 1$), and exclude intermediate-size fires to reduce label ambiguity. Sensitivity to alternative size thresholds is reported in Appendix C. IA Failure prediction is then defined as:

Problem 2.2 (IA Failure Prediction)**.** Given the discovery-day input $\boldsymbol{x}_i(\mathcal{S})$ and binary IA label y_i, the task is to learn a predictor $p_i(\mathcal{S}) = f_{\boldsymbol{\theta}}\big(\boldsymbol{x}_i(\mathcal{S})\big) \approx \Pr(y_i = 1 \mid \boldsymbol{x}_i(\mathcal{S}))$, where $p_i(\mathcal{S}) \in [0, 1]$ is the predicted probability that wildfire event i escapes initial attack. The full benchmark dataset is $\mathcal{D} = \{(\boldsymbol{x}_i(\mathcal{C}), y_i)\}_{i=1}^n$. In WILDFIREIA, $\mathcal{C}$ includes FPA-FOD metadata, FIRMS/VIIRS fire-signal features, gridMET weather and fire-danger features, LANDFIRE vegetation, fuel, and topography features, OpenStreetMap access features, and WorldPop population-context features. Different choices of $\mathcal{S}$ instantiate the full-source, source-ablation, static-source, and weather-history tasks evaluated in later sections. Based on this task design, WILDFIREIA studies five research questions.

RQ1: Full-source predictability. How accurately can public discovery-day sources predict IA failure, and which input representation is most effective?

RQ2: Source necessity. Under the full-input setting, which source groups contribute non-redundant predictive signal, and which become largely redundant once other sources are available?

RQ3: Static-source value. When weather history and discovery-day satellite detections are unavailable, which static sources improve prediction beyond FPA-FOD metadata?

RQ4: Weather-history sufficiency. How much near-discovery weather and fire-danger history is needed for IA failure prediction, and does a multi-day history improve over discovery-day conditions?

RQ5: Containment-duration signal. Can the same discovery-day inputs predict containment duration, or is this post-discovery outcome driven by factors outside the benchmark input contract?

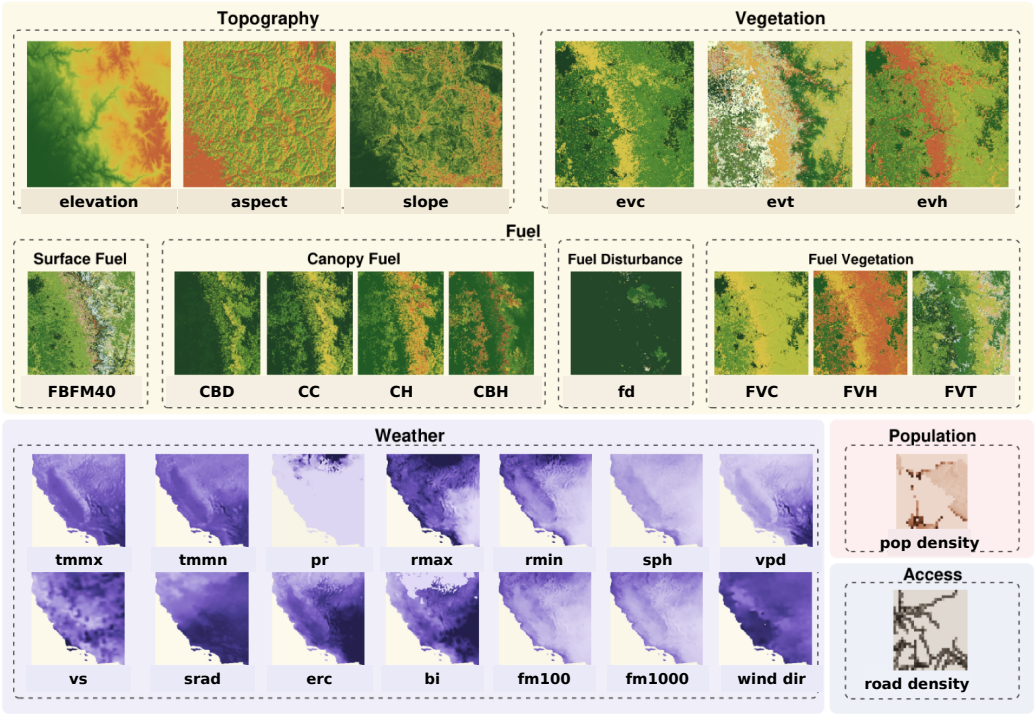


Figure 1: Examples of event-centered source layers used by WILDFIREIA. The visualization shows aligned topography, vegetation, fuel, weather, population, and access channels around a reported wildfire location. These layers are discovery-day or static inputs, not prediction targets.

2.2 DATA SOURCES AND COLLECTION

Table 1 summarizes the public data products used in WILDFIREIA, covering the event backbone, discovery-day fire signal, weather and fire-danger history, static landscape context, access proxies, and population context. Figure 1 illustrates how these event-centered layers are aligned around each reported wildfire before being converted into benchmark features.

Event backbone. WILDFIREIA uses FPA-FOD as the event backbone (Short, 2022), retaining 38,128 naturally caused wildfires from 2016–2020 in the contiguous United States with valid identifiers, discovery dates, coordinates, and positive final sizes. This yields a rare-event task, with IA failure rates of 7.55%, 6.07%, and 8.32% in train, validation, and test.

Discovery-day fire signal. FIRMS/VIIRS provides 375 m active-fire observations used only as discovery-day thermal evidence, not as the event definition (NASA, 2026; Schroeder et al., 2014). We match 1,738 detections to nearby FPA-FOD events and summarize count, fire radiative power, brightness, and patch-level channels; unmatched events receive zero-valued thermal features. The effect of conditioning on FIRMS/VIIRS observability is analyzed in Appendix Figure 5 and Table 7.

Weather. gridMET provides daily weather and fire-danger variables at approximately 4 km resolution (Abatzoglou, 2013). For each event, we extract a $D_i{-}4{:}D_i$ sequence ending on discovery day, including temperature, humidity, wind, energy release component, burning index, and fuel moisture.

Topography. LANDFIRE topography provides 30 m CONUS-wide terrain layers (LANDFIRE, 2026). We use elevation, slope, and aspect as fixed terrain attributes around each reported fire location.

Vegetation. LANDFIRE vegetation provides 30 m CONUS-wide landscape structure and plant-community information (LANDFIRE, 2026). We use Existing Vegetation Type (EVT), Existing Vegetation Cover (EVC), and Existing Vegetation Height (EVH) to characterize local vegetation.

Fuel. LANDFIRE fuel products provide 30 m combustible-material layers over the contiguous United States (LANDFIRE, 2026). We use surface fuel (FBFM40), canopy fuel variables, fuel disturbance, and fuel vegetation layers to represent persistent fuel conditions relevant to fire growth potential.

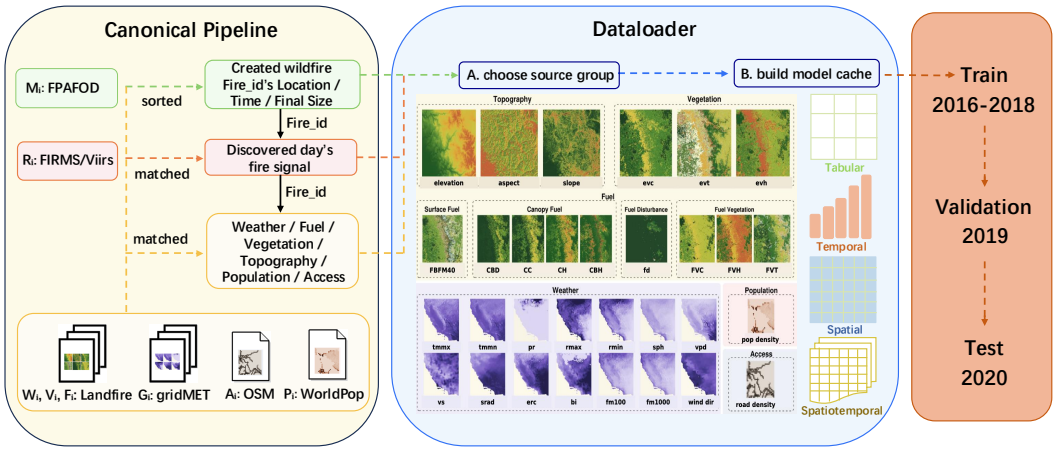


Figure 2: The full construction pipeline aligns raw data into canonicalized data, loads it into a training-ready cache, and defines the train/validation/test setting.

Access. OpenStreetMap provides event and patch level access proxies: drivable-road density and fire-station proximity (Haklay & Weber, 2008). We use a 2020 extract containing 26.0M drivable-road geometries (12.8M km of projected linework) and 27,989 fire-station geometries.

Population. WorldPop provides prior-year annual population density for each fire year, using the preceding year's raster to avoid future information (Tatem, 2017). The annual rasters use 0.000833° cells, approximately 100 m resolution. The 2019 raster contains 430,711 $\times$ 62,976 cells.

2.3 Canonicalization Pipeline

Figure 2 summarizes the construction pipeline used to enforce a common event unit across all sources. Each retained FPA-FOD record defines one benchmark sample with a unique fire_id; its discovery date, location, and fire year determine how external public sources are attached. Raw sources are aligned to these event records and canonicalized into event-level and patch-level artifacts keyed by fire_id. The dataloader then applies the requested source setting and discovery-day rules before converting the canonical artifacts into tabular, temporal, spatial, and spatiotemporal model caches.

Label construction. Each retained FPA-FOD record defines one benchmark sample with a unique fire_id. The event's discovery date, location, and fire year serve as the join key for attaching all external sources. Final fire size and containment timestamps are used solely for label construction and are excluded from model inputs via a forbidden-feature manifest.

