

Short-Horizon Wildfire Growth Forecasting Using Satellite-Derived Fire-Behavior Features

Areen Jain

Gretchen Whitney High School

Abstract

Predicting which active wildfires will grow substantially over the next day can help fire managers direct limited crews and resources. This paper tests whether weather and terrain improve that prediction beyond a fire’s own recent behavior. A dataset of 10,076 fire-days from 320 western United States fires (2018–2021) was built from the NASA Fire Event Data Suite (FEDS) and joined to ERA5 weather, terrain, and gridMET fire-weather indices. Substantial growth was defined in advance as more than 25% relative growth and at least 2 km² of absolute growth in 24 hours. Tree-based models were trained with a fire-level temporal split and compared with a paired, fire-clustered bootstrap. A random forest using only fire-history features reached a test ROC-AUC of 0.911, compared with 0.694 for persistence. Adding weather changed ROC-AUC by -0.002 (95% CI $[-0.007, 0.003]$), and even the actual weather during the forecast window did not help. On quiet days, when history gives the least warning, neither ERA5 nor gridMET gave a significant improvement. These results suggest that satellite-observed fire behavior carries most of the usable signal for next-day growth.

Keywords: wildfire growth; FEDS; VIIRS; ERA5; gridMET; random forest; ablation study; clustered bootstrap

1 Introduction

1.1 Background

Wildfires in the western United States have become larger and more destructive. Warmer temperatures and earlier snowmelt have lengthened the fire season [1], and human-caused

climate change has been linked to a large increase in forest area burned [2]. Fire managers therefore need to know not only where fires are burning, but which ones are likely to grow quickly in the next day.

Traditional spread models, such as Rothermel’s surface fire model [3] and the FARSITE simulator [4], use physical equations that depend on fuel, moisture, wind, and slope. These models need detailed inputs that are often uncertain. Machine learning offers another approach by learning patterns directly from historical fires [5], and the VIIRS satellite sensor now detects active fire at 375 m resolution several times a day [6].

This paper asks: for active large wildfires in the western United States from 2018 to 2021, do weather and terrain improve prediction of substantial 24-hour growth beyond a model that uses only the fire’s current size and recent growth? The paper builds a leakage-safe fire-day dataset, measures how much weather and terrain add using a paired statistical test, and checks whether weather helps more on days after a fire has been quiet.

1.2 Literature Review

Many recent wildfire spread models treat the problem as an image task. Huot et al. [7] created the Next Day Wildfire Spread dataset, which pairs satellite fire masks with weather, terrain, and vegetation layers, and trained convolutional networks to predict the next day’s fire mask. Gerard et al. [8] extended this with WildfireSpreadTS, a multi-day time-series benchmark. These studies show that deep models can use weather layers, but they rarely measure how much of the skill comes from weather compared with the fire’s own recent footprint.

Tabular approaches have also been used. Coffield et al. [9] used decision trees to predict final fire size at ignition and found weather and vegetation to be useful. For growth during a fire, Chen et al. [10] built an object-based system that tracks individual fires every 12 hours from VIIRS detections, which became the basis of the FEDS archive used here. Rapid growth events, or blowups, are especially dangerous, and Werth et al. [11] describe how wind, dry fuels, and terrain can trigger them. This motivates testing whether weather helps most after quiet periods.

Two methodological points shape this study. Tree-based ensembles often match or beat deep learning on medium-sized tabular data [12], which supports a tabular design. Prediction studies can also report inflated skill when future information leaks into training [13]. Because fire-days from the same fire are strongly related, this study splits data by fire and by year and estimates uncertainty by resampling whole fires.

2 Data Description

Fire data come from the NASA FEDS large-fire perimeter archive [10], which reconstructs fire perimeters every 12 hours from VIIRS detections and reports burned area, perimeter, active fire-line length, mean fire radiative power (FRP), and new fire pixels. The 320 longest-lived western United States fires from 2018 to 2021 were kept, 80 from each ignition year. One prediction was made per fire per day at 12:00 UTC, and a fire-day was kept only if the fire was active and a perimeter 24 hours later existed. Hourly weather comes from ERA5 reanalysis [14] through the Open-Meteo archive [15], and terrain from the Copernicus elevation model through Open-Meteo. A sensitivity analysis added daily gridMET data at about 4 km resolution [16], including fuel moisture, energy release component (ERC), and burning index (BI). Fires were split by ignition year: 2018–2019 for training, 2020 for validation, and 2021 for the final test. No fire appears in more than one split, and the test set was scored only once (Table 1). The study population was restricted to long-lived, large western United States fires represented in the FEDS archive. (Figure 1).

