TerraFyre: Spatio-Temporal Transformer Network for Proactive Wildfire Ignition Risk Forecasting from Multispectral Satellite Imagery

Warren M. Xie

Singapore American School
40 Woodlands Street 41, Singapore 738547
xie790031@sas.edu.sg

Abstract

Wildfires are growing in frequency and intensity, demanding predictive models that detect ignition risk early enough to guide mitigation. This work introduces **TerraFyre**, a lightweight spatio-temporal Transformer tailored to 12-frame satellite-image sequences. The architecture couples a Vision Transformer (ViT) that learns local spatial cues with a temporal self-attention stack that captures long-range dependencies across time. Histogram-based inputs (8×32 per frame) are adopted to compress per-band distributions, reduce sensitivity to geolocation jitter and cloud/shadow masking, and reduce I/O volume and computational cost while preserving fuel-state signals over weeks. Experiments use an Indonesian FIRMS test split of 12,073 sequences with 16% positives, reflecting the class imbalance encountered in practice. **TerraFyre** reaches an AUC-ROC of 0.8993 and an AUC-PR of 0.5747 on the held-out test set, surpassing the strongest logistic regression and LSTM baselines on both curves. At a 0.5 decision threshold, the model attains Precision 0.4142 and Recall 0.938. The network trains with batch size 64 on a single GPU while maintaining high AUC scores.

Introduction

Wildfires devastate ecosystems, accelerate climate change, and endanger communities. Timely prediction enables fire agencies to pre-position resources and issue evacuations, yet conventional risk indices (e.g., the Canadian Fire Weather Index (FWI) system) rely on sparse weather stations and hand-crafted thresholds. With widely available active-fire products and moderate-to-high-resolution imagery, modeling short-horizon ignition risk from orbital data is now feasible. Sequence models improve short-horizon prediction, but recurrent networks face long-range dependency limitations and sequential inference costs, motivating a self-attention approach.

TerraFyre, a spatio-temporal Transformer optimized for small multi-spectral image-tile histograms (8×32 per step) and short sequences (12 steps $\approx$ 3 months), unifies patch-level spatial embeddings with temporal self-attention. In short-horizon ignition settings, precursors such as fuel desiccation, live-fuel moisture decline, heat waves, and human activity often accumulate over weekly to subseasonal scales. A 12-week context captures these lagged signals while keeping the receptive field aligned with operational lead times and tractable capacity. The key contributions are:

- **Architecture.** A two-stage design (*PatchEmbedding*→*SimpleViT*→*TemporalTransformer*) unifies local spatial context with global temporal cues.
- **Performance. TerraFyre** attains the top AUC-ROC and AUC-PR on the held-out test set and, at a 0.5 threshold, improves Precision and Recall over the strongest LSTM configuration (B=32) while halving missed events (FN 246 $\rightarrow$ 119).
- **Efficiency.** The architecture trains at batch size 64 on a single GPU while maintaining high AUC scores, enabling high-throughput training without sequential recurrence.

Related Work

Operational fire danger rating systems such as the Canadian Forest Fire Weather Index (FWI) and the Keetch–Byram Drought Index map daily meteorological measurements to scalar risk scores but omit vegetation condition and fine-scale topography, which can yield region-specific bias (Wagner 1987; Keetch and Byram 1968). Related indices and frameworks, including the U.S. National Fire Danger Rating System (2016 update) and the McArthur Forest Fire Danger Index, further illustrate the strengths and limitations of weather-only metrics (Dowdy et al. 2009). Climate analyses also report a global lengthening of fire-weather seasons, reinforcing the need to ingest fuel state and landscape information beyond meteorology (Jolly et al. 2015). This body of work motivates data-driven approaches for short-horizon *ignition* risk rather than long-horizon danger climatology.

The proliferation of moderate to high resolution sensors (e.g., MODIS, Sentinel-1/2, VIIRS) has enabled data-driven pipelines that ingest multispectral imagery and SAR (Drusch et al. 2012; Torres et al. 2012; Hillger et al. 2013). Mature fire products are widely used as supervision or targets: MODIS Collection 6 active fire and burned-area suites, VIIRS 375 m active fire, and associated validations support consistent large-scale labeling (Giglio, Schroeder, and Justice 2016; Giglio et al. 2018; Schroeder et al. 2014; Boschetti et al. 2019; Chuvieco et al. 2019). Beyond official products, deep CNN pipelines have been proposed for

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

burned-area segmentation and near real time mapping from Sentinel-2 and multi-source streams (Knopp et al. 2020; Jiao and Bo 2022), while live fuel moisture can be monitored remotely and incorporated into learning signals (Yebra et al. 2013). These products and measurements underpin labeling schemes that aggregate detections over weekly to subseasonal windows, which aligns with short-horizon ignition risk settings considered here.

