DOES YOUR WILDFIRE PREDICTION MODEL ACTUALLY WORK, OR JUST SCORE WELL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Earth foundation models can be transferred to wildfire prediction, but their reported downstream scores depend on the evaluation choices used to compute them. We formulate wildfire transfer as a fixed-contract evaluation problem and test how evaluation choices affect scores and model rankings. Holding predictions fixed, decision F_1 changes from 5.31% to 66.74% solely with the matching rule; holding frozen features fixed, changing the head-selection metric changes decision F_1 by up to 88.10 percentage points. We introduce the contract dominance ratio and rank instability rate, showing that evaluation choices affect scores more than backbone choice and reverse the relative ranking of up to 46% of model pairs across configurations. We develop WILDFIRE-FM to provide reference performance across six downstream tasks under specified contracts. Under matched contracts, this wildfire reference leads the evaluated frozen backbones on occupancy, spread, and burned-area prediction. The reference provides a baseline for comparing Earth-FM transfer across tasks, while the controlled checks show why these comparisons require explicit evaluation contracts. See implementation details at: https://anonymous.4open.science/r/Wildfire-fm-evaluation-contracts-5AE9.

1 INTRODUCTION

Wildfire prediction supports early warning and resource allocation in disaster response (Goldammer, 1999; Farahmand et al., 2020). However, developing task-specific models is difficult because labeled fire events are sparse relative to the spatial and temporal extent of the prediction domain (Huot et al., 2022; Farahmand et al., 2020). Earth foundation models (Earth FMs) offer a promising alternative. Pretrained on large-scale geophysical data, they learn representations of atmospheric and land-surface dynamics that can be transferred to downstream environmental tasks (Bodnar et al., 2025; Schmude et al., 2024; Nguyen et al., 2023). Their potential for wildfire prediction depends on whether these pretrained representations transfer reliably to fire-related tasks.

To fully evaluate this transfer, a common way is to freeze a pretrained backbone, train a lightweight head for a downstream task, and rank the resulting representations by their task-specific scores. Geospatial benchmarks standardize parts of this process by fixing datasets, splits, and metrics (Lacoste et al., 2023; Rasp et al., 2024; Koh et al., 2021; Yeh et al., 2021), while wildfire benchmarks such as FireCast and Next Day Wildfire Spread define task-specific prediction settings (Radke et al., 2019; Huot et al., 2022). Under this evaluation, a higher downstream score is interpreted as evidence that one pretrained Earth-FM representation transfers better than another.

Nevertheless, this interpretation is valid only if the evaluation choices used to produce the scores are held fixed. For wildfire prediction, the same prediction may be counted as correct or incorrect depending on the spatial and temporal matching rule. The same frozen representation may yield a different downstream result depending on which metric selects its prediction head, while changing the evaluation scope from the full domain to fire-prone regions may alter the relative ordering of backbones. Moreover, occupancy, spread, retrieval, and regression define transfer through different prediction units and metric families. Thus, a reported score is jointly determined by the foundation model and its evaluation contract. Without fixing that contract, score differences and model rankings cannot be attributed solely to Earth-FM transfer quality. Motivated by this attribution gap, we ask:

Can downstream wildfire scores reliably identify which Earth foundation representations transfer better without first fixing the evaluation contract that produces them?

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.

Answering this question requires addressing three challenges.

Challenge 1: Spatial verification sensitivity. Spatial forecast verification has long shown that apparent forecast quality depends on the spatial and temporal tolerance used to match predictions with observations (Ebert, 2009; Gilleland et al., 2009). Furthermore, metric studies show that class imbalance, operating thresholds, and the intended decision can substantially change how the same predictions are evaluated (Lipton et al., 2014; McDermott et al., 2024; Maier-Hein et al., 2024). However, these effects have not been excluded in wildfire Earth-FM transfer while holding either the model outputs or the frozen representations fixed.

Challenge 2: The best Earth FM also depends on the evaluation contract. Prior studies show that benchmark conclusions can change with the selected tasks, hyperparameters, data samples, and evaluation metrics (Bouthillier et al., 2021; Dehghani et al., 2021; Traub et al., 2024). By comparison, existing geospatial benchmarks improve comparability by standardizing a particular evaluation protocol (Lacoste et al., 2023; Marsocci et al., 2024; Simumba et al., 2025), but they do not quantify whether evaluation configuration or backbone identity contributes more to reported wildfire performance, or how often reasonable configurations reverse pairwise Earth-FM rankings.

Challenge 3: Task-dependent transfer. Wildfire occupancy, spread, retrieval, and regression differ in their targets, spatial units, output structures, and metric families (Zamir et al., 2018; Dehghani et al., 2021). A representation that preserves information useful for cell-level fire occurrence may not preserve the event-level or continuous information needed for burned-area regression, retrieval, smoke, or heat prediction. Recent geospatial foundation-model benchmarks likewise show that no single model performs best across all downstream tasks (Marsocci et al., 2024; Simumba et al., 2025). A strong result on one task therefore does not establish general wildfire transfer. The open question is whether the same Earth-FM representation supports consistent conclusions across wildfire task forms, or whether each task supports a narrower claim.

We address these challenges through a fixed-contract evaluation framework. The framework fixes the task, metric, matching rule, evaluation scope, and head family before a model comparison. Fixed-output and fixed-feature checks isolate evaluation effects, while CDR and RIR quantify their influence on Earth-FM scores and rankings. We then develop WILDFIRE-FM, a wildfire-specific foundation model, to provide reference performance across six downstream task forms. We refer to this model as the wildfire reference. Its performance under each contract provides a baseline for comparing general-purpose Earth FMs across wildfire tasks. The paper follows this sequence. Section 2 explains the attribution and reference gaps, and Section 3 confirms their practical effects through fixed-output and fixed-feature checks. Section 4 defines the evaluation contracts, sensitivity measures, wildfire reference, and experimental design. Section 5 evaluates score and ranking sensitivity, transfer across primary and supporting tasks, and robustness across years and regions.

Our contributions are as follows:

- **A novel framework for reliable Earth-FM evaluation.** We formulate wildfire transfer as a fixed-contract evaluation problem and introduce fixed-output and fixed-feature checks together with CDR and RIR. These tools separate changes caused by the foundation-model backbone from changes caused by matching, head selection, and evaluation scope.
- **A wildfire foundation model providing reference performance.** We develop WILDFIRE-FM to provide reference performance across six downstream task forms. The same pretrained backbone is transferred through lightweight task-specific heads. Its results provide a baseline for evaluating general-purpose Earth FMs under matched contracts.
- **A unified benchmark of Earth-FM transfer.** We compare general-purpose Earth FMs with the wildfire reference across six task forms. The reference leads on occupancy, spread, and burned-area prediction, while AlphaEarth approaches it on retrieval, smoke, and extreme heat. These comparisons show that transfer performance varies across downstream tasks and evaluation contracts, and a model's rank within the comparator set also provides limited information about its performance relative to a representation learned from wildfire data.

2 PRELIMINARIES

2.1 ATTRIBUTION IN FROZEN-BACKBONE EVALUATION

Head selection. A frozen-backbone evaluation does not map a representation directly to a reported score. Let f_b denote backbone b, and let $Z_b = f_b(X)$ be its frozen representation of the downstream

inputs X. A lightweight head $h \in \mathcal{A}$ maps this representation to a prediction-score field $\mathrm{s}_{b,h} = h(Z_b)$, where $\mathcal{A}$ is the candidate-head family. The validation component of the metric protocol M selects one of these heads, which we denote by $h^*_{b,M}$. The prediction evaluated at test time is determined jointly by the frozen representation, the candidate-head family, and the criterion used to select a head.

This selection step matters because commonly used validation metrics reward different properties of a predictor. PR-AUC evaluates ranking quality across operating thresholds, whereas decision F_1 evaluates performance at a selected operating point (Lipton et al., 2014; McDermott et al., 2024; Maier-Hein et al., 2024). Consequently, two metric protocols may select different heads from the same candidate family, $h^*_{b,M_{\mathrm{PR}}} \neq h^*_{b,M_{\mathrm{F1}}}$, even though f_b, Z_b, the downstream data, and $\mathcal{A}$ are unchanged. The test score assigned to a frozen representation depends not only on the information contained in Z_b, but also on which candidate predictor is selected to expose that information.

Matching predictions with observations. A second attribution step occurs when the selected prediction is compared with the observations. After applying the validation-selected threshold τ, let $\widehat{P}_\tau$ denote the predicted event set and P the observed event set within evaluation scope Ω. A matching rule Λ determines which prediction–observation pairs count as matches. The decision score is the harmonic mean of matched-prediction precision and matched-observation recall. Exact verification requires agreement at the same cell and time, whereas spatial or temporal tolerance can credit a displaced prediction (Ebert, 2009; Gilleland et al., 2009). Holding $\widehat{P}_\tau$, P, and τ fixed leaves $F_{1,\Lambda}$ conditional on the rule used to define correspondence.

Controlled attribution. Prescribing one head-selection criterion and one matching rule produces a well-defined score under those choices. However, it does not determine how much of that score is attributable to the frozen representation. Establishing this requires controlled comparisons in which the upstream object is held fixed: the predictions for matching-rule comparisons and the frozen representation for head-selection comparisons. We test this in Section 3.

2.2 The Missing Reference in Wildfire Transfer

The required comparison. Frozen-backbone evaluation measures whether information useful for a downstream task can be recovered from a pretrained representation without updating the backbone. Under the notation introduced above, the representation $Z_b = f_b(X)$ remains fixed, and adaptation is restricted to a lightweight head $h \in \mathcal{A}$. The resulting performance characterizes the downstream accessibility of information in Z_b, rather than the performance of a fully fine-tuned or task-specific model, which is central when interpreting the transfer of general-purpose Earth foundation models (Nguyen et al., 2023; Bodnar et al., 2025; Schmude et al., 2024). Existing wildfire predictors do not generally provide a directly comparable reference for this setting. A task-specific predictor may update its complete network, use inputs selected for one wildfire target, or employ a specialized output architecture and training objective (Radke et al., 2019; Huot et al., 2022). Its score reflects both the learned representation and the task-specific optimization used to produce it. Comparing that score with one obtained from a frozen Earth-FM representation would conflate representation quality with differences in trainable capacity, model adaptation, and downstream design.

The limit of relative rankings. Comparisons restricted to general-purpose Earth FMs avoid this particular mismatch, but they provide only a relative ordering within the comparator set. The highest-scoring general-purpose backbone may still omit substantial wildfire-relevant information; its rank alone does not reveal how closely its representation approaches one learned specifically from wildfire-related signals. A domain-specific reference is needed to make this gap observable, but it must be evaluated through the same frozen-representation interface as the general-purpose backbones. Otherwise, an apparent advantage may arise from additional downstream adaptation rather than from the pretrained representation.

A comparable wildfire reference. The reference must also remain comparable across task forms. A wildfire-specific model trained separately for each downstream target would produce a different representation for each comparison, making it impossible to determine whether one pretrained representation transfers consistently. Conversely, a reference established on only one task cannot anchor conclusions for tasks with different prediction units and output structures. Meaningful comparison requires both reliable attribution and a suitable reference. The evaluation must first establish that score differences reflect the frozen representations. Only then can a wildfire-specific frozen representation provide a meaningful reference for general-purpose Earth FMs. Given this, the

next section tests whether these evaluation effects remain material when the predictions or frozen representations are held fixed.

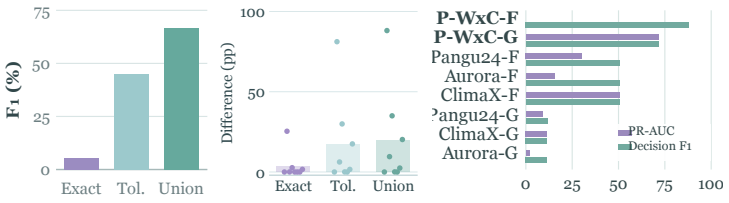
3 Controlled Evidence of Evaluation Sensitivity

We test how much the matching rule and the metric used to choose a downstream head can change a wildfire transfer score when the predictions or frozen features remain unchanged. The two controlled checks change one evaluation choice at a time.

Experimental settings. In the fixed-output check, we apply exact, offset, and union matching to the same occupancy predictions, observations, and validation-selected threshold. In the fixed-feature check, we use either validation PR-AUC or validation decision F_1 to choose a head from the same candidate family for each frozen backbone–scope combination. We call them the PR-AUC-selected head and the decision-F_1-selected head. We evaluate both heads on the same test set with the same threshold-selection and matching rules, and report the test decision-score difference between them. These are two concrete heads chosen from the same candidates, not two different backbones.

Observations and analysis. *Observation 1: The matching rule alone largely changes the reported score.* For the same fixed occupancy predictions, decision F_1 increases from 5.31% under exact matching to 66.74% under union matching, a difference of 61.43 percentage points (Figure 1a). This difference arises entirely from how predicted and observed events are matched. *Observation 2: Changing the head-selection metric largely changes test decision performance.* PR-AUC and decision F_1 frequently select different heads from the same candidate family. Figure 1c orders the eight combinations by the test score of the decision-F_1-selected head. Across these combinations, the resulting test decision-score difference reaches 88.10 percentage points (Figure 1b, Figure 1c). Thus, the same frozen representation can receive substantially different downstream scores depending on whether its head is selected for score ranking across thresholds or for performance at the decision point used downstream.

To summarize, these two checks show that evaluation choices can change individual scores. This leaves a broader question: Are these separate cases, or is evaluation sensitivity an integrated property of Earth-FM comparison? To answer it, the next section introduces fixed evaluation contracts and uses the contract dominance ratio (CDR) and rank instability rate (RIR) to quantify how evaluation contracts affect model differences and pairwise rankings.



(a) Matching rule. (b) Score difference. (c) Different heads.

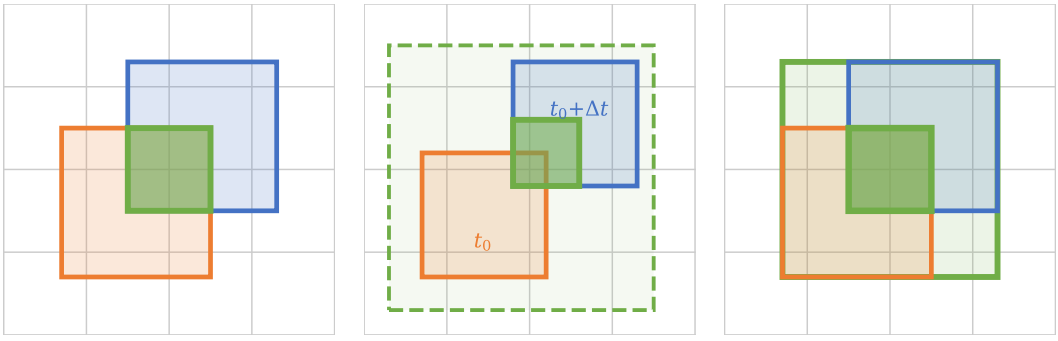
Figure 1: Controlled evidence of evaluation sensitivity. (a) Decision F_1 for the same occupancy predictions under exact, offset, and union matching. (b) Test score differences: dots denote the eight combinations; bars show their means under each matching rule. (c) Test decision F_1 for the head selected by PR-AUC and the head selected by validation decision F_1 across eight backbone–scope combinations. G and F denote global and fire-prone evaluation.

4 Fixed-Contract Evaluation Design

In this section, we define the contracts used to compare Earth FMs and measure evaluation sensitivity. Under each contract, the wildfire reference provides a performance baseline for the comparison.

4.1 Defining the evaluation contract

We define an evaluation contract as $\mathcal{C} = (\mathcal{T}, M, \Lambda, \Omega, \mathcal{A})$. The task $\mathcal{T}$ specifies the dataset, prediction target, and data splits. The metric protocol M specifies the validation metric used to select a head, the test metric used to report performance, and, where required, the threshold-selection rule. The matching rule Λ specifies how predictions correspond to observations. The evaluation scope Ω specifies the spatial units included in evaluation, and $\mathcal{A}$ specifies the family of lightweight heads available to each frozen backbone. A direct comparison between two backbones is made only when all five components are identical. We consider a global scope containing the full spatial domain and a fire-prone scope containing locations with higher historical fire activity. Fire-prone locations are determined using training-period observations and fixed before test evaluation. The two scopes



(a) Exact matching. (b) Tolerated matching. (c) Union matching.