Table 1: Dataset summary and temporal fire-level split.

Split	Years	Fires	Fire-days	Substantial growth (rate)
Train	2018–2019	160	5,472	288 (5.3%)
Validation	2020	80	2,341	205 (8.8%)
Test	2021	80	2,263	224 (9.9%)
Total	2018–2021	320	10,076	717 (7.1%)

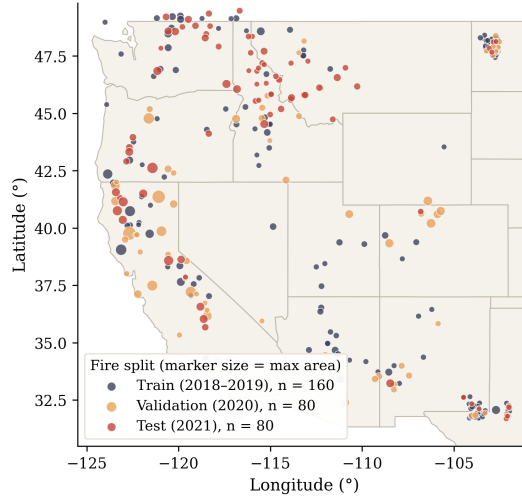


Figure 1: Locations of the 320 study fires by split. Marker size reflects each fire’s maximum burned area.

3 Theoretical Framework

3.1 Problem Formulation

Let $A_i(t)$ be the burned area of fire i at time t in km^2 . The target is substantial growth over the next 24 hours:

$$y_{i,t} = \mathbb{1} \left[\frac{A_i(t+24) - A_i(t)}{A_i(t)} > 0.25 \text{ and } A_i(t+24) - A_i(t) \geq 2 \text{ km}^2 \right]. \quad (1)$$

The relative threshold captures meaningful change for small fires, and the absolute threshold stops tiny perimeter changes from counting as substantial. Recent growth over the past h hours is

$$g_i^{(h)}(t) = \frac{A_i(t) - A_i(t-h)}{\max\{A_i(t-h), 0.01\}}, \quad h \in \{12, 24, 48\}. \quad (2)$$

Figure 2 shows how the two thresholds in Equation 1 divide the fire-days.

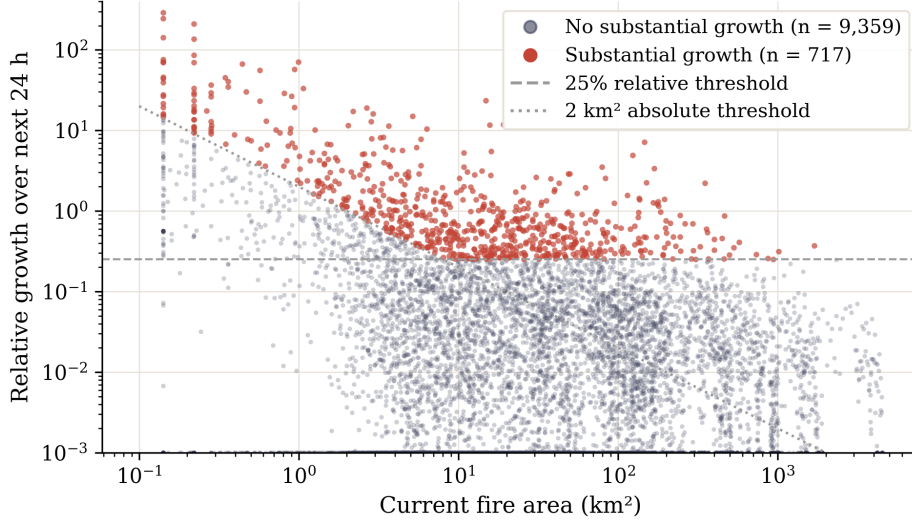


Figure 2: Relative 24-hour growth against current fire area for all 10,076 fire-days. Red points meet both thresholds of the substantial-growth label.

