FIRERESBENCH: AN EVENT-CENTERED BENCHMARK FOR WILDFIRE LIFECYCLE UNDERSTANDING AND OPERATIONAL-RESPONSE FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Wildfire emergency management requires responders to continuously assess evolving fire behavior, incident status, containment progress, and resource deployment using scattered and irregular operational records. Existing wildfire datasets have largely been designed for physical fire detection, burned-area mapping, or spread forecasting, leaving evolving incident observations disconnected from real-world response histories. To bridge this gap, we introduce FIRERESBENCH, an event-centered benchmark that connects wildfire progression with emergency-response records, laying the foundation for wildfire-response intelligence. FireResBench supports two complementary tasks: (1) wildfire lifecycle-state understanding and (2) daily personnel and cost forecasting. To provide fire lifecycle supervision, a two-stage labeling paradigm is adopted. In the first stage, carefully designed labeling rules guide an auto-labeling code pipeline that generates provisional labels. In the second stage, domain scholars verify and correct the labels to obtain the final annotations. The released dataset covers the conterminous United States from January 2017 to December 2020 and contains 33,303 Fire-Day instances from 5,728 incidents. It provides scholar-verified lifecycle labels, temporally valid daily personnel and cost targets, incident-disjoint splits, and carefully designed evaluation protocols. Experiments across representative machine learning, tabular, and deep learning paradigms reveal that accurate forecasting on lifecycle transitions and high-demand Fire-Days remain particularly challenging. By converting fragmented public records into auditable event sequences and reproducible learning tasks, FireResBench establishes a foundation for data-driven wildfire-response modeling. The benchmark and evaluation resources are available at: FIRERESBENCH.

1 INTRODUCTION

Wildfires are no longer isolated environmental disturbances; they increasingly unfold as prolonged, resource-intensive crises shaped by weather extremes, environmental and geographic conditions, and human activities and response actions (Burton et al., 2024; Schug et al., 2023). The scale of this burden became starkly evident, when the 2025 Los Angeles wildfires generated approximately $53 billion overall losses, and became the costliest wildfire disaster on record (Munich Re, 2026). Managing such incidents requires more than merely observing where fire is burning. During an active wildfire, responders must repeatedly interpret evolving fire behaviors, incident status, containment progress, and prior resource deployment, while understanding wildfire developing states and making operational decisions (e.g., personnel and cost) under irregular situational reports (SitReps) (St. Denis et al., 2023; Nguyen et al., 2024; Young et al., 2024). This makes wildfire emergency management a natural but underdeveloped research problem at the intersection of data mining and AI, as it combines heterogeneous evidence integration, wildfire lifecycle understanding, geospatial grounding, operational response modeling, and critical decision support (Huot et al., 2022; Xu et al., 2023; Gerard et al., 2023). However, progress in this field requires benchmarks that represent not only the physical evolution of wildfire states, but also how environmental evidence and operational histories jointly shape real-world response decisions over time.

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.

Existing wildfire datasets and studies have largely focused on factual fire phenomena, including active-fire detection, burned-area mapping, fire perimeter estimation, risk assessment, and spread forecasting (Huot et al., 2022; Xu et al., 2023; Gerard et al., 2023; Li et al., 2024; Zhou et al., 2026). These resources have been essential for understanding where fires occur and how they develop, but they usually remain disconnected from the emergency response side of the problem: how personnel, costs, command structures, and escalation actions evolve as a fire unfolds (Rossi et al., 2022; Young et al., 2024; Nguyen et al., 2026). Conversely, although official response records contain rich operational histories, this information remains fragmented across reports and has not been organized into standardized data representations suitable for deep learning-based modeling and evaluation (St. Denis et al., 2023; Nguyen et al., 2024). Across both line of work, existing studies fail to offer unified temporal incident report sequence, wildfire state supervision, valid daily operational response targets, and evaluation protocols. Such absence hinders progress in data-driven wildfire-response modeling and prevents fair and reproducible comparisons across modeling paradigms. The community therefore lacks a unified benchmark for wildfire-response intelligence that evaluates whether models can understand current incident status and anticipate near-term operational response.

Transforming fragmented wildfire operational response records into a unified benchmark raises three fundamental challenges. ***First***, wildfire situation reports follow operational periods rather than a fixed daily schedule, resulting in unequal reporting intervals (e.g., some incidents have multiple records per day). Constructing coherent data sequence requires consistent temporal normalization. ***Second***, the supervision required for wildfire lifecycle state understanding is not directly available in the source records. Reliable supervision must therefore be derived from heterogeneous, temporally evolving evidence while minimizing annotation ambiguity and subjective labeling decisions. ***Third***, incident-level temporal dependence and highly imbalanced target distributions complicate the evaluation. Lifecycle imbalance and long-tailed resource demand can obscure model behavior under rare or extreme conditions. Addressing these challenges requires the construction of temporally consistent data sequence, reliable supervision, and task-specific evaluation protocols.

We introduce FireResBench, an event-centered wildfire response benchmark that connects evolving wildfire dynamics with real-world operational response histories. FireResBench represents each wildfire incident as a sequence of *Fire-Day* instances, where each instance corresponds to one incident on one calendar date. The current release covers the conterminous United States from 2017/01/01 to 2020/12/31, and contains 33,303 Fire-Day instances from 5,728 incidents. FireResBench develops a standardized modeling protocol, provenance-aware features, curated wildfire lifecycle labels, operational response targets, and task-specific indicators. It enables reproducible comparison across mainstream machine learning, time-series and deep learning paradigms. The main contributions of this work are:

- **We introduce a unified, event-centered, and temporally structured wildfire-response benchmark named FIRERESBENCH.** FireResBench organizes irregular wildfire situation reports from 5,728 incidents into 33,303 standardized Fire-Day instances, providing a common temporal representation for studying wildfire evolution and emergency response.
- **We develop auditable supervision for wildfire lifecycle-state understanding and daily operational-response forecasting.** A two-stage annotation protocol combines rule-guided automatic labeling with domain-scholar verification is introduced to produce lifecycle-state labels. We further develop temporally valid daily personnel and cost targets which provide supervision for operational-response forecasting.
- **We formulate two complementary benchmark tasks with a unified, reproducible evaluation protocol.** The two tasks assess wildfire lifecycle state understanding (classification), daily personnel, and cost forecasting using incident-disjoint splits and task-specific metrics. Evaluation across representative model architectures exposes persistent challenges at lifecycle transitions and during high-demand Fire-Days.

Positioning Against Existing Benchmarks. Table 1 compares representative wildfire datasets and benchmarks by coverage, temporal form, operational-response evidence, response-task supervision, lifecycle-state labels, task-ready protocol, and primary task. Physical forecasting benchmarks such as Next Day Wildfire Spread and WildfireSpreadTS provide standard evaluation protocols, but do not include operational-response histories or associated supervision. IMSR aggregates wildfire resource commitments at regional and incident levels, but is released as daily/weekly summary tables

Table 1: Comparison of representative dataset resources which emphasize fire operational response.

Dataset	Coverage	Temporal Form	Operational Resp. Evidence	Response Task Supervision	Lifecycle State Labels	Task-Ready Protocol	Primary Task
Next Day Wildfire Spread	CONUS 2012-2020	Daily grid pairs	✗	✗	✗	✓	Physical fire-spread prediction
WildfireSpreadTS	USA 2018-2021	Daily event sequences	✗	✗	✗	✓	Multimodal physical fire-spread prediction
FPA FOD-Attributes	USA 1992-2020	One record per incident	✗	✗	✗	✗	Ignition and incident-outcome analysis
IMSR	USA 2007-2021	Daily/weekly wildfire summaries	Aggregated personnel, equipment, and incident priority	✗	✗	✗	Resource-utilization and prioritization analysis
ICS-209-PLUS	USA 1999-2020	Incident-level situation report	Incident-level personnel, costs, containment, and narratives	✗	✗	✗	Incident-progression and operational-response analysis
Fireline Engagement QA/QC	USA 2017-2024	Fireline segments and perimeters	Fireline locations, types, and engagement outcomes	✓	✗	✗	Fireline effectiveness and suppression-outcome analysis
WILDFIREIA	CONUS 2016-2020	Discovery-time incident samples	✗	✓	✗	✓	Initial-attack failure prediction
FireResBench	CONUS 2017-2020	Incident-centered Fire-Day sequences	Fire-Day personnel, aerial resources, costs, containment, command, active-fire dynamics	✓	✓	✓	Wildfire lifecycle state understanding; Daily personnel/cost forecasting

Notes. (1) "Operational Resp. Evidence" denotes the deployed personnel or equipment, incident-management states, response costs, or other operational records. (2) "Response-Task Supervision" denotes labels or prediction targets explicitly associated with wildfire response. (3) "Lifecycle State Labels" indicates whether the dataset provides labels describing wildfire lifecycle states. (4) "Task-Ready Protocol" indicates that the benchmark dataset defines a reusable prediction task, model-ready sample representation, and common evaluation setting.

rather than standardized, continuous incident-centered sequential data. ICS-209-PLUS contains historical situation reports (SitReps) across multiple hazards, but the SitReps are temporally irregular and lack necessary supervision. Fireline Engagement QA/QC focuses on the spatial effectiveness of firelines for fires exceeding 1,000 acres, whereas WILDFIREIA focuses on wildfire initial-attack failure prediction; both therefore address narrow response stages rather than the continuous evolution of wildfire incidents across their full lifecycle. In contrast, FireResBench organizes incident histories into standardized Fire-Day sequences and jointly supplies operational evidence, human scholar-verified lifecycle labels, next-day personnel and daily-cost targets, and a unified evaluation protocol. It enables reproducible study of both wildfire-state understanding and operational-response forecasting. More discussion on related works is provided in Appendix C.

2 FireResBench

We introduce FireResBench from benchmark definition to task-ready supervision. Section 2.1 formalizes the two benchmark tasks and summarizes the dataset composition. Section 2.2 explains how we construct the event-centered Fire-Day representation from irregular source reports and how prediction-time evidence is organized without future leakage. Section 2.3 then details the construction and validation of wildfire lifecycle-state labels and operational-response targets.

