https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

WSDS-CAN: Wildfire Spread Prediction Dataset for Canadian Boreal Forests

Hossein Keshmiri[1], Khan Arif Wahid[1]

[1]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, Canada

Correspondence to: Hossein Keshmiri (hossein.keshmiri@usask.ca)

Abstract. The development of high-fidelity wildfire spread models is contingent upon the availability of spatially and temporally aligned multi-layer datasets. Existing global and continental databases have successfully catalogued fire events; however, they are often constrained by coarse spatial resolutions or absence of environmental variables. Additionally, there is a significant scarcity of machine-learning-ready open-access datasets dedicated to the Canadian landscape. This leaves a critical gap in the data available for modelling fire-front dynamics characteristic of Canadian ecosystems. In this paper, we introduce a comprehensive, multi-layer wildfire perimeter prediction dataset engineered specifically for the Canadian boreal forest, covering the period from 2001 to 2020. Unlike previous catalogues, this dataset includes 2,565 distinct fire events with a minimum area threshold of 1 ha (0.01 km^2), capturing a more inclusive historical record of fire activity. The curation process involves a rigorous integration of final burned geometries from the Canadian National Fire Database (CNFDB) with fire-adapted environmental covariates, including high-frequency meteorological indices, static topographical features, and fuel parameters. To facilitate machine learning applications, the data is processed into a format suitable for tasks such as fire segmentation and perimeter prediction, with spatial resolutions adapted to the scale of individual fire events. By providing granular inputs for both small-scale ignitions and complex fire perimeters, this dataset serves as a foundational resource for advancing predictive modelling and real-time surveillance pipelines in northern forest environments.

1 Introduction

Historically, wildfires have functioned as natural ecological processes vital for ecosystem maintenance; however, compounding climate change impacts have fundamentally altered global fire regimes. The result is a marked increase in the frequency, intensity, and spatial extent of fire events. Recent statistics highlight the severity of this shift: in 2023, Canadian wildfires consumed roughly 15 million hectares—more than twice the previous historical record (Kirchmeier-Young et al., 2024). This trajectory continued across North America throughout 2024 and 2025, with burned areas reaching 46,000 km^2 in Canada and 36,000 km^2 in the United States (Kelley et al., 2025). As wildfires increasingly transcend their natural ecological boundaries to threaten human infrastructure and environmental sustainability, the need for robust, high-resolution data to monitor and predict fire perimeters has never been more critical. The evolution of wildfire spread modelling is deeply rooted in the transition from raw thermal anomalies to refined event-based geometries. At the core of this progression are NASA's

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

Moderate Resolution Imaging Spectroradiometer (MODIS) products: MCD14A1 (Giglio et al., 2016), which provides daily fire masks and thermal hotspots at 1 km resolution, and MCD64A1 (Giglio et al., 2018), which utilizes changes in surface reflectance to map burned area boundaries. While foundational, these raw products require significant post-processing to delineate individual fire lifetimes. To address this, the FRY (Laurent et al., 2018) database integrates MODIS hotspots with burned area products to catalogue global fire patches and functional traits. Similarly, platforms like GlobeFire (Artés et al., 2019) and the Global Fire Atlas (Andela et al., 2019) extract daily perimeters and spread vectors. Building on these methodologies, the Fire Events Delineation (FIRED) database (Balch et al., 2020) employs automated spatiotemporal clustering to provide continuous estimates of spread rates, effectively bridging the gap between pixel-level thermal detections and landscape-scale event analysis across the conterminous United States.

Despite these advancements, accurately predicting fire spread requires capturing the complex, non-linear interactions between a fire and its immediate environment. Wildfire behaviour is governed by high-frequency meteorological shifts, static topographical features, and diverse fuel parameters. Consequently, training effective machine learning (ML) models necessitates spatially and temporally aligned, multi-layer datasets that provide more than just discrete event identification; they must offer a continuous mapping of the climatic and environmental variables that dictate fire dynamics.

One notable effort in this domain is the Next-Day Wildfire Spread dataset proposed by (Huot et al., 2022), which presents multivariate historical data across the contiguous United States from 2012 to 2020. This dataset integrates topography, vegetation, and population density with daily weather and drought indices for 18,845 fire events, with dual snapshots of fire perimeters on successive days (t and $t + 1$). The data is structured to a coarse, fixed resolution of 1km within 64×64-pixel patches which may adversely influence the quality of prediction/detection results. Furthermore, the dataset is specifically designed for the US fire events and may not generalize well to the Canadian boreal forest fires.

Another remarkable effort in this domain is the Canadian Fire Spread Dataset (CFSD) proposed by (Barber et al., 2024), which offers an event-based daily fire progression catalog specifically for the Canadian landscape from 2002 to 2021. By combining high-precision boundaries from the National Burned Area Composite (NBAC) (Skakun et al., 2022) with VIIRS and MODIS active fire detections, the CFSD provides daily-scale progression maps for 3,269 fires, each attributed with 50 environmental covariates including weather metrics, topography, and forest attributes. While the CFSD serves as a robust research-oriented resource for building statistical and machine learning models, its focus is primarily on large-scale regional patterns. The dataset is restricted to fires exceeding 1,000 ha and utilizes a fixed 180-m pixel resolution derived from interpolated satellite hotspots.