Features are grouped into history $\mathcal{H}$, weather $\mathcal{W}$, and geography $\mathcal{G}$. The main question compares nested models $\hat{p}_{\mathcal{H}}$, $\hat{p}_{\mathcal{H}\mathcal{W}}$, and $\hat{p}_{\mathcal{H}\mathcal{W}\mathcal{G}}$. If weather adds information beyond history, $\hat{p}_{\mathcal{H}\mathcal{W}}$ should rank held-out fire-days better than $\hat{p}_{\mathcal{H}}$. A fire-day is called *quiet* when the previous 24 hours showed little growth:

$$q_{i,t} = \mathbb{1} \left[g_i^{(24)}(t) \leq 0.05 \text{ and } |A_i(t) - A_i(t-24)| < 2 \text{ km}^2 \right]. \quad (3)$$

Substantial growth after a quiet day is called a *blowup*. On these days, history gives the least warning, so weather has the best chance to help.

3.2 Random Forests and Gradient Boosting

A random forest [17] averages B decision trees, each trained on a bootstrap sample with a random subset of features at each split:

$$\hat{p}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}). \quad (4)$$

Gradient boosting builds trees in sequence, with each tree correcting the errors of the current model:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta f_m(\mathbf{x}), \quad \hat{p}(\mathbf{x}) = \frac{1}{1 + e^{-F_M(\mathbf{x})}}, \quad (5)$$

where η is the learning rate. XGBoost [18] adds regularization and a second-order approximation of the loss. Both methods capture nonlinear thresholds and interactions, such as a wind effect that only matters for small fires.

3.3 Evaluation Metrics

The primary metric is ROC-AUC, the probability that a randomly chosen positive fire-day is scored higher than a randomly chosen negative one:

$$\text{AUC} = \frac{1}{|P||N|} \sum_{a \in P} \sum_{c \in N} \mathbb{I}[\hat{p}_a > \hat{p}_c]. \quad (6)$$

Because positives are rare, PR-AUC (average precision) is also reported, since it focuses on the minority class [19]. Its no-skill value equals the positive rate.

3.4 Fire-Clustered Paired Bootstrap

Fire-days from the same fire are not independent, so whole fires are resampled [20]. For each of R replicates, a sample $\mathcal{F}_r^*$ of test fires is drawn with replacement and

$$\Delta_r^* = \text{AUC}(\hat{p}_{\mathcal{H}\mathcal{W}}; \mathcal{F}_r^*) - \text{AUC}(\hat{p}_{\mathcal{H}}; \mathcal{F}_r^*) \quad (7)$$

is computed. Scoring both models on the same fires makes the comparison paired. The 95% confidence interval is the 2.5th to 97.5th percentile of $\{\Delta_r^*\}$. If it contains zero, the added features are not shown to help.

3.5 Permutation Importance

The importance of feature j is the drop in test ROC-AUC when its values are shuffled [17]:

$$I_j = \text{AUC}(\hat{p}; \mathbf{X}) - \frac{1}{K} \sum_{k=1}^K \text{AUC}(\hat{p}; \mathbf{X}^{(j,k)}). \quad (8)$$

4 Methodology

4.1 Feature Engineering

All features use only information available at 12:00 UTC on the prediction day (Table 2). Weather features summarize the ERA5 hours in the 24 hours before prediction, plus 7-day and 30-day rainfall. Wind direction is encoded with sine and cosine so that 359° and 1° are treated as close. gridMET values come from the previous calendar day so that no part of the forecast window is included.

Table 2: Feature families.

Family	Features
History (10)	log area; growth over previous 12, 24, 48 h (Eq. 2); absolute 24 h growth; fire age; fire-line length; mean FRP; new fire pixels; perimeter
Weather (10)	max temperature; min humidity; max wind speed; wind direction (sine, cosine); max vapor pressure deficit; rainfall (24 h, 7 d, 30 d); soil moisture
Geography (7 + state)	elevation; slope; roughness; longitude; latitude; day of year (sine, cosine); state
gridMET (13)	temperature; humidity; wind; vapor pressure deficit; rainfall; 100-h and 1000-h fuel moisture; ERC; BI; 3-day means of ERC, 100-h fuel moisture, vapor pressure deficit, and BI

4.2 Models and Training

Two baselines were used: a majority-class predictor and a persistence model that predicts substantial growth tomorrow if the fire grew substantially in the previous 24 hours. Logistic regression, random forest, histogram gradient boosting, and XGBoost were then trained on each feature set using scikit-learn and XGBoost in Python. Tree model settings were chosen by validation ROC-AUC from small grids (for the random forest: 200 trees, depth 6 or 10, and minimum leaf size 3 or 8, with balanced class weights). The random forest had the highest validation ROC-AUC and was used as the primary model. Label thresholds, split years, and the primary comparison were fixed before training. As an upper-bound check, one model was also given the actual weather during the forecast window, which is not available in practice.