2.1 Dataset Overview and Statistics

FireResBench organizes wildfire emergency-management data around an event-centered temporal unit, the *Fire-Day*, defined by an incident identifier and a calendar date. Unlike datasets that treat daily observations as independent samples, FireResBench preserves the ordered evolution of each incident and separates information available on day t from outcomes observed on day $t+1$. Its design is event-centered, provenance-aware, temporally valid, and model-compatible: every feature retains its source, unit, derivation, and quality status, while a common schema supports tabular, time-series, and spatiotemporal learning. This representation directly supports two task-specific learning mappings. For Task A, a classifier $f_A(\cdot)$ maps the current multi-source observations and available

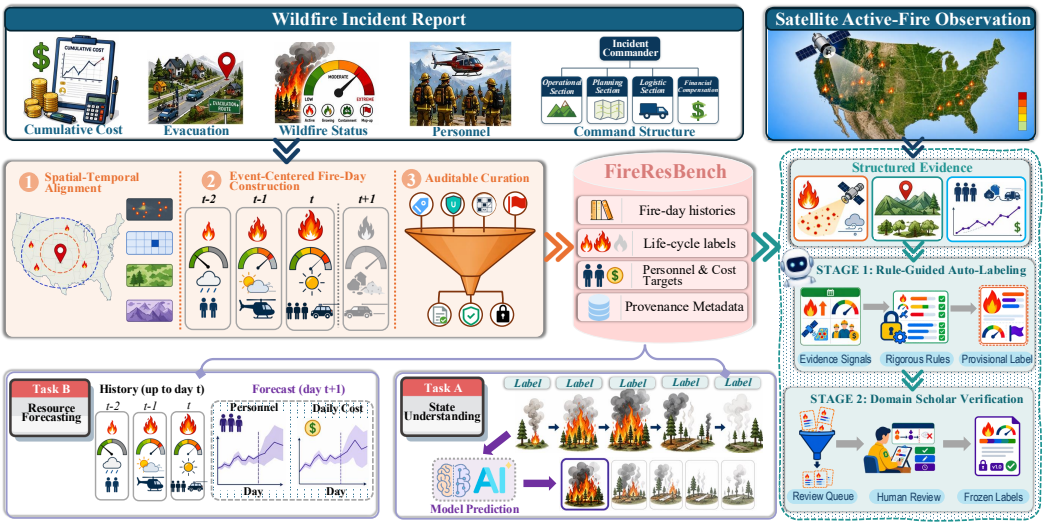


Figure 1: Overview of the FireResBench construction pipeline and benchmark tasks.

incident history to the predicted lifecycle state:

$$\mathbf{x}_{i,t} = [\mathbf{o}_{i,t}, \mathcal{H}_{i,<t}]\,,\ \hat{s}_{i,t} = f_A(\mathbf{x}_{i,t})\,, \quad s_{i,t}, \hat{s}_{i,t} \in \mathcal{S}, \tag{1}$$

where $\mathbf{o}_{i,t}$ contains the observations associated with incident i on day t, $\mathcal{H}_{i,<t}$ contains only incident history available strictly before day t, and $s_{i,t}$ is the curated ground-truth lifecycle state from the five-class set $\mathcal{S}$. For Task B, FireResBench constructs two target-specific forecasting datasets from the temporally normalized Fire-Day sequences. Before model training and evaluation, every retained Fire-Day at t is paired with exactly one record from the same incident on the next calendar day $t+1$. Next-day personnel and daily cost are then modeled by two independent predictors:

$$\begin{aligned} \hat{p}_{i,t+1} &= f_B^{(p)}(\mathbf{x}_{i,t})\,,\ (i,t) \in \mathcal{D}_B^{(p)} \\ \hat{c}_{i,t+1} &= f_B^{(c)}(\mathbf{x}_{i,t})\,,\ (i,t) \in \mathcal{D}_B^{(c)} \end{aligned} \tag{2}$$

where $f_B^{(p)}(\cdot)$ and $f_B^{(c)}(\cdot)$ independently predict next-day personnel $p_{i,t+1}$ and daily cost $c_{i,t+1}$, respectively. $\mathcal{D}_B^{(p)}$ and $\mathcal{D}_B^{(c)}$ denote the corresponding pre-filtered experimental sample sets; they are constructed separately for the two targets, but every included sample satisfies the same strict next-calendar-day pairing. Consequently, temporal validity is enforced during dataset construction rather than through an eligibility variable exposed to the model. Both tasks share the same evidence boundary while differing in their prediction targets: Task A estimates the current lifecycle state, whereas Task B independently forecasts the two next-day operational-response variables.

Table 2 summarizes FireResBench from four perspectives: source-cohort coverage, model-ready composition, task-specific distributions, and incident-history length. The source cohort spans 2017/01/01 to December 2020/12/31, represents all 48 states in the conterminous United States, and contains 36,061 Fire-Days. After coordinate checks and feature alignment, the model-ready dataset comprises 33,303 Fire-Day instances from 5,728 unique incidents, with 216 data and metadata fields. Incident histories vary substantially in duration, with a median of four Fire-Days, a 75th percentile of seven, and a maximum of 117. The task-specific distributions further characterize the learning problems. Task A is relatively balanced across *initial attack*, *rapid escalation*, *extended attack*, *containment*, and *mop-up/monitoring*, which represent 14.24%, 22.13%, 24.47%, 21.72%, and 17.44% of labeled Fire-Days, respectively. Task B exhibits pronounced upper tails: next-day personnel has a median of 57 and a 90th percentile of 306, while daily cost has a median of \$100K and reaches \$1.1M and \$5.2M at the 90th and 99th percentiles.

2.2 Fire-Day Construction and Temporal Evidence Organization

Fire-Day temporal normalization. We first normalize incident identifiers and report timestamps, and define each Fire-Day by the unique key (i,t), where i denotes an incident and t denotes a

calendar date. When multiple reports occur for the same (i, t) pair, they are ordered by the most reliable available timestamp and the latest valid same-day observation is retained as the canonical Fire-Day record. The number and identifiers of contributing source reports are preserved for auditability. Fire-Days are then chronologically ordered within each incident to form an observed event sequence. Missing calendar dates are left unobserved: we neither interpolate report values nor create synthetic Fire-Days.

Table 2: FireResBench dataset composition and task-specific statistics.

Source cohort and coverage			
Statistic			**Value**
Temporal coverage			Jan. 2017-Dec. 2020
Spatial coverage			CONUS; 48 states represented
Source-cohort Fire-Days			36,061
Model-ready dataset			
Statistic			**Value**
Fire-Day instances			33,303
Unique incidents			5,728
Data and metadata fields			216
Task A lifecycle-label distribution			
Lifecycle state			**Share of labeled Fire-Days**
Initial attack			14.24%
Rapid escalation			22.13%
Extended attack			24.47%
Containment			21.72%
Mop-up/monitoring			17.44%
Task B target distribution			
Target	**Median**	**P90**	**P99**
Next-day personnel	57	306	–
Daily cost (USD)	\$100K	\$1.1M	\$5.2M
Incident history length			
Statistic	**Median**	**P75**	**Maximum**
Fire-Days per incident	4	7	117

Source roles and evidence organization. FireResBench assigns each source a distinct role in the benchmark. Wildfire situation reports provide the incident backbone, including reported fire status, containment, impacts, and operational-response histories, and is the source of the numerical targets used in Task B. NASA FIRMS (Schroeder et al., 2014) contributes aligned satellite active-fire evidence that characterizes observed fire activity and supports the construction of Task A supervision. Domain scholars verify and correct the provisional lifecycle annotations for Task A, but do not create or modify the numerical personnel and cost targets for Task B. This separation distinguishes prediction-time evidence from task supervision and preserves the provenance of every released field.

Input Representation and Quality Control. Consistent with Eq. 1, each model input combines the current Fire-Day observation with previously observed Fire-Days from the same incident. Lagged values, temporal differences, and trailing summaries are computed exclusively from records available through day t. The prediction-time view excludes final incident outcomes, supervision-construction metadata, and all future-derived variables. Missingness indicators, source units, provenance, and quality flags are retained explicitly rather than concealed through unconditional imputation. Automated checks enforce unique Fire-Day keys, chronological ordering, valid value ranges, and reproducible feature derivations. Task A and Task B are then generated as task-specific views of this common representation, with their supervision construction described in Section 2.3.

2.3 Task Supervision Construction

Task A: Wildfire lifecycle supervision. The available open-source fire incident reports (ICS-209-PLUS) do not directly provide the lifecycle state labels in $\mathcal{S}$. We therefore develop a two-stage labeling pipeline, as illustrated in Figure 4. ***In Stage I***, researchers and domain scholars freeze the phase definitions, evidence thresholds, rule priority, and temporal constraints, which are then implemented as deterministic autolabeling programs executed by coding agent. For each Fire-Day, the program uses only $\mathbf{o}_{i,t}$ and $\mathcal{H}_{i,<t}$ to derive evidence from incident age, area and growth, containment, FIRMS activity, reported fire behavior, and personnel and cost histories. It assigns one provisional state from *Initial Attack*, *Rapid Escalation*, *Extended Attack*, *Containment*, and *Mop-up/Monitoring*, while retaining the activated evidence, raw rule output, confidence, and transition status. A temporal consistency layer permits stable or forward lifecycle progression and allows a return to rapid escalation only when renewed escalation evidence is observed.

In Stage II, domain scholars review all initial-attack cases, low-confidence or boundary cases, lifecycle transitions, re-escalations, records with missing or conflicting core evidence, and a reproducible phase-stratified audit sample. Each evidence packet truncates the incident history at day

t and excludes future observations, response targets, model outputs, and final incident outcomes. Reviewers may accept, correct, or defer a provisional label; corrections and deferrals receive independent secondary review, and disagreements are resolved by an adjudicating scholar. Corrected incident sequences are re-validated, unresolved cases are excluded from supervised evaluation, and the final annotations are frozen before data splitting and model development. The resulting label $s_{i,t}$ is released as the curated ground-truth lifecycle state together with its rule and annotation. Appendix A specifies the complete evidence definitions, thresholds, transition rules, and review protocol.