Figure 2: Three matching rules applied to the same prediction and label. Blue and orange denote the prediction and observation; green denotes matched support, and the dashed boundary denotes the tolerance region. The prediction and label remain fixed.

answer different questions: global evaluation measures performance over the complete domain, whereas fire-prone evaluation focuses on locations where fires are more likely to occur. For a frozen backbone f_b, we write Z_b for its representation on the downstream data and $h \in \mathcal{A}$ for a lightweight prediction head. The resulting score field and observations are denoted by $\mathbf{s}_{b,h}$ and Y. For occupancy, a validation-selected threshold τ gives the predicted fire set $\widehat{P}_\tau$, while P denotes the observed fire set. A matching rule Λ determines which predicted and observed events may correspond. We use exact matching for the same cell and time and neighborhood matching for bounded spatial or temporal displacement. Let p_Λ and r_Λ be the fractions of predictions and observations with matching support, respectively. Then $F_{1,\Lambda} = 2p_\Lambda r_\Lambda/(p_\Lambda + r_\Lambda)$, with zero assigned when the denominator is zero. Appendix C.4 specifies the support rules used in each experiment. The metric protocol also fixes how a head is chosen. We train the same candidate family for each frozen representation. Validation PR-AUC chooses the head that best ranks positive events across thresholds, whereas validation decision F_1 chooses the head that performs best at the downstream operating point. We refer to these as the PR-AUC-selected head and the decision-F_1-selected head. A valid comparison fixes the candidate family and the selection metric before test evaluation.

4.2 Comparing Earth FMs under fixed contracts

For each contract $\mathcal{C}$, we evaluate every frozen backbone using the same task data, lightweight-head family, training procedure, selection rule, matching rule, and evaluation scope. Let $S_b(\mathcal{C})$ denote the score of backbone b under this contract. Within a fixed contract, backbone identity is the only component that changes, so differences among $\{S_b(\mathcal{C}) : b \in \mathcal{B}\}$ measure the relative downstream performance of the corresponding frozen representations. For the wildfire reference, $S_{\text{ref}}(\mathcal{C})$ is the performance baseline under contract $\mathcal{C}$. We compare each Earth FM with this baseline under the same contract. We instantiate these comparisons across six task forms. Wildfire occupancy and fire spread are the primary tasks because they directly evaluate fire occurrence and spatial propagation, which are central to early warning and resource allocation (Goldammer, 1999; Farahmand et al., 2020). Final burned area, analog retrieval, smoke $PM_{2.5}$, and extreme heat are supporting tasks that test whether the transfer performance observed on the primary fire tasks also appears on other wildfire-related prediction problems. We summarize the test-side components of the task-form contracts in Table 1. For occupancy, we instantiate separate contracts using the exact, offset, and union matching rules defined above. For fire spread, exact F_1, spatial F_1, and AP are reported separately. The corresponding head-selection metric, threshold rule, and candidate-head family are fixed within each model comparison as specified in Section 4.4.

4.3 Measuring contract effects on scores and rankings

The fixed-contract comparisons allow us to test whether the sensitivity found in Section 3 persists across the full comparator set. For a fixed task and test metric, let $\xi \in \Xi$ denote one configuration of the applicable non-backbone contract choices, such as the head-selection metric, matching rule, and evaluation scope. Evaluating every backbone under every configuration produces a model-by-configuration score matrix $q_b(\xi)$.

Table 1: Task-form contracts used in the fixed-contract comparisons. The first two rows contain the primary-task metrics, and the final four rows contain the supporting-task metrics. Arrows indicate whether higher or lower values are better.

Task ($\mathcal{T}$)	**Metric** (M)	**Matching** (Λ)	**Scope** (Ω)
Occupancy	Exact / tol. / union $F_1 \uparrow$	Exact / tol. / union	Global / fire-prone
Fire spread	Exact / spatial $F_1 \uparrow$; AP$\uparrow$	Raster spatial overlap	Spread region
Final burned area	RMSE / MAE$\downarrow$; Spearman $\rho \uparrow$	Event log-area error	All events
Analog retrieval	nDCG@10$\uparrow$; log-RMSE / log-MAE$\downarrow$	Top-k event match	All events
Smoke $PM_{2.5}$	RMSE / MAE$\downarrow$; Pearson $r \uparrow$	Station absolute error	All stations
Extreme heat	RMSE-C / MAE-C$\downarrow$; exceedance $F_1 \uparrow$	$^\circ$C error / exceedance	Heat region

Contract dominance ratio. We first ask whether score variation is associated more strongly with backbone identity or evaluation configuration. We decompose the score matrix as $q_b(\xi) = \mu + \alpha_b + \beta_\xi + \gamma_{b,\xi}$, where α_b is the backbone main effect, β_ξ is the configuration main effect, and $\gamma_{b,\xi}$ captures configuration-specific backbone behavior. We define the contract dominance ratio as $\mathrm{CDR} = SS_\beta/(SS_\beta + SS_\alpha)$. CDR measures the relative magnitude of the configuration main effect compared with the backbone main effect. A value above 0.5 indicates that the configuration accounts for more main-effect variation than backbone identity. We report SS_γ separately because it measures how differently the backbones respond to the same configuration changes. CDR is computed only among configurations using the same test metric, so all entries in a score matrix are comparable.

Rank instability rate. CDR measures score variation but does not show whether the evaluation configuration changes which backbone performs better. We call a backbone pair (b, b') unstable if there exist configurations $\xi, \xi' \in \Xi$ such that $q_b(\xi) > q_{b'}(\xi)$ but $q_b(\xi') < q_{b'}(\xi')$. The rank instability rate is the fraction of comparator pairs that satisfy this condition. Thus, RIR directly measures how often changing the evaluation configuration reverses a pairwise Earth-FM ranking; ties alone are not counted as reversals. Section 5 applies both measures to the wildfire reference and the general-purpose Earth FM comparators.

4.4 Wildfire reference and experimental design

We pretrain WILDFIRE-FM once on weather, active-fire, land-surface, and static environmental inputs, and freeze the resulting backbone for all six downstream tasks. The same representation is transferred through lightweight task-specific heads, so its task scores provide reference performance rather than task-specific end-to-end models. We compare this reference with Prithvi-WxC, Aurora, ClimaX, StormCast, Pangu24, DLWP, FCN, FengWu, FuXi, Pangu-Weather 24h, Pangu-Weather 6h, and AlphaEarth when its representation is compatible with the task contract. Every within-task comparison fixes the data split, candidate heads, head training, validation selection rule, test metric, matching rule, and evaluation scope.

The primary occupancy and spread experiments use a 5-km California grid in EPSG:5070 and a chronological 2024 fire-season split: June–August for training, September for validation, and October 1–30 for testing, with 368, 120, and 120 six-hourly maps. We train on 32×32 crops constructed from HRRR weather fields, validity masks, and static covariates; the 169,854 training crops contain 24,260 positive and 145,594 non-fire examples. Thresholds and heads are chosen on validation data and evaluated once on the held-out test set. Primary results use seeds $\{1, 7, 42, 99, 123\}$. We define fire-prone scopes only from training-period fire frequency. The supporting tasks use task-specific chronological splits at event, station, or grid level.

Table 2: Primary-task ranks under alternative evaluation rules. Green and coral mark improvement and decline from exact; lower is better.

	Occupancy			**Spread**	
Model	**Ex.**	**Tol.**	**Un.**	**Ex.**	**Sp.**
Wildfire ref.	**1**	**1**	**1**	**1**	**1**
AlphaEarth	2	3	4	12	12
FuXi	3	4	5	4	4
ClimaX	4	2	2	3	3
FCN	5	5	3	11	11
Pangu-W 6h	5	6	6	6	6
FengWu	7	8	8	5	7
DLWP	8	7	7	7	5
Aurora	9	9	9	2	2
StormCast	10	10	10	9	9
Pangu24	11	11	11	10	10
Prithvi-WxC	11	12	12	8	8

For robustness, we add a lower-48 cross-year occupancy split using 2022–2023 for training, 2024 for validation, and 2025 for testing, and a six-region leave-one-region-out evaluation in which training tiles that intersect the held-out region are excluded. We provide the complete implementation and compute record in Appendix C.2, report 95% confidence intervals in Appendix Table 31, and detail the temporal and geographic split definitions in

Table 3: Primary fixed-contract transfer results. Occupancy metrics: exact, tolerated, and union F_1 (%). Fire spread metrics: exact F_1, spatial F_1, and AP (%). Each block fixes $\mathcal{T}$, Λ, Ω, $\mathcal{A}$. **Bold** marks the reference row; underlining marks the best non-reference entry per metric. AlphaEarth fire-spread values use the corrected projected-grid cache under the fixed spread contract.

	Occupancy			Fire spread		
Comparator	**Exact** $F_1 \uparrow$	**Tol.** $F_1 \uparrow$	**Union** $F_1 \uparrow$	**Exact** $F_1 \uparrow$	**Spatial** $F_1 \uparrow$	**AP**$\uparrow$
Wildfire ref.	**5.31 ±1.24**	**44.97 ±2.15**	**66.74 ±7.08**	**37.98 ±1.33**	**82.20 ±0.32**	**32.73 ±0.60**
Prithvi-WxC	0.055 ±0.004	7.16 ±0.66	20.19 ±1.83	22.14 ±0.40	58.18 ±0.14	4.83 ±0.17
Aurora	0.07 ±0.01	8.50 ±1.96	23.10 ±4.94	31.08 ±0.40	71.82 ±0.14	17.77 ±0.43
ClimaX	0.35 ±0.08	29.75 ±3.61	60.15 ±7.59	27.99 ± < 0.01	69.06 ± < 0.01	11.17 ±0.23
StormCast	0.056 ±0.011	8.20 ±2.19	22.38 ±5.43	21.95 ±0.37	57.90 ±0.23	8.11 ±1.16
Pangu24	0.055 ±0.008	8.18 ±1.35	22.06 ±3.73	21.87 ±0.42	57.80 ±0.57	9.95 ±1.60
DLWP	0.17 ±0.04	14.91 ±3.24	28.19 ±6.97	22.26 ±1.41	58.71 ±1.91	9.21 ±2.66
FCN	0.28 ±0.08	19.51 ±3.34	40.06 ±9.37	20.60 ±2.36	57.15 ±1.39	6.68 ±1.30
FengWu	0.26 ±0.08	12.00 ±6.02	24.10 ±13.63	22.50 ±1.78	58.20 ±0.84	11.00 ±2.71
FuXi	0.38 ±0.12	21.03 ±4.82	37.29 ±9.45	25.16 ±0.55	59.57 ±0.48	13.55 ±0.38
Pangu-Weather 6h	0.28 ±0.11	17.09 ±4.05	35.64 ±9.03	22.29 ±0.97	58.43 ±0.83	10.63 ±1.46
AlphaEarth	2.06 ±0.44	29.45 ±6.01	37.43 ±9.95	16.19 ±4.79	43.47 ±10.28	12.28 ±1.36

Appendices E.1 and E.2, including the year assignments and regional exclusion rules used to separate training data from the held-out evaluation data.

5 Experiments

We evaluate occupancy with exact, tolerated, and union decision F_1, and spread with exact F_1, spatial F_1, and AP. For the supporting tasks, we use log-RMSE, log-MAE, rank correlation, nDCG@10, Pearson correlation, and threshold-exceedance F_1 as specified by the fixed task contracts in Table 1. We report primary-task means and standard deviations over five seeds, provide full implementation and numerical-precision details in Appendix C.2, give 95% confidence intervals in Appendix Table 31, and record task-by-backbone coverage in Appendix C.1.

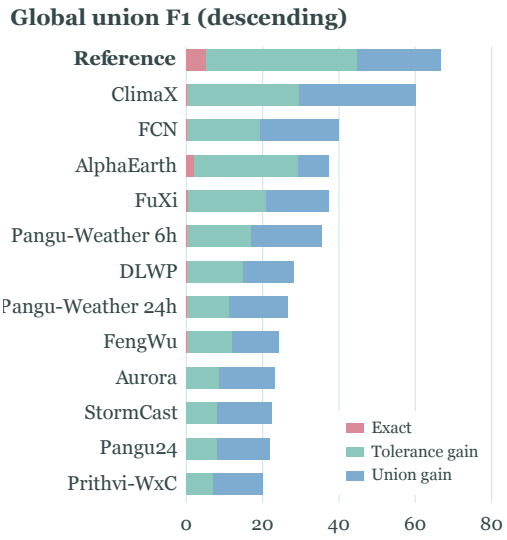


Figure 3: Global occupancy results, ordered by union F_1 (%). The complete scope sweep is reported in Appendix Figure 7.

5.1 Effects of evaluation contracts on scores and rankings (RQ1)

To answer RQ1, we apply CDR and RIR to the occupancy model-by-configuration matrix, summarize global score composition in Figure 3, and report primary-task ranks in Table 2; the matching-only tolerance sweep and all fire-prone scopes appear in the appendix. We make two observations. First, the CDR is 76.71%, so evaluation configuration contributes more main-effect score variation than backbone identity. Second, the RIR is 46.15%, so nearly half of the comparator pairs reverse their winner across configurations. The wildfire reference remains first, but several general-purpose backbones move under alternative occupancy matching rules and spread metrics, and adding scope variation produces more instability than varying matching alone. Thus, comparisons between these backbones depend on both how predictions are matched and which locations are evaluated.

5.2 Primary-task reference performance (RQ2)

To answer RQ2, we compare the wildfire reference with every eligible frozen Earth FM under the fixed occupancy and spread contracts and report the results in Table 3, with the corresponding

	Ref.	WxC	Aurora	ClimaX	Storm	Pangu24	DLWP	FCN	FengWu	FuXi	Pangu6	Alpha
Occupancy Union F1 (%)	**1.00** 66.74	**0.30** 20.19	**0.35** 23.10	**0.90** 60.15	**0.34** 22.38	**0.33** 22.06	**0.42** 28.19	**0.60** 40.06	**0.36** 24.10	**0.56** 37.29	**0.53** 35.64	**0.56** 37.43
Fire spread AP (%)	**1.00** 32.73	**0.15** 4.83	**0.54** 17.77	**0.34** 11.17	**0.25** 8.11	**0.30** 9.95	**0.28** 9.21	**0.20** 6.68	**0.34** 11.00	**0.41** 13.55	**0.32** 10.63	**0.38** 12.28
Burned area log-RMSE	**1.00** 1.16	**0.85** 1.36	**0.62** 1.87	**0.57** 2.03	**0.70** 1.67	**0.84** 1.38	**0.89** 1.31	**0.85** 1.37	**0.85** 1.37	**0.83** 1.41	**0.88** 1.33	**0.47** 2.46
Analog retrieval nDCG@10	**1.00** 0.518	**0.74** 0.386	**0.78** 0.405	**0.80** 0.414	**0.79** 0.408	**0.75** 0.387	**0.77** 0.397	**0.83** 0.432	**0.82** 0.425	**0.83** 0.428	**0.78** 0.402	**0.97** 0.504
Smoke PM2.5 RMSE	**1.00** 4.47	**0.75** 5.93	**0.75** 5.93	**0.75** 5.93	**0.74** 6.02	**0.75** 5.95	**0.75** 5.93	**0.75** 5.93	**0.75** 5.93	**0.75** 5.93	**0.75** 5.93	**1.00** 4.49
Extreme heat RMSE-C	**1.00** 0.218	**0.05** 4.62	**0.01** 18.05	**0.01** 17.65	**0.12** 1.77	**0.04** 4.98	**0.10** 2.27	**0.10** 2.17	**0.10** 2.13	**0.10** 2.13	**0.10** 2.20	**1.00** 0.219

Figure 4: Reference-normalized performance across six fixed task contracts. Cells show normalized score (top) and raw value (bottom); darker shading indicates stronger direction-adjusted performance. Green labels denote higher-is-better metrics and coral labels lower-is-better metrics.

confidence intervals in Appendix Table 31. We make two observations. First, the wildfire reference provides the highest exact, tolerated, and union occupancy F_1, so its advantage persists under strict agreement and bounded displacement, although the strongest general-purpose Earth FM changes with the matching rule. Second, the reference also provides the highest exact F_1, spatial F_1, and AP for fire spread. These complementary metrics establish reference performance for fire occurrence and spatial propagation under their stated contracts.

5.3 Transfer across regression and retrieval tasks (RQ3)

To answer RQ3, we evaluate each backbone separately on burned-area regression, analog retrieval, smoke $PM_{2.5}$, and extreme heat, summarize the direction-adjusted comparison in Figure 4, and report native metrics in Appendix Table 21. We make two observations. First, the wildfire reference provides the strongest reported burned-area performance, so the primary-task advantage extends to an event-level regression task. Second, AlphaEarth approaches the reference on analog retrieval and provides comparable smoke and extreme-heat performance, while the atmospheric backbones remain weaker overall. The gap relative to the reference is therefore task dependent and is clearer for occupancy, spread, and burned area than for the other supporting tasks.

5.4 Robustness across years and regions (RQ4)

To answer RQ4, we repeat occupancy evaluation with the lower-48 cross-year split and the six-region leave-one-region-out design introduced in Section 4.4; we report full results in Appendices E.1 and E.2. We show representative weather inputs, predicted probabilities, targets, and tolerated prediction coverage in Figure 5. We make two observations. First, exact, tolerated, and union matching remain widely separated across years, with an 87.31 percentage-point difference between union and exact F_1, so matching sensitivity is not limited to the California 2024 evaluation. Second, AUROC remains above 0.79 in every held-out region, while fixed-threshold recall varies across regions. The wildfire representation therefore retains ranking performance under geographic holdout, but its operating point remains sensitive to regional calibration.

