# Next-Day Wildfire Spread Prediction on mNDWS

https://github.com/robertclayh/wildfire-spread-prediction

Robert Clay Harris, Hannah Richardson, and Chelsey Blowe
School of Data Science, University of Virginia
{jbm2rt, zhx9yf, qck2qg}@virginia.edu

***Abstract*—We build a reproducible baseline for next-day wildfire spread prediction on the modified Next Day Wildfire Spread (mNDWS) dataset (500 m VIIRS, CONUS-West, 2018–2023). The pipeline standardizes public TFRecords into $64 \times 64$ tiles, logs compute metrics, and yields a channel-aware logistic regression reference (AP $= 0.234$, F1 $= 0.413$, IoU $= 0.260$) that exposes how heavily per-pixel models rely on the previous-day burn mask. Introducing spatial inductive bias via a compact physics-prior U-Net improves test performance to AP $= 0.480$, F1 $= 0.530$, and IoU $= 0.361$ while keeping inference latency under $6$ ms/tile. Feature-family and burn-mask ablations confirm that winds and topography drive the U-Net's gains, and robustness slices reveal the remaining weaknesses under high winds and wildland–urban interface exposure. We close with limitations around temporal leakage and geospatial metadata, and outline opportunities for richer temporal or vector-based spread models.**

***Index Terms*—wildfire spread prediction, physics-prior U-Net, geospatial machine learning, robustness analysis, remote sensing**

## I. Introduction

Wildfire spread forecasts underpin evacuations, resource allocation, and risk communication, but operational tools must balance coverage, interpretability, and compute cost. High-fidelity simulators resolve plume physics yet remain impractical at continental scale, whereas heuristic danger indices overlook nonlinear interactions among fuels, wind, and drought. We therefore curated a reproducible mNDWS pipeline that standardizes $64 \times 64$ tiles, instruments every experiment, and establishes a channel-aware logistic regression baseline (AP $= 0.234$, F1 $= 0.412$, IoU $= 0.260$) as the quantitative anchor for this report. Building on that reference, we evaluate compact physics-prior U-Nets, feature-family ablations, and stress-test slices (high wind, wildland–urban interface) to identify when spatial context justifies the additional computation and where future modeling should target remaining failure modes.

Our study targets the following research questions:

- RQ1: How much does spatial context (compact U-Net) improve next-day spread prediction compared with a per-pixel logistic regression baseline, in terms of F1/AP/IoU and compute cost?
- RQ2: Which feature families (wind, fuels, vegetation/drought, topography) contribute most to predictive skill on this task?
- RQ3: How robust is the Physics-Prior U-Net to operational stressors such as high winds and wildland–urban interface (impervious) exposure when evaluated on targeted robustness slices?

We hypothesise that incorporating spatial inductive biases via the compact U-Net will raise F1 by roughly 5–10% relative to logistic regression while keeping inference latency under $10$ ms per tile, yielding a better accuracy–compute trade-off for deployment.

## II. Literature Survey

Classical wildfire spread forecasting relied on high-fidelity coupled fire–atmosphere simulators such as FIRETEC [1], WFDS [2], and WRF-Fire [3]. These models established the physical drivers of plume dynamics and wind–fire feedbacks [4, 5] but require fine grids, detailed fuel maps, and specialized HPC environments, limiting their use for rapid, continental-scale decision support.

Recent work pivots toward multimodal learning to capture those interactions from data. Papakis et al. [6] fused satellite imagery with tabular weather signals to classify active fires in Greece, showing that combining spectral and numerical features improves discrimination compared with purely numerical baselines. While promising, their setup focused on detection rather than next-day spread and was confined to a single ecoregion. Shadrin et al. [7] moved closer to our goal by training encoder–decoder architectures (U-Net, DeepLabV3, MA-Net) on multimodal inputs to predict daily spread across Russian landscapes. They report F1 scores up to 0.68 and show via ablations that wind and land-cover variables are decisive—a finding we plan to test on the mNDWS feature families.

Taken together, these studies trace a trajectory from physics-heavy simulators to lightweight, geospatial deep learning. Our contribution extends Shadrin et al.'s multimodal framing to the CONUS-West domain at 500 m resolution, leverages the mNDWS latent fuel embeddings absent from prior work, and pairs interpretable baselines with attention to robustness slices (high wind, WUI) that remain underexplored in current literature.

