Spatial Uncertainty in Wildfire Forecasting Using Multi-Modal Earth Observation

Anonymous authors
Paper under double-blind review

Abstract

Accurate wildfire forecasting from remote sensing data is essential for climate resilience and emergency planning. Beyond predictive performance, understanding where and why uncertainty arises is critical for operational trust. We analyze the spatial structure of predictive uncertainty in wildfire spread forecasts using multimodal Earth observation (EO) inputs, including Sentinel-2 vegetation indices and VIIRS thermal reflectance. Using Monte Carlo dropout and deep ensembles, we show that predictive entropy maps exhibit coherent spatial patterns aligned with fire boundaries, unlike randomized baselines. We introduce a novel and interpretable centroid-oriented distance metric that reveals high-uncertainty regions consistently form 20–60 meter buffer zones around predicted firelines. Feature attribution using integrated gradients highlights vegetation condition and recent fire activity as primary drivers of model confidence. Deep ensembles further confirm that these uncertainty estimates are probabilistically well-calibrated across multiple folds. Together, these results suggest that spatial uncertainty in EO-based wildfire forecasting is structured, interpretable, and operationally actionable. The code for all experiments is available on GitHub.[1]

1 Introduction

Wildfires have become an escalating global crisis, intensified by climate change, prolonged droughts, and expanding human development. Most recently in January 2025, Southern California experienced one of its most devastating wildfire events. The wildfires ravaged over 57,000 acres, destroyed more than 18,000 structures, led to the evacuation of over 200,000 residents, and resulted in at least 30 fatalities. Economic losses from these fires are estimated to exceed 250 billion USD, marking them among the costliest natural disasters in U.S. history LACEDC (2025). Wildfires in the European Union have become increasingly frequent and severe, with over 166,000 hectares burned by May 2025—nearly three times the long-term average—driven by climate change and affecting regions beyond the traditional Mediterranean hotspots European Forest Fire Information System (EFFIS) (2025). Globally, regions such as the Amazon, North America, Australia, and parts of Africa have witnessed unprecedented wildfire activity, leading to significant ecological damage, loss of biodiversity, and adverse health effects due to smoke exposure Cunningham et al. (2024); Institute (2024). As climate change continues to exacerbate wildfire risks, there is an urgent need for accurate, high-resolution wildfire forecasting to aid in early response, resource allocation, and risk mitigation. While remote sensing products like VIIRS Schroeder et al. (2014) and MODIS provide near real-time fire detections, they do not forecast how wildfires will evolve in the days to come.

Traditionally, fire spread forecasting has relied on physics-based simulators such as Farsite Finney (1998) and Prometheus Tymstra et al. (2010), which use hand-crafted rules and environmental inputs to simulate fire growth. These tools are interpretable and physically grounded but require fine-grained inputs—like fuel maps and localized weather forecasts—which are difficult to obtain in real time and hard to calibrate to dynamic fire conditions.

Machine learning has emerged as a scalable alternative for fire forecasting Radke et al. (2019); Bolt et al. (2022), learning directly from remote sensing and historical fire data. Shadrin et al. (2024) trained U-Net

[1] `https://github.com/roloccark/wildf-UQ`

and DeepLabV3 models to perform multi-day fire spread segmentation using multimodal remote sensing and meteorological inputs. Their work benchmarks a variety of input combinations and shows strong predictive performance, but does not quantify uncertainty—leaving users without insight into where or why the model might be wrong.

In parallel Huot et al. (2022) introduced the NextDayWildfireSpread dataset, emphasizing single-frame predictions of fire growth based on remote sensing. Gerard et al. (2023) extended this effort by releasing WildfireSpreadTS, a large-scale benchmark for multi-temporal wildfire forecasting across 607 fire events. The dataset incorporates Sentinel-2, VIIRS thermal bands, fire history, meteorology, and slope, enabling testing of temporal models like ConvLSTM Shi et al. (2015b) and U-Net with Temporal Attention Encoder (UTAE Garnot & Landrieu (2022)).

Generative modeling is also being explored for fire forecasting. Shaddy et al. (2024) use a physics-informed GAN (cWGAN) to fuse fire simulations from WRF-SFIRE with satellite observations. Their system outputs ensembles of arrival time maps conditioned on sparse input detections, providing uncertainty-aware predictions that can be used to initialize atmospheric-fire models. However, their approach focuses more on generating plausible initial states than on operational fire masks or end-to-end spatial uncertainty analysis using public datasets.