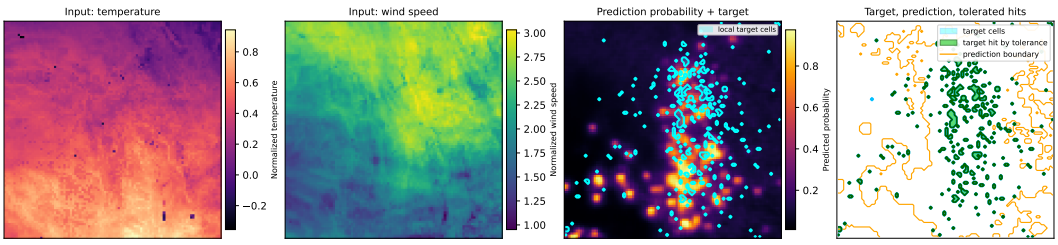


Figure 5: Representative exact-versus-tolerated case from the lower-48 cross-year check. From left: normalized temperature, normalized wind speed, predicted probability with target cells in cyan, and targets with the thresholded prediction boundary in orange and tolerated hits in green. The spatial displacement receives little exact credit but remains matched under the tolerated rule.

6 Related Work

Evaluation choices and model-comparison. Meteorological verification has long shown that spatial tolerance changes whether a forecast is counted as correct (Ebert, 2009; Gilleland et al., 2009). Classification studies likewise show that F_1 depends on the operating threshold and that AUROC and AUPRC behave differently under class imbalance (Lipton et al., 2014; McDermott et al., 2024). Metrics Reloaded connects metric selection to the prediction problem and intended use (Maier-Hein et al., 2024), while broader studies show that randomness, hyperparameter choices, benchmark tasks, and evaluation protocols can change model comparisons (Bouthillier et al., 2021; Dehghani et al., 2021; Lüth et al., 2023; Traub et al., 2024). These studies identify individual sources of evaluation sensitivity but do not provide a controlled analysis of their effects on wildfire Earth-FM transfer. In comparison, we introduce controlled checks, CDR, and RIR to isolate and quantify how evaluation choices change wildfire Earth-FM scores and rankings.

Earth foundation models and wildfire-specific transfer. Data-driven weather models such as DLWP, FourCastNet, GraphCast, Pangu-Weather, FengWu, FuXi, and StormCast learn large-scale atmospheric dynamics for forecasting (Weyn et al., 2020; Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023; Chen et al., 2023a;b; Pathak et al., 2026). ClimaX, Aurora, and Prithvi-WxC support adaptation across weather and climate tasks (Nguyen et al., 2023; Bodnar et al., 2025; Schmude et al., 2024). Remote-sensing models including SatMAE, Scale-MAE, SkySense, and AlphaEarth learn representations from satellite observations for downstream Earth monitoring (Cong et al., 2022; Reed et al., 2023; Guo et al., 2024; Brown et al., 2025). These models are trained for general atmospheric or Earth-observation objectives, while we pretrain one wildfire-specific backbone from weather, active-fire, land-surface, and static environmental inputs and transfer its frozen representation through lightweight heads. Unlike a task-specific wildfire predictor, the same backbone is evaluated across occurrence, spread, retrieval, and environmental-impact tasks. Its performance provides a wildfire reference for comparing general-purpose Earth FMs under matched contracts.

Wildfire tasks and geospatial benchmarks. FireCast and Next Day Wildfire Spread formulate wildfire spread prediction from fire, weather, vegetation, and terrain inputs (Radke et al., 2019; Huot et al., 2022). Recent datasets extend wildfire learning to longer time periods, multiple factors, and satellite time series (Zhao et al., 2025; Xu et al., 2026). Broader benchmarks, including WILDS, SustainBench, GEO-Bench, WeatherBench2, and TorchGeo, standardize datasets, splits, implementations, and metrics across distribution-shift, sustainability, Earth-observation, and weather tasks (Koh et al., 2021; Yeh et al., 2021; Lacoste et al., 2023; Rasp et al., 2024; Stewart et al., 2022). PANGAEA and GEO-Bench-2 further expand GeoFM evaluation across datasets, modalities, resolutions, and task types (Marsocci et al., 2024; Simumba et al., 2025). These benchmarks compare models under prescribed protocols but do not test how the same wildfire outputs and representations are reinterpreted under alternative matching, selection, and scope choices. Our benchmark evaluates wildfire-specific and general-purpose Earth FMs under matched contracts across classification, segmentation, regression, and retrieval and we report all six task forms separately.

7 Conclusion and Future Work

We formulate wildfire transfer as a fixed-contract evaluation problem. Controlled checks show that matching and head selection can substantially change scores without changing the predictions or frozen representation. Across the full comparator set, evaluation configuration accounts for more main-effect score variation than backbone identity and reverses many pairwise rankings. We develop WILDFIRE-FM to provide reference performance across six downstream tasks under specified contracts. This wildfire reference provides a baseline for comparing general-purpose Earth FMs. It leads the evaluated frozen backbones on occupancy, spread, and burned-area prediction, while AlphaEarth approaches it on retrieval, smoke $PM_{2.5}$, and extreme heat. Together, the evaluation contracts and reference performance support comparisons of Earth-FM transfer across wildfire tasks. More broadly, evaluation design should be treated as an integral part of the scientific claim. Reporting each score together with its task and contract is necessary for Earth-FM comparisons to provide interpretable and reproducible evidence of transfer. Future work can extend fixed-contract analysis to additional operational choices, such as calibration and cost-sensitive decision rules, and determine whether similar ranking instability arises in other geospatial prediction problems.

Ethics Statement

Wildfire prediction can influence public safety, emergency response, and resource allocation. We emphasize transparent evaluation as a foundation for responsible use of Earth foundation models. Matching rules, decision thresholds, and evaluation scope determine which events receive credit and which populations contribute to reported performance. Making these choices explicit helps researchers and practitioners relate benchmark results to the decisions they intend to support. Operational applications should combine model outputs with current observations, local expertise, and established emergency-management procedures. Evaluation should reflect the consequences of missed fires, false alerts, and warning delays, together with regional differences in monitoring coverage and response capacity. Regional validation, probability calibration, and ongoing performance monitoring are central to assessing these consequences and supporting equitable resource allocation.

Responsible data use requires compliance with source licenses, attribution requirements, and redistribution terms. Sharing evaluation protocols, processing documentation, and derived results can support reproducibility while respecting these obligations. Communication of model performance should preserve the connection between each score, its evaluation contract, and its intended application, enabling informed assessment of the benefits and risks of downstream use.

AI Use Statement

AI-assisted tools were used solely for grammar checking and presentation-quality review, including the identification of sentence fragments, dangling words, and formatting inconsistencies. The authors reviewed and approved every resulting edit and retain full responsibility for the manuscript.

Reproducibility Statement

We provide evaluation scripts and result records through the anonymous code release linked in this paper. Appendix C.2 describes the available artifacts and compute record; Appendices C.4 and C.7 specify matching settings and reproduce the score decomposition and ranking checks. The temporal and regional splits are given in Appendices E.1 and E.2. Recomputing predictions requires the original input and feature caches in addition to the released aggregate records.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021. URL https://arxiv.org/abs/2103.03098.

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. URL https://arxiv.org/abs/2507.22291.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023a. URL https://arxiv.org/abs/2304.02948.

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*, 6:190, 2023b. doi: 10.1038/s41612-023-00512-1.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, pp. 197–211, 2022. doi: 10.52202/068431-0015.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. URL https://arxiv.org/abs/2107.07002.

Elizabeth E. Ebert. Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting*, 24(6):1498–1510, 2009. doi: 10.1175/2009WAF2222251.1.

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. doi: 10.3390/rs12081252.

Eric Gilleland, David Ahijevych, Barbara G. Brown, Barbara Casati, and Elizabeth E. Ebert. Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5):1416–1430, 2009. doi: 10.1175/2009WAF2222269.1.

Johann Georg Goldammer. Early warning systems for the prediction of and appropriate response to wildfires and related environmental hazards. In *Health Guidelines for Vegetation Fire Events: Background Papers*, pp. 9–70. World Health Organization, 1999. URL http://edoc.mpg.de/373466.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27662–27673, 2024. doi: 10.1109/CVPR52733.2024.02613.

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.

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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

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*, volume 36, pp. 51080–51093, 2023. doi: 10.52202/075280-2223.

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. doi: 10.1126/science.adi2336.

Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*, 2014. URL https://arxiv.org/abs/1402.1892.

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*, volume 36, pp. 9789–9836, 2023. doi: 10.52202/075280-0428.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z.

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. URL https://arxiv.org/abs/2412.04204.

Matthew B. McDermott, Haoran Zhang, Lasse Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44102–44163, 2024. doi: 10.52202/079017-1400.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938, 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. URL https://arxiv.org/abs/2202.11214.

Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, and Michael Pritchard. Kilometer-scale convection-allowing model emulation using generative diffusion modeling. *Science Advances*, 12 (5):eadv0423, 2026. doi: 10.1126/sciadv.adv0423.

David Radke, Anna Hessler, and Dan Ellsworth. FireCast: Leveraging deep learning to predict wildfire spread. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 4575–4581, 2019. doi: 10.24963/ijcai.2019/636.

Stephan Rasp, Stephan Hoyer, Alexander 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. doi: 10.1029/2023MS004019.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4065–4076, 2023. doi: 10.1109/ICCV51070.2023.00378.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E. Phillips, et al. Prithvi WxC: Foundation model for weather and climate. *arXiv preprint arXiv:2409.13598*, 2024. URL https://arxiv.org/abs/2409.13598.

Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longepe, Xiao Xiang Zhu, Hannah Kerner, et al. GEO-Bench-2: From performance to capability, rethinking evaluation in geospatial AI. *arXiv preprint arXiv:2511.15658*, 2025. URL https://arxiv.org/abs/2511.15658.

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 30th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–12, 2022. doi: 10.1145/3557915.3560953.

Jeremias Traub, Till 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*, volume 37, pp. 2323–2347, 2024. doi: 10.52202/079017-0076.

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. doi: 10.1029/2020MS002109.

Zhengsen Xu, Sibo Cheng, Lanying Wang, Hongjie He, Wentao Sun, Jonathan Li, and Lincoln Linlin Xu. BCWildfire: A long-term multi-factor dataset and deep learning benchmark for boreal wildfire risk prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46): 39486–39494, 2026. doi: 10.1609/aaai.v40i46.41299.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon 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: Datasets and Benchmarks Track*, 2021. URL `https://arxiv.org/abs/2111.04724`.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018. doi: 10.1109/CVPR.2018.00391.

Yu Zhao, Sebastian Gerard, and Yifang Ban. TS-SatFire: A multi-task satellite image time-series dataset for wildfire detection and prediction. *Scientific Data*, 12(1):1817, 2025. doi: 10.1038/s41597-025-06271-3.

APPENDIX CONTENTS

A. Evaluation Framework Overview . . . 15
B. Notation and Metric Guide . . . 16
C. Controlled Evidence and Evaluation Details . . . 17
C.1. Model and Task Coverage . . . 17
C.2. Reproducibility and Compute . . . 17
C.3. Reference Implementation and Comparison Design . . . 18
C.4. Matching Implementation and Constant Predictions . . . 18
C.5. Matching Rules and Evaluation Scope . . . 19
C.6. Head Selection on Frozen Representations . . . 21
C.7. Score Decomposition and Ranking Uncertainty . . . 23
D. Complete Task-Level Comparison Matrices . . . 25
D.1. Reference Performance and Task Diagnostics . . . 25
D.2. Interpreting the Task-Level Comparisons . . . 26
D.3. Supporting-Task Overview . . . 27
D.4. Wildfire Occupancy . . . 28
D.5. Fire Spread . . . 29
D.6. Final Burned Area . . . 29
D.7. Analog Retrieval . . . 30
D.8. Smoke $PM_{2.5}$ . . . 30
D.9. Extreme Heat . . . 31
E. Temporal and Geographic Robustness . . . 32
E.1. Cross-Year Evaluation . . . 32
E.2. Region-Disjoint Evaluation . . . 32
E.3. Primary-Result Confidence Intervals . . . 32
F. Discussion: From Benchmark Scores to Transfer Science . . . 34
F.1. Evaluation Contracts Define the Capability Being Measured . . . 34
F.2. Contract Sensitivity as a Representation Profile . . . 35
F.3. Domain Specialization and Adaptation . . . 35
F.4. Connecting Transfer to Operational Decisions . . . 35
F.5. Building Cumulative Evidence Across Studies . . . 35

A Evaluation Framework Overview

The appendix follows the evaluation pipeline. We first define notation and document comparator coverage, the reference implementation, and experimental settings. We then specify matching operations, examine head selection, and quantify score variation and ranking uncertainty. Task diagnostics and complete matrices precede the temporal and geographic checks. The final discussion connects these results to future comparisons and operational evaluation.

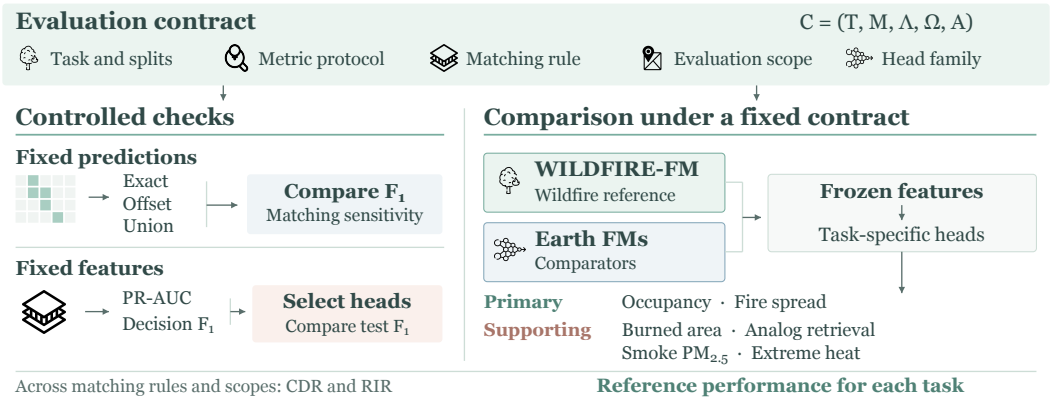


Figure 6: Overview of the evaluation framework. The contract specifies the task, metric protocol, matching rule, evaluation scope, and head family. Controlled checks vary matching or validation head selection with predictions or features held fixed. Model comparisons hold the contract fixed and use WILDFIRE-FM to provide reference performance on six tasks. CDR and RIR summarize score variation and ranking changes across configurations.

Table 4 organizes the interpretation of the evidence by the question each evaluation answers. Event matching determines the form of spatial agreement being measured; head selection determines how frozen features are converted into decisions; task-level evaluation identifies the prediction targets supported by those features.

Table 4: Evaluation questions and the evidence used to answer them. Each row connects an evaluation choice to a distinct aspect of transfer.

Evaluation choice	Scientific question	Evidence to examine
Event matching	How precisely does the prediction locate fire in space and time?	Exact, tolerated, and union scores computed from the same outputs.
Head selection	How does the validation objective translate frozen features into downstream decisions?	Selected-head identities and paired test-score differences.
Task and scope	Which targets and spatial populations are served by a representation?	Per-task metrics, regional evaluation, and scope-specific model rankings.

B NOTATION AND METRIC GUIDE

Table 5 follows the evaluation pipeline from inputs and frozen representations to head selection, event matching, and scalar scores. We use lowercase bold $\mathbf{s}_{b,h}$ for prediction-score fields and $S_b(\mathcal{C})$ for the scalar performance produced by a complete contract. The configuration index $\xi \in \Xi$ tracks changes to evaluation choices, and X denotes downstream inputs.

Table 5: Mathematical notation used in the transfer and fixed-contract evaluation framework.