Fire signal matching. Each FIRMS/VIIRS detection is matched to at most one FPA-FOD event within 3 km matching radius on its reported discovery day. Matching-radius sensitivity is reported in Appendix C. Unmatched detections are discarded; unmatched FPA-FOD events retain zero-valued thermal features and a detection indicator. Detections from $D_i + 1$ onward are excluded. Appendix Tables 7 and 8 assess missing VIIRS observations and stricter pre-discovery-time filtering.

Environmental and contextual features. We extract gridMET weather and fire-danger features through discovery day, and use LANDFIRE fuel, vegetation, and topography as static landscape context. OpenStreetMap provides access and response-proximity features, while prior-year WorldPop rasters provide population context. All features are stored as fire_id-indexed canonical tables.

Spatial artifacts and outputs. For spatial models, we build an event-centered 29 $\times$ 29 grid at 375 m resolution in EPSG:5070 around each discovery location. Raster, vector, and point sources are sampled on this grid to produce spatial and spatiotemporal caches. The canonical outputs include event tables, source-specific features, daily weather records, spatial artifacts, and a master table with fixed splits, input windows, and forbidden-feature rules. Robustness of the source-alignment procedure to discovery-location and date perturbations and to deliberate cross-source misalignment is evaluated in Appendix Tables 11 and 12.

3 TASKS, METRICS & EVALUATION PROTOCOL

In this section we introduce how WILDFIREIA evaluates IA failure and auxiliary containment-duration prediction. The tasks share the same event unit, discovery-day input constraint, and leakage-control policy, while varying the source subset, weather window, representation family, or prediction target. We then specify the corresponding classification and regression metrics.

3.1 EVALUATION TASKS

All tasks use the label y_i, source set $\mathcal{C}$, and input notation $\boldsymbol{x}_i(\mathcal{S})$ from Problem 2.2, and all inputs follow the discovery-day leakage-control rule in Section 2.1.

Task 1: Full-source predictability. This task evaluates whether the complete public discovery-day source stack can predict initial attack failure. Each model receives $\boldsymbol{x}_i(\mathcal{C})$ and predicts $p_i^{\mathrm{IA}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i(\mathcal{C}))$, where $p_i^{\mathrm{IA}} \in [0,1]$ is the predicted probability that event i escapes initial attack.

Task 2: Source necessity under full input. This task measures whether each source group contributes non-redundant information when all other sources are available. For each source group $\mathcal{Q} \in \mathcal{C} \setminus \{\mathcal{M}\}$, the model receives $\boldsymbol{x}_i(\mathcal{C} \setminus \{\mathcal{Q}\})$ and predicts y_i. A large performance drop after removing $\mathcal{Q}$ indicates that the removed source contains information not recovered by the remaining source groups.

Task 3: Static-source value without dynamic observations. This task evaluates which static source remains useful when dynamic observations are unavailable. We remove FIRMS/VIIRS and weather by excluding $\mathcal{R}$ and $\mathcal{W}$, and compare metadata plus one static source group, $\boldsymbol{x}_i(\{\mathcal{M}, \mathcal{Q}\})$, for each $\mathcal{Q} \in \{\mathcal{V}, \mathcal{F}, \mathcal{T}, \mathcal{A}, \mathcal{P}\}$. We also evaluate $\boldsymbol{x}_i(\{\mathcal{M}\})$ as the metadata-only lower bound.

Task 4: Weather-history sufficiency. This task evaluates how much near-discovery weather history is needed for IA failure prediction. Let $\ell \in \{1,2,3,4,5\}$ denote the number of days ending on discovery day D_i, with $\mathcal{W}^{(\ell)} \subseteq \mathcal{W}$ the corresponding gridMET weather and fire-danger window. Each model receives $\boldsymbol{x}_i(\{\mathcal{M}, \mathcal{W}^{(\ell)}\})$ and predicts y_i. Comparing ℓ tests whether longer histories improve over discovery-day weather.

Task 5: Auxiliary containment-duration prediction. This task tests whether discovery-day public inputs can predict containment duration. For events with valid timestamps, we define $h_i = t_i^c - t_i^d$ and $y_i^{\mathrm{TTC}} = \log(1 + h_i)$, and predict $\hat{y}_i^{\mathrm{TTC}} = g_{\boldsymbol{\phi}}(\boldsymbol{x}_i(\mathcal{C}))$ with $\hat{h}_i = \exp(\hat{y}_i^{\mathrm{TTC}}) - 1$. It serves as an auxiliary test of how much containment variation is recoverable from discovery-day public data.

3.2 EVALUATION METRICS

IA failure metrics. For IA failure prediction, we use AUPRC as the primary metric because escaped fires are rare and the benchmark emphasizes ranking high-risk events near the top. We also report AUROC, Recall@5%, F1, Brier score, and ECE. These metrics measure ranking quality, early-warning recall, thresholded classification performance, probability error, and calibration.

Containment-duration metrics. For containment-duration prediction, we use MAE in hours as the primary metric after converting predictions back from log space. We also report RMSE, MedianAE, log-space MAE, $R^2_{\log h}$, and Spearman correlation, covering error magnitude, log-space fit, explained variance, and rank consistency.

4 EXPERIMENTAL SETUP

All experiments use the same canonical WILDFIREIA artifacts. Each sample is one FPA-FOD natural wildfire event indexed by fire_id. We use a chronological split, with 2016–2018 fires for training, 2019 for validation, and 2020 for testing. Across all experiments, we keep the event set, labels, forbidden-feature list, and five-seed protocol fixed, so that differences reflect source choices, representations, or model families rather than changes in benchmark construction.

Model-ready representations. After aligning sources by fire_id, the dataloader applies each task and source setting, removes forbidden fields, fits preprocessing on the training split only, and exports model-ready caches. WILDFIREIA supports tabular, temporal, spatial, and spatiotemporal

Table 2: Full-Source Predictability leaderboard. Values are percentages and shown as mean±std. AUPRC is the primary metric; AUROC and Recall@5% evaluate ranking, F1 evaluates a validation-selected operating point, and Brier/ECE evaluate calibration. Best values are bolded.

Family	Model	AUPRC↑	AUROC↑	Recall@5%↑	F1↑	Brier↓	ECE↓
Tabular	LogisticRegression	43.9 ± 0.0	81.5 ± 0.0	32.9 ± 0.0	42.5 ± 0.0	13.6 ± 0.0	15.8 ± 0.0
	XGBoost	$\mathbf{53.3 \pm 0.3}$	$\mathbf{87.1 \pm 0.1}$	$\mathbf{38.0 \pm 0.6}$	$\mathbf{49.5 \pm 0.6}$	9.1 ± 0.1	13.5 ± 0.1
	MLP	50.5 ± 0.5	83.4 ± 0.5	36.7 ± 0.3	47.1 ± 0.6	6.1 ± 0.1	5.9 ± 0.7
Temporal	GRU	49.5 ± 0.1	82.9 ± 0.4	36.3 ± 0.6	46.0 ± 0.8	5.7 ± 0.0	$\mathbf{0.9 \pm 0.3}$
	TCN	50.4 ± 0.3	83.7 ± 0.5	36.4 ± 0.3	46.7 ± 0.6	5.8 ± 0.0	3.5 ± 0.6
	Transformer	50.6 ± 0.3	84.0 ± 0.4	36.5 ± 0.3	46.8 ± 0.6	5.6 ± 0.0	0.9 ± 0.4
Spatial	ResNet18-UNet	50.2 ± 1.7	84.8 ± 0.7	35.4 ± 1.1	47.2 ± 0.8	5.8 ± 0.3	4.3 ± 1.4
	ResNet50-UNet	49.4 ± 0.6	84.0 ± 0.9	34.9 ± 1.2	44.8 ± 1.5	5.8 ± 0.1	3.3 ± 0.3
	Swin-UNet	51.1 ± 1.2	85.3 ± 0.6	36.5 ± 1.8	46.7 ± 1.9	5.6 ± 0.1	2.6 ± 0.6
	SegFormer	50.6 ± 0.5	84.9 ± 0.9	35.8 ± 0.4	46.6 ± 1.2	5.7 ± 0.1	2.3 ± 0.3
Spatiotemporal	ConvLSTM	38.9 ± 1.7	83.8 ± 0.3	29.7 ± 2.0	40.4 ± 1.0	6.3 ± 0.1	3.3 ± 0.4
	ConvGRU	41.2 ± 1.6	84.1 ± 0.6	31.6 ± 0.9	41.6 ± 1.1	6.2 ± 0.1	3.5 ± 0.2
	ResNet3D	51.3 ± 1.7	84.5 ± 2.0	36.5 ± 1.5	46.7 ± 1.1	5.7 ± 0.1	3.5 ± 1.3
	PredRNN-V2	39.6 ± 1.6	83.4 ± 0.4	30.8 ± 1.1	41.8 ± 2.0	6.3 ± 0.1	2.7 ± 0.5
	UTAE	51.3 ± 1.5	85.1 ± 1.2	36.2 ± 0.8	46.6 ± 1.1	5.7 ± 0.2	3.3 ± 1.2
	SwinLSTM	51.7 ± 1.2	85.7 ± 0.9	37.3 ± 1.3	47.1 ± 2.3	$\mathbf{5.6 \pm 0.1}$	2.4 ± 0.7

representations under a common benchmark protocol; detailed input shapes and baseline families are provided in Appendix E.

Baseline model suite. We evaluate 16 baselines across tabular, temporal, spatial, and spatiotemporal families. The suite covers structured-data models, recurrent and attention-based sequence encoders, convolutional and transformer-based spatial encoders, and recurrent or 3D spatiotemporal architectures (Chen & Guestrin, 2016; Cho et al., 2014; Bai et al., 2018; Vaswani et al., 2017; He et al., 2016; Ronneberger et al., 2015; Liu et al., 2021; Cao et al., 2021; Xie et al., 2021; Shi et al., 2015; Wang et al., 2022; Fare Garnot & Landrieu, 2021; Tang et al., 2023). Segmentation-style architectures are adapted by replacing dense prediction heads with global pooling and a scalar output, so every model predicts either initial attack failure probability or auxiliary containment duration.