Task B: Operational-response target construction. Task B derives two independent regression targets from open wildfire incident SitReps. Let $p_{i,t}$ denote the reported total personnel (TOTAL_PERSONNEL) at day t, and $C_{i,t}$ denote the cumulative cost estimation to date (EST_IM_COST_TO_DATE). For two pieces of data samples from the same incident occurring on consecutive dates, the targets can be obtained by:

$$\begin{aligned} &\text{Next-Day Personnel: } p_{i,t+1}, \\ &\text{Next-Day Cost: } c_{i,t+1} = C_{i,t+1} - C_{i,t} \end{aligned} \tag{3}$$

where the next-day personnel is the actual value and is not differenced, next-day cost is the one-day increment in the reported cumulative cost estimation and should be differenced. Cost supervision requires two temporally auditable cumulative-cost endpoints and an increment within [0,50 million] USD. Missing endpoints are not imputed, and negative revisions are excluded rather than clipped to zero. Human audit is provided to verify field semantics, provenance, anomalous records, and source linkage. The target values, validation status, construction version, and source-report identifiers are retained for audit.

3 Benchmark Evaluation

3.1 Benchmark Tasks

Task A: Wildfire Lifecycle State Understanding. This task evaluates whether a model can identify the current lifecycle state of a wildfire incident using the historical fire response records. For incident i on day t, the input $\mathbf{x}_{i,t}$ comprises the current fire response observations and historical available Fire-Day records of the same incident through day t. The model predicts one of these lifecycle phases: initial attack, rapid escalation, extended attack, containment, or mop-up/monitoring.

Task B: Next-Day Operational-Response Forecasting. Given the same temporally valid input $\mathbf{x}_{i,t}$, Task B defines two independent regression targets: forecasting the required personnel ($p_{i,t+1}$) for the next day, and forecasting the cost ($c_{i,t+1}$) for the next day. The two targets may be trained and evaluated separately or jointly on the subset of FireResBench. In this study, the two targets are separately trained while train/val/test are conducted on the same dataset to ensure fair comparison.

3.2 Research Questions

We organize our evaluation around four research questions. **RQ1.** ***Overall benchmark performance***: How do representative learning paradigms perform on the two FireResBench tasks? **RQ2.** ***Difficulty and failure modes:*** Where do model errors concentrate across wildfire lifecycle phases, phase transitions, and incident-level response trajectories? **RQ3.** ***Feature contribution and historical context:*** Which feature groups are most informative for the two benchmark tasks? How does the historical context length affect model performance? **RQ4.** ***Temporal generalization and missingness robustness:*** Do the evaluated models generalize to future-year incidents and remain reliable under incomplete observations? Experimental settings are in Appendix B.1 due to limited space.

3.3 Setup and Evaluation Protocols

Baseline Methods. ***For Task A***, XGBoost (Chen & Guestrin, 2016) provides a tree-based reference using aggregated historical features, while FT-Transformer (Gorishniy et al., 2021) models tabular representations through feature-wise attention. TCN (Bai et al., 2018) captures temporal dependencies with dilated convolutions, and GRU-D (Che et al., 2018) models feature missingness and elapsed time through learned decay. TFT (Lim et al., 2021) integrates static context and time-varying covariates. TimesNet (Wu et al., 2023) represents multi-period variation in a two-dimensional space.

ShapeFormer (Le et al., 2024) learns discriminative shapelet-aware representations. TSLANet (Eldele et al., 2024) combines adaptive spectral filtering with convolutional modeling. ***For Task B***, representative baselines spanning statistical and learning-based paradigms are evaluated. Historical Average (HA) provides a simple statistical reference. XGBoost and FT-Transformer serve as tree-based and attention-based tabular regressors, respectively. GRU-D and TFT model temporal histories with recurrent decay and attention. TimeXer (Wang et al., 2024) distinguishes target histories from exogenous evidence; TimeMixer++ (Wang et al., 2025) captures patterns across temporal scales and frequency resolutions. S4M (Peng et al., 2025) performs missing-aware forecasting through structured state-space modeling.

Evaluation metrics. ***For Task A***, Weighted-F1 is the primary metric under class imbalance. Balanced accuracy measures average recall across the five phases, while mean Phase Distance measures the ordinal distance between predicted and ground-truth phases. Transition accuracy evaluates classification performance on days when the lifecycle phase changes, expected calibration error (ECE) measures the alignment between prediction confidence and empirical accuracy. ***For Task B***, next-day personnel and daily cost are evaluated independently. Personnel MAE and daily-cost MAE are the primary metrics in their original units; RMSE emphasizes large forecasting errors, while NMAE normalizes the total absolute error by the total ground-truth magnitude for each target. We additionally report peak-personnel, peak-timing, and cumulative-cost MAE at the incident-trajectory level.

Table 3: Task A wildfire lifecycle-state understanding performance.

Method	Overall Classification		Ordinal Error	Boundary & Calibration	
	Weighted-F1↑	Balanced Acc.↑	Phase Dist.↓	Transition Acc.↑	ECE↓
FT-Transformer	**0.841**±0.2%	**0.825**±0.2%	**0.174**±0.2%	**0.847**±0.1%	0.036±0.3%
TSLANet	0.805±0.3%	0.814±0.1%	0.242±0.2%	0.801±1.5%	0.050±0.2%
GRU-D	0.802±1.1%	0.815±0.7%	0.248±1.2%	0.790±0.2%	0.067±1.2%
TFT	0.824±0.4%	0.819±0.3%	0.205±0.6%	0.813±0.9%	0.042±0.9%
ShapeFormer	0.792±0.3%	0.802±0.9%	0.266±0.5%	0.721±1.6%	0.073±1.8%
XGBoost	0.787±0.8%	0.793±0.4%	0.267±1.4%	0.752±1.7%	0.061±1.0%
TimesNet	0.783±0.5%	0.786±1.3%	0.276±0.3%	0.709±1.8%	**0.023**±0.2%
TCN	0.771±0.9%	0.784±0.7%	0.283±1.1%	0.700±1.4%	0.058±1.0%

*Note: All results are reported as mean ± std. **Bold** and underlined values indicate the best and 2nd-best performance, respectively.*

Table 4: Task B next-day operational-response forecasting performance.

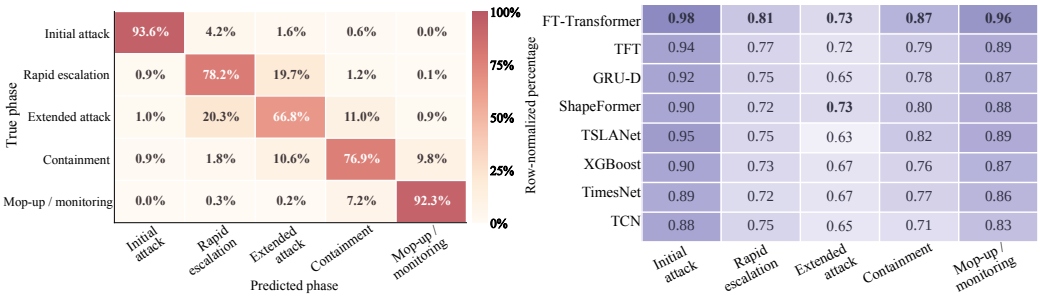
Model	Personnel Forecasting			Daily-Cost Forecasting (K/USD)		
	MAE↓	NMAE↓	RMSE↓	MAE↓	NMAE↓	RMSE↓
HA	65.13±0.00	0.4485±0.0000	83.36±0.00	193.5K±0.0K	0.8643±0.0000	577.3K±0.0K
FT-Transformer	**22.76**±0.20	**0.1802**±0.0014	**45.93**±0.63	**82.6K**±1.1K	**0.5385**±0.0022	**242.3K**±4.3K
TFT	25.69±0.27	0.1936±0.0019	55.75±1.22	91.6K±2.4K	0.5762±0.0245	456.7K±5.1K
TimeMixer++	30.68±0.25	0.2278±0.0023	64.95±0.96	114.5K±2.7K	0.5910±0.0238	479.3K±5.4K
XGBoost	27.42±0.28	0.1958±0.0023	56.47±1.21	93.5K±2.8K	0.5861±0.0184	436.3K±5.7K
S4M	34.72±0.32	0.2473±0.0029	70.59±1.41	120.4K±5.1K	0.6277±0.0325	507.5K±6.3K
TimeXer	32.10±0.28	0.2263±0.0021	62.99±1.25	118.4K±3.3K	0.6120±0.0217	516.2K±6.5K
GRU-D	29.14±0.33	0.2125±0.0026	58.80±1.39	102.9K±2.7K	0.6067±0.0019	493.8K±6.4K

*Note: All results are reported as mean ± std. **Bold** and underlined values indicate the best and 2nd-best performance, respectively.*

3.4 Experimental Results and Analysis

RQ1: Overall Benchmark Performance. ***Lifecycle understanding.*** As shown in Tables 3, FT-Transformer provides the strongest results for lifecycle-state understanding, attaining a Weighted-F1 of 0.841 while also leading balanced accuracy, phase distance, and transition accuracy. TFT ranks second on each of these measures. Agreement across class-frequency-aware, class-balanced, ordinal, and transition-specific criteria indicates that this advantage is not an artifact of a single metric. TimesNet produces the lowest calibration error despite weaker classification scores, showing that confidence calibration and discriminative performance remain distinct challenges on Task A.

Operational-response forecasting. As shown in Table 4, FT-Transformer leads the targets in Task B, yielding a personnel MAE of 22.76 and a daily-cost MAE of $82.6K while substantially improv-



(a) (b)

Figure 2: Task A diagnostics: (a) Row-normalized confusion matrix. (b) Per-class F1 scores.

ing over the historical-average reference. Its advantage extends from MAE to RMSE, indicating that it is not limited to typical Fire-Days but also reduces the influence of larger misses. The remaining baselines do not preserve a single ordering across personnel and cost: TFT is generally the closest competitor, whereas XGBoost is second best for daily-cost RMSE. More broadly, specialized temporal forecasters do not automatically outperform tabular attention or tree-based models. Given the short incident histories in FireResBench, these results suggest that resolving interactions among heterogeneous current and recent Fire-Day variables is at least as important as imposing a more elaborate long-range temporal architecture.