Despite the operational risks involved, no prior work has investigated uncertainty quantification in high-resolution wildfire forecasting. Most existing approaches are fully deterministic, producing binary or probabilistic predictions without expressing model confidence—leaving critical questions of when and where the model may fail unanswered. This omission is especially concerning for wildfire response, where uncertainty-aware decision-making is essential for frontline planning, containment strategies, and risk assessment. In this paper, we take a first step toward addressing this gap by analyzing the spatial structure of predictive uncertainty in Earth observation-based wildfire forecasts. Using Monte Carlo dropout and attribution techniques, we investigate where uncertainty arises, how it aligns with vegetation and fire morphology, and how it could guide the construction of interpretable buffer zones to support triage and operational planning.

2 Why This Study

Our work focuses on the operational quantification of uncertainty in deep learning-based, pixel-wise wildfire spread prediction. While prior efforts have tackled deterministic segmentation, event-level uncertainty, probabilistic fire danger indices, surrogate-assisted physical modeling, and generative reconstruction of fire histories, these approaches often lack the spatial resolution or reliability needed for on-the-ground decision-making. We aim to directly benchmark and interpret spatial uncertainty in multi-day, high-resolution spread forecasts using real-world wildfire events. By emphasizing calibrated and interpretable uncertainty at the pixel level, our work bridges the gap between state-of-the-art machine learning and the practical needs of fire managers and scientists operating in high-risk environments.

3 Methods

Model We use the UTAE model (Garnot & Landrieu, 2020), a transformer-based spatiotemporal encoder-decoder architecture designed for multitemporal satellite image time series. UTAE has previously shown strong performance on change detection and land cover segmentation tasks using Sentinel-2 data, and is well-suited for the wildfire spread forecasting setting, where temporal patterns are key. As a baseline comparison, we also experiment with a ConvLSTM model (Shi et al., 2015a), a recurrent architecture tailored for spatiotemporal prediction. ConvLSTM has the ability to capture local temporal dependencies using convolutional operations, and has been used in wildfire modeling in prior work (Burge et al., 2022).

Dataset We conduct all experiments on the publicly available WildfireSpreadTS dataset[2] (Gerard et al., 2023), which provides spatial-temporal cubes of 64×64 patches centered on active wildfire regions. Each sample consists of 5 days of multimodal input features (Sentinel-2 reflectance bands, meteorological variables,

