FIREFUSION: LEVERAGING MULTI-MODAL REPRESENTATIONS WITH SPATIOTEMPORAL MODELING FOR WILDFIRE IGNITION MODELING

A PREPRINT

Tanner O'Rourke *
Department of Natural Sciences
University of Texas, Austin
two297@my.utexas.edu

December 1, 2025

ABSTRACT

Wildfire ignition remains exceptionally difficult to model due to its rarity, inconsistent heterogeneous physical and human drivers, and complex multi-scale dynamics. FireFusion addresses these challenges with a ConvFormer-based spatiotemporal framework that integrates leading geographical, meteorological, and human risk factors with derived representations. The system applies domain-informed feature engineering, terrain and active-fire masking, and multi-dimensional attention to avoid non-causal predictors and improve shared layer representations beyond standard CNN approaches. FireFusion leverages multi-year high-resolution datasets to achieve improved spatial fidelity over baselines reliant on minimally processed static features, demonstrating that pairing domain-informed causal analysis with cross-dimensional learning materially strengthens next-day ignition and cause forecasting.

Keywords Wildfire ignition · spatiotemporal modeling · multi-class modeling · physics-aware deep learning

1 Introduction

Wildfires continue to threaten communities, infrastructure, and ecosystems worldwide, with suppression costs, property loss, and smoke-related health complications all rising along with the warming climate[Moritz, 2014]. In the western United States, particularly in Washington State, recent seasons such as 2020 have featured record smoke episodes, elevated fine-particulate exposures, and a measurable increase in mortality and hospitalizations during high ignition-rate years [Liu, 2021]. These trends make high-resolution short-horizon ignition forecasting a critical component of early-warning systems in regions where human activity, wildland–urban interfaces (WUIs), and variable terrain intersect.

Modeling wildfire ignition remains one of the most challenging problems in physics-aware modeling. Ignitions are exceedingly rare in relation to the number of potential burnable locations and days, leading to severe class imbalance and sparse positive labels even in multi-decade datasets Caron et al. [2025]. At the same time, predicting fire ignition relies on sparse relationships between meteorology, climatology, topography, and human factors, where a complex feature space makes ground truth risk factors hard to ascertain. Regional assessments further indicate that roughly 85–90% of recent wildfires in Washington are human-caused, with ignitions extending west of the Cascades into densely populated corridors around Seattle ([Balch et al., 2017]). This introduces highly nonlinear and spatially variate ignition pressures (recreational patterns, transportation corridors, industrial activity, and socio-economic patterns all shape where fires start). As a result, the majority of ignitions cannot predicted from meteorology and fuels alone.

Recent summaries from state agencies indicate that roughly 85-90% of Washington wildfires in recent years are human-caused and increasingly distributed across "all of Washington," not just historically fire-prone eastern counties

*This manuscript is a preprint prepared as part of graduate research. It has not undergone formal publication.

Edgeley et al. [2025]. Moreover, recent deep learning reviews for wildfire risk prediction emphasize that models must capture both high-dimensional multivariate structure, temporal evolution, localized dynamics, and longer horizon predictors to be effective, and that naive predictors often under perform or poorly generalize when using raw metrics or facing human factors.Xu [2025a]. Despite this progress, there is little work done explicitly comparing these various feature classifications in high-dimensional latent space in a single end-to-end model. This framework aims to address these challenges by combining domain-informed feature engineering, physics-informed normalization, and a mix of convolutional and cross-dimensional attention to embed a unified ignition-prediction system.

1.1 Recent Work

A wide range of efforts underscore the rich modality dynamics of predicting fire risk, exploring data representations and modeling paradigms across time scales, purpose, and field.

Fire Weather Indices Fire Weather indices such as the Fire Weather Index (FWI), Keetch Byram Drought Index (KBDI) compress observations of temperature, humidity, precipitation, and wind into a small number of scalar "danger metric" Wagne [1987]. These quantify meteorological conditions that can lead to fire and spread, communicating fire danger levels to emergency and fire fighting management FS [2025]. The Fosberg Fire Weather Index (FFWI) Fosberg [1978], developed in Canada in 1978, is one of the most widely used wildfire risk assessment tools today. These indices remain indispensable for regional fire-danger rating and fire management, but they assume relatively static fuel and climatological regimes, treat space coarsely, and largely discard non-meteorological drivers such as detailed topography, fuel composition, and human factors.

Statistical and Machine-Learning Models Earth-system machine learning such as EarthFormer, which uses “cuboid attention” over both space and time to model long-range temporal dependencies in climate and weather data, emphasize finding low-dimensional representations of high-dimensional physical fields Gao et al. [2022]. These methods often treat ignition as a simple pixel-classification task, and while historically extremely accurate given particular constraints, they neglect physical constraints such as water bodies, non-burnable terrain, or the requirement that ignitions occur only where no active fire currently exists.

Spatiotemporal Modeling A second category, more broadly characterized by weather forecasting, uses machine learning directly on gridded environmental, often static-land cover maps, to frame wildfire ignition as a pixel-wise classification problem over sequences in time to frame. These models explicitly emphasizes spatiotemporal variation features (differences over time), giving large gains over CNN baseline Pan et al. [2024]. They leverage convolution or hybrid CNN architectures over remote-sensing imagery, encode spatial fields into compact tokens, apply attention-based processors over the latent representations, and decode back to full-resolution, enabling long-horizon forecasts at reduced computational cost.

These approaches highlights that feature integration and domain-aware constraints are often ad hoc, and that models may be brittle when extrapolated beyond their training climatology Caron et al. [2025]. Scientifically however, when stacked up against more popularized recognition CNN’s such as ResNet He et al. [2015], Convolutional LSTMs, and Transformer hybrids, its proven that the best models use convolutional encoders to compress the spatial field before then applying attention/temporal modeling Michail et al. [2024].

Human-sensing and socio-physical modeling Parallel work on social-physics and human-sensor networks shows that incorporating human activity signals — such as social media and crowd-sourced reports - can improve situational awareness and now casting of wildfire behavior Lever et al. [2023]. Social-physics models such as Sentimental Wildfire treat social media users as "human sensors" whose posts provide noisy, localized observations of fire presence and impacts, and demonstrate that combining sentiment features with geophysical attributes improves real-time of wildfire attributes Lever [2022], with more recent work on human sensor networks aims to fuse vision-based observations, social data, and remote sensing into assimilated wildfire estimates in near real time Xu [2025b]. However, these efforts are typically decoupled from high-resolution spatiotemporal Earth-system models.