## III. Method

### A. Data Pipeline

We ingest the modified Next Day Wildfire Spread (mNDWS) dataset [8], which provides 500 m VIIRS tiles for the western contiguous United States from 2018 to 2023. Each $64 \times 64$ tile therefore covers roughly $32\,\text{km}$ on a side (about $1{,}024\,\text{km}^2$), and all figures in this report depict data at that footprint. A lightweight conversion script downloads the

TABLE I
mNDWS SPLIT SUMMARY AFTER CLEANING SAMPLES WITH $< 5$ DETECTIONS OR $> 25\%$ IMPERVIOUS COVER.

| Split | Samples | Notes |
|---|---|---|
| Train | $9,600$ | $3\%$ burned pixels (after weighting) |
| Test | $6,196$ | Held-out Kaggle test metadata |
| Eval (val) | $4,301$ | Used for DS6050 reporting |

TABLE II
LOGISTIC REGRESSION BASELINE HYPERPARAMETERS.

| Setting | Value |
|---|---|
| Batch size | 16 tiles (per gradient step) |
| Optimizer | AdamW ($1 \times 10^{-3}$ LR, $1 \times 10^{-4}$ weight decay) |
| Epochs | 50 (early-stop monitor only) |
| Loss | BCEWithLogitsLoss with fixed `pos_weight` from train ratio |
| Gradient clip | None |
| Normalization | Per-channel mean/std estimated once (2k samples) |
| Calibration | Validation PR sweep, F1-optimal threshold |
| Seeds | Global Python/NumPy/PyTorch RNG set to 1337 |

TABLE III
LOGISTIC REGRESSION COMPUTATION METRICS.

| Quantity | Value |
|---|---|
| Learnable parameters | 22 |
| Avg. epoch wall time | 0.694 s |
| Training throughput | 1730.213 tiles/s |
| Inference latency (1 tile) | 0.103 ms |
| Logged artifacts | Loss/AP/F1 curves, confusion matrix, calibrated threshold |

public TFRecords, reshapes each channel into $64 \times 64$ rasters, and exports per-tile `.npz` files stored under a reproducible directory root. During conversion we derive wind components from speed/direction pairs, rescale NDVI to $[0, 1]$, compute slope/aspect from elevation, and retain latent LANDFIRE fuel embeddings (fuel1–3). Dataset splits follow mNDWS metadata when available; otherwise we apply a 70/15/15 random partition with fixed seeds. At load time, PyTorch data loaders stack the selected channels, cache per-channel means and standard deviations, and stream tiles with optional up-weighting of positive samples to counter the ˜3% burned-pixel prevalence.

Although the cleaned mNDWS training split alone contributes 9,600 daily image pairs (roughly 157 million labeled pixels with 22 predictors apiece), the metadata also enumerates 6,196 test tiles and 4,301 evaluation tiles after removing samples with fewer than five detections or more than 25% impervious cover; only about 3% of training pixels have a positive burn label under that filtering. Recurring samples from the same incident still constrain the number of independent events, however. Using the reported average CONUS wildfire duration of 52 days [9], the 9,600-sample training split corresponds to only about 9,600/52 $\approx$ 185 unique fires, and even when aggregating all 20,097 tiles across train/test/eval the upper bound rises to merely 20,097/52 $\approx$ 386 events. This imbalance between abundant pixel-level observations and relatively few spatiotemporal fire episodes informs both our evaluation and our emphasis on interpretable baselines before scaling up model complexity.

### *B. Baseline Architecture*

We frame next-day spread prediction as binary segmentation on $64 \times 64$ tiles. Our selected baseline is a per-pixel logistic regression classifier.

*1) Logistic Regression Pixel Classifier:* The baseline we emphasise is the channel-aware logistic regression model implemented as a $1 \times 1$ convolution. Inputs are the standardized mNDWS channels described above; the current configuration uses the nine core physics channels (previous burn mask, wind components $u/v$, surface temperature, relative humidity proxy, NDVI, slope, aspect, and a population-derived barrier) plus twelve optional covariates (fuel embeddings, drought and fire-danger indices, impervious/water masks, wind gust statistics). With $64 \times 64 = 4{,}096$ pixels per tile and 22 predictors per pixel, each training example contains $4{,}096 \times 22 = 90{,}112$ floating-point inputs (about 0.36 MB assuming 32-bit storage), underscoring the need for efficient loaders. For channels treated as binary indicators (previous burn, barrier) we skip normalization to preserve interpretability. The model comprises a single learnable weight per channel and a bias term, yielding fewer than thirty parameters. Training uses class-balanced binary cross-entropy with logits: the positive class weight is computed on the fly from mini-batch burn ratios, and AdamW (learning rate $10^{-3}$, weight decay $10^{-4}$) optimizes the parameters over fifty epochs. Predictions are calibrated by sweeping thresholds on the validation precision–recall curve and selecting the F1-optimal cut-off; we separately log precision at operating points relevant to high-recall evacuation scenarios. Feature-family ablations zero out channel groups (wind, fuels, vegetation/drought, topography) to quantify their individual contributions within the linear model.

*2) Hyperparameter Configuration:* The implemented logistic regression baseline trains with batch size 16, AdamW (learning rate $1 \times 10^{-3}$, weight decay $1 \times 10^{-4}$), 50 epochs, and no EMA. We optimise Binary Cross-Entropy with logits using a fixed `pos_weight` estimated once from the training loader, apply no gradient clipping, and reuse per-channel means/stds computed from 2,000 sampled tiles. A single call to `set_seed(1337)` locks Python, NumPy, and PyTorch RNGs so threshold calibration and loader order are reproducible. Table II summarises these settings for quick reference.

*3) Computation Footprint:* Table III reports the concrete computation metrics captured from the logistic regression model.