RQ2: Difficulty and Failure Modes. ***Lifecycle phases and boundaries.*** Figure 2 localizes the remaining errors in Task A. Initial attack and mop-up/monitoring are recognized reliably across models, whereas the intermediate phases, especially extended attack, are consistently harder. For FT-Transformer, the per-class F1 decreases from 0.98 for initial attack to 0.73 for extended attack. The confusion matrix shows that errors largely remain within neighboring phases, most visibly between rapid escalation and extended attack and between extended attack and containment. Models therefore recover the broad ordering of wildfire progression but struggle to determine precisely when an incident crosses an operational boundary. The per-class results reinforce this interpretation: despite their different architectures, the baselines share a similar weakness in the less stable middle of the lifecycle. Appendix B.2 further examines how different models perform near the lifecycle phase transition point: TimesNet and TSLANet deteriorate on day 0, while XGBoost, TCN, and GRU-D can improve when contemporaneous threshold cues become explicit. The difficulty is consequently not on the transition day itself, but the heterogeneous and partially observed evidence by which different transitions become detectable.

Table 5: Incident-level recovery of wildfire response trajectories in Task B.

Model	Peak Personnel MAE↓	Peak Timing MAE (days)↓	Cumulative Cost MAE (million)↓
HA	66.819±0.000	1.530±0.000	0.729±0.000
FT-Transformer	**22.885±0.470**	**0.631±0.025**	**0.425±0.011**
TFT	25.916±0.350	0.763±0.024	0.442±0.015
GRU-D	31.082±0.649	1.320±0.026	0.608±0.027
XGBoost	29.031±0.475	1.206±0.020	0.584±0.019
TimeXer	33.764±0.563	1.246±0.018	0.629±0.035
TimeMixer++	31.258±0.412	1.307±0.032	0.592±0.024
S4M	29.512±0.327	1.069±0.018	0.564±0.027

Note: Peak-personnel MAE is reported in persons. All results are reported as mean ± std. ***Bold*** *and underlined values indicate the best and 2nd-best performance, respectively.*

Demand tails and incident trajectories. Figure 3 shows that forecasting difficulty is concentrated in the upper tail of operational demand. Errors remain comparatively small and tightly grouped across low- and moderate-demand Fire-Days, then rise rapidly and separate across models in the highest strata. The results indicate that none of the evaluated paradigms resolves the rare-demand regime. The trajectory-level results in Table 5 lead to a compatible but more operationally specific conclusion. FT-Transformer and TFT remain the two strongest methods for recovering peak personnel, peak timing, and cumulative cost, so their day-level advantage is not merely an averaging effect. Changes are seen in the middle of the ranking, while S4M is more competitive on peak timing and cumulative cost than on pointwise forecasting. The shift reveals that forecasting next-day demand magnitude and reconstructing incident operational trajectories favor different model behaviors.

RQ3: Feature Contribution and Temporal Context. The ablations in Appendix B.4 reveal task-specific evidence dependencies. In Task A, removing fire-condition features produces the largest degradation, lowering Weighted-F1 by 0.198-0.254, While removing FIRMS data from the complete representation has more modest effect. Task B relies primarily on operational history, remove them increases personnel NMAE by 0.210-0.253 and cost NMAE by 0.048-0.116. Thus, the shared Fire-Day representation supports different mechanisms: fire conditions distinguish lifecycle states, while recent operational records anchor short-term resource demand. All primary results retain the complete feature set. Appendix B.5 shows that increasing the observed Fire-Days H from 1 to 4 or 6 improves Task A performance and reduces Task B errors, but gains saturate near $H = 6$. At $H = 8$, model performance on Task A generally plateaus, while cost MAE increases across five representative models, suggesting that longer time windows may incorporate outdated information rather than provide additional evidence. These results support six Fire-Days as a practical default and favor selecting temporally relevant evidence over assuming that longer histories are always beneficial.

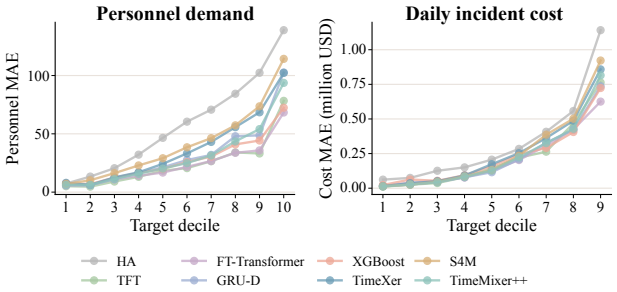


Figure 3: Task B next-day forecasting MAE stratified by personnel and daily-cost target deciles.

RQ4: Temporal Generalization and Missingness Robustness. The Temporal-OOD results in Appendix B.6 reveal model- and target-dependent transfer. Changes span improvements to clear degradations, showing that no baseline model is globally stable. Because the IID and Temporal-OOD protocols evaluate different incident cohorts, negative changes may partly reflect cohort difficulty rather than stronger extrapolation. Thus, random incident-disjoint evaluation alone cannot fully characterize temporal generalization. Appendix B.7 complements this finding with fixed-cohort missingness tests. Random field masking causes dose-dependent degradation, with the Weighted-F1 of Task A falling to 0.396-0.533 at 50% masking. Randomly deleting complete historical Fire-Days is generally less harmful and more model-dependent, indicating that within-day feature loss and missing temporal observations pose distinct challenges. Removing entire feature groups further shows that operational evidence is crucial for personnel and cost forecasting, whereas Task A relies on a model-specific mixture of cues.

4 Conclusions

We present FireResBench, an unified event-centered benchmark for studying wildfire lifecycle evolution and operational response from temporally organized incident records. It comprises 5,728 wildfire incidents across the CONUS United States between 2017 and 2020. FireResBench provides two complementary tasks: wildfire lifecycle-state understanding with rule-guided, domain-scholar-verified annotations, and the forecasting of assigned personnel and daily cost using temporally valid targets and incident-disjoint splits. Experiments across representative machine learning, tabular-based, and temporal deep learning paradigms establish strong reference performance for both benchmark tasks, while showing that no single paradigm is uniformly reliable across all evaluation dimensions. Beyond aggregate task metrics, our evaluation examines phase-specific behavior, lifecycle transitions, predictive calibration, resource-demand strata, and complete incident trajectories. We found that errors concentrate around adjacent lifecycle boundaries, with the intermediate extended-attack phase remaining notably harder to distinguish. Forecasting errors rise sharply for high-demand Fire-Days, and pointwise rankings do not fully determine recovery of personnel peaks, peak timing, and cost. These findings position FireResBench as both a reproducible evaluation resource and a testbed for boundary- and trajectory-aware wildfire response modeling. Future work will expand FireResBench into a broader foundation for wildfire-response intelligence by extending its temporal and geographic coverage, modeling rare and extreme operational demands, and enabling trustworthy human-AI collaboration in complex emergency-management settings.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and language editing to improve clarity and readability. In addition, LLM agents were used in the first stage of FireAgentBench's two-stage annotation process to implement and execute the rigorous autolabeling rules designed by the scholar team and to perform automated consistency checks. The agents did not design the annotation rules or determine the final labels. The provisional labels were reviewed and, when necessary, corrected by domain scholars. Apart from these uses, generative AI tools did not contribute to the study design, experimental evaluation, or interpretation of the results. The authors reviewed all AI-assisted outputs and take responsibility for the dataset annotations and the final content.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. doi: 10.48550/arXiv.1803.01271.

Sunniva Bloem, Alison C. Cullen, John T. Abatzoglou, Linda O. Mearns, and Erin Belval. Drivers of international fire management personnel deployed to the united states. *International Journal of Wildland Fire*, 33(3), 2024. doi: 10.1071/WF23093.

Chantelle Burton, Seppe Lampe, Douglas I. Kelley, Wim Thiery, Stijn Hantson, Nikos Christidis, Lukas Gudmundsson, Matthew Forrest, Eleanor Burke, Jinfeng Chang, Huilin Huang, Akihiko Ito, Sian Kou-Giesbrecht, Gitta Lasslop, Wei Li, Lars Nieradzik, Fang Li, Yang Chen, James Randerson, Christopher P. O. Reyer, and Matthias Mengel. Global burned area increasingly explained by climate change. *Nature Climate Change*, 14(11):1186–1192, 2024. doi: 10.1038/s41558-024-02140-w.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

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*, pp. 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.

Alison C. Cullen, Brian R. Goldgeier, Erin Belval, and John T. Abatzoglou. Characterising ignition precursors associated with high levels of deployment of wildland fire personnel. *International Journal of Wildland Fire*, 33(8), 2024. doi: 10.1071/WF23182.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, and Xiaoli Li. TSLANet: Rethinking transformers for time series representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12409–12428. PMLR, 2024.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74515–74529. Curran Associates, Inc., 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Bin-Wei Hsu. Integrating explainable ai and multi-objective optimization for wildfire resource allocation under fixed staffing constraints: A decision-oriented framework. *Scientific Reports*, 16(1), 2026. doi: 10.1038/s41598-026-50836-0.

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.

Xuan-May Le, Ling Luo, Uwe Aickelin, and Minh-Tuan Tran. ShapeFormer: Shapelet transformer for multivariate time series classification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2024. doi: 10.1145/3637528.3671862.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 1428–1442. Curran Associates, Inc., 2024. doi: 10.52202/079017-0045.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Munich Re. Climate change presses on: Devastating wildfires and intense thunderstorms exacerbate losses for insurers, January 2026. Natural disaster figures for 2025. Accessed: 2026-07-17.

Dung Nguyen, Erin J. Belval, Yu Wei, Karen C. Short, and David E. Calkin. Dataset of united states incident management situation reports from 2007 to 2021. *Scientific Data*, 11(1), 2024. doi: 10.1038/s41597-023-02876-8.

Dung Nguyen, Erin J. Belval, Yu Wei, and David E. Calkin. A historical analysis of factors driving the daily prioritization of wildland fires in california. *International Journal of Disaster Risk Reduction*, 132:105967, 2026. doi: 10.1016/j.ijdrr.2025.105967.

Jing Peng, Meiqi Yang, Qiong Zhang, and Xiaoxiao Li. S4M: S4 for multivariate time series forecasting with missing values. In *International Conference on Learning Representations*, pp. 49369–49395, 2025.

David Rossi, Olli-Pekka Kuusela, and Christopher Dunn. A microeconometric analysis of wildfire suppression decisions in the western united states. *Ecological Economics*, 200:107525, 2022. doi: 10.1016/j.ecolecon.2022.107525.