4.3 Follow-Up Analyses

Two follow-ups used the same split. First, random forests with and without weather were compared on subsets of test days: quiet, active, young fires (7 days old or less), high-wind days, and quiet days on young fires. Second, models were trained and tested only on quiet days, comparing history alone with history plus ERA5, gridMET, or both.

5 Results and Discussion

5.1 Test Set Performance

Table 3 and Figure 3 show locked test results for the primary random forest and the baselines. All three feature sets reach a ROC-AUC of about 0.91, far above persistence (0.694) and the majority baseline (0.500). PR-AUC is also well above the no-skill value of 0.099. The same pattern held for every model type: XGBoost reached 0.913 with history alone and 0.913 with history and weather, while logistic regression was weaker (0.886), suggesting nonlinear relationships.

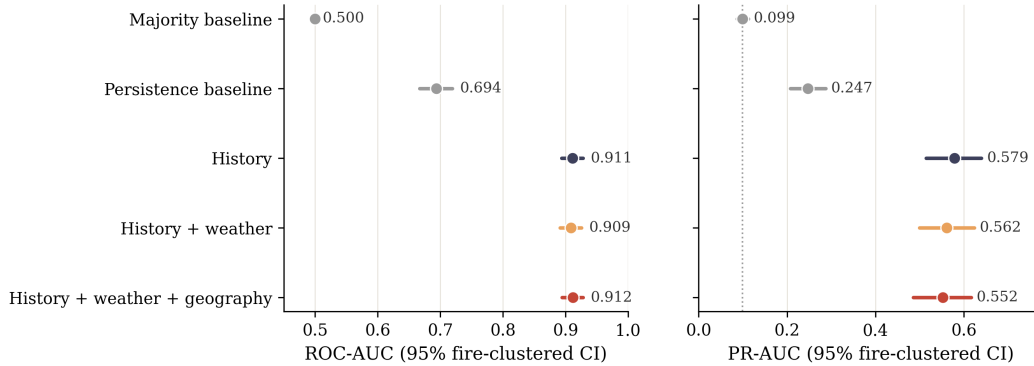


Figure 3: Test ROC-AUC and PR-AUC with fire-clustered 95% confidence intervals.

Table 3: Locked test performance (80 fires, 2,263 fire-days) with fire-clustered 95% confidence intervals.

Model	Features	ROC-AUC [95% CI]	PR-AUC [95% CI]
Majority	—	0.500	0.099
Persistence	—	0.694 [0.667, 0.719]	0.247 [0.207, 0.287]
Random forest	$\mathcal{H}$	0.911 [0.893, 0.928]	0.579 [0.515, 0.640]
Random forest	$\mathcal{HW}$	0.909 [0.891, 0.925]	0.562 [0.500, 0.624]
Random forest	$\mathcal{HWG}$	0.912 [0.894, 0.928]	0.552 [0.486, 0.617]

5.2 Do Weather and Terrain Add Value?

Table 4 gives the primary result. Adding weather to history changed ROC-AUC by -0.002 with an interval of $[-0.007, 0.003]$, and adding geography changed it by $+0.003$ with an interval of $[-0.001, 0.007]$. Both intervals contain zero and are narrow, so large gains are ruled out. Even when given the actual weather during the next 24 hours, the model reached only 0.901, no better than history alone. This means the null result is not simply because weather is hard to forecast.

Table 4: Paired fire-clustered bootstrap comparisons for the random forest (2,000 replicates).

Comparison	$\Delta\text{ROC-AUC}$	95% CI
Weather beyond history ($\mathcal{HW} - \mathcal{H}$)	-0.002	$[-0.007, 0.003]$
Geography beyond weather ($\mathcal{HWG} - \mathcal{HW}$)	$+0.003$	$[-0.001, 0.007]$
All context beyond history ($\mathcal{HWG} - \mathcal{H}$)	$+0.001$	$[-0.006, 0.006]$

5.3 Which Features Matter?

Permutation importance (Figure 4) shows that the model relies on fire-behavior features. Mean FRP (0.017), fire-line length (0.017), perimeter (0.014), and 24-hour growth (0.014) were most important. The strongest weather feature, maximum wind speed, had an importance of only 0.001. FRP and fire-line length describe how intense and widespread burning is at the moment of prediction, so they likely already reflect the combined effect of weather, fuels, and terrain.