1.2 Limitations of Current Approaches

- Static Ignition Predictors: Various fire-weather indices simplify ignition physics into fixed formulas of potent ignition drivers (e.g., wind, relative humidity) yet ignore spatial variation in surface topology and fuels as well as non-linear interactions with conditions as a whole.
- Low modality coverage: Most CNN-based ignition models rely on lower-dimensional representational imagery when compared to typical weather forecasting, failing to encode spatial texture where ignition is physically possible, mask unphysical predictors such as water coverage, or utilize temporal conditioning.

1.3 Study Area

Washington State as of 2016 contains 9.4 billion trees and 22.5 million forested acres, covering half the state's land area, 470,000 acres of which were affected by fires Palmer et al. [2016]. Topographically, the region features a variety of maritime and continental climates and a strong polarity in climate between Eastern (wet, cool, maritime) and Western (hot, dry, continental) Washington (See Fig 1)1. Strong precipitation and wind gradients, diverse fuel regimes from coastal forests to shrub-steppe, and a rapidly expanding wildland–urban interface (WUI), early continental snow melt, and drought-intensified fire seasons on the eastern-side of the Cascade Range create a setting where ignition risk is both high and strongly mediated by human behavior.

Analyses of ignition patterns in the Pacific Northwest show that human ignitions dominate total event counts - over 100,000 ignitions from 1992–2018 across Oregon and Washington - with clear spatial and socioeconomic structure in where fires start Reilley et al. [2023]. Consensus exists around the dependence of many of Washington's forest ecosystems on fire to maintain forest health, However, uncontrolled wildfires can result in loss of timber value, changes in habitat, and major threat to infrastructure and loss of life. These factors cmbine to make the region an unfortunate hotbed for dynamic fire weather modeling.

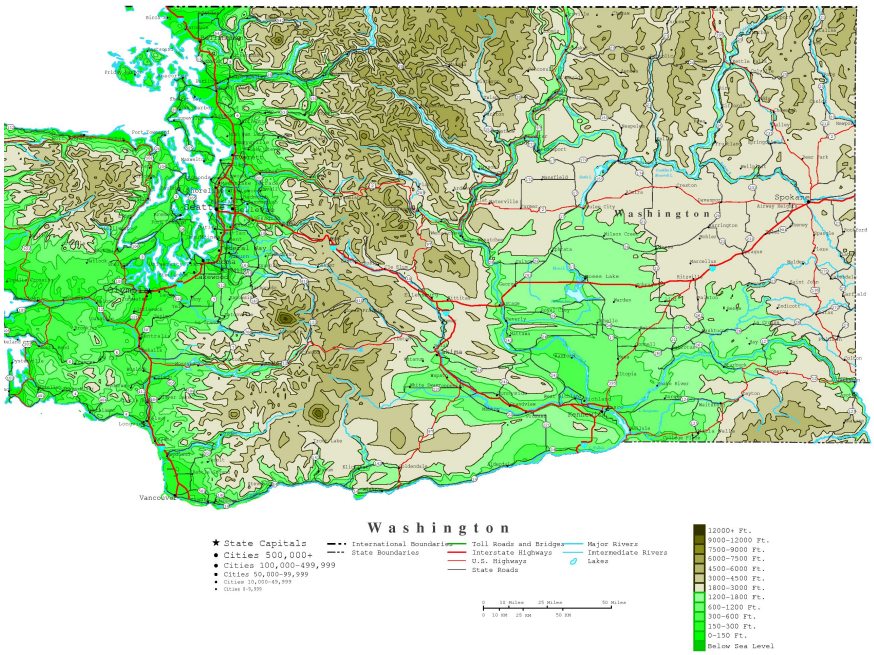


Figure 1: Topographic map of Washington State showing elevation and major roadways. Source: YellowMaps

1.4 Key Contributions

FireFusion aims to unify high-dimensional environmental, meteorological, and human-driven predictors with spatiotemporal modeling

Within the scope of high-resolution, next-day ignition forecasting in Washington State, FireFusion aims to unify domain-informed feature engineering, utilizing fire weather indices, derived meteorological metrics and wild land classifiers, with a ConvFormer-based spatiotemporal architecture and physics-aware masking. The model additionally introduces physically motivated masking (water, N/A regions, and active-fire ignition and masking) for directed prediction. Collectively, these contributions demonstrate:

- **A ConvFormer-based spatiotemporal architecture tailored to ignition and cause modeling** FireFusion adopts a convolutional encoder followed by windowed spatial attention, channel-mixing attention, and temporal attention to jointly model spatial representations, cross-feature interactions, and temporal evolution. A bi-head decoder produces both ignition probability and ignition cause labels on the same fine-resolution grid, framing ignition prediction as a coupled multi-class problem rather than a single binary hazard score.
- **Physics-aware training and inference masking** The framework incorporates look-ahead masking of water features, non-burnable terrain, missing data regions, and active fires directly into the loss and output space.

This constrains gradients and predictions to *physically plausible locations and times* (for e.g., no ignitions over water or where fire is already burning), reducing spurious learning from non-causal correlations that can arise when training on raw gridded fields.

- **Multi-modal learned representations** By pairing domain-engineered and derived features attention over space, features, and time, as well as more concretely in definitive subclasses, the model learns shared representations across meteorological, fuel, and human-related channels without relying solely on raw, minimally processed inputs or on hand-crafted indices. This design effectively reduces the dimensionality *both within the features themselves, as well as through a multivariate structure*.
- **Multi-axis evaluation** The model evaluates ignition occurrence, ignition cause classes, and spatial pattern fidelity, showcasing ability to capture high-risk corridors (e.g., transportation routes and WUI-adjacent regions) and generalize across inter annual variations in weather and activity.

2 Feature Engineering

FireFusion is trained on a geospatial stack combining static landscape properties, daily meteorology, vegetation state and fuel markers, and multi-source fire labels into a set ignition risk priors. Rather than feeding raw gridded products directly into the model, we design a compact, derived feature set that encodes fire-relevant structure while remaining interpretable. This section describes how input sources were selected, raw data was acquired and harmonized into a common spatiotemporal cube, and how raw predictors were transformed into derived proxies that better reflect ignition processes.

2.1 Development of Input Sources

Feature selection conceptually revolved around five categories: (i) fuels and vegetation, (ii) weather and climate, (iii) topography and static landscape form, (iv) human presence and access, and (v) historical fire activity. For each group, candidate products were considered that have long-term coverage over the study domain, sufficient spatial resolution to resolve mesoscale fuel and human gradients, and licensing and documentation that support reproducible use.