Wilfrid Schroeder, Patricia Oliva, Louis Giglio, and Ivan A. Csiszar. The new VIIRS 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143:85–96, 2014. doi: 10.1016/j.rse.2013.12.008.

Franz Schug, Avi Bar-Massada, Amanda R. Carlson, Heather Cox, Todd J. Hawbaker, David Helmers, Patrick Hostert, Dominik Kaim, Neda K. Kasraee, Sebastián Martinuzzi, Miranda H. Mockrin, Kira A. Pfoch, and Volker C. Radeloff. The global wildland–urban interface. *Nature*, 621(7977):94–99, 2023. doi: 10.1038/s41586-023-06320-0.

Lise A. St. Denis, Karen C. Short, Kathryn McConnell, Maxwell C. Cook, Nathan P. Mietkiewicz, Mollie Buckland, and Jennifer K. Balch. All-hazards dataset mined from the us national incident management system 1999–2020. *Scientific Data*, 10(1), 2023. doi: 10.1038/s41597-023-01955-0.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0015.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Chen Xu, Yao Xie, Daniel A. Zuniga Vazquez, Rui Yao, and Feng Qiu. Spatio-temporal wildfire prediction using multi-modal data. *IEEE Journal on Selected Areas in Information Theory*, 4: 302–313, 2023. doi: 10.1109/JSAIT.2023.3276054.

Kevin Young, Erin Belval, Karin Riley, and Peng Gao. Predicting daily firefighting personnel deployment trends in the western united states. *Journal of Environmental Management*, 370:122705, 2024. doi: 10.1016/j.jenvman.2024.122705.

Nan Zhou, Huandong Wang, Jiahao Li, Han Li, Yali Song, Qiuhua Wang, Yong Li, and Xinlei Chen. Firesentry: A multi-modal spatio-temporal benchmark dataset for fine-grained wildfire spread forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2935–2946, 2026. doi: 10.1145/3770854.3785696.

A WILDFIRE LIFECYCLE STATE ANNOTATION GUIDELINE

Lifecycle-Supervision Pipeline Overview. Figure 4 illustrates the construction of Task A supervision, from temporally bounded Fire-Day evidence and deterministic pre-annotation to domain-scholar review and final label freezing. We develop a two-stage labeling pipeline. ***In Stage I***, domain scholars freeze the phase definitions, evidence thresholds, rule priority, and temporal constraints, which are then implemented as deterministic autolabeling programs executed by coding agent. For each Fire-Day, the program uses only $\mathbf{o}_{i,t}$ and $\mathcal{H}_{i,<t}$ to derive evidence from incident age, area and growth, containment, FIRMS activity, reported fire behavior, and personnel and cost histories. It assigns one provisional state from *Initial Attack*, *Rapid Escalation*, *Extended Attack*, *Containment*, and *Mop-up/Monitoring*, while retaining the activated evidence, raw rule output, confidence, and transition status. A temporal consistency layer permits stable or forward lifecycle progression and allows a return to rapid escalation only when renewed escalation evidence is observed.

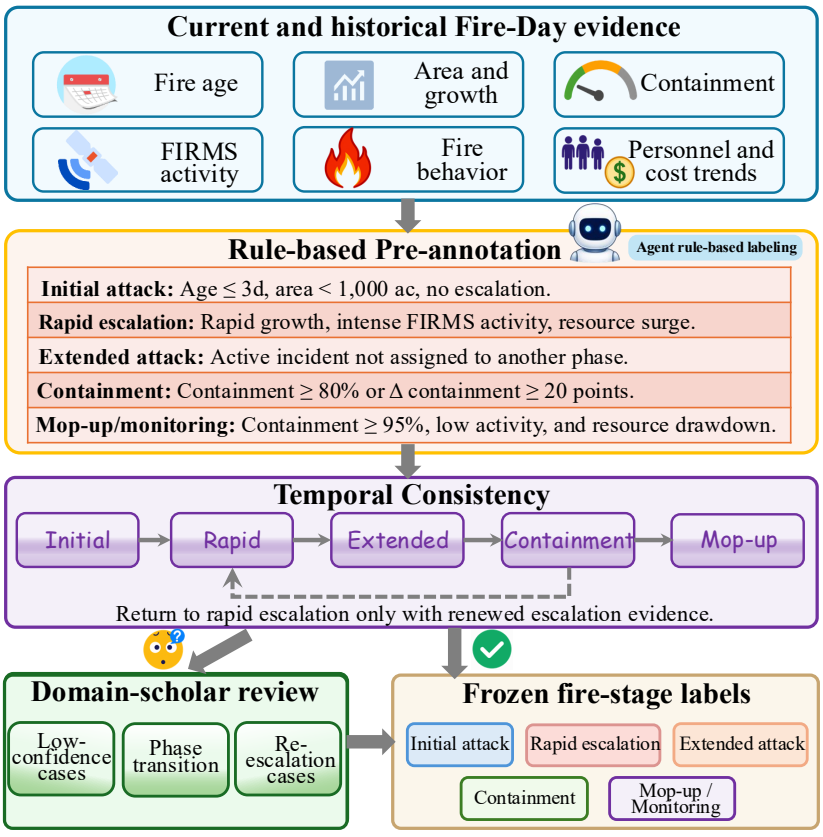


Figure 4: Task A lifecycle-supervision pipeline.

In Stage II, domain scholars review all initial-attack cases, low-confidence or boundary cases, lifecycle transitions, re-escalations, records with missing or conflicting core evidence, and a reproducible phase-stratified audit sample. Each evidence packet truncates the incident history at day t and excludes future observations, response targets, model outputs, and final incident outcomes. Reviewers may accept, correct, or defer a provisional label; corrections and deferrals receive independent secondary review, and disagreements are resolved by an adjudicating expert. Corrected incident sequences are re-validated, unresolved cases are excluded from supervised evaluation, and the final annotations are frozen before data splitting and model development. The resulting label $s_{i,t}$ is released as the curated ground-truth lifecycle state together with its rule and annotation.

Stage I: Rule-Based Lifecycle Pre-Annotation

Evidence boundary. For each incident, Fire-Day records are sorted chronologically. The pre-annotation procedure uses only the observations available on the current Fire-Day and earlier records from the same incident. It never uses next-day response targets, subsequent fire development, final incident size, or final containment outcome. Changes are computed relative to the most recent preceding Fire-Day, denoted by t^-.

Derived evidence signals. Let A_t denote reported area, ΔA_t new area, q_t containment, P_t personnel, and $r_t = \Delta A_t / \max(A_{t^-}, 1)$ the relative growth rate. The following signals are computed before assigning a provisional phase:

- *Rapid growth* is active when a previous Fire-Day exists and either (i) $\Delta A_t \geq 500$ acres and $r_t \geq 0.20$, (ii) $\Delta A_t \geq 2{,}000$ acres, or (iii) the absolute fire-spread-rate indicator is at least 0.20.
- *Intense satellite activity* is active when the 5-km FIRMS count is at least 30 or the summed fire radiative power is at least 500. A lower active-FIRMS threshold, defined as at least 10 detections or summed FRP of at least 100, is used when identifying renewed escalation.
- *High-severity fire behavior* is active when the current report contains indicators such as crowning, extreme behavior, running, spotting, torching, or wind-driven fire.
- *Resource surge* is active when personnel increases by at least 25, personnel grows by at least 50%, average daily cost reaches \$250K, or the reported cost increment reaches \$250K.
- *Low activity* requires $\Delta A_t \leq 25$ acres, no more than one 5-km FIRMS detection, summed FRP no greater than 5, and no high-severity fire-behavior signal.
- *Resource drawdown* is active when personnel decreases by at least 15 or by at least 20% relative to the previous available Fire-Day.

Provisional phase assignment. The rules are evaluated in the following priority order so that late-stage evidence takes precedence over weaker early-stage cues:

1. *Mop-up/monitoring*: assign this phase when $q_t \geq 95\%$, current activity is low, and at least one supporting late-stage condition holds: resource drawdown, $P_t \leq 20$, or incident age of at least seven days.
2. *Containment*: assign this phase when $q_t \geq 80\%$ or containment has increased by at least 20 percentage points since t^-, provided that neither rapid growth nor intense FIRMS activity is present.
3. *Rapid escalation*: assign this phase when $q_t < 80\%$ and at least one escalation signal is present: rapid growth, intense FIRMS activity, high-severity fire behavior, or resource surge.
4. *Initial attack*: assign this phase when incident age is at most three days, $A_t \leq 1{,}000$ acres, $q_t < 80\%$, and none of the escalation signals is active.
5. *Extended attack*: assign this phase to the remaining operationally active Fire-Days that do not satisfy the preceding initial-attack, escalation, containment, or mop-up conditions.

Temporal constraints. The normal lifecycle order is:

$$\text{Initial attack} \rightarrow \text{Rapid escalation} \rightarrow \text{Extended attack} \rightarrow \text{Containment} \rightarrow \text{Mop-up/monitoring}.$$

A label may remain unchanged or advance by one phase between consecutive available Fire-Days. If the raw rule output skips multiple phases, it is advanced by only one phase and marked as a smoothed transition. A backward transition is otherwise held at the preceding phase. The sole exception is a return to rapid escalation after prior containment of at least 80%, which is permitted when renewed expansion, active FIRMS detections, high-severity behavior, or a resource surge is observed.

Confidence and audit fields. Each output record retains the provisional label, a rule-derived confidence score, and a transition flag. Initial attack receives confidence 0.82, extended attack 0.68, and mop-up/monitoring 0.90. Rapid-escalation confidence ranges from 0.81 to 0.91 according to the number of active escalation signals, while containment confidence ranges from 0.82 to 0.86. Smoothed forward transitions and backward-held labels are capped at 0.65 and 0.60, respectively; confirmed re-escalations receive confidence of at least 0.80. These scores are routing indicators for expert review, not calibrated probabilities.

Stage II: Human Review, Correction, and Label Freezing

Review principle. Human review verifies whether a provisional label is supported by the evidence available through Fire-Day t. Reviewers must not inspect model predictions, next-day response labels, subsequent

fire development, or the final incident outcome. The objective is to correct unsupported annotations, not to force balanced class frequencies or retrospectively optimize benchmark performance.