[2] `https://github.com/SebastianGer/WildfireSpreadTS`

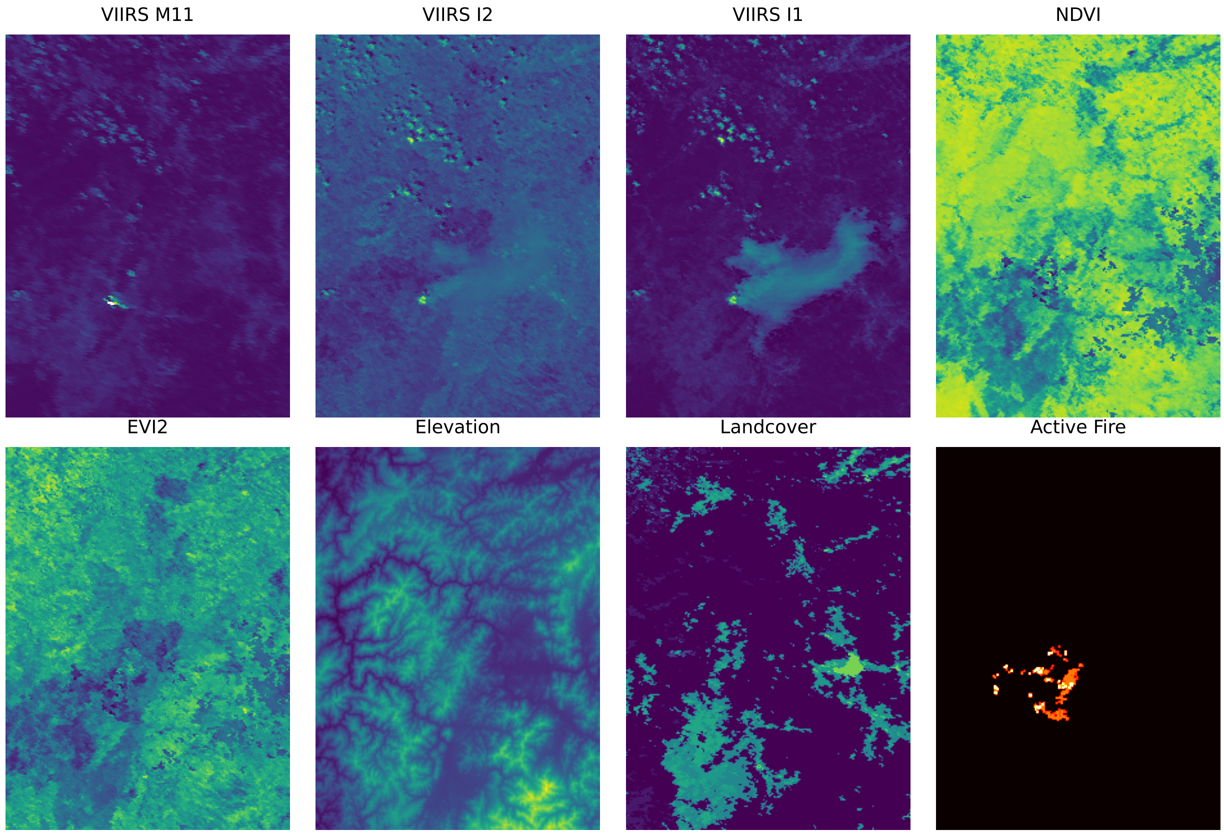


Figure 1: Example input channels from a single sample at prediction time, including Sentinel-2 bands, NDVI, EVI2, and active fire features. These inputs are provided as a 5-day sequence to the model.

NDVI, slope, and other static features), and a binary burn mask for a future day as target. The dataset includes fires from 2018 to 2021 across diverse regions. Following the original benchmark protocol, we perform 12-fold cross-validation over all year-based train/val/test permutations to account for inter-annual variability and covariate shift. A sample of the input modalities used by the model—including reflectance bands, vegetation indices, and active fire features—is visualized in Figure 1. The dataset includes 607 wildfire events across the western United States from January 2018 to October 2021, with a total of 13,607 daily images. These fires span diverse ecosystems and terrain across states like California, Oregon, and Washington.

Training, Evaluation and Uncertainty Quantification The model is trained in a sliding-window fashion. For each fire event, we extract 5-day sequences as input and predict the binary burn mask at the 6th day. Spatial crops (128×128) are used to batch variable-sized regions. Following the original benchmark configuration, we train our model using only five vegetation-based input channels: VIIRS bands M11, I2, I1, NDVI, and EVI2. No meteorological or static terrain features are used. We report Average Precision (AP) on test folds, along with probabilistic calibration metrics described below. We use Monte Carlo (MC) Dropout (Gal & Ghahramani, 2016), performing 20 stochastic forward passes with dropout active at test time. We compute the per-pixel mean and variance of predicted probabilities. Post-hoc temperature scaling (Guo et al., 2017) is applied on validation folds to calibrate the logits. We evaluate the quality of probabilistic predictions using three standard metrics: (1) Negative Log-Likelihood (NLL), which penalizes miscalibrated and overconfident predictions; (2) Brier Score (Brier, 1950), a proper scoring rule for binary probabilistic outputs; and (3) Expected Calibration Error (ECE, (Guo et al., 2017)), which measures the gap between predicted confidence and empirical accuracy across probability bins. These metrics are widely used in uncertainty quantification studies for segmentation and scientific forecasting tasks. The code for the experiments outlined is available on Github[3].

4 Results

Given that the model is driven purely by vegetation-based features, we interpret high uncertainty as a reflection of ambiguous fuel signatures — e.g., sparse or transitional NDVI zones, or conflicting spectral responses.