Training and leakage control. All preprocessing is fit on the training split only and then fixed for validation and test events. Before training, the dataloader excludes final fire size, containment timestamps, MTBS identifiers, label-derived variables, and all detections from $D_i + 1$ onward. Discovery-day FIRMS/VIIRS features are allowed; unmatched events are retained with zero-valued thermal features and a detection indicator.

5 EXPERIMENTS

We organize the experiments around the five research questions defined in Section 2.1. All experiments share the same event unit, chronological split, leakage policy, and evaluation protocol, making results comparable across model families, source combinations, and weather windows.

Full-Source Predictability. To answer **RQ1**, we evaluate all discovery-day environmental and contextual inputs across tabular, temporal, spatial, and spatiotemporal model families. Table 2 reports the full metric suite over five seeds, and Figure 3 shows the AUPRC leaderboard by representation family. We find that: (1) The full-source setting contains a clear predictive signal: all major representation families substantially outperform a trivial rare-event baseline, showing that initial attack failure can be predicted from public discovery-day and static contextual variables. (2) The task behaves primarily as event-level risk prediction rather than dense spatial forecasting. Higher-capacity spatial and spatiotemporal models do not automatically outperform structured event-level models, and the best results are obtained by XGBoost. (3) Patch-based models are competitive, indicating that local landscape and discovery-day spatial context are useful, but they do not replace aggregated event-level summaries. At the same time, the best Recall@5% remains limited, showing that public discovery-day inputs do not fully determine suppression outcome. Appendix Table 9 reports performance heterogeneity across U.S. Census regions.

Table 3: Source Necessity under Full Input for initial attack escape prediction. Each cell is AUPRC in percent after removing one source from the full input, reported as mean±standard deviation over five random seeds. The final column reports the unablated full-input result.

Model	w/o FIRMS	w/o Weather	w/o Vegetation	w/o Fuel	w/o Topography	w/o Access	w/o Population	Full
XGBoost	38.3 ± 0.5	50.5 ± 0.5	$\mathbf{53.6 \pm 0.3}$	52.1 ± 0.3	53.1 ± 0.2	53.6 ± 0.4	51.2 ± 0.4	53.3 ± 0.3
Transformer	32.3 ± 0.2	49.2 ± 0.3	$\mathbf{51.4 \pm 0.3}$	49.7 ± 0.3	50.7 ± 0.2	50.2 ± 0.5	50.3 ± 0.6	50.6 ± 0.3
Swin-UNet	30.4 ± 1.0	48.8 ± 1.2	48.2 ± 1.5	49.5 ± 1.1	50.6 ± 0.9	46.2 ± 1.1	50.0 ± 1.4	$\mathbf{51.1 \pm 1.2}$
SwinLSTM	33.4 ± 1.1	48.1 ± 0.8	51.3 ± 0.7	$\mathbf{52.2 \pm 0.4}$	51.0 ± 1.0	50.6 ± 0.9	51.6 ± 1.1	51.7 ± 1.2

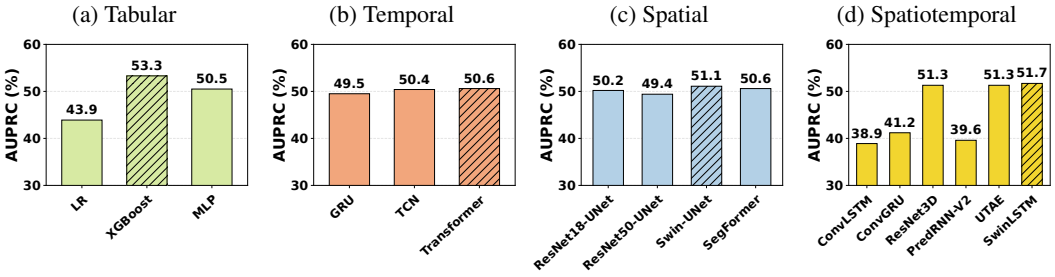


Figure 3: Full-Source Initial Attack Predictability leaderboard. Bar heights report test AUPRC (%). Hatched bars indicate the best-performing model.

Source Necessity under Full Input. To answer **RQ2**, we perform leave-one-source-out ablations and compare each setting with the full-source protocol. This measures marginal necessity: a large AUPRC drop indicates that the removed source provides signal not recovered by the remaining inputs. Table 3 shows that FIRMS/VIIRS is the least redundant source, as removing discovery-day thermal evidence causes the largest degradation across representative models. Weather provides a smaller but consistent gain, suggesting partial overlap with other contextual variables. Vegetation, fuel, topography, access, and population have marginal or even slightly negative effects in the full-input setting, indicating that their information is partly covered once FIRMS/VIIRS, weather, metadata, and other contextual sources are available.

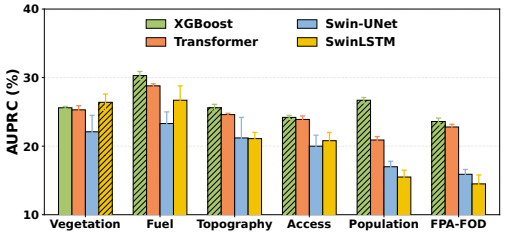
Static-source value without dynamic observations. To answer **RQ3**, we remove both dynamic sources and evaluate two conditions: FPA-FOD metadata alone as the reference baseline, and FPA-FOD metadata combined with one static source group at a time. This fallback setting measures which static information remains useful without weather history or discovery-day satellite evidence. Figure 4a shows that fuel provides the largest and most consistent improvement over metadata alone across representative models, suggesting that combustible landscape structure provides meaningful background risk signal. Gains from vegetation, topography, access, and population are smaller and more model-dependent. Together with Section 5.2, these results show that source value depends on the available observation context: fuel contributes useful fallback information even though its marginal contribution is limited when the full source stack is available. Nevertheless, performance with individual static sources remains substantially below the full-input results. Static context can therefore support early risk assessment when dynamic observations are unavailable, but offers only a partial substitute for current fire signal and near-discovery weather observations.

Weather-History Sufficiency. To answer **RQ4**, we vary the weather and fire-danger window from discovery day to the full $D_i{-}4{:}D_i$ history while keeping the initial attack failure target fixed. Each model receives FPA-FOD metadata and weather history, isolating the predictive value of extending this window. Figure 4b shows that longer weather histories do not consistently improve prediction: the best or near-best AUPRC values occur with one or two days of weather, while adding three to five days provides little benefit. This suggests that, under the current event-level target and gridMET feature design, the useful weather signal is concentrated near discovery day. A short weather window thus provides a compact baseline, although longer antecedent conditions may remain physically relevant and could require different temporal representations to yield additional predictive value.

Table 4: Containment-Duration prediction over five random seeds. Models predict log containment hours; MAE, RMSE, and MedianAE are reported after converting predictions back to hours. Lower is better for error metrics, and higher is better for $R^2_{\log h}$ and Spearman correlation.

Family	Model	MAE h↓	RMSE h↓	MedianAE h↓	log MAE↓	$R^2_{\log h}$↑	Spearman↑
Tabular	RidgeRegression	39.3 ± 0.0	129.1 ± 0.0	8.4 ± 0.0	1.319 ± 0.000	0.058 ± 0.000	0.345 ± 0.000
	XGBoost	$\mathbf{35.1 \pm 0.0}$	$\mathbf{121.3 \pm 0.2}$	8.2 ± 0.1	$\mathbf{1.197 \pm 0.002}$	0.172 ± 0.004	$\mathbf{0.411 \pm 0.002}$
	MLP	35.5 ± 0.3	125.6 ± 0.8	$\mathbf{4.9 \pm 0.2}$	1.209 ± 0.011	0.176 ± 0.016	0.396 ± 0.005
Temporal	GRU	36.4 ± 0.3	126.2 ± 3.1	7.3 ± 0.1	1.222 ± 0.006	$\mathbf{0.177 \pm 0.011}$	0.375 ± 0.007
	TCN	36.1 ± 0.3	124.7 ± 0.7	6.6 ± 0.3	1.226 ± 0.003	0.157 ± 0.007	0.376 ± 0.005
	Transformer	36.0 ± 0.3	124.1 ± 1.3	6.9 ± 0.5	1.216 ± 0.010	0.168 ± 0.015	0.377 ± 0.007
Spatial	ResNet18-UNet	36.0 ± 0.3	125.2 ± 0.4	8.1 ± 0.1	1.231 ± 0.006	0.150 ± 0.013	0.365 ± 0.011
	ResNet50-UNet	37.2 ± 2.3	128.5 ± 2.6	8.4 ± 0.5	1.263 ± 0.014	0.110 ± 0.018	0.338 ± 0.009
	Swin-UNet	36.9 ± 0.4	128.7 ± 6.0	8.3 ± 0.5	1.274 ± 0.026	0.114 ± 0.028	0.333 ± 0.024
	SegFormer	36.9 ± 0.6	126.9 ± 2.8	8.3 ± 0.4	1.263 ± 0.021	0.106 ± 0.025	0.337 ± 0.019
Spatiotemporal	ConvLSTM	36.5 ± 0.1	125.8 ± 0.5	7.8 ± 0.4	1.251 ± 0.009	0.136 ± 0.013	0.341 ± 0.008
	ConvGRU	36.4 ± 0.1	125.5 ± 0.4	7.8 ± 0.2	1.243 ± 0.004	0.150 ± 0.008	0.353 ± 0.005
	ResNet3D	37.5 ± 2.3	127.4 ± 1.2	8.4 ± 0.6	1.266 ± 0.042	0.126 ± 0.055	0.356 ± 0.022
	PredRNN-V2	36.4 ± 0.1	125.6 ± 0.4	7.9 ± 0.3	1.243 ± 0.008	0.145 ± 0.009	0.353 ± 0.008
	UTAE	37.6 ± 2.1	131.7 ± 3.8	6.6 ± 1.9	1.300 ± 0.060	0.080 ± 0.114	0.309 ± 0.019
	SwinLSTM	36.3 ± 0.2	125.6 ± 0.7	7.8 ± 0.6	1.231 ± 0.014	0.151 ± 0.016	0.365 ± 0.010

(a) Static-source ablation



(b) Weather-history sufficiency

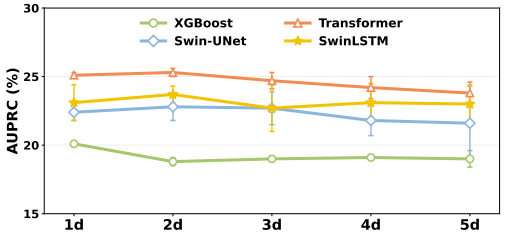


Figure 4: **Left:** Static-source ablation across representative models. **Right:** Weather-history sufficiency for initial attack escape prediction. Bars and lines report mean test AUPRC (%) over five seeds, with error bars indicating one standard deviation.

