# Does Your Wildfire Prediction Model Actually Work, or Just Score Well?

**Yangshuang Xu**[*]
Florida State University
yx21e@fsu.edu

**Yuyang Dai**[*]
Florida State University
yd26@fsu.edu

**Liling Chang**
Florida State University
liling.chang@fsu.edu

**Qi Wang**
Northeastern University
wangqi@vt.edu

**Yushun Dong**[†]
Florida State University
yd24f@fsu.edu

## Abstract

Wildfire prediction is important for early warning and resource allocation, yet existing Earth foundation models (Earth FMs) are pretrained for general atmospheric and geophysical objectives rather than wildfire forecasting. To address this gap, we introduce WILD**FIRE**-FM, the first foundation model pretrained specifically for wildfire prediction using weather, active-fire observations, topography, vegetation, and static environmental data. However, introducing a domain-specific backbone alone does not solve the evaluation problem: wildfire events are sparse in space and time, making transfer conclusions highly sensitive to matching rules and evaluation settings. To address this problem, we introduce a fixed-contract evaluation framework with two controlled checks: a fixed-output check for matching-rule effects and a fixed-feature check for head-selection effects. Under matched contracts, we compare WILD**FIRE**-FM with ten Earth-FM baselines across occupancy, spread, retrieval, and regression tasks. Our results show that wildfire transfer conclusions depend strongly on evaluation design and task formulation. We hope this framework and WILD**FIRE**-FM provide a foundation for future wildfire-specific Earth-FM research and benchmarking. Our code is available at https://anonymous.4open.science/r/Wildfire-fm-evaluation-contracts-5AE9/.

## 1 Introduction

Wildfire prediction is critical for early warning and resource allocation in disaster response [10; 7]. As extreme fire events grow more frequent and severe, accurate forecasting of wildfire occurrence and spread is becoming increasingly important [29; 13]. Recent Earth foundation models (Earth FMs), pretrained on large-scale atmospheric and geophysical data [2; 35; 23], provide transferable representations for Earth-system dynamics and have shown strong performance across weather and remote-sensing tasks. However, wildfire dynamics depend on complex interactions among weather, vegetation, topography, fuel conditions, and active-fire behavior, which are not explicitly modeled during pretraining in existing general-purpose Earth FMs. This mismatch raises a natural question: can representations learned for general atmospheric or geophysical objectives transfer reliably to wildfire forecasting, and how should we measure that transfer?

Answering this question requires solving two intertwined problems. The first is that existing Earth FMs are not pretrained specifically for wildfire dynamics, but instead adapted to wildfire tasks

[*]Equal contribution.
[†]Corresponding author.

Preprint.

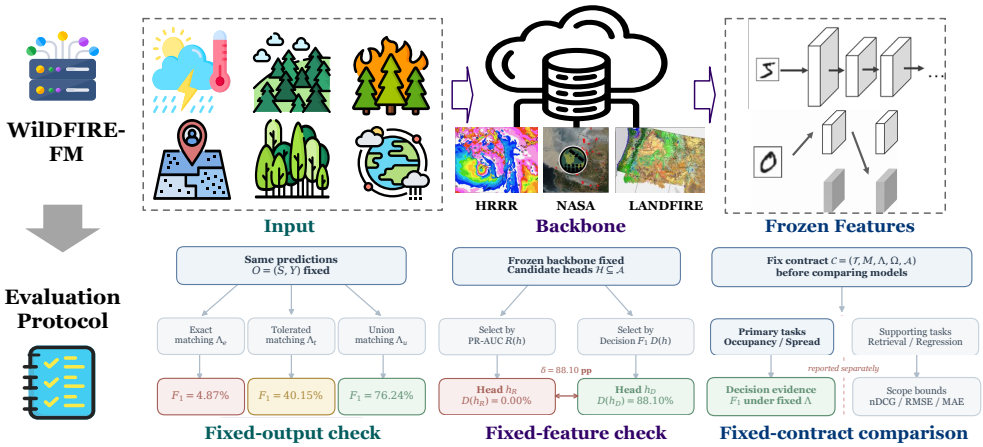


Figure 1: Overview of **WILDFIRE-FM** and **Evaluation Protocol** in this paper.

after general-purpose pretraining. To address this limitation, we introduce **WILDFIRE-FM**, the first foundation model pretrained specifically for wildfire prediction using fire-relevant multimodal data, including regional weather dynamics, active-fire observations, topography, vegetation, and static environmental context. By incorporating wildfire-specific signals directly during pretraining, WILD**FIRE**-FM learns representations aligned with the physical processes underlying fire behavior rather than relying on transfer from general atmospheric or geophysical objectives.

The second problem is evaluation: even with a domain-specific model, reliably comparing WILD**FIRE**-FM against transferred general-purpose Earth FMs remains difficult. Wildfire events are sparse in space and time [6; 9], making transfer conclusions highly sensitive to three sources of evaluation variability. *First,* matching rules determine what counts as a correct prediction. Early warning systems tolerate spatial offsets that post-fire damage assessment cannot, so different matching rules can produce substantially different F1 scores from the same model outputs [6; 9]. *Second,* head-selection metrics determine which lightweight adapter is chosen on top of a frozen representation. Ranking metrics such as PR-AUC and decision metrics such as F1 can favor different heads from the same frozen features [20; 37]. *Third,* wildfire task forms operate over different prediction units and metric families. Occupancy prediction, spread forecasting, burned-area regression, and analog retrieval therefore produce scores that are not directly comparable even under the same backbone [34; 8]. Related protocol sensitivity has also been observed in active learning [18] and selective classification [37]. We show that these effects become particularly severe in wildfire transfer evaluation, where sparse events and heterogeneous task forms amplify evaluation instability [19].

This evaluation instability makes reliable comparison between WILD**FIRE**-FM and existing Earth FMs fundamentally difficult. Standard geospatial benchmarks such as GEO-Bench [14], WeatherBench2 [31], WILDS [12], SustainBench [41], and TorchGeo [36] standardize datasets, splits, and metrics, but do so for tasks with dense, balanced labels where matching-rule sensitivity is not a primary concern. Wildfire studies such as FireCast [30] and Next Day Wildfire Spread [11] apply the same report-and-compare paradigm directly to sparse fire events without controlling for matching-rule choice, head-selection metric, or task-form comparability, the three sources of instability identified above. As a result, scores reported under different implicit protocol choices are not directly comparable, even when the underlying predictions are identical.

Based on the *limitations of prior work*, our contributions are as follows (see Figure 1).

- **Wildfire-specific foundation model.** We introduce WILD**FIRE**-FM, the first foundation model pretrained specifically for wildfire prediction using multimodal wildfire data spanning weather, active-fire observations, topography, vegetation, and static environmental context. Unlike general Earth FMs adapted after pretraining, WILD**FIRE**-FM learns wildfire representations directly from fire-relevant processes during pretraining.

- **Fixed-contract evaluation framework.** We formulate wildfire Earth-FM transfer as a fixed-contract evaluation problem, defining a contract $\mathcal{C} = (\mathcal{T}, M, \Lambda, \Omega, \mathcal{A})$ that specifies the task, metric, matching rule, evaluation scope, and lightweight-head family before comparison. We introduce two controlled checks: a *fixed-output check* for matching-rule effects and a

*fixed-feature check* for head-selection effects, enabling evaluation artifacts to be separated from representation quality.

- **Systematic wildfire transfer benchmark.** Under fixed contracts, we compare WILD**FIRE**-FM against ten general-purpose Earth FMs across six wildfire task forms. Our results show that wildfire transfer conclusions are highly sensitive to evaluation design and strongly task-dependent across occupancy, spread, retrieval, and regression settings.

## 2 WILDFIRE-FM Reference Backbone

WILD**FIRE**-FM is a wildfire-specialized regional backbone trained on fire-relevant multimodal data for wildfire prediction. Existing general-purpose Earth FMs are pretrained for atmospheric and geophysical objectives [15], or for remote-sensing objectives [32], so wildfire-relevant information enters only indirectly through those objectives. In contrast, WILD**FIRE**-FM is trained with weather, active-fire observations, topography, vegetation, and static environmental context, so its representation is learned from inputs tied directly to wildfire behavior. This design makes WILD**FIRE**-FM a strong wildfire-specific backbone whose features are shaped by signals directly relevant to fire occurrence and spread. It provides a task-aligned regional model trained directly for wildfire prediction. It also serves as an empirical anchor for interpreting how transferred Earth FMs behave under matched evaluation contracts. This section describes the data resources and training strategy used to build WILD**FIRE**-FM as an in-domain reference backbone. The fixed-contract protocol used to compare it with transferred Earth FMs is defined separately in Section 3.

### 2.1 Data Resources

We group the resources by their role in the study: dynamic weather inputs, occupancy supervision, static context, and event-level resources for supporting tasks. Source and terms-of-use notes for the external data and model assets used in this study are summarized in Appendix Table 16.

**Dynamic weather inputs.** The weather inputs come from a California regional dataset built from NOAA High-Resolution Rapid Refresh (HRRR) fields [24; 25]. The data are placed on a projected 5 km grid in EPSG:5070. Each time map uses weather fields every 6 hours and predicts wildfire occupancy at a 12-hour lead. The variables include near-surface temperature and dew point, wind, CAPE, surface pressure, boundary-layer height, visibility, precipitation rate, and accumulated precipitation.

**Occupancy supervision.** Wildfire supervision comes from NASA FIRMS active-fire detections [21]. The detections are mapped to the same grid as the weather fields. WILD**FIRE**-FM is trained on gridded occupancy labels derived from these detections. This defines the occupancy target used by the reference backbone throughout the primary experiments.

**Static context.** Static context describes landscape and exposure factors that do not change at the weather time step. These variables are LANDFIRE fire-behavior fuel model [16], LANDFIRE canopy cover [17], Wildfire Risk to Communities housing-unit density [39], and LandScan population [26]. Together with validity masks for the weather and static fields, the occupancy input has 16 channels: 10 weather fields, two validity masks, and four static layers for regional fire prediction.

**Event-level resources.** Event-level resources are used for supporting burned-area and analog tasks, not as occupancy labels for WILD**FIRE**-FM. These resources include WFIGS incident and perimeter attributes [22] and MTBS burned-area and burn-severity records [38]. They provide event-scale outcomes and incident metadata for supporting tasks in the experiments and appendix analyses.

### 2.2 Training Strategy

**Model and data split.** WILD**FIRE**-FM uses a compact U-Net [33] that maps gridded weather and static inputs to wildfire predictions. Its primary output is fire occupancy on the common spatial grid. Data are split by time: June–August 2024 for training, September 2024 for validation, and October 2024 for testing. This yields 368 training time maps, 120 validation time maps, and 120 test time maps. Temporal splitting keeps later fire outcomes out of earlier training periods.

**Fire-aware tile training.** Training is performed on $32\times32$ tiles sampled from the time maps. The tiles include fire-centered regions and non-fire context, so the model sees both sparse fire labels and surrounding background conditions. This sampling reduces the dominance of empty cells without

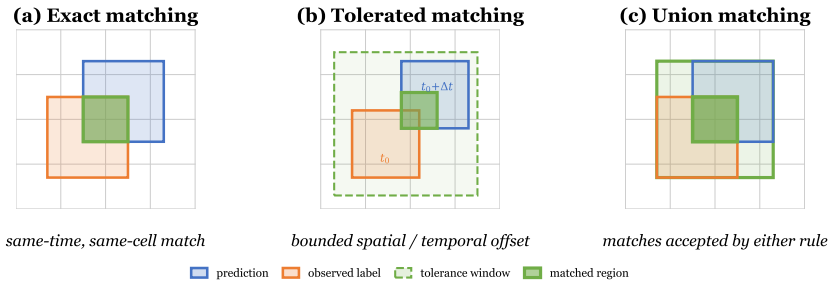


Figure 2: Matching rules for one fixed occupancy output. (a) Exact matching counts only same-time, same-cell overlap. (b) Tolerated matching accepts bounded spatial or temporal offsets. (c) The union reading counts matches accepted by either rule.

removing non-fire examples from the training distribution. Class-weighted binary cross-entropy is used for the primary occupancy target to further balance sparse positives.