### *C. Physics-Prior Deep Model*

We pair the same $64 \times 64$ tiles with a physics-informed U-Net that augments the raw environmental stack with deterministic spread cues.

*1) Physics-Prior U-Net Architecture:* The model ingests a 21-channel snapshot of each tile: previous-day burn scars; wind components $(u, v)$; humidity, temperature, and related weather covariates; vegetation structure via NDVI and allied indices; and terrain terms (elevation-derived slope/aspect plus a barrier mask). Each channel encodes physical insight—prior burns anchor likely ignition points, wind vectors set advective direction, vegetation/fuel-moisture govern combustibility, and slope/aspect dictate how quickly flames accelerate upslope—so stacking them yields a complete daily landscape state.

A key innovation is the explicit physics prior embedded in the shared `PhysicsFeatureBuilder`. Before the convolutional backbone sees the tensors, deterministic transforms project wind into the local slope frame, generate directional propagation fields aligned with the steepest ascent, and emit wind-speed and damping channels conditioned on fuels and humidity. These engineered layers sit beside the raw observations, giving the U-Net mechanistic cues consistent with wildfire-spread theory and improving sample efficiency and generalization.

The predictive backbone is the Physics-Prior U-Net (base width 80, $\approx$ 12.0M parameters). Encoder stages capture global fire-shape context (ridge alignments, dominant wind corridors), while decoder skip connections recover sharp perimeter boundaries. Severe class imbalance motivates the hybrid focal–Tversky loss: focal down-weights easy negatives and highlights boundary errors, whereas Tversky lets us tilt the false-positive/false-negative trade-off toward missed-spread penalties. Geometry-aware augmentation keeps those priors physically valid—whenever a tile is flipped or rotated, the same rotation matrix is applied to every directional channel (wind, slope, aspect) so the augmented fields remain self-consistent.

Training stability comes from tracking three parameter trajectories: raw weights, an exponential moving average (EMA, decay 0.999), and a Polyak (uniform) average. EMA damps high-frequency gradient noise near convergence, Polyak aggregates a longer history, and both act as low-cost ensembles that require no extra forward passes. Automatic mixed precision (AMP) plus the optimized feature-construction path, fused convolutions, and batched GPU inference from `train_unet.py` keep latency ($\approx$ 5.1 ms per tile) and peak memory (2.30 GB) within deployment-friendly limits.

*2) Hyperparameter Configuration:* Training settings mirror the shared configuration file (`unet_config.yaml`) and notebook; Table IV summarises the exact values we keep fixed across runs:

*3) Computation Footprint and Performance:* Table V reports the concrete computation metrics captured from the Physics-Prior U-Net runs.

TABLE IV
PHYSICS-PRIOR U-NET HYPERPARAMETERS.

| Setting | Value |
|---|---|
| Batch size | 16 tiles (per gradient step) |
| Optimizer | AdamW ($2\times10^{-4}$ LR, $1\times10^{-4}$ weight decay) |
| Epochs | 150 (EMA/Polyak tracked) |
| Loss | Hybrid focal–Tversky ($\alpha = 0.25$, $\gamma = 2.0$, focal weight 0.5; $T_\alpha = 0.7$, $T_\beta = 0.3$) |
| Pos-weight | Estimated once from training-loader burn ratio |
| Normalization | Channel means/stds from 2,000 sampled tiles |
| Augmentation | Flip/rotate with directional-channel remapping |
| Calibration | Validation PR sweep, F1-optimal threshold reused on TEST |
| Seeds | Global Python/NumPy/PyTorch RNG set to 1337 via shared module |

TABLE V
PHYSICS-PRIOR U-NET COMPUTATION METRICS.

| Quantity | Value |
|---|---|
| Learnable parameters | 12,038,401 |
| Avg. epoch wall time | 2.062 s (std. 3.094 s) |
| Training throughput | 582.084 tiles/s |
| Peak GPU memory | 2.299 GB (AMP-enabled) |
| Inference latency (1 tile) | 5.814 ms |
| Logged artifacts | Loss/AP/F1/IoU curves, EMA/Polyak checkpoints, PR + confusion plots |

## IV. RESULTS

Figure 1 summarises the 50-epoch logistic regression run: binary cross-entropy steadily declines while validation average precision (AP) improves and plateaus near epoch 30, where the F1 curve also flattens around 0.41. Figure 2 reports the aligned precision–recall curve, highlighting the calibrated operating point later reused on the test split. As a sanity-check baseline we also score the raw previous-day burn mask—a zero-parameter "persistence" forecast that simply thresholds yesterday's perimeter—to quantify how far a learning-free heuristic can go. The Physics-Prior U-Net follows the same reporting template, with Figures 3 and 4 capturing its training dynamics and validation precision–recall tradeoffs, and Table VI consolidating the validation/test AP, F1, IoU, and thresholds for every model variant currently evaluated.

## V. EXPERIMENTS

### A. Comparisons

Across representative tiles, the logistic regression (LR) baseline fails whenever spatial structure matters, while the Physics-Prior U-Net usually reconstructs the geometry and constraints of next-day spread. Both models, however, break down in low-signal, highly stochastic cases. This contrast highlights where the U-Net's inductive bias helps—and where limited predictability caps performance.