The review and correction procedure consists of the following steps:

1. *Construct the review queue.* The queue contains all initial-attack annotations; all records with confidence no greater than 0.68; all smoothed, backward-held, re-escalation, and phase-changing transitions; and records with missing or conflicting core evidence. Boundary cases are also included when containment is within five percentage points of the 80% or 95% thresholds, or when growth, FIRMS, personnel, or cost signals fall within 10% of a rule threshold. From the remaining records, a fixed 10% audit sample is drawn while stratifying by provisional phase and incident.
2. *Prepare the evidence packet.* For each queued Fire-Day, the reviewer receives the current and preceding Fire-Day observations, original units, missingness and match-status fields, source-report identifiers, triggered rule clauses, provisional label, confidence score, and transition flag. The incident timeline is truncated at day t so that later observations cannot influence the decision.
3. *Apply consistent phase distinctions.* Initial attack is retained only for an early, comparatively small incident without strong escalation evidence. Rapid escalation requires evidence of an active increase in physical fire activity, operational intensity, or resource demand, whereas extended attack represents sustained activity without a current escalation spike. Containment requires clear control progress without dominant renewed growth, while mop-up/monitoring requires near-complete containment together with low activity and late-stage or drawdown evidence.
4. *Record the review decision.* The reviewer selects one of three actions: *accept* the provisional label; *correct* it to another lifecycle phase with a documented evidence rationale; or *defer* the case when the available evidence is insufficient or internally contradictory. Corrections must reference observable Fire-Day evidence and use one of the predefined reason codes: `threshold_boundary`, `conflicting_evidence`, `missing_evidence`, `temporal_inconsistency`, `source_error`, or `other`. Reviewers must not infer unreported conditions from later outcomes.
5. *Independently verify corrections.* Every proposed correction and deferred case is independently reviewed by a second expert. The second expert additionally reviews a random 20% subset of accepted cases. Disagreements are resolved by an adjudicating expert using the same evidence boundary and phase definitions. Agreement before adjudication is reported using raw agreement and Cohen's κ.
6. *Revalidate the incident sequence.* After correction, the complete label sequence is checked for skipped phases, unsupported backward transitions, and inconsistencies around the corrected Fire-Day. The temporal rules do not automatically overwrite an expert decision; instead, any resulting inconsistency is returned for adjudication together with the adjacent Fire-Days.
7. *Freeze and version the annotations.* The final label, original provisional label, review action, reason code, and adjudication outcome are retained in an immutable annotation log. Unresolved cases are excluded from supervised evaluation rather than assigned a speculative label. Annotation rules and final labels are frozen before train-validation-test splitting and model development. Each release reports the reviewed, corrected, excluded, and adjudicated counts, together with agreement statistics, rule version, and file checksum.

B Supplementary Experimental Results

This appendix expands the diagnostic experiments summarized in RQ2-RQ4. Unless stated otherwise, all comparisons use the same task definitions, target construction, eligibility rules, and incident-level separation described in the main paper. The purpose of these analyses is not to introduce additional benchmark tasks, but to determine when the aggregate results hold, where they fail, and which properties of the Fire-Day representation account for the observed behavior.

B.1 Experimental Settings

All experiments use incident-disjoint splits so that Fire-Day records from the same wildfire never appear in both training and evaluation sets. Unless otherwise specified, we use a fixed 80%/10%/10% for train/validation/test set split, and provide each model with the current Fire-Day together with up to 6-day historical observed Fire-Day features. Each experiment is repeated 10 runs under the same data split, and we report the mean and standard deviation. All the deep learning baseline models are implemented in PyTorch 2.11.0 with CUDA 12.8 and trained on an NVIDIA RTX PRO 6000

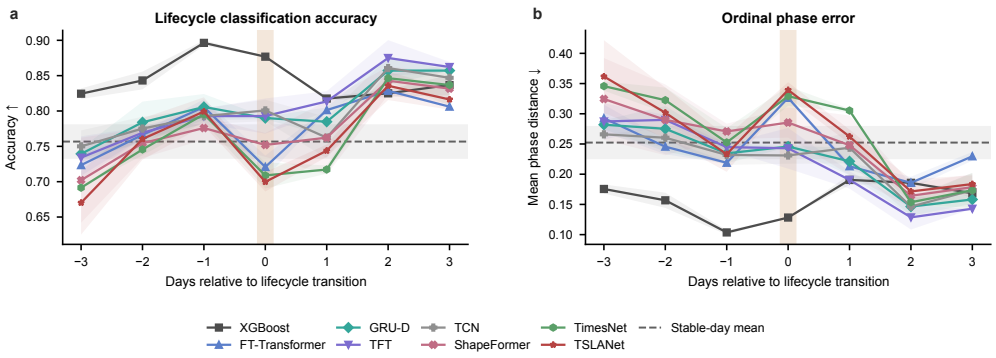


Figure 5: Task A performance around expert-verified lifecycle transitions. Panel (a) reports classification accuracy and panel (b) reports mean ordinal phase distance. Day 0 is the first observed Fire-Day carrying the changed lifecycle label. Curves and shaded regions summarize repeated evaluations; the dashed line and gray band denote performance on stable Fire-Days outside the transition windows.

Blackwell Max-Q GPU with 96GB memory. ***For Task A***, all neural models are trained as five-class classifiers using class-weighted cross-entropy, with method-specific auxiliary losses retained where applicable, and selected by validation weighted-F1. ***For Task B***, Both targets are trained and evaluated separately on the same Fire-Day pairs with valid personnel and cost targets, using identical incident-disjoint splits. Learning-based models are trained with standardized log-transformed supervision, and all reported errors are computed after inverse transformation to the original units. Open-source baseline implementations are adapted only at the data-loading, input-projection, and prediction-head layers when necessary, while preserving their core backbone designs to support reproducibility.

B.2 Lifecycle-Transition Diagnostics

We align every expert-verified phase change to the first observed Fire-Day carrying the new label and define that observation as day 0. Performance is then evaluated from three observed Fire-Days before to three observed Fire-Days after the transition. The comparison includes 881 transitions. From day -3 through day $+3$, the respective support is 94, 169, 386, 881, 244, 140, and 98 Fire-Days, while the stable reference contains 1,456 Fire-Days from 314 incidents outside these windows. Because incident reports are irregular, offsets denote preceding or following *observed* Fire-Days and are never filled by interpolation. This choice preserves the information pattern that a deployed model would encounter, while the support imbalance requires cautious interpretation of the outer offsets.

Findings. Figure 5 shows that transition behavior is systematic but not uniform across models. TimesNet and TSLANet lose accuracy and incur larger phase distance on day 0 relative to stable Fire-Days, consistent with a lag in resolving newly changed states. FT-Transformer also deteriorates modestly, from 0.748 stable-day accuracy to 0.721 at the boundary, with phase distance increasing from 0.262 to 0.326. By contrast, XGBoost, TCN, and GRU-D improve on day 0; XGBoost reaches 0.877 accuracy and 0.128 phase distance, compared with 0.805 and 0.196 on stable days. These models can exploit sharp contemporaneous cues–for example a threshold-crossing change in containment, fire growth, or operational intensity–when such cues are present. Several temporal models recover strongly after the boundary, particularly by day $+2$, once the new regime is supported by more than one observation.

The result refines the phase-level confusion analysis in the main paper. Transition Fire-Days are not intrinsically and equally difficult. Difficulty depends on how the phase change is expressed in the available evidence and on the inductive bias of the model. A strong benchmark method should therefore do more than smooth the incident trajectory: it should detect abrupt, evidence-supported changes without becoming unstable when reports are sparse. Because the outer offsets contain fewer

incidents, we emphasize the day-0 pattern and the broad recovery trend rather than small point-to-point differences at -3 or $+3$.

B.3 Incident-Level Operational-Response Diagnostics

The main Task B metrics evaluate one-day-ahead predictions independently. Operational planning, however, also depends on whether a method reconstructs consequential properties of the complete response trajectory. We therefore aggregate test predictions within each incident and measure error in peak personnel, the date of the personnel peak, and cumulative incident cost. These measures are deliberately complementary: a model can achieve a low average next-day error while smoothing away the peak, shifting it in time, or accumulating a persistent cost bias.

Findings. Table 5 shows that FT-Transformer remains strongest on all three incident-level measures, followed by TFT. FT-Transformer reduces the HA peak-personnel error from 66.819 to 22.885, peak-timing error from 1.530 to 0.631 days, and cumulative-cost error from 0.729 to 0.425 million dollars. Its day-level advantage is therefore not merely an averaging effect: it transfers to operationally meaningful properties of the incident trajectory. The relative ranking below the leading pair changes more noticeably. S4M, for example, is not among the strongest pointwise forecasters in Table 4, yet its peak-timing and cumulative-cost errors are more competitive than those of several sequence baselines. Conversely, low next-day error does not guarantee equally strong recovery of every trajectory functional. This ranking shift supports reporting both levels of evaluation and discourages selecting models solely from an aggregate Fire-Day metric.

B.4 Evidence-Source Ablation

We evaluate representative baselines under current-day, history-only, single-source, combined-source, and leave-one-group-out input configurations. Task A loss is the full-input Weighted-F1 minus the reduced-input Weighted-F1; Task B loss is the reduced-input NMAE minus the full-input NMAE. Positive values therefore always indicate worse performance than the complete representation. The lifecycle labels are supervision for Task A and are never included among the input features. These experiments are diagnostic: every primary benchmark result continues to use the full released feature set.

Task A. Figure 6a identifies fire-condition features as the principal source of lifecycle discrimination. Removing them lowers Weighted-F1 by 0.198-0.254 across XGBoost, FT-Transformer, GRU-D, and TFT. A FIRMS-only representation performs much worse than the complete input, with losses of 0.280-0.396, showing that active-fire detections do not by themselves define an operational lifecycle state. At the same time, omitting FIRMS from the otherwise complete evidence causes only a 0.020-0.029 loss. The two observations are compatible: satellite activity corroborates growth and persistence but is largely useful in combination with incident-report variables such as containment, fire status, and operational change. Response history and missingness flags contribute little in isolation to the aggregate Task A score, although they may remain useful for particular incidents or transition types.