**Spatial-support training objective.** Wildfire labels can shift by a few grid cells because detections, weather fields, and static layers are aligned on a common grid. To reduce sensitivity to these small displacements during training, the occupancy target is dilated by two grid cells. An auxiliary spatial-support output is trained for the same neighborhood alongside the primary occupancy output. At test time, WILD**FIRE**-FM is scored under the same task-specific evaluation contracts as the transferred Earth-FM backbones in Section 3, ensuring matched comparison conditions.

# 3 Evaluation Design

## 3.1 Wildfire Output Records and Fire Sets

**Output record.** A wildfire prediction model produces scores over spatial units and forecast times, which are compared against observed fire activity to compute performance. We formalize this comparison as a *wildfire output record* $\mathcal{O} = (S, Y)$, where the score field $S = \{s_{i,t}\}$ contains model scores over spatial units $i$ and times $t$, and the label field $Y = \{y_{i,t}\}$ contains the corresponding observations. For occupancy tasks, $y_{i,t} \in \{0, 1\}$ indicates whether fire is observed at $(i, t)$.

**Predicted and observed fire sets.** To evaluate $\mathcal{O}$, the score field is thresholded at $\tau$ to produce a predicted fire set $\hat{P}_\tau = \{(i, t) : s_{i,t} \geq \tau\}$, while the observed fire set is $P = \{(i, t) : y_{i,t} = 1\}$. The pair $(\hat{P}_\tau, P)$ is evaluated under a matching rule. Given a matching rule, true positives (TP), false positives (FP), and false negatives (FN) are computed from matched and unmatched elements, and the decision F1 score is $\text{F1} = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$. The same $(\hat{P}_\tau, P)$ can yield different TP, FP, and FN counts under different matching rules without changing model outputs, motivating the fixed-output check in Section 3.4.

**Matching rules.** A matching rule specifies when a predicted unit-time pair in $\hat{P}_\tau$ is considered a match to an observed pair in $P$ [6; 9]. Because wildfire applications tolerate different levels of spatial and temporal error, we define three matching rules for occupancy outputs. *(1) Exact matching*: requires agreement in both spatial unit and forecast time. *(2) Tolerated matching*: accepts predictions within a fixed spatial or temporal neighborhood defined by the evaluation contract $\mathcal{C}$. *(3) Union matching*: accepts predictions satisfying either exact or tolerated matching. Figure 2 illustrates these rules for a fixed output. Because the output record is held constant, any score difference is attributed solely to the matching rule.

## 3.2 Evaluation Contract

A wildfire transfer score depends not only on the model, but also on the evaluation choices used to compute it [18]. Changing the matching rule $\Lambda$, metric $M$, or evaluation scope $\Omega$ changes what the score measures even when model outputs are fixed.

We define an *evaluation contract* as the tuple $\mathcal{C} = (\mathcal{T}, M, \Lambda, \Omega, \mathcal{A})$, where $\mathcal{T}$ denotes the task, $M$ the metric, $\Lambda$ the matching rule, $\Omega$ the evaluation scope, and $\mathcal{A}$ the allowed lightweight-head family. Two transfer scores are comparable only when all five components are identical. The evaluation

scope $\Omega$ is particularly important in wildfire settings. A global scope evaluates the full spatial domain, including many fire-inactive regions that can mask differences between models. A fire-prone scope restricts evaluation to regions with higher historical fire activity. We report both scopes separately rather than averaging across them. Fixed matching-rule, task-form, and scope parameters are reported in Appendix Tables 3, 4, and 5.

### 3.3 Task-Form Contracts

Contract components depend on task form. We distinguish *primary* and *supporting* tasks based on whether they directly evaluate wildfire decisions. Occupancy and fire spread are primary tasks because they evaluate spatial fire outputs under matching or overlap rules. Retrieval, burned-area regression, smoke $PM_{2.5}$, and extreme heat are supporting tasks because they use different prediction units and metric families. Their results provide complementary evidence rather than direct substitutes for occupancy and spread evaluation [34].

For primary tasks, multiple metrics are reported for the same output under different contracts. For occupancy, exact F1 requires same-cell same-time agreement, tolerated F1 accepts predictions within a spatial or temporal neighborhood, and union F1 accepts predictions satisfying either rule. For fire spread, exact F1 evaluates raster-cell agreement, spatial F1 evaluates region overlap between $\hat{B}$ and $B$ [9], and AP summarizes ranking quality across thresholds. These metrics are reported separately rather than aggregated because they measure different aspects of the same prediction task. Figure 3 summarizes the contract map across all six task forms.

### 3.4 Controlled Checks

We isolate the two instability sources with two checks. Each check fixes all contract components except one, so any difference is attributed solely to that component.

**Fixed-output check.** The fixed-output check isolates matching-rule effects by holding the output record $\mathcal{O} = (S, Y)$ and all other contract components fixed while varying only $\Lambda$. For the same occupancy record, we compute F1 under exact, tolerated, and union matching. Any score difference is therefore attributed solely to the matching rule. If matching rules alone shift F1 by tens of percentage points on the same output, then comparing models under different $\Lambda$ conflates model quality with evaluation design.

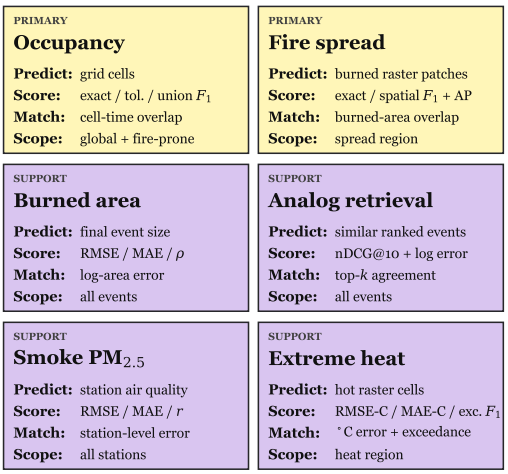


Figure 3: Evaluation contract map for the six fixed-contract tasks. Yellow boxes denote **primary** decision tasks; purple boxes denote **supporting** tasks.

**Fixed-feature check and selection regret.** The fixed-feature check isolates head-selection effects by holding the frozen feature source, $\mathcal{T}$, $\Omega$, $\Lambda$, and candidate head family $\mathcal{H} \subseteq \mathcal{A}$ fixed while varying only the selection metric. Let $R(h)$ denote the ranking score of head $h$ and $D(h)$ its decision score. Ranking-based selection chooses $h_R = \arg\max_{h \in \mathcal{H}} R(h)$, while decision-based selection chooses $h_D = \arg\max_{h \in \mathcal{H}} D(h)$. We define *selection regret* as the decision-score gap incurred by using a ranking metric as a proxy for a decision metric during head selection: $\delta = D(h_D) - D(h_R) \geq 0$ [20; 37]. When $\delta > 0$, the ranking metric selects a head with lower decision performance under the same frozen representation, indicating that the observed gap arises from metric misalignment rather than from representation quality.The head family used in fixed-feature comparisons is summarized in Appendix Table 15.

**Fixed-contract transfer comparison.** After the controlled checks establish that matching-rule and selection-metric effects are non-trivial, Earth-FM backbones are evaluated under a shared contract $\mathcal{C}$. Entries are compared only when they satisfy the same $(\mathcal{T}, M, \Lambda, \Omega, \mathcal{A})$ tuple. Supporting tasks test

whether occupancy and spread patterns generalize across task forms and provide additional evidence when transfer orderings are preserved.

# 4 Experiments

We address three research questions under the fixed-contract framework defined in Section 3. **RQ1:** Under fixed outputs, does the matching rule determine whether a wildfire model appears usable? **RQ2:** Under fixed features, does ranking-based head selection lose decision performance? **RQ3:** Under fixed task contracts, do model comparisons remain consistent across task forms?

## 4.1 Experimental Setup

**Task instances.** We instantiate the six task-form contracts defined in Section 3.3. Occupancy and fire spread serve as primary tasks because they evaluate spatial fire outputs under matching or overlap rules and align with the decision structure of early warning systems [10; 7]. The four supporting tasks, *final burned area, analog retrieval, smoke $PM_{2.5}$, and extreme heat*, use different prediction units and metric families; their results bound rather than replace primary decision evidence.

**Compared backbones.** The frozen Earth-FM comparator set includes Prithvi-WxC [35], Aurora [2], ClimaX [23], StormCast [27], DLWP [40], FCN [28], FengWu [4], FuXi [5], Pangu-Weather [1], and AlphaEarth [3]. WILD**FIRE**-FM serves as the wildfire-specialized reference backbone.

**Protocol.** For each comparison, the contract $\mathcal{C} = (\mathcal{T}, M, \Lambda, \Omega, \mathcal{A})$ is fixed before reporting test scores. Thresholds and morphology parameters are selected on validation data and held fixed at test time. Stochastic components are evaluated over five seeds and reported as mean $\pm$ standard deviation; deterministic fixed-output checks have zero seed variance by construction. Entries outside a fixed contract are omitted from main tables and documented in the appendix. For error metrics lower is better ($\downarrow$); for F1, AP, nDCG, and correlation metrics higher is better ($\uparrow$). Appendix Table 14 summarizes the seed-level checks behind the reported mean-with-std convention.

## 4.2 Matching-Rule Sensitivity Under Fixed Output (RQ1)

To answer RQ1, we conduct a fixed-output check on occupancy and fire spread tasks, holding the score field $S$, label field $Y$, threshold, and all other operating choices fixed while varying only the matching rule $\Lambda$ across exact, tolerated, and union settings. Occupancy results are reported in Figure 7 under both global and fire-prone scopes. The same progression is applied to fire spread outputs. Complete occupancy sweeps and predicted-positive rates are reported in Appendix Tables 6 and 8.

Because both tasks involve spatially sparse targets, fire-active cells for occupancy, burned raster patches for spread, the operational assumptions encoded in $\Lambda$ directly govern what the model is being asked to get right, making matching-rule choice a substantive experimental setting rather than a post hoc evaluation detail. The fixed-output results reveal a pattern that goes beyond score differences: matching-rule choice determines whether a model appears viable for wildfire decision tasks at all. Under exact matching, which requires same-cell same-time agreement, the majority of frozen Earth-FM backbones produce F1 scores that are effectively near zero, rendering them indistinguishable from an uninformative baseline and suggesting they have no practical utility for the task. As the matching rule relaxes to tolerated and then union matching, both of which reflect operationally realistic assumptions for early warning systems, where a prediction displaced by a few

| | Occupancy top-5% | | Spread region |
|---|---|---|---|
| | Ex->Tol | Tol->Un | Ex->Sp |
| **WildFIRE-FM** | #2->2 | #2->1 | #1->1 |
| Prithvi-WxC | #7->8 | #8->8 | #4->4 |
| Aurora | #10->11 | #11->10 | #2->2 |
| ClimaX | #9->3 | #3->2 | #3->3 |
| StormCast | #11->10 | #10->9 | #6->6 |
| DLWP | #4->5 | #5->3 | #9->9 |
| FCN | #5->6 | #6->4 | #11->11 |
| FengWu | #6->9 | #9->11 | #10->10 |
| FuXi | #3->4 | #4->5 | #5->5 |
| Pangu-Weather | #8->7 | #7->7 | #7->8 |
| AlphaEarth | #1->1 | #1->6 | #8->7 |

Figure 4: **Primary-task rank changes (RQ1).** Cells show rank before→after. Green/red/gray mark moving up/down/no change; darker green or red marks a larger move. Following Section 3.3, Ex/Tol/Un are occupancy exact, tolerated, and union matching; Sp is spread spatial-overlap $F_1$.