A CNN applied to Sentinel-1 SAR time series has tracked wildfire progression in near real time (Ban et al. 2020), and the public *WildfireSpreadTS* dataset enabled supervised next-day spread modeling over 607 U.S. fires (Gerard, Zhao, and Sullivan 2023). Sequence models improve ignition prediction accuracy, for example an LSTM trained on multimodal drivers surpassing an FWI baseline (Kondylatos et al. 2022). Additional spatio-temporal lines include U-Net families for spread and burned-area forecasting and SAR/InSAR progression mapping (Khennou and Akhloufi 2023; Michaelides et al. 2024). Much of this literature targets spread segmentation or per-fire progression; the present work instead targets short-horizon *ignition* risk from compact histogram tiles.

Building on success in Earth observation, *Earthformer* introduced cuboid attention and reported state-of-the-art precipitation nowcasting (Gao et al. 2022). Self-supervised and fully attentional models such as SatMAE and TSViT improve general-purpose spatio-temporal representation learning on satellite time series (Cong et al. 2022; Tarasiou, Chavez, and Zafeiriou 2023). For fire specifically, TeleViT fuses local imagery with global climate teleconnections and improves subseasonal burned-area forecasts (Prapas et al. 2023). Closest to this work, recent efforts move from weather-only "fire danger" toward data-driven fire activity/ignition prediction by integrating fuel and ignition proxies. A notable example is ECMWF's Probability of Fire (PoF) system, which shows that adding fuel and ignition information improves global fire-activity forecasts over weather-only indices (Di Giuseppe et al. 2025; McNorton et al. 2024). These approaches generally address generic satellite forecasting or long-horizon burned-area dynamics, whereas the scope here is short-window ignition risk with a fully self-attentional encoder tailored to compact histogram inputs.

Spatio-Temporal Transformer for Satellite Imagery

Architecture Overview

An input sequence of $T = 12$ satellite frames is denoted $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\} \in \mathbb{R}^{T\times C\times H\times W}$, with C=1, H=8, and W=32. Each frame $\mathbf{X}_t \in \mathbb{R}^{C\times H\times W}$ is divided into $P\times P$ non-overlapping patches (P=2), yielding $N = \frac{HW}{P^2} = 64$ patches. For patch index $p \in \{1, \ldots, N\}$, $\mathbf{x}_{t,p} = \mathrm{vec}\big(\mathbf{X}_t^{(p)}\big) \in \mathbb{R}^{CP^2}$ is flattened and projected via $\mathbf{s}_{t,p} = \mathbf{W}_{\text{patch}}\,\mathbf{x}_{t,p},\ \mathbf{W}_{\text{patch}} \in \mathbb{R}^{d\times CP^2},\ d \in \{64, 128\}$.

Table 1: Building blocks and key hyperparameters of **TerraFyre**.

Component	Function	Parameters
PatchEmbedding	$\mathbf{X}_t \mapsto \mathbf{S}_t^{(0)}$	P=2, $d\in\{64, 128\}$
SimpleViTEncoder	$\mathbf{S}_t^{(0)} \mapsto \mathbf{t}_t$	depth L_s=2, heads h_s=4
TemporalTransformer	$\{\mathbf{t}_t\} \mapsto \mathbf{h}_T$	depth L_t=2, heads h_t=4
MLP Head	$\mathbf{h}_T \mapsto \hat{y}$	$d \to 64 \to 1$

Spatial Stream

A learnable class token $\mathbf{s}_{\text{CLS}} \in \mathbb{R}^d$ is prepended to the patch tokens and combined with 2-D positional encodings $\mathbf{\Pi}_{\text{2D}} \in \mathbb{R}^{(N+1)\times d}$, resulting in

$$\mathbf{S}_t^{(0)} = \big[\mathbf{s}_{\text{CLS}};\, \mathbf{s}_{t,1}; \ldots; \mathbf{s}_{t,N}\big] + \mathbf{\Pi}_{\text{2D}}. \tag{1}$$

$$\mathbf{Q}_t^{(\ell)} = \mathbf{S}_t^{(\ell)}\mathbf{W}_Q^{(\ell)}, \mathbf{K}_t^{(\ell)} = \mathbf{S}_t^{(\ell)}\mathbf{W}_K^{(\ell)}, \mathbf{V}_t^{(\ell)} = \mathbf{S}_t^{(\ell)}\mathbf{W}_V^{(\ell)}. \tag{2}$$

$$\mathcal{A}_t^{(\ell)} = \mathrm{softmax}\left(\frac{\mathbf{Q}_t^{(\ell)}\mathbf{K}_t^{(\ell)\top}}{\sqrt{d/h_s}}\right), \tag{3}$$

$$\mathbf{S}_t^{(\ell+1)} = \mathcal{A}_t^{(\ell)}\mathbf{V}_t^{(\ell)} + \mathrm{FFN}\big(\mathbf{S}_t^{(\ell)}\big), \qquad \ell \in \{0, 1\}. \tag{4}$$

The CLS embedding, after passing through the final Feed-Forward Network (FFN) of the second spatial self-attention layer, $\mathbf{t}_t \triangleq \mathbf{S}_t^{(2)}[0] \in \mathbb{R}^d$, serves as the spatial descriptor for frame t.