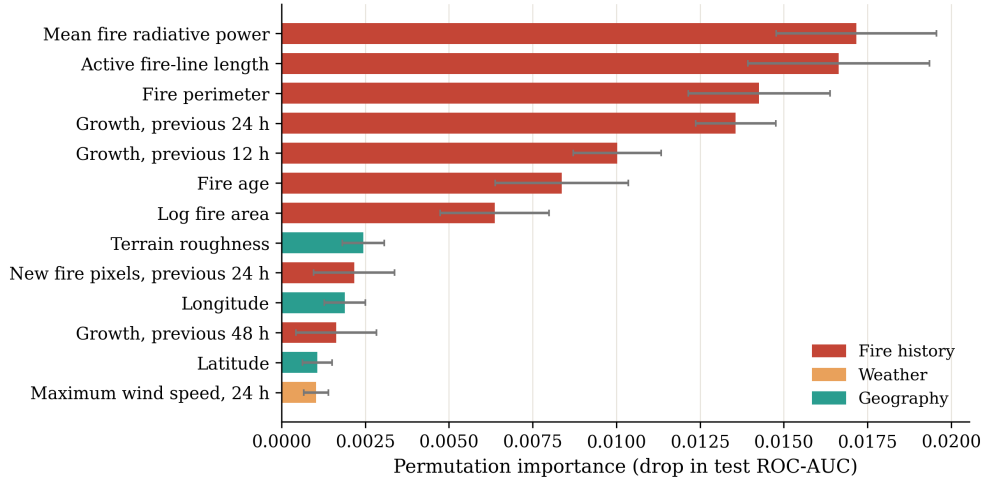


Figure 4: Permutation importance for the full-feature random forest on the test set.

5.4 When Does Weather Matter?

About half of the test fire-days were quiet (1,067 of 2,263), and 39 of these were followed by substantial growth. Within quiet days, the history-only model still ranked fire-days well (ROC-AUC 0.892), so the fire’s current intensity and size carry warning signs even when it has not recently grown. Adding weather raised quiet-day ROC-AUC to 0.920 ($\Delta = +0.027$, $[-0.001, 0.049]$), the largest weather effect in the study, but the interval still touches zero (Figure 5). On active days, weather slightly lowered ROC-AUC ($\Delta = -0.007$, $[-0.015, -0.0002]$). This interval barely excludes zero, and six subsets were tested without correction, so it should be treated as exploratory. One explanation is that when a fire is actively growing, its history already explains most of the variation and extra weather features mainly add noise.

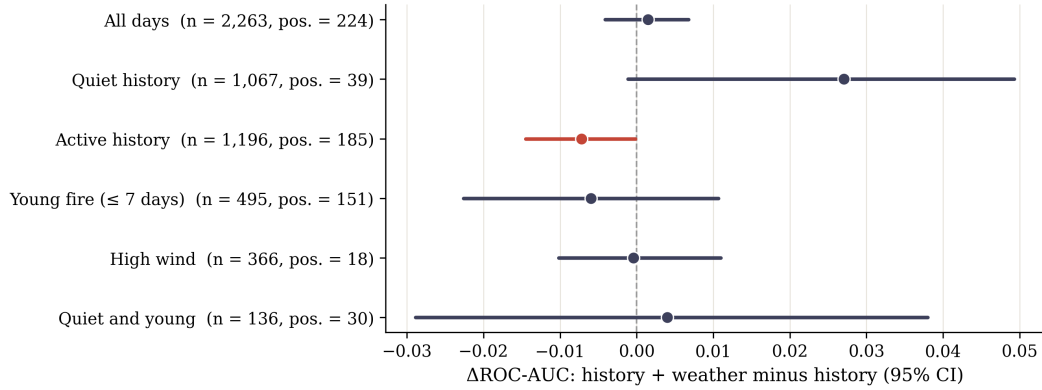


Figure 5: Paired weather effect by test subset. Red marks an interval that excludes zero.