Table 1: **Primary fixed-contract transfer results (RQ1).** Occupancy metrics: exact, tolerated, union $F_1$ (%). Fire spread metrics: exact $F_1$ and spatial $F_1$ (%). Each block fixes $\mathcal{T}$, $\Lambda$, $\Omega$, and $\mathcal{A}$. Upward arrows indicate that larger values are better. **Bold** marks the best value per metric. **Tol.** = Tolerated

| | Occupancy | | | Fire spread | |
|---|---|---|---|---|---|
| **Comparator** | **Exact $F_1$ ↑** | **Tol. $F_1$ ↑** | **Union $F_1$ ↑** | **Exact $F_1$ ↑** | **Spatial $F_1$ ↑** |
| WILD**FIRE**-FM | $0.4546_{\pm 0.1412}$ | $29.7484_{\pm 1.2868}$ | $59.0656_{\pm 2.7372}$ | $\mathbf{37.6700_{\pm 0.9800}}$ | $\mathbf{80.9700_{\pm 2.0200}}$ |
| Prithvi-WxC | $0.0552_{\pm 0.0039}$ | $7.1649_{\pm 0.6557}$ | $20.1853_{\pm 1.8299}$ | $22.3500_{\pm 3.4500}$ | $65.2600_{\pm 1.0700}$ |
| Aurora | $0.0656_{\pm 0.0094}$ | $8.5009_{\pm 1.9594}$ | $23.1037_{\pm 4.9418}$ | $30.8757_{\pm 0.1343}$ | $71.7329_{\pm 0.0141}$ |
| ClimaX | $0.3480_{\pm 0.0754}$ | $\mathbf{29.7535_{\pm 3.6073}}$ | $\mathbf{60.1506_{\pm 7.5865}}$ | $27.9853_{\pm 2.0532}$ | $69.0634_{\pm 2.3832}$ |
| StormCast | $0.0626_{\pm 0.0119}$ | $8.1951_{\pm 2.1895}$ | $22.3817_{\pm 5.4294}$ | $14.8387_{\pm 7.5791}$ | $55.7568_{\pm 21.3003}$ |
| DLWP | $0.1693_{\pm 0.0419}$ | $14.9148_{\pm 3.2446}$ | $28.1901_{\pm 6.9658}$ | $5.9335_{\pm 10.0712}$ | $22.8587_{\pm 22.3750}$ |
| FCN | $0.2829_{\pm 0.0839}$ | $19.5061_{\pm 3.3412}$ | $40.0604_{\pm 9.3701}$ | $3.1798_{\pm 2.6598}$ | $15.6203_{\pm 12.4531}$ |
| FengWu | $0.2613_{\pm 0.0757}$ | $12.0050_{\pm 6.0239}$ | $24.1022_{\pm 13.6293}$ | $5.5189_{\pm 9.0883}$ | $18.4774_{\pm 22.4703}$ |
| FuXi | $0.3774_{\pm 0.1212}$ | $21.0323_{\pm 4.8211}$ | $37.2888_{\pm 9.4470}$ | $19.9909_{\pm 2.1364}$ | $56.1826_{\pm 3.0412}$ |
| Pangu-Weather | $0.2755_{\pm 0.1089}$ | $17.0909_{\pm 4.0477}$ | $35.6386_{\pm 9.0327}$ | $11.2583_{\pm 11.0719}$ | $32.5081_{\pm 25.4969}$ |
| AlphaEarth | $\mathbf{2.0606_{\pm 0.4404}}$ | $29.4476_{\pm 6.0064}$ | $37.4286_{\pm 9.9458}$ | $11.0995_{\pm 3.6088}$ | $32.8316_{\pm 7.4634}$ |

grid cells still triggers the correct response, the same frozen representations recover substantial decision performance, with several backbones crossing from near-zero to practically meaningful F1 levels. This transition is not a marginal score improvement: it is a qualitative change in whether a model can be considered usable. The same pattern holds for fire spread under region-level matching relaxation, where strict raster-cell agreement again suppresses performance for most backbones while spatial tolerance restores it. The implications for prior wildfire transfer claims are significant: papers that report model performance under a single implicit matching rule, which is common practice given that sparse decision targets almost always require some form of tolerance [6; 9], may be drawing viability conclusions that are entirely dependent on an undisclosed protocol choice. A model claimed to perform well under one tolerance assumption may be completely unusable under a stricter one, and vice versa. Matching rule cannot be treated as an evaluation detail; it is an experimental setting that must be fixed, reported, and justified as part of any wildfire transfer claim. Additional spread AP values under fixed scopes are reported in Appendix Table 9.

### 4.3 Head-Selection Sensitivity Under Fixed Features (RQ2)

To answer RQ2, we conduct a fixed-feature check on occupancy and fire spread tasks, holding the frozen feature source, $\mathcal{T}$, $\Omega$, $\Lambda$, and candidate head family $\mathcal{H} \subseteq \mathcal{A}$ fixed while varying only the selection metric between PR-AUC-based and decision-F1-based selection. The resulting selection regret $\delta = D(h_D) - D(h_R)$ measures the decision-score loss induced by metric misalignment. Occupancy results are reported in Figure 5 under both global and fire-prone scopes. Full per-seed and per-head details are reported in Appendix D, and the exact, tolerated, and union regret breakdown is provided in Appendix Table 7.

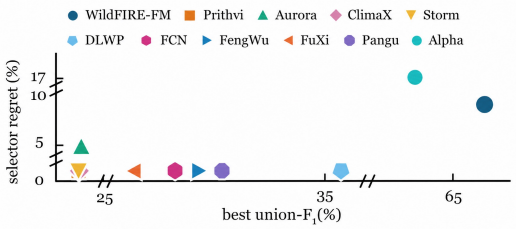


Figure 5: **Head-selection regret under fixed features (RQ2).** Each point is one backbone; selection regret $\delta$ follows Section 3.4 under global-scope union-$F_1$.

The fixed-feature results show that head-selection metrics introduce substantial backbone-dependent variation that is not explained by representation quality alone. Some backbones exhibit near-zero regret, indicating agreement between PR-AUC and decision-F1 selection, while others show large regret concentrated in specific scope-matching settings. Regret is generally larger under the global scope, where severe fire imbalance amplifies misalignment between ranking and decision metrics [20]. Restricting evaluation to fire-prone scopes typically reduces regret by concentrating evaluation on fire-relevant regions. A similar pattern appears for fire spread, where ranking and decision metrics

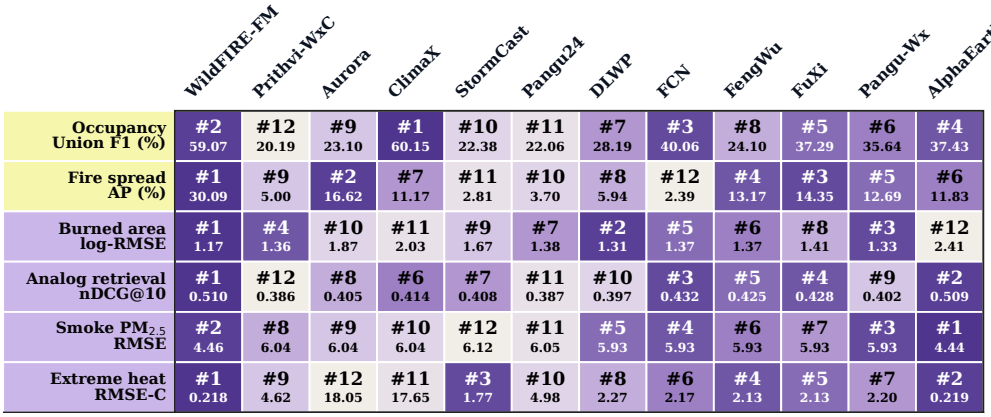


Figure 6: **Rank map for supporting task comparison (RQ4).** Each row fixes one task contract $\mathcal{C}$ and ranks the eligible backbones within that contract. The figure shows rank changes across task forms; native metric values are reported in Table 2.

can favor different heads under the same frozen representation. These results show that selection metrics must be aligned with the evaluation objective as part of the evaluation contract [37].

### 4.4 Supporting Task Checks (RQ3)

To answer RQ3, we evaluate all backbones across the four supporting task contracts, *burned area, analog retrieval, smoke PM$_{2.5}$,* and *extreme heat*, and examine whether the reference-versus-frozen ordering established under primary tasks generalizes across task forms. A rank overview across all six contracts is provided in Figure 6, which maps backbone-by-task rank positions and makes cross-task ordering shifts directly visible. Native metric values are reported in Table 2. Additional supporting-task diagnostics are reported in Appendix Tables 10, 11, 12, and 13.

The supporting task results produce three qualitatively distinct patterns relative to the primary findings. Burned area largely preserves the reference-versus-frozen ordering seen under occupancy and spread: WILD**FIRE**-FM leads frozen entries on log-RMSE and Spearman $\rho$, suggesting that the representational advantage of wildfire-specific pretraining generalizes to event-scale regression under a different metric family, providing convergent evidence for the primary claim. Analog retrieval and smoke PM$_{2.5}$ show a different pattern, with AlphaEarth matching WILD**FIRE**-FM closely on both tasks while atmospheric FMs show near-zero correlation on smoke PM$_{2.5}$, indicating that retrieval and air-quality signals are captured comparably by a general remote-sensing backbone, and that the primary occupancy advantage does not extend uniformly to these task forms. Extreme heat exhibits the largest variance across the comparator set, with atmospheric FMs ranging from near-reference performance to near-complete failure depending on backbone pretraining domain, while AlphaEarth again matches WILD**FIRE**-FM closely. The scale of this variance is itself informative: aggregating scores across task forms without respecting contract boundaries would produce rankings dominated by scale artifacts in the extreme heat block rather than by transfer quality. Taken together, these results establish that supporting tasks bound rather than extend the primary claim, they provide useful evidence about where backbone families generalize and where they do not, but they cannot substitute for primary decision task evaluation, and their results must within their own task-form contracts.

**Pattern 1: primary pattern preserved (burned area).** WILD**FIRE**-FM leads all frozen entries on log-RMSE and Spearman $\rho$. The ordering observed under occupancy and spread is preserved under burned-area regression despite the different prediction unit and metric family.

**Pattern 2: primary pattern bounded (analog retrieval and smoke PM$_{2.5}$).** For analog retrieval, AlphaEarth matches WILD**FIRE**-FM (nDCG@10 $= 0.51 \pm 0.04$ vs. $0.51 \pm 0.03$). For smoke PM$_{2.5}$, AlphaEarth also matches WILD**FIRE**-FM on MAE and Pearson $r$, while atmospheric Earth FMs show near-zero correlation. These results show that the occupancy-and-spread ordering does not fully extend to all supporting tasks once AlphaEarth is included.