Symbol	Meaning	Symbol	Meaning
Backbones, representations, and outputs			
$\mathcal{B}$	Set of evaluated frozen backbones.	b, b'	Backbone indices in $\mathcal{B}$.
$\mathcal{T}$	Downstream prediction task, including its data and split.	n	Index of a downstream example.
(i, t)	Spatial unit i at forecast time t.	(j, u)	Observed spatial unit j at time u.
x_n	Input for downstream example n; $n = (i, t)$ for gridded tasks.	f_b	Frozen backbone associated with model b.
$z_n^{(b)}$	Representation of input x_n produced by f_b.	Z_b	Complete frozen representation produced by backbone b.
$\mathcal{A}$	Candidate family of lightweight prediction heads.	h	One prediction head in $\mathcal{A}$.
K	Number of candidate heads in $\mathcal{A}$.	$s_n^{(b,h)}$	Score produced by backbone b and head h for example n.
$\mathbf{s}_{b,h}$	Complete score field produced by backbone b and head h.	y_n	Observed target for example n.
Y	Complete observation field.	X	Complete downstream input collection.
Fire sets and matching			
τ	Decision threshold selected on validation data.	$\widehat{P}_\tau$	Set of predicted fire events after thresholding.
P	Set of observed fire events.	$B, \widehat{B}$	Observed and predicted burned regions for fire spread.
Λ	Prediction–observation matching rule.	Λ_{exact}	Same-cell, same-time matching relation.
Λ_{offset}	Matching relation for displaced events within the spatial and temporal tolerances.	Λ_{union}	Combined matched support, specified per experiment.
$d_{\text{s}}(i, j)$	Spatial distance between units i and j.	$r_{\text{s}}, r_{\text{t}}$	Fixed spatial and temporal matching tolerances.
A_Λ, B_Λ	Matched prediction and observation sets under Λ.	p_Λ, r_Λ	Matched-prediction precision and matched-observation recall.
TP_Λ	Number of exact true positives.	$\text{FP}_\Lambda, \text{FN}_\Lambda$	Exact false positives and false negatives.
$F_{1,\Lambda}$	Decision F_1 obtained under matching rule Λ.		
Head selection and evaluation contracts			
M_{sel}	Validation metric used to select a prediction head.	M_{eval}	Metric used to evaluate the selected head.
$h^*_{b,M}$	Head selected for backbone b by the validation component of M.	$M_{\text{PR}}, M_{\text{F1}}$	Protocols using validation PR-AUC or decision F_1 for head selection.
$S_b(\mathcal{C})$	Scalar performance of backbone b under contract $\mathcal{C}$.	Test score difference	Decision-F_1 of the M_{F1}-selected head minus that of the M_{PR}-selected head on the same test set.
$\mathcal{C}$	Evaluation contract $(\mathcal{T}, M, \Lambda, \Omega, \mathcal{A})$.	M	Metric protocol, including selection, evaluation, and threshold rules.
Ω	Evaluation scope, such as global or fire-prone.	ξ	One configuration of evaluation choices.
Ξ	Set of evaluated configurations.	$q_b(\xi)$	Scalar test score of backbone b under configuration ξ.
Contract sensitivity			
μ	Grand mean of the score matrix.	α_b	Backbone main effect.
β_ξ	Evaluation-configuration main effect.	$\gamma_{b,\xi}$	Backbone–configuration interaction.
SS_α	Sum of squares for the backbone main effect.	SS_β	Sum of squares for the configuration main effect.
SS_γ	Sum of squares for the backbone–configuration interaction.	CDR	Contract dominance ratio, $SS_\beta/(SS_\beta + SS_\alpha)$.
RIR	Fraction of backbone pairs whose winner reverses across configurations.		

Table 6 groups the metric vocabulary used in the task tables. Error metrics retain the units or transformations of their targets; correlations describe association; ranking and thresholded metrics describe different ways of recovering relevant events. Upward arrows mark larger preferred values, and downward arrows mark smaller preferred values. Each caption states whether a score is expressed as a fraction or a percentage.

Table 7: Model/task coverage for the shared six-task comparator set. R denotes a reported result. Pangu6 denotes Pangu-Weather 6h. The additional Pangu-Weather 24h occupancy result appears in Table 22.

Task	Reference	Prithvi-WxC	Aurora	ClimaX	StormCast	Pangu24	DLWP	FCN	FengWu	FuXi	Pangu6	AlphaEarth
Wildfire occupancy	R	R	R	R	R	R	R	R	R	R	R	R
Fire spread	R	R	R	R	R	R	R	R	R	R	R	R
Final burned area	R	R	R	R	R	R	R	R	R	R	R	R
Analog retrieval	R	R	R	R	R	R	R	R	R	R	R	R
Smoke $PM_{2.5}$	R	R	R	R	R	R	R	R	R	R	R	R
Extreme heat	R	R	R	R	R	R	R	R	R	R	R	R

Table 6: Metric abbreviations used in the experiments.

Abbreviation	Meaning	Abbreviation	Meaning
F_1	Harmonic mean of precision and recall.	IoU	Intersection over union.
AP	Average precision.	AUROC	Area under the receiver operating characteristic curve.
PR-AUC	Area under the precision–recall curve.	MAE	Mean absolute error.
RMSE	Root mean squared error.	RMSE-C	Root mean squared error measured in degrees Celsius; MAE-C is mean absolute error in the same units.
nDCG@10	Normalized discounted cumulative gain at rank 10.	ρ	Spearman rank correlation.
r	Pearson correlation coefficient.		

C Controlled Evidence and Evaluation Details

We establish the experimental records before interpreting score changes. Matching definitions identify the measured agreement; head-selection checks and matrix statistics then quantify sensitivity under those definitions.

C.1 Model and Task Coverage

Table 7 records the twelve entries shared across all six tasks: the wildfire reference and eleven general-purpose comparators. Each transfer comparison uses a frozen backbone and the common within-task candidate-head family, training procedure, and evaluation rules. The AlphaEarth spread results use the projected-grid feature cache aligned with the spread targets. The occupancy matrix additionally reports Pangu-Weather 24h in Table 22.

The shared rows make it possible to follow a backbone across occupancy, spread, event-level regression, retrieval, and environmental outcomes. The task-specific matrices retain the native metrics required to interpret those transitions.

C.2 Reproducibility and Compute

The reproducibility record links each aggregate result to its evaluation definition and result-level artifacts. The manuscript specifies task targets, data splits, label spaces, matching tolerances, spatial populations, validation selection, decision thresholds, and candidate-head families. The linked code release contains check scripts and per-head and per-seed records. This manuscript package contains the LaTeX tables and the additional score-decomposition script.

These artifacts expose exact, tolerated, and union F_1, PR-AUC, selected-head identities, top-1 selection agreement, and paired test-score arithmetic. Reconstructing a table proceeds from candidate-head records to validation selection, test evaluation, and aggregation. This chain provides an executable record of how each reported decision score is obtained. The released occupancy summaries use $\{1, 7, 42, 99, 123\}$ for seven entries and $\{7, 17, 37, 57, 97\}$ for six; the explicit task-specific spread table lists seeds 1, 7, and 99. Each diagnostic records its aggregation count, and the LORO study uses three seeds for each of six regional holdouts.

Means and standard deviations describe repeated-run results. Fixed-output checks reuse one prediction field, and fixed-feature checks reuse one representation while changing the specified evaluation component. The confidence-interval tables additionally summarize uncertainty in the estimated means. Values reported as < 0.0001 give the display bound for the recorded standard deviation.

Result-level reconstruction uses the released scripts and CSV records. Access to raw wildfire inputs and large feature arrays follows redistribution review and storage planning; end-to-end recomputation requires those input artifacts alongside the result-level package.

The experiments use a shared Slurm-managed cluster. Tabular regression, retrieval, and same-feature head controls use CPU workers with 4–8 cores, 24–64 GB host memory, and 2–4-hour scheduling limits. Spread U-Net training and threshold calibration use a single B200 GPU, eight CPU cores, 96 GB host memory, and a four-hour scheduling limit. The recorded seed/check waves have approximately 78 CPU job-hours and 12 GPU job-hours of scheduled budget. This accounting covers the recorded seed/check waves. Exploratory runs form a separate compute category. The LORO study comprises eighteen independently trained models together with preprocessing and result aggregation.

C.3 Reference Implementation and Comparison Design

Occupancy implementation. The released occupancy implementation uses `UNetSmallFlex`, with encoder widths 32, 64, 128, and 256, a 512-channel bottleneck, and four upsampling stages with skip connections. Each block contains two 3×3 convolutions, group normalization, and ReLU. The configuration uses dropout 0.1 and 1×1 output heads. With 16 input channels and the auxiliary head, this configuration has 7,766,818 trainable parameters. The training template specifies AdamW, learning rate 2×10^{-4}, weight decay 10^{-4}, batch size 32, and ten epochs. Its objective combines weighted binary cross-entropy with an auxiliary spatial-support loss of weight 0.5 and radius two. The run summary records positive-class weight 10 and initialization without an earlier checkpoint. These details describe the released occupancy implementation; the task-specific matrices below describe the downstream comparisons.

Input information and time splits. The released 16-channel input specification contains ten weather fields, two validity masks, fuel and canopy fields, and two exposure channels. The no-exposure configuration zeros the exposure channels. FIRMS supplies the occupancy targets in this specification. The fire-history inputs described in the robustness experiments belong to their corresponding input settings. The released cache template uses a lower-48 5-km projected grid and a 12-hour target offset, with June–August 2024 for training, September for validation, and October for testing. The main comparison specifies its California evaluation domain, while Appendices E.1 and E.2 describe the separate temporal and spatial holdouts. We distinguish the target time, the input observation time, and the period used to fit a representation when interpreting these splits.

The role of the wildfire reference. The reference provides a domain-trained performance baseline for the six task forms. Its cross-task profile gives meaning to the relative ordering of general-purpose comparators: a model can rank highly among those comparators while retaining a substantial gap to the reference, or approach the reference closely on a different task. We report this relationship under each downstream contract. Input channels, architecture, and earlier training data also shape the representation available to the head. Matching the downstream contract controls the comparison procedure; matching these upstream factors supports a separate attribution of gains to domain training.

The input-matched comparisons discussed in Appendix F.3 follow directly from this distinction between downstream evaluation and the information available to a model.

Complementary reference models. The constant-head controls measure performance available without spatially varying features. The task-trained U-Net in Table 17 measures a different training route: optimizing a network directly for spread. Its close exact and spatial F_1, higher mean AP, and stronger seed-99 result show that task-specific training can match or exceed the reference on particular metrics. These comparisons establish useful reference points for metric behavior, frozen transfer, and task-specific optimization.

C.4 Matching Implementation and Constant Predictions

Coverage-based precision and recall. We make the released occupancy implementation explicit. Let $U_\Lambda(P)$ be the support generated from observations by a matching rule and define $U_\Lambda(\widehat{P}_\tau)$ analogously. The matched sets are

$$A_\Lambda = \widehat{P}_\tau \cap U_\Lambda(P) \cap \Omega, \qquad B_\Lambda = P \cap U_\Lambda(\widehat{P}_\tau) \cap \Omega.$$

Precision is $p_\Lambda = |A_\Lambda|/|\widehat{P}_\tau \cap \Omega|$ and recall is $r_\Lambda = |B_\Lambda|/|P \cap \Omega|$. Each evaluated cell contributes once to its corresponding count, and one cell can provide support to several cells in the other set. The implementation constructs support before applying the scope mask. Empty denominators receive zero. This definition reduces to ordinary binary F_1 under exact matching; neighborhood matching uses two matched counts, which can differ.

Settings recorded for each experiment. The fixed-output scope summary records spatial tolerance as `t0_s8` and union as `t3_s8`. Spatial dilation uses a square max-pooling neighborhood: radius eight means eight grid cells along either coordinate. On the 5-km grid this is 40 km along each axis. Three six-hour steps allow an absolute temporal offset of 18 hours. Temporal matching concerns agreement at verification time, separate from the model's forecast lead time.

The head-selection runner uses temporal radius three and spatial radius eight for its tolerated score. Its `comprehensive_union` score combines spatiotemporal, spatial-buffer, boundary, and coarse-grid matched support. Buffer and boundary radii are eight, and the coarse-grid factor is eight. The fire-prone head-selection scope uses the training-defined top 20% mask. The cross-year record uses spatial radius three and temporal radius three at six-hour spacing, corresponding to 15 km along each axis and 18 hours. These source settings identify distinct experiments; the shared shorthand "union" alone is insufficient to reconstruct them.

Constant predictions as a reference for the metric. For an all-positive prediction over an evaluated domain with observation prevalence π, exact F_1 is $2\pi/(1+\pi)$. If neighborhood support covers a fraction u of the domain and every observed cell has predicted support, neighborhood F_1 becomes $2u/(1+u)$. Broad support can yield a large score without spatial discrimination. The constant-head union score of 50.55% in the head-selection records is a measured example under that experiment's contract. Its interpretation depends on the head's thresholded predictions and the evaluated support.

The cross-year union score of 95.19% should be read together with exact overlap, spatial-only scores, and the tolerance sweep in Appendix E.1. A high neighborhood score measures agreement at the specified verification scale. Constant, persistence, and climatological predictions provide complementary reference values for that scale when evaluated on the same masks, threshold protocol, and time indices.

Reading the three scores together. The exact score describes agreement at the original cell and time. The tolerated score credits agreement at its stated neighborhood scale, and the union score credits the combined support. As support expands, more predicted and observed cells can contribute to precision and recall without any change to the prediction field. The constant-head result makes the role of support especially clear because its decisions carry no spatial ranking within the scope. We retain exact and neighborhood scores together to distinguish precise localization from broader agreement. The controlled comparison measures how much the reported performance changes along these verification scales; the fixed-contract comparison then uses a specified scale for every model in a task.

C.5 Matching Rules and Evaluation Scope

The next step holds the feature source fixed and examines how spatial and temporal correspondence changes the comparison between prediction heads. Table 8 separates exact agreement, spatial tolerance, and two temporal-tolerance settings. The heads share an exact F_1 of 0.0012 at the displayed precision. The shallow adapter scores higher under spatial tolerance, while the linear probe scores higher under both spatiotemporal settings. This transition identifies the verification scale at which their relative performance changes.

Training-defined spatial populations. The scope-progression table below extends the comparison from a fixed feature source to the occupancy comparator set. Its global, top-5%, top-10%, and top-20% scopes are defined using training-period fire activity. The scope controls which spatial population contributes to the score, while the matching rule controls which predictions receive credit within that population.

Figure 7 provides the corresponding visual comparison. A common model order, determined by global union F_1, allows each backbone to be followed across scopes. Exact overlap, tolerance gain,

Table 8: Matching-rule sweep for linear and shallow heads on fixed Prithvi-WxC features. Scores are fractional F_1 values. The temporal and spatial columns identify the tolerance components varied in each setting.

Matching rule	Temporal	Spatial	Linear $F_1 \uparrow$	Shallow $F_1 \uparrow$
exact matching	none	none	0.0012	0.0012
spatial tolerance	none	yes	0.0388	**0.0390**
short temporal + spatial	short	yes	**0.1055**	0.0987
long temporal + spatial	long	yes	**0.1738**	0.1635

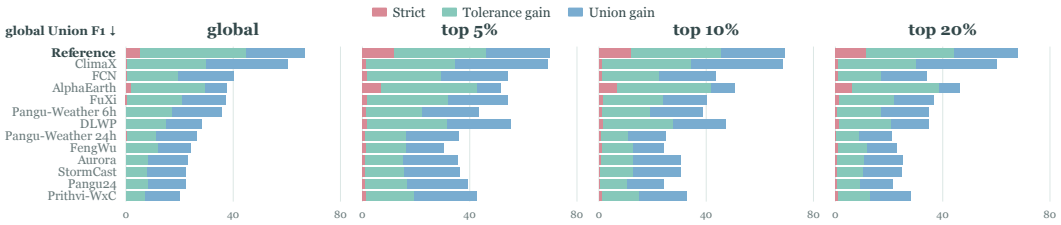


Figure 7: Occupancy across global and training-defined fire-prone scopes. Models follow global union-F_1 order. Bars show exact overlap, tolerance gain, and union gain for the global, top-5%, top-10%, and top-20% scopes.

and union gain identify the contribution of successive matching views. The full configuration analysis in the main text associates 76.71% of the combined backbone and configuration main-effect variation with configuration; 46.15% of backbone pairs reverse their ordering across configurations. The scope sweep locates these changes in the spatial populations being evaluated.

Table 9: Occupancy scores across global and fire-prone scopes for the fixed-output scope sweep. Global uses the full validation/test domain; top-k rows use train-defined fire-prone masks from historical fire frequency. Tolerance is spatial-only; union adds temporal and spatial matching. Difference is union minus strict. Rows report five-seed mean with small std; values are percentages.