5.5 Quiet Days with Fire-Weather Indices

Because quiet days showed the most promise, the final analysis trained models only on quiet days and added gridMET fuel moisture and fire-danger indices (Figure 6). History alone reached 0.885, history plus ERA5 0.912, history plus gridMET 0.894, and all three 0.899. Neither weather source significantly beat history: ERA5 gave $+0.025$ ($[-0.011, 0.054]$) and gridMET gave $+0.009$ ($[-0.027, 0.038]$). gridMET was significantly worse than ERA5 (-0.017 , $[-0.030, -0.003]$), likely because it is a daily product taken from the previous day, while ERA5 covers the 24 hours ending at prediction time. FRP, perimeter, and fire-line length remained the most important features.

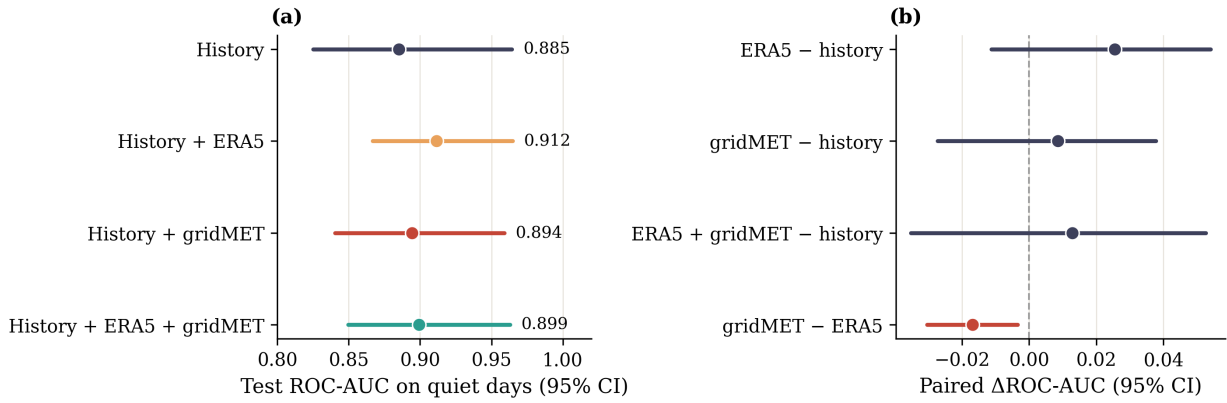


Figure 6: Quiet-day models: (a) test ROC-AUC for each feature set and (b) paired differences. Red marks an interval that excludes zero.

6 Conclusion

6.1 Findings

Within the selected FEDS population of long-lived western United States fires from 2018–2021, fire-history variables provided strong next-day ranking performance, while the weather and terrain representations evaluated here did not provide a statistically distinguishable improvement in ROC-AUC. This result does not imply that weather is unimportant to wildfire growth; rather, the available fire-behavior measurements captured most of the predictive information detectable by these models under the present sampling and evaluation design.

This paper tested whether weather and terrain improve 24-hour prediction of substantial wildfire growth beyond a fire’s recent history. Using 320 FEDS fires and a fire-level temporal split, a random forest with only history features reached a test ROC-AUC of 0.911, far above persistence (0.694). Adding ERA5 weather (-0.002) or terrain ($+0.003$) did not significantly change performance, and even perfect knowledge of forecast-window weather did not help. FRP and fire-line length were the most important features. On quiet days, history still ranked fire-days well, and neither ERA5 nor gridMET gave a significant improvement.

6.2 Limitations

FEDS perimeters are reconstructed from satellite detections, so clouds and smoke can hide growth. The study covers only large western fires from 2018 to 2021, and the test set is a single year of 80 fires. Weather comes from reanalysis rather than local stations, and no fuel

maps or firefighting data were used. Blowups are rare, so their intervals are wide, and the subset analyses were not corrected for multiple comparisons. Finally, the results describe prediction, not the causal effect of weather on fire.

6.3 Future Work

Future work could add firefighting effort from incident reports, since suppression may explain why some intense fires do not grow. Fuel-type maps and hourly high-resolution weather could test whether finer data change the weather result, especially on quiet days. More years and regions would provide more blowups and tighter intervals. Most importantly, this study shows why wildfire models should report a history-only baseline, so that skill coming from the fire’s own behavior is not credited to weather.

