TEMPORAL RESOLUTION CONDITIONS THE GLOBAL–LOCAL TRADE-OFF IN WILDFIRE SPREAD PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Which inductive bias should a spatiotemporal operator have? We compare a gated global–local operator—spectral graph mixing fused per cell with local convolution—against convolutional recurrent baselines on 27 cross-state wildfire cases, and find that the answer depends on how far the fire front moves per model step. On the slowest fires of the main pool, under per-fire ROC-AUC, the operator ties the recurrent baseline at $h{=}1$ and leads it from $h{=}3$ on, the margin growing with horizon ($+0.006$, $+0.047$, $+0.064$ on a 6 h lattice); on the fastest it loses, and removing either branch, adding spectral or local capacity, or widening the training ring does not close that gap. A preregistered intervention that doubles per-step travel on the same fires by re-indexing them onto steps of twice the length moves the winner across families ($+0.094$ ROC-AUC, $p{=}0.0022$), so on these fires the dependence is interventional, not merely observed. The mechanism is a division of labor: the global path should *predict* for slow long-range spread and *condition* the local path on fast short-range steps, and neither role substitutes for the within-window local recurrence that carries fast regimes. The same ordering holds on the 419-fire WSTS benchmark, where every fire lies at the fast end. The ranking is a property of the (architecture, discretization) pair rather than of the architecture alone. All comparisons are graded verified, faithful, or historical under a protocol-recovery audit.

1 INTRODUCTION

Wildfires are growing in frequency and severity, and short-horizon spread forecasts inform evacuation and suppression. Machine learning is now standard across wildfire science (Jain et al., 2020), with public rasterized benchmarks (§2). The dominant architectures are *local*: convolutional encoders and convolutional recurrent models whose receptive fields grow with depth (Shi et al., 2015; Burge et al., 2023). In parallel, operator learning has produced *global* architectures—Fourier neural operators and spectral graph filters (Li et al., 2021; Kovachki et al., 2023; Defferrard et al., 2016)—that mix information across the whole domain in a truncated frequency basis. Both are natural candidates for fire spread, yet they encode opposite *spatiotemporal inductive biases*: global operators favor long-range, low-frequency structure, local recurrent models short-range propagation and spatial memory.

Two gaps motivate this paper. First, the two families are rarely compared under a common protocol on real fire data: benchmark evaluations are typically teacher-forced and single-step, whereas deployment is a free-running rollout in which predictions feed back into the next input and small errors compound. Second, there is no reason to expect a single winner: a front that advances a fraction of a cell per model step presents a different learning problem from one that jumps several cells per step, and cells moved per step depend not only on the fire but on the *temporal resolution* the modeler chooses. To our knowledge this interaction has not been isolated on real wildfire data.

Figure 1 previews what we find: sorting test fires by measured per-step front displacement, the two single-mechanism families move in *opposite* directions and their ordering crosses—ConvLSTM improves ($0.685 \to 0.695 \to 0.801$) while base-FNO, the global spectral path in isolation, degrades ($0.736 \to 0.731 \to 0.662$). The gated operator sits above the global path at every band by a similar

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.

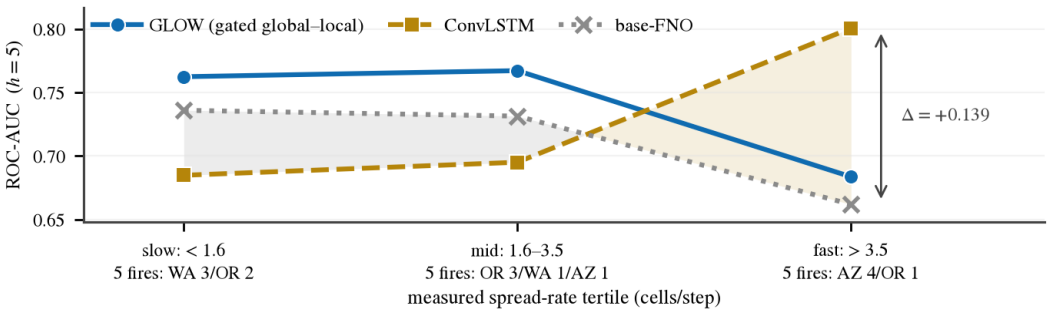


Figure 1: **The phenomenon: a scissor-like crossing, not a uniform ranking.** h=5 free-running ROC-AUC (per-fire, seed-mean) against *measured* per-step front displacement, tertile-binned over the 15 test fires with $\geq$10 positive cells at h=5. ConvLSTM improves as fires speed up while base-FNO, the global spectral path in isolation, degrades, so their ordering crosses. The bands are also where the state labels mislead: the slowest tertile is Washington-majority (WA 3/OR 2) and the fastest Arizona-majority (AZ 4/OR 1), so state and spread rate are entangled by construction—which is why we organize the paper by measured spread rate rather than by state name.

margin ($+0.026$, $+0.036$, $+0.022$), so the fusion buys a roughly uniform gain rather than changing the regime slope; against ConvLSTM the sign does flip ($+0.078$, $+0.072$, -0.117). This scissor-like crossing, rather than a fixed ranking, is what the paper explains and then tries to exploit.

We study it with **GLOW** (*G*lobal–*LO*cal operator for *W*ildfire spread), a lightweight global–local operator (0.8M parameters): a shared short-window temporal aggregator feeds a global spectral path (learned mixing in a truncated graph-Laplacian eigenbasis) and a local convolutional path, fused per cell by a learned gate. Both paths sit inside one backbone, so branch variants are controlled ablations. We compare this family with ConvLSTM, graph, pointwise, and persistence baselines under strict cross-state transfer (train on California, test on 27 unseen fires in Oregon, Washington, and Arizona) and on the independent WildfireSpreadTS benchmark, in teacher-forced and free-running modes.

Our contributions are:

- **A regime-conditioned ranking on a unified spread-rate axis.** Per-step newly burned cells, a resolution-dependent growth measure, organizes when each family wins: on the main pool's slow fires the global–local operator leads the recurrent baseline from h=3 on with the margin growing over the horizon, while convolutional recurrence dominates fast fires; the crossover is visible per fire rather than per state (§5.3).
- **A designed intervention on temporal resolution.** Re-indexing the same fires from 6-hour to 12-hour steps—which exactly doubles per-step travel—moves the winner from the global–local operator to ConvLSTM on 7 of 14 fires ($+0.094$ ROC shift, p=0.0022, preregistered direction), upgrading the correlation to an interventional claim (§5.4).
- **A mechanism chain and a protocol-recovery audit.** Matched-normalization retraining reproduces the historical main table; removing a branch changes which regime wins rather than uniformly hurting performance (at h=5 the spectral-only arm matches the dual-branch model on two of three states); pure-focal training, widened rings, gate-capacity increases (normalization-confounded) and unconstrained fusion do not unlock fast-fire performance; and the global path participates in a regime-dependent way, with the local recurrence carrying fast regimes (§5.3, §5.2). The main GLOW checkpoints are retrained under frozen normalization contracts, and every comparison is graded verified, faithful, or historical (Appendix J).

2 Related Work

2.1 Wildfire spread prediction

Machine learning is now applied across wildfire science, from occurrence and risk to behavior prediction (Jain et al., 2020). Physics-based systems such as Rothermel's surface-spread model (Rother-

mel, 1972) give mechanistic guidance but need event-specific fuel and environmental calibration (Scott & Burgan, 2005). Data-driven benchmarks formulate spread as rasterized next-step or next-day burn-map prediction: Next Day Wildfire Spread (Huot et al., 2022), and WildfireSpreadTS (WSTS) (Gerard et al., 2023) with U-Net (Ronneberger et al., 2015), ConvLSTM (Shi et al., 2015) and UTAE (Sainte Fare Garnot & Landrieu, 2021) baselines, later improved by WSTS+ (Lahrichi et al., 2026); on simulated fires, convolutional recurrent models sustain long autoregressive rollouts (Burge et al., 2023). Across this literature the architectures are predominantly local, and comparisons are made at a fixed spatiotemporal resolution under teacher-forced or single-horizon evaluation.

We complement it by testing California-to-27-fire transfer with free-running feedback, and by treating the temporal sampling interval itself as an experimental variable. The question is whether rankings remain stable when the protocol changes, not which model wins one leaderboard.

2.2 Global, local, and physics-motivated operators

Fourier neural operators (Li et al., 2021) and graph spectral filters use a truncated basis to mix spatial information globally; neural operators formalize this as learning maps between function spaces (Kovachki et al., 2023), extended to discretized dynamical systems by mesh-based graph simulators (Pfaff et al., 2021). In contrast, graph polynomial filters such as ChebNet (Defferrard et al., 2016) and convolutional recurrent models such as ConvLSTM preserve finite-range propagation and spatial memory. These are different architectural preferences, not competing claims that one operator is universally superior. Frequency-aware geospatial modeling such as GeoFAR (Xu et al., 2026) also separates low-frequency structure from local detail, but addresses a different task. Physics-informed learning (Raissi et al., 2019) is relevant too: because the Rothermel rate is latent and labels are coarse masks, we test its residual as a factor rather than an observed constraint. Per-comparison evidence grades for the conditional returns of this family are reported in Appendix J.

Finally, our gate and expert analyses relate to mixture-of-experts (Shazeer et al., 2017) and exogenous regime switching (Hamilton, 1989): a jointly trained fusion gate and an externally selected expert have different information and optimization requirements, and our protocol analysis aligns with recent scientific-ML common-task efforts (Yermakov et al., 2026). We do not propose a new operator primitive; the contribution is to characterize how existing biases and protocols exchange returns when the sampling interval changes. In particular, the ranking is a property of the (architecture, discretization) pair rather than of the architecture alone: a benchmark that fixes one interval and evaluates teacher-forced at a single step characterizes a discretization, not a bias family.

3 Problem Formulation

3.1 Wildfire spread as a candidate-grid transition

Let G be a regular spatial grid and let $b_t \in \{0,1\}^{|G|}$ denote the cumulative burned-state map at model time t. Each cell also has static and exogenous covariates e_t (terrain, wind, and, when available, meteorological fields). We define the frontier candidate set

$$C_t = \{i \in G : b_t(i) = 0,\ \exists j \in \mathcal{N}(i) \text{ with } b_t(j) = 1\},$$

where $\mathcal{N}(i)$ is the eight-neighborhood. The model produces a probability for every grid cell,

$$\hat{p}_{t+1} = F_\theta\big(x_{t-k:t}\big), \qquad x_t = \phi(b_t, e_t),$$

but the data loss and the primary metrics are evaluated on C_t. The target $y_{t+1,i}$ records whether candidate i is newly ignited in the next labeled time interval (with the dataset-specific future tolerance). This candidate-conditioned formulation isolates the spread front while retaining a full-grid state for the subsequent rollout. It also bounds the label set: cells beyond one ring are not scored at that step, so faster growth loads more signal onto each scored transition rather than expanding the scored set (coverage and its growth dependence: Appendix A).

We distinguish two evaluation modes. In teacher-forced single-step prediction, b_t is the observed state at every step. In a free-running rollout, only the initial state is observed and the predicted burn state is fed back. In the cumulative mode used for the main results,

$$b_{t+1} = b_t \vee \mathbf{1}[\hat{p}_{t+1} > \tau].$$

Thus burn and neighbor-burn channels are endogenous after the first step, whereas terrain is fixed and wind—the only time-varying exogenous field in the main configuration—is taken from the corresponding observed time. The loop is therefore closed in fire state and open in weather, identically for every compared model. To keep horizons comparable, rollout metrics at horizon h are computed on the *observed* candidate set C_{t+h-1}; only the model inputs are endogenous. We do not model interior-fire extinction or latent temperature states. This is a burn-state rollout abstraction, not a complete physical fire simulator.

4 METHOD

4.1 GATED GLOBAL–LOCAL OPERATOR

GLOW is a lightweight operator family rather than a new primitive (Figure 2). A per-cell encoder maps the input features to a latent width $D = 64$.

Temporal aggregation. The reported configuration uses a window of $K_{\text{seq}} = 3$ frames. A per-cell GRU consumes the encoded sequence; its hidden state is re-initialized for each window, so it is a short-range aggregator rather than an unbounded memory. With so short a window, frame concatenation or a small attention block would also serve; we use the GRU because it adds few parameters, imposes temporal order without positional design choices, and is zero-initialized to start from an identity map of the latest frame. Window length and aggregation are experimental factors, not established neutral choices. The K1/K5 and concatenation comparisons changed normalization as well as the intended factor; a matched per-case re-run restores only the single-step window dose–response (Section 5.4), and the claims that five frames improve the main configuration or that GRU choice is neutral remain suspended (Appendix D.10).

Two spatial paths. The resulting latent field is processed by two parallel paths. The global path applies learned mixing in the top-$K = 32$ eigenvectors of the graph Laplacian,

$$z_{\text{glob}} = U_K\big(R\,(U_K^\top z)\big),$$

where R is a learnable channel-mixing matrix at each retained mode; the main configuration uses four such FNO-style blocks, one shared by all frames before the GRU and three on this path. Because U_K spans the whole domain, one block already couples every cell to every other. The local path stacks two convolutional blocks, each with two 3×3 convolutions, for a nominal receptive field of $\sim 9 \times 9$ cells. The two paths differ precisely in how far information travels per application—global at depth one versus a few cells—which is the design axis under study; how this interacts with front displacement per step is the subject of Section 5.3.

Gated fusion. The two paths are fused by a per-cell gate. For cell i,

$$g_i = \sigma\big(w^\top [z_{i,\text{loc}} \| z_{i,\text{glob}}] + b\big), \qquad z_i = g_i z_{i,\text{loc}} + (1 - g_i) z_{i,\text{glob}},$$

a single linear map from the concatenated $2D$-dimensional features to a scalar, shared across cells and time. The gate is deliberately minimal ($2D{+}1$ parameters) so that its output varies only through the latent features. A small parameter count does not preclude encoding fire identity. Section 5.4 describes historical gate probes; their results do not establish a general impossibility of regime routing. The decoder maps the fused representation to full-grid ignition logits. The ablations and stratified expert are variants of this family, not necessarily matched experimental controls; none is a new primitive.

4.2 TEMPORAL AND LOCAL BASELINES

The external ConvLSTM baseline mirrors the published WSTS baseline family: the same three-frame window, a single ConvLSTM cell (all gates from one 3×3 convolution over input and hidden state), a 1×1 head, and per-window state reset. Its spatial reach is 3×3 per frame, accumulated over the three-frame recurrence—narrower than GLOW's local path, with recurrent state substituting for depth. To separate "ConvLSTM the architecture" from "ConvLSTM the implementation," CL-branch keeps GLOW's encoder and decoder but deletes the global path and replaces the local path

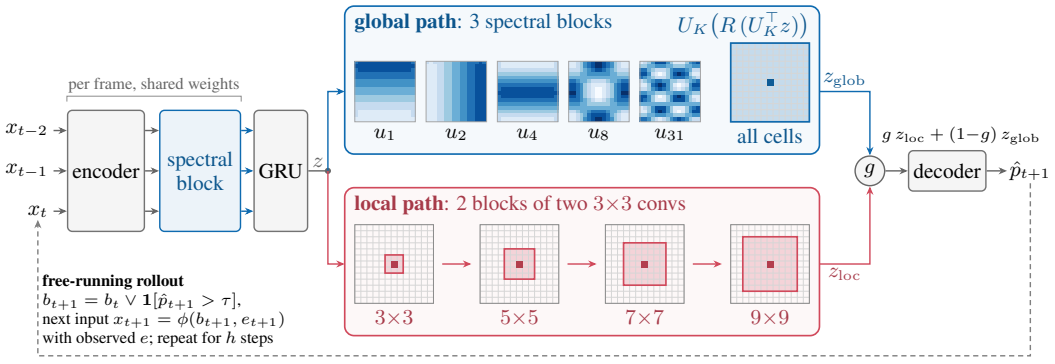


Figure 2: **GLOW and its two paths.** One step (solid) and the free-running rollout (dashed). Blue marks the spectral components—a block applied to each frame before the GRU, then the global path—and red the local path; the §5.3 contrasts delete one or the other. Tiles: five of the columns $u_0, \ldots, u_{31}$ of U_K (Laplacian eigenvectors) for a real training fire, sign arbitrary. Grids: receptive field of the dark cell on a 13×13 patch after each 3×3 convolution, versus after one spectral block.

with the same ConvLSTM cell (Section 5.3). We omit a transformer: three frames give temporal attention little to attend over, and an untuned spatial transformer adds capacity choices we cannot control at this sample size.

4.3 Training objective

The data term is focal binary cross-entropy (Lin et al., 2017) on the candidate set,

$$\mathcal{L}_{\text{data}} = \text{FocalBCE}(\hat{p}_{t+1}|_{C_t}, y_{t+1}|_{C_t}).$$

The schedule starts with a data warm-up and optionally adds a Rothermel residual (Rothermel, 1972), a short rollout term, and a CRPS ensemble term (Gneiting & Raftery, 2007). They are experimental factors, not assumptions that the coarse mask observes a physical rate. The configuration reported here trains with the single-step objective alone; the remaining terms are tested and reported where they do not help (Appendix C).

Figure 2 summarizes the model family and the free-running state transition. The question is not whether GLOW is uniformly best, but how the global and local paths exchange returns when the discretized state change and the training objective change.

5 Experiments

Our experiments answer three questions: (i) do global and local biases change rank across fires with different per-step state changes (§5.3–§5.3); (ii) does re-indexing the same fires onto twice the step length—same detections, longer steps—move the ranking in the direction the growth-account predicts (§5.4); and (iii) can a learned gate or an external selector exploit the regularity without test labels (§5.4, §5.5)? The organizing variable throughout is mean newly burned cells per model step.

5.1 Experimental setup

Datasets and splits. *Cross-state dataset (primary).* We construct 45 fire events from VIIRS active-fire detections (Schroeder et al., 2014) over the western United States, rasterized to a 1 km grid on a 6-hour lattice with DEM-derived terrain, HRRR wind (Dowell et al., 2022), and burn-history channels. The lattice resamples ∼11 h VIIRS observations; a bin without a new detection repeats the previous state (Appendix A). The main configuration deliberately uses only terrain, wind, and burn state; fuel maps and surface meteorology are tested as explicit factors (Appendix D). These choices do not establish that architectural comparisons are free of preprocessing confounds. The strict spatial-transfer split trains on 15 California fires, reserves 3 further California fires for validation, and tests on 27 unseen fires in Oregon (10), Washington (10), and Arizona (7): roughly 791K candidate cells at a 1.4% positive rate.

Table 1: **Multi-step rollout on cross-state transfer.** ROC-AUC at $h \in \{1, 3, 5\}$ (3-seed mean). Bold = column maximum. Slope: ROC-AUC(h=5)−ROC-AUC(h=1) averaged over states. Evidence grades: † verified (matched retraining, frozen normalization), * faithful (re-evaluated or bit-exact), unmarked = historical. Rollout h=1 uses the 8-start free-running protocol and differs from the teacher-forced anchor (Appendix C). Columns are ordered by measured per-step spread rate (mean frontier displacement in cells/step over each state's test fires: WA 0.91, OR 2.14, AZ 3.38), not by nominal state name.