Backbone	Scope	Strict F_1 ↑	Tolerance F_1 ↑	Union F_1 ↑	Difference ↑
Reference	global	5.31 ± 1.24	44.97 ± 2.15	66.74 ± 7.08	61.43 ± 8.13
	top 5%	11.94 ± 2.66	46.04 ± 5.20	69.84 ± 7.52	57.89 ± 8.62
	top 10%	11.78 ± 2.49	45.67 ± 4.82	69.21 ± 7.44	57.44 ± 8.65
	top 20%	11.42 ± 2.26	44.20 ± 4.18	67.80 ± 7.64	56.38 ± 8.87
Prithvi-WxC	global	0.055 ± 0.004	7.160 ± 0.660	20.190 ± 1.830	20.130 ± 1.830
	top 5%	1.410 ± 1.160	19.260 ± 4.500	42.580 ± 4.550	41.170 ± 3.480
	top 10%	1.240 ± 1.320	14.880 ± 8.440	32.690 ± 13.210	31.450 ± 11.910
	top 20%	1.150 ± 1.380	13.150 ± 9.460	28.130 ± 15.290	26.980 ± 13.920
Aurora	global	0.070 ± 0.010	8.500 ± 1.960	23.100 ± 4.940	23.040 ± 4.930
	top 5%	0.990 ± 0.930	15.130 ± 6.080	35.480 ± 11.020	34.500 ± 10.370
	top 10%	0.780 ± 1.050	12.740 ± 6.560	30.530 ± 10.880	29.750 ± 9.870
	top 20%	0.670 ± 1.100	10.530 ± 7.430	24.940 ± 12.580	24.280 ± 11.490
ClimaX	global	0.350 ± 0.080	29.750 ± 3.610	60.150 ± 7.590	59.800 ± 7.550
	top 5%	1.290 ± 0.110	34.580 ± 2.380	69.220 ± 5.720	67.920 ± 5.730
	top 10%	1.250 ± 0.160	34.330 ± 2.290	68.570 ± 5.540	67.320 ± 5.550
	top 20%	1.030 ± 0.270	30.210 ± 4.290	60.060 ± 7.570	59.040 ± 7.590
StormCast	global	0.056 ± 0.011	8.200 ± 2.190	22.380 ± 5.430	22.320 ± 5.420
	top 5%	0.960 ± 0.800	15.320 ± 5.530	36.190 ± 9.730	35.230 ± 9.180
	top 10%	0.730 ± 0.930	12.670 ± 6.330	30.470 ± 10.650	29.750 ± 9.750
	top 20%	0.580 ± 0.910	10.420 ± 7.340	24.660 ± 12.400	24.080 ± 11.500
Pangu24	global	0.055 ± 0.008	8.180 ± 1.350	22.060 ± 3.730	21.990 ± 3.730
	top 5%	1.060 ± 0.910	16.670 ± 4.710	39.140 ± 7.020	38.070 ± 6.490
	top 10%	0.720 ± 1.090	10.450 ± 8.190	24.120 ± 16.140	23.400 ± 15.200
	top 20%	0.660 ± 1.110	9.080 ± 8.200	21.220 ± 14.380	20.560 ± 13.310
AlphaEarth	global	2.060 ± 0.440	29.450 ± 6.010	37.430 ± 9.950	35.370 ± 10.030
	top 5%	6.910 ± 0.850	42.880 ± 4.610	51.740 ± 8.730	44.830 ± 9.080
	top 10%	6.640 ± 0.990	41.900 ± 5.950	50.570 ± 10.010	43.930 ± 9.920
	top 20%	6.190 ± 1.130	38.830 ± 7.500	46.380 ± 12.170	40.190 ± 11.680
DLWP	global	0.170 ± 0.040	14.910 ± 3.240	28.190 ± 6.970	28.020 ± 6.930
	top 5%	1.810 ± 0.480	31.720 ± 3.290	55.460 ± 5.290	53.650 ± 5.480
	top 10%	1.610 ± 0.600	27.660 ± 5.920	47.130 ± 8.010	45.520 ± 7.790
	top 20%	1.520 ± 0.900	20.940 ± 4.800	34.930 ± 7.850	33.410 ± 7.880
FCN	global	0.280 ± 0.080	19.510 ± 3.340	40.060 ± 9.370	39.780 ± 9.340
	top 5%	1.620 ± 0.510	29.380 ± 2.760	54.300 ± 7.410	52.680 ± 7.440
	top 10%	1.180 ± 0.510	22.420 ± 3.980	43.450 ± 9.250	42.270 ± 9.030
	top 20%	1.000 ± 0.430	16.980 ± 3.940	34.090 ± 8.260	33.090 ± 7.930
FengWu	global	0.260 ± 0.080	12.000 ± 6.020	24.100 ± 13.630	23.840 ± 13.570
	top 5%	1.570 ± 0.360	16.280 ± 3.700	30.110 ± 5.010	28.540 ± 4.770
	top 10%	1.240 ± 0.530	12.950 ± 5.610	24.190 ± 8.690	22.940 ± 8.190
	top 20%	1.120 ± 0.500	11.950 ± 5.070	22.790 ± 7.910	21.670 ± 7.440
FuXi	global	0.380 ± 0.120	21.030 ± 4.820	37.290 ± 9.450	36.910 ± 9.430
	top 5%	2.030 ± 0.680	31.890 ± 4.730	53.930 ± 8.380	51.900 ± 8.690
	top 10%	1.650 ± 0.730	24.010 ± 5.780	40.210 ± 9.930	38.560 ± 9.770
	top 20%	1.360 ± 0.680	21.950 ± 5.860	36.730 ± 10.030	35.370 ± 9.920
Pangu-Weather 24h	global	0.280 ± 0.140	11.340 ± 5.220	26.580 ± 12.360	26.300 ± 12.370
	top 5%	1.070 ± 0.310	16.440 ± 4.420	36.000 ± 11.620	34.930 ± 11.380
	top 10%	0.680 ± 0.370	10.950 ± 6.120	25.110 ± 14.420	24.430 ± 14.100
	top 20%	0.470 ± 0.290	8.710 ± 3.540	20.910 ± 8.550	20.440 ± 8.450
Pangu-Weather 6h	global	0.280 ± 0.110	17.090 ± 4.050	35.640 ± 9.030	35.360 ± 9.080
	top 5%	1.370 ± 0.310	22.220 ± 6.860	43.420 ± 13.240	42.060 ± 13.060
	top 10%	1.090 ± 0.350	18.930 ± 5.930	38.530 ± 11.720	37.440 ± 11.530
	top 20%	0.880 ± 0.360	17.020 ± 5.490	34.570 ± 10.290	33.680 ± 10.130

C.6 Head Selection on Frozen Representations

We begin with the selection step between a representation and its downstream prediction. For each feature source and scope, Z_b remains fixed while the validation criterion selects $h^*_{b,M_{\mathrm{PR}}}$ or $h^*_{b,M_{\mathrm{F1}}}$. Both selected heads are evaluated on the same test data using the same matching and threshold rules. The tables report the test decision-score difference in the direction of the decision-F_1 selection.

Paired-head comparison. Table 10 starts with a linear probe and a shallow adapter. The Prithvi-WxC fire-prone row makes the selection mechanism explicit: PR-AUC selects the linear head, whereas union F_1 selects the shallow adapter, producing an 88.10 percentage-point test difference. The corresponding global row selects the shallow adapter under both criteria. Thus, the same feature source can exhibit different selection behavior across spatial populations. Aurora provides

repeated-seed examples, and the ClimaX and Pangu24 rows show how selection agreement and score differences vary across feature sources.

Single-run failure and repeated-run evidence. The 88.10-point difference comes from one recorded Prithvi-WxC fire-prone comparison. Its linear head has zero decision F_1 under all three rules, while its PR-AUC exceeds the shallow head's value. Zero F_1 establishes zero credited overlap at the selected operating point. The case shows how ranking quality and thresholded decisions can disagree. Its single run supplies an example of the failure, with the repeated Aurora and Pangu24 rows providing separate evidence about variability. Selection on validation data also permits a negative difference on test data.

Why the selection objectives can disagree. PR-AUC summarizes precision and recall across score thresholds. Decision F_1 evaluates a particular set of thresholded predictions under the chosen matching rule. A head can preserve useful ordering over part of its score range while producing poor overlap at the selected operating point. The Prithvi-WxC case exhibits this separation: the linear head's PR-AUC is 14.00 against 12.21 for the shallow adapter, while their union decision scores are 0.00 and 88.10. The candidate family makes this outcome observable, and the selection metric determines which candidate supplies the reported prediction. The repeated-head comparisons describe how this relationship varies across feature sources, scopes, and runs. Together, these checks connect the individual failure case to the selection step in the evaluation contract.

Relation to the primary comparison. The same-feature controls compare particular candidate heads and matching implementations. The primary table aggregates a separate fixed-contract transfer evaluation. Prithvi-WxC's 71.52% global score in the head control and 20.19% in the primary table belong to these different experimental records. The candidate family, selected operating point, support rule, scope, and run count accompany each result. Appendix C.4 spells out the implementation differences.

Table 10: Same-feature occupancy paired-head check. Each row fixes the feature source and scope, then compares a linear probe against a shallow adapter. The final columns compare the PR-AUC choice with the union-F_1 choice and report the resulting union-F_1 score difference once per feature/scope pair. Values are percentages; seeded rows show mean F_1 triplets and report score-difference uncertainty over five seeds.

Feature	Scope	Linear probe		Shallow adapter		PR-AUC/union-F_1 choice	Score difference ↓
		PR-AUC↑	Exact/tolerated/union F_1 ↑	PR-AUC↑	Exact/tolerated/union F_1 ↑		
Prithvi-WxC	global	1.47	0.00/0.00/0.00	3.17	6.64/58.66/71.52	shallow / shallow	0.00
	fire-prone	14.00	0.00/0.00/0.00	12.21	25.38/81.17/88.10	linear / shallow	88.10
Aurora	global	0.02 ± 0.01	0.05/6.88/10.35	0.02 ± 0.01	0.05/7.61/11.55	differs 3/5	1.08 ± 1.82
	fire-prone	2.28 ± 0.87	3.65/38.43/44.21	2.20 ± 0.76	4.10/43.74/50.55	differs 2/5	6.35 ± 13.28
ClimaX	global	0.02	0.05/7.40/11.25	0.03	0.05/7.57/11.50	differs 1/5	0.05 ± 0.11
	fire-prone	2.21 ± 0.07	4.10/43.74/50.55	2.33 ± 0.19	4.10/43.74/50.55	aligned choices	0.00
Pangu24	global	0.03 ± 0.01	0.04/7.44/11.26	0.03 ± 0.01	0.05/7.56/11.48	differs 3/5	0.17 ± 0.18
	fire-prone	2.50 ± 1.12	3.66/43.68/50.42	3.19 ± 1.10	4.10/43.74/50.55	aligned	0.00

Expanded candidate family. Table 11 extends the choice to five candidates: a constant prior, linear probe, pixel MLP, shallow adapter, and wide shallow adapter. This expansion tests how the selection criterion interacts with the set of available predictors. For Aurora, the global count rises from a paired-head choice difference in three of five seeds to a score gap exceeding the stated threshold in all five expanded-family runs. These counts describe their respective events: differing selected identities in the paired-head table and a thresholded test-score difference in the expanded-family table.

The fire-prone rows expose the interaction between feature source and spatial scope: Aurora and Pangu24 have larger mean differences than their global rows, while ClimaX has equal fire-prone decision scores. The Prithvi-WxC row records one fixed-feature comparison; the Aurora, ClimaX, and Pangu24 rows summarize five runs each.

Per-head diagnostics. Table 12 expands the same per-head records into selection diagnostics. Reading the selected identities alongside their test decision scores connects the validation objective to the final decision outcome. This trace also links the paired-head and expanded-family summaries to the candidate-level measurements used to construct them.

Table 11: Five-candidate occupancy head selection on fixed features. Counts give runs in which the PR-AUC-selected head's test union F_1 is more than 0.005 below the decision-F_1-selected head on the fractional scale (0.5 percentage points). Mean and maximum score differences are in percentage points. Denominators specify the number of runs summarized in each row.

Feature	Scope	Count ↓	Mean selection regret ↓	Max selection regret ↓	Reading
Prithvi-WxC	fire-prone	$1/1$	88.10	88.10	largest gap
Aurora	global	$5/5$	9.56 ± 4.00	11.58	repeated gap
Aurora	fire-prone	$4/5$	34.99 ± 22.85	50.55	repeated gap
ClimaX	global	$0/5$	0.07 ± 0.15	0.33	below the count threshold
ClimaX	fire-prone	$0/5$	0.00	0.00	equal decision scores
Pangu24	global	$1/5$	2.45 ± 5.11	11.58	one seed differs
Pangu24	fire-prone	$2/5$	20.22 ± 27.69	50.55	two seeds differ

Table 12: Selection-regret check with task, matching rule, scope, and head set held fixed. Here D is union-F_1 percentage; regret is $D(h_D) - D(h_R)$. Seeded rows report mean with small std; single-run rows report one value.

Feature source	Ω	PR-AUC-selected $D(h_R)$ ↑	Decision-selected $D(h_D)$ ↑	Regret $D(h_D) - D(h_R)$ ↓
Prithvi-WxC	global	71.52	71.52	0.00
Prithvi-WxC	fire-prone	0.00	88.10	88.10
Aurora	global	1.80 ± 4.03	11.37 ± 0.14	9.56 ± 4.00
Aurora	fire-prone	15.56 ± 22.85	50.55	34.99 ± 22.85
ClimaX	global	11.37 ± 0.16	11.43 ± 0.17	0.07 ± 0.15
ClimaX	fire-prone	50.55	50.55	0.00
Pangu24	global	9.12 ± 5.10	11.57 ± 0.04	2.45 ± 5.11
Pangu24	fire-prone	30.34 ± 27.70	50.56 ± 0.02	20.22 ± 27.69

Table 13: Selection-regret scores under the available matching rules. The main table uses union-F_1. This table repeats the same PR-AUC-selection comparison for exact, tolerated, and union matching. Values are percentages or score differences; seeded rows list the selected head types across seeds.

Feature	Scope	Matching rule	PR-AUC-selected score ↑	Decision-selected score ↑	Selection regret ↓	PR-AUC-selected head type(s)	Decision-selected head type(s)
Prithvi-WxC	global	exact F_1	6.64	6.64	0.00	shallow adapter	shallow adapter
Prithvi-WxC	global	tolerated F_1	58.66	58.66	0.00	shallow adapter	shallow adapter
Prithvi-WxC	global	union-F_1	71.52	71.52	0.00	shallow adapter	shallow adapter
Prithvi-WxC	fire-prone	exact F_1	0.00	25.38	25.38	linear probe	shallow adapter
Prithvi-WxC	fire-prone	tolerated F_1	0.00	81.17	81.17	linear probe	shallow adapter
Prithvi-WxC	fire-prone	union-F_1	0.00	88.10	88.10	linear probe	shallow adapter
Aurora	global	exact F_1	0.01 ± 0.03	0.05 ± 0.01	0.04 ± 0.02	linear/pixel MLP	constant/linear/shallow
Aurora	global	tolerated F_1	1.25 ± 2.79	7.48 ± 0.09	6.23 ± 2.77	linear/pixel MLP	constant/shallow
Aurora	global	union-F_1	1.80 ± 4.03	11.37 ± 0.14	9.56 ± 4.00	linear/pixel MLP	constant/shallow
Aurora	fire-prone	exact F_1	1.43 ± 1.99	4.10	2.68 ± 1.99	linear/pixel MLP/wide	constant
Aurora	fire-prone	tolerated F_1	13.80 ± 20.00	43.74	29.94 ± 20.00	linear/pixel MLP/wide	constant
Aurora	fire-prone	union-F_1	15.56 ± 22.85	50.55	34.99 ± 22.85	linear/pixel MLP/wide	constant
ClimaX	global	exact F_1	0.05	0.05	0.00	shallow/wide	shallow
ClimaX	global	tolerated F_1	7.48 ± 0.11	7.53 ± 0.11	0.05 ± 0.10	shallow/wide	shallow
ClimaX	global	union-F_1	11.37 ± 0.16	11.43 ± 0.17	0.07 ± 0.15	shallow/wide	shallow
ClimaX	fire-prone	exact F_1	4.10	4.10	0.00	linear/shallow/wide	constant
ClimaX	fire-prone	tolerated F_1	43.74	43.74	0.00	linear/shallow/wide	constant
ClimaX	fire-prone	union-F_1	50.55	50.55	0.00	linear/shallow/wide	constant
Pangu24	global	exact F_1	0.04 ± 0.02	0.05	0.01 ± 0.02	pixel MLP/shallow/wide	shallow
Pangu24	global	tolerated F_1	6.00 ± 3.36	7.62 ± 0.02	1.62 ± 3.36	pixel MLP/shallow/wide	shallow
Pangu24	global	union-F_1	9.12 ± 5.10	11.57 ± 0.04	2.45 ± 5.11	pixel MLP/shallow/wide	shallow
Pangu24	fire-prone	exact F_1	2.47 ± 2.26	4.11 ± 0.03	1.64 ± 2.25	linear/pixel MLP/shallow/wide	constant/linear
Pangu24	fire-prone	tolerated F_1	26.25 ± 23.97	43.75 ± 0.02	17.50 ± 23.96	linear/pixel MLP/shallow/wide	constant/linear
Pangu24	fire-prone	union-F_1	30.34 ± 27.70	50.56 ± 0.02	20.22 ± 27.69	linear/pixel MLP/shallow/wide	constant/linear

Selection differences across matching rules. Table 13 evaluates the selected-head comparison across the matching family. Each matching column gives a different spatial interpretation of the two heads' outputs. Reading across a row shows how the decision-score difference changes when the evaluator credits exact agreement or bounded displacement; reading down a column compares feature sources under that common rule.