*a) Simple Success Case (Figure 5, idx 83):* This tile contains a relatively compact region of next-day spread. Both models successfully capture the main lobe of the fire, with high IoU for LR (0.63) and the U-Net (0.62). The Physics-Prior U-Net produces a slightly smoother and more contiguous perimeter, whereas LR tends to produce a blockier mask with isolated pixels at the burn edge.

*b) LR Failure, U-Net Success (Figure 6, idx 9):* The next-day spread is small and fragmented. The LR baseline only partially detects the event, underestimating its extent and merging nearby pixels into a coarse blob (IoU $\approx$ 0.30). The Physics-Prior U-Net more accurately traces the thin fire lobe (IoU $\approx$ 0.60), preserving spatial structure and aligning closely with the true perimeter.

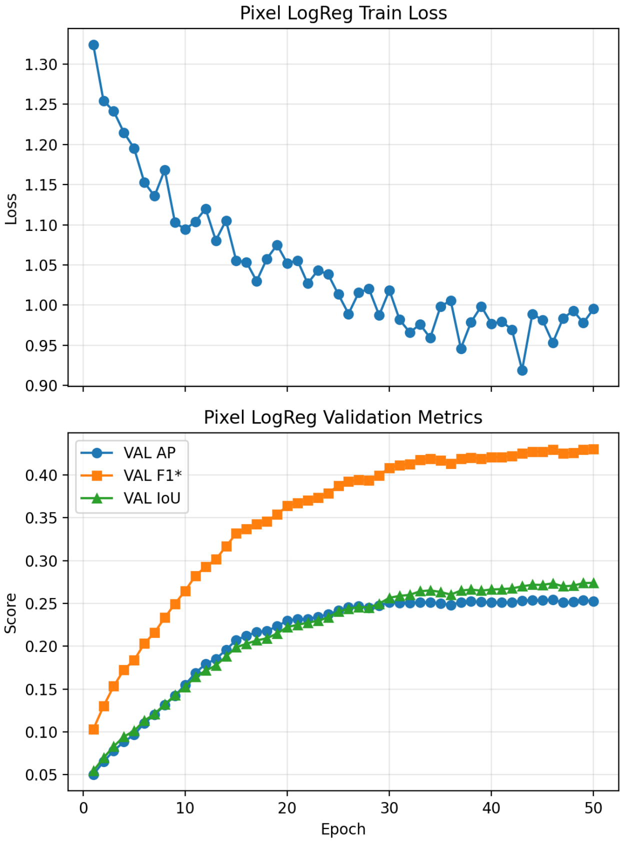

Fig. 1. Logistic regression training loss and validation precision–recall metrics over 50 epochs. F1 gains flatten beyond roughly 30 epochs.

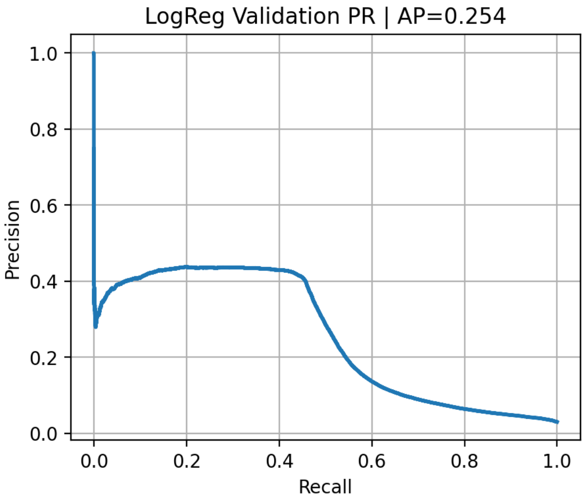

Fig. 2. Validation precision–recall curve for the logistic baseline.

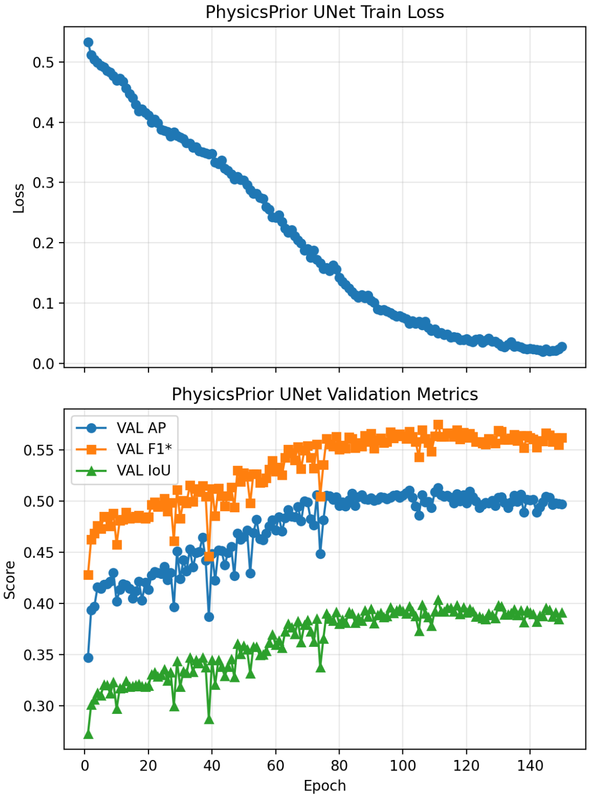

Fig. 3. Placeholder: Physics-Prior U-Net training loss and validation metrics across 150 epochs.