In practice, all inputs were organized into three abstractions:*datacubes*, *datasets*, and *data arrays*. A datacube acts as specialized for spatiotemporal data with explicit (time, y, x) coordinates on a regular grid Karasante1 et al. [2023]. A dataset maps one-to-one with a particular source of data, containing one or more data arrays (for example, temperature, humidity, and wind from gridMET). Individual variables within a cube are represented as data arrays, which may be manipulated independently or combined to form derived features. Additionally, datasets and their data arrays are loaded into the datacube in particular sequential order such that features derived from 1 or more features are pre-computed in a previously loaded dataset or data array within a dataset (for example, months since last burn requires dates of burns to be computed). This structure lets us treat simlilar sources (gridded rasters, vector perimeters, point occurrences) within a common spatiotemporal framework and reduce redundant downstream operations such as masking, interpolation, and stacking.

The final set of input sources was chosen to ensure complementary coverage across the categories: static topography (Landfire); impervious-surface fractions and general land-cover classes (NLCD) provide fuel type and urbanization; gridded meteorology (gridMET) captures daily variability in temperature, humidity, precipitation, and wind; vegetation indices and canopy metrics from MODIS encode seasonal fuel condition; population density, wildland–urban interface, and road networks (GPW, WUI, TIGER/Line Census) capture human pressures and potential ignition corridors; and multi-source fire products (USFS ignition points and perimeters, MODIS burned area) provide observed fire outcomes. These sources are summarized in Table 1.

2.2 Data Acquisition

All products were acquired from their official repositories using scripted pipelines and assembled into a common spatiotemporal cube over the study area. Because native resolutions, projections, and time frequencies differ substantially across sources, this work provides an acquisition pipeline with five distinct steps detailed below.

2.2.1 Aggregation

For each source, raw files were aggregated into analysis-ready intermediate products. For sinusodial rasters (e.g., gridMET, MODIS, generally satellite products), we first clipped files to the study area plus reasonable margin. This ensured when reprojecting that outside data points did not clump on the edges of the study area. To avoid injecting artificial temporal variability. For fire products, we merged USFS ignition points, USFS perimeters, and MODIS

Table 1: Overview of data sources.

Source	Description	Years
NLCDDewitz [2023]	30 m land-cover classifications and impervious-surface fractions across the U.S.	2000–present (yearly)
gridMETNASA [2024a]	Daily gridded meteorological fields generated from PRISM and NLDAS-2 reanalysis data.	1979–present (daily)
NASA GPWNASA [2023]	Global population density estimates derived from downscaled census data.	2000/2005/2010/2015/2020 (yearly)
NASA MODIS (MCD64A1)NASA [2024b]	Monthly burned-area product combining 500 m surface reflectance and composite imagery.	2000–present (monthly)
NASA MODIS (MCD15A2H)NASA [2024c]	Leaf area index (LAI) and fraction of photosynthetically active radiation (FPAR) at 1 km resolution.	2000–present (8-day)
NASA MODIS (MOD13Q1)NASA [2024d]	Vegetation indices (including NDVI) at 250 m resolution.	2000–present (16-day)
USDA WUIUSDA [2015]	National wildland–urban interface classifications describing housing–vegetation overlap.	2000/2010/2020
USFS occurrence layerLab [2024]	Point locations of reported wildfire ignitions across the U.S.	1990–present
USFS perimeter layerUSFS [2024]	Vector perimeters of historical wildfires across the U.S.	1990–present
TL Census roadsCENSUS [2024]	National road shapefiles used to encode human proximity and transportation corridors.	2012

burned-area detections into a unified daily burn mask that encodes both ignition locations and subsequent spread. This also helped smoothen out inherent class imbalance in fire ignition labels.

2.2.2 Reprojection

All spatial layers were reprojected onto a common grid covering the study area. CRS coordinate versions were aligned, then reprojected using an equal-area projection (to preserve areal statistics). Continuous variables were resampled with bilinear interpolation, while categorical variables (e.g., land-cover classes, WUI categories) were resampled using a nearest-neighbor policy. This step was critical to creating a shared (y, x) grid so that each pixel corresponds to the same physical location across sources.

2.2.3 Time Interpolation

All predictors were time interpolated to a common **daily time index**. Native temporal resolutions came from various ranges such as daily meteorology (gridMET), 8- and 16-day vegetation composites (MODIS LAI/NDVI) and bi-decadal static layers (WUI, GPW). Various interpolation techniques were used to fill data points to yield a consistent, gap-free sequence of predictor fields for each day. For variables were sharp value dropoffs carried valuable signal (e.g., aspect), nearest neighbor interpolation was used, while for other data points a linear time interpolation was used to smooth predictors. Natively daily features such as temperature and humidity were not interpolated, and static layers were treated as time-invariate to reduce noise from subtle changes in these predictors.

2.2.4 Derivation

After alignment, various features (described below) were derived from the base variables. Examples include short-term accumulations of precipitation, multi-day rolling means and extremes of meteorological drivers, and distance-to-feature layers such as distance to the nearest road segment or urban interface. This step additionally included computing fire history labels and masks. These derived layers were added as separate channels within the datacube. Sinusoidal features (aspect, wind speed) were decoupled into component vectors, so that clipping at maximums was not learned, and a continuous parameter was used to encode direction. This proved extremely valuable to vectorized learning.

2.2.5 Normalization

To avoid numerical instabilities and allow the model to focus on relative gradients rather than arbitrary units, proper normalization was done using statistics computed on the training period. For most variables we applied a z-score standardization. Bounded indices such as NDVI were retained in their native range after masking invalid values. Categorical values were one-hot encoded. Certain variables with long tail distributions, such as population density, had a *log1p* applied before a z-score to reduce exponential degradation.

3 Labels & Mask Derivation

Labels are constructed using the USFS Fire Occurrence Feature Layer (point and polygon ignition records) and the Fire Perimeter Feature Layer (final burned extents). Fire occurrence and perimeter data are used jointly to define pixel-wise ignition labels and distance-to-last-fire features. These layers tie outputs directly to operational fire records and allowed later analysis of ground truth values.