C.7 Score Decomposition and Ranking Uncertainty

Decomposition of the reported matrix. We reconstruct the 13×12 occupancy matrix from the displayed means in Table 9. Its columns combine three matching rules and four scopes. With $B = 13$

and $K = 12$, we compute

$$\mu = \frac{1}{BK}\sum_{b,\xi} q_b(\xi), \qquad \alpha_b = \frac{1}{K}\sum_{\xi} q_b(\xi) - \mu, \qquad \beta_\xi = \frac{1}{B}\sum_{b} q_b(\xi) - \mu, \tag{1}$$

$$SS_\alpha = K\sum_b \alpha_b^2, \qquad SS_\beta = B\sum_\xi \beta_\xi^2, \qquad SS_\gamma = \sum_{b,\xi}(q_b(\xi) - \mu - \alpha_b - \beta_\xi)^2. \tag{2}$$

We report all three components in Table 14. The CDR denominator contains the two main effects; the interaction accounts for 9.10% of total variation in this matrix. These are descriptive sums of squares on aggregate scores, with units of squared percentage points.

Table 14: Decomposition of the displayed occupancy matrix.

Component	Sum of squares	Share of total (%)
Backbone, SS_α	11604.83	21.17
Configuration, SS_β	38222.11	69.73
Interaction, SS_γ	4990.88	9.10

Interpretation of CDR. Neighborhood verification already establishes that tolerance changes spatial forecast scores (Ebert, 2009; Gilleland et al., 2009). Here we measure the size of that dependence in a specified transfer comparison. Increasing the support region can raise every model's score and enlarge SS_β. CDR consequently depends on the chosen configuration range. Monotonicity alone leaves its value undetermined: for an additive matrix $q_b(\xi) = u_b + v_\xi$, an increasing v_ξ permits either main effect to dominate as their relative variation changes. The measured CDR describes this benchmark matrix, while SS_γ describes departures from a shared response to configuration. Neither quantity establishes an intrinsic property of a backbone or statistical significance.

Connecting score changes to ranking changes. The decomposition separates two mechanisms. A shared increase under a more permissive configuration contributes to β_ξ. For two backbones, however, the score difference is

$$q_b(\xi) - q_{b'}(\xi) = \alpha_b - \alpha_{b'} + \gamma_{b,\xi} - \gamma_{b',\xi}.$$

The common configuration effect cancels. Reversing the ordering requires the difference between the interaction terms to cross the backbone-effect gap. This explains why CDR, the interaction sum of squares, and RIR describe complementary properties of the same matrix. A large shared score increase can coexist with stable rankings, while a smaller model-specific response can reverse a pair whose scores are close. The scope sweep connects these responses to particular matching rules and evaluated populations.

Ranking sensitivity beyond near ties. The reported RIR counts 36 of 78 pairs with opposite mean-score orderings in two configurations. We additionally require differences exceeding a margin ϵ in both directions. Table 15 reports this check and a separate interval-based screen. For the latter, we reconstruct marginal 95% Student-t intervals as $\bar{q} \pm t_{0.975,4}s/\sqrt{5}$ from the displayed five-run means and standard deviations. A pair passes only when the intervals are disjoint in opposite directions in two configurations.

Table 15: Sensitivity of ranking reversals to the comparison criterion. The interval screen is descriptive and uses rounded table entries.

Criterion	Reversing pairs	Rate (%)
Opposite mean orderings	36/78	46.15
Difference > 0.1 pp in both directions	34/78	43.59
Difference > 0.5 pp in both directions	25/78	32.05
Difference > 1.0 pp in both directions	22/78	28.21
Disjoint marginal 95% intervals in both directions	2/78	2.56

Uncertainty and independent samples. The interval screen and a paired significance test answer different questions. Marginal intervals omit covariance between predictions on shared examples and the search across configurations. A paired reversal analysis requires aligned predictions and resampling units, with multiplicity control across the tested pairs and configurations. Events or temporal blocks are suitable units when nearby cells share the same fire. Seed variability measures training variability conditional on the observed test set. The 46.15% result counts descriptive changes in ordering; 2.56% passes the marginal-interval screen. The available aggregates support these two statements separately.

D Complete Task-Level Comparison Matrices

The matrices organize the results by prediction unit: cell–time occupancy, spatial spread, event-level area, event retrieval, station concentrations, and temperature fields. Each subsection introduces the metric interpretation, reports the complete task table, and explains the resulting pattern. Together, they provide the native-scale values underlying the cross-task comparison.

D.1 Reference Performance and Task Diagnostics

The wildfire reference connects the evaluation analysis to a domain-trained representation. Table 16 summarizes its relationship to the general-purpose frozen comparators using representative metrics from the complete matrices. The six rows establish a task profile: occupancy and burned-area prediction favor different general-purpose comparators, while AlphaEarth approaches the reference on retrieval, smoke, and heat.

Table 16: Wildfire reference and the leading general-purpose frozen comparator on one representative metric per task. Entries retain the means and standard deviations from the task matrices.

Task	Metric	Reference	Comparator	Comparator score
Occupancy	Union F_1 (%) ↑	66.74 ± 7.08	ClimaX	60.15 ± 7.59
Spread	Exact F_1 (%) ↑	37.98 ± 1.33	Aurora	31.08 ± 0.40
Burned area	log-RMSE ↓	1.1649 ± 0.0139	DLWP	1.3070 ± 0.0980
Analog retrieval	nDCG@10 ↑	$0.5178 \pm < 0.0001$	AlphaEarth	$0.5036 \pm < 0.0001$
Smoke $PM_{2.5}$	RMSE ↓	4.4684 ± 0.0060	AlphaEarth	4.4899 ± 0.0042
Extreme heat	RMSE-C ↓	0.2179 ± 0.0048	AlphaEarth	$0.2187 \pm < 0.0001$

Task-specific spread training. Table 17 places the frozen wildfire reference beside a task-specific U-Net trained for spread. The three listed U-Net seeds have mean exact and spatial F_1 close to the reference, and higher mean AP. Seed 99 exceeds the reference mean on all three metrics. This comparison connects domain-pretrained transfer to task-specific optimization and shows how ranking quality and thresholded spatial agreement can favor different training approaches.

Table 17: Task-specific U-Net spread runs and the wildfire reference. The U-Net summary aggregates the three listed seeds (1, 7, and 99). Metrics use the fractional scale. Each U-Net threshold is validation-selected. The reference row reproduces the five-seed reference summary.

Entry	Threshold	Exact F_1 ↑	Spatial F_1 ↑	AP↑	Comparison summary
Wildfire reference	validation-selected	0.3798 ± 0.0133	0.8220 ± 0.0032	0.3273 ± 0.0060	frozen-reference summary
U-Net seed 1	0.90	0.3768	0.8187	0.3319	higher U-Net AP
U-Net seed 7	0.85	0.3483	0.8140	0.3280	higher reference exact F_1
U-Net seed 99	0.90	0.3802	0.8258	0.3409	higher U-Net means
U-Net mean with std	per-seed selection	0.3684 ± 0.0175	0.8195 ± 0.0059	0.3336 ± 0.0066	close exact and spatial means

Seed-level summaries. Table 18 collects repeated-seed results for the reference and the Aurora head-selection check. Burned-area error and correlation jointly describe event-level prediction, while smoke and heat combine continuous errors with association or exceedance scores. The Aurora row records a larger standard deviation than mean score difference, showing that seed-level selection behavior varies substantially around its aggregate. The five-seed spread aggregate reproduces the reference values. The three-seed task-specific U-Net summary appears in Table 17.

Table 18: Repeated-seed summaries for the wildfire reference and the Aurora paired-head check. Values are means with standard deviations. The Aurora score difference is in percentage points; the spread and heat F_1 values use the fractional scale.

Task check	Seeds	Metric 1	Metric 2	Reading
Burned-area reference	5	log-RMSE↓ 1.1649 ± 0.0139	log-MAE↓ 1.0355 ± 0.0114; Spear.↑ 0.5884 ± 0.0338	error and event ordering
Smoke $PM_{2.5}$ reference	5	RMSE↓ 4.4684 ± 0.0060	MAE↓ 2.4127 ± 0.0016; r↑ 0.6360 ± 0.0013	continuous prediction and association
Extreme-heat reference	5	RMSE-C↓ 0.2179 ± 0.0048	MAE-C↓ 0.1787 ± 0.0020; exceed. F_1 ↑ 0.9541 ± 0.0164	continuous and exceedance scores
Spread aggregate	5	exact F_1 ↑ 0.3798 ± 0.0133	spatial F_1 ↑ 0.8220 ± 0.0032; AP↑ 0.3273 ± 0.0060	matches the reference summary
Aurora paired-head replication	5	fire-prone score difference ↓ 6.35 ± 13.28	PR and union choices differ in 2/5 seeds	seed-dependent head selection

Retrieval input composition. Table 19 examines the feature groups used for analog retrieval. The boundary-field ablation has the largest nDCG@10 in this diagnostic, while the early-deployment feature set has the smallest log-MAE and the largest Spearman correlation. The WFIGS-field ablation has the smallest log-RMSE. These metrics emphasize different aspects of retrieved-event similarity. The early-deployment row supplies the reference values used in the main retrieval comparison; the categorical-field ablation has nDCG@10 equal to AlphaEarth's displayed value, with its own regression errors and event-ordering correlation.

Table 19: Analog-retrieval feature-group diagnostic. Each row identifies the feature group removed from its evaluated feature set and reports retrieval relevance, log-area error, and event-ordering correlation.

Feature set	Removed columns	nDCG@10↑	log-MAE↓	log-RMSE↓	Spearman↑	Reading
Full feature set	all columns retained	0.5215	1.0929	1.2725	0.5011	full-input reference point
Boundary-field ablation	`BurnBndLat/Lon`	0.5256	1.0424	1.2006	0.5529	highest nDCG@10
WFIGS-field ablation	acres/date/distance/match fields	0.5237	0.9805	1.1758	0.5963	lowest log-RMSE
Early-deployment feature set	boundary + WFIGS side info	0.5178	0.9683	1.1769	0.6004	lowest log-MAE; highest Spearman
Categorical-field ablation	incident/state/county/match type	0.5036	1.0183	1.2361	0.4779	nDCG@10 matches AlphaEarth

Task-specific decision views. Table 20 translates the supporting-task outputs into acceptable-error bands, retrieval hits, and exceedance events. For burned area, widening the log-error band increases event coverage; for retrieval, a wider log-area relevance band increases the fraction of queries with a matching analog among the top five. Smoke pairs concentration-error coverage with event scores at three concentration thresholds. Heat pairs Celsius-error coverage with a fixed 27°C event view. Together, these columns connect continuous or ranked outputs to the decisions each task serves.

Table 20: Supporting-task error bands and decision views. Continuous tasks use target-specific acceptable-error bands and exceedance thresholds; retrieval uses top-k relevance bands. Coverage, hit rates, and event scores in the additional-view column are percentages. Primary-view errors retain their stated units, and primary heat exceedance F_1 is fractional.

Supporting task	Primary view	Additional views	Reading
Final burned area	log-RMSE↓ 1.1649 ± 0.0139; log-MAE↓ 1.0355 ± 0.0114	$\lvert\Delta \log A\rvert \le$ 0.5/1.0/1.5: 21.1/42.1/78.9	event-level log-area error bands
Analog retrieval	nDCG@10↑ 0.5178; best log-delta↓ 0.1901	hit@5 within log-delta 0.5/1.0/1.5: 84.2/94.7/100.0	top-k and graded-relevance tolerance
Smoke $PM_{2.5}$	RMSE↓ 4.4684 ± 0.0060; MAE↓ 2.4127 ± 0.0016	within 2/5/10 μg m^{-3}↑: 59.2/89.9/97.9; event-F_1 ↑ at 12/35/55: 48.8/20.3/7.7	station regression with exceedance views
Extreme heat	RMSE-C↓ 0.2179 ± 0.0048; MAE-C↓ 0.1787 ± 0.0020; exceed. F_1 ↑ 0.9541 ± 0.0164	within 0.25/0.5/1.0°C↑: 75.1/98.8/99.9; 27°C event-F_1 ↑: 96.0	temperature accuracy and exceedance detection

D.2 Interpreting the Task-Level Comparisons

From a task score to an explanation. The full matrices pair error magnitude with event ordering, association, or thresholded decisions. Reading these metrics together helps distinguish the behavior of an evaluated prediction pipeline from a hypothesis about the information in its backbone. The feature ablations and repeated-run summaries then provide more direct comparisons of particular components.

The reference-normalized map summarizes the distance to a common task baseline. Nearby normalized cells can arise from different predictions with similar aggregate errors. We use the native metrics and feature-group comparisons below to connect that summary to the behavior of each task. This is why the six-task evaluation retains separate prediction units, metrics, and reference values throughout the comparison.

Heat error and temperature variation. Aurora and ClimaX have heat RMSE values of 18.05 and 17.65 °C in Table 27, while the reference and AlphaEarth are close in RMSE and exceedance F_1.

RMSE combines systematic bias with error in temperature variation. For residuals $e_i = \widehat{y}_i - y_i$, the identity $\mathrm{RMSE}^2 = \bar{e}^{\,2} + n^{-1}\sum_i (e_i - \bar{e})^2$ separates these contributions. The released AlphaEarth heat script explicitly converts the source temperature from kelvin to Celsius by subtracting 273.15. That conversion documents its target construction. The existing table also shows that FCN and FengWu have similar RMSE values near 2.1 °C but exceedance F_1 values of 0.9257 and 0.0481. Continuous error and exceedance performance capture different properties of the predictions. We retain both views when comparing heat transfer under the specified task contract.

Smoke accuracy and association. The atmospheric smoke rows have RMSE near 5.93 and Pearson correlations near zero. Together, these values show weak recovery of the observed concentration variation under the evaluated heads. A predictor close to a constant can have this profile, as can a varying predictor with little association with the target. The released smoke runner includes Ridge and ElasticNet candidates, with additional tree-based regressors in the larger candidate family. Deterministic fitting on fixed features can also produce repeated values across nominal seed changes. The small reported standard deviations characterize repeatability of the evaluated pipeline. The error and correlation columns characterize its accuracy and association on the evaluated station–time observations.

Shared information and retrieval ablations. In Table 19, the second column lists removed features. The early-deployment row removes boundary and WFIGS side information; its nDCG@10 is 0.5178 compared with 0.5215 for the full feature set. Its log-RMSE also improves from 1.2725 to 1.1769. These changes show that the retained features support the reported retrieval performance after removal of those field groups. AlphaEarth's nearby nDCG@10 and the similar smoke and heat scores establish comparable task-level performance. The feature-group comparisons give a more specific account of retrieval behavior: removing a field group can affect event relevance and numerical area error differently. This is consistent with reporting each task and metric separately in the fixed-contract comparison.

Event-level interpretation. The burned-area coverage and retrieval hit rates in Table 20 summarize the evaluated events and queries. Widening an error or relevance band increases coverage, which connects a continuous score to the number of cases meeting an application criterion. Each event contributes $100/n$ percentage points to coverage over n events, so small event sets produce visibly discrete changes. The sampling units are distinct: a burned-area target is an event, a retrieval target is a query, and the retrieval library contains candidate events. The coverage table describes these event-level outcomes, while the repeated-seed summaries describe variation across fitted models. We keep these two forms of evidence distinct when interpreting the task profile.

Projected-grid spread evaluation. The AlphaEarth spread values in the primary and complete matrices use the corrected projected-grid cache identified in their captions. Projection, resolution, pixel origin, and patch indexing determine which feature vector corresponds to each burned-raster cell. This correspondence is part of the input-to-label alignment underlying exact overlap, spatial overlap, and AP. We present the current-cache results consistently across these metrics.

D.3 Supporting-Task Overview

Table 21 brings together the four supporting tasks. The upper block compares event magnitude and analog relevance; the lower block compares continuous environmental outcomes and thresholded heat events. The wildfire reference has the strongest means across the burned-area and retrieval metrics. AlphaEarth approaches the reference on retrieval and environmental outcomes, with a lower smoke MAE and a slightly higher heat exceedance F_1. The subsequent task-specific tables retain the same numbers and unpack the different information conveyed by each metric.

Table 21: Supporting-task metric matrix. Top: final burned area and analog retrieval. Bottom: smoke $PM_{2.5}$ and extreme heat. Each task fixes its data split, candidate-head family, selection procedure, and evaluation rules. Bold identifies the best mean per metric. Error metrics retain their native or log-transformed units; correlations, nDCG, and heat exceedance F_1 use fractional values. Standard deviations below the displayed precision appear as < 0.0001.