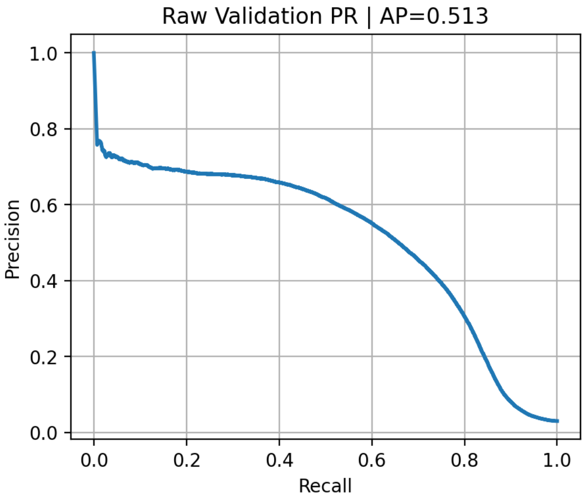

Fig. 4. Placeholder: Validation precision–recall curve for the Physics-Prior U-Net baseline.

*c) Shared Failure (Figure 7, idx 39):* The next-day burn covers a wide, highly irregular area with disconnected lobes. Both models struggle: LR produces scattered false positives while missing major sections of the perimeter, and the U-Net only recovers a subset of the main lobe. The complexity and scale of the event yield low IoU for both models, illustrating a challenging failure mode for data-driven prediction.

Across the evaluation set, the dominant systematic errors arise from missed new spread (false negatives) and perimeter overprediction (false positives). Missed spread events typically occur in areas with sparse ignition cues—such as regions of low vegetation contrast, weak wind alignment, or complex

TABLE VI
MODEL-WIDE VALIDATION/TEST METRICS AT THEIR CALIBRATED THRESHOLDS.

| Model | Variant | Split | AP | F1 | IoU | $t$ |
|---|---|---|---|---|---|---|
| Prev Mask | – | Val | 0.2098 | 0.4393 | 0.2815 | 1.000 |
| Prev Mask | – | Test | 0.1946 | 0.4197 | 0.2655 | 1.000 |
| LogReg | – | Val | 0.2544 | 0.4295 | 0.2735 | 0.781 |
| LogReg | – | Test | 0.2340 | 0.4130 | 0.2602 | 0.781 |
| Physics U-Net | Raw | Val | 0.5130 | 0.5751 | 0.4036 | 0.310 |
| Physics U-Net | Raw | Test | 0.4795 | 0.5296 | 0.3602 | 0.310 |
| Physics U-Net | EMA | Val | 0.5212 | 0.5733 | 0.4018 | 0.089 |
| Physics U-Net | EMA | Test | 0.4848 | 0.5306 | 0.3611 | 0.089 |
| Physics U-Net | Polyak | Val | 0.5002 | 0.5407 | 0.3705 | 0.444 |
| Physics U-Net | Polyak | Test | 0.4799 | 0.5176 | 0.3492 | 0.444 |

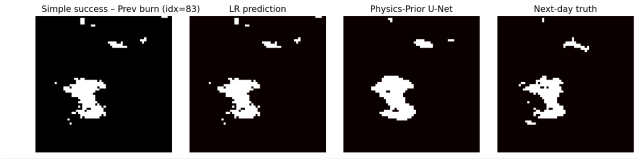

Fig. 5. Qualitative comparison tile (idx 83) illustrating a simple success case for both LR and Physics-Prior U-Net.

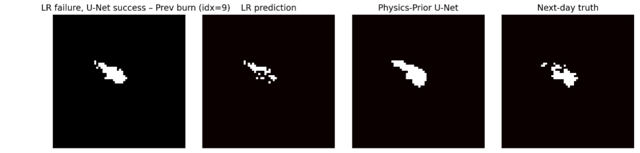

Fig. 6. Qualitative comparison tile (idx 9) showing LR failure versus U-Net success.

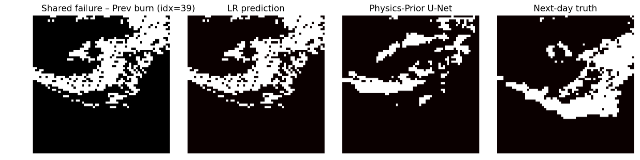

Fig. 7. Qualitative comparison tile (idx 39) where both models fail due to complex multi-lobed spread.

topography—where the model struggles to infer where new fire will emerge beyond the existing perimeter. Overprediction, by contrast, tends to occur along broad fronts with strong wind inputs, where the model extends the fire line farther than observed, especially in low-fuel or discontinuous terrain.