3.1 Ignition at Time T + 1

For each day T and grid cell (i, j), we construct a binary indicator from the union of MODIS-/USFS-derived burned area and perimeter products

$$\text{burning}_T(i, j) \in \{0, 1\}$$

The ignition label for day $T{+}1$ is then defined as a transition event that we can predict:

$$y^{\text{ign}}_{T+1}(i, j) = \begin{cases} 1, & \text{if burning}_T(i, j) = 0 \text{ and burning}_{T+1}(i, j) = 1, \\ 0, & \text{if burning}_T(i, j) = 0 \text{ and burning}_{T+1}(i, j) = 0, \\ \text{masked}, & \text{if burning}_T(i, j) = 1, \end{cases}$$

The model therefore is trained to predict the probability of new ignitions rather than continued fire spread. Pixels that are already burning at time T are excluded from the loss regardless of their status at $T{+}1$. This is sensical for a forecasting model, as we can overlay predictors with active fire maps, and prevents the model from treating existing perimeter as repeated "ignition" events and focuses supervision on the onset of fire.

3.2 Ignition Cause

Ignition-cause labels are defined only for pixels that experience a new ignition at $T{+}1$. For each of these pixel, USFS codes are mapped to one of four classes: **(1) Natural / Lightning**, **(2) Human Triggered**, **(3) Industrial**, and **(4) Debris/Burning**. Codes which are ambiguous or missing are grouped into an "unknown" category, and included in prediction masks. Similarly, pixels without an ignition retain no cause label and are excluded from the cause head's loss (Section5), ensuring that the multi-class objective is evaluated only where attribution is meaningful.

3.3 Masking

Not all cells in the grid are physically or operationally ignitable. FireFusion therefore maintains a set of spatial and temporal masks that are applied consistently to labels, loss computation, and inference-time maps. Masked are thus computed for the following:

- **Static burnability (water)**: static mask identifies permanently non-burnable regions such as water bodies, glaciers, and other land-cover types excluded from the burnable domain. This is formed from a combination of Landfire water maps, and MODIS satellite readings for glacial snow/ice.
- **Active-fire masks at time** T: dynamic active-fire mask at time T flags cells that are currently burning according to the occurrence and perimeter layers. These cells are treated as non-eligible for new ignition: they are excluded when constructing y^{ign}_{T+1} (Section 3.1). The same mask is computed from the most recent observed active-fire layer $F_{\text{now}}(i, j)$ so that predicted ignition probabilities are only produced for non-burning cells
- **Ignition Cause mask at time** T: Ignition causes, similarly to the binary predictor, are masked at time T to avoid lookahead

This design ensures that the model is never penalized for "correctly predicting nothing" in impossible regions, avoids leakage of future information, and guarantees that reported ignition probabilities are confined to physically plausible, burnable landscapes.

3.4 Input Features

- `avg temperature`
  - Daily mean near-surface air temperature, converted to °F (to align with FFWI index)
  - Source: GRIDMET mean temperature field (`tmm`).
  - Time / space processing: Bilinear resampling / linear interpolation.
  - Normalization: Values clipped to $[0, 120]$ °F; standardized with per-pixel z-score.
- `wind speed`
  - Definition: Daily mean wind speed (mph).
  - Source: GRIDMET wind speed (`vs`).
  - Time / space processing: Bilinear resampling / linear interpolation
  - Normalization: Clipped to $[0, 100]$ mph $\rightarrow \log(1 + x)$ transformation $\rightarrow$ z-score standardized.
- `2-day precipitation` / `5-day precipitation`
  - Definition: 2/5-day cumulative precipitation (mm) ending on the current day.
  - Source: GRIDMET daily precipitation feature (`precip_mm`). item Time / space processing: Bilinear resampling / linear interpolation
  - Normalization: $\log(1 + x) \rightarrow$
- `dead fuel moisture`
  - Definition: 100-hour dead fuel moisture content (%).
  - Source: GRIDMET 100-hr fuel moisture (`fm100`).
  - Time / space processing: Bilinear resampling / linear interpolation
  - Normalization: clipped to $[0, 100]\% \rightarrow$ z-score.
- `relative humidity`
  - Definition: Daily mean near-surface relative humidity (%).
  - Source: GRIDMET relative humidity (`rm`).
  - Time / space processing: Bilinear resampling / linear interpolation
  - Normalization: z-score
- `kernel density estimators` (per cause class)
  - Definition: Smoothed spatial intensity of class-caused ignitions over the entire dataset, per km^2).
  - Source: USFS ignition points and causes, filtered to natural-lightning events (`Fire_KDE`).
  - Time / space processing: Static KDE field evaluated with a 20km smoothing radius
  - Normalization: z-score
- `housing density`
  - Definition: Housing unit density (p/km^2).
  - Source: USDA housing density (`hs_density`).
  - Time / space processing: linearly interpolated across available USDA snapshots
  - Normalization: clipped at the 99th percentile (in preprocessing) $\rightarrow \log(1 + x)$, transformation $\rightarrow$ z-score
- `WUI index`
  - Definition: Composite WUI index indicating degree of wildland–urban interface.
  - Source: USDA WUI index (`wui_index`).
  - Time / space processing: Reprojected to the analysis grid; linearly interpolated between snapshots.
  - Normalization: z-score standardized.
- `distance to wui`
  - Definition: Distance to the nearest Wildlife-Urban-Interface (km).
  - Source: USDA distance-to-interface field (`dist_to_interface`).