	Burned area			Analog retrieval		
Backbone	**log-RMSE↓**	**log-MAE↓**	**Spearman↑**	**nDCG@10↑**	**log-RMSE↓**	**log-MAE↓**
Wildfire ref.	1.1649 ± 0.0139	1.0355 ± 0.0114	0.5884 ± 0.0338	$0.5178 \pm < 0.0001$	$1.1769 \pm < 0.0001$	$0.9683 \pm < 0.0001$
Prithvi-WxC	1.3630 ± 0.0681	1.2435 ± 0.0668	0.1799 ± 0.3002	0.3857 ± 0.0189	1.3908 ± 0.0938	1.2585 ± 0.0865
Aurora	1.8658 ± 0.2009	1.6717 ± 0.1245	-0.1156 ± 0.2982	0.4046 ± 0.0144	1.3659 ± 0.0792	1.2596 ± 0.0968
ClimaX	2.0300 ± 0.2103	1.8443 ± 0.1528	-0.2515 ± 0.2688	0.4143 ± 0.0191	1.4526 ± 0.0926	1.2441 ± 0.1446
StormCast	1.6679 ± 0.1438	1.4745 ± 0.1134	0.1830 ± 0.1969	0.4076 ± 0.0094	1.3663 ± 0.0781	1.2371 ± 0.1078
Pangu24	1.3820 ± 0.0638	1.2369 ± 0.0967	0.2309 ± 0.1358	0.3866 ± 0.0148	1.4057 ± 0.0769	1.1840 ± 0.1052
DLWP	1.3070 ± 0.0980	1.1769 ± 0.0834	0.4888 ± 0.1368	0.3972 ± 0.0146	1.5351 ± 0.0802	1.3196 ± 0.0781
FCN	1.3693 ± 0.0885	1.2599 ± 0.0723	0.3484 ± 0.1662	0.4316 ± 0.0134	1.4604 ± 0.1035	1.2351 ± 0.0586
FengWu	1.3715 ± 0.1011	1.2604 ± 0.0820	0.3221 ± 0.2004	0.4246 ± 0.0237	1.4179 ± 0.0986	1.2233 ± 0.0915
FuXi	1.4068 ± 0.1011	1.3023 ± 0.0789	0.2663 ± 0.2561	0.4279 ± 0.0212	1.4290 ± 0.0929	1.2236 ± 0.0961
Pangu-Weather 6h	1.3280 ± 0.0735	1.2081 ± 0.0607	0.4141 ± 0.1573	0.4017 ± 0.0245	1.4235 ± 0.0731	1.2225 ± 0.0847
AlphaEarth	$2.4583 \pm < 0.0001$	$2.1010 \pm < 0.0001$	$-0.2965 \pm < 0.0001$	$0.5036 \pm < 0.0001$	$1.2048 \pm < 0.0001$	$1.0188 \pm < 0.0001$

	Smoke $PM_{2.5}$			Extreme heat		
Backbone	**RMSE↓**	**MAE↓**	**Pearson r ↑**	**RMSE-C↓**	**MAE-C↓**	**Exceed. F_1 ↑**
Wildfire ref.	4.4684 ± 0.0060	2.4127 ± 0.0016	0.6360 ± 0.0013	0.2179 ± 0.0048	0.1787 ± 0.0020	0.9541 ± 0.0164
Prithvi-WxC	$5.9280 \pm < 0.0001$	$3.7304 \pm < 0.0001$	$0.0237 \pm < 0.0001$	4.6225 ± 0.0192	2.6315 ± 0.0128	0.8693 ± 0.0023
Aurora	$5.9281 \pm < 0.0001$	$3.7268 \pm < 0.0001$	$0.0189 \pm < 0.0001$	18.0474 ± 0.0708	15.3747 ± 0.0594	0.0951 ± 0.0038
ClimaX	$5.9299 \pm < 0.0001$	$3.7292 \pm < 0.0001$	$0.0007 \pm < 0.0001$	17.6492 ± 0.0347	14.4938 ± 0.0319	0.7684 ± 0.0068
StormCast	$6.0160 \pm < 0.0001$	$3.8189 \pm < 0.0001$	$0.0177 \pm < 0.0001$	1.7671 ± 0.2145	1.3507 ± 0.1576	0.9073 ± 0.0189
Pangu24	$5.9454 \pm < 0.0001$	$3.7479 \pm < 0.0001$	$0.0367 \pm < 0.0001$	4.9829 ± 0.3035	3.6298 ± 0.2130	0.7698 ± 0.0249
DLWP	5.9289 ± 0.1031	3.7331 ± 0.0088	0.0303 ± 0.0060	2.2662 ± 0.1106	1.7153 ± 0.0748	0.9156 ± 0.0112
FCN	5.9277 ± 0.1033	3.7345 ± 0.0088	0.0312 ± 0.0062	2.1657 ± 0.1800	1.6033 ± 0.1039	0.9257 ± 0.0096
FengWu	5.9297 ± 0.1032	3.7395 ± 0.0088	0.0304 ± 0.0063	2.1266 ± 0.1589	1.5801 ± 0.1004	0.0481 ± 0.0459
FuXi	5.9319 ± 0.1029	3.7398 ± 0.0088	0.0299 ± 0.0061	2.1282 ± 0.0969	1.5759 ± 0.0719	0.2268 ± 0.0623
Pangu-Weather 6h	5.9270 ± 0.1036	3.7320 ± 0.0088	0.0301 ± 0.0060	2.2045 ± 0.1483	1.6307 ± 0.0889	0.0199 ± 0.0062
AlphaEarth	4.4899 ± 0.0042	2.4009 ± 0.0020	0.6315 ± 0.0010	$0.2187 \pm < 0.0001$	$0.1795 \pm < 0.0001$	0.9542 ± 0.0107

D.4 WILDFIRE OCCUPANCY

Table 22 reports occupancy across the three matching rules. Each row follows one backbone from exact localization to tolerated and union correspondence. The reference has the highest mean under all three rules. Among the general-purpose comparators, AlphaEarth has the highest exact score, while ClimaX has the highest tolerated and union scores. This switch identifies different strengths in precise localization and recovery of spatially or temporally displaced events.

Table 22: Occupancy results under exact, tolerated, and union matching. Values are mean F_1 percentages with standard deviations. Bold marks the largest mean in each column.

Entry	Exact F_1 ↑	Tolerated F_1 ↑	union-F_1 ↑
Wildfire reference	5.31 ± 1.24	44.97 ± 2.15	66.74 ± 7.08
`Prithvi-WxC`	0.055 ± 0.004	7.16 ± 0.66	20.19 ± 1.83
`Aurora`	0.07 ± 0.01	8.50 ± 1.96	23.10 ± 4.94
`ClimaX`	0.35 ± 0.08	29.75 ± 3.61	60.15 ± 7.59
`StormCast`	0.056 ± 0.011	8.20 ± 2.19	22.38 ± 5.43
`Pangu24`	0.055 ± 0.008	8.18 ± 1.35	22.06 ± 3.73
AlphaEarth	2.06 ± 0.44	29.45 ± 6.01	37.43 ± 9.95
`DLWP`	0.17 ± 0.04	14.91 ± 3.24	28.19 ± 6.97
`FCN`	0.28 ± 0.08	19.51 ± 3.34	40.06 ± 9.37
`FengWu`	0.26 ± 0.08	12.00 ± 6.02	24.10 ± 13.63
`FuXi`	0.38 ± 0.12	21.03 ± 4.82	37.29 ± 9.45
`Pangu-Weather 24h`	0.28 ± 0.14	11.34 ± 5.22	26.58 ± 12.36
`Pangu-Weather 6h`	0.28 ± 0.11	17.09 ± 4.05	35.64 ± 9.03

D.5 FIRE SPREAD

Table 23 combines exact overlap, spatial agreement, and AP. The wildfire reference leads the frozen comparison on all three metrics, and Aurora has the strongest general-purpose means across those columns. Other comparators change relative position between thresholded and ranked views: AlphaEarth's AP exceeds ClimaX's, while ClimaX has higher exact and spatial F_1. These complementary views describe both the spatial extent of a chosen prediction and the ordering of its probability scores.

Table 23: Fire-spread results. Exact F_1, spatial F_1, and AP are percentages; entries show means with standard deviations over five seeds. Bold marks the largest mean per metric.

Entry	Exact $F_1 \uparrow$	Spatial $F_1 \uparrow$	AP↑
Wildfire reference	37.98 ± 1.33	82.20 ± 0.32	32.73 ± 0.60
`Prithvi-WxC`	22.14 ± 0.40	58.18 ± 0.14	4.83 ± 0.17
`Aurora`	31.08 ± 0.40	71.82 ± 0.14	17.77 ± 0.43
`ClimaX`	$27.99 \pm < 0.01$	$69.06 \pm < 0.01$	11.17 ± 0.23
`StormCast`	21.95 ± 0.37	57.90 ± 0.23	8.11 ± 1.16
`Pangu24`	21.87 ± 0.42	57.80 ± 0.57	9.95 ± 1.60
`DLWP`	22.26 ± 1.41	58.71 ± 1.91	9.21 ± 2.66
`FCN`	20.60 ± 2.36	57.15 ± 1.39	6.68 ± 1.30
`FengWu`	22.50 ± 1.78	58.20 ± 0.84	11.00 ± 2.71
`FuXi`	25.16 ± 0.55	59.57 ± 0.48	13.55 ± 0.38
`Pangu-Weather 6h`	22.29 ± 0.97	58.43 ± 0.83	10.63 ± 1.46
AlphaEarth	16.19 ± 4.79	43.47 ± 10.28	12.28 ± 1.36

D.6 FINAL BURNED AREA

Table 24 evaluates event magnitude through log-RMSE and log-MAE, together with event ordering through Spearman correlation. The wildfire reference combines the smallest errors with the largest correlation. DLWP is the leading general-purpose comparator on all three measures, followed by Pangu-Weather 6h on these same means. AlphaEarth's negative correlation indicates an inverse ordering of predicted and observed event sizes in this evaluation. Its different behavior on retrieval below shows why event regression and analog search deserve separate treatment.

Table 24: Full reported matrix for final burned area. Event-level regression keeps one log-error contract across the wildfire reference, frozen comparators, and AlphaEarth.

Entry	log-RMSE↓	log-MAE↓	Spearman↑
Wildfire reference	1.1649 ± 0.0139	1.0355 ± 0.0114	0.5884 ± 0.0338
`Prithvi-WxC`	1.3630 ± 0.0681	1.2435 ± 0.0668	0.1799 ± 0.3002
`Aurora`	1.8658 ± 0.2009	1.6717 ± 0.1245	-0.1156 ± 0.2982
`ClimaX`	2.0300 ± 0.2103	1.8443 ± 0.1528	-0.2515 ± 0.2688
`StormCast`	1.6679 ± 0.1438	1.4745 ± 0.1134	0.1830 ± 0.1969
`Pangu24`	1.3820 ± 0.0638	1.2369 ± 0.0967	0.2309 ± 0.1358
`DLWP`	1.3070 ± 0.0980	1.1769 ± 0.0834	0.4888 ± 0.1368
`FCN`	1.3693 ± 0.0885	1.2599 ± 0.0723	0.3484 ± 0.1662
`FengWu`	1.3715 ± 0.1011	1.2604 ± 0.0820	0.3221 ± 0.2004
`FuXi`	1.4068 ± 0.1011	1.3023 ± 0.0789	0.2663 ± 0.2561
`Pangu-Weather 6h`	1.3280 ± 0.0735	1.2081 ± 0.0607	0.4141 ± 0.1573
AlphaEarth	$2.4583 \pm < 0.0001$	$2.1010 \pm < 0.0001$	$-0.2965 \pm < 0.0001$

Table 25: Full reported matrix for analog retrieval under the early-deployment retrieval contract used in the main text.

Entry	nDCG@10↑	log-RMSE↓	log-MAE↓
Wildfire reference	$0.5178 \pm < 0.0001$	$1.1769 \pm < 0.0001$	$0.9683 \pm < 0.0001$
`Prithvi-WxC`	0.3857 ± 0.0189	1.3908 ± 0.0938	1.2585 ± 0.0865
`Aurora`	0.4046 ± 0.0144	1.3659 ± 0.0792	1.2596 ± 0.0968
`ClimaX`	0.4143 ± 0.0191	1.4526 ± 0.0926	1.2441 ± 0.1446
`StormCast`	0.4076 ± 0.0094	1.3663 ± 0.0781	1.2371 ± 0.1078
`Pangu24`	0.3866 ± 0.0148	1.4057 ± 0.0769	1.1840 ± 0.1052
`DLWP`	0.3972 ± 0.0146	1.5351 ± 0.0802	1.3196 ± 0.0781
`FCN`	0.4316 ± 0.0134	1.4604 ± 0.1035	1.2351 ± 0.0586
`FengWu`	0.4246 ± 0.0237	1.4179 ± 0.0986	1.2233 ± 0.0915
`FuXi`	0.4279 ± 0.0212	1.4290 ± 0.0929	1.2236 ± 0.0961
`Pangu-Weather 6h`	0.4017 ± 0.0245	1.4235 ± 0.0731	1.2225 ± 0.0847
AlphaEarth	$0.5036 \pm < 0.0001$	$1.2048 \pm < 0.0001$	$1.0188 \pm < 0.0001$

D.7 Analog Retrieval

Table 25 evaluates both the relevance of the top-ranked analogs and the log-area errors associated with retrieval. The reference and AlphaEarth form the leading pair on all three metrics. Their nDCG@10 means are close, while the log-MAE and log-RMSE columns quantify the associated event-size agreement. FCN has the next-highest nDCG@10, whereas Aurora and StormCast attain smaller log-RMSE than FCN. Thus, the ranking of relevant analogs and the numerical accuracy of retrieved event properties provide complementary descriptions of retrieval quality.

D.8 Smoke $PM_{2.5}$

Table 26 pairs concentration errors with Pearson correlation. The wildfire reference has the smallest RMSE and highest correlation; AlphaEarth has the smallest MAE and closely aligned RMSE and correlation. The remaining comparators cluster around larger errors and correlations close to zero. The contrast connects average concentration accuracy to the strength of association across observations and highlights why the error and correlation columns should be read together.

Table 26: Full reported matrix for smoke $PM_{2.5}$. Station regression uses one continuous metric family with RMSE, MAE, and Pearson r.

Entry	RMSE↓	MAE↓	Pearson r ↑
Wildfire reference	4.4684 ± 0.0060	2.4127 ± 0.0016	0.6360 ± 0.0013
`Prithvi-WxC`	$5.9280 \pm < 0.0001$	$3.7304 \pm < 0.0001$	$0.0237 \pm < 0.0001$
`Aurora`	$5.9281 \pm < 0.0001$	$3.7268 \pm < 0.0001$	$0.0189 \pm < 0.0001$
`ClimaX`	$5.9299 \pm < 0.0001$	$3.7292 \pm < 0.0001$	$0.0007 \pm < 0.0001$
`StormCast`	$6.0160 \pm < 0.0001$	$3.8189 \pm < 0.0001$	$0.0177 \pm < 0.0001$
`Pangu24`	$5.9454 \pm < 0.0001$	$3.7479 \pm < 0.0001$	$0.0367 \pm < 0.0001$
`DLWP`	5.9289 ± 0.1031	3.7331 ± 0.0088	0.0303 ± 0.0060
`FCN`	5.9277 ± 0.1033	3.7345 ± 0.0088	0.0312 ± 0.0062
`FengWu`	5.9297 ± 0.1032	3.7395 ± 0.0088	0.0304 ± 0.0063
`FuXi`	5.9319 ± 0.1029	3.7398 ± 0.0088	0.0299 ± 0.0061
`Pangu-Weather 6h`	5.9270 ± 0.1036	3.7320 ± 0.0088	0.0301 ± 0.0060
AlphaEarth	4.4899 ± 0.0042	2.4009 ± 0.0020	0.6315 ± 0.0010

D.9 Extreme Heat

Table 27 combines Celsius-regression accuracy with validation-selected exceedance detection. The reference and AlphaEarth have closely aligned error and exceedance scores. Several other backbones have similar continuous errors and markedly different exceedance F_1: FCN and FengWu have RMSE-C values near 2.1, while their exceedance scores are 0.9257 and 0.0481. The two views expose different properties of temperature estimates and threshold crossings. Prithvi-WxC uses the native MERRA2 and climatology batch construction for its feature extraction.

Table 27: Extreme-heat results. RMSE-C and MAE-C are in degrees Celsius; exceedance F_1 is fractional. Entries show means with standard deviations, and bold marks the best mean per metric.

Entry	RMSE-C↓	MAE-C↓	Exceed. F_1 ↑
Wildfire reference	0.2179 ± 0.0048	0.1787 ± 0.0020	0.9541 ± 0.0164
`Prithvi-WxC`	4.6225 ± 0.0192	2.6315 ± 0.0128	0.8693 ± 0.0023
`Aurora`	18.0474 ± 0.0708	15.3747 ± 0.0594	0.0951 ± 0.0038
`ClimaX`	17.6492 ± 0.0347	14.4938 ± 0.0319	0.7684 ± 0.0068
`StormCast`	1.7671 ± 0.2145	1.3507 ± 0.1576	0.9073 ± 0.0189
`Pangu24`	4.9829 ± 0.3035	3.6298 ± 0.2130	0.7698 ± 0.0249
`DLWP`	2.2662 ± 0.1106	1.7153 ± 0.0748	0.9156 ± 0.0112
`FCN`	2.1657 ± 0.1800	1.6033 ± 0.1039	0.9257 ± 0.0096
`FengWu`	2.1266 ± 0.1589	1.5801 ± 0.1004	0.0481 ± 0.0459
`FuXi`	2.1282 ± 0.0969	1.5759 ± 0.0719	0.2268 ± 0.0623
`Pangu-Weather 6h`	2.2045 ± 0.1483	1.6307 ± 0.0889	0.0199 ± 0.0062
AlphaEarth	$0.2187 \pm < 0.0001$	$0.1795 \pm < 0.0001$	0.9542 ± 0.0107

E TEMPORAL AND GEOGRAPHIC ROBUSTNESS

We separate training variability from transfer across years and regions. The following records specify each split and its uncertainty summary.