Table 2: **Supporting task-metric matrix (RQ3).** Top: final burned area and analog retrieval. Bottom: smoke $PM_{2.5}$ and extreme heat. Each block fixes $\mathcal{T}$, $\Lambda$, and $\Omega$; backbone column is shared across paired tasks. WILD**FIRE**-FM row is separated by a rule as the empirical anchor. **Bold** marks the best value per metric. For error metrics lower is better (↓); for $F_1$, nDCG, and $r$ higher is better (↑).

| | Burned area | | | Analog retrieval | | |
|---|---|---|---|---|---|---|
| **Backbone** | **log-RMSE↓** | **log-MAE↓** | **Spearman↑** | **nDCG@10↑** | **log-RMSE↓** | **log-MAE↓** |
| WILD**FIRE**-FM | $\mathbf{1.1657}_{\pm 0.0126}$ | $\mathbf{1.0423}_{\pm 0.0081}$ | $\mathbf{0.6298}_{\pm 0.0338}$ | $\mathbf{0.5099}_{\pm 0.0336}$ | $\mathbf{1.1977}_{\pm 0.1029}$ | $\mathbf{1.0043}_{\pm 0.0759}$ |
| Prithvi-WxC | $1.3630_{\pm 0.0681}$ | $1.2435_{\pm 0.0668}$ | $0.1799_{\pm 0.3002}$ | $0.3857_{\pm 0.0189}$ | $1.3908_{\pm 0.0938}$ | $1.2585_{\pm 0.0865}$ |
| Aurora | $1.8658_{\pm 0.2009}$ | $1.6717_{\pm 0.1245}$ | $-0.1156_{\pm 0.2982}$ | $0.4046_{\pm 0.0144}$ | $1.3659_{\pm 0.0792}$ | $1.2596_{\pm 0.0968}$ |
| ClimaX | $2.0300_{\pm 0.2103}$ | $1.8443_{\pm 0.1528}$ | $-0.2515_{\pm 0.2688}$ | $0.4143_{\pm 0.0191}$ | $1.4526_{\pm 0.0926}$ | $1.2441_{\pm 0.1446}$ |
| StormCast | $1.6679_{\pm 0.1438}$ | $1.4745_{\pm 0.1134}$ | $0.1830_{\pm 0.1969}$ | $0.4076_{\pm 0.0094}$ | $1.3663_{\pm 0.0781}$ | $1.2371_{\pm 0.1078}$ |
| DLWP | $1.3070_{\pm 0.0980}$ | $1.1769_{\pm 0.0834}$ | $0.4888_{\pm 0.1368}$ | $0.3972_{\pm 0.0146}$ | $1.5351_{\pm 0.0802}$ | $1.3196_{\pm 0.0781}$ |
| FCN | $1.3693_{\pm 0.0885}$ | $1.2599_{\pm 0.0723}$ | $0.3484_{\pm 0.1662}$ | $0.4316_{\pm 0.0134}$ | $1.4604_{\pm 0.1035}$ | $1.2351_{\pm 0.0586}$ |
| FengWu | $1.3715_{\pm 0.1011}$ | $1.2604_{\pm 0.0820}$ | $0.3221_{\pm 0.2004}$ | $0.4246_{\pm 0.0237}$ | $1.4179_{\pm 0.0986}$ | $1.2233_{\pm 0.0915}$ |
| FuXi | $1.4068_{\pm 0.1011}$ | $1.3023_{\pm 0.0789}$ | $0.2663_{\pm 0.2561}$ | $0.4279_{\pm 0.0212}$ | $1.4290_{\pm 0.0929}$ | $1.2236_{\pm 0.0961}$ |
| Pangu-Weather | $1.3280_{\pm 0.0735}$ | $1.2081_{\pm 0.0607}$ | $0.4141_{\pm 0.1573}$ | $0.4017_{\pm 0.0245}$ | $1.4235_{\pm 0.0731}$ | $1.2225_{\pm 0.0847}$ |
| AlphaEarth | $2.4068_{\pm 0.2841}$ | $2.0822_{\pm 0.2371}$ | $-0.3428_{\pm 0.1716}$ | $0.5086_{\pm 0.0440}$ | $1.2158_{\pm 0.1310}$ | $1.0350_{\pm 0.1018}$ |

| | Smoke $PM_{2.5}$ | | | Extreme heat | | |
|---|---|---|---|---|---|---|
| **Backbone** | **RMSE↓** | **MAE↓** | **Pearson $r$ ↑** | **RMSE-C↓** | **MAE-C↓** | **Exceed. $F_1$ ↑** |
| WILD**FIRE**-FM | $4.4646_{\pm 0.0060}$ | $2.4108_{\pm 0.0016}$ | $\mathbf{0.6368}_{\pm 0.0013}$ | $\mathbf{0.2179}_{\pm 0.0043}$ | $\mathbf{0.1787}_{\pm 0.0018}$ | $0.9541_{\pm 0.0164}$ |
| Prithvi-WxC | $6.0382_{\pm 0.0828}$ | $3.7301_{\pm 0.0055}$ | $0.0243_{\pm 0.0045}$ | $4.6225_{\pm 0.0192}$ | $2.6315_{\pm 0.0128}$ | $0.8693_{\pm 0.0023}$ |
| Aurora | $6.0384_{\pm 0.0828}$ | $3.7265_{\pm 0.0055}$ | $0.0193_{\pm 0.0043}$ | $18.0474_{\pm 0.0708}$ | $15.3747_{\pm 0.0594}$ | $0.0951_{\pm 0.0038}$ |
| ClimaX | $6.0402_{\pm 0.0828}$ | $3.7290_{\pm 0.0055}$ | $0.0004_{\pm 0.0029}$ | $17.6492_{\pm 0.0347}$ | $14.4938_{\pm 0.0319}$ | $0.7684_{\pm 0.0068}$ |
| StormCast | $6.1230_{\pm 0.0830}$ | $3.8182_{\pm 0.0073}$ | $0.0183_{\pm 0.0041}$ | $1.7671_{\pm 0.2145}$ | $1.3507_{\pm 0.1576}$ | $0.9073_{\pm 0.0189}$ |
| DLWP | $5.9289_{\pm 0.1031}$ | $3.7331_{\pm 0.0088}$ | $0.0303_{\pm 0.0060}$ | $2.2662_{\pm 0.1106}$ | $1.7153_{\pm 0.0748}$ | $0.9156_{\pm 0.0112}$ |
| FCN | $5.9277_{\pm 0.1033}$ | $3.7345_{\pm 0.0088}$ | $0.0312_{\pm 0.0062}$ | $2.1657_{\pm 0.1800}$ | $1.6033_{\pm 0.1039}$ | $0.9257_{\pm 0.0096}$ |
| FengWu | $5.9297_{\pm 0.1032}$ | $3.7395_{\pm 0.0088}$ | $0.0304_{\pm 0.0063}$ | $2.1266_{\pm 0.1589}$ | $1.5801_{\pm 0.1004}$ | $0.0481_{\pm 0.0459}$ |
| FuXi | $5.9319_{\pm 0.1029}$ | $3.7398_{\pm 0.0088}$ | $0.0299_{\pm 0.0061}$ | $2.1282_{\pm 0.0969}$ | $1.5759_{\pm 0.0719}$ | $0.2268_{\pm 0.0623}$ |
| Pangu-Weather | $5.9270_{\pm 0.1036}$ | $3.7320_{\pm 0.0088}$ | $0.0301_{\pm 0.0060}$ | $2.2045_{\pm 0.1483}$ | $1.6307_{\pm 0.0889}$ | $0.0199_{\pm 0.0062}$ |
| AlphaEarth | $\mathbf{4.4403}_{\pm 0.0488}$ | $\mathbf{2.3992}_{\pm 0.0056}$ | $0.6347_{\pm 0.0066}$ | $0.2194_{\pm 0.0039}$ | $0.1800_{\pm 0.0014}$ | $\mathbf{0.9542}_{\pm 0.0107}$ |

**Pattern 3: primary pattern bounded with large variance (extreme heat).** AlphaEarth matches WILD**FIRE**-FM on RMSE-C and remains close on exceedance F1, while atmospheric FMs range from RMSE-C $= 1.77$ (StormCast) to $18.05$ (Aurora). This large spread indicates that aggregated scores across task forms would be dominated by scale artifacts rather than transfer quality, reinforcing the need for per-contract reporting established in Section 3.

*Answer to RQ3:* Figure 6 and Table 2 show that burned area preserves the primary reference-versus-frozen pattern under a different metric family. Analog retrieval, smoke $PM_{2.5}$, and extreme heat fourcast-net AlphaEarth matches or approaches WILD**FIRE**-FM on these tasks, indicating that the primary occupancy and spread claims do not extend uniformly across all task forms.

## 5 Conclusion

We introduced WILD**FIRE**-FM, the first foundation model pretrained specifically for wildfire prediction using fire-relevant multimodal data. Our results show that wildfire forecasting requires representations aligned with wildfire dynamics rather than transfer alone from general atmospheric or geophysical pretraining. At the same time, our study shows that reliable wildfire transfer evaluation is substantially more difficult than standard benchmark settings suggest. Wildfire transfer conclusions depend strongly on matching rules, head-selection metrics, and task form, and scores computed under different evaluation settings are not directly comparable. These effects become particularly pronounced in sparse spatiotemporal prediction settings such as wildfire forecasting. We therefore introduced a fixed-contract evaluation framework for wildfire Earth-FM transfer. By explicitly specifying the task, metric, matching rule, evaluation scope, and head family before comparison, fixed-contract evaluation enables more controlled and interpretable comparison across wildfire tasks and models. We hope WILD**FIRE**-FM and the fixed-contract framework provide a foundation for future wildfire-specific Earth FMs, transfer benchmarks, and decision-oriented evaluation protocols. More broadly, our research provides a reliable system to guide real-world intervention and resource allocation at the intersection of AI for environmental decision-making.

**Limitations.** The conclusions apply to the task forms, scopes, evaluation rules, and comparator eligibility decisions used in this study. The evaluation covers selected wildfire decision tasks and supporting retrieval and regression task forms. They provide task-form evidence rather than a single score across all wildfire-related prediction tasks.

## References


[1] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *Nature*, 619(7970):533–538, 2023.

[2] Cristian Bodnar et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024.

[3] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

[4] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

[5] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

[6] Elizabeth E. Ebert. Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting*, 24(6):1498–1510, 2009.

[7] Alireza Farahmand, E Natasha Stavros, John T Reager, and Ali Behrangi. Introducing spatially distributed fire danger from earth observations (fdeo) using satellite-based data in the contiguous united states. *Remote Sensing*, 12(8):1252, 2020.

[8] Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. *Advances in Neural Information Processing Systems*, 36:74515–74529, 2023.

[9] Eric Gilleland, David Ahijevych, Barbara G Brown, and Elizabeth E Ebert. Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5):1416–1430, 2009.

[10] Johann Georg Goldammer. Early warning systems for the prediction of and appropriate response to wildfires and related environmental hazards. In *Early Warning Systems for Natural Disaster Reduction*, 1999.

[11] Fantine Huot, R Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire prediction using deep learning. *arXiv preprint arXiv:2206.08930*, 2022.

[12] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, pages 5637–5664, 2021.

[13] Vassiliki Kotroni, Constantinos Cartalis, Silas Michaelides, Julia Stoyanova, Filippos Tymvios, Antonis Bezes, Theodoros Christoudias, Stavros Dafis, Christos Giannakopoulos, Theodore M. Giannaros, et al. Disarm early warning system for wildfires in the eastern mediterranean. *Sustainability*, 12(16):6670, 2020.

[14] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. GEO-Bench: Toward foundation models for earth monitoring. In *Advances in Neural Information Processing Systems*, 2023.

[15] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[16] LANDFIRE. LANDFIRE 40 Fire Behavior Fuel Models. `https://landfire.gov/fuel/fbfm40`, 2026. Accessed: 2026-05-05.