Auxiliary Containment-Duration Signal. To answer **RQ5**, we replace the initial attack label with log containment hours to test the boundary of the discovery-day benchmark contract. Table 4 shows XGBoost as the strongest baseline, indicating that public discovery-day inputs contain some severity signal. However, the low $R^2_{\log h}$ suggests that containment duration is only weakly explained by these inputs and is largely shaped by post-discovery resources, tactics, weather evolution, and reporting practices. This reinforces our main design choice: WILDFIREIA is best suited for early event-level triage, while containment duration serves as an auxiliary stress test of the benchmark's limits.

6 CONCLUSION

We introduce WILDFIREIA, a reproducible AI benchmark for event-level wildfire initial attack failure prediction from public U.S. national environmental data. WILDFIREIA fixes the event unit, discovery-day input contract, leakage policy, temporal split, metrics, and model-ready representations, enabling fair comparison across tabular, temporal, spatial, and spatiotemporal models. Experiments show that discovery-day public data contains useful early-risk signal, while also revealing source-specific contributions and the limits of containment-duration prediction.

AI USE STATEMENT

Generative AI tools were used for writing and editing assistance, literature and reference search and verification, and feedback on experimental design, robustness analysis, and interpretation of results. Generative AI tools were also used for limited code editing and debugging. All AI-assisted content and code were reviewed and tested by the authors, and all experiments were executed by the authors.

REFERENCES

John T. Abatzoglou. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131, 2013. doi: https://doi.org/10.1002/joc.3413. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3413.

John T. Abatzoglou and A. Park Williams. Impact of anthropogenic climate change on wildfire across western us forests. *Proceedings of the National Academy of Sciences*, 113(42):11770–11775, 2016. doi: 10.1073/pnas.1607171113.

M. Cecilia Arienti, Steven G. Cumming, and Stan Boutin. Empirical models of forest fire initial attack success probabilities: the effects of fuels, anthropogenic linear features, fire weather, and management. *Canadian Journal of Forest Research*, 36(12):3155–3166, 2006. doi: 10.1139/x06-188.

Cong Bai, Feng Sun, Jinglin Zhang, Yi Song, and Shengyong Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi: 10.1109/LGRS.2022.3162882.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018. URL https://arxiv.org/abs/1803.01271.

Jennifer K. Balch, Bethany A. Bradley, John T. Abatzoglou, R. Chelsea Nagy, Emily J. Fusco, and Adam L. Mahood. Human-started wildfires expand the fire niche across the united states. *Proceedings of the National Academy of Sciences*, 114(11):2946–2951, 2017. doi: 10.1073/pnas.1617394114.

Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations, 2016. URL https://arxiv.org/abs/1511.06432.

Shashank Bhardwaj. Predicting the containment time of california wildfires using machine learning, 2025. URL https://arxiv.org/abs/2512.09835.

Leo Breiman. Random forests. *Machine Learning*, 2001. URL https://doi.org/10.1023/A:1010933404324.

GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Kristy Butler, Elena Tartaglia, Jason Rennie, Stephen Deutsch, and Nick McCarthy. An application-specific approach to modelling the probability of unsuccessful initial attack on wildfires in victoria, australia. *bioRxiv*, 2025. doi: 10.1101/2025.07.31.668020. URL https://www.biorxiv.org/content/early/2025/08/02/2025.07.31.668020.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021. URL https://arxiv.org/abs/2105.05537.

Adrián Cardil, Adrián Jiménez-Ruano, Santiago Monedero, Phillip SeLegue, Macarena Ortega, Raúl Quilez, Jeff Fuentes, Geoff Marshall, Robert Clark, Tim Chavez, Sergio de Miguel, Tomás Quiñones, Fernando Bezares, Scott Purdy, Kristen Allison, Tim Metzger, Chris Waters, Midhun Mohan, and Joaquín Ramirez. Assessing the suppression difficulty of wildland fires for initial attack response. *International Journal of Wildland Fire*, 34(12):WF24160, 2025. doi: 10.1071/WF24160.

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao. Mau: A motion-aware unit for video prediction and beyond. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26950–26962. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e25cfa90f04351958216f97e3efdabe9-Paper.pdf.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. URL https://arxiv.org/abs/1802.02611.

Mengxuan Chen, Li Li, Zou Ziheng, Fang Wang, Jinxiao Zhang, Runmin Dong, Juepeng Zheng, and Haohuan Fu. Seasonbench-ea: A multi-source benchmark for seasonal prediction and numerical model post-processing in east asia. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4977. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/dabdd0930e28cbda8da623129f6971b8-Paper-Datasets_and_Benchmarks_Track.pdf.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Santos Daniel Chicas and Jonas Østergaard Nielsen. Who are the actors and what are the factors that are used in models to map forest fire susceptibility? a systematic review. *Natural Hazards*, 114: 2417–2434, 2022. doi: 10.1007/s11069-022-05495-5.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. URL https://arxiv.org/abs/1406.1078.

Janice L. Coen, Marques Cameron, John Michalakes, Edward G. Patton, Philip J. Riggan, and Kara M. Yedinak. Wrf-fire: Coupled weather–wildland fire modeling with the weather research and forecasting model. *Journal of Applied Meteorology and Climatology*, 52(1):16–38, January 2013. ISSN 1558-8432. doi: 10.1175/jamc-d-12-023.1. URL http://dx.doi.org/10.1175/JAMC-D-12-023.1.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 1995.

D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 07 1958. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1958.tb00292.x. URL https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

Weizhi Deng, Jun Wang, Meng Zhou, Xi Chen, Xiaodong Wu, Huanxin Zhang, Jason B. Cohen, Jing Wei, Arlindo da Silva, Guy P. Brasseur, Claire Granier, and Laurence Rouil. Fires reverse progress toward ozone air quality standards in the united states. *Science*, 392(6802):1088–1092, 2026. doi: 10.1126/science.aed3197. URL https://www.science.org/doi/abs/10.1126/science.aed3197.

Philip E. Dennison, Simon C. Brewer, James D. Arnold, and Max A. Moritz. Large wildfire trends in the western united states, 1984–2011. *Geophysical Research Letters*, 41(8):2928–2933, 2014. doi: 10.1002/2014GL059576.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4852–4861, 2021. doi: 10.1109/ICCV48922.2021.00483.

Jean-Baptiste Filippi, Frédéric Bosseur, and Damien Grandi. Forefire: Open-source code for wildland fire spread models. In *Advances in Forest Fire Research*, pp. 275–282. Imprensa da Universidade de Coimbra, 2014. doi: 10.14195/978-989-26-0884-6_29.

Mark A. Finney. *FARSITE: Fire Area Simulator-model development and evaluation*. 1998. doi: 10.2737/rmrs-rp-4. URL http://dx.doi.org/10.2737/RMRS-RP-4.

Mark A. Finney. An overview of flammap fire modeling capabilities. In Patricia L. Andrews and Bret W. Butler (eds.), *Fuels Management—How to Measure Success: Conference Proceedings*, Proceedings RMRS-P-41, pp. 213–220, Fort Collins, CO, 2006. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. URL https://research.fs.usda.gov/treesearch/25948.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting, 2023. URL https://arxiv.org/abs/2207.05833.

Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate*, 30(14), 2017. doi: 10.1175/JCLI-D-16-0758.1. URL https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74515–74529. Curran Associates, Inc., 2023. doi: 10.52202/075280-3258. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ebd545176bdaa9cd5d45954947bd74b7-Paper-Datasets_and_Benchmarks.pdf.

Rafik Ghali and Moulay A. Akhloufi. Deep learning approaches for wildland fires using satellite remote sensing data: Detection, mapping, and prediction. *Fire*, 6(5), 2023a. ISSN 2571-6255. doi: 10.3390/fire6050192. URL https://www.mdpi.com/2571-6255/6/5/192.

Rafik Ghali and Moulay A. Akhloufi. Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. *Remote Sensing*, 15(7), 2023b. ISSN 2072-4292. doi: 10.3390/rs15071821. URL https://www.mdpi.com/2072-4292/15/7/1821.

Louis Giglio, Wilfrid Schroeder, and Christopher O. Justice. The collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment*, 178:31–41, 2016. doi: 10.1016/j.rse.2016.02.054.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Cécile Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

Pingfan Hu, Rachel Tanchak, and Qingsheng Wang. Developing risk assessment framework for wildfire in the united states – a deep learning approach to safety and sustainability. *Journal of Safety and Sustainability*, 1(1):26–41, 2024. ISSN 2949-9267. doi: https://doi.org/10.1016/j.jsasus.2023.09.002. URL https://www.sciencedirect.com/science/article/pii/S2949926723000033.

Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. doi: 10.1109/TGRS.2022.3192974.

W. Matt Jolly, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, and David M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6:7537, 2015. doi: 10.1038/ncomms8537.