Complementary evaluation scales. The California comparison fixes a season and test population, allowing matching, scope, and head selection to be varied on common examples. The 120 six-hourly October maps provide repeated observations of that population; nearby cells and consecutive maps can share fire events. The cross-year split changes the test year, while LORO changes the evaluated geography and excludes intersecting training tiles. These checks examine different forms of transfer. Together, the fixed-population controls and the temporal and geographic checks connect evaluation sensitivity to the conditions under which the predictions are used.

E.1 CROSS-YEAR EVALUATION

The cross-year experiment evaluates transfer across calendar years on a shared lower-48 grid. The setup table below records March–October data from 2022–2025 at 5-km resolution in EPSG:5070. Training uses 2022–2023, validation uses 2024, and testing uses 2025, with 1,952, 980, and 976 matched temporal-map instances. Counts follow intersected FIRMS and HRRR availability; the October records for 2022, 2023, and 2025 extend through October 30. All five seeds share the candidate-head family and validation threshold rule.

The setup separates the years used for learning, model selection, and testing while preserving the geographic grid. This makes temporal transfer the central comparison and provides a complementary axis to the geographic holdouts below.

Table 29 applies the three matching rules to the cross-year occupancy outputs. Exact, tolerated, and union F_1 are $7.88 \pm 2.71\%$, $47.06 \pm 1.72\%$, and $95.19 \pm 2.42\%$, respectively. The table also provides bootstrap intervals for interpreting these estimates.

Paired seed-wise comparisons connect the three scores directly. The tolerated-minus-exact gain is 39.18 percentage points (bootstrap 95% CI $[38.29, 40.05]$); union-minus-exact is 87.31 points ($[86.47, 88.30]$); and union-minus-tolerated is 48.13 points ($[47.25, 48.98]$). Each paired sign-flip test gives a one-sided $p = 0.0312$ and a two-sided $p = 0.0625$. The paired gains show that the distinction between exact localization and tolerated correspondence persists through the year-based split.

E.2 REGION-DISJOINT EVALUATION

The geographic experiment uses the lower-48 2024 occupancy setup with weather, static covariates, and lagged FIRMS observations. Each regional model trains on 32×32 tiles lying fully outside its held-out latitude–longitude mask and is evaluated on valid cells inside that mask. Training tiles remain fully separated from the held-out mask, including its boundary.

The six masks are California/Pacific ($32.0°$–$42.5°$N, $125.0°$–$113.5°$W), Pacific Northwest ($42.0°$–$49.5°$N, $125.0°$–$111.0°$W), Rockies/Intermountain ($31.0°$–$49.5°$N, $116.0°$–$102.0°$W), Southeast/Florida ($24.0°$–$32.5°$N, $88.0°$–$79.5°$W), Plains/Central ($28.0°$–$49.5°$N, $105.0°$–$90.0°$W), and Eastern U.S. ($25.0°$–$49.5°$N, $90.0°$–$66.0°$W).

Table 30 reports means and standard deviations over three seeds per mask. Its ranking, recall, and predicted-area measures describe complementary aspects of geographic transfer.

AUROC ranges from 0.7992 ± 0.0439 in Plains/Central to 0.9034 ± 0.0150 in Southeast/Florida. Top-1% recall spans $23.95 \pm 2.86\%$ in the Eastern U.S. and $68.50 \pm 1.94\%$ in the Rockies/Intermountain region. The representation maintains ranked discrimination across all six holdouts, while the thresholded recall and predicted-area profiles reveal regional differences in the mapping from scores to decisions. This combination motivates explicit calibration studies across regional fire prevalence and environmental regimes.

E.3 PRIMARY-RESULT CONFIDENCE INTERVALS

Table 31 presents the primary occupancy and spread means with two-sided 95% confidence intervals across the same five seeds as the main comparison. Standard deviations describe variation among

Table 28: Lower-48 cross-year robustness setup for the occupancy check.

Item	Value
Geographic scope	lower-48 CONUS
Date span	March–October 2022–2025
Split	train 2022/2023, validation 2024, test 2025
Grid and projection	5 km, EPSG:5070
Samples	3908 total; 1952 train / 980 val / 976 test
Inputs	HRRR weather, FIRMS history, and static covariates
Head family	same lightweight head family across seeds
Threshold rule	validation exact F_1 selects the operating threshold

Table 29: Five-seed lower-48 cross-year occupancy contract summary. Exact is retained as a strict diagnostic anchor, while tolerated and union are the decision-relevant readings.

Reading	Mean $\pm$ std (%)	Bootstrap 95% CI (%)
Exact	7.88 ± 2.71	[5.73, 9.93]
Tolerated	47.06 ± 1.72	[45.55, 48.29]
Union	95.19 ± 2.42	[93.05, 96.73]

runs; confidence intervals describe uncertainty in their estimated mean. Reading the interval widths alongside the per-metric ranking connects the average comparison to training-seed variation.

Table 30: Three-seed LORO occupancy robustness across six held-out regions. Models train outside and evaluate inside each mask; values are mean ± SD over seeds $\{7, 42, 99\}$. AUROC and PR-AUC are threshold-free; top-1% recall measures positives in the highest-scored 1% of cells, while recall and predicted area use threshold 0.10. Except for AUROC, values are percentages.

Held-out region	AUROC↑	PR-AUC↑	Top-1% recall↑	Recall@0.10↑	Pred. area@0.10↓
California/Pacific	0.8986 ± 0.0143	3.6283 ± 0.4037	41.2725 ± 4.4406	44.7917 ± 4.7337	2.0442 ± 1.3536
Pacific Northwest	0.8642 ± 0.0277	17.8807 ± 0.7759	61.1818 ± 2.9223	75.1202 ± 15.1595	21.2648 ± 28.5915
Rockies/Intermountain	0.8995 ± 0.0163	17.3278 ± 2.7126	68.4966 ± 1.9388	64.9116 ± 2.7555	0.6068 ± 0.1591
Southeast/Florida	0.9034 ± 0.0150	1.3689 ± 0.1082	28.6058 ± 0.2734	51.3460 ± 30.8009	10.4995 ± 14.8820
Plains/Central	0.7992 ± 0.0439	2.2962 ± 0.4684	31.5604 ± 5.0182	65.8487 ± 13.2612	24.9983 ± 19.4363
Eastern U.S.	0.8640 ± 0.0046	0.7452 ± 0.0893	23.9518 ± 2.8563	42.0546 ± 37.8204	10.5803 ± 17.4068

Table 31: Primary fixed-contract results with 95% confidence intervals. Each cell reports mean ± SD above and the 95% CI below, in percent. Two-sided t intervals use the five seeds in Table 3; the two spread SDs reported as < 0.01 use 0.01 as a conservative upper bound.

	Occupancy			Fire spread		
Comparator	**Exact F_1**	**Tol. F_1**	**Union F_1**	**Exact F_1**	**Spatial F_1**	**AP**
Wildfire ref.	5.31 ± 1.24 CI [3.77, 6.85]	44.97 ± 2.15 CI [42.30, 47.64]	66.74 ± 7.08 CI [57.95, 75.53]	37.98 ± 1.33 CI [36.33, 39.63]	82.20 ± 0.32 CI [81.80, 82.60]	32.73 ± 0.60 CI [31.99, 33.47]
Prithvi-WxC	0.055 ± 0.004 CI [0.050, 0.060]	7.16 ± 0.66 CI [6.34, 7.98]	20.19 ± 1.83 CI [17.92, 22.46]	22.14 ± 0.40 CI [21.64, 22.64]	58.18 ± 0.14 CI [58.01, 58.35]	4.83 ± 0.17 CI [4.62, 5.04]
Aurora	0.07 ± 0.01 CI [0.058, 0.082]	8.50 ± 1.96 CI [6.07, 10.93]	23.10 ± 4.94 CI [16.97, 29.23]	31.08 ± 0.40 CI [30.58, 31.58]	71.82 ± 0.14 CI [71.65, 71.99]	17.77 ± 0.43 CI [17.24, 18.30]
ClimaX	0.35 ± 0.08 CI [0.251, 0.449]	29.75 ± 3.61 CI [25.27, 34.23]	60.15 ± 7.59 CI [50.73, 69.57]	$27.99\pm < 0.01$ CI [27.98, 28.00]	$69.06\pm < 0.01$ CI [69.05, 69.07]	11.17 ± 0.23 CI [10.88, 11.46]
StormCast	0.056 ± 0.011 CI [0.042, 0.070]	8.20 ± 2.19 CI [5.48, 10.92]	22.38 ± 5.43 CI [15.64, 29.12]	21.95 ± 0.37 CI [21.49, 22.41]	57.90 ± 0.23 CI [57.61, 58.19]	8.11 ± 1.16 CI [6.67, 9.55]
Pangu24	0.055 ± 0.008 CI [0.045, 0.065]	8.18 ± 1.35 CI [6.50, 9.86]	22.06 ± 3.73 CI [17.43, 26.69]	21.87 ± 0.42 CI [21.35, 22.39]	57.80 ± 0.57 CI [57.09, 58.51]	9.95 ± 1.60 CI [7.96, 11.94]
DLWP	0.17 ± 0.04 CI [0.120, 0.220]	14.91 ± 3.24 CI [10.89, 18.93]	28.19 ± 6.97 CI [19.54, 36.84]	22.26 ± 1.41 CI [20.51, 24.01]	58.71 ± 1.91 CI [56.34, 61.08]	9.21 ± 2.66 CI [5.91, 12.51]
FCN	0.28 ± 0.08 CI [0.181, 0.379]	19.51 ± 3.34 CI [15.36, 23.66]	40.06 ± 9.37 CI [28.43, 51.69]	20.60 ± 2.36 CI [17.67, 23.53]	57.15 ± 1.39 CI [55.42, 58.88]	6.68 ± 1.30 CI [5.07, 8.29]
FengWu	0.26 ± 0.08 CI [0.161, 0.359]	12.00 ± 6.02 CI [4.53, 19.47]	24.10 ± 13.63 CI [7.18, 41.02]	22.50 ± 1.78 CI [20.29, 24.71]	58.20 ± 0.84 CI [57.16, 59.24]	11.00 ± 2.71 CI [7.64, 14.36]
FuXi	0.38 ± 0.12 CI [0.231, 0.529]	21.03 ± 4.82 CI [15.05, 27.01]	37.29 ± 9.45 CI [25.56, 49.02]	25.16 ± 0.55 CI [24.48, 25.84]	59.57 ± 0.48 CI [58.97, 60.17]	13.55 ± 0.38 CI [13.08, 14.02]
Pangu-Weather 6h	0.28 ± 0.11 CI [0.143, 0.417]	17.09 ± 4.05 CI [12.06, 22.12]	35.64 ± 9.03 CI [24.43, 46.85]	22.29 ± 0.97 CI [21.09, 23.49]	58.43 ± 0.83 CI [57.40, 59.46]	10.63 ± 1.46 CI [8.82, 12.44]
AlphaEarth	2.06 ± 0.44 CI [1.51, 2.61]	29.45 ± 6.01 CI [21.99, 36.91]	37.43 ± 9.95 CI [25.08, 49.78]	16.19 ± 4.79 CI [10.24, 22.14]	43.47 ± 10.28 CI [30.71, 56.23]	12.28 ± 1.36 CI [10.59, 13.97]

F Discussion: From Benchmark Scores to Transfer Science

The experimental sections establish how evaluation choices alter scores and how the wildfire reference compares with general-purpose representations. The broader research opportunity is to use those relationships to study what a representation makes accessible, at which spatial scales, through which adaptation procedures, and for which decisions. This discussion develops that opportunity into a sequence of testable research directions.

F.1 Evaluation Contracts Define the Capability Being Measured

An evaluation contract specifies the capability to which a score refers. Exact occupancy evaluates cell–time localization. Tolerated correspondence evaluates recovery within a neighborhood. Retrieval evaluates the ordering of useful precedents, while regression evaluates numerical properties of events or environmental conditions. Each score answers a particular question about how a representation supports a task. The contract supplies the scientific meaning of the measurement.

This perspective gives matching sensitivity a constructive role. A response curve across spatial and temporal tolerances can reveal the scale at which a model captures predictive structure. A model whose score rises rapidly under a small spatial tolerance may encode regional fire conditions together with localization error. A model whose score improves primarily with temporal tolerance may capture event development with a timing offset. These are testable hypotheses about the structure of prediction errors. A future study could vary spatial and temporal tolerances independently, associate matched events with displacement vectors and forecast lead times, and compare the resulting profiles

with known uncertainty in observations and grid alignment. This would turn a verification sweep into a diagnostic of predictive scale.

F.2 Contract Sensitivity as a Representation Profile

The score matrix $q_b(\xi)$ describes how each backbone responds to evaluation choices. Its configuration main effect captures changes shared across models, and its interaction term $\gamma_{b,\xi}$ captures model-specific responses. CDR and RIR provide compact summaries of score movement and ranking reversals; the interaction matrix provides a richer object for studying representation behavior. Backbones with similar response profiles may share predictive scales, calibration patterns, or dependence on the prevalence of fire in the evaluated population.

A useful next study would relate these response profiles to pretraining resolution, input modality, temporal context, and downstream head capacity. For example, the hypothesis that high-resolution land representations support precise localization can be tested against matching-response curves, while weather-dominated representations can be examined for gains at broader spatial or temporal scales. Clustering the response profiles and evaluating the groups on held-out regions and tasks would test whether contract sensitivity predicts transfer behavior. Benchmark variation would then become an empirical signal for understanding which information is accessible from a representation.

F.3 Domain Specialization and Adaptation

A wildfire-specific reference adds a domain-trained point of comparison to the general-purpose model set. Its value lies in locating tasks where domain specialization yields a substantial performance difference and tasks where general-purpose representations approach that point of comparison. The cross-task profile motivates a deeper question: how much downstream adaptation is required to recover each type of wildfire-relevant information?

A factorial study could compare general-purpose and wildfire-specific pretraining using matched input channels, architecture, pretraining compute, downstream examples, and evaluation contracts. Within each condition, head capacity could increase from linear probes to shallow adapters, followed by parameter-efficient and full fine-tuning with explicit compute budgets. A gap that closes as head capacity increases would support an explanation based on the accessibility of predictive information. A gap that closes with backbone updates would point toward the importance of representation adaptation. Learning curves across label budgets would further measure the sample efficiency of each route. This design connects the wildfire reference to a mechanistic study of specialization, adaptation cost, and data efficiency.

F.4 Connecting Transfer to Operational Decisions

The regional results motivate a distinction between ranking events and assigning decision-ready probabilities. A score ordering can remain useful across regions while the threshold needed for a given recall, alert area, or false-alarm rate changes with local prevalence and environmental conditions. Calibration forms a natural connection between representation transfer and regional decision making.

Future experiments could combine spatial holdouts with validation-based calibration and explicit response budgets. One study could fix an alert-area budget and measure fire-event recall across regions and forecast horizons. Another could assign operationally motivated costs to missed events, false alerts, and warning delay, then examine how the preferred backbone and head change with those costs. Calibration methods could be compared using the same regional validation data and evaluated on a later season. Together, these experiments would connect the matching, scope, and head-selection components of a contract to the information and resources available to a decision maker.

The resulting evidence could inform model selection for early warning and resource allocation by making the consequences of each evaluation choice explicit. Operational utility would be measured through event coverage, alert burden, and lead time, alongside the predictive metrics used to construct the models.

F.5 Building Cumulative Evidence Across Studies

Contract-aware evaluation also creates an opportunity for cumulative benchmarking. A machine-readable record can connect each score to its task, data version, split, spatial population, matching parameters, selection rule, and head family. Prediction-level records can support repeated verification;

candidate-head records can support alternative selection analyses; per-seed records can support paired comparisons and uncertainty estimates. These artifacts make the path from representation to published score traceable.

An interesting extension is a shared evaluation service that re-scores stored predictions across a predefined contract grid and reports performance profiles together with pairwise ranking stability. Contract choices derived from scientific or operational requirements could be registered before evaluating new backbones. Cross-study synthesis could then identify conclusions that persist across evaluation settings and map conclusions that vary to the specific decisions responsible for that variation. In this view, progress in Earth foundation models is measured through increasingly clear evidence about which capabilities transfer, how efficiently they can be adapted, and where they contribute to useful environmental decisions.