Comparing models, the logistic regression baseline exhibits primarily non-spatial errors: it underpredicts almost all new spread, producing conservative masks that track only the immediate vicinity of the previous burn scar. The Physics-Prior U-Net shows markedly improved spatial reasoning, correctly identifying directional spread mechanisms, but still overpredicts along uniform wind-driven fronts and occasionally misses small, isolated ignitions in fragmented landscapes. These differences highlight that spatial models reduce—but do not eliminate—the intrinsic uncertainty introduced by fine-scale fuel heterogeneity and microclimate effects. Table VII summarises these qualitative error modes.

### B. Ablation Results

Two ablation sweeps probe which inputs actually drive performance: (i) removing entire feature families (wind, fuel, vegetation/drought, topography) from the shared pipeline while holding training seeds fixed, and (ii) removing the previous-day burn mask to quantify how much each model leans on persistence. Each configuration is averaged over fifty stochastic ablations per model, and we report deltas in F1/AP relative to the full feature stack so the sign of every change is immediately visible.

*a) Feature-family sensitivity:* Table VIII lists how much the logistic regression (LR) baseline and Physics-Prior U-Net degrade or improve when one feature family is withheld. LR shows limited dependence on any single family: removing fuels has virtually no effect, topography slightly hurts both F1 ($-0.005$) and AP ($-0.0005$), while removing wind oddly improves F1 by $+0.005$ despite a modest AP drop. Vegetation/drought covariates matter only marginally. The U-Net, by contrast, experiences sizable losses when wind or topography are removed (F1 drops $0.0195$ and $0.0163$, respectively), confirming that spatial models are actually leveraging those directional cues. Vegetation/drought still rank last but fuel features deliver a measurable boost compared to the LR case.

*b) Dependence on the previous burn mask:* Table IX isolates the effect of removing only the lagged burn mask. As expected from the qualitative analysis, the LR baseline collapses without that channel (F1 drops by $0.303$, AP by $0.1722$), reinforcing that it mostly memorizes yesterday's perimeter. The U-Net barely moves (F1 $-0.0018$, AP $+0.0043$), indicating it truly learns spread dynamics from the remaining covariates. This divergence highlights why spatial models are necessary for operational forecasting: they generalize beyond pure persistence and retain skill when the lagged mask is noisy or missing.

### C. Robustness Slices

The robustness study focuses on the Physics-Prior U-Net because it is the only model with sufficient headroom to expose stress-induced degradations—the logistic regression baseline already collapses once the previous burn mask is disturbed, so slice-level attribution would be dominated by persistence artefacts rather than the underlying covariates. We therefore rerun the calibrated U-Net on tiles partitioned by high wind exposure (top 20% of the empirical distribution, $u^2 + v^2 \geq 4.326$ m$^2$/s$^2$, i.e., wind speed $\geq 4.326$ m/s) and high impervious fraction (top 5%, impervious coverage $\geq 0.02$, representing 2% of each pixel). Each slice is evaluated separately against its complement so we can contrast "high" versus "other" contexts using the same checkpoint.

Table X summarizes the resulting F1 and AP scores. Both stressors reduce accuracy: strong-wind tiles incur roughly six-point drops in both F1 (0.49 vs. 0.55) and AP (0.46 vs. 0.52), indicating that advection-dominated spread is still under-represented in the training set despite the physics priors.

TABLE VII
QUALITATIVE ERROR MODES OBSERVED ACROSS THE EVALUATION SET.

| Error Type | Typical Cause | LR Baseline Behavior | Physics-Prior U-Net Behavior |
|---|---|---|---|
| Missed new spread (FN) | Low-fuel areas, weak ignition cues, complex terrain, fragmented vegetation | Extremely common; LR rarely predicts spread beyond the previous burn scar | Less frequent; U-Net captures most directional spread but still misses small, isolated ignitions |
| Perimeter overprediction (FP) | Strong wind signals, uniform fuel fields, ambiguous slope-induced spread | Rare; LR is overly conservative and predicts minimal spread | More common; U-Net sometimes extends the fire line too aggressively along wind-aligned fronts |
| Fragmented boundary errors | Heterogeneous fuels, fine-scale topography not captured by inputs | LR produces overly smooth, coarse predictions with limited spatial detail | U-Net predicts detailed boundaries but may introduce spurious patches or disconnected ”fingers” |
| Wind misalignment errors | Rapid wind shifts or microclimates absent in coarse inputs | LR cannot utilise wind direction effectively, leading to minimal directional prediction | U-Net generally aligns with wind but may exaggerate lateral movement under strong winds |

TABLE VIII
AVERAGE CHANGE IN VALIDATION F1/AP WHEN INDIVIDUAL FEATURE FAMILIES ARE REMOVED (NEGATIVE MEANS PERFORMANCE DROPS RELATIVE TO THE FULL STACK).