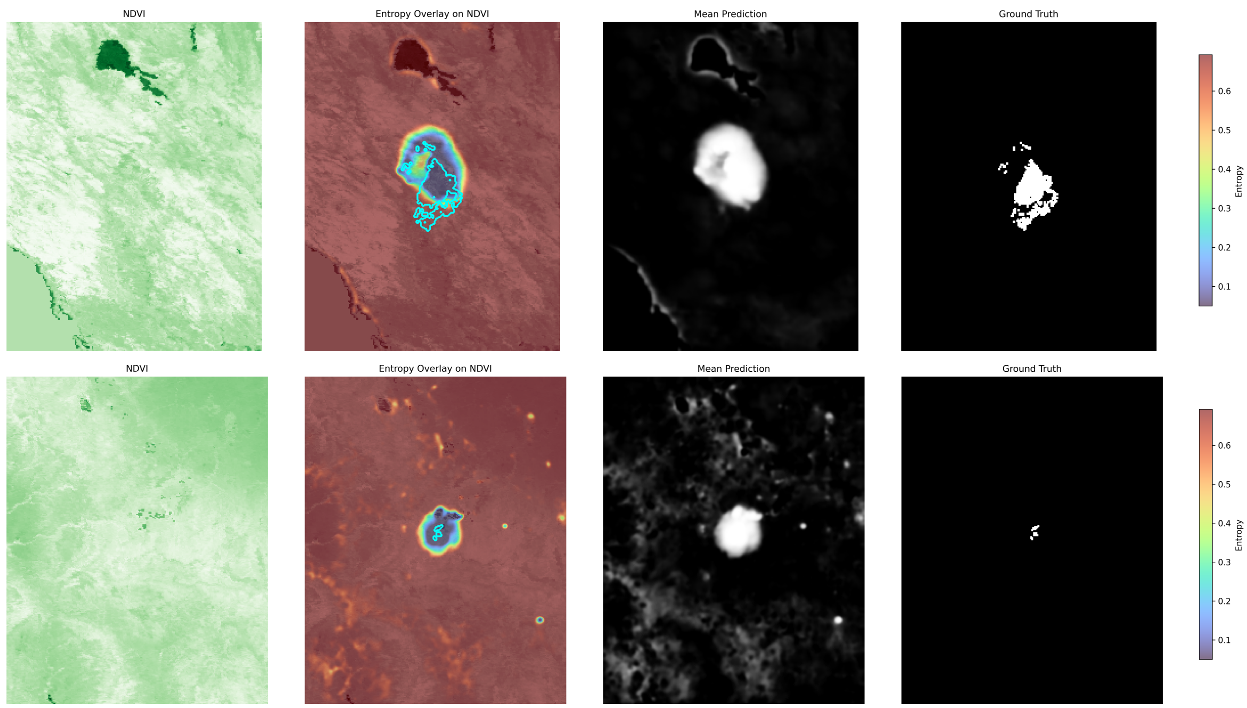


Figure 2: Qualitative comparison of model predictions and predictive uncertainty for three fire events: large (top) and small (bottom). Each row shows: NDVI input, entropy overlaid on NDVI with ground truth contour, mean prediction across 20 MC Dropout samples, and ground truth burn mask. These span a range of ground-truth sizes: approximately 125.6 acres (large),and 5.2 acres (small), corresponding to 1271 and 53 burned pixels respectively.

4.1 Feature Group Ablations

To better understand the contribution of different input feature categories, we conduct a systematic ablation study by training UTAE models using isolated subsets of inputs. Each configuration retains only one feature group—vegetation, weather, land cover, or topography—while preserving active fire indicators to ensure temporal grounding. Table 1 summarizes the 12-fold average precision for each setup. We also include a ConvLSTM model trained on vegetation and active fire to provide architectural contrast. As shown in Table 1, the UTAE model using vegetation and active fire inputs achieves the highest average precision (0.378 ± 0.083), outperforming both ablated UTAE variants and the ConvLSTM baseline (0.304 ± 0.093). Accordingly, all downstream uncertainty analyses in this study—including spatial structure, buffer zone characterization, and feature attribution—are conducted using the UTAE model trained on vegetation inputs and active fire.

[3]https://github.com/roloccark/wildf-UQ

Table 1: Mean Average Precision (AP) across 12 folds for different feature groups using UTAE. For comparison, ConvLSTM trained on vegetation + active fire is also included.

Feature Group	**Mean AP**
Vegetation + active fire	0.378 ± 0.083
Weather + active fire	0.323 ± 0.078
Land cover + active fire	0.319 ± 0.092
Topography + active fire	0.317 ± 0.082
ConvLSTM (veg. + active fire)	0.304 ± 0.093

Benchmarking Calibration for Trustworthy Spatial Uncertainty To ensure that our uncertainty estimates are trustworthy, we evaluate calibration using three standard metrics: Expected Calibration Error (ECE), Brier Score, and Negative Log-Likelihood (NLL). Our primary analyses—such as spatial uncertainty overlays, buffer zone estimation, and feature attribution—are based on Monte Carlo (MC) Dropout (Gal & Ghahramani, 2016) with 20 stochastic forward passes. To benchmark its calibration robustness, we also train a Deep Ensemble of 5 independently initialized models for each of the 12 folds. As shown in Table 2, the ensemble achieves consistently stronger calibration across all metrics—lower ECE (0.512), Brier Score (0.265), and NLL (0.731)—compared to the MC Dropout baseline. This comparison strengthens our confidence that the uncertainty maps derived from MC Dropout are probabilistically reliable, providing a sound foundation for the spatial analyses that follow.

Table 2: Calibration metrics for MC Dropout and Deep Ensemble (12-fold averages). Lower is better.