Temporal Stream

Stacking descriptors forms $\mathbf{Z} = \big[\mathbf{t}_1; \ldots; \mathbf{t}_T\big] \in \mathbb{R}^{T\times d}$. With temporal encodings $\mathbf{\Pi}_{\text{time}} \in \mathbb{R}^{T\times d}$, $\mathbf{H}^{(0)} = \mathbf{Z} + \mathbf{\Pi}_{\text{time}}$. Two temporal self-attention layers update

$$\mathbf{H}^{(\ell+1)} = \mathrm{MHA}\big(\mathbf{H}^{(\ell)}\big) + \mathrm{FFN}\big(\mathbf{H}^{(\ell)}\big), \quad \ell = 0, 1.$$

Frames are stored most-recent-first; accordingly, the temporal summary is $\mathbf{h}_{\text{last}} = \mathbf{H}^{(2)}[0] \in \mathbb{R}^d$ (first position, most-recent week), which is fed to an MLP–sigmoid head to produce $\hat{y} \in (0, 1)$. For completeness, a temporal [CLS] token or mean-pooling can replace $\mathbf{H}^{(2)}[0]$; unless noted otherwise, results use the first-position summary.

Imbalance-Aware Loss

To counter severe class imbalance, false negatives are penalized exponentially:

$$\mathcal{L}(y, \hat{y}) = \big\|y - \hat{y}\big\|_1 \exp\Big(\alpha\,[y - \hat{y}]\Big), \qquad \alpha = \tfrac{100}{30}. \tag{5}$$

The gradient with respect to $\hat{y}$ is

$$\nabla_{\hat{y}}\mathcal{L} = -\,e^{\alpha(y-\hat{y})}\,\Big(\mathrm{sgn}(y - \hat{y}) + \alpha\,|y - \hat{y}|\Big),$$

which grows rapidly as $\hat{y} \to 0$ when y=1, encouraging high recall. A detailed mathematical analysis of Eq. (5) is provided in the Appendix (complementary material), including

gradient, curvature, conditional risk, precision–recall operating bounds, and guidance on choosing α.

Compared with BCE and class-weighted BCE, the proposed loss increases false-negative penalties exponentially at equal error magnitude while damping confident negatives, which shifts the operating point toward higher recall without tuning class weights. Unlike focal loss, which symmetrically scales cross-entropy by confidence factors, the proposed loss applies a signed exponential tilt that is stronger on hard positives and weaker on easy negatives, improving recall at fixed precision. These properties were chosen to address the severe positive sparsity typical of ignition labels.

Experiments

Dataset

Input–label pairs use a one-year look-back $[t{-}52\text{wk},\ t]$ to predict $[t{+}4\text{wk},\ t{+}5\text{wk}]$ (Yang, Lupascu, and Meel 2021). Weekly features are compiled over 52 weeks; the model uses only the latest 12 ($T{=}12$). Each $8\,\text{km} \times 8\,\text{km}$ tile yields weekly Landsat 7 ETM+ histograms (8 spectral channels: B1–B5, B7, and B6 in low/high gain; the 15 m panchromatic B8 is excluded), with 32 bins per channel. Stacking across channels forms an 8×32 (*bands*$\times$*bins*) 2D array that is treated as a single-channel image. The 12 most recent *valid* weeks ($T{=}12$, stored most-recent-first) are retained; weeks with no valid pixels (orbit gaps, SLC-off) are ignored during weekly aggregation, and tiles with fewer than 12 valid weeks are dropped.

Labels use FIRMS hotspots aggregated to the same $8\,\text{km} \times 8\,\text{km}$ grid and aligned to the Landsat tile layout; the *maximum* confidence in the target window defines a continuous target $y \in [0, 1]$ (Schroeder et al. 2014; Giglio, Schroeder, and Justice 2016), yielding a per-tile map (hereafter *FireLabel*). Tiles are split 8:1:1 (train/val/test) by location, assigning all sequences from a tile to exactly one split; stratified sampling preserves the corpus-level positive prevalence (approximately 16%) across splits. Tensors are stored as $(T, 1, 8, 32)$.

Evaluation Protocol

A tile is labeled positive if the maximum FIRMS hotspot confidence within the $8\,\text{km} \times 8\,\text{km}$ tile during the target window $[t{+}4\text{wk},\ t{+}5\text{wk}]$ is at least 0.5; otherwise it is labeled negative. This binarized ground truth is used for all confusion-matrix–based metrics. For AUC-ROC and AUC-PR, continuous predictions $\hat{y} \in (0, 1)$ are evaluated against the binary ground truth across all thresholds. Unless otherwise noted, predicted probabilities are thresholded at 0.5 when reporting Accuracy, Precision, Recall, F1, and confusion-matrix counts. Threshold-dependent metrics reflect the chosen operating point and the associated application trade-offs.

Baselines

Logistic regression (LogReg). For each cell, the per-timestep (weekly) histogram mean and variance are concatenated to form a $T \times 2$ feature vector, which is then flattened and passed to a logistic classifier.