Hari Katuwal, Michael S. Hand, Matthew P. Thompson, Crystal S. Stonesifer, and David E. Calkin. Predict and attack (or don't): An econometric approach to large wildfire early detection and suppression effectiveness. In *2018 Agricultural and Applied Economics Association Annual Meeting*, Washington, D.C., 2018. doi: 10.22004/ag.econ.274304.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Rey Koki, Michael McCabe, Dhruv Kedar, Josh Myers-Dean, Annabel Wade, Jebb Stewart, Christina Kumler-Bonfanti, and Jed Brown. Smokeviz: A large-scale satellite dataset for wildfire smoke detection and segmentation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1383. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3b2116b636f64553d7fe33d87a701b06-Paper-Datasets_and_Benchmarks_Track.pdf.

Kennedy Korkola, Melanie Wheatley, Jennifer Beverly, Patrick M. A. James, and Mike Wotton. A comparative analysis of wildfire initial attack containment objectives and modelling strategies in ontario, canada. *International Journal of Wildland Fire*, 33(12):WF24104, 12 2024. ISSN 1049-8001. doi: 10.1071/WF24104. URL https://doi.org/10.1071/WF24104.

Saad Lahrichi, Jake Bova, Jesse Johnson, and Jordan Malof. Improved wildfire spread prediction with time-series data and the wsts+ benchmark, 2025. URL https://arxiv.org/abs/2502.12003.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

LANDFIRE. LANDFIRE: Landscape fire and resource management planning tools. https://www.landfire.gov/, 2026. Accessed 2026.

Yohan Lee, Jeremy S. Fried, Heidi J. Albers, and Robert G. Haight. Deploying initial attack resources for wildfire suppression: Spatial coordination, budget constraints, and capacity constraints. *Canadian Journal of Forest Research*, 43(1):56–65, 2013. doi: 10.1139/cjfr-2011-0433.

Bronte Sihan Li and Ryan Rad. Wildfire spread prediction in north america using satellite imagery and vision transformer. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 1536–1541, 2024. doi: 10.1109/CAI59869.2024.00278.

Yanzhi Li, Keqiu Li, Guohui Li, Zumin Wang, Changqing Ji, Lubo Wang, Die Zuo, Qing Guo, Feng Zhang, Manyu Wang, and Di Lin. Sim2real-fire: A multi-modal simulation dataset for forecast and backtracking of real-world forest fire. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1428–1442. Curran Associates, Inc., 2024. doi: 10.52202/079017-0045. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/02e978a2cc9a1d0d4376a7deb01db612-Paper-Datasets_and_Benchmarks_Track.pdf.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015. URL https://arxiv.org/abs/1411.4038.

Mohammad Marjani, Seyed Ali Ahmadi, and Masoud Mahdianpari. Firepred: A hybrid multi-temporal convolutional neural network model for wildfire spread prediction. *Ecological Informatics*, 78:102282, 2023. doi: 10.1016/j.ecoinf.2023.102282. URL https://www.sciencedirect.com/science/article/pii/S1574954123003114.

Mohammad Marjani, Masoud Mahdianpari, Seyed Ali Ahmadi, Emadoddin Hemmati, Fariba Mohammadimanesh, and Mohammad Saadi Mesgari. Application of explainable artificial intelligence in predicting wildfire spread: An aspp-enabled cnn approach. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. doi: 10.1109/LGRS.2024.3417624.

NASA. Fire information for resource management system (firms). https://firms.modaps.eosdis.nasa.gov/, 2026. Accessed: February 13, 2026.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction, 2024. URL https://arxiv.org/abs/2402.00712.

National Interagency Fire Center. National fire news. https://www.nifc.gov/fire-information/nfn, 2026. Accessed: 2026-06-07.

National Wildfire Coordinating Group. Nwcg glossary of wildland fire, pms 205. https://www.nwcg.gov/publications/pms205/nwcg-glossary-of-wildland-fire-pms-205, 2026. Accessed: September 20, 2026.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. URL https://arxiv.org/abs/1804.03999.

Y. Pourmohamad, J. T. Abatzoglou, E. J. Belval, E. Fleishman, K. Short, M. C. Reeves, N. Nauslar, P. E. Higuera, E. Henderson, S. Ball, A. AghaKouchak, J. P. Prestemon, J. Olszewski, and M. Sadegh. Physical, social, and biological attributes for improved understanding and prediction of wildfires: Fpa fod-attributes dataset. *Earth System Science Data*, 16(6):3045–3060, 2024. doi: 10.5194/essd-16-3045-2024. URL https://essd.copernicus.org/articles/16/3045/2024/.

David Radke, Anna Hessler, and Dan Ellsworth. Firecast: Leveraging deep learning to predict wildfire spread. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 4575–4581. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/636. URL https://doi.org/10.24963/ijcai.2019/636.

Eghbal Rashidi, Hugh R. Medal, and Aaron Hoskins. An attacker-defender model for analyzing the vulnerability of initial attack in wildfire suppression. *Naval Research Logistics*, 65(6):449–464, 2018. doi: 10.1002/nav.21792.

Marcos Rodrigues, Fermín Alcasena, and Cristina Vega-García. Modeling initial attack success of wildfire suppression in catalonia, spain. *Science of The Total Environment*, 666:915–927, 2019. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2019.02.323. URL https://www.sciencedirect.com/science/article/pii/S0048969719308319.

Francisco Rodríguez y Silva, Christopher D. O'Connor, Matthew P. Thompson, Juan Ramón Molina Martínez, and David E. Calkin. Modelling suppression difficulty: current and future applications. *International Journal of Wildland Fire*, 29(8):739–751, May 2020. ISSN 1448-5516. doi: 10.1071/wf19042. URL http://dx.doi.org/10.1071/WF19042.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E. Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik Hamann, Campbell Watson, Manil Maskey, Tsengdar J. Lee, Juan Bernabe Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate, 2024.

Wilfrid Schroeder, Patricia Oliva, Louis Giglio, and Ivan A. Csiszar. The new viirs 375m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143:85–96, 2014. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2013.12.008. URL https://www.sciencedirect.com/science/article/pii/S0034425713004483.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015. URL https://arxiv.org/abs/1506.04214.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model, 2017. URL https://arxiv.org/abs/1706.03458.

Assaf Shmuel and Eyal Heifetz. Global wildfire susceptibility mapping based on machine learning models. *Forests*, 13(7), 2022. ISSN 1999-4907. doi: 10.3390/f13071050. URL https://www.mdpi.com/1999-4907/13/7/1050.

Karen C. Short. Spatial wildfire occurrence data for the united states, 1992–2020 [fpa_fod_20221014], 2022.

Samriddhi Singla, Ayan Mukhopadhyay, Michael Wilbur, Tina Diao, Vinayak Gajjewar, Ahmed Eldawy, Mykel J Kochenderfer, Ross Shachter, and Abhishek Dubey. Wildfiredb: An open-source dataset connecting wildfire occurrence with relevant determinants. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/3fe94a002317b5f9259f82690aeea4cd-Paper-round2.pdf.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021. URL https://arxiv.org/abs/2105.05633.

Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm:improving spatiotemporal prediction accuracy using swin transformer and lstm, 2023. URL https://arxiv.org/abs/2308.09891.

Andrew J. Tatem. Worldpop, open data for spatial demography. *Scientific Data*, 4:170004, 2017.

U.S. Geological Survey and USDA Forest Service. Monitoring trends in burn severity (mtbs). https://www.mtbs.gov/, 2026. Accessed: February 13, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning, 2022. URL https://arxiv.org/abs/2103.09504.

A. L. Westerling, H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam. Warming and earlier spring increase western u.s. forest wildfire activity. *Science*, 313(5789):940–943, 2006. doi: 10.1126/science.1128834. URL https://www.science.org/doi/abs/10.1126/science.1128834.

Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarseer: Versatile spatio-temporal dynamical systems modeling in one model, 2024. URL https://arxiv.org/abs/2312.08403.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021. URL https://arxiv.org/abs/2105.15203.

Yiqing Xu, Kaiwen Zhou, and Fuquan Zhang. Modeling wildfire initial attack success rate based on machine learning in liangshan, china. *Forests*, 14(4):740, 2023. doi: 10.3390/f14040740.

Yu Zhao, Sebastian Gerard, and Yifang Ban. Ts-satfire: A multi-task satellite image time-series dataset for wildfire detection and prediction, 2024. URL https://arxiv.org/abs/2412.11555.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2021. URL https://arxiv.org/abs/2012.15840.

SUPPLEMENTARY MATERIALS

A METRIC DEFINITIONS

Task 1 Classification Metrics. For the initial attack failure task, let p_i^{IA} denote the predicted failure probability and $y_i^{\mathrm{IA}} \in \{0, 1\}$ denote the binary label.

The area under the precision–recall curve is

$$\mathrm{AUPRC} = \int_0^1 \mathrm{Precision}(r)\, dr.$$

Precision and recall at threshold τ are

$$\mathrm{Precision}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FP}(\tau)}, \qquad \mathrm{Recall}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}.$$

AUROC is

$$\mathrm{AUROC} = \Pr(p_i^+ > p_j^-),$$

where p_i^+ is the predicted risk for a failed initial attack event and p_j^- is the predicted risk for a successful initial attack event.