Task B. The forecasting panels exhibit a different hierarchy. Removing response history increases personnel NMAE by 0.210-0.253 for every model and cost NMAE by 0.048-0.116. Historical personnel and cost levels provide a strong local anchor because tomorrow's demand is conditionally persistent, while fire-condition and FIRMS groups mainly refine that anchor. History-only inputs are nevertheless insufficient, most visibly for FT-Transformer, whose personnel NMAE increases by 0.389 and cost NMAE by 0.236. Thus, the operational trajectory carries most of the short-horizon signal, but contemporaneous incident conditions are needed to recognize deviations from persistence. A few cells are slightly negative, indicating that a reduced configuration occasionally improves a particular model. Their small and inconsistent magnitude suggests redundancy, optimization sensitivity, or finite-sample variation rather than a general case for deleting those features.

Taken together, the three panels validate the benchmark's multi-purpose design. A single standardized Fire-Day interface supports both tasks, yet the tasks interrogate different evidence pathways: state understanding depends mainly on physical and status descriptors, whereas response forecasting

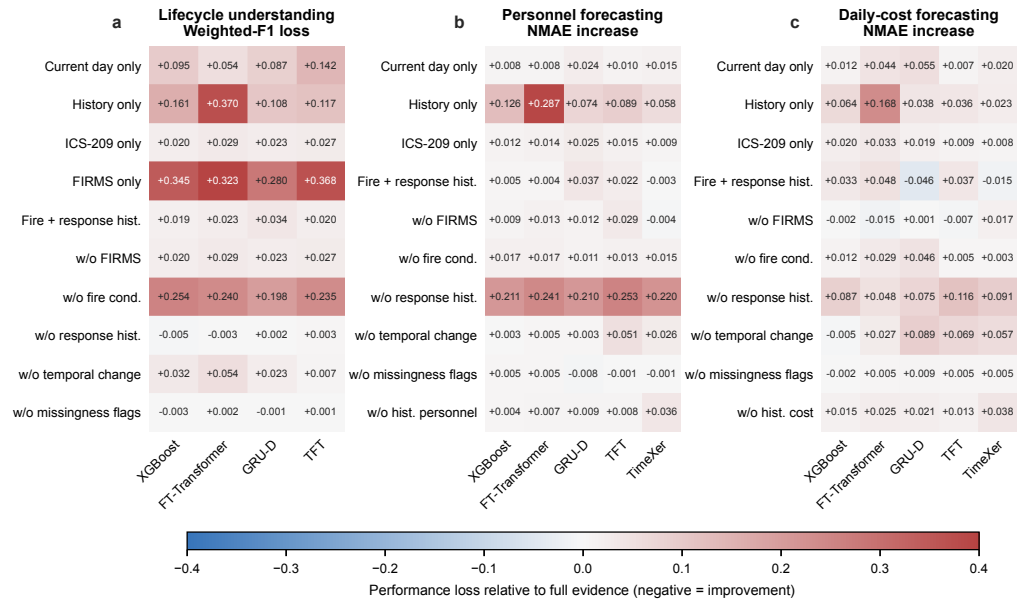


Figure 6: Contribution of dataset feature groups to FireResBench. Panel (a) reports loss in Task A Weighted-F1 relative to full evidence; panels (b) and (c) report the increase in NMAE for personnel and daily-cost forecasting. Positive values denote degradation, and negative values denote improvement under the corresponding reduced configuration.

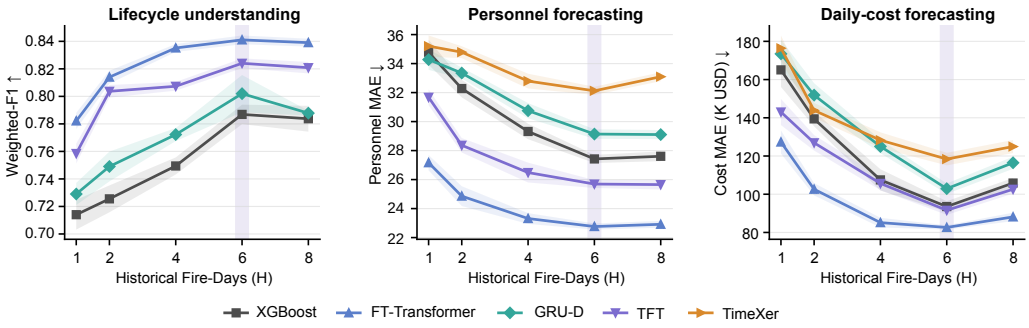


Figure 7: Sensitivity to historical context length. Panel (a) reports Task A Weighted-F1; panels (b) and (c) report Task B personnel and daily-cost MAE.

depends mainly on the evolving operational record. This separation also explains why performance on one task should not be used as a proxy for the other.

B.5 Historical-Window Sensitivity

We vary the number of observed historical Fire-Days over $H \in \{1, 2, 4, 6, 8\}$ while retaining the same eligibility and split logic. Because reports do not arrive at a perfectly fixed interval, H counts earlier observed Fire-Days rather than consecutive calendar days. The primary benchmark uses $H = 6$.

Findings. Figure 7 shows a consistent benefit from moving beyond a single observation. FT-Transformer's Task A Weighted-F1 rises from 0.782 at $H = 1$ to 0.841 at $H = 6$, and the other representative models show the same broad improvement. For Task B, all five models reduce personnel and cost MAE as the window grows to six observations. FT-Transformer, for example, lowers personnel MAE from 27.16 to 22.76 and cost MAE from \$127.3K to \$82.6K. These gains confirm

Table 6: Incident and Fire-Day counts in the Temporal-OOD protocol.

Task	Split	Incidents	Fire-Days
Task A	Train (2017-2018)	3,404	19,559
Task A	Validation (2019)	1,015	4,670
Task A	Test (2020)	1,308	9,073
Task B	Train (2017-2018)	3,405	19,560
Task B	Validation (2019)	1,015	4,670
Task B	Test (2020)	1,308	9,073

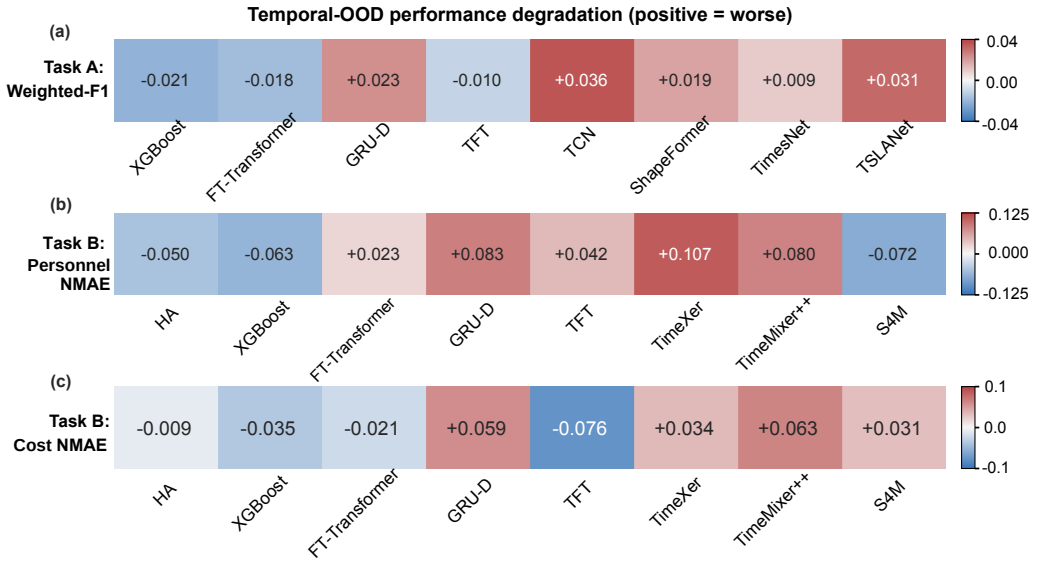


Figure 8: Performance change from Incident-IID to Temporal-OOD evaluation. Panel (a) reports Task A Weighted-F1 degradation. Panels (b) and (c) report Task B personnel and daily-cost NMAE degradation. Positive values indicate worse Temporal-OOD performance.

that a Fire-Day should be interpreted within its incident history rather than as an isolated tabular row.

The relationship is not monotonic beyond the default. At $H = 8$, Task A performance plateaus or decreases for most models, and cost MAE increases for every evaluated baseline relative to $H = 6$. TimeXer also begins to deteriorate in personnel forecasting. Longer histories may contain stale operational regimes, larger reporting gaps, and more missing fields, so extra observations can increase sequence length without adding current information. The saturation pattern is therefore itself a property of irregular emergency records, not merely a hyperparameter artifact. It motivates selective temporal aggregation and recency-aware architectures, and supports $H = 6$ as a balanced common context rather than claiming it is universally optimal for every incident.

B.6 Temporal-OOD Generalization

Temporal-OOD evaluation trains on 2017-2018 incidents, validates on 2019 incidents, and tests on 2020 incidents. Assignments are incident-level, ensuring that no incident crosses temporal partitions. Table 6 reports the eligible support. The nearly identical Task A and Task B counts make the two analyses comparable at the dataset level, while the OOD test cohort remains distinct from the randomly selected Incident-IID test cohort.