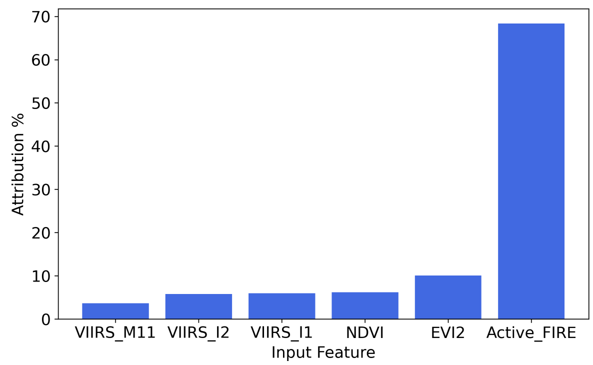
Metric	**MC Dropout**	**Deep Ensemble**
ECE	0.536 ± 0.015	$\mathbf{0.512 \pm 0.018}$
Brier Score	0.294 ± 0.012	$\mathbf{0.265 \pm 0.009}$
NLL	0.805 ± 0.020	$\mathbf{0.731 \pm 0.023}$

Qualitative Uncertainty Patterns. Figure 2 shows entropy overlays and mean predictions for three representative fire events. The three fire examples shown in Figure 2 span a range of ground-truth sizes: approximately 125.6 acres (large) and 5.2 acres (small) corresponding to 1271 and 53 burned pixels respectively. In larger fires, uncertainty is sharply localized near the fire perimeter, while in smaller or fragmented fires, it appears more diffuse and spatially ambiguous. This shows that uncertainty often gathers around the edges of fires—places where the model has to guess how far the fire might have spread, especially when the burn area is small or broken into patches.

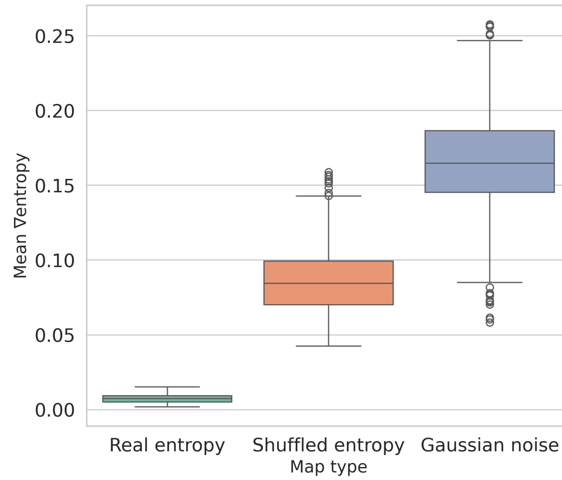
Feature Attribution. To identify which input features most influence uncertainty, we use Integrated Gradients on a ResNet He et al. (2016) surrogate model trained for 50 epochs to approximate the UTAE model's mean predictions. The surrogate model achieves a high fidelity with $R^2 = 0.81$, indicating strong alignment with the original model and validating its use for interpretability. As shown in Figure 3a, attribution is dominated by recent fire activity and vegetation indices (NDVI, EVI2), with relatively lower contributions from thermal reflectance bands (VIIRS-M11, I1, I2). This ranking aligns with biophysical intuition: model uncertainty peaks in sparsely vegetated or transitional fuel zones, where spectral signals are ambiguous or weak, leading to higher predictive hesitation (Archibald et al., 2018).

Spatial Coherence of Uncertainty. If uncertainty is meaningful, it shouldn't appear as random speckling across the map—it should follow real spatial patterns, such as fire boundaries. To test whether this is the case, we analyze the smoothness of the entropy maps by computing their spatial gradient norms: a simple measure of how quickly uncertainty values change across space. We then compare these real uncertainty maps to two baselines: (1) maps where entropy values are randomly shuffled, and (2) synthetic maps generated using Gaussian noise.

As shown in Figure 3b, the real entropy maps have much lower spatial gradients than either baseline, meaning they are smoother and more spatially coherent. This supports the idea that the system's uncertainty is not



(a) Feature importance scores computed using Integrated Gradients on a CNN surrogate. Active fire presence dominates attribution, followed by vegetation indices (NDVI, EVI2). Thermal bands are less influential in determining predictive confidence.



(b) Comparison of spatial gradient norms across entropy maps. Real uncertainty maps are smoother than noise or shuffled baselines, confirming that model uncertainty is spatially coherent and semantically aligned with fire fronts.

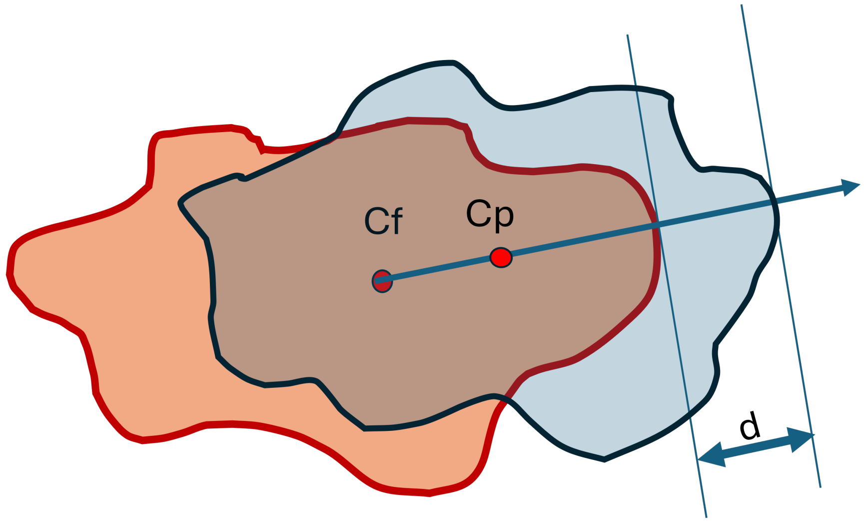
Figure 3: Left: Feature attribution via Integrated Gradients. Right: Spatial coherence of model uncertainty demonstrated through gradient norms.