7 Supplemental Materials

The complete Python code, dataset, and instructions on how to run the codes are available in the project’s Github Repository for reproducibility and further exploration. The link is: <https://github.com/areenjain09/fire-detection>

8 Acknowledgments

This work would not have been possible without the guidance of Dr. Olga Korosteleva, Professor of Statistics at California State University, Long Beach, whose mentorship and deep expertise shaped both the direction of this project and my growth as a researcher. I am equally grateful to the teachers and mentors who have invested in my academic development over the years. To my family, whose encouragement and presence has been a constant source of strength throughout this journey, I owe more than words can express.

References

- [1] A. L. Westerling, H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam, “Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity,” *Science*, vol. 313, no. 5789, pp. 940–943, 2006.

- [2] J. T. Abatzoglou and A. P. Williams, “Impact of Anthropogenic Climate Change on Wildfire Across Western US Forests,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. 11770–11775, 2016.
- [3] R. C. Rothermel, “A Mathematical Model for Predicting Fire Spread in Wildland Fuels,” USDA Forest Service, Research Paper INT-115, 1972.
- [4] M. A. Finney, “FARSITE: Fire Area Simulator—Model Development and Evaluation,” USDA Forest Service, Research Paper RMRS-RP-4, 1998.
- [5] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan, “A Review of Machine Learning Applications in Wildfire Science and Management,” *Environmental Reviews*, vol. 28, no. 4, pp. 478–505, 2020.
- [6] W. Schroeder, P. Oliva, L. Giglio, and I. A. Csizsar, “The New VIIRS 375 m Active Fire Detection Data Product: Algorithm Description and Initial Assessment,” *Remote Sensing of Environment*, vol. 143, pp. 85–96, 2014.
- [7] F. Huot, R. L. Hu, N. Goyal, T. Sankar, M. Ihme, and Y.-F. Chen, “Next Day Wildfire Spread: A Machine Learning Dataset to Predict Wildfire Spreading From Remote-Sensing Data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.
- [8] S. Gerard, Y. Zhao, and J. Sullivan, “WildfireSpreadTS: A Dataset of Multi-Modal Time Series for Wildfire Spread Prediction,” in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, vol. 36, 2023.
- [9] S. R. Coffield, C. A. Graff, Y. Chen, P. Smyth, E. Foufoula-Georgiou, and J. T. Randerson, “Machine Learning to Predict Final Fire Size at the Time of Ignition,” *International Journal of Wildland Fire*, vol. 28, no. 11, pp. 861–873, 2019.
- [10] Y. Chen, S. Hantson, N. Andela, S. R. Coffield, C. A. Graff, D. C. Morton, L. E. Ott, E. Foufoula-Georgiou, P. Smyth, M. L. Goulden, and J. T. Randerson, “California Wildfire Spread Derived Using VIIRS Satellite Observations and an Object-Based Tracking System,” *Scientific Data*, vol. 9, 249, 2022.
- [11] P. A. Werth, B. E. Potter, C. B. Clements, M. A. Finney, S. L. Goodrick, M. E. Alexander, M. G. Cruz, J. A. Forthofer, and S. S. McAllister, “Synthesis of Knowledge

of Extreme Fire Behavior: Volume I for Fire Managers,” USDA Forest Service, General Technical Report PNW-GTR-854, 2011.

- [12] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?” in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, vol. 35, pp. 507–520, 2022.
- [13] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, “Leakage in Data Mining: Formulation, Detection, and Avoidance,” *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 1–21, 2012.
- [14] H. Hersbach et al., “The ERA5 Global Reanalysis,” *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.
- [15] P. Zippenfenig, *Open-Meteo.com Weather API*, Zenodo, 2023. doi:10.5281/zenodo.7970649. Available at: <https://open-meteo.com/>.
- [16] J. T. Abatzoglou, “Development of Gridded Surface Meteorological Data for Ecological Applications and Modelling,” *International Journal of Climatology*, vol. 33, no. 1, pp. 121–131, 2013.
- [17] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [19] T. Saito and M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” *PLoS ONE*, vol. 10, no. 3, e0118432, 2015.
- [20] C. A. Field and A. H. Welsh, “Bootstrapping Clustered Data,” *Journal of the Royal Statistical Society: Series B*, vol. 69, no. 3, pp. 369–390, 2007.