Findings. Figure 8 reveals that the effect of temporal shift depends on both model and target. Task A degradation ranges from -0.021 for XGBoost to $+0.036$ for TCN. Personnel forecasting spans -0.072 for S4M to $+0.107$ for TimeXer, while cost forecasting spans -0.076 for TFT to $+0.063$ for

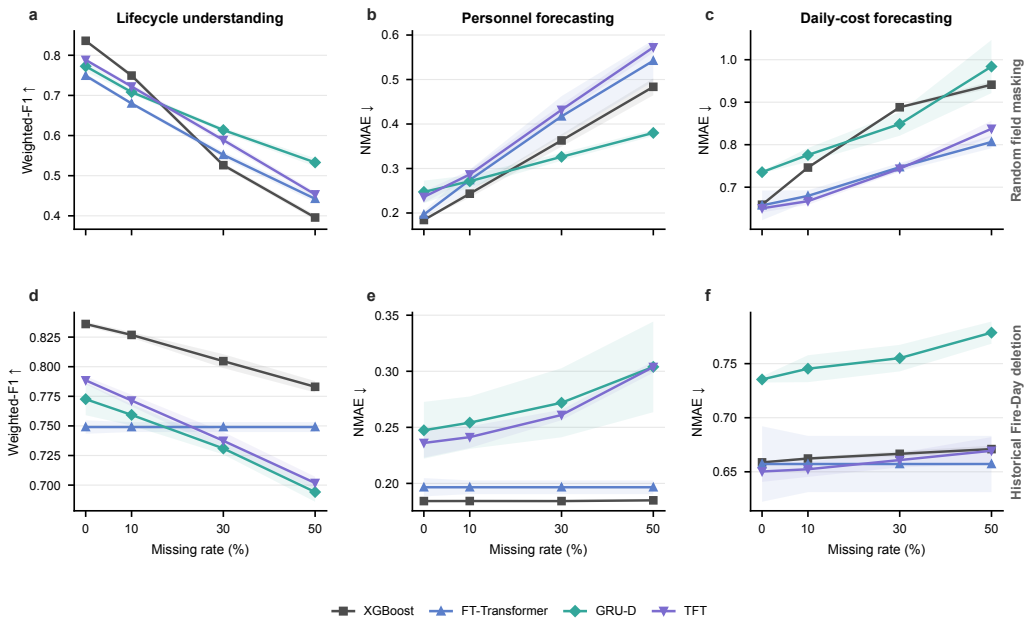


Figure 9: Robustness to stochastic test-time missingness. Panels (a)-(c) report Task A Weighted-F1 and Task B personnel and cost NMAE under random input-field masking; panels (d)-(f) report the corresponding metrics under random deletion of historical Fire-Days. Curves and shaded regions summarize repeated evaluations.

TimeMixer++. Several models that are competitive under Incident-IID evaluation deteriorate on at least one future-year target, and no architecture is uniformly stable across lifecycle, personnel, and cost outcomes. In particular, temporal sophistication alone does not guarantee temporal transfer: the learned dynamics can be sensitive to changes in incident composition, reporting practice, or response scale.

Negative values require a bounded interpretation. They indicate better aggregate performance on the Temporal-OOD cohort than on the IID cohort, not that distribution shift is beneficial for every incident. Because the cohorts are not paired and may differ in intrinsic difficulty, the heatmap measures the combined consequence of future-year composition and model transfer. The important benchmark conclusion is therefore not that one signed value proves robustness, but that rankings and degradation directions vary enough to justify a dedicated temporal protocol. Reporting only a random incident-disjoint split would conceal this instability.

B.7 Robustness to Missing Observations

All models are trained with the complete benchmark input; missingness is introduced only at test time. We first apply stochastic perturbations by independently masking input fields or deleting historical Fire-Days at rates of 10%, 30%, and 50%. Every nonzero setting summarizes ten evaluations (two model seeds and five masking seeds). We then remove complete evidence groups to emulate structured sensor or reporting loss. Finally, we stratify naturally observed samples into completeness quartiles. The first two analyses are controlled interventions on a fixed test cohort; the quartile analysis is observational and may be confounded by incident composition.

Stochastic missingness. Figure 9 shows a clear dose response under random field masking. At 50% masking, Task A Weighted-F1 falls to 0.396-0.533 across the four representative models. Personnel NMAE rises as high as 0.572 and cost NMAE as high as 0.984. The common deterioration indicates that within-day variables are complementary: even architectures designed for missing or temporal data cannot reconstruct arbitrary losses of contemporaneous evidence. Deleting entire historical Fire-Days is generally less destructive. FT-Transformer is almost unchanged in several

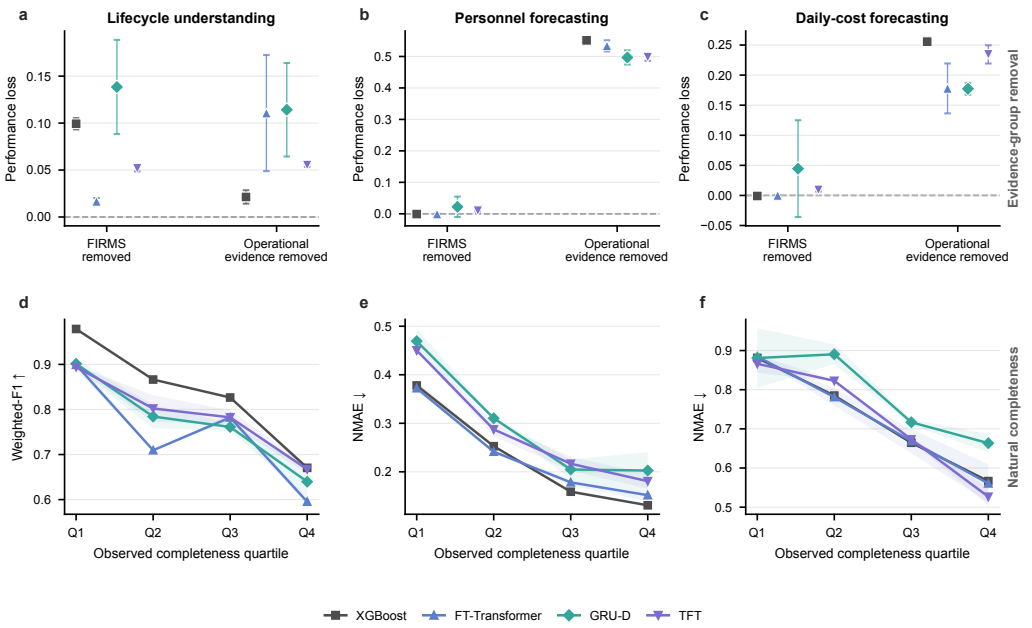


Figure 10: Robustness under structured and naturally occurring missingness. Panels (a)-(c) report performance loss after removing FIRMS or operational evidence as a complete feature group. Panels (d)-(f) report Task A Weighted-F1 and Task B personnel and cost NMAE across quartiles of naturally observed input completeness. Positive loss denotes degradation.

settings, while GRU-D and TFT are more sensitive because their sequence representations depend directly on the availability of historical steps. The contrast demonstrates that a single missing-rate statistic is insufficient; scattered field loss and missing temporal observations alter the problem in different ways.

Structured and natural missingness. Figure 10 confirms that the identity of the absent evidence matters more than its raw volume. Removing operational evidence increases personnel NMAE by roughly 0.50-0.55 and cost NMAE by 0.18-0.26, whereas removing FIRMS causes only small changes for these forecasting targets. Task A is more model-dependent: some baselines rely strongly on FIRMS, while others extract most lifecycle information from the operational report. This controlled result is consistent with the evidence ablation, but answers a distinct robustness question because the trained full-input model must cope with an unavailable source at test time.

Performance across natural-completeness quartiles does not reproduce the intervention curves uniformly. Personnel and cost forecasting generally improve with more complete observations, but Task A can show the opposite ordering for some models. This does not contradict the masking experiment. Naturally incomplete records may originate from systematically different incidents, such as shorter, smaller, or operationally simpler fires. Consequently, completeness may be correlated with the target distribution and sample difficulty. We therefore use the quartiles to document real-data heterogeneity, not to estimate the causal effect of missingness. Collectively, the robustness experiments establish a practical boundary for FireResBench: current models tolerate moderate loss of historical steps better than broad within-day or source-level loss, and future methods should combine missing-aware representations with explicit uncertainty estimates.

C Related Works

C.1 Wildfire Datasets and Spatiotemporal Modeling

Recent wildfire datasets and spatiotemporal models have substantially advanced data-driven understanding of fire occurrence, spread, and remote-sensing dynamics. The Next Day Wildfire Spread

dataset aligns historical fire masks with topography, vegetation, weather, drought, and population variables to support one-day-ahead spread prediction from gridded remote-sensing inputs (Huot et al., 2022). Subsequent benchmark resources further enrich the temporal and modal structure of wildfire prediction. WildfireSpreadTS introduces multi-modal daily time series for active wildfire spread forecasting across U.S. fire events (Gerard et al., 2023), while Sim2Real-Fire studies forecast and backtracking with large-scale simulated fire scenarios and real-world test cases (Li et al., 2024). FireSentry moves toward fine-grained local dynamics by collecting synchronized visible and infrared UAV videos, environmental measurements, and validated fire masks for high-resolution spread forecasting (Zhou et al., 2026). However, these efforts primarily focus on the physical fire process, such as fire boundary detection, burned-area mapping, and fire spread prediction. They generally do not connect evolving wildfire observations with operational response decisions. FireResBench is positioned in this gap: it organizes wildfire conditions and operational-response histories into standardized, incident-centered Fire-Day sequences, enabling wildfire lifecycle state understanding and next-day personnel and daily-cost forecasting.

C.2 Emergency Response and Resource Modeling

A separate line of work studies wildfire response records, suppression decisions, personnel deployment, and resource allocation. ICS-209-PLUS provides a large-scale all-hazards incident dataset mined from the U.S. National Incident Management System, including situation reports, incident status, impacts, and response-related variables for 1999-2020 (St. Denis et al., 2023). The U.S. Incident Management Situation Reports dataset further captures daily national and regional wildfire activity, preparedness levels, fire priorities, and committed suppression resources (Nguyen et al., 2024). Building on such operational records, recent empirical studies have examined suppression decisions in the western United States (Rossi et al., 2022), ignition precursors associated with high personnel deployment (Cullen et al., 2024), international fire management personnel sharing (Bloem et al., 2024), daily firefighting personnel deployment trends (Young et al., 2024), and factors driving daily fire prioritization (Nguyen et al., 2026). Optimization-oriented work has also begun to connect wildfire risk prediction with decision support for staffing-constrained resource allocation (Hsu, 2026). These studies demonstrate that response data contain rich signals about institutional priorities, fire conditions, exposure, and operational constraints. Yet most existing work is designed as a data descriptor, econometric analysis, prediction model, or decision-support pipeline for a specific response question. Such work does not provide a unified benchmark that organizes irregular incident reports into standardized Fire-Day sequences, defines lifecycle-state supervision and daily operational response targets, and supports reproducible comparisons across machine learning and temporal deep learning models under a common evaluation protocol. FireResBench therefore complements prior response modeling by turning irregular incident reports into event-centered standardized sequences for benchmark evaluation.