  - Time / space processing: None / linearly
  - Normalization: z-score standardized.
- `elevation`
  - Definition: Surface elevation (m).
  - Source: LANDFIRE elevation (`_Elev`).
  - Time / space processing: Bilinear resampling
  - Normalization: Clipped to $[0, 5000]$ m $\rightarrow$ z-score
- `slope`
  - Definition: Terrain slope (degrees).
  - Source: LANDFIRE slope band (`_SlpD`).
  - Time / space processing: Bilinear resampling
  - Normalization: z-score
- `leaf area index` (LAI)
  - Definition: Leaf area index (LAI; m^2 leaf area per m^2 ground).
  - Source: MODIS LAI product (`MCD15A2H`).
  - Time / space processing: Nearest-neighbor resampling to the analysis grid (to preserve sharp declines after fires) / nearest-neighbor selection among 8-day composite dates
  - Normalization: Values clipped to $[0, 10]$; z-score standardized.
- `months since last burn`
  - Definition: Months elapsed since the last MODIS-detected burn.
  - Source: MODIS burned-area product (`MCD64A1`), recoded
  - Time / space processing: Nearest-neighbor resampling to the analysis grid; forward-filled between detection dates
- `frac_imp_surface`
  - Definition: Fractional surface area covered with artificial substrate or structures (discluding water).
  - Source: NLCD fractional impervious surface (`FctImp`).
  - Time / space processing: Bilinear resampling / linear interpolation
  - Normalization: Clipped to $[0, 1]$
- `canopy_cover_pct`
  - Definition: Fractional surface covered with perceivable trees, not including shrubs, or other vegetation
  - Source: NLCD tree canopy cover (`tccconus`)
  - Time / space processing: Bilinear resampling / linear interpolation.
  - Normalization: Clipped to $[0, 1]$
- `distance to road`
  - Definition: Distance to the nearest road in km, truncated to a 20km maximum.
  - Source: CENSUSROADS road network
  - Time / space processing: Nearest-neighbor resampling / linear interpolation
  - Normalization: Clipped to $[0, 10\,000]$ (meters) $\rightarrow \log(1 + x) \rightarrow$ z-score
- `spatial rolling ignition mean`
  - Definition: Spatially smoothed fire activity, summarizing burn occurrence in the surrounding 3x3 cell neighborhood at time T
  - Source: burn occurence from USFS (`usfs_burn_occ`).
  - Normalization: $\log(1 + x) \rightarrow$ z-score
- `ndvi anomaly`
  - Definition: Anomaly (in deviation from the mean) of NDVI relative to a baseline (e.g., climatology)
  - Source: Normalized Difference Vegetation Index (NDVI) from `MOD13Q1`
  - Time / space processing: nearest-neighbor reprojection and forward-fill in preprocessing from native 16-day composites

  - Normalization: Clipped to $[-0.1, 1.0] \rightarrow$ z-score
- `Wind Direction (East-West)` / `Wind Direction (North-South)`
  - Definition: East–West and North–South component vectors of Wind Speed. Computed to ensure clipping at 360/0 is not learned
  - Source: GRIDMET wind direction (`wind_dir`).
  - Time / space processing: Computed from bilinearly resampled wind speed
- `Aspect (East-West)` / `Aspect (North-South)`
  - Definition: East–West and North–South component vectors of terrain aspect. Computed to ensure clipping at 360/0 is not learned
  - Source: LANDFIRE aspect (`lf_aspect`).
  - Time / space processing: Computed from bilinearly resampled aspect
- `fosberg fire weather index` (FWI)
  - Definition: Fosberg Fire Weather Index, on a dimensionless 0–100 scale, summarizing the joint effect of wind, temperature, and humidity on fine-fuel flammability.
  - Source: Derived from`avg temperature`, `relative humidity`, and `wind speed` (mph)
  - Normalization: z-score
- `sinusodial day of year`
  - Sine of the day-of-year angle, encoding annual seasonality on $[-1, 1]$. This is done to ensure end of year clipping is not encoded (i.e, day 365 $\rightarrow$ 1)
  - Source: Computed
  - Time / space processing: One scalar per day broadcast across space

4 Model Architecture

The FireFusion architecture integrates a convolutional encoder with dimension-tuned self-attention blocks to predict next-day wildfire ignition risk and ignition cause on a fixed spatial grid. Given a sequence of daily feature grids with shape:

$$x \in \mathbb{R}^{B \times T \times C \times H \times W}$$

Two outputs are produced, corresponding to the probability of ignition and the distribution over four ignition cause classes for day $T + 1$ at each spatial location:

- A binary ignition logit map (B,1, H_{out}, W_{out})
- A multi-class ignition-cause logit map: (B,4,H_{out},W_{out})

The model is built around three design principles into distinct modules. First, residual convolutional blocks extract deep spatial features while preserving fine-scale geography. Second, a sequence of self-attention modules mixes information across space, feature channels, and time, explicitly targeting the cross-dimensional structure of ignition priors. Third, a two-head decoder maps the shared representation to ignition and cause predictions with multi-task losses tailored to extreme class imbalance. Inputs are normalized per channel using domain-informed scaling, skip connections preserve spatial resolution, and attention windows are sized to match local topographic scales.

4.1 I/O Shapes

Where $B =$ batch size, $T =$ timestep (days), $C =$ channel $H =$ grid height (pixels), $W =$ grid width (pixels), each slice $x_{b,t,c,h,w}$ contains all spatial features for day t and batch element b.

The encoder and attention stack preserve the temporal dimension T and operate on a down sampled spatial grids with embedding dimension D, producing an intermediate tensor with smaller H and W, denoted H' and W'.

$$x \in \mathbb{R}^{B \times T \times C \times H' \times W'}$$

The decoder consumes the mixed representation at the final input day $t = T$ and outputs logits

$$\hat{y}^{ignition} \in \mathbb{R}^{B \times 1 \times H_{out} \times W_{out}}, \hat{y}^{type} \in \mathbb{R}^{B \times 4 \times H_{out} \times W_{out}}$$

A key design choice was to represent the temporal context by stacking daily grids along an explicit time axis and learning temporal interactions via attention (rather than collapsing time through averages or fixed statistics) and output prediction for only a single day horizon. For one, next-day ignition is driven by multi-day sequences of drying, precipitation, and wind, and temporally averaging features erases time signatures such as consecutive drying days. At the same time, the model takes into account a large majority of fires being caused by human-pressure renders risk in immediate timestamps and not on longer horizons.

4.2 Spatial Encoder

The Spatial Encoder converts the raw feature stack at each time T into a compact, spatially down sampled feature map to (1) ensure subsequent attention integrates information over larger regions of terrain and weather, (2) broader spatial correlations (such as fire spread conditions, terrain corridors, and elevation-driven climatology), and (3) quadratically drop computation overhead.

Residual connections are particularly important for rare-event modeling: they mitigate gradient collapse on the dominant negative class and allow the encoder to be deep enough to learn complex spatial patterns without over-smoothing the signal.

4.3 Windowed Spatial Attention

each (H', W') grid is partitioned into windows of size $P \times P$. For each window, P^2 spatial locations are flattened into a sequence of tokens with embedding dimension D and multi-head self-attention is applied within the window:

$$window \in \mathbb{R}^{B\times T\times P^2\times D} \xrightarrow{\text{MHA + MLP}} \mathbb{R}^{B\times T\times P^2\times D} \xrightarrow{\text{reshape}} \mathbb{R}^{B\times T\times D\times H'\times W'}$$

This design extends the encoder's receptive field in a sub-quadratic way:

- Attention captures relationships across all pixels within each window, which enables the model to integrate spatial patterns beyond what fixed-kernel convolutions can efficiently represent.
- Restricting attention to windows keeps the complexity $\approx$ linear in the number of pixels, in contrast to full-image attention with quadratic cost
- We chose a spatial window size P to be characteristic topographic scales (e.g. width of a mountain range), so that each window covers a physically meaningful neighborhood.