random, but instead aligned with meaningful structures—particularly the firefront—where prediction is inherently more difficult.

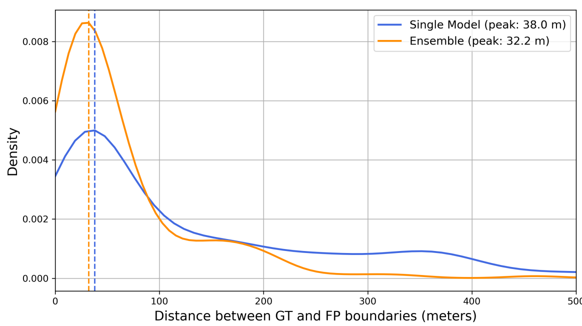
Centroid-Aligned Boundary Distance as a Proxy for Fireline Uncertainty. To better understand the spatial structure of false positives in fireline predictions, we introduce a *centroid-aligned boundary distance* metric. For each test instance, we compute the centroid of the ground truth burn mask (C_f) and the centroid of the predicted fireline (C_p), using the mean prediction from 20 MC dropout samples thresholded at 0.95. We then trace a straight line between C_f and C_p, and identify the nearest points along this axis where the predicted and ground truth firelines terminate. The distance between these two edge points serves as a localized estimate of spatial prediction error.

Aggregating this distance across test samples yields a proxy for the *operational buffer zone*—the typical separation between where the model expects the fire to be and where it actually was. We find that these distances tend to concentrate around 25–35 meters, indicating that model predictions are generally offset by approximately 1.5 to 3.5 Sentinel-2 pixels relative to the observed fireline. This level of spatial deviation is operationally meaningful, particularly in the vicinity of the active fire edge where suppression activities are planned and executed (Morvan & Dupuy, 2001; Thompson et al., 2016). Given that Sentinel-2 bands used in our models have spatial resolutions of 10–20 meters (Drusch et al., 2012), the centroid-aligned boundary distance provides a practically interpretable and spatially grounded complement to traditional pixel-wise accuracy metrics.

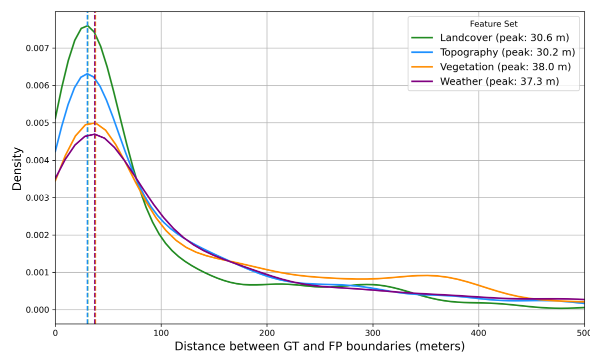
Figure 4a provides a visual summary of this procedure. Figure 4b shows the density of boundary distances for a single MC-dropout model and a deep ensemble, both trained on vegetation-related features. The ensemble exhibits a sharper and more compact uncertainty profile. Figure 4c overlays the distance distributions for models trained on different feature subsets (vegetation, landcover, topography, and weather), highlighting how input modality affects the model's spatial error structure.



(a) Schematic of boundary distance computation.



(b) Single vs Deep Ensemble (Vegetation model).



(c) Distance KDEs across feature variants.

Figure 4: Visualizing the centroid-aligned fireline buffer zone. (**a**) Centroid-aligned boundary distance is computed between predicted and ground truth fire masks. (**b**) Distance KDEs for a single MC-dropout model and deep ensemble (vegetation inputs). (**c**) KDEs across different input feature groups.

5 Discussion

Limitations This study centers on uncertainty modeling in deep learning-based wildfire forecasting using the WildfireSpreadTS dataset. While our main spatial uncertainty analyses focus on the UTAE architecture, we also evaluate a ConvLSTM baseline (see Table 1) and construct a deep ensemble of UTAE models to test robustness across initialization. This cross-architecture and multi-model setup increases confidence that observed uncertainty patterns are not artifacts of a single model or seed. Broader generalization to other architectures (e.g., transformers, diffusion models) remains open for future exploration.