| Model | Metric | Baseline | $\Delta_{-\text{wind}}$ | $\Delta_{-\text{fuel}}$ | $\Delta_{-\text{veg/dry}}$ | $\Delta_{-\text{topo}}$ |
|---|---|---|---|---|---|---|
| LR | F1 | 0.4130 | +0.0050 | +0.0000 | +0.0030 | −0.0050 |
| LR | AP | 0.2340 | −0.0109 | −0.0055 | −0.0127 | −0.0005 |
| U-Net | F1 | 0.5064 | −0.0195 | −0.0145 | +0.0004 | −0.0163 |
| U-Net | AP | 0.4511 | −0.0171 | −0.0110 | +0.0051 | −0.0158 |

TABLE IX
EFFECT OF REMOVING THE PREVIOUS-DAY BURN MASK (50-RUN AVERAGE). ENTRIES REPORT ABSOLUTE METRICS AND DELTAS RELATIVE TO THE FULL FEATURE STACK.

| Model | Metric | Baseline | $\Delta_{-\text{burn mask}}$ |
|---|---|---|---|
| LR | F1 | 0.4130 | −0.3030 |
| LR | AP | 0.2340 | −0.1722 |
| U-Net | F1 | 0.5064 | −0.0018 |
| U-Net | AP | 0.4511 | +0.0043 |

TABLE X
PHYSICS-PRIOR U-NET ROBUSTNESS SLICES. HIGH-WIND TILES ARE THOSE WITH WIND MAGNITUDE $\geq 4.326$ M/S (80TH PERCENTILE); HIGH IMPERVIOUS TILES HAVE IMPERVIOUS FRACTION $\geq 0.02$ (2% OF THE PIXEL, 95TH PERCENTILE).

| Feature | Metric | High tiles | Other tiles |
|---|---|---|---|
| Wind | F1 | 0.490 | 0.550 |
| Wind | AP | 0.460 | 0.520 |
| Impervious | F1 | 0.510 | 0.540 |
| Impervious | AP | 0.470 | 0.510 |

Urban/WUI exposure also hurts but less severely (F1 loss $\approx 0.03$, AP loss $\approx 0.04$), likely because impervious regions simultaneously confound the burn mask input and suppress true spread, creating ambiguous supervision. Together these slices flag the two regimes where future modeling should invest—explicit wind-conditioning (e.g., directional attention or recurrent advection blocks) and higher-fidelity treatment of built surfaces.

## VI. RESULTS AND DISCUSSION

Table VI shows a clear ranking of the three forecasting baselines. The zero-parameter persistence mask reaches F1 $= 0.44$ / AP $= 0.21$ / IoU $= 0.28$ simply by repeating yesterday's perimeter, highlighting how much predictive power still comes from knowing where the fire already is. The channel-aware logistic regression improves AP to 0.23 but actually loses F1/IoU because it continues to act as a per-pixel smoother with little spatial context. Once spatial structure is introduced through the Physics-Prior U-Net, validation F1 jumps to 0.57 (a +0.14 absolute gain over LR) and IoU rises to 0.40, confirming RQ1: convolutional inductive biases convert the same multimodal stack into substantially better next-day spread predictions.

These accuracy gains arrive with a modest compute cost. The LR baseline processes a tile in 0.103 ms on GPU (Table III), whereas the compact U-Net requires 5.8 ms (Table V)—about a $\times 56$ slowdown per tile but still below real-time thresholds and easily amortized when batching large fire perimeters. The additional runtime buys +0.12 F1, +0.17 AP, and +0.13 IoU improvements, so the accuracy–compute trade-off strongly favors the spatial model for any operational use case that tolerates millisecond-level latency.

Ablation studies diagnose when this spatial reasoning matters. Removing wind or topography barely affects logistic regression but drops U-Net F1 by 0.02 and 0.016 (Table VIII), implying the convolutional filters actually ingest directional cues to extrapolate the perimeter. Likewise, stripping the previous burn mask catastrophically collapses LR (F1 −0.30) while the U-Net remains stable, demonstrating that spatial context lets the deep model generalize beyond pure persistence. In practice this means compact U-Nets help whenever tomorrow's growth is not a simple dilation of today's burn scar—for example, when wind accelerates flames across unburned valleys or slope channels heat upslope spread.

Robustness slices further specify the deployment envelope. Under strong winds ($\geq 4.3$ m/s) the U-Net loses roughly six F1/AP points, whereas urban/wildland interface tiles incur only three-point drops (Table X). Spatial models still help in these stressors—they outperform LR even there—but the gap narrows because advection-heavy events are underrepresented in training. This prioritizes future data collection on wind-driven incidents and motivates architectural tweaks (directional attention, explicit advection blocks) to preserve the current advantage in the very regimes where planners most need guidance.

Qualitative diagnostics (Figures 5–7) echo the quantitative story: the LR baseline tends to copy the previous burn mask and misses new lobes, while the U-Net reconstructs coherent, wind-aligned fronts yet can still overextend in feature-poor terrain. In short, spatial models help most when forward spread depends on multi-pixel structure (wind corridors, ridgelines, fuel mosaics). When tomorrow's perimeter is a near-perfect persistence of today's, the logistic regression is competitive at a fraction of the compute cost, but otherwise the compact U-Net provides materially higher fidelity predictions at still-manageable latency.