[17] LANDFIRE. LANDFIRE Forest Canopy Cover. `https://landfire.gov/fuel/cc`, 2026. Accessed: 2026-05-05.

[18] Carsten T. Lüth, Till J. Bungert, Lukas Klein, and Paul F. Jäger. Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment. In *Advances in Neural Information Processing Systems*, 2023.

[19] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

[20] Matthew B. McDermott, Haoran Zhang, Lasse Hyldig Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. In *Advances in Neural Information Processing Systems*, 2024.

[21] NASA Earthdata. Fire Information for Resource Management System (FIRMS). `https://www.earthdata.nasa.gov/data/tools/firms`, 2026. Accessed: 2026-05-05.

[22] National Interagency Fire Center. Wildland Fire Interagency Geospatial Services (WFIGS): Current Perimeters. `https://data-nifc.opendata.arcgis.com/datasets/nifc::wfigs-current-perimeters/about`, 2026. Accessed: 2026-05-05.

[23] Tung Nguyen, Johannes Brandstetter, Aditya Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[24] NOAA National Centers for Environmental Information. Rapid Refresh / High-Resolution Rapid Refresh. `https://www.ncei.noaa.gov/products/weather-climate-models/rapid-refresh-update`, 2026. Accessed: 2026-05-05.

[25] NOAA National Centers for Environmental Prediction Environmental Modeling Center. High-Resolution Rapid Refresh (HRRR). `https://rapidrefresh.noaa.gov/hrrr/`, 2026. Accessed: 2026-05-05.

[26] Oak Ridge National Laboratory. LandScan Global 2024. `https://landscan.ornl.gov/`, 2024. Accessed: 2026-05-05.

[27] Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, et al. Kilometer-scale convection-allowing model emulation using generative diffusion modeling. *Science Advances*, 12(5):eadv0423, 2026.

[28] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[29] Paul D Pickell, Nicholas C Coops, Colin J Ferster, Christopher W Bater, Karen D Blouin, Mike D Flannigan, and Jinkai Zhang. An early warning system to forecast the close of the spring burning window from satellite-observed greenness. *Scientific Reports*, 7(1):14190, 2017.

[30] David Radke, Anna Hessler, and David Ellsworth. Firecast: Leveraging deep learning to predict wildfire spread. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4575–4581, 2019.

[31] Stephan Rasp, Stephan Hoyer, Alex Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

[32] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Stefano Ermon, and Ruslan Salakhutdinov. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*, 2023.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[34] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

[35] Johannes Schmude, Sujit Roy, Paulina Trofimova, Karthik Ramesh, Bethany Lusch, Harikumar Kesa, Shraddha Singh, Phil Chen, Zhuohan Liu, Shubhankar Parashar, et al. Prithvi wxc: Foundation model for weather and climate. *arXiv preprint arXiv:2409.13598*, 2024.

[36] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. Torchgeo: Deep learning with geospatial data. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2022.

[37] Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, 2024.

[38] U.S. Geological Survey and USDA Forest Service. Monitoring Trends in Burn Severity (MTBS). `https://www.mtbs.gov/`, 2025. Accessed: 2026-05-05.

[39] USDA Forest Service. Wildfire Risk to Communities: Housing Unit Density Image Service. `https://catalog.data.gov/dataset/wildfire-risk-to-communities-housing-unit-density-image-service-fac22`, 2026. Accessed: 2026-05-05.

[40] Jonathan A Weyn, Dale R Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020.

[41] Christopher Yeh, Chenlin Meng, Sijing Wang, Anne Driscoll, Erik Rozi, Peng Liu, Jae Yong Lee, Marshall Burke, David B. Lobell, and Stefano Ermon. SustainBench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Advances in Neural Information Processing Systems*, 2021.

## Appendix Contents



*Retention rule.* Appendix tables are retained when they add contract parameters, controlled-check arithmetic, task-specific non-main metrics, seed summaries, eligibility checks, or protocol details. Full task matrices and reference-summary tables that repeat the main result tables are not repeated here.

# A Evaluation Contract Specifications

## A.1 Matching Rule Definitions

The three matching rules used across occupancy task forms are defined as follows.

**Exact matching.** A predicted unit-time pair $(i, t) \in \widehat{P}_\tau$ is counted as a true positive if and only if the same pair appears in the observed fire set $P = \{(i, t) : y_{i,t} = 1\}$. This is the strictest rule and yields the lowest $F_1$ for any fixed output.

**Tolerated matching.** A predicted pair $(i, t)$ is counted as correct if there exists an observed pair $(i', t') \in P$ such that $\|i - i'\|_\infty \leq k$ and $|t - t'| \leq \Delta t$, where $k$ is the spatial tolerance in grid cells and $\Delta t$ is the temporal tolerance in forecast steps. Both parameters are fixed as part of the evaluation contract $\mathcal{C}$ before scoring.

**Union matching.** A predicted pair is counted as a true positive if it satisfies either exact or tolerated matching. The resulting union-$F_1$ provides an upper bound on decision performance under the chosen tolerance.

**Fixed parameter values.** For occupancy, the spatial tolerance is $k = 8$ grid cells. The temporal tolerance is $\Delta t = 3$ forecast steps for union matching and $\Delta t = 0$ for spatial-only tolerance. The threshold $\tau$ is selected on validation strict-$F_1$ before test scoring. For fire spread, the spatial tolerance is $k = 4$ grid cells, $\Delta t = 0$, and the threshold is selected on validation spatial $F_1$.

Table 3 records the fixed matching-rule parameters.

Table 3: Matching-rule values used in the evaluation contracts.

| Parameter | Occupancy | Fire spread |
|---|---|---|
| $k$ | 8 cells | 4 cells |
| $\Delta t$ | 3 for union; 0 spatial-only | 0 |
| $\tau$ | val. strict $F_1$ | val. spatial $F_1$ |

## A.2 Task-Form Contract Parameters

Table 4 lists fixed scoring values not shown in the main contract map.

Table 4: Fixed scoring values used by each task-form contract.

| $\mathcal{T}$ | Scoring | Validation | $\Omega$ |
|---|---|---|---|
| Occupancy | $k = 8$, $\Delta t = 3$; exact/tol./union $F_1$ | val. strict $F_1$ | global; top-5/10/20% fire-prone |
| Fire spread | $k = 4$, $\Delta t = 0$; exact/spatial $F_1$, AP | val. spatial $F_1$ | spread-region cells |
| Final burned area | log-RMSE, log-MAE, Spearman $\rho$ | val. log-RMSE | test events |
| Analog retrieval | nDCG@10; retrieved-event log error | val. nDCG@10 | test events |
| Smoke $PM_{2.5}$ | RMSE, MAE, Pearson $r$; exceedance 35 | val. RMSE | test stations |
| Extreme heat | RMSE-C, MAE-C, exceedance $F_1$ | val. threshold 27/30/33°C | heat-region stations |

## A.3 Evaluation Scope Definitions

**Global scope.** Evaluation covers all spatial units in the domain, including fire-inactive regions. This scope can mask model differences on fire-relevant locations because inactive cells inflate true-negative counts.

**Fire-prone scope.** Evaluation is restricted to grid cells in the top-$k$% of historical fire activity. We report results for top-5%, top-10%, and top-20% cutoffs. The cutoff thresholds are derived from the training period and held fixed at test time.

**Spread region scope.** For fire spread tasks, evaluation is restricted to the predicted and observed burned raster patches. Only cells within the union of $\widehat{B}$ and $B$ contribute to metric computation.

**Fixed scope sizes.** The global scope contains 8,085,000 test cells. The fire-prone top-5%, top-10%, and top-20% scopes contain 404,280, 808,560, and 1,617,000 test cells, respectively. The spread-region scope is event-specific and uses the union of $\widehat{B}$ and $B$.

Table 5: Scope values used in the evaluation contracts.

| $\Omega$ | **Definition** | **Units** |
|---|---|---|
| Global | full domain | 8,085,000 test cells |
| Fire-prone top-5% | top 5% by training-period fire frequency | 404,280 test cells |
| Fire-prone top-10% | top 10% by training-period fire frequency | 808,560 test cells |
| Fire-prone top-20% | top 20% by training-period fire frequency | 1,617,000 test cells |
| Spread region | union of $\widehat{B}$ and $B$ | event-specific cells |

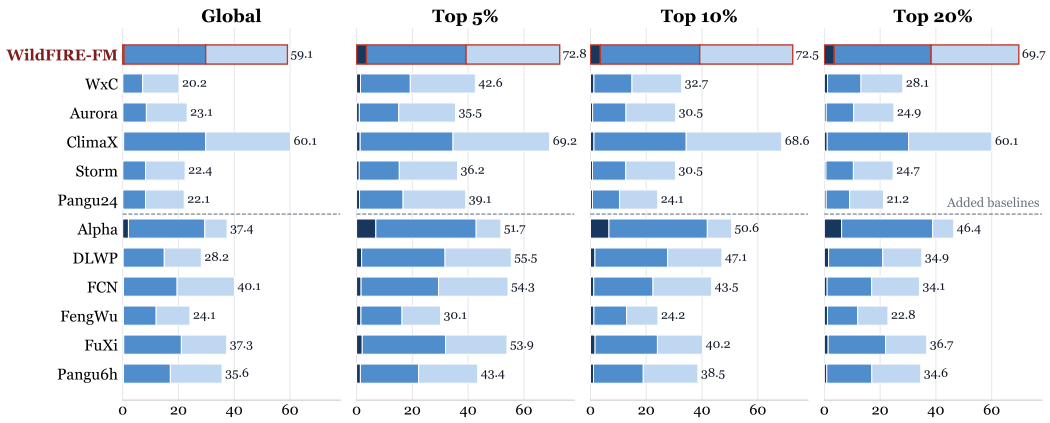


Figure 7: **Matching-rule sensitivity in fire-prone occupancy (RQ1).** Each row holds the score field $S$, label field $Y$, threshold, and $\Omega$ fixed, and changes only $\Lambda$. Legend: ■ strict $F_1$, ■ added $F_1$ from spatial tolerance, ■ added $F_1$ from union matching, red outline WILD**FIRE**-FM, and dashed line original weather FMs vs. added baselines. The horizontal axis is $F_1$ in percent.

# B Controlled Check Details

## B.1 Fixed-Output Check: Full Sweep

The fixed-output check holds the score field $S$ and label field $Y$ fixed and varies only $\Lambda$. Table 6 reports the full global and fire-prone sweep for all retained backbones. The same table is the numeric counterpart to Figure 7.

## B.2 Fixed-Feature Check: Selection Summary

The paper appendix keeps the fixed-feature result at the selection-summary level. The full per-head rows are retained in the supplementary CSV files and are not repeated as a manuscript table because many degenerate heads produce identical zero decision scores. The supplementary selection rows report the decision-score loss after changing only the head-selection metric.

## B.3 Selection Regret Under Matching Rules

The fixed-feature check trains the same head family $\mathcal{H}$ on a fixed feature source and changes only the selection metric. Table 7 reports the same selection comparison under exact, tolerated, and union matching. Here, $h_R$ is selected by PR-AUC and $h_D$ is selected by the decision metric. The reported regret is $D(h_D) - D(h_R)$. Exact zero entries mean the two selectors give the same decision score for all five seeds.

## B.4 Additional Value Tables

Table 8 reports the predicted-positive rate behind the occupancy $F_1$ sweep.

Tables 9–13 report additional values that are not repeated in the main tables. Each table fixes the task $\mathcal{T}$ and reports either a different $\Omega$, metric, or event subset.