Although the dataset includes fire events spanning a wide range of eco regions across the continental U.S.—from forested landscapes to grasslands—it may not fully extrapolate to other geographies such as the Mediterranean or Australia. Testing under such regimes remains an important avenue for validating robustness.

Our uncertainty quantification focuses solely on epistemic uncertainty, modeled via Monte Carlo Dropout and deep ensembles. Aleatoric uncertainty, which can arise from label noise, cloud cover, or unobserved ignitions, is not modeled here but is a natural extension.

The definition of the uncertainty buffer zone using centroid-aligned boundary distance is intuitive and consistent across folds, but it implicitly assumes centroids are a meaningful geometric representation. This may break down for fragmented or non-convex fires, where alternative shape-aware methods could yield better approximations.

Our input feature set is limited to vegetation indices and active fire bands. This constraint is supported by ablation results, which show that including additional inputs like weather, topography, or land cover reduces performance. We speculate this degradation may arise from temporal misalignment, redundancy, or noise. Importantly, the spatial resolution of these feature groups also differs—vegetation indices and land cover layers are high-resolution, while weather variables are typically coarse (e.g., 2.5–10 km grids). For a pixel-level segmentation task, this resolution mismatch introduces a valuable modeling challenge and may explain why weather-derived features failed to help. Exploring fusion strategies that account for spatial scale differences remains a compelling direction for future work.

Future Work Currently, we are extending this work to explore multi-fold and multi-model comparisons of uncertainty dynamics. Another important direction involves improving how high-dimensional multimodal EO inputs are encoded. With data streams including spectral bands, vegetation indices, terrain layers, and fire history, the input space can be both redundant and noisy. We are investigating more effective compression strategies—such as bottlenecked attention, sparse fusion layers, and contrastive pretraining—to ensure that the most informative features drive both prediction and uncertainty, while reducing model complexity and overfitting risk. Testing whether the observed uncertainty patterns generalize to different geographic regions, fuel types, or fire regimes would validate their robustness. Evaluating uncertainty under domain shift (e.g., cross-continent generalization) or in out-of-distribution conditions could also expose failure modes.

6 Conclusion

This work presents a systematic analysis of spatial uncertainty in high-resolution, Earth observation-based wildfire forecasting. Using Monte Carlo dropout and deep ensembles, we demonstrate that predictive entropy does not manifest as scattered noise, but instead forms coherent spatial structures aligned with fire perimeters and vegetation gradients. These patterns are quantitatively validated: entropy gradients are smoother than randomized baselines, and uncertainty consistently forms narrow, spatially meaningful bands—typically 20–60 meters wide—around predicted firelines.

To formalize this observation, we introduce a novel and interpretable centroid-oriented boundary distance metric that quantifies the spatial offset between predicted and ground-truth firelines. This metric reveals a consistent uncertainty buffer zone and offers a practical proxy for operational planning. Through feature attribution, we find that vegetation health and recent fire activity are the strongest drivers of predictive confidence, reinforcing the spatial and temporal grounding of model uncertainty. Additionally, we observe that uncertainty zones scale modestly with fire size, suggesting that predictive uncertainty reflects localized ambiguity rather than arbitrary noise.

Our analyses span multiple architectures and model instances, increasing confidence that the observed uncertainty behaviors are not artifacts of a specific model but reflect generalizable patterns tied to fire morphology and feature dynamics. Overall, these results highlight that spatial uncertainty carries interpretable and actionable structure. Rather than being discarded as noise, it can be embraced as a signal—indicating where model hesitation, boundary ambiguity, or further scrutiny may be warranted. As EO-based systems become increasingly operational, such structured uncertainty maps may support more robust, risk-aware wildfire response and decision-making.

Broader Impact Statement

Wildfires represent a growing global threat, exacerbated by climate change, land-use patterns, and increased human activity in fire-prone areas. This work contributes to the development of transparent and trustworthy AI models for high-resolution wildfire forecasting by introducing methods to quantify and spatially interpret predictive uncertainty. Our goal is to improve decision-making for fire managers, emergency responders, and policymakers through spatially calibrated uncertainty estimates that identify where model predictions may be unreliable.

By revealing areas of model hesitation—particularly near fire boundaries and in transitional vegetation zones—our approach may help inform triage strategies, resource allocation, and evacuation planning. The incorporation of uncertainty into operational workflows could reduce overreliance on overconfident predictions, thereby enhancing safety and trust.