LSTM. Following Yang, Lupascu, and Meel (2021), every frame is embedded with a single $3{\times}3$ convolution (32 filters) followed by $\tanh$ activation and global average pooling, yielding a length-32 vector. A one-layer LSTM with 64 hidden units processes the $T{=}12$ vectors and outputs an ignition-risk score through a linear-sigmoid head.

Both baselines are trained for 50 epochs with batch sizes $B \in \{1, 2, 4, 8, 16, 32, 64\}$, Adam optimizer and an initial learning rate of 10^{-3}, decayed by 0.5 on plateau.

Hyperparameter Selection and Final Training

A 2×2 grid is conducted over the embedding dimension $d \in \{64, 128\}$ and learning rate $\eta \in \{10^{-4},\ 5 \times 10^{-5}\}$. Each candidate is trained for 5 warm-up epochs on the training split and evaluated on the validation split. The configuration with the highest F_1 on validation, ($d{=}128,\ \eta{=}10^{-4}$), is then re-trained on the union of train and validation for a fixed 50 epochs without early stopping, and the test split is evaluated once. Early stopping with a patience of 10 is used only during hyperparameter selection.

Training recipe and regularization. Unless otherwise specified, optimization uses Adam with default $(\beta_1, \beta_2) = (0.9, 0.999)$, no weight decay, mini-batch size 64, and dropout $p = 0.1$ inside Transformer layers. The MLP head has no additional dropout. All runs use the same random seed, deterministic CuDNN settings, and per-worker seeding for data loaders. The learning-rate choices and early-stopping policy follow the validation protocol described above; no learning-rate schedule is used during the final 50-epoch retraining (**TerraFyre** fixed LR; baselines plateau decay to mitigate small-B underfitting).

Batch positive-absence probability. With a positive rate $p \approx 0.16$, the probability that a mini-batch of size B contains no positives is $(1 - p)^B$: $B{=}8 \to 0.248$, $16 \to 0.061$, $32 \to 3.8{\times}10^{-3}$, $64 \to 1.4{\times}10^{-5}$, explaining improved stability at larger B under the recall-tilted loss.

Results

Table 2 summarizes test performance. **TerraFyre** achieves the highest curve-based metrics on the held-out Indonesian FIRMS test set, with AUC-ROC of 0.8993 and AUC-PR of 0.5747. At an operating threshold of 0.5, **TerraFyre** attains Precision of 0.4142 and Recall of 0.9384 (FN = 119). Threshold-dependent metrics such as Precision, Recall, Accuracy, and confusion-matrix counts vary with the chosen operating point, whereas AUC-ROC and AUC-PR are threshold-independent and therefore provide the primary basis for comparison.

Relative to the strongest LSTM configuration (batch size 32), Precision increases by 5.2 percentage points (0.3623 $\to$ 0.4142) and Recall increases by 6.6 percentage points (0.8727 $\to$ 0.9384), reducing missed events from 246 to 119 at the same 0.5 decision threshold. High AUC is retained at batch size 64 on a single GPU, indicating favorable scaling at this operating point.

To assess qualitative differences among modeling strategies, a visualization was conducted comparing reconstructed

Table 2: Test performance. For **LogReg/LSTM**, the best batch size is selected on **validation** F_1 over $B \in \{1, 2, 4, 8, 16, 32, 64\}$; the chosen setting is retrained on train(+val) and evaluated once on the held-out test split. **TerraFyre** is reported at B=64 due to small-batch instability under the recall-tilted loss.

Model	TP	TN	FP	FN	Acc	Prec.	Rec.	AUC-ROC	AUC-PR
LogReg (best-batch)	1481	**7984**	**2156**	452	**0.7840**	0.4072	0.7662	0.7794	0.3525
LSTM (best-batch)	1687	7171	2969	246	0.7337	0.3623	0.8727	0.8600	0.5275
TerraFyre (B=64)	**1814**	7574	2566	**119**	0.7776	**0.4142**	**0.9384**	**0.8993**	**0.5747**

fire labels with the outputs of three representative models: logistic regression, LSTM, and a Spatio-Temporal Transformer. Figure 1 shows side-by-side comparisons for five FIRMS–Landsat scenes: `LE07_113061` (2019-08-19), `LE07_114063` (2019-08-26), `LE07_115061` (2019-08-01), `LE07_115062` (2019-08-01), and `LE07_115066` (2019-08-01). Columns are ordered as reconstructed FireLabel, logistic regression, LSTM, and Spatio-Temporal Transformer. Inference used *batch size 1* for logistic regression, *batch size 32* for the best-performing LSTM checkpoint, and *batch size 64* for the Spatio-Temporal Transformer. All predictions are normalized to the $[0, 1]$ range, while the reconstructed FireLabel uses a global min–max scale across scenes for consistent visual contrast.

Under this common scaling, the logistic regression model produces spatially smooth yet overly generalized heatmaps, missing localized fire-prone hotspots that appear in the reconstructed reference. The limitation is consistent with a linear decision surface applied to tile-level features without temporal context, which tends to dilute sharp boundaries.