For threshold-based reporting, the operating threshold is selected on the validation set:

$$\tau^\star = \arg\max_\tau \frac{2\,\mathrm{Precision}_{\mathrm{val}}(\tau)\,\mathrm{Recall}_{\mathrm{val}}(\tau)}{\mathrm{Precision}_{\mathrm{val}}(\tau) + \mathrm{Recall}_{\mathrm{val}}(\tau)}.$$

The predicted binary label is then

$$\hat{y}_i^{\mathrm{IA}} = \mathbf{1}[p_i^{\mathrm{IA}} \geq \tau^\star].$$

Balanced accuracy is

$$\mathrm{BalancedAcc} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right).$$

For top-k operational triage metrics, let $\mathcal{S}_k$ be the top $k\%$ of test events ranked by predicted failure probability. Then

$$\text{Recall@}k\% = \frac{\sum_{i\in\mathcal{S}_k} y_i^{\text{IA}}}{\sum_{i\in\mathcal{D}_{\text{test}}^{\text{IA}}} y_i^{\text{IA}}}, \qquad \text{Precision@}k\% = \frac{\sum_{i\in\mathcal{S}_k} y_i^{\text{IA}}}{|\mathcal{S}_k|}.$$

Brier score is

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(p_i^{\text{IA}} - y_i^{\text{IA}})^2.$$

Binary cross-entropy is

$$\text{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i^{\text{IA}}\log p_i^{\text{IA}} + (1-y_i^{\text{IA}})\log(1-p_i^{\text{IA}})\right].$$

Expected calibration error is

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{n}\left|\text{acc}(B_m) - \text{conf}(B_m)\right|,$$

where B_m is the m-th confidence bin, $\text{acc}(B_m)$ is the observed positive frequency in that bin, and $\text{conf}(B_m)$ is the average predicted confidence.

Task 2 Regression Metrics. For the containment-time task, let h_i denote true containment hours, $\hat{h}_i$ denote predicted containment hours after converting from log space, and $\hat{y}_i^{\text{TTC}}$ denote the predicted log-space target.

Mean absolute error in hours is

$$\text{MAE}_h = \frac{1}{n}\sum_{i=1}^{n}|\hat{h}_i - h_i|.$$

Root mean squared error in hours is

$$\text{RMSE}_h = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{h}_i - h_i)^2}.$$

Median absolute error in hours is

$$\text{MedianAE}_h = \text{median}_i\left(|\hat{h}_i - h_i|\right).$$

Log-space MAE is

$$\text{MAE}_{\log} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i^{\text{TTC}} - y_i^{\text{TTC}}|.$$

Log-space RMSE is

$$\text{RMSE}_{\log} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^{\text{TTC}} - y_i^{\text{TTC}})^2}.$$

Spearman correlation is

$$\rho_s = \text{corr}\left(\text{rank}(\hat{h}), \text{rank}(h)\right).$$

The log-space coefficient of determination is

$$R^2_{\log h} = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\text{TTC}} - \hat{y}_i^{\text{TTC}}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\text{TTC}} - \bar{y}^{\text{TTC}}\right)^2}.$$

B Related Work

Initial attack, suppression effectiveness, and suppression difficulty. Initial attack is usually evaluated as a suppression-system outcome, and prior studies show that the definition of "success" depends on containment objective, final-size threshold, response standard, and agency context (Rodrigues et al., 2019; National Wildfire Coordinating Group, 2026; Korkola et al., 2024; Xu et al., 2023; Butler et al., 2025; Katuwal et al., 2018). Accordingly, operational studies model initial attack with variables that are not always available in public national data, including response time, resource deployment, station capacity, aircraft use, and local suppression objectives (Lee et al., 2013; Katuwal et al., 2018; Rashidi et al., 2018; Cardil et al., 2025). Such work also motivates our access and terrain features, because suppression difficulty is affected by fire behavior, terrain difficulty, roads, response opportunities, and suppression capacity rather than by weather alone (Rodríguez y Silva et al., 2020; Cardil et al., 2025; Finney, 1998; 2006). WildfireIA differs from these by fixing a public event-level contract across the United States and excluding operational variables that are not always available near discovery day (Short, 2022; Pourmohamad et al., 2024; Korkola et al., 2024).

Containment duration and post-discovery outcomes. Containment time is related to initial attack failure, but it is not the same target: a fire can escape the first response and still be contained quickly, or it can remain active because of terrain, fuel, weather, resource limits, or reporting conventions (Bhardwaj, 2025; Rodríguez y Silva et al., 2020; Cardil et al., 2025). Machine-learning work on containment duration therefore treats the target as a continuous or transformed time variable rather than as a binary escape label (Bhardwaj, 2025). Due to that distinction, WildfireIA reports containment time as an auxiliary regression task and keeps the initial attack failure task as the primary ranking problem (BRIER, 1950; Guo et al., 2017; Korkola et al., 2024).

Public fire records and environmental covariates. FPA-FOD provides a national event backbone with discovery time, location, cause, final size, and containment fields, while recent FPA-FOD-Attributes work demonstrates the value of attaching physical, biological, social, and administrative covariates to fire occurrence records (Short, 2022; Pourmohamad et al., 2024). Weather and fire danger are commonly derived from gridded meteorological or reanalysis products such as gridMET, ERA5, and MERRA-2, while vegetation, fuel, and terrain features are commonly derived from LANDFIRE and related landscape data (Abatzoglou, 2013; Hersbach et al., 2020; Gelaro et al., 2017; LANDFIRE, 2026; Finney, 1998; 2006). Similarly, access and exposure variables can be attached from vector and demographic products such as OpenStreetMap and WorldPop, but these sources have different spatial resolution, update frequency, and missingness patterns (Haklay & Weber, 2008; Tatem, 2017; Chicas & Nielsen, 2022; Shmuel & Heifetz, 2022). Thus, our canonicalization pipeline follows the public-data integration direction of prior occurrence and susceptibility work while changing the prediction target from ignition or susceptibility to initial attack outcome (Singla et al., 2021; Hu et al., 2024; Shmuel & Heifetz, 2022; Pourmohamad et al., 2024).

Remote sensing for active fire detection, burned area, and fire mapping. Satellite products are central to modern wildfire monitoring, but they measure thermal activity or burned-area evidence rather than initial attack success (Schroeder et al., 2014; Giglio et al., 2016; NASA, 2026; U.S. Geological Survey & USDA Forest Service, 2026). Deep-learning studies and reviews have covered active-fire detection, smoke detection, burned-area segmentation, and satellite-based fire mapping, and these tasks often use image-level or pixel-level labels rather than incident-level suppression outcomes (Ghali & Akhloufi, 2023a;b; Koki et al., 2025; Zhao et al., 2024). Because small or newly discovered fires may lack a matched VIIRS detection because of orbital timing, cloud, smoke, geolocation uncertainty, or intensity limits, WildfireIA treats discovery-day VIIRS as an optional feature and retains events without matched detections (Schroeder et al., 2014; Giglio et al., 2016; Korkola et al., 2024). This choice prevents the benchmark from becoming a satellite-detection benchmark and keeps the sample unit aligned with the fire report (Short, 2022; NASA, 2026; Pourmohamad et al., 2024).

Wildfire spread prediction and fire simulators. Physics-based and coupled fire models such as FARSITE, FlamMap, ForeFire, and WRF-Fire represent fire growth, spread, or fire-atmosphere interaction under explicit environmental assumptions (Finney, 1998; 2006; Filippi et al., 2014; Coen et al., 2013). Data-driven spread benchmarks such as WildfireDB, Next Day Wildfire Spread, WildfireSpreadTS, WSTS+, and Sim2Real-Fire instead evaluate models that predict future fire masks, burned cells, or spread trajectories from remote-sensing and environmental grids (Singla et al., 2021;

Table 5: Representative prior work and how WILDFIREIA differs. The table groups related work by target, because the target determines the valid sample unit, split, and leakage constraints.

Line of work	Examples	Typical target	Difference from WILDFIREIA
Initial attack outcome	Korkola et al.; Xu et al.; Butler et al. (Korkola et al., 2024; Xu et al., 2023; Butler et al., 2025)	IA success or escape	Regional or agency-specific definitions; no shared U.S. multi-modal cache
Suppression difficulty	Lee et al.; Katawal et al.; Cardil et al. (Lee et al., 2013; Katawal et al., 2018; Cardil et al., 2025)	Response success or difficulty	Often uses operational variables not consistently public nationally
Containment duration	Bhardwaj (Bhardwaj, 2025)	Time to containment	Regression target; not a primary IA failure benchmark
Fire occurrence and susceptibility	FPA-FOD-Attributes; WildfireDB; global susceptibility maps (Pourmohamad et al., 2024; Singla et al., 2021; Shmuel & Heifetz, 2022)	Ignition, occurrence, or susceptibility	Predicts where fires occur, not whether a discovered fire escapes IA
Active fire and burned-area mapping	VIIRS/MODIS/MTBS and remote-sensing DL reviews (Schroeder et al., 2014; Giglio et al., 2016; U.S. Geological Survey & USDA Forest Service, 2026; Ghali & Akhloufi, 2023a;b)	Thermal detection or burned pixels	Sensor or pixel labels, not incident-level suppression outcomes
Wildfire spread benchmarks	FireCast, Next Day Wildfire Spread, WildfireSpreadTS, WSTS+, Sim2Real-Fire (Radke et al., 2019; Huot et al., 2022; Gerard et al., 2023; Lahrichi et al., 2025; Li et al., 2024)	Future fire mask or spread trajectory	Forecasts spatial propagation rather than initial attack failure

Table 6: Sensitivity to alternative size-based IA thresholds across four representative models. Each model entry reports AUPRC/AUROC as five-seed means.

Rule (ha)	Excluded	Pos. rate	XGBoost	Transformer	Swin-UNet	SwinLSTM
$\leq 5/\geq 25$	486	10.8%	59.8/87.1	56.3/85.6	54.9/85.1	55.9/85.9
$\leq 10/\geq 50$	414	8.3%	53.3/87.1	50.6/84.0	51.1/85.3	51.7/85.7
$\leq 10/\geq 100$	549	6.3%	51.2/86.5	48.6/85.7	47.6/86.1	48.9/86.9
$\leq 20/\geq 100$	347	6.1%	47.7/85.4	46.1/84.5	46.1/86.0	45.9/85.7
$\leq 10/> 10$	0	14.1%	58.1/85.9	55.6/84.0	53.6/82.7	55.4/84.4

Huot et al., 2022; Gerard et al., 2023; Lahrichi et al., 2025; Li et al., 2024). These resources are important for fire-growth modeling, yet their target is future spatial propagation rather than whether an incident remains below an initial attack size objective (Radke et al., 2019; Marjani et al., 2023; Li & Rad, 2024; Marjani et al., 2024). Accordingly, WILDFIREIA complements spread benchmarks by using an event-level label, a chronological split, and a forbidden-feature policy built around discovery-day suppression prediction (Korkola et al., 2024; Xu et al., 2023; Butler et al., 2025).