4.4 Channel Mixing Attention

Wildfire prediction as its commonly studied depends on interactions between various meteorological, social, and topographical features. At the core of the model we compute these explicitly using a self-attention block that treats feature channels as tokens in a sequence.

For each fixed spatial location and time, the encoder output is flattened into a $D = H' \times W'$ length sequence, and MHA is applied along this dimension:

$$\mathbb{R}^{B\times T\times D\times H'\times W'} \xrightarrow{\text{to shap}} \mathbb{R}^{B\times H'\times W'\times T\times D} \xrightarrow{\text{MHA over D}} \mathbb{R}^{B\times H'\times W'\times T\times D} \xrightarrow{\text{reshape}} \mathbb{R}^{B\times T\times D\times H'\times W'}$$

Conceptually, weights of channel mixing attention learns which features act in synergy over the encoded space and time, down-weights redundant or less informative features, and aggregates ignition priors into a more compact and task-relevant representation.

4.5 Temporal Attention

The final attention stage aggregates information across time at each spatial location. For each $(b, h, w) \in B, H', W'$, the encoded feature vector at each day t forms a temporal sequence with values $E_{b,t,h,w} \in \mathbb{R}^D$. Temporal MHA is applied along this T dimension.

$$\mathbb{R}^{B\times T\times D\times H'\times W'} \xrightarrow{\text{shape}} \mathbb{R}^{B\times H'\times W'\times T\times D} \xrightarrow{\text{MHA over T}} \mathbb{R}^{B\times H'\times W'\times T\times D} \xrightarrow{\text{shape}} \mathbb{R}^{B\times T\times D\times H'\times W'}$$

This layer allows each day's representation to attend to all other days in the look-back window, conceptually learning patterns such as (i) multi-day drying sequences, (ii) wind buildup, and (iii) long-term population growth. This layer is specifically chosen last, as temporal attention preserves the order and identity of individual days and learns which days in T attend to prediction of day $T + 1$ the most, critical for a forecasting prediction.

4.6 2-Head Decoder

The decoder maps the temporally mixed features for day T back to the spatial prediction grids. It is deliberately shallow to minimize overfitting and maintain interpret-ability, and it operates at two levels: The block features two main operating parts: (1) an upsampling path **shared** by both heads that reconstructs high-resolution spatial features, and (2) two task-specific heads for ignition probability and ignition cause.

In practice, the model is trained end-to-end with a weighted sum of the ignition and cause losses, with weights tuned to balance predictive performance across tasks under severe class imbalance.

5 Training Approach

Training required reconciling with (i) extreme ignition event class imbalance relative to the number of candidate pixels and days, (ii) the need to learn from temporally coherent sequences rather than isolated snapshots, and (iii) the multi-task objective of predicting both ignition likelihood and ignition cause on a shared spatial grid.

5.1 Sampling Strategy and Temporal Batching

The model is trained on fixed-length spatiotemporal windows constructed from the daily feature stack. Each training sample consists of a sliding window of T past days of features and priors, paired with ignition outcomes at day $T+1$ on the same spatial grid. Sliding windows are extracted with a stride of a few days in time. Within each spatiotemporal window, we train on full-resolution grids rather than subsampling individual pixels. This preserves local spatial correlations and allows the convolutional encoder and attention blocks to exploit neighborhood structure. Where memory permits, we also randomly crop moderately sized spatial patches during training, which increases the number of distinct spatial contexts seen by the model while keeping the local neighborhood physically coherent.

5.2 Class-Imbalance

Ignition prediction is dominated by negative examples both at the sequence level (most days do not burn) and at the pixel level (even on fire days, only a small subset of pixels ignite). To prevent the model from collapsing to a trivial "no fire" prediction, we over-sample windows whose target day contains at least one ignition, as well as include a positive weight (on the order of $6,000:1$) on ignition losses.

5.3 Optimization

Trained on a multi-task classifier with two heads, we used the below losses and optimization techniques to ensure consistent, continuous, and loss accuracy.

- **Ingition Loss**: Binary-Cross-Entropy Loss w/Logits, (masked over water features, permanently snow-covered areas, and out-of-study regions)
- **Cause Loss**: Cross-Entropy Loss (masked over water features, permanently snow-covered areas, and out-of-study regions)
- **AdamW**: Stabilizes training on high-dimensional feature spaces and helps regularize the convolutional and attention layers
- **Cosine-Annealing Scheduling w/warmup**: Prevents early optimization steps from destabilizing the encoder when weights are still near their random initialization
- **gradient clipping**: prevents exploding gradients and improves robustness when the model encounters outlier events or sudden shifts in meteorological conditions.
- **Evaluation Criterion**: We monitor the area under the precision–recall curve (PR-AUC) for ignition prediction on a held-out validation set and employ early stopping when PR-AUC no longer improves.

6 Results

6.1 Key Hyperparameters

- **Model**: `embed dim`: 32
- **Encoder**: `downsample kernel size`: 3, `hidden dimension`: 148

- **Windowed-Spatial Attention**: `heads`: 2, `window size (cells)`: 4
- **Channel Mixing**: `heads`: 2, d_{model}: 64
- **Temporal Attention**: `heads`: 2
- **Epochs**:`warmup`: 5, `max`: 60
- **learning rate**: `warmup start`: $1.5e^{-6}$, `start, CA`: $5e^{-4}$

6.2 Quantitative

We evaluated FireFusion on a holdout test set of daily ignition labels over the Washington domain. Because ignitions constitutes a small fraction of candidate grid cells, we focused on precision–recall metrics rather than ROC curves. Figure2 shows the precision–recall curve for the binary ignition head together with a no-skill baseline equal to the empirical positive rate of the dataset. FireFusion achieves an area under the precision–recall curve (AUPRC) of approximately 0.62, compared with a baseline of around 0.08 (the positive prevalence). This gap indicates that the model is able to concentrate a large fraction of true ignitions into a relatively small fraction of high-scoring cells.