However, there are potential risks. Misinterpretation of low uncertainty as a guarantee of safety could be dangerous, especially in out-of-distribution regions. Our models are trained on U.S. fire data and rely on vegetation-based inputs, which may not generalize to other ecosystems. Additionally, we focus on epistemic uncertainty, without modeling aleatoric factors such as sensor noise or unobserved ignitions.

To mitigate these limitations, we recommend human-in-the-loop deployment, transparent communication of model assumptions, and expanded testing across geographies. We also encourage future work to involve diverse stakeholders—including frontline responders and indigenous communities—in evaluating the operational value and limitations of spatial uncertainty maps. Ultimately, this work aims to support safer, more robust wildfire management systems that align with public and environmental benefit.

References

S. Archibald, C. E. R. Lehmann, C. M. Belcher, W. J. Bond, R. A. Bradstock, A.-L. Daniau, K. G. Dexter, E. J. Forrestel, M. Greve, T. He, S. I. Higgins, W. A. Hoffmann, P. M. Holmes, W. M. Jolly, K. P. Kirkman, L. Mucina, S. Pau, I. C. Prentice, P. W. Rundel, and R. J. Williams. Biological and geophysical feedbacks with fire in the earth system. *Environmental Research Letters*, 13(3):033003, 2018. doi: 10.1088/1748-9326/aa9ead.

Andrew Bolt et al. A spatio-temporal neural network forecasting approach for emulation of firefront models. In *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2022.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, 1950.

John Burge, Matthew R Bonanni, R Lily Hu, and Matthias Ihme. Recurrent convolutional deep neural networks for modeling time-resolved wildfire spread behavior. *arXiv preprint arXiv:2210.16411*, 2022.

Alex Cunningham, Sarah Jones, and Ravi Patel. State of wildfires 2023–2024. *Earth System Science Data*, 16:3601–3620, 2024.

Matthias Drusch, Umberto Del Bello, Serge Carlier, Olivier Colin, Victor Fernandez, Ferran Gascon, Bernhard Hoersch, Claudio Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

European Forest Fire Information System (EFFIS). Wildfire season 2025: Early trends and eu response. `https://effis.jrc.ec.europa.eu/`, 2025. Accessed: 2025-05-29.

Mark A Finney. Farsite: Fire area simulator–model development and evaluation. *USDA Forest Service Research Paper RMRS-RP-4*, 1998.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pp. 1050–1059, 2016.

Vivien Sainte Fare Garnot and Loic Landrieu. Satellite image time series classification with temporal attention and joint segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 135–144, 2020.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *arXiv preprint arXiv:2107.07933*, 2022.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. WildfirespreadTS: A dataset of multi-modal time series for wildfire spread prediction. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=RgdGkPRQ03`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Fantine Huot et al. 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.

World Resources Institute. The latest data confirms: Forest fires are getting worse. *WRI Insights*, 2024. Accessed: 2025-05-19.

LACEDC. Impact of 2025 la wildfires. `https://laedc.org/wpcms/wp-content/uploads/2025/02/LAEDC-2025-LA-Wildfires-Study.pdf`, 2025. Accessed: 2025-05-19.

D. Morvan and J.L. Dupuy. Modeling of fire spread through a forest fuel bed using a multiphase formulation. *Combustion and Flame*, 127(1):1981–1994, 2001. ISSN 0010-2180. doi: 10.1016/S0010-2180(01)00302-9. URL `https://www.sciencedirect.com/science/article/pii/S0010218001003029`.

David Radke, Anna Hessler, and Dan Ellsworth. Firecast: Leveraging deep learning to predict wildfire spread. In *IJCAI*, pp. 4575–4581, 2019.

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.

Bryan Shaddy, Deep Ray, Angel Farguell, Valentina Calaza, Jan Mandel, James Haley, Kyle Hilburn, Derek V. Mallia, Adam Kochanski, and Assad Oberai. Generative algorithms for fusion of physics-based wildfire spread models with satellite data for initializing wildfire forecasts. *Artificial Intelligence for the Earth Systems*, 3(3):e2023AIES087, 2024.

I. Shadrin, V. Yadykin, N. Zavyalova, et al. Wildfire spreading prediction using multimodal data and deep neural network approach. *Scientific Reports*, 14(1), 2024.

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*, pp. 802–810, 2015a.

Xingjian Shi et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, pp. 802–810, 2015b.

Matthew P Thompson, David E Calkin, Mark A Finney, Alan A Ager, and Julie W Gilbertson-Day. Risk management: Core principles and practices, and their relevance to wildland fire. *US Department of Agriculture, Forest Service, Rocky Mountain Research Station*, RMRS-GTR-315, 2016.

Cordy Tymstra et al. Development and structure of prometheus: the canadian wildland fire growth simulation model. 2010.