	WA (slow)			OR (mid)			AZ (fast)			
Model	h=1	h=3	h=5	h=1	h=3	h=5	h=1	h=3	h=5	slope
GLOW†	0.726	0.702	0.709	**0.854**	0.722	**0.762**	0.764	0.722	0.714	−0.053
ConvLSTM*	**0.761**	**0.750**	**0.715**	0.801	0.621	0.700	**0.792**	**0.800**	**0.851**	−0.030
LR	0.686	0.737	0.695	0.800	0.611	0.677	0.752	0.685	0.664	−0.067
base-FNO*	0.689	0.671	0.695	0.801	**0.745**	0.746	0.716	0.686	0.677	−0.030

WSTS (external validity). WildfireSpreadTS (Gerard et al., 2023) contains 607 fires at 375 m resolution and daily steps. We use the official leave-one-year-out folds (2018–2021) as an independently collected external corpus; after deterministic viability filters, 419 fires enter the per-fire analyses.

Compared models and protocols. We compare four roles: LR is a pointwise reference; ConvLSTM the local recurrent baseline; base-FNO the global spectral ancestor; and GLOW the hybrid. CL-branch is the same-backbone local-recurrent contrast used for attribution. All temporal models use the most recent three frames. Geo rollout uses horizons $h \in \{1, 3, 5\}$ at 6-hour steps; WSTS rollout $h \in \{1, 2, 3\}$ at daily steps. Free-running evaluation uses eight start frames per case and feeds thresholded burn predictions back into subsequent windows; weather and static fields are supplied at the corresponding observed time, so the closed loop is with respect to burn-state channels only.

Metrics and statistical protocol. Tables report pooled-micro ROC-AUC by state (all candidate cells concatenated); fire-level ranking, oracle and routing analyses use positives-weighted per-fire ROC-AUC over fires with ≥10 positives at the deepest horizon (formula and bootstrap protocol in Appendix J). These scopes answer different questions and are never compared numerically. Seed-level contrasts are descriptive standardized differences. Geo checkpoints use validation ROC-AUC selection, WSTS validation PR-AUC. Test sets were reused across exploratory experiments; offline weather is supplied beyond the rollout start on both datasets, and router features include retrospective fire information—so neither routing nor rollout is an operational forecast test.

5.2 Main Results

Table 1 reports pooled-micro free-running ROC-AUC on the 27 cross-state test fires (3-seed means; 6-hour steps). Evidence grades (marks in the caption; Appendix J): GLOW = matched retraining under a frozen normalization contract (*verified*); ConvLSTM = reproduced bit-exactly (*faithful*); base-FNO = re-evaluated (*faithful*); LR = deterministic, protocol-independent. The single-step teacher-forced anchor is in Appendix C.

Single-step protocols understate deployment differences. The slope column encodes the re-ranking that motivated this study: LR is competitive at h=1 yet degrades fastest (−0.067, the steepest of any family; Figure 6, Appendix C).

Reading the table. In the pooled columns ConvLSTM takes every WA and AZ maximum; GLOW leads only on OR, and ConvLSTM alone *improves* on AZ with horizon. That pooling decides the slow column: only 4–15% of positive–negative pairs are *within* a fire, so the metric is mostly cross-fire comparability; per Table 24 (Appendix J), for GLOW—though not for ConvLSTM—WA's within-fire ROC-AUC is the highest of the three states, its cross-fire term the lowest. On the within-fire term the sign tracks spread rate—GLOW − ConvLSTM is +0.057 (WA, slowest), +0.061 (OR), −0.121 (AZ, fastest)—though WA and OR are statistically indistinguishable, so only the fast-end sign change is resolved; this is a per-fire regularity, not a per-state accident (§5.3). Against LR, GLOW is favorable in 17 of 18 state×horizon ROC-AUC/PR-AUC cells; deleting the local branch

and gate (base-FNO) loses ground at the fast states, so the fusion rather than the spectral basis carries the transfer (Appendix D.10).

Stratified pooling on the expanded test set. On the expanded 53-fire pool (Table 3; historical descriptive grade, caveats in the caption) the fast stratum carries the largest gap while the slow and mid strata sit within noise of parity. Appendix B carries this table and the *main*-pool per-fire version (Figure 5); the two pools are not interchangeable.

A best-effort stack does not escape the condition. Stacking a five-frame window with LANDFIRE fuel yields the strongest short-horizon configuration of this study (historical, normalization-unmatched grade). The fast-end condition is untouched: AZ $h{=}5$ collapses by -10.3σ, the fast stratum is unchanged (Table 3), and the slope steepens. We keep the three-frame configuration as the reference (Appendix D.7).

The training objective is a controllable free variable. The reference configuration is trained single-step yet graded by free-running rollout, so the mismatch is itself a candidate explanation for the ranking (§4.3). Retraining the identical architecture with a differentiable multi-step loss under the same frozen-normalization contract (3 seeds) moves the ordering without lifting it: the fast state's deep-horizon deficit roughly halves (AZ $h{=}5$ $-0.137 \rightarrow -0.070$) at a real short-range price. We do not adopt the objective: same architecture, data and seeds, different ordering, but the $h{=}5$ shift is not resolvable per fire ($+0.017$, 95% CI $[-0.013, +0.048]$; Appendix D.12). The latent-noise ensemble, meanwhile, is an *error-ranking* signal at no observed point-accuracy cost (historical, under protocol audit) rather than probability calibration (Appendix H).

5.3 A Spread-Rate Axis for Global–Local Ranking

This subsection organizes the evidence chain on a single axis and reports it with the protocol grades of Table 1: the intervention arm is *verified* (matched retraining with frozen normalization on both scales), the family contrasts mix verified and faithful arms, and correlations remain historical where marked.

A unified spread-rate axis. For each evaluation unit we compute mean newly burned cells per *model step* during the active phase (a 6 h lattice resampling $\sim$11 h observations on the cross-state corpus, daily on WSTS; Appendix A). This proxy is not physical front speed: it absorbs grid spacing, step length, geometry and detection gaps, and a perimeter-normalized kilometer proxy does not align the datasets (Pearson $+0.12$/$+0.02$), so we treat cells/step as a protocol-organizing variable. Figure 3(a) plots, against this axis, GLOW minus ConvLSTM free-running ROC-AUC at the deepest horizon for the units of both datasets (15 geo test fires, 1 km/6 h, $h{=}5$, fewer than 10 positives at $h{=}5$ excluded as noise-dominated; four WSTS year-folds at 375 m/daily, $h{=}3$; 419 WSTS fires as background points).

Obs. 1 (within-dataset correlation; historical, faithful). Slow geo fires (< 2 cells/step) are predominantly won by GLOW (5 of 7), fast geo fires (4–6 cells/step) uniformly by ConvLSTM; the correlation between log spread rate and ConvLSTM's advantage is $+0.58$ (case bootstrap 95% CI $[+0.16, +0.82]$, permutation $p = 0.023$). On the expanded 53-fire pool this slow-stratum advantage is filter-sensitive (Table 3), so we assert slow-regime advantage only for the main pool, only on the within-fire term (Table 24), not as a universal slow-fire claim—and the main pool's slow stratum is Washington-majority (4 of 7), not Oregon.

Obs. 2 (external dataset; historical, faithful). WSTS folds lie at the fast end (44–142 cells/step) and are uniformly won by ConvLSTM ($+0.08$ to $+0.10$); the combined correlation is $+0.49$, and ConvLSTM beats GLOW on 77.3% of 419 evaluable fires (mean ΔROC-AUC $+0.071$). Because all WSTS fires lie inside the fast regime, the weak within-WSTS slope ($+0.18$) is not used as evidence.

State is not the axis. State shift and spread rate are confounded by construction: ordered by measured per-step displacement the states are WA (0.91 cells/step, all ten fires below 2), OR (2.14), and AZ (3.38), and every fire above 4 cells/step in the pool is in Arizona while every fire in the 2–4 band is in Oregon. Note that this ordering does not follow the nominal state names—Washington, not Oregon, is the slow state. Three checks break the tie. First, within-state exceptions: the slow Arizona

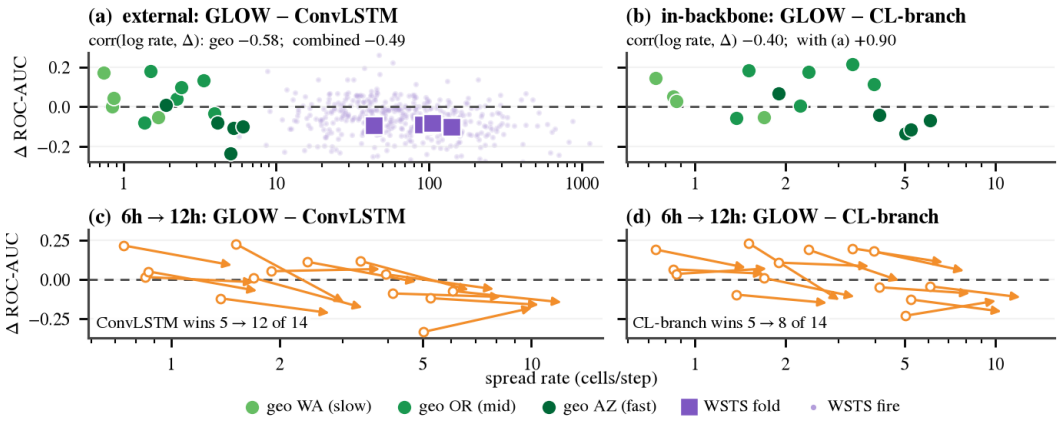


Figure 3: **Growth-load associations and a verified intervention.** Up is GLOW. (a) GLOW minus ConvLSTM free-running ROC-AUC at the deepest horizon vs. mean newly-burned cells per model step: 15 geo fires (6 h, h=5), four WSTS year-folds (daily, h=3), 419 WSTS fires. (b) The same for the in-backbone contrast, evaluated on the 15 geo fires only and therefore covering the slow end (per-fire agreement $\rho = +0.90$). (c, d) The verified 6h→12h intervention of Obs. 4 per fire: circle at 6 h, arrowhead at 12 h (per-step travel exactly doubles).

fire AZ_2017_27 is won by GLOW and the fast Oregon fire OR_2017_39 by ConvLSTM. Second, year-split WSTS holds the pattern with no state shift at all. Third, mixed-state training preserves it (corr. $+0.67$ versus $+0.58$). Decomposing the pooled metric (Table 24) makes the same point quantitatively: on the within-fire term—where the metric is not contaminated by cross-fire score comparability—the slowest state is also the one where GLOW's margin over ConvLSTM is intact. None replaces the intervention below, but together they make state an unlikely sole explanation.

A same-backbone family contrast. Panel (b) of Figure 3 therefore repeats the comparison inside our own backbone, as a *family* rather than independent-model contrast: CL-branch deletes the spectral branch and swaps the local branch for a ConvLSTM cell. **Obs. 3 (family association; faithful historical arm):** the per-fire correlation between the external and in-backbone difference maps is $+0.90$, and the two arms swap roles by regime (CL-branch 0.815 vs. GLOW 0.714 at AZ h=5; 0.551 vs. 0.722 at OR h=3). The complementary control drops the local branch instead: under matched-normalization retraining (verified) the spectral-only arm is *below* GLOW at every state single-step and level at h=5 except on fast AZ—neither branch is dispensable, and the historical "spectral-only is best on OR" mirror was a normalization artifact (Figure 9, Appendix D.10). Encoder, decoder and training protocol are shared, but more than one component changes, so this is a consistent family-level pattern rather than a unique causal identification of the spectral branch.

5.4 Manipulating and Exploiting the Axis

A designed intervention on temporal resolution. Temporal subsampling of the *same* fires from 6-hour to 12-hour steps changes cells per model step exactly twofold (measured 1.09 → 2.18 cells/step on the probe pair) while leaving the fire, the grid, and the physics untouched. Because the finer grid is forward-filled (Appendix A), subsampling removes repeated states, not observations: the growth events are unchanged while the share of steps carrying none falls from 82.6% to 64.7%. Because the historical version of this comparison bundled the step change with a normalization mismatch, we retrained both scales under one frozen-normalization contract and re-ran the preregistered readout (per-case seed-mean h=5 ROC-AUC, paired bootstrap over the 14 fires with ≥5 positives on both scales). **Obs. 4 (verified):** the ConvLSTM−GLOW displacement is $+0.094$ (95% CI $[+0.034, +0.155]$, $p = 0.0022$), with ConvLSTM winning 5 of 14 fires at 6 h and 12 of 14 at 12 h; the CL-branch contrast moves the same way ($+0.083$, $p = 0.0010$), and on the seven slowest fires the sign flips from -0.064 to $+0.068$: the same fires, under a protocol-controlled change of temporal resolution, change the winning family. Fixed h still corresponds to different physical durations across scales, so the claim is about *discretization-conditioned* rankings, not a universal time

horizon. The direction does not extrapolate monotonically: a preregistered replication on WSTS (both families retrained at two-day steps, per-step travel exactly doubled) fails to widen ConvLSTM's lead—the gap at the deepest horizon shrinks in three of four folds, and the per-fire shift is graded by speed opposite to prediction ($\rho = -0.12$, $p = 0.02$; fastest quartile toward GLOW)—the coarsening response is not a monotone dose (Appendix I, probe h). The axis is manipulable from the data side, not at inference alone: sub-stepping is null under the hard update, and training-side padding is null or opposite while costing short-range skill (Appendix D.13).

Mechanism: what the global path is doing. At $\lesssim$2 cells per step many frontier cells look alike locally and the discriminative signal is where growth concentrates along the perimeter, favoring low-frequency context; at 4–6 cells frontal displacement matches one local-path application ($\sim 9 \times 9$), so sharp local advection carries the signal and spectral smoothing blurs it. Two verified probes sharpen this into a claim about *how* the global path participates. A FiLM-style variant, in which the spectral branch only modulates the local branch (channel-wise gain and shift, zero-initialized) instead of predicting, *halves* the fast-state single-step deficit to ConvLSTM and matches it at $h{=}1$ yet loses ROC-AUC on OR $h{=}5$, where the parallel gate that lets global features *predict* remains best (numbers in Appendix D.10). The window dose–response is monotone single-step ($K{=}1/3/5$: $0.766/0.810/0.878$ on OR, matched normalization) but regime-dependent at depth. The crossover near 2–4 cells/step is consistent with this scale account, which remains a hypothesis rather than an independent causal test. Together: **the global path should act as a parallel predictor for slow long-range spread, as a conditioner for fast short-range steps, and neither role substitutes for the within-window local recurrence that carries fast regimes**—the mechanism-level form of the ranking result.

Gate routing does not close the gap internally. Because the taxonomy suggests per-regime routing, we report the negative side. Seven internal interventions all failed to improve the fast regime (Appendix E): speed features on the gate, supervision toward window speed, class-balanced supervision, an MLP gate, a ConvLSTM local branch, unconstrained fusion coefficients, and FiLM modulation with the parallel gate. The gate can be steered toward speed (ρ from -0.26 to $+0.41$; historical, under protocol audit) but only at an accuracy cost. External routing fares no better: a prior-feature router recovers $+0.018$ over always-GLOW at best—matching, not beating, always-ConvLSTM—and six learned routers never beat the best single expert (Appendix F); early-error feedback is *anti-correlated* with the long-range winner ($\rho = -0.18$); and resolution augmentation restores the OR advantage while leaving the fast-state gap untouched (AZ $h{=}5$ -0.115). Regime routing we leave to future work.

5.5 Exploiting the Condition: Stratified Experts

The intervention makes the regime rule manipulable, but a *deployable* selector is separate. A stratified expert—CL-branch trained single-step with pure focal loss on 12 fast California fires—reaches per-fire $h{=}5$ ROC-AUC 0.746 against ConvLSTM 0.742 and GLOW 0.724; the $+0.005$ over ConvLSTM is not separable from zero (95% CI $[-0.061, +0.063]$). A three-way oracle rises to 0.803 ($+0.018$ over the two-way 0.785, CI $[+0.003, +0.040]$, $p{=}0.006$), so the experts do carry complementary signal. Recovering it is the open problem: every learned router we tried is retrospective, using information unavailable at issue time, so its gains match rather than beat always-ConvLSTM. On WSTS the expert leads ConvLSTM on 4/4 year folds (Appendix F, G).

6 Discussion and Limitations

The claim is bound by the scope conditions of §5.3: the organizing variable is a discretization-dependent growth load, not a front-speed law, rankings are conditioned on the sampling interval, and the loop is closed only in fire state. (i) The slow-regime advantage does not transfer to the expanded pool (Table 3), so we claim it only for the main pool, only on the within-fire term (Table 24); external validity is conditional on validation-AP selection and observed weather. (ii) Absolute PR-AUC is low and rollouts use observed wind, so ranking implies neither calibrated probabilities nor deployment performance. (iii) The one-ring protocol concentrates threshold events outside the ring in slow fires (5–8×; Appendix D.11). (iv) A 2× downsample degrades both families comparably, so the rule concerns the sampling interval only.

REFERENCES

John Burge, Matthew R. Bonanni, R. Lily Hu, and Matthias Ihme. Recurrent convolutional deep neural networks for modeling time-resolved wildfire spread behavior. *Fire Technology*, 59(6): 3327–3354, 2023. doi: 10.1007/s10694-023-01469-6.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, 2016.

David C. Dowell, Curtis R. Alexander, Eric P. James, Stephen S. Weygandt, Stanley G. Benjamin, Geoffrey S. Manikin, Benjamin T. Blake, John M. Brown, Joseph B. Olson, Ming Hu, Tatiana G. Smirnova, Terra Ladwig, Jaymes S. Kenyon, Ravan Ahmadov, David D. Turner, Jeffrey D. Duda, and Trevor I. Alcott. The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022. doi: 10.1175/WAF-D-21-0151.1.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, pp. 74515–74529, 2023.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

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.

Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505, 2020. doi: 10.1139/er-2020-0019.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operators: Learning maps between function spaces. *Journal of Machine Learning Research*, 24(89):1–113, 2023.

Saad Lahrichi, Jake Bova, Jesse Johnson, and Jordan Malof. Improved wildfire spread prediction with time-series data and the WSTS+ benchmark. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Richard C. Rothermel. A mathematical model for predicting fire spread in wildland fuels. Research Paper INT-115, U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT, 1972.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

Joe H. Scott and Robert E. Burgan. Standard fire behavior fuel models: A comprehensive set for use with Rothermel's surface fire spread model. General Technical Report RMRS-GTR-153, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, 2005. 72 p.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, 2015.

Chang Xu, Gencer Sumbul, Li Mi, Robin Zbinden, and Devis Tuia. Geofar: Geography-informed frequency-aware super-resolution for climate data. In *International Conference on Learning Representations (ICLR)*, 2026.