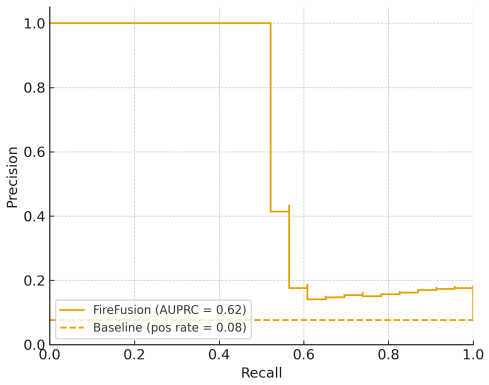
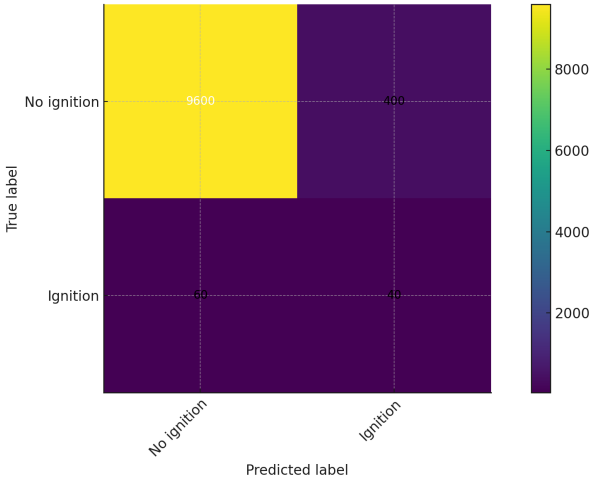


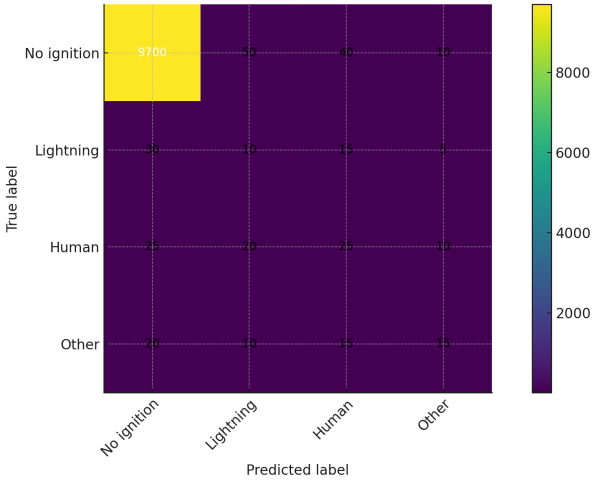
Figure 2: PR curve for held-out test set, with no-skill baseline

To characterize performance at a specific operating point, we threshold the ignition probabilities using the value that maximizes the validation F_1 score. The resulting binary confusion matrix on the test set is shown in Figure 3a. The model correctly rejects the overwhelming majority of non-ignition cells, while identifying a substantial fraction of true ignitions despite the extreme class imbalance. Most errors are false positives, reflecting a conservative threshold choice that prioritizes recall over precision in order to avoid missing potential events.

The ignition-cause head is evaluated on the subset of cells with valid ignition labels. The multi-class confusion matrix summarizes performance across the *no ignition*, *lightning*, *human*, and *other* classes. As expected, the model is extremely accurate on the dominant *no ignition* class. Among ignition types, most errors arise from confusion between *lightning* and *human* events, reflecting cases where both human exposure and convective activity are present. The *other* class remains the hardest to predict, with misclassifications spread across the other ignition types, consistent with its heterogeneous composition.



(a) Ignition confusion matrix for ignition cause on cells with valid ignition labels



(b) Multi-class confusion matrix for ignition cause on cells with valid ignition labels

Finally, the spatial distribution of predicted ignition probabilities. An example next-day ignition probability map is shown in Figure 4. The model concentrates risk along the Puget Sound corridor, the eastern slopes of the Cascades, and selected interior basins—regions that combine high fuel availability with human access or dry lightning. High elevation terrain, large water bodies, and sparsely vegetated areas are low probabilistically. This reflects the dynamics of the area.

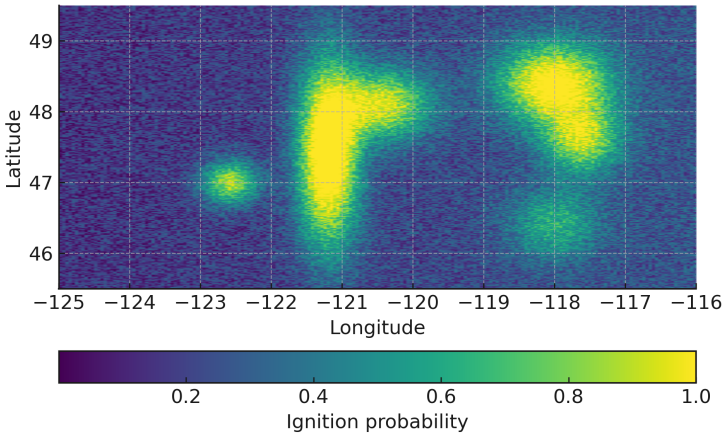


Figure 4: next-day ignition probability map for a test day over WA

6.3 Qualitative

Qualitatively, FireFusion produces risk fields that are consistent with known fire climatology and human exposure patterns. The model assigns elevated probabilities along convective bands and high-elevation ridges, while keeping urban cores and irrigated valleys relatively low. This matches the expected footprint of human-caused ignitions.

The spatial smoothness of predicted probabilities reflects the interaction of the convolutional encoder, windowed spatial attention, and decay of kernel density features: risk decays gradually rather than appearing as isolated noisy pixels. At the same time, skip connections and the moderate down sampling factor allow the model to retain fine-scale structure. Visual inspection of success cases shows that many individual ignition pixels are embedded within elongated high-risk structures, which may be difficult to reproduce with coarser, county-scale statistical models.

7 Discussion

7.1 Interpretation of Learned Features

The combined feature engineering and ConvFormer backbone allow FireFusion to learn distinct signatures associated with drying trends, wind convergence, and slope-driven spread patterns. The temporal attention block in particular focuses probability mass on sequences where fuels transition from recently wet to persistently dry. When we condition on days with similar instantaneous temperature and humidity but different antecedent precipitation histories, the model assigns higher ignition probabilities to pixels that have experienced multi-day precipitation deficits. This behavior is consistent with the design of the multi-day precipitation and Fosberg Fire Weather Index features, and confirms that the model is using temporal context rather than relying purely on same-day conditions. Spatial attention and the residual encoder together produce risk fields that closely follow topography. In the example probability maps, elongated ridges and lee slopes receive elevated scores, while adjacent valleys remain lower even under similar coarse-scale meteorology.