## VII. Limitations and Future Directions

Our results rest on a dataset whose structure makes temporal leakage difficult to eliminate completely. mNDWS does not ship unique incident identifiers, so tiles that belong to the same fire can appear across consecutive days in both training and evaluation splits despite nominally disjoint indices. Because we sample tiles independently, the logistic regression and U-Net might partially memorize a fire's trajectory if adjacent days from the same event appear in both train and eval. Future releases should tag each tile with a fire ID or ignition timestamp so splits can enforce incident-level separation.

The absence of geolocation metadata inside the converted $64 \times 64$ rasters creates another limitation. Fires routinely span multiple tiles on the same day (e.g., an east-west front that crosses tile boundaries), yet our pipeline cannot stitch those tiles into a coherent geospatial footprint. This prevents us from reasoning about cross-tile spread or capturing interactions with large-scale features such as valleys or metropolitan areas. Adding explicit latitude/longitude corners or tile adjacency graphs would allow models to reason about perimeters that grow beyond a single raster.

Even with 20,097 tiles, the dataset represents only a few hundred unique fires (Section III), meaning the true sample size for independent events remains small. This constrains generalization to rare regimes (wind-driven crown fires, complex smoke plumes) and shows up in the robustness slices where high-wind tiles underperform. Gathering more annotated incidents, especially from years with extreme weather, should be a priority.

Finally, our modeling paradigm treats next-day spread as binary segmentation. While intuitive, this view misses richer temporal dynamics such as the vector magnitude and direction of the advancing front. A promising extension is to recast the task as predicting per-pixel motion fields or a coarse vector describing forward velocity, heading, and volumetric growth. Such a formulation could share components with optical-flow or trajectory forecasting models, capture uncertainty more naturally, and plug directly into downstream simulators for evacuation planning. Combining that vector-based view with longer temporal context (multiple previous days instead of one) and explicit temporal encoding (transformers, autoregressive decoders) would further reduce leakage and improve realism.

## VIII. Member Contributions

- Robert Clay Harris: Managed the codebase and GitHub repository, drove code cleanup and QA, produced the the robustness-slice analysis, and synthesized and edited the manuscript.
- Hannah Richardson: Led the feature-family ablation studies, authored the GitHub README, contributed to code QA/cleanup, and handled targeted manuscript revisions.
- Chelsey Blowe: Conducted the preliminary logistic regression and U-Net modeling runs and produced the qualitative model comparisons.

## References

[1] R. R. Linn, J. M. Reisner, J. J. Colman, and J. Winterkamp, "FIRETEC: A physics-based model for coupled fire–atmosphere interaction," *International Journal of Wildland Fire*, vol. 11, no. 4, pp. 233–246, 2002. DOI: https://doi.org/10.1071/WF02006.

[2] W. Mell, S. L. Manzello, A. Maranghides, D. Butry, and R. G. Rehm, "The wildland–urban interface fire problem: Current approaches and research needs," *International Journal of Wildland Fire*, vol. 19, no. 2, pp. 238–251, 2009. DOI: https://doi.org/10.1071/WF07131.

[3] J. L. Coen, M. Cameron, J. Michalakes, E. G. Patton, P. J. Riggan, and K. Yedinak, "WRF-Fire: Coupled weather–wildland fire modeling with the Weather Research and Forecasting model," *Journal of Applied Meteorology and Climatology*, vol. 52, no. 1, pp. 16–38, 2013. DOI: https://doi.org/10.1175/JAMC-D-12-023.1.

[4] A. L. Sullivan, "Wildland fire behaviour modelling: Review of models, evaluation, and applications," *International Journal of Wildland Fire*, vol. 18, no. 4, pp. 369–386, 2009. DOI: https://doi.org/10.1071/WF06143.

[5] A. Bakhshaii and E. A. Johnson, "A review of wildland fire spread modelling, 1990–2019," *International Journal of Wildland Fire*, vol. 28, no. 11, pp. 827–842, 2019. DOI: https://doi.org/10.1071/WF19002.

[6] I. Papakis, V. Linardos, and M. Drakaki, "A Multimodal Ensemble Deep Learning Model for Wildfire Prediction in Greece Using Satellite Imagery and Multi-Source Remote Sensing Data," *Remote Sensing*, 17, no. 19: 3310, 2025. DOI: https://doi.org/10.3390/rs17193310.

[7] D. Shadrin, S. Illarionova, F. Gubanov, K. Evteeva, M. Mironenko, I. Levchunets, R. Belousov, and E. Burnaev, "Wildfire spreading prediction using multimodal data and deep neural network approach," *Scientific Reports*, vol. 14, no. 2606, 2024. DOI: https://doi.org/10.1038/s41598-024-52821-x.

[8] G. Hulsey, "Modified Next Day Wildfire Spread (mNDWS)," Kaggle Dataset, Sep. 2024. https://www.kaggle.com/datasets/georgehulsey/modified-next-day-wildfire-spread.

[9] U.S. Department of Agriculture Climate Hubs, "Wildfire," U.S. Department of Agriculture, Climate Hubs. [Online]. Available: https://www.climatehubs.usda.gov/taxonomy/term/398. Accessed: Oct. 2025.