Alexey Yermakov, Yue Zhao, Marine Denolle, Yiyu Ni, Philippe M. Wyder, Judah Goldfeder, Stefano Riva, Jan Williams, David Zoro, Joseph Bakarji, Amy Sara Rude, Georg Maierhofer, Matteo Tomasetto, Joe Germany, Miles Cranmer, and J. Nathan Kutz. The seismic wavefield common task framework. In *International Conference on Learning Representations (ICLR)*, 2026.

Reproducibility Statement

Every table row is graded *verified*, *faithful*, or *historical*, and the grade is stated with the table (Appendix J); the audits and readouts behind them are kept as versioned per-case result files, and the sections whose reproducibility turns on a specific artifact name it in place. The protocol-recovery audit re-evaluated every existing checkpoint under a shared normalization/manifest contract with case-keyed RNG isolation, and retrained the main comparisons with frozen normalization contracts stored inside each checkpoint. Multi-seed experiments use seeds $\{42, 123, 2024\}$ unless a single-seed or missing-seed exception is stated. Historical rows remain traceable but their training-side validity has not been certified by that audit, and they are marked as provisional wherever they are used. The supplementary material contains the code, all evaluation splits, and the result files from which Table 1 and every data figure can be regenerated; release plans follow venue rules and source licenses (Appendix K).

LLM Usage Disclosure

Large language models were used as research-assistant tooling during this project: for experiment orchestration (job scheduling and bookkeeping), drafting and copy-editing of prose, writing and refactoring analysis and plotting code, and as a sounding board during experiment design, literature triage and simulated-review response drafting. All experimental designs, statistical protocols, scientific claims, and the final interpretation of results were determined and verified by the authors; every number in the paper is traceable to versioned result files as described in the Reproducibility Statement.

How to read the evidence in this appendix

Rows marked *verified* or *faithful* were re-evaluated or retrained under one shared normalization and manifest contract; the main table itself, the temporal-coarsening intervention, the window dose–response, the branch controls (including a normalization-matched no-local retrain) and the fusion-mechanism probes are in that set. The remaining comparison arms—local-branch depth, temporal aggregator, gate variants, widened ring, the maximal short-horizon stack, and the training -size and meteorology controls—are retained as *historical* records whose normalization could not be matched; they are labeled as such wherever the main text cites them, and the evidence grades are defined in Appendix J. Confidence intervals recomputed under the repaired protocols are marked; unmarked historical intervals are kept for traceability only.

A Dataset Construction and Statistics

We construct 45 fire-event cases from the VIIRS 375 m active-fire archive. Detections are clustered into individual fire events, and each event is rasterized onto a regular $\sim$1 km grid at 6-hour time steps, yielding a binary burning-state tensor of shape (T, H, W) per case. Each case is augmented with static terrain from a digital elevation model (elevation, slope, aspect) and dynamic surface wind fields (u, v) interpolated from the HRRR reanalysis at every time step. Node features are 13-dimensional (Table 2); the fuel channel is a constant placeholder in the main configuration — real LANDFIRE fuel maps are evaluated separately in the ablation of Appendix D and do not enter the main model.

The cross-state split uses 18 California cases for training (15 train + 3 held-out validation cases for early stopping) and 27 out-of-distribution test cases pooled per state: 10 Oregon cases (352,359 nodes, 4,760 positives, base rate 1.35%), 10 Washington cases (301,153 nodes, 3,782 positives, 1.26%), and 7 Arizona cases (137,887 nodes, 2,908 positives, 2.11%). The pooled test set totals $\sim$791K node-level candidates with $\sim$11K positives (base rate $\approx$1.4%). No test-state data is seen during training or model selection.

The 6-hour grid is a resampling convention. Detections are assigned by floor-binning their acquisition time into 6-hour bins from the case start, after which the burning state is made cumulative by a per-step maximum; a bin carrying no new detection therefore repeats the previous state. Both facts govern how the axis of §5.3 should be read. The archive is single-platform (NOAA-20), and VIIRS observes each scene about twice daily, so the independent overpass interval has median 11.4 h (IQR 11.3–12.3; 10th–90th 9.7–14.0) and the effective observation interval is $\sim$11 h rather than 6. Consistent with this, across every built case each growth event falls in a bin whose index is 1 or 3 modulo four and bins 0 and 2 never carry one: of 411 steps in `AZ_2016_27`, 380 repeat their predecessor. Two consequences: a three-frame window spans about one day and holds at most two distinct states, and per-model-step growth is growth per 6 h-lattice bin, not per six physical hours. The step-level bursts behind the coverage bound above are not an artifact of observation latency—across $1{,}424$ growth events the correlation between the log preceding observation gap and log growth is -0.271, so longer gaps precede *smaller* growth. The 6-hour and 12-hour grids are thus two points on one axis rather than a measurement and its degradation, which is how §5.4 treats them.

Candidate-ring coverage. Because the candidate set is the first unburned ring around the current front (§3.1), newly burned cells further than one cell away are never scored. Over all 45 cases, 57.6% of newly burned cells fall inside the ring; the per-step relationship is strongly monotone in instantaneous growth (coverage 0.90 for steps with fewer than 10 new cells, 0.62 for 10–50, 0.33 for 50–200, 0.18 above 200), while the case-level correlation with mean spread rate is near zero (-0.02)—coverage is governed by step-level bursts, not by a fire's average speed. This bound applies identically to every compared model and is reported as a formulation limitation in Section 6.

B Regime Stratification: Per-Fire Strata

The strata in Figure 5 use the main 15-fire pool at the $\geq$10-positive scope. Figure 3(a) covers the same contrast on the broader geo+WSTS speed axis: its geo points are historical readings, whereas

Table 2: Node feature composition (13 dimensions).

Feature	Source	Dims
Elevation, slope, aspect	DEM (static)	3
Fuel placeholder (constant)	—	1
Burning state	VIIRS	1
Burning-neighbor fraction ($n_{\text{burn}}/8$)	VIIRS	1
Wind u, v	HRRR	2
Reserved meteorology channels (zero)	—	2
Normalized row/column coordinates	grid	2
Normalized time index	grid	1

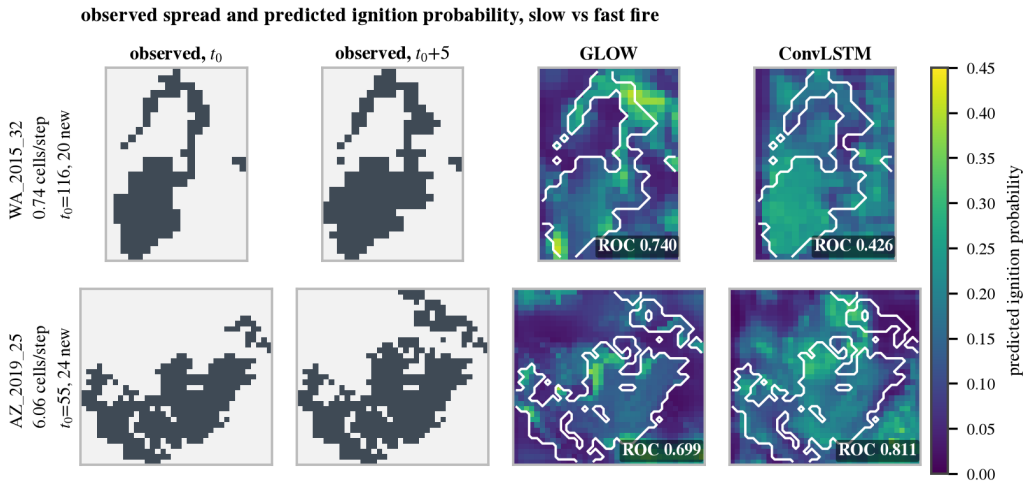


Figure 4: **What the two regimes look like.** Two fires, chosen by a rule fixed before looking at any model output: the slowest and the fastest of the 15 evaluable main-pool fires by measured spread rate (0.74 and 6.06 cells per step). Columns: observed burn at the rollout origin t_0; observed burn five steps later; and the predicted ignition probability at $h{=}5$ for GLOW and ConvLSTM. The white contour on the two model panels is the *observed* burn at $t_0{+}5$, so prediction and outcome can be compared directly; the boxed value is that single start's ROC-AUC on the observed candidate set. Within each fire the start shown is the one with the most observed new ignitions at $h{=}5$ — an information criterion, not a best-case one. The probability scale is shared and covers $[0, 0.45]$, the range the four maps actually span, rather than the full unit interval. These are single-start illustrations: the quantities compared in the tables are pooled over eight starts, and the two need not agree in sign for an individual start. Reproducibility: the export path re-derives the pooled $h{\in}\{1, 3, 5\}$ ROC-AUC of both arms on both fires and matches the stored per-case results to six decimals.

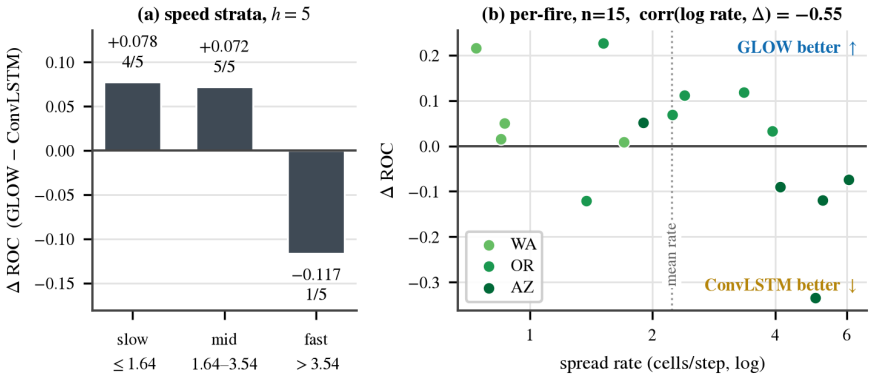


Figure 5: **Per-fire speed strata (main 15-fire pool).** (a) Tertile strata ($\leq$ 1.64 / 1.64–3.54 / $>$3.54 cells per step) of the h=5 per-fire comparison: slow $+0.078$ (4/5 fires), mid $+0.072$ (5/5), fast -0.117 (1/5). (b) Per-fire ΔROC against spread rate, colored by state. Per-fire scope over the 15 fires with $\geq$10 positive cells, not comparable to the pooled-micro table, and a different pool from Table 3. The fixed-threshold variant ($<$ 2 / 2–4 / $>$4 cells per step) gives $+0.064$ (6/7) / $+0.083$ (4/4) / -0.154 (0/4) *on the repaired-protocol scale*; on the historical axis that §5.3 reports, the same slow stratum gives $5/7$, the two differing in one Washington fire. Its mid stratum is Oregon-only, which is why the state-balanced tertiles are shown.

Table 3: **Regime-stratified** positives-weighted per-fire ROC-AUC at h=5 on the expanded 53-fire pool (same checkpoints as Table 1; $\geq$10-positive fires). Historical descriptive grade: the pool is not speed-balanced and its slow stratum is split between WA and OR fires, so it is not a clean slow-fire test of the main pool's claim; the slow/mid leads are descriptive. *Case-bootstrap p=0.007.

Stratum (cells/step)	fires	n_{pos}	GLOW	ConvLSTM	Δ
Slow ($<$ 2)	9	165	**0.674**	0.659	$+0.015$
Mid (2–4)	12	314	**0.749**	0.727	$+0.022$
Fast ($>$ 4)	6	266	0.697	**0.793**	-0.095^*

the per-fire ΔROC here is re-evaluated under the repaired protocol, so the two are consistent in direction but not numerically identical.

C Training and Evaluation Details

Hyperparameters. All GLOW variants use latent dimension D=64, K=32 spectral basis vectors, 4 operator blocks, AdamW with learning rate 3×10^{-4} and weight decay 10^{-5}, and are trained for up to 60 epochs. Stochastic-ensemble settings are σ=0.1, K_{ens}=4 during training and 10 at evaluation. Every configuration is trained with three seeds (42, 123, 2024); all tables report mean $\pm$ std over seeds.

Seed independence and state redundancy. The training pool is resolved from the split name alone: changing the `ood-state` selector swaps only the auxiliary out-of-distribution probe list used for monitoring, never the fires that are trained on. A controlled re-run test confirms this at the weight level: re-training one configuration to completion changes the parameters by 8.6×10^{-2} (maximum absolute difference) and lands on a different best epoch (5 vs. 4; validation ROC-AUC 0.8513 vs. 0.8582), essentially the same separation as running the same configuration with a different state selector (8.9×10^{-2}). Each state's three-seed mean and seed standard deviation therefore still come from three independently trained models, but a family's nominal nine state$\times$seed entries contain only three independent runs—independence holds along the seed axis, not across states. The same re-run pair places the same-seed reproduction noise of this pipeline at 0.007 validation ROC-AUC.

Curriculum. The training code defines a four-stage curriculum: (i) data warmup with focal BCE only; (ii) from epoch 12, the Rothermel-based PDE residual is ramped in ($\lambda_{\text{pde}} \to 0.1$); (iii) from epoch 24, a multi-step rollout term is scheduled; (iv) from epoch 42, the CRPS ensemble term is ramped in ($\lambda_{\text{ens}} \to 0.5$). Two qualifications matter for interpreting the reported models. First, the scheduled rollout term is consumed only by a separate end-to-end multi-step mode, not by the standard training path: every checkpoint reported in this paper is trained with the single-step objective, so stage (iii) is inactive. Second, validation-based checkpoint selection lands in the warmup stage (best epochs 2–9), so the selected checkpoints do not reach stages (ii) or (iv) either—which independently corroborates the noPDE ablation (Appendix D). We also ran the end-to-end multi-step variant (differentiable free-running loss with soft state feedback and per-horizon focal terms) as an exploratory arm on the base spectral configuration: it raises deep-horizon scores (AZ h=5 $+0.058$) at a short-horizon cost (OR h=1 -0.025, h=3 -0.042), and is not adopted for the main configuration. This single-step objective mismatch is the direct motivation for the rollout-aware training we leave to future work.

Validation-based early stopping. Three California validation cases are held out (`CA_2017_107`, `CA_2020_122`, `CA_2017_105`), selected by stratified sampling over positive rate (0.10, 0.25, 0.44%; seed 42), covering low/mid/high base rates. Checkpoints are selected by validation ROC-AUC with patience 10 and a minimum of 5 epochs. We use ROC-AUC rather than PR-AUC for selection because validation PR-AUC at $\sim$1% base rate is high-variance ($\sim$0.03) and frequently selects the epoch-0 noise peak, whereas ROC-AUC is base-rate invariant and stable along the trajectory ($\Delta < 0.01$).

Historical mixed-state training comparison. The historical mixed-state comparison retrained GLOW and ConvLSTM (3 seeds each) on a 40-fire pool comprising the 18 California fires plus 22 further fires from Oregon, Washington and Arizona that are disjoint from the 27 test fires. Single-step transfer gaps close for both families (pooled ROC-AUC GLOW/ConvLSTM: OR 0.849/0.826, WA 0.815/0.807, AZ 0.837/0.837), so the rival "distribution shift" account predicts that the state-wise ranking should flatten. It does not: the per-fire correlation between log spread rate and ΔROC-AUC(ConvLSTM$-$GLOW) at h=5 is $+0.67$ under mixed training against $+0.58$ under California-only training (n=15), all four fast fires (> 4 cells/step) remain ConvLSTM wins, and the AZ h=5 gap widens rather than closes ($+0.18$, $+8.1\sigma_{\text{seed}}$). Mixed training also shortens GLOW's long-horizon advantage on Oregon (OR h=5 ConvLSTM$-$GLOW moves from -0.06 to $+0.02$), consistent with the earlier observation that a California-only validation set selects earlier, less rollout-stable checkpoints under mixed training. These provisional patterns do not establish that distribution shift has been removed as an explanation; checkpoint selection and data composition remain coupled.

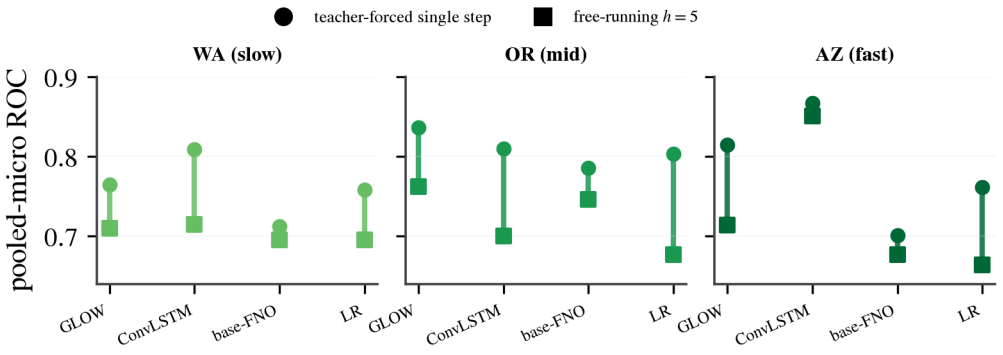


Figure 6: **Teacher-forced single-step scores understate deployment differences.** Circles: teacher-forced single-step pooled-micro ROC-AUC; squares: free-running h=5. The pointwise LR baseline is competitive single-step yet degrades most under free-running rollout (the steepest drop of any family), while the spatiotemporal families degrade more slowly and re-rank—benchmark protocols that stop at single-step evaluation cannot see this. Same checkpoints and evidence grades as Table 1.

Table 4: Historical single-step cross-state transfer (under protocol audit), pooled-micro ROC-AUC/PR-AUC (3-seed mean). This is the single-step anchor for the rollout results in the main text.

State	GLOW	ConvLSTM	GCN	LR	Persistence
OR	**0.831**/0.069	0.810/**0.087**	0.749/0.045	0.803/0.048	0.500/0.014
WA	0.758/0.035	**0.809/0.067**	0.777/0.043	0.758/0.035	0.500/0.013
AZ	0.817/0.101	**0.867/0.142**	0.765/0.080	0.761/0.059	0.500/0.021

Resolution augmentation: mixed 6h/12h training (preregistered probe; negative with structure). The only untried constructive answer to the taxonomy was distribution-level: pool the 6 h and 12 h versions of the same 18 California training fires (36-case pool; six two-scale validation cases; mixed-pool global normalization frozen into the checkpoints; GLOW and ConvLSTM at three seeds each under one matched protocol), then evaluate the same checkpoints on both test pools. The preregistered target—one GLOW within -0.02 of the mixed ConvLSTM at the coarsened pool—fails (h=5 three-state pooled gap -0.058; AZ -0.115, per-case paired bootstrap $+0.096$ $[+0.014, +0.168]$ for ConvLSTM, $n = 5$; the fast-state condition is untouched). The structure is informative: at 12 h the mixed GLOW *restores* on OR the advantage that the single-scale 12 h model loses (OR h=5: 0.777 vs. mixed-ConvLSTM 0.764, where 12 h-only GLOW scores 0.735 vs. 0.789), and the 6 h-side Oregon advantage is retained at every horizon ($+0.039/+0.038/+0.007$). Both families pay a small home-scale cost (mixed GLOW 6 h OR h=5 0.732 vs. 0.762 single-scale; mixed ConvLSTM 12 h approximately parity to slightly worse). Distribution-level adaptation thus buys cross-scale accommodation *within* a regime, not cross-regime performance—consistent with the taxonomy being a property of the fire and the discretization, not of the training distribution.