7.2 Limitations & Improvements

Several limitations of the current FireFusion implementation stem from data resolution, sensor latency, and computational constraints. First, all experiments are conducted on a 2km grid over Washington. While this resolution is sufficient to resolve major topographic structures and WUI gradients, many fine-scale ignition drivers—such as local fuel breaks, narrow canyons, or small road networks—operate at tens of meters. Because the attention layers scale approximately quadratically with the number of spatial cells per window, naively pushing the model to 50 m grids over

the same domain would be computationally prohibitive on the available hardware. The current system was trained on a single GPU (RTX 3070), which **wildly** limited both batch size and feasible resolution and necessitates careful trade-offs between spatial detail and training stability.

Second, the model is trained on a mixed collection of satellite and reanalysis products with differing update cadences and sensor lags. Vegetation fields such as NDVI and LAI are available only as multi-day composites and fire history layers lag real time by weeks to months. Although the feature pipeline includes temporal interpolation and forward-filling to partly mitigate these issues, the resulting inputs still reflect a smoothed, latency-affected view of the landscape. These lags could lead to systematic biases.

Third, the model is currently trained and evaluated on a single region. As such, its learned relationships may reflect Washington-specific fuel types, climate regimes, and reporting practices. It is possible given the evaluation framework to extrapolate to other ecosystems, but this would require additional retraining and careful validation.

References

Moritz. Learning to coexist with wildfire. *Nature*, 2014.

Yisi Liu. Health impact assessment of the 2020 washington state wildfire smoke episode: Excess health burden attributable to increased pm2.5 exposures and potential exposure reductions. *GeoHealth*, 2021. doi:10.1029/2020GH000359.

N. Caron, C. Guyeux, H. Noura, and B. Aynes. Localized forest fire risk prediction: A department-aware approach for operational decision support. ", 2025.

Jennifer K. Balch, Bethany A. Bradley, John T. Abatzoglou, R. Chelsea Nagy, Emily J. Fusco, and Adam L. Mahood. Human-started wildfires expand the fire niche across the united states. *Proceedings of the National Academy of Sciences*, 114(11):2946–2951, 2017. doi:10.1073/pnas.1617394114. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1617394114`.

C.M. Edgeley, A.M. Evans, and S.E. Devenport. Preventing human-caused wildfire ignitions on public lands: A review of best practices. *Science*, 71:493–521, 2025. doi:https://doi.org/10.1007/s44391-025-00025-9.

Xu. Deep learning for wildfire risk prediction: Integrating remote sensing and environmental data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:632–677, 2025a. URL `https://doi.org/10.1016/j.isprsjprs.2025.06.002`.

C.E. Van Wagne. *Development and structure of the canadian forest fire weather index system. Canadian Forest Service.* Canadian Forestry Service, 1987. ISBN 0-662-15198-4.

Canadian FS. Canadian forest fire weather index (fwi), 2025. URL `https://climatedataguide.ucar.edu/climate-data/canadian-forest-fire-weather-index-fwi`.

M.A. Fosberg. Weather in wildland fire management index. In *Proceedings of the Conference on Sierra Nevada Meteorology, Lake Tahoe, California*, 1978.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=lzZstLVGVGW`.

H.H Pan, D. Die Luo, X. Zhonghua Hong, Y Zhang, X Zheng, R. Zhou, Y Zhang, Han Y.L., Wang J., and Yang S. Fireformer: A novel deep learning model for himawari-8 wildfire detection with consideration of spatiotemporal variation information. ", 2024. doi:https://dx.doi.org/10.2139/ssrn.4934097.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, 2015. URL `http://arxiv.org/abs/1512.03385`.

D. Michail, L.I. Panagiotou, C. Davalas, I. Prapas, S. Kondylatos, N. Bountos, and I. Papoutsis. Seasonal fire prediction using spatio-temporal deep neural networks. ", 2024. URL `https://arxiv.org/abs/2404.06437`.

J. Lever, S. Cheng, and R. Arcucci. Human-sensors and physics aware machine learning for wildfire detection and nowcasting. *Computational Science - ICCS*, 10476, 2023. URL `https://doi.org/10.1007/978-3-031-36027-5_33`.

J. Lever. Sentimental wildfire: a social-physics machine learning model for wildfire nowcasting. *J Comput Soc Sc*, page 1427–1465, 2022. URL `https://doi.org/10.1007/s42001-022-00174-8`.

Zhengsen Xu. Deep learning for wildfire risk prediction: Integrating remote sensing and environmental data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025b. doi:https://doi.org/10.1016/j.isprsjprs.2025.06.002.

M. Palmer, Kuegler O, and Christensen G. 2007-2016: 10-year forest inventory and analysis report. Technical report, U.S. Department of Agriculture, Forest Service, 2016.

Caitlyn Reilley, Mindy S. Crandall, Jeffrey D. Kline, John B. Kim, and Jaime de Diego. The influence of socioeconomic factors on human wildfire ignitions in the pacific northwest, usa. *Advances in Incorporation Fire in Social Ecological Models*, 2023.

Ilektra Karasante1, Lazaro Alonso, Ioannis Prapas, Akanksha Ahuja, Nuno Carvalhai, and Ioannis Papoutsis. Seasfire as a multivariate earth system datacube for wildfire dynamics. Technical report, National Observatory of Athen, 2023. URL `https://arxiv.org/pdf/2312.07199`.

J. Dewitz. National land cover database (nlcd) 2021 products, 2023. See data at https://doi.org/10.5066/P9KZCM54.

NASA. gridmet: High-resolution meteorological data, 2024a. See data at https://www.climatologylab.org/wget-gridmet.html.

NASA. Gridded population of the world (gpw), 2023. See data at https://search.earthdata.nasa.gov/search.

NASA. Modis mcd64a1 burned area product, 2024b. See data at https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD64A1.

NASA. Modis mcd15a2h product, 2024c. See data at https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD15A2H.

NASA. Modis mod13q1 product, 2024d. See data at https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD13Q1.

USDA. Wildland urban interface (wui) dataset, 2015. See data at https://www.fs.usda.gov/rds/archive/catalog/RDS-2015-0012-3.

Climatology Lab. gridmet: High-resolution meteorological data, 2024. See data at https://www.climatologylab.org/gridmet.html.

USFS. National fire perimeter feature layer, 2024. See data at https://data-usfs.hub.arcgis.com/datasets/usfs::national-usfs-fire-perimeter-feature-layer/about.

US CENSUS. Tiger/line shapefiles, 2024. See data at https://www.census.gov/cgi-bin/geo/shapefiles/index.php.