The LSTM-based model (batch 32) incorporates sequential dependencies across image tiles and timesteps, yielding higher sensitivity in regions where temporal accumulation is informative. Nevertheless, the model still struggles to resolve fine-grained spatial heterogeneity, and predicted boundaries remain less crisp when local texture is decisive. The LSTM maps appear broadly warm because the 3×3 convolution + global average pooling compresses spatial detail per tile and the recurrent state propagates elevated activations across weeks. With imbalanced labels, BCE-trained LSTMs trade precision for recall, which combined with min–max visualization, manifests as widespread red regions.

The Spatio-Temporal Transformer (batch 64) most closely aligns with the reconstructed FireLabel by jointly modeling local spatial structure and long-range temporal relations via self-attention. Predictions exhibit sharper transitions and better localization of high-risk patches. The comparison indicates that explicitly modeling spatio-temporal interactions improves fidelity over conventional baselines, shifting the operating point toward better localization without sacrificing global consistency.

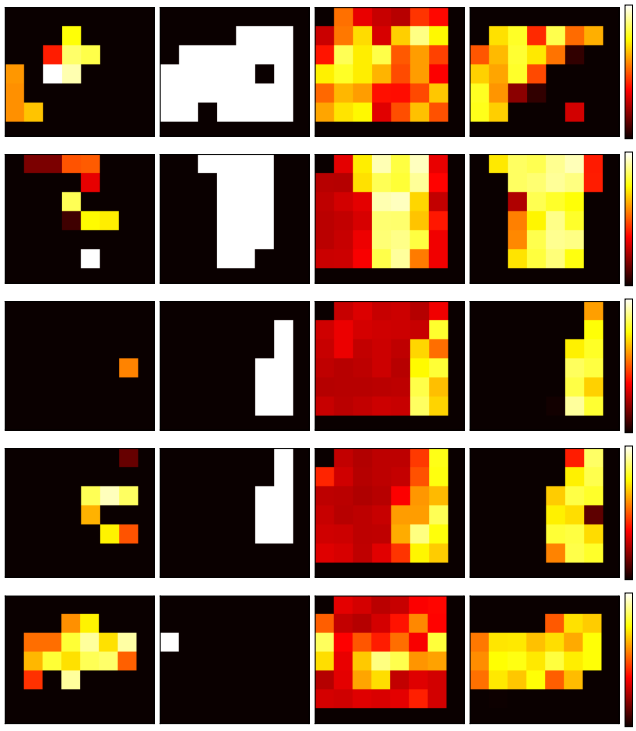

Figure 1: Reconstructed fire labels and model inference comparisons on multiple Landsat satellite scenes from the FIRMS dataset. The first column shows the continuous-valued FireLabel maps reconstructed from reference data, while the second through fourth columns present the predicted fire risk distributions from three approaches in order: a baseline logistic regression model, an LSTM-based temporal model, and a spatio-temporal Transformer. All predictions are normalized to the $[0, 1]$ range and visualized using a hot colormap.

Conclusion

TerraFyre, a lightweight spatio-temporal Transformer, surpasses logistic-regression and LSTM baselines on AUC-based wildfire-ignition metrics in the 12-step 8×32 Indonesian FIRMS setting. It achieves the best AUC-ROC and AUC-PR, reduces missed events relative to the strongest LSTM, and serves as a compact spatio-temporal baseline that can be paired with stronger backbones. Current limitations—regional generalization gaps, tile spillover, and label biases—will be addressed via region-adaptive training and leakage-robust splits. Future work includes higher-resolution inputs, deeper architectures benchmarked against stronger spatio-temporal models, and integration of meteorological priors and other modalities, together with uncertainty estimation and cross-biome adaptation toward operational deployment.

References

Ban, Y.; Zhang, P.; Nascetti, A.; Bevington, A. R.; and Wulder, M. A. 2020. Near real-time wildfire progression monitoring with Sentinel-1 SAR time series and deep learning. *Scientific reports*, 10(1): 1322.

Boschetti, L.; Roy, D. P.; Giglio, L.; Huang, H.; Zubkova, M.; and Humber, M. L. 2019. Global validation of the collection 6 MODIS burned area product. *Remote sensing of environment*, 235: 111490.

Chuvieco, E.; Mouillot, F.; Van der Werf, G. R.; San Miguel, J.; Tanase, M.; Koutsias, N.; García, M.; Yebra, M.; Padilla, M.; Gitas, I.; et al. 2019. Historical background and current developments for mapping burned area from satellite Earth observation. *Remote Sensing of Environment*, 225: 45–64.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

Di Giuseppe, F.; McNorton, J.; Lombardi, A.; and Wetterhall, F. 2025. Global data-driven prediction of fire activity. *Nature Communications*, 16(1): 2918.

Dowdy, A. J.; Mills, G. A.; Finkele, K.; and De Groot, W. 2009. Australian fire weather as represented by the McArthur forest fire danger index and the Canadian forest fire weather index.

Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120: 25–36.