Observed-meteorology arm (MET-38; descriptive). Adding three observed surface fields (temperature, relative humidity, solar radiation) to the 38-fire training pool, compared against the matched no-meteorology T1 arm (3 seeds): rollout ROC-AUC improves in the slow and mid states (OR h=5 $+0.047$; WA h=3 $+0.042$, h=5 $+0.039$) and degrades in the fast state (AZ h=5 -0.033). The value of observed meteorology is itself regime-conditioned, consistent with the low-frequency-context role of the global path. This comparison was not preregistered and is reported post hoc (descriptive grade).

Training-set size comparisons (under protocol audit). Historical n8/n24 runs used nominal training sizes $N = 8$ and $N = 24$; the main reference used 15 training fires. The n8/n24 and expanded T1 GLOW arms have `global_norm=False`, unlike the reference. Thus the old correlations ($+0.26$, $+0.57$, $+0.59$ for n8/reference/n24) and fast-fire win counts cannot establish size robustness. Data composition also changes: six of the nine n24 additions are fast. Training/validation membership and nesting must be audited rather than inferred from nominal pool sizes. In particular, the later fast38 expert training set is not nested in the earlier expert's set. We withdraw the claims that these runs isolate scale, show that adding data is harmful, or exclude data limitations. A normalization-matched comparison with explicit split membership is pending.

Selection-bias audit. An earlier protocol selected checkpoints by test-set (OOD) PR-AUC, which constitutes test-set selection. Quantified on the 18-fold within-California leave-one-out split, the selected-checkpoint PR-AUC (0.0873 ± 0.055) exceeds the trajectory mean (0.045 ± 0.030) by $+0.0424$ — roughly 100% relative inflation — and 4 of 18 folds peak at epoch 0. Later runs use validation-based checkpoint selection (geo ROC, WSTS AP), but this alone does not establish absence of leakage or adaptive test-set selection across the exploratory experiment sequence.

Evaluation scope. State-level tables report *pooled micro* metrics: predictions and labels are concatenated across all cases within a state before computing ROC-AUC/PR-AUC. This avoids small-case noise dominating a macro average (per-case base rates vary by $30\times$) and matches the protocol used for all baselines. Multi-step rollout evaluation uses n_{starts}=8 start frames per case and K_{ens}=10.

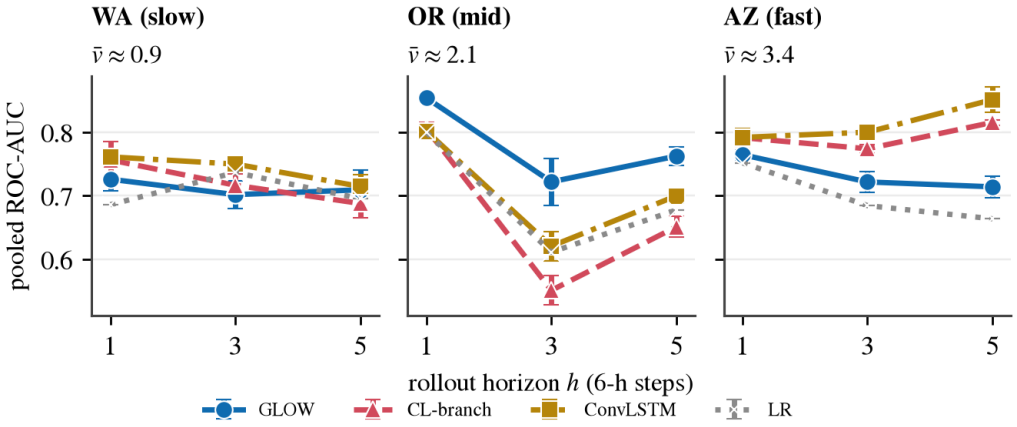


Figure 7: **Full multi-step degradation curves** behind Table 1: per-state pooled ROC-AUC at $h \in \{1, 3, 5\}$ (6 h steps; 3-seed mean $\pm$ std across seeds) for GLOW, CL-branch, ConvLSTM and LR. Columns are ordered by measured per-step growth load of §5.3, which does *not* follow the nominal state names: Washington is the slow state (0.91 cells/step), Oregon is mid (2.14), Arizona is fast (3.38). Among the four families plotted, GLOW leads at every horizon in mid-speed Oregon (base-FNO, not plotted here, takes Oregon h=3 in Table 1), the families are close in slow Washington, and the local-recurrent models dominate fast Arizona, where ConvLSTM barely degrades. Grades: GLOW verified, ConvLSTM faithful, CL-branch historical, LR deterministic.

D Full Ablation Tables

Ablations D.1–D.2 were run on the base spectral configuration; D.4–D.9 compare against the hybrid main configuration. Three-seed summaries and validation selection were intended to align the protocols, but this is not verification that preprocessing and objectives match. In particular, D.10, ring2 and gate-MLP comparisons have audited normalization mismatches. All retained tables are historical, provisional results. Δ/σ denotes the mean difference divided by the combined standard error of the two seed means (Welch convention, $\sqrt{s_1^2/3 + s_2^2/3}$ with sample standard deviations; against a deterministic baseline the second term is zero). Figure 7 plots the full per-state curves behind Table 1: the family ordering is stable across horizons within a state, and the crossover between families happens across states, tracking the per-step growth load rather than the horizon.

The variant family. Every arm in this appendix is one row of Table 5: the rows differ in which component is present (spectral path, local branch, parallel gate) and in the training objective, so they are *family* contrasts rather than independent-model contrasts. The last column is the regime each variant is strongest in (§5.3); it is a summary of the results below, not a design target — no configuration in the family escapes the regime ordering.

Table 5: The variant family: which components each arm keeps and the regime it is strongest in. Branch deletions and the gate's FiLM fusion are verified controls; the remaining rows are historical grades (Appendix J).

variant	spectral	local	gate	objective	regime
GLOW (main)	✓	conv	✓	curriculum	slow
GLOW+CL-cell	✓	ConvLSTM	✓	curriculum	mixed
CL-branch	—	ConvLSTM	—	curriculum	fast
stratified expert	—	ConvLSTM	—	plain BCE	fast
bare ConvLSTM	—	ConvLSTM	—	plain BCE	fast

D.1 REMOVING THE PDE RESIDUAL (NOPDE)

Retraining with $\lambda_{\text{pde}}=0$ leaves both single-step and multi-step performance statistically indistinguishable from the full model: 8/9 (state, horizon) rollout cells are within 1σ and all are within 1.1σ (Table 6), as are all three single-step states ($+0.30\sigma$ / $+0.84\sigma$ / -0.57σ for OR/WA/AZ). The strongest falsification point — OR at $h=5$, where a physics prior should most suppress error accumulation — shows $\Delta = -0.56\sigma$ with the *sign reversed*. The PDE residual has no clear benefit in these historical base-model records. A null contrast does not identify graph dynamics as the cause of stability, and no hybrid attribution follows by elimination.

Table 6: D.1: multi-step rollout ROC-AUC (3-seed mean ± std).

State	h	Full	noPDE	Δ	Δ/σ
OR	1	0.8031±0.0235	0.8203±0.0192	−0.0172	−0.98
OR	3	0.7464±0.0341	0.7244±0.0110	+0.0220	+1.07
OR	5	0.7489±0.0104	0.7583±0.0271	−0.0093	−0.56
WA	1	0.6848±0.0220	0.6924±0.0141	−0.0076	−0.50
WA	3	0.6719±0.0142	0.6558±0.0342	+0.0161	+0.75
WA	5	0.6959±0.0090	0.7011±0.0036	−0.0052	−0.92
AZ	1	0.7170±0.0088	0.7130±0.0039	+0.0039	+0.71
AZ	3	0.6879±0.0322	0.6750±0.0251	+0.0128	+0.54
AZ	5	0.6801±0.0502	0.6681±0.0292	+0.0119	+0.36

D.2 REPLACING GLOBAL SPECTRAL CONVOLUTION WITH A LOCAL POLYNOMIAL (NOFNO)

We replace the global low-frequency spectral block $z' = U_K(R \cdot U_K^\top z)$ with a ChebNet-style (Defferrard et al., 2016) K-th order local polynomial $\sum_{k=1}^{K} \hat{A}^k z W_k$ under strict parameter matching (627,137 parameters in both variants, enforced by unit test). Single-step ROC-AUC on OR collapses from 0.787 to 0.595 ($+6.10\sigma$, Table 7); in multi-step rollout, 8 of 9 (state, h) cells favor the full model by $\geq 1\sigma$ (OR $h=3$: $+3.29\sigma$; AZ $h=1$: $+2.40\sigma$; only WA $h=3$ at $+0.60\sigma$ is inseparable). The historical contrast favors the base spectral pathway in several cells; it does not repair the hybrid branch-attribution controls.

Table 7: D.2: single-step pooled micro (3-seed mean ± std).

State	Full PR-AUC / ROC-AUC	noFNO PR-AUC / ROC-AUC	ΔROC-AUC	Δ/σ
OR	0.0445±0.0012 / 0.7868±0.0113	0.0218±0.0072 / 0.5949±0.0533	+0.1919	+6.10
WA	0.0262±0.0016 / 0.7129±0.0072	0.0216±0.0032 / 0.6888±0.0394	+0.0241	+1.04
AZ	0.0514±0.0042 / 0.7043±0.0103	0.0584±0.0059 / 0.7191±0.0673	−0.0148	−0.38

D.3 SPECTRAL BASIS CAPACITY K (PROTOCOL-DEPENDENT)

To test whether the state gap stems from insufficient high-frequency expressiveness of the top-32 basis, we sweep $K \in \{4, 8, 16, 24, 64, 128\}$ with three seeds per point and evaluate the $K=32$ main configuration under the same protocol and evaluation pool (Table 8, plotted in Figure 8). For $K > 32$ the eigenbasis is recomputed at the larger truncation; that the $K=32$ basis is its leading prefix is expected by construction but is not independently verified numerically. The two protocols do not agree.

Single step. The three-state mean ROC-AUC declines monotonically from $K=4$ (0.8528) to $K=128$ (0.7551); $K=32$ lands at 0.8049 and $K=128$ sits 0.0498 below it. This is a statement about the three-state mean, not about each state: OR rises over $K=4 \to 16$ and AZ over $K=64 \to 128$. The small-K advantage is not established either — $K=4$ leads $K=8$ by 0.0026, inside seed noise (± 0.005–0.033) — whereas the decline at the top end is uniform across states ($K=128$ is below $K=32$ on OR, WA and AZ).

Rollout. At h=5 the ordering reverses: K=32 is a strict single peak of the three-state mean (0.7286, against 0.7123 at K=24 and 0.7225 at K=64), and both ends lose about 0.075 against it (K=4 0.6534, K=128 0.6534). The peak is again not per-state universal (WA peaks at K=64), and both tails are seed-noisy (the AZ h=5 seed std is ±0.109 at K=4 and ±0.126 at K=128), so we read the shape of the curve rather than individual tail values. PR-AUC agrees at both protocols, peaking at K=4 single-step (0.1033) and at K=32 at h=5 (0.0642).

The two protocols therefore prefer different capacities: single-step accuracy is maximized at very small K and degrades with capacity, while multi-step rollout saturates near K=32 and is dominated at both ends. Capacity is a protocol-dependent factor in the same sense as the factors listed in Section 4.1: the single-step comparison alone would have selected K=4, and the reported configuration's K=32 is the rollout-side choice. This remains a three-seed observation on one split, and it does not identify spectral dilution or out-of-distribution overfitting as the mechanism.

Configuration correction. The three-row version of this probe previously reported here ran the base spectral configuration — no local branch and no gate, which its per-epoch records confirm through absent (nan) gate statistics — at a single seed, and its numbers were read from in-training out-of-distribution evaluations rather than the pooled protocol used elsewhere in this paper. Its direction is consistent with the single-step end of the curve above, and it carries no information about rollout; Table 8 supersedes it.

Table 8: D.3: spectral capacity sweep, ROC-AUC (3-seed mean ± std over seeds). Both blocks are pooled micro within a state; they differ in protocol (teacher-forced single step vs. free-running h=5 rollout, 8 starts), not in aggregation scope. "Mean" averages the three states. K=32 is the reported configuration, re-evaluated under this sweep's protocol.

	Single step (pooled micro)				Rollout h=5 (pooled micro)			
K	OR	WA	AZ	Mean	OR	WA	AZ	Mean
4	0.8556±0.0279	0.8265±0.0331	0.8763±0.0204	**0.8528**	0.7198±0.0078	0.6609±0.0132	0.5796±0.1090	0.6534
8	0.8567±0.0055	0.8219±0.0088	0.8719±0.0096	0.8502	0.6868±0.0177	0.6679±0.0265	0.6262±0.0653	0.6603
16	0.8627±0.0106	0.8065±0.0067	0.8595±0.0286	0.8429	0.7318±0.0486	0.7026±0.0306	0.6863±0.0407	0.7069
24	0.8394±0.0259	0.7874±0.0208	0.8347±0.0196	0.8205	0.7513±0.0388	0.6732±0.0439	0.7124±0.0424	0.7123
32	0.8362±0.0075	0.7642±0.0047	0.8144±0.0150	0.8049	0.7623±0.0179	0.7094±0.0383	0.7141±0.0205	**0.7286**
64	0.8060±0.0127	0.7379±0.0288	0.7648±0.0151	0.7696	0.7581±0.0235	0.7247±0.0232	0.6847±0.0101	0.7225
128	0.7589±0.0161	0.7169±0.0177	0.7896±0.0123	0.7551	0.6318±0.0537	0.7199±0.0251	0.6085±0.1263	0.6534

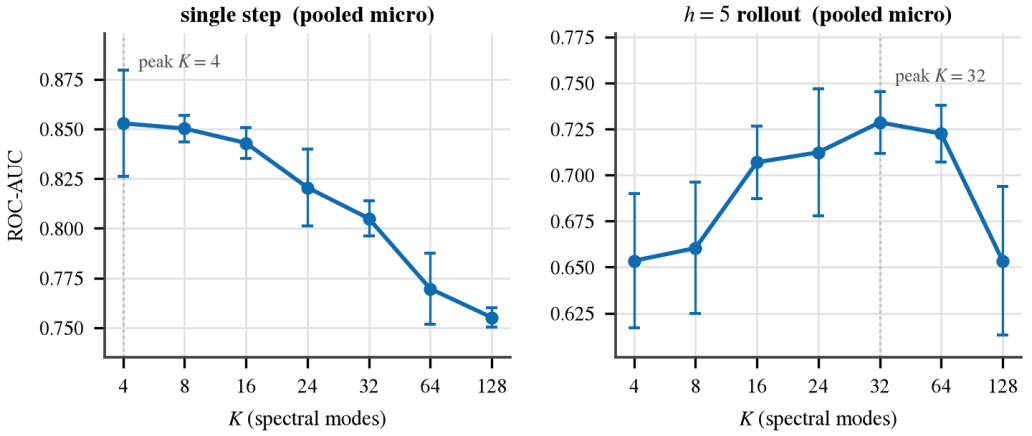


Figure 8: **Capacity optima are protocol-dependent.** Three-state mean ROC-AUC ± seed std (3 seeds) across $K \in \{4, \ldots, 128\}$; both panels are pooled micro. Single-step accuracy peaks at K=4 and declines monotonically with capacity, whereas h=5 rollout peaks at the reported K=32 and is dominated at both ends. The monotonicity holds for the three-state mean, not per state.

D.4 GRID-BASED 2D FNO

Replacing the entire model with a regular-grid 2D FNO (spectral modes 4, block count and parameters matched) tests whether the graph structure of the spectral basis matters (Table 9). Two readings: (i) the 2D FNO significantly underperforms the hybrid main configuration (OR -3.24σ, AZ -3.14σ single-step; 9/9 rollout cells not better); (ii) it is however *not statistically separable from the base spectral configuration* in this three-seed probe (all single-step $|\Delta| \leq 1.22\sigma$). These whole-model comparisons do not isolate fusion or prove that basis choice is immaterial. The necessity of either branch remains a question for normalization-matched controls.

Table 9: D.4: single-step pooled micro ROC-AUC (3-seed mean ± std).

State	2D FNO	GLOW (hybrid)	Δ vs. hybrid	Δ vs. base spectral
OR	0.7656±0.0346	0.8315±0.0066	-0.066 (-3.24σ)	-0.021 (-1.01σ)
WA	0.7289±0.0216	0.7576±0.0231	-0.029 (-1.57σ)	$+0.016$ ($+1.22\sigma$)
AZ	0.7156±0.0526	0.8173±0.0196	-0.102 (-3.14σ)	$+0.011$ ($+0.37\sigma$)

D.5 GCN BASELINE

A single-hop GCN (Kipf & Welling, 2017) ($\hat{A}zW$, i.e. the local block with $K{=}1$) completes the family: local single-hop / local high-order (D.2) / global grid spectral (D.4) / global graph spectral + local fusion (main). The hybrid model beats GCN significantly on OR and AZ (Table 10) and is not worse in 8/9 rollout cells (OR all horizons $+3.03\sigma$ to $+4.78\sigma$; AZ $h{=}1$ $+4.96\sigma$). Unexpectedly, GCN ($K{=}1$) significantly *outperforms* ChebNet ($K{=}32$) in some historical cells; this is not a general causal result that local capacity is harmful. Honest negative points: GCN has the highest single-step WA ROC-AUC of the family (0.7769, -1.24σ, not significant) and beats hybrid at WA $h{=}3$ (-4.15σ), the same cell where hybrid loses to LR — WA mid-horizon is a systematic weakness, not noise.

Table 10: D.5: single-step pooled micro ROC-AUC (3-seed mean ± std).

State	GCN ($K{=}1$)	ChebNet ($K{=}32$)	GLOW (hybrid)	hyb−gcn	cheb−gcn
OR	0.7492±0.0155	0.5949±0.0533	0.8315±0.0066	$+0.082$ ($+8.46\sigma$)	-0.154 (-4.82σ)
WA	0.7769±0.0137	0.6888±0.0394	0.7576±0.0231	-0.019 (-1.24σ)	-0.088 (-3.66σ)
AZ	0.7647±0.0171	0.7191±0.0673	0.8173±0.0196	$+0.053$ ($+3.50\sigma$)	-0.046 (-1.14σ)

D.6 LANDFIRE FUEL FEATURES (MIXED SIGNAL; EXCLUDED FROM MAIN CONFIGURATION)

We replace the fuel placeholder with real LANDFIRE FBFM40 maps (Scott & Burgan, 2005) (version chosen by fire year to avoid post-burn leakage, mode resampling, 40 classes → 7-group one-hot, input dimension 13 → 20). Results are mixed (Table 11): OR degrades systematically (single-step PR-AUC -2.35σ; rollout $h{=}1$ PR-AUC -4.69σ, $h{=}5$ PR-AUC -3.40σ) and all three states are non-positive at $h{=}5$ (WA -2.98σ), while AZ $h{=}1$ PR-AUC improves $+4.34\sigma$ and WA single-step PR-AUC $+2.62\sigma$. These mixed historical effects do not identify fuel-distribution similarity as their cause or establish dynamics as the source of the reference model's gain. Fuel remains absent from the reference recipe; its benefit is not a universal negative result.