Table 6: Occupancy $F_1$ scores across global and fire-prone scopes. Global uses the full validation/test domain; top-$k$ rows use train-defined fire-prone masks from historical fire frequency. Values are percentages from the same validation-selected strict threshold. Tolerance is spatial-only; union adds temporal and spatial matching. $\Delta$ is union minus strict. Cells report five-seed mean with std in small type.

| Backbone | $\Omega$ | Strict $F_1$ ↑ | Tol. $F_1$ ↑ | Union $F_1$ ↑ | $\Delta$ ↑ |
|---|---|---|---|---|---|
| WILD**FIRE**-FM | global | $0.4546_{\pm 0.1412}$ | $29.7484_{\pm 1.2868}$ | $59.0656_{\pm 2.7372}$ | $58.6109_{\pm 2.6945}$ |
| | top 5% | $3.5604_{\pm 0.8809}$ | $39.2617_{\pm 1.4011}$ | $72.8280_{\pm 2.5784}$ | $69.2676_{\pm 1.9960}$ |
| | top 10% | $3.5575_{\pm 0.8799}$ | $39.1665_{\pm 1.3906}$ | $72.5204_{\pm 2.5670}$ | $68.9629_{\pm 1.9888}$ |
| | top 20% | $3.5300_{\pm 0.8700}$ | $38.2849_{\pm 1.2952}$ | $69.7228_{\pm 2.4664}$ | $66.1928_{\pm 1.9273}$ |
| Prithvi-WxC | global | $0.0552_{\pm 0.0039}$ | $7.1649_{\pm 0.6557}$ | $20.1853_{\pm 1.8299}$ | $20.1301_{\pm 1.8297}$ |
| | top 5% | $1.4119_{\pm 1.1635}$ | $19.2636_{\pm 4.5019}$ | $42.5793_{\pm 4.5495}$ | $41.1674_{\pm 3.4846}$ |
| | top 10% | $1.2376_{\pm 1.3201}$ | $14.8780_{\pm 8.4429}$ | $32.6913_{\pm 13.2085}$ | $31.4536_{\pm 11.9053}$ |
| | top 20% | $1.1520_{\pm 1.3770}$ | $13.1512_{\pm 9.4556}$ | $28.1319_{\pm 15.2866}$ | $26.9800_{\pm 13.9224}$ |
| Aurora | global | $0.0656_{\pm 0.0094}$ | $8.5009_{\pm 1.9594}$ | $23.1037_{\pm 4.9418}$ | $23.0382_{\pm 4.9325}$ |
| | top 5% | $0.9859_{\pm 0.9299}$ | $15.1337_{\pm 6.0821}$ | $35.4834_{\pm 11.0192}$ | $34.4975_{\pm 10.3728}$ |
| | top 10% | $0.7790_{\pm 1.0453}$ | $12.7381_{\pm 6.5558}$ | $30.5305_{\pm 10.8842}$ | $29.7515_{\pm 9.8656}$ |
| | top 20% | $0.6655_{\pm 1.1043}$ | $10.5304_{\pm 7.4309}$ | $24.9444_{\pm 12.5844}$ | $24.2790_{\pm 11.4943}$ |
| ClimaX | global | $0.3480_{\pm 0.0754}$ | $29.7535_{\pm 3.6073}$ | $60.1506_{\pm 7.5865}$ | $59.8026_{\pm 7.5454}$ |
| | top 5% | $1.2937_{\pm 0.1086}$ | $34.5791_{\pm 2.3772}$ | $69.2186_{\pm 5.7215}$ | $67.9249_{\pm 5.7263}$ |
| | top 10% | $1.2522_{\pm 0.1602}$ | $34.3341_{\pm 2.2852}$ | $68.5713_{\pm 5.5377}$ | $67.3191_{\pm 5.5538}$ |
| | top 20% | $1.0287_{\pm 0.2686}$ | $30.2140_{\pm 4.2857}$ | $60.0650_{\pm 7.5674}$ | $59.0363_{\pm 7.5891}$ |
| StormCast | global | $0.0626_{\pm 0.0119}$ | $8.1951_{\pm 2.1895}$ | $22.3817_{\pm 5.4294}$ | $22.3191_{\pm 5.4178}$ |
| | top 5% | $0.9573_{\pm 0.8011}$ | $15.3219_{\pm 5.5337}$ | $36.1857_{\pm 9.7331}$ | $35.2284_{\pm 9.1816}$ |
| | top 10% | $0.7284_{\pm 0.9280}$ | $12.6669_{\pm 6.3290}$ | $30.4748_{\pm 10.6527}$ | $29.7464_{\pm 9.7494}$ |
| | top 20% | $0.5795_{\pm 0.9104}$ | $10.4157_{\pm 7.3437}$ | $24.6598_{\pm 12.3973}$ | $24.0803_{\pm 11.4988}$ |
| DLWP | global | $0.1693_{\pm 0.0419}$ | $14.9148_{\pm 3.2446}$ | $28.1901_{\pm 6.9658}$ | $28.0208_{\pm 6.9257}$ |
| | top 5% | $1.8054_{\pm 0.4835}$ | $31.7231_{\pm 3.2923}$ | $55.4596_{\pm 5.2920}$ | $53.6542_{\pm 5.4752}$ |
| | top 10% | $1.6110_{\pm 0.5999}$ | $27.6581_{\pm 5.9216}$ | $47.1269_{\pm 8.0111}$ | $45.5158_{\pm 7.7927}$ |
| | top 20% | $1.5248_{\pm 0.8987}$ | $20.9403_{\pm 4.7971}$ | $34.9301_{\pm 7.8471}$ | $33.4054_{\pm 7.8760}$ |
| FCN | global | $0.2829_{\pm 0.0839}$ | $19.5061_{\pm 3.3412}$ | $40.0604_{\pm 9.3701}$ | $39.7775_{\pm 9.3423}$ |
| | top 5% | $1.6231_{\pm 0.5064}$ | $29.3769_{\pm 2.7626}$ | $54.3033_{\pm 7.4089}$ | $52.6801_{\pm 7.4389}$ |
| | top 10% | $1.1777_{\pm 0.5118}$ | $22.4217_{\pm 3.9803}$ | $43.4510_{\pm 9.2513}$ | $42.2734_{\pm 9.0251}$ |
| | top 20% | $0.9962_{\pm 0.4315}$ | $16.9792_{\pm 3.9371}$ | $34.0859_{\pm 8.2616}$ | $33.0897_{\pm 7.9275}$ |
| FengWu | global | $0.2613_{\pm 0.0757}$ | $12.0050_{\pm 6.0239}$ | $24.1022_{\pm 13.6293}$ | $23.8410_{\pm 13.5736}$ |
| | top 5% | $1.5695_{\pm 0.3592}$ | $16.2763_{\pm 3.7024}$ | $30.1055_{\pm 5.0103}$ | $28.5360_{\pm 4.7696}$ |
| | top 10% | $1.2427_{\pm 0.5333}$ | $12.9503_{\pm 5.6052}$ | $24.1854_{\pm 8.6854}$ | $22.9427_{\pm 8.1863}$ |
| | top 20% | $1.1192_{\pm 0.5023}$ | $11.9508_{\pm 5.0745}$ | $22.7860_{\pm 7.9115}$ | $21.6668_{\pm 7.4438}$ |
| FuXi | global | $0.3774_{\pm 0.1212}$ | $21.0323_{\pm 4.8211}$ | $37.2888_{\pm 9.4470}$ | $36.9114_{\pm 9.4327}$ |
| | top 5% | $2.0307_{\pm 0.6800}$ | $31.8944_{\pm 4.7331}$ | $53.9308_{\pm 8.3822}$ | $51.9001_{\pm 8.6878}$ |
| | top 10% | $1.6542_{\pm 0.7316}$ | $24.0128_{\pm 5.7784}$ | $40.2140_{\pm 9.9307}$ | $38.5597_{\pm 9.7744}$ |
| | top 20% | $1.3646_{\pm 0.6773}$ | $21.9548_{\pm 5.8601}$ | $36.7314_{\pm 10.0289}$ | $35.3668_{\pm 9.9223}$ |
| Pangu-Weather | global | $0.2755_{\pm 0.1089}$ | $17.0909_{\pm 4.0477}$ | $35.6386_{\pm 9.0327}$ | $35.3630_{\pm 9.0774}$ |
| | top 5% | $1.3656_{\pm 0.3064}$ | $22.2222_{\pm 6.8613}$ | $43.4234_{\pm 13.2383}$ | $42.0578_{\pm 13.0599}$ |
| | top 10% | $1.0931_{\pm 0.3535}$ | $18.9337_{\pm 5.9329}$ | $38.5325_{\pm 11.7221}$ | $37.4394_{\pm 11.5261}$ |
| | top 20% | $0.8844_{\pm 0.3601}$ | $17.0172_{\pm 5.4859}$ | $34.5688_{\pm 10.2932}$ | $33.6844_{\pm 10.1334}$ |
| AlphaEarth | global | $2.0606_{\pm 0.4404}$ | $29.4476_{\pm 6.0064}$ | $37.4286_{\pm 9.9458}$ | $35.3679_{\pm 10.0271}$ |
| | top 5% | $6.9133_{\pm 0.8450}$ | $42.8790_{\pm 4.6087}$ | $51.7449_{\pm 8.7321}$ | $44.8315_{\pm 9.0763}$ |
| | top 10% | $6.6366_{\pm 0.9901}$ | $41.8981_{\pm 5.9454}$ | $50.5712_{\pm 10.0057}$ | $43.9346_{\pm 9.9156}$ |
| | top 20% | $6.1908_{\pm 1.1330}$ | $38.8325_{\pm 7.4966}$ | $46.3833_{\pm 12.1697}$ | $40.1925_{\pm 11.6788}$ |

Table 7: Selection-regret values under exact, tolerated, and union matching. Values are percentage-point regret from selecting $h_R$ by PR-AUC instead of $h_D$ by the decision metric. Rows report mean with small std over five seeds; 0.0000 denotes exact zero regret.

| Feature | $\Omega$ | Exact regret | Tolerated regret | Union regret |
|---|---|---|---|---|
| WILD**FIRE**-FM | global | 0.0000 | $8.7830_{\pm 9.6705}$ | $8.7830_{\pm 9.6705}$ |
| WILD**FIRE**-FM | fire-prone | 0.0000 | $3.4027_{\pm 3.2045}$ | $3.4027_{\pm 3.2045}$ |
| Prithvi-WxC | global | 0.0000 | 0.0000 | 0.0000 |
| Prithvi-WxC | fire-prone | 0.0000 | 0.0000 | 0.0000 |
| Aurora | global | $0.0200_{\pm 0.0267}$ | $9.8520_{\pm 12.9878}$ | $9.8520_{\pm 12.9878}$ |
| Aurora | fire-prone | $0.8203_{\pm 1.8341}$ | $14.3919_{\pm 32.1219}$ | $14.3919_{\pm 32.1219}$ |
| ClimaX | global | $0.0003_{\pm 0.0004}$ | $0.1296_{\pm 0.1775}$ | $0.1296_{\pm 0.1775}$ |
| ClimaX | fire-prone | 0.0000 | 0.0000 | 0.0000 |
| StormCast | global | 0.0000 | 0.0000 | 0.0000 |
| StormCast | fire-prone | 0.0000 | 0.0000 | 0.0000 |
| DLWP | global | 0.0000 | 0.0000 | 0.0000 |
| DLWP | fire-prone | $0.0770_{\pm 0.1100}$ | $4.3266_{\pm 4.3323}$ | $4.3266_{\pm 4.3323}$ |
| FCN | global | 0.0000 | 0.0000 | 0.0000 |
| FCN | fire-prone | $0.0006_{\pm 0.0013}$ | $1.1680_{\pm 1.9872}$ | $1.1680_{\pm 1.9872}$ |
| FengWu | global | 0.0000 | 0.0000 | 0.0000 |
| FengWu | fire-prone | $0.0691_{\pm 0.1191}$ | $0.5222_{\pm 0.6239}$ | $0.5222_{\pm 0.6239}$ |
| FuXi | global | 0.0000 | 0.0000 | 0.0000 |
| FuXi | fire-prone | 0.0000 | $0.1084_{\pm 0.1729}$ | $0.1084_{\pm 0.1729}$ |
| Pangu-Weather | global | 0.0000 | 0.0000 | 0.0000 |
| Pangu-Weather | fire-prone | $0.0728_{\pm 0.1179}$ | $0.1849_{\pm 0.3263}$ | $0.1849_{\pm 0.3263}$ |
| AlphaEarth | global | 0.0000 | $17.2217_{\pm 8.8492}$ | $17.2217_{\pm 8.8492}$ |
| AlphaEarth | fire-prone | 0.0000 | $3.8804_{\pm 5.9483}$ | $3.8804_{\pm 5.9483}$ |