Gao, Z.; Shi, X.; Wang, H.; Zhu, Y.; Wang, Y. B.; Li, M.; and Yeung, D.-Y. 2022. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35: 25390–25403.

Gerard, S.; Zhao, Y.; and Sullivan, J. 2023. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. *Advances in Neural Information Processing Systems*, 36: 74515–74529.

Giglio, L.; Boschetti, L.; Roy, D. P.; Humber, M. L.; and Justice, C. O. 2018. The Collection 6 MODIS burned area mapping algorithm and product. *Remote sensing of environment*, 217: 72–85.

Giglio, L.; Schroeder, W.; and Justice, C. O. 2016. The collection 6 MODIS active fire detection algorithm and fire products. *Remote sensing of environment*, 178: 31–41.

Hillger, D.; Kopp, T.; Lee, T.; Lindsey, D.; Seaman, C.; Miller, S.; Solbrig, J.; Kidder, S.; Bachmeier, S.; Jasmin, T.; et al. 2013. First-light imagery from Suomi NPP VIIRS. *Bulletin of the American Meteorological Society*, 94(7): 1019–1029.

Jiao, L.; and Bo, Y. 2022. Near real-time mapping of burned area by synergizing multiple satellites remote-sensing data. *GIScience & Remote Sensing*, 59(1): 1956–1977.

Jolly, W. M.; Cochrane, M. A.; Freeborn, P. H.; Holden, Z. A.; Brown, T. J.; Williamson, G. J.; and Bowman, D. M. 2015. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature communications*, 6(1): 7537.

Keetch, J. J.; and Byram, G. M. 1968. *A drought index for forest fire control*, volume 38. US Department of Agriculture, Forest Service, Southeastern Forest Experiment Station.

Khennou, F.; and Akhloufi, M. A. 2023. Improving wildland fire spread prediction using deep U-Nets. *Science of Remote Sensing*, 8: 100101.

Knopp, L.; Wieland, M.; Rättich, M.; and Martinis, S. 2020. A deep learning approach for burned area segmentation with Sentinel-2 data. *Remote Sensing*, 12(15): 2422.

Kondylatos, S.; Prapas, I.; Ronco, M.; Papoutsis, I.; Camps-Valls, G.; Piles, M.; Fernández-Torres, M.-Á.; and Carvalhais, N. 2022. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17): e2022GL099368.

McNorton, J. R.; Di Giuseppe, F.; Pinnington, E.; Chantry, M.; and Barnard, C. 2024. A global probability-of-fire (PoF) forecast. *Geophysical Research Letters*, 51(12): e2023GL107929.

Michaelides, R. J.; Siegfried, M. R.; Lovekin, J.; Berry, K.; Dugan, B.; and Roth, D. L. 2024. Wildfire Progression Time Series Mapping With Interferometric Synthetic Aperture Radar (InSAR). *IEEE Geoscience and Remote Sensing Letters*, 21: 1–5.

Prapas, I.; Bountos, N.-I.; Kondylatos, S.; Michail, D.; Camps-Valls, G.; and Papoutsis, I. 2023. Televit: Teleconnection-driven transformers improve subseasonal to seasonal wildfire forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3754–3759.

Schroeder, W.; Oliva, P.; Giglio, L.; and Csiszar, I. A. 2014. The New VIIRS 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143: 85–96.

Tarasiou, M.; Chavez, E.; and Zafeiriou, S. 2023. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10418–10428.

Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. 2012. GMES Sentinel-1 mission. *Remote sensing of environment*, 120: 9–24.

Wagner, C. v. 1987. *Development and structure of the Canadian forest fire weather index system*. Number 35 in Forestry Technical Report. Canadian Forestry Service, Headquarters, Ottawa.

Yang, S.; Lupascu, M.; and Meel, K. S. 2021. Predicting forest fire using remote sensing data and machine learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14983–14990.

Yebra, M.; Dennison, P. E.; Chuvieco, E.; Riaño, D.; Zylstra, P.; Hunt Jr, E. R.; Danson, F. M.; Qi, Y.; and Jurdao, S. 2013. A global review of remote sensing of live fuel moisture content for fire danger assessment: Moving towards operational products. *Remote Sensing of Environment*, 136: 455–468.

Appendix A: Qualitative Analysis of Spatio-Temporal Attention

Visualization Setup and Random Sampling

Figure S2 visualizes **TerraFyre**'s temporal and spatial self-attention on two *randomly sampled* examples from the held-out test set: one non-fire sequence (top row) and one fire sequence (bottom row). For each example, the first column shows the one-dimensional temporal self-attention of the most recent week over the past 12 weeks, and the second column shows the full 12×12 temporal self-attention matrix. The third column depicts the spatial self-attention map over the 8×32 histogram for the most recent week, obtained from the CLS-to-patch attention scores of the spatial transformer. The fourth column displays the corresponding input histogram itself. All attention maps are averaged over layers and heads for readability.