Models for tabular, temporal, spatial, and spatiotemporal inputs. Initial attack prediction requires comparing models across heterogeneous input forms, because event reports, weather sequences, landscape patches, and patch sequences encode different parts of the same incident (Short, 2022; Abatzoglou, 2013; LANDFIRE, 2026; Haklay & Weber, 2008; Tatem, 2017). For tabular data, logistic regression, support-vector machines, random forests, gradient boosting, LightGBM, XGBoost, and multilayer perceptrons are common structured-data baselines, and probabilistic evaluation further requires calibration-aware metrics (Cox, 1958; Cortes & Vapnik, 1995; Breiman, 2001; Ke et al., 2017; Chen & Guestrin, 2016; Rumelhart et al., 1986; Guo et al., 2017; BRIER, 1950; Lakshminarayanan et al., 2017). For temporal data, recurrent, convolutional, and attention-based sequence models provide complementary ways to encode short weather histories, while weather and Earth-system forecasting benchmarks motivate careful handling of spatiotemporal covariates (Hochreiter & Schmidhuber, 1997; Bai et al., 2018; Vaswani et al., 2017; Gao et al., 2023; Schmude et al., 2024; Nathaniel et al., 2024; Chen et al., 2025). For spatial and spatiotemporal patches, convolutional encoders, U-Net-style decoders, attention U-Nets, DeepLab-style decoders, transformer segmentation backbones, ConvLSTM/ConvGRU units, PredRNN, TrajGRU, MAU, UTAE, Rainformer, and SwinLSTM-style modules provide established templates for processing gridded or video-like environmental inputs (He et al., 2016; Ronneberger et al., 2015; Oktay et al., 2018; Chen et al., 2018; Cao et al., 2021; Xie et al., 2021; Shi et al., 2015; Ballas et al., 2016; Wang et al., 2022; Shi et al., 2017; Chang et al., 2021; Fare Garnot & Landrieu, 2021; Bai et al., 2022; Tang et al., 2023). Our baseline suite adapts these architectures to scalar event-level prediction by replacing dense prediction heads with global pooling and a scalar output, thereby keeping the comparison focused on the same initial attack and containment labels (Long et al., 2015; Strudel et al., 2021; Zheng et al., 2021; Dosovitskiy et al., 2021; Wu et al., 2024).

C DISCUSSION AND LIMITATIONS

Sensitivity to the size-based IA threshold. The canonical $10/50$ ha rule follows prior size-based initial-attack studies (Rodrigues et al., 2019; Xu et al., 2023). Because this rule is a size-based benchmark definition, we evaluate whether the predictive results and relative model performance remain stable under alternative success and escape thresholds. We repeat the analysis for four representative models and additionally include a no-gap rule that retains all test events. Table 6 reports AUPRC/AUROC under each threshold configuration.

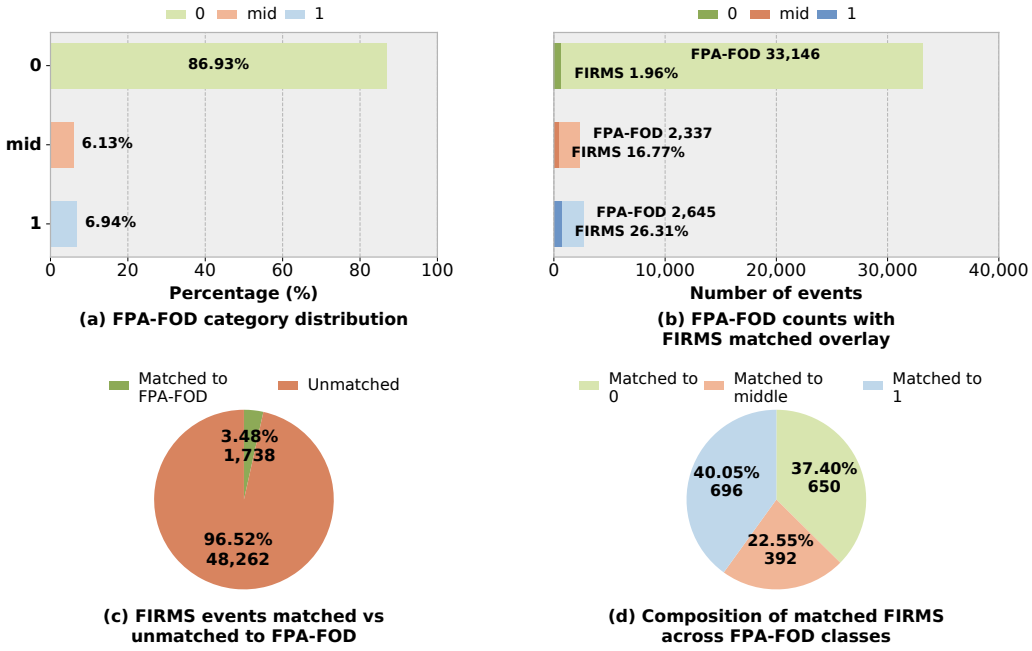


Figure 5: Distribution of FPA-FOD wildfire events and their overlap with FIRMS/VIIRS detections. **(a)** FPA-FOD class proportions. **(b)** Event counts and FIRMS matching rates by class. **(c)** Proportions of FIRMS events matched and unmatched to FPA-FOD. **(d)** Class composition of matched FIRMS events. FIRMS matching rates differ substantially across classes, so the matched subset does not preserve the full FPA-FOD class distribution. We therefore use FIRMS/VIIRS as an input feature source rather than to define the benchmark samples.

The qualitative ordering is broadly stable across threshold choices: XGBoost remains the strongest of the four representative models under all five rules, while the neural baselines remain comparatively close to one another. The no-gap rule retains all 6,582 test events and yields the same overall conclusion, showing that the benchmark signal is not dependent on excluding the intermediate-size fires. Absolute AUPRC values vary with the class prevalence and should therefore not be compared across threshold rules as if the target distribution were unchanged.

FPA-FOD events, not FIRMS detections, define the benchmark. A central design choice in WILDFIREIA is that the sample unit is the FPA-FOD natural wildfire event, not the FIRMS/VIIRS active-fire detection. This choice is necessary because the benchmark target is an event-level initial attack outcome: whether a reported wildfire ultimately remains small or escapes initial attack. FIRMS/VIIRS provides valuable discovery-day thermal evidence, but it does not provide a complete event catalog, a final fire size, or a containment outcome. Therefore, we treat FIRMS/VIIRS as an input source attached to FPA-FOD events rather than as the dataset definition.

Figure 5 illustrates why this distinction matters. The FPA-FOD event distribution is highly imbalanced: the initial attack success class accounts for most retained natural wildfire events, while escaped fires are rare. However, discovery-day FIRMS/VIIRS matching is much more frequent for larger or more extreme events. Only a small fraction of class-0 events have matched discovery-day FIRMS/VIIRS evidence, whereas the matched fraction is substantially higher for intermediate and escaped-fire events. Thus, restricting the benchmark to FIRMS-matched events changes the event population and makes the label distribution less rare-event-like. A model trained and evaluated only on this subset can appear much more accurate, but it is solving an easier and less representative problem.

Performance without matched FIRMS/VIIRS detections. To evaluate performance when discovery-day satellite evidence is unavailable, we separately examine test events without a

Table 7: Performance conditional on discovery-day FIRMS/VIIRS availability. Results are reported as mean±standard deviation over five seeds.

Group	N	Positives	Prevalence	AUPRC	AUROC
No 3 km VIIRS match	5,865	352	6.0%	29.4±0.1	83.2±0.2
With 3 km VIIRS match	303	161	53.1%	88.9±0.6	86.7±0.4

Table 8: Sensitivity to FIRMS/VIIRS timing across four representative models. Each model entry reports AUPRC/AUROC as five-seed means. "Matched" reports the number of events with eligible VIIRS evidence across all splits.

VIIRS protocol	Matched	XGBoost	Transformer	Swin-UNet	SwinLSTM
Same-day	1,738	53.3/87.1	50.6/84.0	51.1/85.3	51.7/85.7
Pre-discovery-time	293	38.1/85.3	31.7/81.7	29.4/80.6	32.1/81.2
No VIIRS	0	38.3/85.2	31.8/81.7	30.5/82.0	29.5/79.8

FIRMS/VIIRS match within the canonical 3 km radius. Table 7 compares this subgroup with events that have matched discovery-day VIIRS detections.

Events without matched discovery-day VIIRS detections form a substantially harder subgroup. XGBoost achieves an AUPRC of 29.4 on the no-match subset, despite weather and the remaining benchmark sources still being available. The matched and unmatched groups have very different positive-class prevalence, so their AUPRC values should not be interpreted as the isolated contribution of VIIRS. Instead, this diagnostic shows that satellite observability is strongly associated with the difficulty and composition of the evaluated event population.

Sensitivity to FIRMS/VIIRS timing. The primary WILDFIREIA protocol uses a discovery-day information contract: FIRMS/VIIRS detections are matched when they occur on the same reported calendar day as the FPA-FOD discovery date, while detections from $D_i + 1$ and later are excluded. Thus, same-day VIIRS acquisitions are not required to occur before the reported discovery time t_i^d.

To quantify the effect of this timing choice, we compare the same-day protocol with a stricter protocol that retains only VIIRS detections acquired at or before t_i^d, as well as a no-VIIRS baseline. We evaluate all four representative models under the same training and evaluation protocol, changing only the VIIRS input. Table 8 reports AUPRC/AUROC as five-seed means on the full test set.