Table 8: For fixed occupancy $\mathcal{T}$, this table reports predicted-positive rate. Values are percentages under the same validation-selected strict threshold. Scopes $\Omega$ are fixed before test scoring; cells report five-seed mean with std in small type.

| **Backbone** | $\Omega =$**global** | $\Omega =$**top 5%** | $\Omega =$**top 10%** | $\Omega =$**top 20%** |
|---|---|---|---|---|
| WILD**FIRE**-FM | $1.6808_{\pm 0.3684}$ | $3.0619_{\pm 1.0925}$ | $1.5310_{\pm 0.5463}$ | $0.7655_{\pm 0.2732}$ |
| Prithvi-WxC | $61.9711_{\pm 30.9101}$ | $57.4117_{\pm 47.8987}$ | $58.4565_{\pm 51.0897}$ | $58.9788_{\pm 52.6991}$ |
| Aurora | $55.5849_{\pm 19.7524}$ | $57.2238_{\pm 35.3400}$ | $68.7942_{\pm 37.6958}$ | $67.2891_{\pm 38.3991}$ |
| ClimaX | $5.6763_{\pm 3.9261}$ | $24.0091_{\pm 9.2816}$ | $11.8450_{\pm 4.5067}$ | $5.7442_{\pm 4.1341}$ |
| StormCast | $60.6507_{\pm 17.4895}$ | $57.6017_{\pm 35.2921}$ | $68.0766_{\pm 37.3899}$ | $67.8397_{\pm 39.2410}$ |
| DLWP | $4.3221_{\pm 1.5619}$ | $9.4001_{\pm 5.0807}$ | $4.9700_{\pm 3.6849}$ | $1.9198_{\pm 1.4678}$ |
| FCN | $1.5202_{\pm 1.3446}$ | $4.7856_{\pm 2.9409}$ | $2.7257_{\pm 1.6353}$ | $0.8368_{\pm 0.2358}$ |
| FengWu | $0.4277_{\pm 0.4830}$ | $0.6004_{\pm 0.3041}$ | $0.2609_{\pm 0.1935}$ | $0.1501_{\pm 0.1206}$ |
| FuXi | $0.4505_{\pm 0.2773}$ | $2.9315_{\pm 2.6392}$ | $0.5197_{\pm 0.6074}$ | $0.3621_{\pm 0.4346}$ |
| Pangu-Weather | $1.0801_{\pm 1.1308}$ | $2.0549_{\pm 2.1893}$ | $1.4029_{\pm 1.4739}$ | $1.0103_{\pm 1.1084}$ |
| AlphaEarth | $0.0691_{\pm 0.0499}$ | $0.2826_{\pm 0.1497}$ | $0.1524_{\pm 0.0770}$ | $0.0656_{\pm 0.0414}$ |

Table 9: For fixed spread $\mathcal{T}$ and strict $\Lambda$, this table reports AP under three $\Omega$ scopes: full test, top-5% train-fire area, and top-10% train-fire area. Values are percentages; cells report mean with small std.

| Backbone | full $\Omega$ AP | top-5% $\Omega$ AP | top-10% $\Omega$ AP |
|---|---|---|---|
| WILD**FIRE**-FM | $30.0197_{\pm 1.5651}$ | $40.7452_{\pm 2.0542}$ | $37.4096_{\pm 1.8731}$ |
| Prithvi-WxC | $4.8319_{\pm 0.1731}$ | $12.6086_{\pm 0.4468}$ | $8.7051_{\pm 0.1889}$ |
| Aurora | $17.7723_{\pm 0.4293}$ | $30.3106_{\pm 0.9404}$ | $26.4732_{\pm 0.6932}$ |
| ClimaX | $11.1726_{\pm 0.2337}$ | $25.7871_{\pm 1.2896}$ | $19.9977_{\pm 1.2217}$ |
| StormCast | $8.1147_{\pm 1.1569}$ | $18.5461_{\pm 1.1727}$ | $14.1286_{\pm 1.2956}$ |
| DLWP | $9.2142_{\pm 2.6587}$ | $19.3346_{\pm 2.3922}$ | $14.9788_{\pm 2.6696}$ |
| FCN | $6.6774_{\pm 1.3001}$ | $16.7396_{\pm 3.2955}$ | $11.9308_{\pm 2.3881}$ |
| FengWu | $11.0046_{\pm 2.7092}$ | $21.1506_{\pm 1.2163}$ | $17.0113_{\pm 1.5778}$ |
| FuXi | $13.5507_{\pm 0.3840}$ | $22.5434_{\pm 0.4100}$ | $19.1964_{\pm 0.3943}$ |
| Pangu-Weather | $10.6250_{\pm 1.4643}$ | $19.8294_{\pm 1.3044}$ | $15.8013_{\pm 1.1602}$ |
| AlphaEarth | $12.2847_{\pm 1.3562}$ | $22.8692_{\pm 0.4915}$ | $18.2992_{\pm 1.2110}$ |

Table 10: For fixed final-area $\mathcal{T}$ and $\Omega$, this table reports median log error and acre-scale errors in addition to the main log-RMSE/log-MAE/Spearman metrics. Cells report mean with small std.

| Backbone | log median AE | acre median AE | acre MAPE |
|---|---|---|---|
| WILD**FIRE**-FM | $1.0235_{\pm 0.0982}$ | $4504.0692_{\pm 459.0483}$ | $1.4525_{\pm 0.0254}$ |
| Prithvi-WxC | $1.2184_{\pm 0.2107}$ | $5375.8770_{\pm 788.7906}$ | $1.9517_{\pm 0.2875}$ |
| Aurora | $1.4547_{\pm 0.0301}$ | $9904.9483_{\pm 457.4260}$ | $6.8728_{\pm 3.0026}$ |
| ClimaX | $1.6841_{\pm 0.1818}$ | $18130.4820_{\pm 3248.3873}$ | $8.2373_{\pm 2.8540}$ |
| StormCast | $1.4522_{\pm 0.1519}$ | $11155.7881_{\pm 2020.8656}$ | $4.6142_{\pm 1.1500}$ |
| DLWP | $1.0952_{\pm 0.1306}$ | $4406.9315_{\pm 303.0944}$ | $1.7357_{\pm 0.3625}$ |
| FCN | $1.1688_{\pm 0.1139}$ | $5166.9993_{\pm 213.0333}$ | $2.0800_{\pm 0.4004}$ |
| FengWu | $1.1589_{\pm 0.1772}$ | $5137.2822_{\pm 628.7543}$ | $2.0944_{\pm 0.4545}$ |
| FuXi | $1.1855_{\pm 0.0612}$ | $5697.7117_{\pm 796.8785}$ | $2.4411_{\pm 0.5567}$ |
| Pangu-Weather | $1.1221_{\pm 0.1470}$ | $5092.3621_{\pm 483.8243}$ | $1.9571_{\pm 0.3113}$ |
| AlphaEarth | $1.7459_{\pm 0.6057}$ | $15110.7573_{\pm 7106.3417}$ | $9.7398_{\pm 2.7425}$ |

Table 11: For fixed retrieval $\mathcal{T}$ and $\Omega$, this table reports nDCG@5, best log gap, and rank $\rho$ in addition to the main nDCG@10/log-error metrics. Cells report mean with small std.

| Backbone | nDCG@5 | best log gap | rank $\rho$ |
|---|---|---|---|
| WILD**FIRE**-FM | $0.5175_{\pm 0.0445}$ | $0.1868_{\pm 0.0285}$ | $0.6019_{\pm 0.1460}$ |
| Prithvi-WxC | $0.3591_{\pm 0.0107}$ | $0.2151_{\pm 0.0594}$ | $0.1514_{\pm 0.1489}$ |
| Aurora | $0.4423_{\pm 0.0210}$ | $0.1551_{\pm 0.0437}$ | $0.2162_{\pm 0.1856}$ |
| ClimaX | $0.4151_{\pm 0.0293}$ | $0.2129_{\pm 0.0653}$ | $0.1587_{\pm 0.2831}$ |
| StormCast | $0.3960_{\pm 0.0240}$ | $0.1714_{\pm 0.0310}$ | $0.1258_{\pm 0.1625}$ |
| DLWP | $0.3795_{\pm 0.0274}$ | $0.1944_{\pm 0.0807}$ | $-0.3865_{\pm 0.2802}$ |
| FCN | $0.4250_{\pm 0.0112}$ | $0.1856_{\pm 0.0846}$ | $-0.1357_{\pm 0.2571}$ |
| FengWu | $0.4228_{\pm 0.0310}$ | $0.1870_{\pm 0.0858}$ | $-0.1926_{\pm 0.2194}$ |
| FuXi | $0.4544_{\pm 0.0356}$ | $0.2171_{\pm 0.0806}$ | $-0.1367_{\pm 0.2885}$ |
| Pangu-Weather | $0.3988_{\pm 0.0506}$ | $0.1901_{\pm 0.0838}$ | $-0.1970_{\pm 0.2216}$ |
| AlphaEarth | $0.5276_{\pm 0.0531}$ | $0.1782_{\pm 0.0454}$ | $0.4639_{\pm 0.2802}$ |

Table 12: For fixed smoke $\mathcal{T}$ and station $\Omega$, this table reports RMSE, MAE, and 90th-percentile absolute error on test rows with observed $PM_{2.5} \geq 35$; std uses a row bootstrap over those rows. Cells report mean with small std.

| Backbone | high-smoke RMSE | high-smoke MAE | high-smoke 90th AE |
|---|---|---|---|
| WILD**FIRE**-FM | $47.4870_{\pm 0.6346}$ | $34.3954_{\pm 0.7654}$ | $65.6213_{\pm 3.8778}$ |
| Prithvi-WxC | $57.2224_{\pm 1.7268}$ | $47.3871_{\pm 0.3153}$ | $74.9666_{\pm 3.2381}$ |
| Aurora | $57.2752_{\pm 1.7248}$ | $47.4368_{\pm 0.3149}$ | $75.0755_{\pm 3.1074}$ |
| ClimaX | $57.2828_{\pm 1.7239}$ | $47.4407_{\pm 0.3140}$ | $75.1012_{\pm 3.0777}$ |
| StormCast | $56.6512_{\pm 1.7517}$ | $46.7914_{\pm 0.3281}$ | $74.0794_{\pm 3.4707}$ |
| DLWP | $57.0075_{\pm 1.7359}$ | $47.1971_{\pm 0.3198}$ | $74.4936_{\pm 3.3826}$ |
| FCN | $57.0582_{\pm 1.7339}$ | $47.2401_{\pm 0.3187}$ | $74.6431_{\pm 3.1982}$ |
| FengWu | $57.0158_{\pm 1.7357}$ | $47.1957_{\pm 0.3194}$ | $74.5652_{\pm 3.2871}$ |
| FuXi | $56.9622_{\pm 1.7371}$ | $47.1508_{\pm 0.3201}$ | $74.3278_{\pm 3.4435}$ |
| Pangu-Weather | $57.1282_{\pm 1.7307}$ | $47.3050_{\pm 0.3170}$ | $74.6830_{\pm 3.2375}$ |
| AlphaEarth | $48.0665_{\pm 0.7904}$ | $35.6088_{\pm 0.7341}$ | $66.7613_{\pm 3.9235}$ |