Non-Fire Example

The top row of Figure S2 corresponds to a non-fire sequence and illustrates a broadly distributed temporal and spatial focus. In the left-most panel, the temporal self-attention of the most recent week is relatively flat across the 12-week window, with only a modest elevation around weeks 5–7. This pattern indicates that, in the absence of fire activity, many weeks contribute comparable amounts of evidence, and the model effectively aggregates long-term temporal context rather than relying on a single "critical" week. The adjacent 12×12 temporal self-attention matrix reinforces this interpretation: a strong diagonal band appears, showing that each week attends mostly to itself and its immediate neighbors, which is consistent with smooth, slowly varying dynamics and the use of local temporal continuity as a cue for normal behavior. In the spatial domain, the third panel shows that attention is not uniform across the 8×32 histogram. Several narrow bands in the low-to-mid spectral range receive moderate attention, while the remaining bins are comparatively suppressed, suggesting that **TerraFyre** monitors a subset of spectro-spatial regions as characteristic signatures of normal scenes. The final panel reveals that the actual input histogram for this week exhibits strong energy primarily in those same bands, with large parts of the histogram nearly empty; the overlap between high-energy bins and attention hotspots suggests that the model selectively focuses on physically meaningful parts of the spectrum while largely ignoring empty or noisy regions in non-fire conditions.

Fire Example

The bottom row of Figure S2 shows a fire sequence and exhibits a markedly more selective attention pattern. In the temporal 1D bar plot, the most recent week assigns a disproportionately large fraction of its attention to one or two specific weeks, while many other weeks receive only small weights. This sharply peaked profile implies that, when fire activity is present, the prediction is anchored on a small number of "event weeks" where the signal changes abruptly, rather than on a diffuse average over the entire 12-week history. The corresponding temporal self-attention matrix displays a distinct vertical band in addition to the diagonal trend: many different query weeks attend strongly to the same key week, indicating the emergence of a global reference week against which the rest of the sequence is interpreted. In the spatial attention map, attention becomes even more localized than in the non-fire example. A few patches in the 8×32 grid form bright hotspots that absorb most of the attention mass, while the remaining bins receive only weak attention. The right-most panel shows that these hotspots coincide with extremely high-energy regions of the input histogram, where fire-related signals dominate. This alignment suggests that, for fire sequences, **TerraFyre** focuses on localized, high-intensity spectro-spatial patterns that are characteristic of fire activity and uses them as primary evidence for a positive prediction.

Comparative Discussion

A direct comparison between the two rows in Figure S2 highlights class-dependent attention strategies that offer an intuitive explanation for **TerraFyre**'s predictions. For the non-fire case, temporal self-attention is broadly distributed across the 12-week window, and the corresponding attention matrix is dominated by a smooth diagonal band with only mild off-diagonal structure. This behavior is consistent with an interpretation in which the model checks for consistency over time: similar weeks reinforce each other via mutual attention, and the absence of a strongly emphasized event week reflects the lack of abrupt changes. Spatially, attention in the non-fire example concentrates on a small number of bands that carry most of the energy but still spreads across adjacent bins, suggesting that the model monitors a region rather than a single point in the spectrum and remains robust to minor shifts or noise.

In contrast, the fire example activates a highly selective mode. Temporally, attention collapses onto one or a few weeks, and the attention matrix exhibits strong vertical and horizontal bands centered on these weeks, indicating that they serve as anchors for interpreting the entire sequence. This difference implies that **TerraFyre** does not apply a uniform temporal pooling strategy; instead, it switches from a "distributed evidence" mode in non-fire conditions to an "event-centric" mode when fire signals are present. Spatially, the fire example shows attention hotspots that are both more intense and more localized than in the non-fire case, tightly aligned with the bins of highest energy in the histogram. This contrast suggests that **TerraFyre** distinguishes fire from non-fire not merely by the overall level of activity but by detecting specific spectro-spatial patterns that are rare under normal conditions. The figure therefore provides a qualitative but concrete rationale for the model's ability to separate fire and non-fire sequences based on how it allocates attention in both temporal and spatial dimensions.

Limitations and Future Work

The analysis in this section is intentionally qualitative and restricted to two randomly chosen examples, and all visualizations rely on attention weights averaged over layers and heads. This design facilitates interpretation but may conceal layer-specific roles, head specialization, or rare but important behavior that only appears in particular regions or