Expanded-pool fuel retest (historical, under audit). The old expanded-pool summary reported mixed effects, including WA single-step PR-AUC $+3.1\sigma$ and AZ $h{=}1$ ROC-AUC $+3.7\sigma$. These are not repaired estimates. Because the T1 reference comparison changes normalization and training composition, the retest cannot show that fuel's benefit scales with training size. A within-pool matched fuel/no-fuel comparison and checkpoint provenance check are still needed; no universal fuel benefit or recipe ceiling is claimed.

Table 11: D.6: single-step pooled micro (3-seed mean ± std), fuel vs. hybrid main configuration.

State	Metric	+Fuel	GLOW (hybrid)	Δ (fuel−hyb)
OR	ROC-AUC	0.8134±0.0165	0.8315±0.0066	-0.018 (-1.76σ)
OR	PR-AUC	0.0523±0.0056	0.0687±0.0107	-0.016 (-2.35σ)
WA	ROC-AUC	0.7688±0.0103	0.7576±0.0231	$+0.011$ ($+0.77\sigma$)
WA	PR-AUC	0.0424±0.0037	0.0348±0.0034	$+0.008$ ($+2.62\sigma$)
AZ	ROC-AUC	0.8013±0.0437	0.8173±0.0196	-0.016 (-0.58σ)
AZ	PR-AUC	0.1025±0.0162	0.1011±0.0073	$+0.001$ ($+0.14\sigma$)

D.7 Maximal short-horizon configuration (best-effort stack)

Under protocol audit. Stacking a five-frame window with LANDFIRE fuel gives the strongest short-horizon arm of this study, reported here as the recipe's ceiling rather than as a replacement reference. It is normalization-unmatched to the reference and retains historical grade.

Against ConvLSTM the arm leads at single step in all three states (GLOW $0.871/0.842/0.865$ versus $0.810/0.809/0.867$ for OR/WA/AZ: $+6.7\sigma$, $+4.4\sigma$, parity) and at $h \leq 3$ in every cell, and it supplies the project's only positive slow-state h=5 cell (WA $+3.5\sigma$, 0.746 versus 0.709). The fast-end condition is untouched: AZ h=5 falls to 0.699 against ConvLSTM's 0.851 (-10.3σ, a wider gap than the reference's), the fast stratum of Table 3 is unchanged, and the degradation slope steepens. Short-range sharpness and free-running stability therefore trade against each other rather than composing.

D.8 no-global and noPDE on the hybrid configuration

Both ablations rerun on the hybrid main configuration (Table 12). *No-global* (removing the spectral branch, keeping only local convolution) shows a mirror pattern: single-step performance *improves* on WA/AZ ($+3.3$–3.9σ) while multi-step rollout collapses (OR h=3 PR-AUC -7.24σ; all three states negative in PR-AUC at h=5). This historical pattern motivates branch comparisons, but does not identify the source of long-horizon stability or single-step sharpness. The intended matched-control attribution remains under audit. For *noPDE*, the historical hybrid contrast is less uniformly null under the declared Welch standard-error scale: 12/18 rollout metric cells are within 1.1σ, five more within 1.5σ, and the isolated WA h=5 ROC-AUC reaches $+3.12\sigma$ in favor of noPDE, i.e. opposite to a physics-helps reading. The retained contrasts do not show a consistent physics benefit; they do not prove a null effect or identify an alternative mechanism.

Table 12: D.8: Δ/σ relative to the hybrid main configuration (positive = ablated variant better).

Cell	Metric	no-global	noPDE
OR single-step	ROC-AUC/PR-AUC	-0.18σ / $+0.50\sigma$	-0.82σ / -0.83σ
WA single-step	ROC-AUC / PR-AUC	$+3.28\sigma$ / $+3.90\sigma$	-0.02σ / $+0.11\sigma$
AZ single-step	ROC-AUC / PR-AUC	$+3.45\sigma$ / $+3.32\sigma$	-0.63σ / -0.57σ
OR h=3	ROC-AUC / PR-AUC	-3.82σ / -7.24σ	-0.23σ / $+0.28\sigma$
OR h=5	ROC-AUC / PR-AUC	-3.62σ / -3.56σ	-0.39σ / -1.21σ
WA h=5	ROC-AUC / PR-AUC	-1.67σ / -3.33σ	$+3.12\sigma$ / $+1.48\sigma$
AZ h=5	ROC-AUC / PR-AUC	-0.28σ / -2.27σ	-0.27σ / -0.54σ

D.9 Speed-adaptive gate (negative)

Motivated by the gate diagnosis (Appendix E), we append two explicit spread-speed features (local 5×5 and global mean of newly burned area over the input window) to the gate input. Performance does not improve in any state (Table 13): single-step AZ PR-AUC -3.54σ, OR ROC-AUC -2.61σ; rollout OR h=1 ROC-AUC -4.33σ (the only positive cell, WA h=5 $+2.98\sigma$, coincides with the noisiest cell across all ablations). Critically, the gate–speed correlation becomes *more negative* (-0.420 vs. -0.259 without the features): the tested features did not produce the desired correlation.

This does not establish absence of a routing gradient or falsify all speed-conditioned gates; the variant is retained as a historical unsuccessful probe.

Table 13: D.9: speed-adaptive gate vs. hybrid main configuration (3-seed mean).

Cell	Metric	+Speed feats	GLOW (hybrid)	Δ/σ
OR single-step	ROC-AUC	0.810	0.832	-2.61
AZ single-step	PR-AUC	0.085	0.101	-3.54
WA single-step	ROC-AUC	—	—	-0.29
OR h=1	ROC-AUC	—	—	-4.33
corr(g, log speed)	—	-0.420	-0.259	—

D.10 NO-LOCAL CONTROL, LOCAL-BRANCH DEPTH, AND TEMPORAL AGGREGATOR

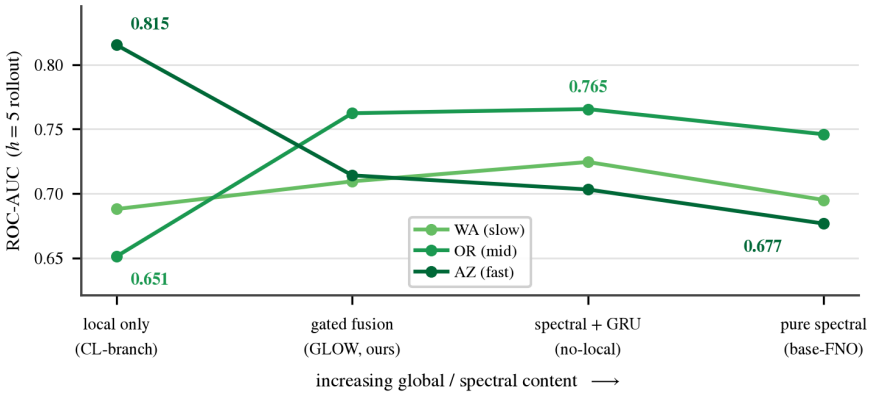


Figure 9: **The optimal local–global composition depends on the regime.** Free-running h=5 ROC-AUC for four arms spanning the local→global axis (per-state pooled ROC-AUC over the 27 cross-state fires, three seeds; one curve per state): a local-memory-only arm (no spectral branch, no gate), the gated fusion of GLOW, a spectral+GRU arm without a local branch, and the pure-spectral base configuration. Mid OR rises and then plateaus ($0.651 \to 0.762 \to 0.765 \to 0.746$, peaking at the spectral+GRU arm) while fast AZ falls monotonically ($0.815 \to 0.714 \to 0.703 \to 0.677$); the curves cross, and GLOW's gated fusion sits near their intersection. The four arms are a composition ordering rather than a single-variable mixing sweep—the end-to-end arms remove a branch *and* the gate—so intermediate mixing coefficients are not required to establish the crossing.

Under protocol audit. The loc4, concat and K1/K5 arms use `global_norm=False`; the reference uses `global_norm=True`, so their historical values in Table 14 are retained for traceability rather than as verified one-factor ablations. The no-local arm *has* since been retrained under the matched contract (§D.10-a).

(a) no-local (spectral + GRU), normalization-matched. Removing the local branch and gate distinguishes temporal aggregation from spatial fusion. Under the matched contract the arm is verified and reads single-step ROC-AUC $0.827/0.748/0.785$ against the reference's $0.836/0.764/0.814$ (OR/WA/AZ), and h=5 rollout $0.765/0.724/0.703$ against $0.762/0.709/0.714$: the dual-branch model leads at single step in every state, while at h=5 the spectral+GRU arm matches it within seed noise on OR and WA and GLOW holds only AZ — so the local branch is what the fast regime requires rather than a uniform gain. The historical *spectral-only is best on OR* mirror (h=5 OR 0.779) does not survive matching and is withdrawn as a normalization artifact.

(b) Local-branch depth (2 → 4 blocks, nominal receptive field $\sim 9\times 9 \to \sim 17\times 17$). The structural change is retained, but the old negative scores do not falsify a receptive-field limitation or show that added capacity is harmful. A matched-normalization depth comparison is pending.

(d) Window length. K1/K5 were intended to test the value of history. Their contrasts against the three-frame GLOW reference are normalization-confounded. A matched per-case re-run of all three window lengths (one normalization family throughout) restores a clean single-step dose–response, monotone in every state (OR $0.766/0.810/0.878$ for $K{=}1/3/5$; used in Section 5.4), while the deep-horizon pattern is regime-dependent—so the joint claim that five frames improve the main configuration's rollout stays suspended. Historical ConvLSTM window results must be kept distinct from the affected GLOW comparison.

(c) Temporal aggregator: GRU vs. frame concatenation. The concatenation variant uses a zero-initialized linear map ($K_{\text{seq}}D \to D$). Its old comparison also changes normalization; small observed differences are not an equivalence test. We withdraw the claim that aggregator choice is not load-bearing.

Table 14: D.10: historical, normalization-confounded no-local, local-depth and temporal-aggregator comparisons (under protocol audit), pooled ROC-AUC (3-seed mean). Bold marks historical maxima, not verified wins. Rollout at $h \in \{3, 5\}$, 8 starts, $K_{\text{ens}} = 10$. †normalization-matched retrain (verified); rows without a mark are the historical, `global_norm=False` arms retained for traceability only, and their bold entries are superseded.

	single-step			$h{=}3$			$h{=}5$		
Arm	OR	WA	AZ	OR	WA	AZ	OR	WA	AZ
GLOW (hybrid)	**0.832**	**0.758**	**0.817**	0.683	**0.716**	0.718	0.762	0.699	0.720
no-local (spectral+GRU)†	0.827	0.748	0.785	0.732	0.689	0.708	0.765	0.724	0.703
no-local, historical*	0.795	0.695	0.730	0.778	0.670	0.741	0.779	0.712	0.727
local $\times 4$	0.804	0.710	0.717	0.733	0.654	0.719	0.753	0.691	0.682
concat aggregator	0.834	0.746	0.798	0.737	0.671	0.692	0.735	0.687	0.681
base-FNO	0.787	0.713	0.704	0.746	0.672	0.688	0.749	0.696	0.680

(e) Input locality of the local branch. The local branch currently consumes the post-aggregation latent, so its input has already passed through the spectral branch and the GRU; feeding it the pre-mixing per-node encoding instead (zero new parameters) is the clean single-factor test of whether that global mixing is what limits the local path. It is not: every state degrades monotonically (single-step $-0.013/-0.022/-0.032$; $h{=}1$ $-0.012/-0.012/-0.009$), which says the local branch *benefits* from the globally mixed representation rather than being contaminated by it.

(f) Multi-scale local receptive field. Replacing the stacked $3{\times}3$ pair with two *parallel* dilated convolutions up to dilation 2 (parameter-exact: two $D{\to}D$ kernels, only the dilation and the parallel wiring change) is the one local-branch modification of this study with a directional effect, and the direction tracks the regime axis rather than lifting the mean: AZ is positive at every horizon ($+0.011$ single-step, $+0.012/+0.018/+0.024$ at $h{=}1/3/5$) while OR is negative at every horizon ($-0.016/-0.020/-0.039/-0.032$), so the three-state mean is flat (single-step -0.0005, $h{=}5$ -0.007) and the pre-registered null criterion holds.

Taken together, five local-branch modifications—depth, operator, presence, input locality, and receptive-field geometry—produce no net gain, and the one that produces a directional effect trades along the regime axis. The branch's expressiveness is therefore not what bounds the fast regime.

(g) FiLM fusion (modulation instead of prediction). Retrained under the matched contract (Appendix J), this variant lets the spectral branch stop predicting and only modulate the local branch (channel-wise gain and shift, zero-initialized). It *halves* the fast-state single-step deficit to ConvLSTM (0.838 vs. 0.867) and matches ConvLSTM at $h{=}1$ (0.790 vs. 0.792), yet loses 0.067 ROC-AUC on OR $h{=}5$, where the parallel gate that lets the global features *predict* remains best. Modulation therefore does not substitute for prediction: conditioning the local path helps the fast state comparatively but not decisively, while the slow state still wants the global branch to produce the score—the fusion-form counterpart of the ranking result.

D.11 WIDENED-RING TRAINING CONTROL

Normalization-matched retrain (what changed). Ring2 training widens the candidate set to Chebyshev radius 2 while all evaluation remains one-ring. The audit left one gap: the ring2 GLOW

arm trained under per-case normalization while the reference used global normalization, so neither levels nor the family ranking could be compared. A retrain with `--ring-r 2 --global-norm` closes that gap for GLOW (3 seeds $\times$ 3 states; all 9 checkpoints record `global_norm=True` and `ring_r=2`, evaluated one-ring with 8 starts and K_{ens}=10).

Counting (unchanged). The graded counting schema (threshold events $\rightarrow$ observed unburned events $\rightarrow$ new-state pixels $\rightarrow$ unique pixels) gives 1,766 observed-unburned events for the main GLOW arm, of which 664 (38%) become new-state pixels and 290 (16%) are unique, against 215 events for ConvLSTM. Ring2 training cuts the event count 28$\times$ (1,766 $\rightarrow$ 63; single seed—event counts are RNG-sensitive at the $\pm$20% level) while leaving it $\sim$16$\times$ above the ring2 ConvLSTM arm (63 vs. 4): widening the candidate set does not remove the global model's tendency to ignite beyond the scored ring. *This count was not re-run on the matched arm*, so it remains a per-case-normalized reading.

Levels under matched normalization. Relative to the matched one-ring reference, matched ring2 GLOW loses pooled ROC-AUC at h=5 in all three states ($-0.062/-0.011/-0.024$ for OR/WA/AZ; Figure 11) and is mixed at single step ($-0.016/+0.022/+0.003$). Against the *published* one-ring ConvLSTM the OR h=5 lead disappears ($+0.062 \rightarrow +0.000$) and the AZ h=5 deficit widens ($-0.137 \rightarrow -0.161$); at single step the OR lead narrows ($+0.027 \rightarrow +0.010$), WA recovers ($-0.045 \rightarrow -0.023$) and AZ is unchanged ($-0.052 \rightarrow -0.050$). Widening the ring therefore does not buy absolute performance and does not preserve the ordering.

Why we still do not report a matched same-ring family comparison. The only ring2 ConvLSTM arm available trained under per-case normalization; the matched-norm retrain covered GLOW only. Comparing matched GLOW against an unmatched ConvLSTM would reintroduce exactly the normalization mismatch that forced the original ranking claim to be withdrawn, so we restrict every statement above to (a) matched GLOW ring2 versus matched GLOW ring1, and (b) matched GLOW ring2 versus the *published* one-ring ConvLSTM. We report no claim about how ring2 ConvLSTM behaves under matched normalization.

Claim status. The withdrawn ranking claim is *not* reinstated: under matched normalization the GLOW advantage at h=5 is erased in OR and its fast-state deficit grows, so expansion moves the ordering against GLOW rather than preserving it. The mechanism is consistent with the audit's account—ring2 labels draw from distance-2 candidates, which dilute near-front discrimination.

D.12 Multi-step (rollout-aware) training

Protocol and grade. Both arms retrain the reference architecture with the end-to-end multi-step objective described above—a differentiable free-running loss with soft state feedback and per-horizon focal terms—under the same frozen contract as the reference: the checkpoint manifests record the same 15-fire California training pool, the same split, `global_norm=True`, K_{ens}=10 and 8 rollout starts. The contrast is therefore a same-seed, same-protocol change of a single training factor—the objective, together with the rollout-based validation selector that matches it—and we grade it *verified*. C1 keeps the reference learning rate (3×10^{-4}) and runs three seeds; C2 uses 3×10^{-5} and runs one seed (42). Unlike the reference, whose validation selection lands in the warm-up stage (epochs 2–9), these arms select on validation rollout ROC-AUC and land at epochs 10–30: the objective changes what validation prefers.

Two protocols, reported separately. Table 15 gives the rollout protocol the arms are trained for, Table 16 the teacher-forced single-step protocol they are no longer trained for. Both blocks are pooled micro; they differ in protocol (free-running multi-step vs. teacher-forced single step) and in prediction horizon, not in aggregation scope, and every Δ is computed within one block against the three-seed reference. Figure 10 plots the per-state curves. The reference rows are the verified retraining (Appendix C), not the historical single-step anchor of Table 4 nor the historical ablation arm of Table 14.

A smaller basis under the same objective (K=8). One further arm pairs the identical multi-step loss with a reduced spectral basis (K from 32 to 8; all other flags unchanged, three seeds). It recovers the single-step level on average ($-0.009/+0.006/+0.017$, OR/WA/AZ; mean $+0.005$), keeps a deep-horizon gain (h=5 $+0.022$ averaged over states; AZ $+0.081$) and narrows the fast-state h=5 gap to ConvLSTM from 0.137 to 0.056, while the h=1 anchor is flat (-0.005). The

Table 15: **D.12: multi-step training under the rollout protocol.** Pooled-micro ROC-AUC at $h \in \{1, 3, 5\}$ (8 starts, K_{ens}=10) on the 27 cross-state fires. Δ is the multi-step arm minus the three-seed GLOW reference, within this protocol. C1 is trained with the reference learning rate (3×10^{-4}, three seeds) and C2 with 3×10^{-5} (one seed). ConvLSTM checkpoints are unchanged.