Table 13: For fixed heat $\mathcal{T}$ and heat-region $\Omega$, this table reports precision and recall for the exceedance label used by the main $F_1$. Cells report mean with small std.

| Backbone | precision | recall |
|---|---|---|
| WILD**FIRE**-FM | $0.9767_{\pm 0.0117}$ | $0.9330_{\pm 0.0299}$ |
| Prithvi-WxC | $0.8260_{\pm 0.0030}$ | $0.9173_{\pm 0.0033}$ |
| Aurora | $0.5920_{\pm 0.0347}$ | $0.0517_{\pm 0.0020}$ |
| ClimaX | $0.7397_{\pm 0.0099}$ | $0.7994_{\pm 0.0051}$ |
| StormCast | $0.8840_{\pm 0.0237}$ | $0.9320_{\pm 0.0165}$ |
| DLWP | $0.9429_{\pm 0.0085}$ | $0.8899_{\pm 0.0167}$ |
| FCN | $0.9408_{\pm 0.0097}$ | $0.9111_{\pm 0.0127}$ |
| FengWu | $0.3808_{\pm 0.2719}$ | $0.0266_{\pm 0.0267}$ |
| FuXi | $0.3262_{\pm 0.1262}$ | $0.1810_{\pm 0.0481}$ |
| Pangu-Weather | $0.1159_{\pm 0.0743}$ | $0.0112_{\pm 0.0032}$ |
| AlphaEarth | $0.9824_{\pm 0.0040}$ | $0.9278_{\pm 0.0178}$ |

## C Comparator Eligibility Notes

All numeric comparator rows in Tables 1 and 2 are included only after the task form, metric, matching rule, scope, and head family are fixed. The appendix does not repeat those full matrices. The key eligibility rule is simple: reported rows satisfy the same contract as the row block in which they appear, while excluded rows are excluded because their representation or output form does not satisfy that contract.

**Reading rule.** Exact-only, tolerated, union, ranking, retrieval, and regression scores answer different questions. The fixed-contract reading is therefore to compare entries only within one row block and not to average across task forms.

# D Seeded Audits

## D.1 Seed Robustness Summary

Table 14 summarizes stochastic checks used to support the reported mean-with-std convention. It is not a replacement for the main fixed-contract result tables.

Table 14: Seed summaries for stochastic checks. Values report mean with small std over completed seeds.

| $\mathcal{T}$ check | Seeds | Primary value | Other value(s) | Reading |
|---|---|---|---|---|
| Final burned area | 5 | log-RMSE $1.1657_{\pm 0.0126}$ | log-MAE $1.0423_{\pm 0.0081}$; Spear. $0.6298_{\pm 0.0338}$ | stable across seeds |
| Smoke $PM_{2.5}$ | 5 | RMSE $4.4646_{\pm 0.0060}$ | MAE $2.4108_{\pm 0.0016}$; $r$ $0.6368_{\pm 0.0013}$ | stable at table precision |
| Extreme heat | 5 | RMSE-C $0.2179_{\pm 0.0043}$ | MAE-C $0.1787_{\pm 0.0018}$; exceed. $F_1$ $0.9541_{\pm 0.0164}$ | stable across seeds |
| Fire spread | 5 | exact $F_1$ $37.6700_{\pm 0.9800}$ | spatial $F_1$ $80.9700_{\pm 2.0200}$; AP $30.0900_{\pm 1.2500}$ | stable across seeds |
| Aurora paired-head check | 5 | fire-prone score diff. $6.3500_{\pm 13.2800}$ | PR-AUC and union choices differ in 2/5 seeds | variable across seeds |

# E Lightweight Head and Adaptation Details

All frozen-transfer comparisons use the same five lightweight head architectures applied on top of the frozen backbone representations. Table 15 summarises each head family, its architecture, approximate parameter count, and the adaptation procedure used.

Table 15: Lightweight head architectures used in the fixed-contract transfer comparisons. All heads are trained from random initialisation on the frozen backbone features. Parameter counts are approximate and depend on the feature dimensionality of each backbone.

| $\mathcal{A}$ **head** | **Architecture** | **Approx. params** | **Notes** |
|---|---|---|---|
| Constant prior | Outputs a fixed bias vector, ignoring input features. | Output dimension only | Provides a degenerate baseline; selected when backbone features carry no useful signal. |
| Linear probe | Single linear layer mapping backbone features to output. No nonlinearity. | $d \times c + c$ | Standard frozen-representation baseline. |
| Pixel MLP | Two-layer MLP applied independently per spatial unit. | $d \times h + h \times c$ | Captures per-pixel nonlinearity; ignores spatial context. |
| Shallow adapter | Two-layer MLP with a spatial context window; uses $3 \times 3$ convolution before the linear output. | $9dh + hc$ | Balances local spatial context with parameter efficiency. |
| Wide adapter | Shallow adapter with wider hidden dimension. | $9dH + Hc$ | Higher capacity variant; can overfit on small fire-event sets. |

**Training protocol.** Each occupancy head-control run uses seeds $\{1, 7, 42, 99, 123\}$, the five heads listed above, and the fixed variants identity, erode-r1, and close-r1. The spread U-Net reference is trained for 4 epochs. The threshold $\tau$ is selected on the validation split by maximising union-$F_1$ (for occupancy) or spatial $F_1$ (for spread) and held fixed at test time. Morphology parameters (spatial tolerance $k$, temporal tolerance $\Delta t$) are fixed as part of the evaluation contract and are not tuned after validation.

**Head selection procedure.** For each (feature source, scope, seed) tuple, all five heads are trained independently. The PR-AUC-based selector picks $h_R = \arg\max_{h \in \mathcal{H}} R(h)$ on the validation set; the decision-based selector picks $h_D = \arg\max_{h \in \mathcal{H}} D(h)$ on the same set. The selection regret $\delta = D(h_D) - D(h_R) \geq 0$ is computed on the held-out test set.

# F Limitations

The conclusions apply to the task forms, scopes, evaluation rules, and comparator eligibility decisions used in this study. The evaluation covers selected wildfire decision tasks and supporting retrieval and regression task forms. Comparator eligibility is fixed before metric values are interpreted. This eligibility rule keeps each comparison within one task-form contract. It also leaves some model and task pairs outside the evaluated comparison set by design.

The transfer comparison uses frozen backbones with lightweight heads. The results therefore describe frozen-backbone transfer under the allowed head families in each contract. Full fine-tuning, alternative adaptation procedures, and broader head families are outside the evaluated scope. The task-specific reference baselines serve as empirical anchors for same-contract comparison. WILD**FIRE**-FM is a regional wildfire reference for the reported California fixed-contract experiments.

The supporting retrieval and regression checks bound the primary spatial decision claim. They provide task-form evidence rather than a single score across all wildfire-related prediction tasks. The analysis focuses on the reported metric families, matching rules, and fixed comparison choices. Operational response rules, intervention costs, and deployment policies are part of wildfire early-warning use contexts [10; 29; 7]. They are outside the scope of this evaluation study and are not inferred from the reported scores.

# G Reproducibility and Evaluation Artifacts

## G.1 External Assets and Terms of Use

We use external datasets and model assets only for research evaluation. Access to each asset follows the original provider's portal, license, or terms of use; this submission does not imply that all assets are openly redistributable. We do not redistribute raw external datasets, provider-hosted embeddings, or third-party model weights. Table 16 records the source and terms-of-use status used to interpret reproducibility.

Table 16: External assets used by the study and their source or terms-of-use status.

| Asset family | Use in this study | Source and terms-of-use note |
|---|---|---|
| NOAA HRRR fields [24; 25] | Dynamic weather inputs for WILD**FIRE**-FM and transfer tasks. | NOAA provider terms and citation requirements apply. |
| NASA FIRMS [21] | Active-fire occupancy supervision. | NASA Earthdata/FIRMS access terms and citation requirements apply. |
| LANDFIRE and WRC layers [16; 17; 39] | Static fuel, canopy, and exposure context. | Original geospatial-product provider terms and citations apply. |
| LandScan [26] | Static population context. | ORNL/LandScan source-specific access terms apply; raw data are not redistributed. |
| WFIGS and MTBS [22; 38] | Event-level resources for burned-area and analog tasks. | Original incident/perimeter-product provider terms and citations apply. |
| External Earth-FM baselines [35; 2; 23; 27; 40; 28; 4; 5; 1; 3] | Frozen comparator representations or task-model baselines. | Original model-provider licenses and access terms apply; third-party weights are not redistributed. |

This note supports the NeurIPS checklist and identifies the files that support the reported claims. This file statement does not imply full raw-data release. The main claims can be checked from the manuscript contracts, metric definitions, and per-head result files, even if full raw-data release is delayed or limited. Sections 3 and 4 specify the contract components used by the main claims: task definition, split logic, label space, tolerance parameters, scope definitions, threshold or operating-point rules, and lightweight-head set.

The supplementary source includes the check scripts, per-head and per-seed CSV result files, and LaTeX result tables for the expanded check and matching-rule support. These files expose exact $F_1$, tolerated $F_1$, union-$F_1$, PR-AUC, per-head selection, top-1 agreement, and selection-regret arithmetic. The manuscript also includes full figure and table reproduction values in result tables and appendix tables. These files provide a runnable check of the selection-regret arithmetic and the table-construction logic from fixed per-head rows. The seeded occupancy check uses seeds $\{1, 7, 42, 99, 123\}$, and the spread task-specific U-Net check uses repeated seeds; reported error bars are standard deviations over the completed seeded runs. Full raw wildfire inputs and large feature arrays are not released at submission because redistribution and storage constraints require a separate review.

For stochastic results, the paper reports mean with standard deviation over repeated seeds. For fixed-output or fixed-feature controls, the table uses one fixed output or feature set; the changed item is the matching rule or selection metric.

The reported experiments use two resource classes on a shared Slurm-managed cluster. Tabular retrieval/regression checks and same-feature head controls run on CPU workers with 4 to 8 cores, 24 to 64 GB host memory, and 2 to 4 hour wall-clock limits. Spread U-Net training and threshold calibration run on single-GPU jobs with one B200 GPU, 8 CPU cores, 96 GB host memory, and a 4 hour wall-clock limit. The seed/check waves reported in the appendix correspond to roughly 78 CPU job-hours and 12 GPU job-hours of scheduled wall-clock budget; exploratory runs are not included in the reported compute accounting.

The raw-data limitation is separate from the selection-regret files. The supplementary source is sufficient to inspect the selection-regret arithmetic and reproduce the reported tables. Full end-to-end

recomputation from raw wildfire inputs is not included at submission because redistribution review is still required. The broader impact is evaluation-facing rather than operational. Better reading of wildfire transfer evidence can reduce overconfident benchmark claims, while misread transfer results could still encourage inappropriate reliance on models with low decision scores. For that reason, the paper keeps its claims wildfire-centered, decision-task specific, and explicitly separate from any predictive deployment recommendation.