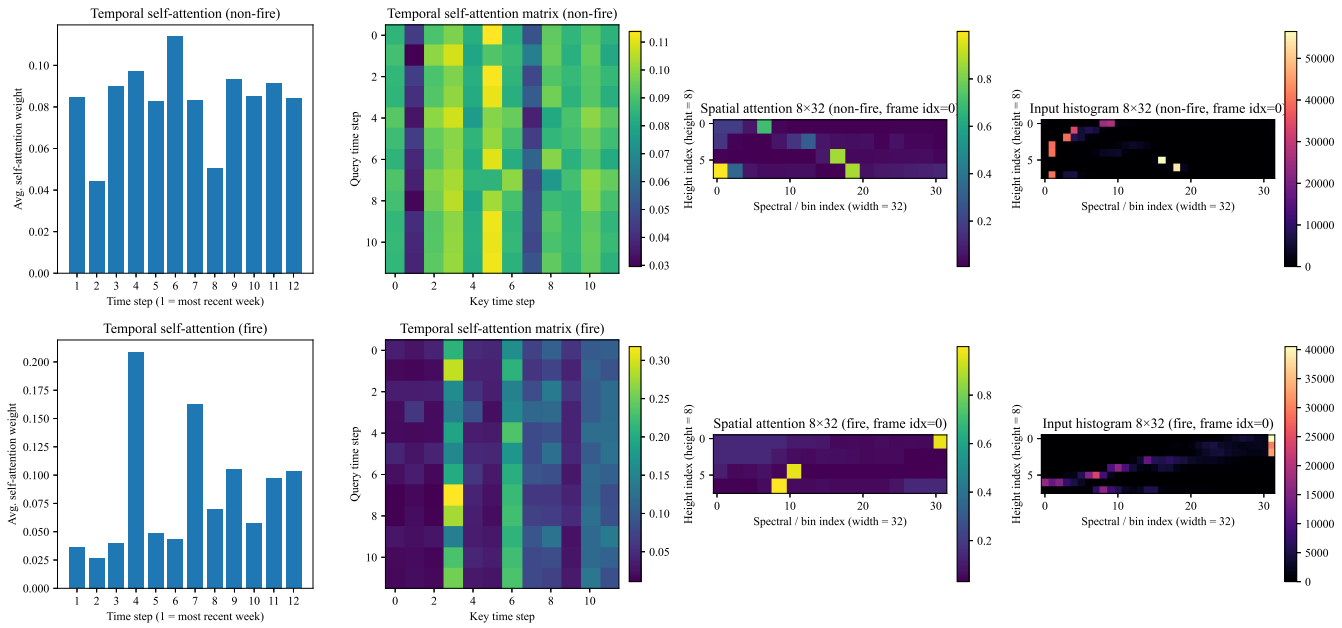

Figure S2: Temporal and spatial self-attention patterns of **TerraFyre** for two randomly sampled test examples: a non-fire case (top row) and a fire case (bottom row). The first column shows the 1D temporal self-attention of the most recent week over the 12-week history, the second column visualizes the full $T \times T$ temporal self-attention matrix, the third column plots the CLS-to-patch spatial self-attention over the 8×32 histogram of the most recent week, and the fourth column displays the corresponding input histogram, enabling direct comparison between attention hotspots and the underlying signal energy.

seasons. A more exhaustive study could include layer-wise and head-wise inspection of attention patterns, statistical aggregation over large numbers of samples to obtain confidence intervals for class-conditional attention distributions, and comparisons across different geographic regions, land-cover types, or sensor configurations. Extending the analysis to additional interpretability tools, such as integrated gradients, SHAP, or counterfactual perturbations applied separately along the temporal and spectral axes, would further clarify which features drive **TerraFyre**'s decisions and how robust those decisions are to distribution shifts. Such in-depth investigations are left for future work.

Appendix B: Meteorological Integration and Deployment

This section sketches potential extensions of **TerraFyre** to incorporate meteorological information as well as a simple operational deployment scenario. In addition to multispectral histogram inputs, tile-level meteorological features can be extracted from global reanalysis or numerical weather prediction products (e.g., ERA5, GFS) and aggregated to the same weekly cadence as the 12-step input sequence. For each tile and week, a compact weather feature vector (e.g., summarizing temperature, humidity, wind speed, and rainfall) can then be combined with the existing spatio-temporal representation in two main ways. One option is to append the weekly weather feature vector as an additional time token in the temporal Transformer, allowing the model to attend jointly over optical and meteorological signals along the sequence. An alternative is to construct a separate MLP or Transformer branch that processes the sequence of weekly weather vectors into a dedicated meteorological embedding, which is then fused with the **TerraFyre** representation via late fusion (e.g., concatenation followed by a small fusion MLP at the sequence or tile-embedding level). A practical starting point is to restrict the weather vector to a small set of physically motivated covariates (e.g., weekly mean or maximum temperature, cumulative rainfall, mean relative humidity, and wind speed) and a low-dimensional embedding (e.g., 8–32 channels), postponing more expressive cross-modal attention mechanisms to future large-scale experiments.

A corresponding operational scenario envisions **TerraFyre** being run on a regular schedule (e.g., every 3–7 days) as new satellite and meteorological data become available. For each update, the model produces ignition-risk scores for all monitored tiles in a region containing on the order of 10^3–10^5 tiles, and the top $k\%$ highest-risk tiles (typically yielding tens to a few hundred tiles per update when $k \in [1, 5]$) are highlighted and passed to downstream early-warning or dispatch systems. These tiles can be visualized on standard geospatial dashboards so that agencies can rapidly inspect emerging hotspots and focus limited field resources on a small, high-risk subset of the domain. Decision thresholds for alerting can be selected on the basis of validation PR and ROC curves, using operating points that prioritize high recall of true ignition events (e.g., targeting 90–95% recall, even at the cost of lower precision), in line with the objectives of conservative early-warning systems.