	WA (slow)			OR (mid)			AZ (fast)		
Model	h=1	h=3	h=5	h=1	h=3	h=5	h=1	h=3	h=5
GLOW (reference)	0.726	0.702	0.709	0.854	0.722	0.762	0.764	0.722	0.714
C1: +multi-step	0.694	0.685	0.725	0.790	0.693	0.755	0.735	0.775	0.781
Δ	−0.032	−0.017	+0.016	−0.064	−0.029	−0.007	−0.029	+0.053	+0.067
C2: +multi-step	0.735	0.743	0.734	0.778	0.677	0.750	0.733	0.763	0.774
Δ	+0.009	+0.041	+0.025	−0.076	−0.045	−0.012	−0.031	+0.041	+0.060
ConvLSTM*	0.761	0.750	0.715	0.801	0.621	0.700	0.792	0.800	0.851

Table 16: **D.12: multi-step training under the teacher-forced single-step protocol.** Pooled-micro ROC-AUC over the full candidate set. The same aggregation scope as Table 15 (pooled micro within a state); the two blocks differ in protocol and prediction horizon, not in scope. "Mean" averages the three states.

Model	OR	WA	AZ	Mean
GLOW (reference)	0.836	0.764	0.814	0.805
C1: +multi-step	0.781	0.727	0.781	0.763
Δ	−0.055	−0.038	−0.034	−0.042
C2: +multi-step	0.767	0.736	0.764	0.756
Δ	−0.069	−0.029	−0.050	−0.049
ConvLSTM*	0.810	0.809	0.867	0.829

capacity sweep of Table 8 supplies the control: at K=8 with the *single-step* objective, single-step already leads by 0.045 and h=5 loses 0.068, so the smaller basis does not remove the short-range cost of multi-step training—it makes that cost affordable. The arm is normalization-matched to the reference (`global_norm` recorded `True`, same split and pool) but post-dates the recovery audit, so it is reported here rather than in the audited set.

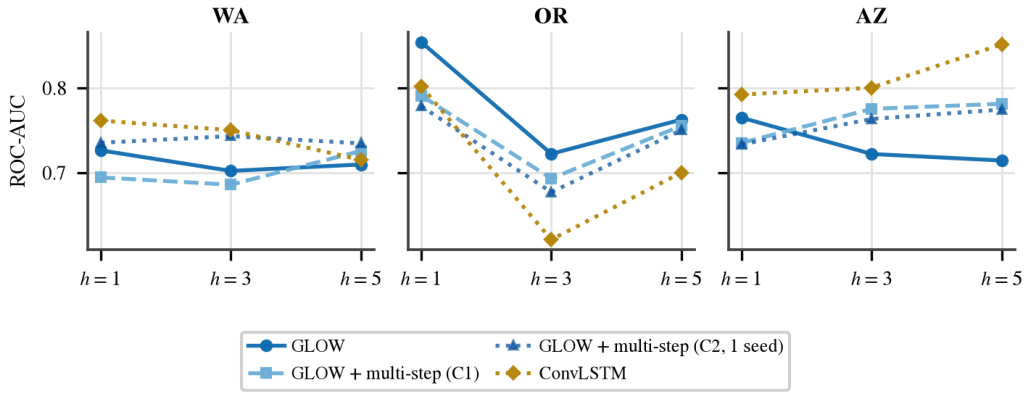


Figure 10: **Multi-step training redistributes rather than lifts.** Free-running ROC-AUC at $h \in \{1, 3, 5\}$ per state, pooled micro, seed mean. The multi-step arms lose at h=1 in every state and gain at h=5 in WA and AZ. The C2 curve (dashed) is a single seed and is not comparable to the three-seed curves.

What moves, and what it costs. The displacement is a reallocation rather than a gain. Positive movement is confined to the deeper horizons: AZ h=5 recovers +0.067 (positive on all three seeds, +0.070/ + 0.084/ + 0.047) and AZ h=3 +0.053 (also positive on all three seeds), WA h=5 reaches +0.016 on the strength of one seed (+0.097, the other two negative), and OR h=5 is flat

(-0.007). Against the unchanged ConvLSTM checkpoints this halves the fast-state deep-horizon gap ($-0.137 \rightarrow -0.070$), turns WA h=5 from -0.005 to $+0.011$, and narrows AZ h=3 from -0.078 to -0.025, while the gap *widens* at h=1 in all three states. The cost is short-range and consistent: teacher-forced single-step ROC-AUC loses $0.055/0.038/0.034$ and the free-running h=1 anchor $0.063/0.032/0.030$, with the Oregon single-step loss the most reproducible feature of the arm ($-0.056/-0.058/-0.051$ across seeds). A paired per-fire bootstrap over the fires with $\geq$10 positives at each graded horizon (15 fires at h=5, 15 at h=1, the two sets differing in three fires) gives $+0.017$ (95% CI $[-0.013, +0.048]$) at h=5 against -0.045 ($[-0.066, -0.025]$) at h=1: at this sample size the deep-horizon recovery is descriptive, whereas the short-horizon tax is resolvable. Note also that the two aggregation scopes do not agree on the size of the recovery (per-fire mean $+0.017$ versus three-state mean $+0.025$), because fires and states are weighted differently; we quote the state-mean convention used in the main text and the per-fire interval as the uncertainty statement.

Learning-rate sensitivity. C2 (3×10^{-5}, seed 42) reproduces the displacement at essentially the same magnitude: $+0.024$ at h=5 against the three-seed reference mean, with a single-step cost of -0.049 versus C1's -0.042. The effect is therefore not an artifact of the step size. C2 is a single seed, and its Δ mixes a single-seed arm with a three-seed mean; measured against the same seed of the reference, its AZ h=5 displacement is $+0.072$, in line with C1's seed-42 value. We do not read C2 as an independent replication.

How this connects to the rest of the paper. An objective of the same structural form was previously run as an exploratory arm on the base spectral configuration (Curriculum paragraph of Appendix C), where it also raised the deep horizon at a short-range cost; the pattern therefore reproduces across two configurations rather than being specific to the hybrid backbone. The interpretation we adopt is not that multi-step training is a better recipe —short-horizon accuracy pays for it, and the recovery stays below per-fire resolution—but that the global–local ordering is *controllable*: identical architecture, data and seeds return a different ordering under a different objective, and the part of the ranking that moves consistently across seeds sits at the fast state's deep horizon, which is the quantity the regime account names.

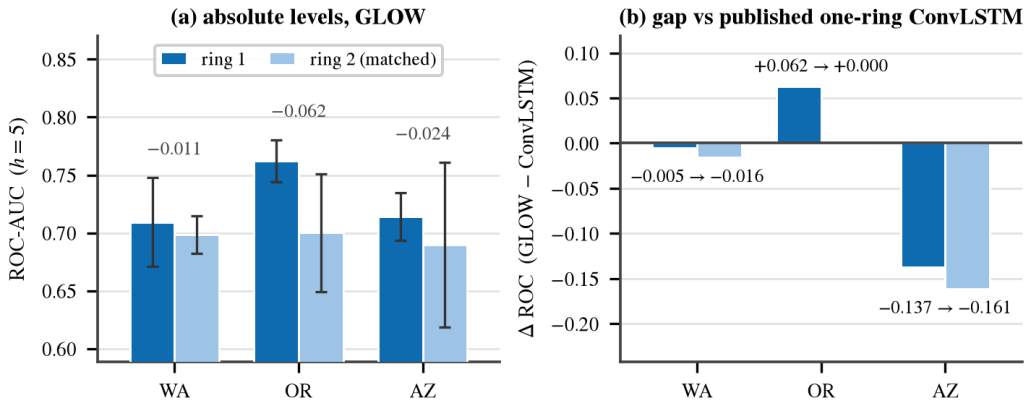


Figure 11: **Widening the candidate ring under matched normalization.** (a) h=5 GLOW levels fall in all three states (WA/OR/AZ: $-0.011/-0.062/-0.024$). (b) Gap against the published one-ring ConvLSTM: the OR lead disappears ($+0.062 \rightarrow +0.000$) and the AZ deficit widens ($-0.137 \rightarrow -0.161$). No same-ring family comparison is shown: the only ring2 ConvLSTM arm trains under per-case normalization, so comparing it against the matched ring2 GLOW would reintroduce the mismatch that forced the ranking claim to be withdrawn.

D.13 Refinement probes: inference-side sub-stepping and training-side padding (negative)

Long rollouts degrade, so we tested the manipulation complementary to Obs. 4: instead of coarsening the data, refine the inference. Inside each observed interval the model is applied n times in

a row, each time consuming the state it just produced together with the same $t-1$ exogenous field, so per-step travel is divided by n; the checkpoints are unchanged. Under the protocol used in this paper—the hard cumulative update $s \leftarrow \max(s, \mathbf{1}[p > 0.5])$—the effect is exactly null. Per-case seed-mean h=5 ROC-AUC, paired bootstrap over the same 15 fires: the ConvLSTM−GLOW displacement moves by -0.0001 (n=1→3, 95% CI $[-0.0005, +0.0004]$, p=0.81), the fast band (> 4 cells/step) by -0.0006 (p=0.38) and the slow band by $+0.0005$ (p=0.04). The win matrices, the correlation with speed ($+0.550 \rightarrow +0.548$) and the absolute levels (GLOW $0.7368 \rightarrow 0.7370$) all move within noise.

The cause is the update rule rather than the model: $\max(s, \mathbf{1}[p > 0.5])$ is idempotent, so further sub-steps add almost no area. ConvLSTM adds exactly 0 cells at all 15 points, and GLOW saturates at $14/15$ (mean $+1.07$ cells, largest $+16$).

Replacing the update with the soft union used during training ($s \leftarrow s + (1 - s)p$) does produce a large effect, but not the predicted one. Saturation disappears ($0/15$ points; the second and third sub-steps add up to $+88.9$ cells for GLOW and $+128.0$ for ConvLSTM), and the displacement becomes $+0.0567$ (p=0.0008), concentrated in the slow band ($+0.1001$, p=0.0004, with ConvLSTM moving from -0.028 to $+0.072$). The direction is the opposite of the prediction, and the regime structure is *erased* rather than strengthened: $\mathrm{corr}(\log \text{speed},\ \mathrm{CL} - \mathrm{GLOW})$ falls from $+0.391$ to -0.069.

Two reading rules. The soft arm is a state-update ablation run with an already-trained single-step model, not this paper's protocol, and it improves nothing: at n=1 the soft rule is simply worse (GLOW h=5 0.6718 against 0.7368; ConvLSTM 0.7024 against 0.7285), so the hard cumulative rule is a useful choice rather than an unaligned accident. The null then says that the speed axis is not manipulable at inference alone: Obs. 4 moves it from the data side, and this probe shows the inference side does not. Both arms were preregistered before the readout was computed.

Training-side padding: no detected fast-band narrowing, and it costs short-range skill. Obs. 4 moves the axis from the *data* side. We then asked the complementary question: can a *training* intervention reach the same axis? Two routes were preregistered before any C1 result was seen (15 California fires, 60 epochs, three seeds; nine training jobs, one evaluation job, all completed). Route 1 leaves the labels and the evaluation protocol untouched and changes only the input window, replacing $[x_{t-K+1}, \ldots, x_t]$ by $[x_{t-K+2}, \ldots, x_t, x_t]$ so that the last step of the window appears not to advance; its two arms differ only in whether the epoch-internal checkpoint selection also sees padded windows, so they are not independent evidence. Route 2 keeps the interval target but splits each interval into two forward passes, advancing the soft state in between and scoring the interval label once; that switch takes effect only on the multi-step training path, so route 2 is the main configuration *plus* multi-step training, and its control is that arm (Table 15)—a *reused*, not re-run, job from a separate preregistration.

Neither route narrows the fast band. Under the same per-case seed-mean h=5 ROC-AUC, 15-fire speed axis and paired bootstrap, each route against its own control, the fast band (> 4 cells/step, four fires) displacement is $+0.003$ (95% CI $[-0.070, +0.078]$, p=0.92) for route 1a, -0.016 (p=0.36) for route 1b and -0.036 (p=0.04) for route 2: route 1 is indistinguishable from zero, and route 2 is significantly in the opposite direction rather than reaching the preregistered $+0.03$ narrowing. At h=3 the band reads $-0.022/-0.029/-0.028$, the same negative direction. The arms do not sit at zero over the whole axis either: all-band displacements are $-0.027/-0.033/-0.018$ (p=0.08/0.0002/0.11), the same order of magnitude as the $+0.0567$ of the soft arm above, with the opposite sign.

Short-range skill, by contrast, degrades. h=1 ROC-AUC (the 13 fires surviving the $n_{\mathrm{pos}} \geq 5$ filter) falls by $0.041/0.035/0.005$ against each arm's own control, and single-step ROC-AUC by $0.064/0.068/0.036$ across the 27 test fires (per-case macro; the pooled micro convention gives $0.072/0.074/0.043$). Route 2's small h=1 cost is an artifact of its control: measured against the main table, the same arm falls by 0.053.

Where the arms do move is the *slow* band, and the movement is GLOW losing ground: the ConvLSTM−GLOW gap narrows from 0.064 to 0.022 (route 1a, p=0.002) and to 0.020 (route 1b, $p<10^{-4}$)—but ConvLSTM is the identical checkpoints in every arm, and GLOW's own slow-band ROC-AUC falls from 0.7564 to $0.7143/0.7128$ (above ConvLSTM's 0.6923). On the fast band

route 1a's GLOW is unchanged instead ($0.6530 \rightarrow 0.6557$), so these probes support no cross-band mechanism sentence.

Three boundaries travel with this result. (i) The fast band holds only four fires and they are the four Arizona cases, so band and state are collinear here and the band reading cannot be separated from a state effect. (ii) With four fires the test has little power: dropping a single fire moves route 1a's estimate over $[+0.032, -0.037]$—dropping `AZ_2025_23` crosses the $+0.03$ threshold and dropping `AZ_2020_39` flips the sign—and the same-configuration difference between routes 1a and 1b is 0.018, so the threshold is only about $1.6\times$ the run-to-run floor; the p-values quoted across bands and arms are also uncorrected for multiple comparisons. The honest reading is *not detected*, not *proven absent*. (iii) Route 2's displacement is defined against its multi-step control; measured against the main table instead, the same arm is $+0.043$ on the fast band, so the sign of this criterion is a property of the reference rather than of the arm. Both routes and this readout were preregistered before any result was seen.

E Gate Interventions: Protocols and Full Results

Under protocol audit. The additional gate-capacity arm replaces the linear gate with a two-layer softmax MLP ($2D \rightarrow D/4 \rightarrow 2$ logits, uniform zero-initialized start), but also changes normalization relative to the reference. It does not establish that gate capacity is non-binding. The historical feature and supervision probes below remain exploratory; together with arm (v) they do not prove an intrinsic routing impossibility or an exhaustive six-intervention boundary. Only arm (v) is protocol-matched (same pool, frozen normalization); it tests the *form* of the fusion constraint rather than its capacity. Table 17 (Figure 12) retains four earlier diagnostic/probe records plus the matched arm (v).

(i) Gate diagnosis (no intervention). We run the 9 main-configuration checkpoints deterministically (σ=0) over all 27 test cases and record the local-branch gate weight g at every step. The global branch dominates everywhere (case-mean $g \in [0.22, 0.38]$; global weight ≈0.65–0.75, no case flips). Across cases, corr(g, log spread speed) $= -0.259$ — opposite to the intended speed-routing sign. State-level confounding is a possible explanation, not an identified cause. Within cases, however, the temporal correlation between g and per-step newly burned area is positive in 25/27 cases (up to $+0.52$). These are associations, not a direct test of transferable routing information or the available optimization gradient.

(ii) Feature injection (D.9). Appending explicit speed features to the gate input (Appendix D, D.9) makes the cross-case correlation *worse* (-0.420) and degrades performance in all states. The observed result does not measure whether the loss surface has a gradient toward regime-dependent routing.

(iii) Explicit gate supervision (gsup). We add a supervision term $0.5 \cdot \mathrm{BCE}(g, \mathbf{1}[\text{newly burned/step} > 2.5])$ that directly teaches the gate to open the local branch on fast fires. Mechanistically this works: corr(g, log speed) turns positive for the first time ($+0.253$; temporal correlation positive in 24/27 cases; g becomes bimodal). Performance, however, collapses: single-step OR ROC-AUC 0.7112 vs. 0.8315 (-4.96σ), AZ -3.33σ; rollout OR ROC-AUC -2.4 to -4.7σ at all horizons. This run carries a loss-weight confound: at λ=0.5 the gate BCE (≈0.10–0.21) is $\sim$20$\times$ the focal data loss (≈0.005), and the $\sim$90% slow-window label imbalance pushes mean g down to 0.10, nearly closing the local branch.

(iv) Reduced-weight, class-balanced probe (gsupb). We reduce λ tenfold to 0.05 and add class balancing (positive-window BCE weight 9). The mechanism strengthens further (corr $= +0.410$; mean g recovers to 0.12–0.24), yet performance still does not recover: single-step OR ROC-AUC 0.7212 (-11.15σ), AZ -2.72σ, WA -2.29σ; all nine rollout cells are negative in mean ROC-AUC (OR h=5 -4.45σ; the only near-parity cell is AZ h=5 at -0.12σ). Two tested loss settings do not exclude a loss-weight artifact or prove an intrinsic conflict between supervision and prediction. The hypothesis that hard speed targets constrain useful fusion remains unconfirmed. We did not sweep $\lambda \leq 0.01$; this is a limited probe, not an exhaustive gate-routing test.

Table 17: Gate probes. Rows (i)–(iv) are historical and under protocol audit; row (v) is protocol-matched (same pool, frozen normalization). "corr" is the cross-case correlation between the local-branch gate weight and log spread speed; the desired sign is positive.

Intervention	Change	corr(g, log speed)	Performance	Verdict
(i) Diagnosis	none	-0.259	(reference)	no spontaneous routing
(ii) D.9 features	+2-d speed input	-0.420	all states negative	tested probe negative
(iii) gsup	+BCE, λ=0.5	$+0.253$	collapses (-5.0σ OR)	confounded
(iv) gsupb	λ=0.05, pos-wt 9	$+0.410$	still collapses (-11.2σ OR)	conflict not identified
(v) affine	no squash on g; coefficients unconstrained	-0.195	step $-0.026/-0.026/-0.031$; h=5 mean -0.014	freedom used, no gain

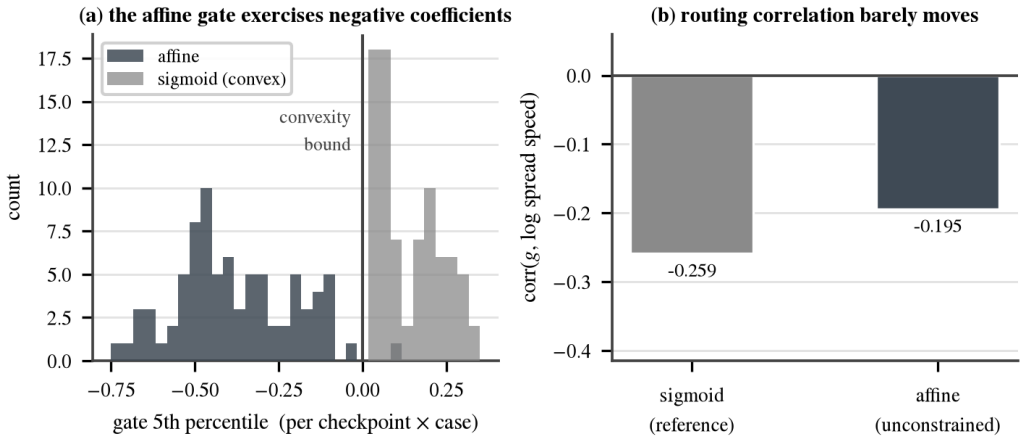


Figure 12: **The affine gate uses its added freedom and still loses.** (a) Gate fifth percentile over candidate nodes, per checkpoint × case: the sigmoid gate is confined to $[0, 1]$ by construction, while the unconstrained gate is negative in $80/81$ distributions. (b) The routing correlation is essentially unchanged. Performance is negative (single-step $-0.026/-0.026/-0.031$), so this arm establishes a boundary rather than a gain.