Only 293 of the 1,738 events with same-day VIIRS evidence have a matched acquisition at or before the reported discovery time. Across all four models, the strict pre-discovery-time protocol provides no consistent improvement over removing VIIRS entirely, whereas the same-day protocol yields substantially higher AUPRC. This indicates that the predictive gain associated with FIRMS/VIIRS primarily reflects same-day thermal evidence rather than observations strictly available before the reported discovery time.

We therefore interpret FIRMS/VIIRS as a discovery-day source in the primary benchmark, with the strict pre-discovery-time protocol serving as a sensitivity analysis. This timing sensitivity applies specifically to FIRMS/VIIRS; day-D_i gridMET variables are daily aggregates and remain part of the discovery-day information contract rather than a strict pre-discovery-time contract.

Regional performance. We additionally evaluate XGBoost performance across U.S. Census regions on the 2020 test split. Table 9 reports event counts, positive counts, prevalence, AUPRC, and AUROC by region.

Performance varies substantially across regions. The Northeast estimate is not reliable because the test split contains only one positive event, so region-level results should be interpreted together with sample size and prevalence.

Initial attack failure remains a rare-event prediction problem. The full FPA-FOD-defined benchmark preserves the operational difficulty of initial attack prediction: escaped fires are rare,

Table 9: Regional performance of XGBoost on the 2020 test split.

Region	N	Positives	Prevalence	AUPRC	AUROC
Midwest	543	18	3.3%	38.6±1.0	91.4±0.3
Northeast	581	1	0.2%	3.9±1.1	95.5±1.3
South	765	99	12.9%	66.5±0.3	89.1±0.1
West	4,279	395	9.2%	50.7±0.4	84.0±0.1

Table 10: Sensitivity to the FIRMS/VIIRS spatial matching radius.

Radius	Matched events	Matched escapes	Escape recall
0.5 km	1,186	462	17.5%
1 km	1,446	581	22.0%
3 km	1,738	696	26.3%
5 km	1,892	724	27.4%

and many reported fires remain small. This is why we use AUPRC as the primary metric and report calibration metrics in addition to thresholded classification scores. High accuracy on the full benchmark would be easy to obtain by predicting the majority class, but such a model would be useless for identifying fires that require escalation. The benchmark therefore emphasizes rare-failure prioritization rather than aggregate correctness.

Containment duration is a harder auxiliary target. The auxiliary containment-duration analysis shows that the same discovery-day public source stack contains signal about time to containment, but only partially explains the target. The best models achieve modest $R^2_{\log h}$, around 0.18, indicating that early public variables explain a limited fraction of containment-duration variance. This result is consistent with the nature of the target. Containment duration depends not only on early weather, fuel, topography, access, population context, and satellite fire signal, but also on suppression tactics, crew and aircraft availability, changing weather after discovery, incident command decisions, reporting conventions, and containment criteria. Many of these drivers are not fully observed in our public discovery-day input stack. We therefore interpret containment-duration prediction as an auxiliary stress test rather than as the primary benchmark objective.

Limitations of public event records. FPA-FOD provides a nationwide event backbone, but it is still a reporting-derived dataset. Reported discovery locations may be approximate, final fire size may be recorded with agency-specific conventions, and containment timestamps may be missing or inconsistently reported. Our thresholded initial attack label reduces some ambiguity by separating clearly small fires from clearly escaped fires, but it does not eliminate label noise. Events in the intermediate 10–50 ha range are marked missing for the binary initial attack target to avoid forcing ambiguous cases into either class.

Sensitivity to FIRMS/VIIRS matching radius. FIRMS/VIIRS detections are matched to same-day FPA-FOD events in EPSG:5070. Each detection is assigned to the single nearest eligible event within the matching radius and is never duplicated across fires. A match is flagged as ambiguous when the second-nearest eligible event is within 500 m. We evaluate matching radii of 0.5, 1, 3, and 5 km; Table 10 reports the resulting match counts and escaped-fire coverage.

The 3 km radius captures 92% of the events matched at 5 km. Expanding the radius from 3 to 5 km adds only 154 matched events and 28 escaped fires, increasing escaped-fire recall by only 1.1 percentage points while permitting more distant associations. We therefore retain 3 km as the canonical matching radius.

Sensitivity to discovery-location and date perturbations. We test weather-join robustness to errors in reported discovery location and date by perturbing test-event coordinates by 500 m, 1 km, and 5 km and re-sampling gridMET features. Separately, we shift discovery dates by one day and rebuild the weather windows. Table 11 reports XGBoost performance.

Table 11: Sensitivity to perturbations in reported discovery location and date.

Measurement-error test	Re-sampled source	AUPRC	AUROC
Original	none	53.3	87.1
Jitter 500 m	gridMET + lat/lon	53.6	87.2
Jitter 1 km	gridMET + lat/lon	53.6	87.2
Jitter 5 km	gridMET + lat/lon	53.2	87.2
Date +1 day	gridMET window	52.9	86.7

Table 12: Source-alignment permutation stress test. Selected source features are permuted across test events to break event-level alignment. Results are reported as mean±standard deviation over 5 seeds.

Stress test	Permuted features	AUPRC	AUROC
No permutation	0	53.3±0.3	87.1±0.1
Permute VIIRS same-day	10	28.0±0.7	82.4±0.2
Permute weather window	95	43.0±0.9	81.0±0.3
Permute static context	4,546	31.6±1.1	72.7±1.0
Permute all external joins	4,651	8.9±0.2	52.5±1.1

AUPRC changes by at most 0.4 points under coordinate perturbations of up to 5 km or a one-day shift in discovery date, indicating that the weather join is not highly sensitive to small errors in reported discovery metadata. This is consistent with the approximately 4 km spatial resolution of gridMET and the use of multi-day weather summaries. This diagnostic specifically tests robustness of the weather alignment; LANDFIRE, OpenStreetMap, WorldPop, and other static sources remain fixed at the reported event coordinates.

Source-alignment permutation stress test. To evaluate the dependence of benchmark performance on correct event-level source alignment, we permute selected source groups across test events while leaving the trained model unchanged. Table 12 reports XGBoost performance under each perturbation.

Performance degrades substantially when source groups are deliberately misaligned across events. Permuting static context reduces AUPRC from 53.3 to 31.6, while permuting all external joins reduces AUPRC to 8.9 and AUROC to 52.5. This stress test represents an intentionally severe alignment failure rather than typical reporting noise, but it confirms that correct event-level alignment across sources is essential to benchmark performance.

Limitations of source alignment. All non-FPA-FOD sources are aligned to events through reported discovery date, location, and year. This makes the benchmark reproducible, but it also introduces uncertainty. FIRMS/VIIRS detections depend on overpass timing, cloud and smoke conditions, sensor geometry, and the strictness of the matching radius. gridMET represents coarse weather fields rather than exact fireground conditions. LANDFIRE fuel and vegetation layers are treated as static landscape context, which does not fully capture annual disturbance, treatment, or fuel-moisture dynamics. OSM roads and fire stations are public proxies for access and response proximity, not direct measurements of dispatch decisions, travel time, crew availability, or suppression capacity. WorldPop provides human-settlement context but does not directly encode evacuation priority, asset value, or wildland-urban-interface structure.

Scope of the current benchmark. The current version focuses on natural wildfire events from 2016–2020 and uses a fixed chronological split. This scope reduces ignition-source heterogeneity and supports future-year evaluation, but it does not cover all human-caused wildfire dynamics. Extending the benchmark to all-cause wildfires, adding explicit suppression-resource records, incorporating post-discovery weather evolution, or using high-resolution perimeter growth would enable richer operational questions. However, those additions would also change the information contract. The present benchmark intentionally focuses on what can be attached to a reported natural wildfire near discovery day using reproducible public data.

Table 13: Model-ready representations and baseline families in WILDFIREIA.

Family	Input shape	Models
Tabular	$N \times F$	logistic regression, XGBoost, MLP
Temporal	$N \times T \times F_t$ plus $N \times F_s$	GRU, TCN, Transformer
Spatial	$N \times C \times 29 \times 29$	ResNet18-UNet, ResNet50-UNet, Swin-UNet, SegFormer
Spatiotemporal	$N \times T \times C \times 29 \times 29$	ConvLSTM, ConvGRU, ResNet3D, PredRNN-V2, UTAE, SwinLSTM

D ETHICS STATEMENT

This work uses publicly available environmental, geospatial, and wildfire records and does not involve human subjects or personally identifiable information. WildfireIA is intended to support research toward practical wildfire risk assessment and initial-attack decision support, but its outputs should be interpreted in light of several limitations. In particular, the size-based initial-attack failure label is a benchmark proxy rather than a complete representation of operational outcomes; discovery-day data alignment may not exactly match the information available at every real-world decision time; and the availability and quality of sources such as satellite observations may vary across fires, regions, and operating conditions. Geographic and temporal distribution shifts may further affect model performance outside the benchmark setting. These limitations do not preclude practical use, but they motivate careful validation before deployment. Operational use should therefore combine model outputs with domain expertise, local context, uncertainty awareness, and human oversight rather than treat benchmark predictions as stand-alone decisions.

E REPRODUCIBILITY ARTIFACTS

The released artifacts follow the same end-to-end workflow as the benchmark: raw public data are first transformed into canonical event-level and spatial representations, which are then converted into task- and model-specific caches. Models are trained and evaluated from these fixed representations under the benchmark protocol, with run configurations, predictions, evaluation metrics, and model checkpoints saved for each experiment. Summary utilities aggregate these outputs across tasks, model families, and random seeds to reproduce the reported tables and figures. This allows reported results to be traced from the final tables and figures back to the corresponding experiment outputs. Table 13 summarizes the model-ready input shapes and the baseline architectures used for the tabular, temporal, spatial, and spatiotemporal representations.

F REPRODUCIBILITY STATEMENT

We provide the data-construction pipeline, preprocessing and model-training implementations, experiment configurations, and evaluation artifacts required to reproduce the benchmark. Dataset construction and evaluation protocols are described in Sections 2–4, with robustness analyses and additional reproducibility details in Appendices C–D and the anonymized supplementary materials.