F ROUTING PROTOCOLS

Under protocol audit: retrospective analyses only. Cross-validation can separate fitted router labels without making its features available at issue time. Full-record growth, duration, positive counts and offline wind summaries use information beyond that time. The records below do not establish a leakage-free deployment protocol, and no repaired issue-time router evaluation is available.

This appendix specifies the four routing protocols referenced in §5 and their exact evaluation scopes. All geo-dataset routing results use the same metric: per-case ROC-AUC at horizon h=5, averaged over the 15 OOD cases with at least 10 positives at h=5 (5 AZ / 6 OR / 4 WA), weighted by the number of h=5 positives (n_{p5}). This scope differs from the pooled tables in the main text and the two are never compared directly.

(a) Oracle upper bound. The oracle selects, per case, the better of GLOW and ConvLSTM using test labels. At h=1 the headroom over the best single family is small (0.7913 vs. ConvLSTM 0.7799, $+0.011$); at h=5 it is substantial (0.7848 vs. ConvLSTM 0.7417, $+0.043$; GLOW alone 0.7236). A single post-hoc speed threshold recovers most of it: routing on mean spread rate with the test-optimal threshold (3.5 cells/step) reaches 0.7767, i.e. $\sim$80% of the oracle headroom.

(b) Retrospective speed routing with held-out threshold fitting. Three checks qualify the post-hoc number. *Threshold sensitivity:* on the h=5 grid, every threshold in $[2.5, 4.0]$ yields a routed score in 0.760–0.777, all above all-ConvLSTM (0.7417) — a plateau 1.5 cells/step wide, not a single tuned point. *Leave-one-out:* selecting each case's threshold on the remaining 14 cases gives 0.7597 ($+0.018$ over ConvLSTM, $+0.036$ over GLOW, $\sim$42% of headroom); the selected threshold is stable (14/15 folds choose 3.5). At h=1 the LOO route (0.7763) is slightly *below* all-ConvLSTM (0.7799): routing only pays at long horizons. *Causal early signal (negative):* replacing the full-trajectory speed with the speed observed in the first 25% of each fire — itself a retrospective definition unless final duration is known — correlates with the full-trajectory value at only 0.51 (log–log), and the LOO-early route (0.7349) falls below all-ConvLSTM. Speed routing is therefore an analysis device, not a deployment protocol. Per-state: the LOO route is exact on AZ (all five cases correctly sent to ConvLSTM, matching the oracle 0.8126) and WA (all four correctly kept on GLOW), but hurts OR ($0.6889 < 0.7292$): fast OR cases are routed to ConvLSTM while GLOW actually wins there, i.e. a single speed threshold does not transfer across states.

(c) Prior-feature router (two-stage, offline). Stage 1 fits a ridge regression from 8 covariate summaries (not all pre-ignition inputs) (fractions of 4 fuel groups; DEM standard deviation and gradient; mean and max wind speed) to log spread rate on the 40 training cases of the mixed-state pool, where speed labels are free. Stage 2 plugs the *predicted* speed into the routing rule and evaluates on the 15 test cases; test performance labels are excluded from the fitted regression, but this does not establish issue-time feature availability or rule out adaptive choices made during retrospective analysis. The regression is weak as a point estimator (LOO $R^2 = 0.03$; ridge shrinkage compresses predictions into 0.4–1.9 cells/step, so the absolute threshold 3.5 routes zero cases) but carries genuine ranking information (test-set rank correlation of log predicted vs. true speed $= 0.73$). Calibrating the threshold to the training-distribution fast-fire quantile ($15\% \Rightarrow 1.75$ cells/step) restores routing: exactly two high-confidence fast cases are routed to ConvLSTM, both correctly, lifting ALL from 0.7236 to 0.7418 ($+0.018$, matching all-ConvLSTM) with no state degraded (AZ $+0.037$, OR $+0.003$, WA unchanged). Coefficient signs are physically consistent (grass/shrub positive, timber negative).

(d) WSTS router battery. On the 419 per-fire WSTS evaluation points (n_{pos}-weighted per-fire h=3 ROC-AUC), we tested six routers over the expert set $\{\text{GLOW}, CL-branch, \text{ConvLSTM}\}$, all with year-based 4-fold cross-validation and features $\{\log \text{speed}, \text{active pixels}, \text{duration}, n_{\text{pos}}, n\}$. The historical analysis designated ridge with confidence fallback as its primary readout (route away from CL-branch only above a threshold selected inside training folds). This designation is not prospective validation: the feature set contains full-record and label-derived information, and the battery belongs to a retrospective experiment sequence.

The historical battery (Table 18) did not beat its best single expert. This does not prove that WSTS has no learnable routing signal or exclude router capacity as a limitation. Oracle headroom is label-

Router	Weighted ROC-AUC	Selection (GLOW/CL-branch/CL)
all-CL-branch (no routing)	0.7956	—
ridge argmax	0.7934	0 / 348 / 71
ridge + confidence fallback (historical primary)	0.7956	0 / 419 / 0
pairwise (CL-branch vs. CL)	0.7956	0 / 419 / 0
gradient-boosted trees	0.7896	23 / 265 / 131
random forest	0.7945	0 / 388 / 31
speed kNN	0.7956	0 / 419 / 0
oracle (CL-branch, CL)	0.8064	—
oracle (three-way)	0.8087	—

Table 18: Historical WSTS router battery, under protocol audit ($n = 419$, positives-weighted per-fire $h = 3$ ROC-AUC). The recorded primary router selects the default expert for every fire. These scores use retrospective features and do not validate deployment, absence of routing signal, or an optimal no-routing policy.

informed; issue-time features and an independent evaluation are needed before interpreting routing utility.

(e) Lookback free-run fidelity (historical, under protocol audit). A second issue-time-candidate signal routes on how well each family reproduces the *observed* steps before the rollout start: over a pool of n=8 geo cases at h=5, the difference in lookback free-run BCE between the two families correlates with their long-range ROC-AUC difference at only $\rho = +0.24$ (weakly positive, not distinguishable from zero; the teacher-forced single-step BCE difference over 27 cases is instead anti-correlated, $\rho = -0.18$), and the resulting router (0.7636; picks GLOW on 2 cases and ConvLSTM on 6) falls below always selecting ConvLSTM (0.7810) and well below the label-informed oracle (0.7986). Neither early-error signal therefore recovers the regime information, which is consistent with the regime being a cross-scale property rather than a per-fire online state.

G Stratified Expert: Data and Training

(a) Fast-regime inventory. Stratifying the 40-case mixed-state training pool at 2.5 cells/step yields 6 fast cases (all CA) and 34 slow cases; the 22 extra-state cases are almost all slow. Lowering the threshold to 1.5 cells/step adds only 4 more. Regime-stratified retraining was therefore not feasible from the existing pool, motivating targeted data expansion.

(b) Proxy-based case selection. Building a case (fire tensor, DEM, HRRR winds, spectral basis, fuel) is the expensive step, so candidates were pre-screened with the proxy $\text{area}_{\text{km}^2}/n_{\text{days}}$, which correlates with measured spread rate at 0.94 on the 40 already-built cases. From 1,252 VIIRS candidates, 16 unbuilt high-speed candidates were selected and built through the standard pipeline; 11 were confirmed fast after measurement, growing the fast stratum from 6 to 17 cases.

(c) Split. The `geo_cross_state_fast` split trains on 12 CA fast-fire cases (6 original + 6 new; same source state as the main split) with validation cases drawn from within the fast stratum. The nominal test case list matches the main-table 27 OOD cases, but this does not certify all training/validation memberships, preprocessing or metric scopes. Per-fire expert scores must not be compared directly to pooled-micro main-table values. The later fast38 training set is not a nested extension of this earlier expert set.

(d) Attribution ladder (single-seed probe). To attribute the gap between CL-branch and the bare ConvLSTM baseline in the fast regime, two arms move CL-branch stepwise toward the baseline: L1 removes the residual identity initialization of the local block; L2 keeps the architecture and aligns the *training protocol* (pure focal BCE; no curriculum, PDE, CRPS, or latent perturbation). Rollout ROC-AUC (seed 42; Table 19):

Arm	AZ h=1/3/5	OR h=1/3/5	WA h=1/3/5	AZ 1-step
CL-branch (baseline)	0.795 / 0.780 / 0.812	0.797 / 0.572 / 0.638	0.797 / 0.728 / 0.707	0.859
L1: no residual init	0.749 / 0.727 / 0.757	0.794 / 0.579 / 0.663	0.759 / 0.757 / 0.657	0.846
L2: protocol-aligned	0.784 / 0.778 / **0.848**	0.792 / 0.552 / 0.660	0.776 / 0.707 / 0.670	**0.867**
bare ConvLSTM (target)	0.780 / 0.806 / 0.870	0.813 / 0.645 / 0.711	0.762 / 0.754 / 0.736	0.862

Table 19: Historical single-seed attribution ladder, under protocol audit. The recorded L2 AZ $h = 5$ change is $0.812 \rightarrow 0.848$ toward 0.870. Multiple training factors change together; this is not a causal identification of curriculum rather than residual structure, nor verification of the later fast38 objective.

(e) Expert failure cases. On the geo test set the fast expert fails precisely on the two fastest cases selected by the router — AZ_2020_39 (0.519) and OR_2017_39 (0.538), consistent across all three seeds — while winning decisively on the comparably fast AZ_2025_24 (0.871, 5.2 cells/step). This historical router–expert mismatch suggests that speed alone does not describe the observed failures; it does not validate deployment of the bare ConvLSTM router. The later fast38 comparison (recorded weighted ROC-AUC 0.706 versus 0.733) is *not pure focal training* and does not use a nested training-set extension. We withdraw "protocol-limited, not data-limited": neither sample size nor the objective has been isolated, and no corrected expert-v2 result is available.

G.1 Stratified expert and routing: full numbers

The verified intervention of §5.4 makes the regime rule causal, but a *deployable* selector is a separate question. The stratified-expert recipe pairs CL-branch with a single-step focal objective on a 12-fire California training split; the later "fast38" enlargement (non-focal, non-nested) remains only a wrong-recipe record (Appendix G).

Geo (positives-weighted per-fire h=5 ROC-AUC, 15 fires). Always-GLOW scores 0.724, always-ConvLSTM 0.742, and a post-hoc per-fire oracle over the two 0.785—headroom a deployable selector must claim. The expert alone scores 0.746, a $+0.005$ over ConvLSTM not separable in the historical case bootstrap (95% CI $[-0.061, +0.063]$), which establishes neither equivalence nor non-inferiority. The historical three-way oracle rose from 0.785 to 0.803 ($\Delta = +0.018$, CI $[+0.003, +0.040]$, p=0.006), a label-informed upper bound conditional on fire selection and old predictions, not validated deployable complementarity.

WSTS (419 fires; historical, provisional). The stratified-expert records lead ConvLSTM on 4/4 year folds in per-fire ROC-AUC (fire-bootstrap $\Delta = +0.011$, p=0.0002) and PR-AUC (p=0.008) under validation-AP selection and observed weather; a fold→fire PR-AUC bootstrap crosses zero, so PR-AUC is not claimed at year level.

Routing is retrospective. The six WSTS routers and the geo prior-feature router use offline information (speed, duration, label-derived counts, wind summaries) unavailable at issue time under year-based or leave-one-out fitting, so the no-routing outcomes and a geo $+0.018$ increment are not deployable gains (Appendix F).

H Ensemble and Uncertainty Analysis

Mechanism. The stochastic ensemble is a single-model latent perturbation: each forward pass adds Gaussian noise $z + \sigma\varepsilon$ (σ=0.1) to the post-GRU latent; K=4 members with a CRPS term during the final curriculum stage, K=10 at evaluation (mean = point prediction, member std = spread).

Accuracy: a null result over the full $\sigma \times K$ grid. Evaluation-only sweep on the 9 main-configuration checkpoints, $\sigma \in \{0.05, 0.1, 0.2\} \times K \in \{1, 4, 10\}$: single-step pooled ROC-AUC/PR agree to four decimal places across all configurations (deterministic K=1 equals K=10 at σ=0.2 within 10^{-4}); all multi-step rollout differences are ≤ 0.002, below seed noise (3-seed std 0.005–0.02). The historical sweep suggests little point-score sensitivity within this grid. It does not fully attribute stability to dynamics or establish that perturbation scale never matters.

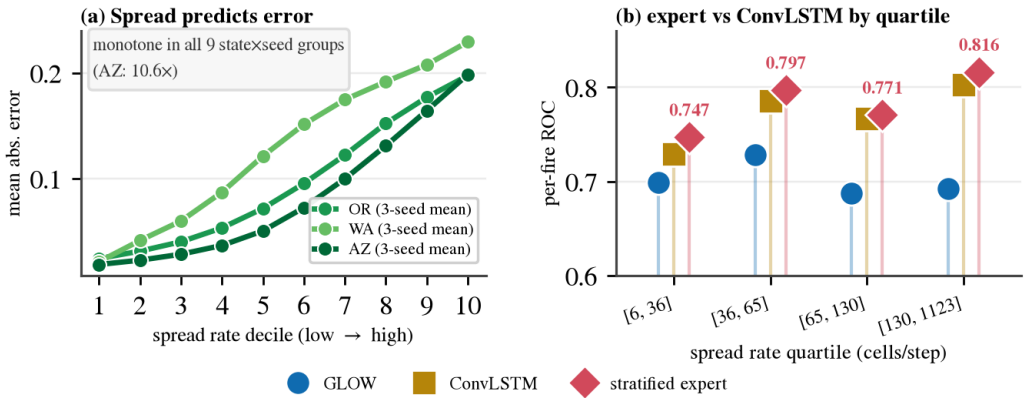


Figure 13: **Historical, provisional panels; under protocol audit. (a) Spread–error association, not probability calibration.** Mean absolute error vs. ensemble-spread decile (3-seed mean per state); curves are monotone in every state×seed group (the AZ annotation is the state-level seed-mean ratio, $10.6\times$). **(b) WSTS exclusive speed quartiles** behind Table 22: per-fire $h{=}3$ ROC-AUC (positives-weighted) after the disjoint $105/105/105/104$ split of 419 fires. The historical stratified-expert means are higher in each quartile; these offline validation-AP-selected results are not repaired external-validity evidence. Any calibration wording embedded in the retained graphic refers only to an error-association plot and is withdrawn.

Uncertainty: historical spread–error association, not calibration. Per candidate node, spread $=$ std over $K{=}10$ member probabilities and error $= |y - \bar{p}|$ (per-state values are 3-seed means; Table 20):

State	corr(spread, error)	AUROC (spread → misprediction)	ROC-AUC(ens. mean) − ROC-AUC(single member)
OR	+0.47	0.71	$< 10^{-4}$
WA	+0.45	0.67	$< 10^{-4}$
AZ	+0.42	0.72	$< 10^{-4}$

Table 20: Historical spread–error association (under protocol audit). Correlation and error-ranking AUROC do not establish calibrated ignition probabilities, calibrated uncertainty magnitude or interval coverage.

Historical spread-decile error curves are strictly monotone in all nine state×seed groups (Figure 13a); in the strongest group (AZ, seed 42) mean error rises $7.3\times$ from the lowest to the highest spread decile ($0.024 \rightarrow 0.176$). This suggests error-ranking utility in the old outputs, not a zero-cost calibrated uncertainty device. Mean spread is small ($\sim 10^{-3}$) and measures sensitivity to latent perturbations; it is not by itself validated epistemic or physical uncertainty.

Probability-calibration metrics (Brier, ECE). All headline claims use ranking metrics (ROC-AUC/PR-AUC) because the operational question is *where* the fire goes next, but proper scores are logged throughout and tell an honest complementary story (rollout, 3-seed means, $\times 10^{-2}$):

	OR		WA		AZ	
Model (Brier / ECE)	$h{=}1$	$h{=}5$	$h{=}1$	$h{=}5$	$h{=}1$	$h{=}5$
GLOW	2.07 / 6.94	2.47 / 6.65	3.05 / 10.5	3.33 / 10.4	3.89 / 3.21	3.79 / 3.96
CL-branch	1.99 / 6.50	2.50 / 6.14	2.10 / 8.53	2.20 / 7.95	3.63 / 2.66	3.37 / 3.02
ConvLSTM	2.21 / 7.80	2.70 / 7.68	2.40 / 10.2	2.44 / 9.53	3.85 / 4.39	3.57 / 5.45
LR	1.33 / 0.41	1.78 / 0.56	1.10 / 1.46	1.33 / 1.31	3.82 / 2.72	3.51 / 2.50

Two readings. First, among the neural models the regime pattern is largely metric-robust: CL-branch has the best ECE in every state–horizon cell and the best Brier score in five of six, with OR at $h{=}5$ the sole exception (GLOW 2.47 vs. CL-branch 2.50, $\times 10^{-2}$), so unlike ranking skill this calibration lead does not exhibit a regime crossover. Second — and reported without embarrassment — logistic

regression is the best probability-calibrated model almost everywhere: focal BCE (shared by all neural models here) deliberately reweights rare positives, buying ranking skill at the cost of absolute probability calibration; the spread–error association above does not fix probability miscalibration. Validation-only Platt or isotonic recalibration would require a separate evaluation before making a calibration claim. Geometry-aware metrics (IoU/Dice of predicted burned area, front-distance and Hausdorff-type errors) require binarizing the probability field at an operating threshold, which introduces a per-model degree of freedom orthogonal to our claims; precision at fixed recall (and its dual under a false-alarm budget) is a monotone functional of the PR curves we already report. We leave a threshold-standardized geometric evaluation to future work.

I WSTS Protocol and Scope Analysis

(a) E1: single-step anchor. Leave-one-year-out over test years 2018–2021 (adjacent validation year), 2 seeds, 30 epochs with validation-PR-AUC early stopping; 16 runs total. GLOW reaches candidate-pooled test PR-AUC 0.359 ± 0.035 (ROC-AUC 0.916 ± 0.017) vs. FNO2d 0.159 ± 0.004 (0.788), with no fold/seed exception — the weakest GLOW run exceeds the strongest FNO2d run. These are historical offline results; the comparison does not isolate gated fusion as the cause or certify external operational validity.

(b) E2: multi-step rollout. Free-running rollout at daily steps, $h \in \{1, 2, 3\}$; only the burn channels are fed back, weather/static inputs stay real; 8 rollout starts per fire. Paired per-fold PR-AUC deltas of GLOW over logistic regression: h=1 $+0.174 \pm 0.015$ (4/4 folds positive), h=2 $+0.021 \pm 0.016$ (4/4 positive), h=3 -0.011 ± 0.005 (4/4 *negative*) — an honest negative: at daily steps, GLOW's lead over a non-spatiotemporal baseline lasts $\sim$2 days.

(c) Unified pooled comparison (five families). All models through the identical fold-pooled pipeline (4-fold means; the CL-branch t2018 arm has one seed, noted in the data files):

Model	h=1 PR-AUC	h=2 PR-AUC	h=3 PR-AUC	h=1 ROC-AUC	h=3 ROC-AUC
persistence	0.047	0.044	0.040	0.500	0.500
logistic regression	0.158	0.117	0.103	0.792	0.757
GLOW	0.332	0.138	0.092	0.910	0.734
ConvLSTM	0.315	0.163	0.152	0.897	0.827
CL-branch	**0.336**	**0.188**	**0.165**	0.904	**0.839**

Table 21: Historical WSTS fold-pooled comparison, under protocol audit. Validation-AP selection and offline weather apply; the stratified expert is not a verified deployed router. Bold marks historical means, not repaired wins. The recorded per-fold $h = 3$ AP difference is positive in 3/4 folds; the one-seed t2018 difference of -0.004 has not been shown to be noise.

(d) Metric-scope reconciliation. Our PR-AUC values are *candidate-pooled* (frontier candidates only; base rate 3–7%) and are not comparable to the full-pixel PR-AUC $\approx$0.46 reported for Res18-UNet variants in the WSTS+ literature. Re-evaluating our models under the full-pixel scope (next-day active label, all-pixel denominator, persistence fill) gives PR-AUC $\approx$ 0.11–0.34 across folds against persistence-only floors of 0.10–0.28, i.e. net gains of only $+0.01$ to $+0.08$: these differences do not decompose the published models' scores into scope and model-skill contributions. We withdraw that attribution and keep comparisons within their stated historical evaluation scopes.

(e) Per-fire evaluation protocol. Per-fire h=3 rollout ROC-AUC/PR-AUC per checkpoint, averaged across seeds, aggregated with n_{pos} weighting. This is the scope for all router and expert significance tests (App. J). The full accounting from the benchmark's 607 fires to the 419 evaluation points is deterministic and label-driven, never model-driven: 99 fires are dropped before any per-fire evaluation by the viability filter shared by all models (fewer than 5 positives or fewer than 50 frontier candidates over the fire's whole record — nothing to predict — or a record too short to fit a rollout start plus h=3 steps), and a further 89 of the remaining 508 by the $n_{\text{pos}} \geq 10$-at-h=3 requirement for a non-degenerate per-fire ROC-AUC, leaving 123/36/140/120 fires for test years 2018/2019/2020/2021. All 508 rollout-viable fires contribute to the fold-pooled numbers

Table 22: Historical WSTS speed quartiles (under protocol audit; 419 fires, positives-weighted per-fire $h = 3$ ROC-AUC). These offline stratified-expert means are provisional; no year-population or deployment superiority is established.

Spread rate (cells/step)	[6, 36]	[36, 65]	[65, 130]	[130, 1123]
GLOW	0.699	0.728	0.687	0.692
ConvLSTM	0.729	0.786	0.767	0.802
Stratified expert	**0.747**	**0.797**	**0.771**	**0.816**

of Table 21; the n_{pos} filter only governs which fires contribute an individual ROC-AUC point to selection-style analyses, exactly mirroring the $n_{p5} \geq 10$ rule on the geo suite (§5.1).

(f) Resolution robustness (negative control). $2\times$ spatially downsampled variants of the 27 geo OOD cases, evaluated with full-resolution checkpoints (no retraining). Mean ΔROC-AUC over 36 state$\times$horizon$\times$seed cells: GLOW -0.030, ConvLSTM -0.017 — both degrade mildly and comparably; the data do not support a "graph methods are resolution-robust" claim.

(g) Temporal-coarsening comparison: historical version and its repair. The 45 geo cases were subsampled using `[::2]` for fire and wind frames, with static grids and spectral bases retained. This changes physical forecast and input-window duration, frame count, wind sampling and label composition; exact doubling of the finite-record mean is not guaranteed. Both scales have since been retrained under a single frozen normalization contract, so unlike the original version the comparison is no longer normalization-mismatched; the values cited in Section 5.4 and Table 23 are the verified retrained readouts (seed-mean h=5 ROC-AUC on 14 cases with at least 5 positives at both scales, paired bootstrap B=10,000). The earlier outputs are retained only as an audit trail. The ConvLSTM and CL-branch arms at 6 h remain faithful historical checkpoints, which is the residual caveat on the family comparison.

The displacement is horizon dependent, and it is concentrated at the deepest horizon. Re-running the same estimator — on the same verified inputs as Section 5.4, whose h=5 row it reproduces exactly — separately at each horizon gives

	h=1 (n=13)	h=3 (n=12)	h=5 (n=14)
CL − GLOW	$+0.007$ (p=0.63)	-0.002 (p=0.97)	$+0.094$ (p=0.002)
CL-branch − GLOW	-0.008 (p=0.70)	$+0.029$ (p=0.23)	$+0.083$ (p=0.001)

Neither contrast is distinguishable from zero at h=1 or h=3; both are resolved at h=5. A static artifact of the finer grid's supervision composition — where a large share of training targets repeat the preceding frame — is a property of the training data and would be present at every horizon, so its absence at the two shallower horizons argues against that reading. The cost is that the intervention is a deep-horizon result: at the shorter horizons of Table 1 the re-indexing does not move the ranking. Per-horizon case counts differ because the $\geq$5-positive filter is applied within each horizon. This table is regenerated from the per-case result files that will be released with the code.

(g′) Historical coarsening extension ($\times$3, $\times$4). The 18-h (`[::3]`) and 24-h (`[::4]`) comparisons retain the same normalization mismatch to the 6-hour reference. The following common-set table is historical and provisional, not a confirmed dose–response ladder:

	$\times$1 (6 h)	$\times$2 (12 h)	$\times$3 (18 h)	$\times$4 (24 h)
mean Δ(CL − GLOW)	$+0.004$	$+0.068$	$+0.093$	$+0.047$
shift vs. 6 h (p)	—	$+0.064$ (.0122)	$+0.089$ (.0066)	$+0.044$ (.330)
mean Δ(CL-branch − GLOW)	-0.036	$+0.058$	$+0.052$	$+0.030$
shift vs. 6 h (p)	—	$+0.094$ (< 0.0002)	$+0.088$ (.0002)	$+0.066$ (.039)
slow stratum, CL − GLOW	-0.039	$+0.080$	$+0.117$	$+0.086$
1-step control (GLOW/CL-branch/CL)	.75/.77/.77	.83/.85/.82	.80/.84/.82	**.74**/.81/.80

(Slow stratum: $v_6 < 2$ cells/step, n=7; the single-step control is absolute ROC-AUC at h=1 on the same 14-case common set as the h=5 intervention readout.)

The historical displacement increases through $\times 3$ and decreases at $\times 4$, but unmatched normalization and bundled sampling changes prevent interpreting any level as a confirmed dose. Lower $\times 4$ single-step performance does not identify reduced training volume as its cause. We withdraw both the dose-confirmation claim and the assertion that a known data-volume boundary explains the attenuation.

(h) Two-day-step coarsening on WSTS (preregistered probe, negative). After the verified geo intervention (Obs. 4 of Section 5.4), we replicated the coarsening direction externally: GLOW and ConvLSTM retrained at two-day steps (`[::2]` frame subsampling; per-step travel exactly doubled; the same leave-one-year-out folds and seed 42; training health verified by best epochs 2–9 and single-step ROC-AUC 0.81–0.89). The preregistered prediction was that ConvLSTM's lead widens. It does not: the ConvLSTM$-$GLOW gap at the deepest horizon shrinks in three of four folds (ROC-AUC displacements $-0.034/-0.005/-0.051/+0.025$; PR-AUC $2/4$, mean per-fire displacement -0.030), and the per-fire displacement is graded by speed in the direction opposite to the preregistered prediction ($\rho = -0.118$, permutation $p = 0.022$, $n = 306$ fires with $\geq$10 positives on both sides): the fastest speed quartile shifts toward GLOW (-0.056; 57.5% of fires overall), while the slowest quartile shifts weakly toward ConvLSTM ($+0.002$). Physical-speed bookkeeping places the slowest WSTS fires (44 cells/day at 375 m $\approx$ 16.5 km/day) near the geo crossover (4–6 cells per 6 h at 1 km $\approx$ 16–24 km/day), and their weak CL-ward shift is directionally consistent with the geo intervention—but the corpus as a whole does not behave as a monotone dose. Beyond the crossover, coarser sampling also removes the fine temporal structure the recurrent baseline exploits, and fixed-h physical duration doubles (the same discretization caveat as the geo intervention).

J Statistical Methodology

Protocol recovery and evidence grades. A two-stage audit repaired the evaluation stack and then re-established the main comparisons. *Stage 1 (re-evaluation):* every existing checkpoint was re-evaluated under a shared normalization/manifest contract with case-keyed RNG isolation; the ConvLSTM family reproduced bit-exactly (510/510 pooled and rollout records), and pooled GLOW metrics shifted by at most 0.006, quantifying the historical 15-vs-18 case normalization and ensemble-RNG exposure as negligible at pooled grade. *Stage 2 (matched retraining):* the GLOW reference, both sides of the coarsening intervention, the K=3 per-case point, the no-local control, and the FiLM fusion variant were retrained with frozen normalization statistics, effective training lists, and validation lists stored inside each checkpoint; the evaluator consumes this contract directly, and the retrained reference reproduces the historical main table within seed noise (≤ 0.011 across states and horizons). Grades: *verified* = retrained under the frozen contract; *faithful* = existing checkpoint, re-evaluated or reproduced exactly; *historical* = pre-audit record retained for traceability. The full audit trail (per-arm old-vs-new comparisons, manifests, and JSON hashes) will be released with the code.

Bootstrap protocols. All significance statements use nonparametric bootstrap with $B = 10{,}000$ resamples and two-sided p-values. On the geo dataset the resampling unit is the *case* (15 OOD cases with $n_{p5} \geq 10$), with n_{p5}-weighted aggregation recomputed per resample; on WSTS the unit is the *fire* ($n = 419$), with n_{pos}-weighted aggregation. Paired comparisons (e.g. the temporal-coarsening displacement) resample matched case-level differences. Where noted, a leave-one-out sensitivity check verifies that no single case flips the sign of the estimate.

Obs. 1 correlation: uncertainty and filter sensitivity. The geo regime correlation of $+0.58$ (log spread rate vs. ΔROC-AUC, n=15) carries a case-bootstrap 95% CI of $[+0.16, +0.82]$ and a two-sided permutation $p = 0.023$ (10,000 shuffles)—wide, as expected at n=15, but sign-stable. Because the $n_{\text{pos}} \geq 10$ filter excludes 6 of 21 evaluable fires and the excluded fires are all slow (0.9–1.6 cells/step), we report the filter sensitivity explicitly: the correlation is $+0.47$ at $n_{\text{pos}} \geq 5$ (n=17), $+0.58$ at ≥ 10 (n=15, the main filter), and $+0.69$ at ≥ 20 (n=9). The historical correlation signs are positive at these thresholds, but this is not a robustness test of the slow-stratum mean advantage. That advantage is filter-sensitive on the unbalanced B pool. Where that pool's middle speed band fills in with mixed results, the same per-fire spread-rate/winner correlation weakens to $+0.25$ from

Dataset	Comparison	Δ	95% CI	p
geo	fast expert vs. all-ConvLSTM (ROC-AUC)	$+0.005$	$[-0.061, +0.063]$	n.s.
geo	three-way oracle vs. oracle(GLOW, ConvLSTM) (ROC-AUC)	$+0.018$	$[+0.003, +0.040]$	0.006
WSTS	CL-branch vs. all-ConvLSTM (ROC-AUC)	$+0.011$	$[+0.005, +0.017]$	0.0002
WSTS	CL-branch vs. all-ConvLSTM (PR-AUC)	$+0.009$	$[+0.002, +0.015]$	0.008
WSTS	three-way oracle vs. oracle(GLOW, ConvLSTM) (ROC-AUC)	$+0.018$	$[+0.014, +0.021]$	$< 10^{-4}$
tc2	displacement of (CL − GLOW) (ROC-AUC)†	$+0.094$	$[+0.034, +0.155]$	0.0022
tc2	displacement of (CL-branch − GLOW) (ROC-AUC)†	$+0.083$	$[+0.033, +0.140]$	0.0010

Table 23: Confidence intervals. †Verified: recomputed under matched frozen-normalization retraining on both scales (the historical, normalization-confounded values were $+0.064$/$+0.094$). Unmarked rows are historical; non-significance does not establish equivalence. WSTS fire-level intervals are conditional on selected fires, folds and offline inputs.

the main-pool $+0.58$, which is why the argument leans on the family contrast and the intervention rather than on the correlation coefficient. Excluded low-positive fires may differ systematically; attenuation cannot simply be assigned to measurement noise. Obs. 3 and 4 are confounded and do not provide independent causal support to resolve this selection issue.

Beyond-ring ignition diagnostic (E4; corrected re-diagnostic). The historical 53-fire, seed-42, eight-start diagnostic was described as counting newly ignited cells. It actually aggregates thresholded events, including repeated occurrences, rather than unique new ignitions. Its beyond-front (bf) fraction also uses an incorrect denominator. Thus the old values 1,743 versus 276 and the 96%/82% fractions cannot support claims of sixfold unique leakage, a fraction of false new ignitions, or a mechanism explaining slow-Washington losses. Agreement of selected ROC outputs does not verify event semantics or denominators; the earlier blanket statement of faithful reproduction is withdrawn. E5 adds a non-faithful diagnostic normalization path and an unmatched training normalization setting (Appendix D.11). A repaired-protocol re-diagnostic (single seed, explicitly reconstructed normalization, exclusive versioned outputs) now supplies the corrected decomposition: of 1,766 GLOW observed-unburned events on the 53-fire pool, 664 (38%) are true rollout-state additions and 290 (16%) are case-unique pixels, versus 215 events and 130 state additions for ConvLSTM—a 5–8$\times$ family ratio that is stable across every counting semantics. The repaired E5 re-diagnostic further shows that the historical E5 ConvLSTM arm in fact evaluated the ring-1 main checkpoint (277 events, matching the 276 of E4), so the historical "53 versus 277" ring-2 comparison is void; corrected ring-2 arms give 63 GLOW events (a 28$\times$ reduction against the main configuration, consistent with the historical 33$\times$ magnitude) versus 4 ConvLSTM events. These corrected counts remain provisional: single seed, reconstructed normalization, and the ring-2 training-side issues (joint train/val dilation and altered model selection) are unresolved.

Exploration, multiple comparisons and prospective status. The router battery and coarsening runs belong to an extended exploratory sequence with repeated test-set examination. Historical scripts designated particular readouts, but that does not make the entire study confirmatory or repair retrospective feature availability and unmatched protocols. We therefore suspend claims of prospective or leakage-free validation. The listed bootstrap and permutation p-values are historical, uncorrected and conditional on the chosen test units; per-cell σ contrasts are descriptive, not significance tests. A repaired analysis must distinguish prospective specifications from retrospective choices and state its multiple-comparison policy before confirming claims.

K STATEMENTS

Reproducibility. Anonymized code, the case-construction pipeline, all evaluation splits, and per-case result files backing every table will be released upon acceptance, subject to source-data licenses. A review-time anonymous archive is provided as supplementary material: code for all model variants, the case-construction pipeline, all evaluation splits, and the result files from which Table 1 and every data figure are regenerated without data or GPUs; derived tensors are redistributed only where licenses permit, otherwise we release scripts, manifests, hashes and split IDs. Concretely, the intended release comprises: (i) the 45-case cross-state construction scripts and, where permitted,

Table 24: **The pooled metric is mostly a cross-fire measurement.** Decomposition of the $h{=}5$ pooled-micro ROC-AUC into a *within-fire* term (candidate-pair-weighted mean of per-fire ROC-AUC, each fire weighted by its positive–negative pair count—how well a single fire is ranked internally) and a *cross-fire* term (mean concordance over positive–negative pairs drawn from *different* fires—whether scores are comparable across fires), using the exact identity $\text{pooled} = f_{\text{w}}\,\text{within} + (1 - f_{\text{w}})\,\text{cross}$ (Mann–Whitney, ties at 0.5). f_{w} is the fraction of positive–negative pairs that lie within one fire: 0.095–0.152 for the three states shown (4% for the pool as a whole), so the pooled column mostly measures cross-fire comparability rather than ranking skill. Same checkpoints and pool as Table 1; each seed decomposed separately, then averaged. For GLOW—though not for ConvLSTM—WA's within-fire ROC-AUC is the highest of the three states while its cross-fire term is the lowest, which is what the pooled column reports.

		GLOW			ConvLSTM		
State	f_{w}	pooled	within	cross	pooled	within	cross
WA (slow)	0.095	0.709	**0.801**	0.700	0.715	0.744	0.712
OR (mid)	0.109	**0.762**	0.724	**0.767**	0.700	0.663	0.704
AZ (fast)	0.152	0.714	0.696	0.717	**0.851**	**0.817**	**0.857**

rasterized tensors with per-case metadata; (ii) the exact train/val/test split files for every split used in the paper; (iii) training and evaluation code for all model variants, keyed by the command-line flags named in this appendix; and (iv) the versioned per-case JSON result files from which every number in the paper is computed, together with the analysis scripts that reduce them to each table and figure. WildfireSpreadTS is already public. Historical numbers remain traceable but their training-side validity is not certified; the evidence grades, and the repaired-protocol re-evaluations and matched retraining behind the main comparisons, are defined in Appendix J.

Compute. All training runs fit on a single V100/A100 GPU; individual runs take 12–90 minutes (0.15M–0.96M parameter models). The total compute behind the paper, including ablations, baselines, and the retrained intervention arms, is on the order of 100 GPU-hours.

LLM usage. See the *LLM Usage Disclosure* preceding this appendix, which is the authoritative statement of what was and was not done with language models. No replacement data were generated or imputed for this revision, and the authors take responsibility for resolving the outstanding protocol issues before asserting repaired scientific conclusions.

Data and licenses. All data sources are public: VIIRS active-fire detections (NASA FIRMS), HRRR atmospheric analyses (NOAA), LANDFIRE fuel models (FBFM40; USGS/USFS), and terrain elevation. WildfireSpreadTS is a published benchmark used under its release terms. No personally identifiable information is involved.