Although these foundational datasets have enabled broad-scale analysis, a gap remains for high-resolution, multi-layer integration capable of capturing both emerging ignitions and complex fire perimeters. Moreover, the majority of machine-learning-ready datasets are focused on other regions specially the contiguous United States, with a comparatively small portion of open-access data dedicated to the Canadian landscape. In this paper, we introduce a multi-layer wildfire perimeter prediction dataset engineered specifically for the Canadian boreal forest. By cataloguing fire events greater than or equal to 1 ha ($0.01 km^2$) from 2001 to 2020, our dataset provides a more inclusive historical record that encompasses smaller, high-

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

frequency fire events often omitted from larger catalogues. Furthermore, by coupling discrete fire geometries with fire-adapted environmental covariates, this dataset provides the granular inputs necessary for training ML models for precise fire segmentation and spread prediction tasks. Table 1 summarizes the details of our proposed dataset compared with other most related datasets.

Table 1. Comparison of most related open-access datasets.

Dataset	Spatial extent	Temporal range	Minimum fire size	Pixel size	Burned masks	Environmental variables
FRY	Global	2005-2011	$1km^2$	500m	Final masks	None
GlobeFire	Global	2001 to present	$0.21km^2$	500m	Daily masks	None
Global Fire Atlas	Global	2003-2016	$0.21km^2$	500m	Daily masks	None
Copernicus Daily Burnt Area	Global	2023-2025	NA	300m	Daily masks	None
FIRED	US	2001-2019	$0.21km^2$	500m	Daily masks	None
Next day Wildfire Spread	US	2012 - 2020	$1km^2$	1km	Two masks at days t and $t+1$	Meteorological and terrain
CFSDS	Canada	2002 - 2021	$10km^2$	180m	Daily masks	Meteorological and terrain
WSDS-CAN	Canada	2001 - 2020	$0.01km^2$	Varies based on fire size (>30m)	Final masks	Meteorological and terrain

2 Data Sources and Methods

The curation of this dataset integrates multiple geospatial data sources. First, we acquired final fire burned area polygons from the (National Fire Database Fire Polygon Data, 2025) also known as the Canadian National Fire Database (CNFDB) maintained by the Canadian Wildland Fire Information System (CWFIS), which catalogs the start and end dates, final burned masks, and total sizes of fire events across Canada from 1917 to 2024. Each fire instance is identified by a unique ID. Due to historical constraints in auxiliary terrain data (e.g., land cover), the dataset was restricted to events occurring between 2001 and 2020 with a minimum burned area of 0.01 km^2. Figure 1 illustrates the distribution of these events across Canada during the study period. To ensure suitability for fire detection and spread prediction tasks, we excluded records lacking confirmed start or end dates. Furthermore, we constrained the temporal scope to events lasting between 1 and 60 days.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

Finally, we removed instances where burned areas were represented by primitive circular geometries—typically an artifact of measurement limitations—resulting in a final catalog of 2,565 distinct fire events.

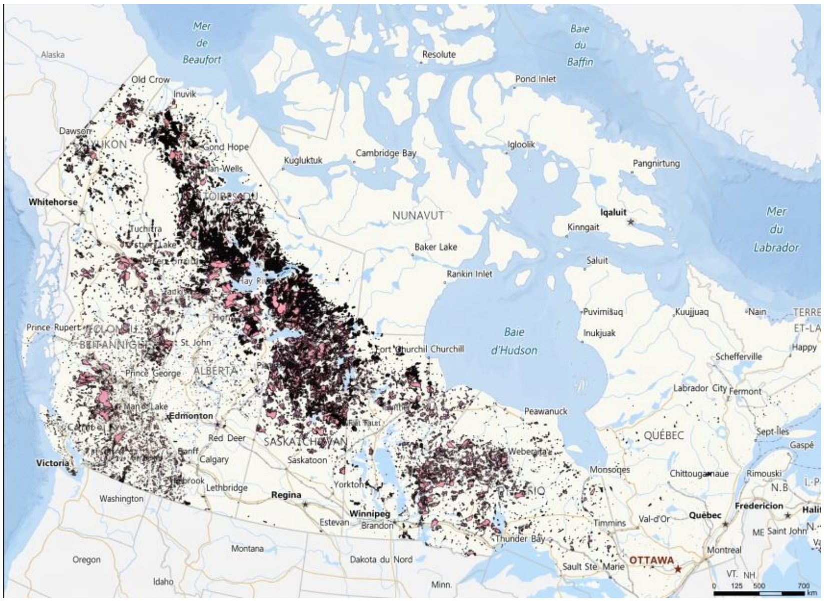


Figure 1. Distribution of fire events greater than 1 ha during 2001 and 2020 across Canada. Image created using CNFDB data in QGIS software.

Because developing ML models for fire spread prediction requires a thorough characterization of the conditions driving fire evolution, the integration of environmental covariates is essential. To optimize data handling and reduce the computational complexity of the preprocessing phase, we aggregated these terrain and meteorological data at the provincial and territorial levels.

Topographical features were sourced from the (Copernicus GLO-30 Digital Elevation Model, 2025) with the spatial resolution of 30 m. Secondary spatial variables, including slope and aspect, were subsequently derived from this DEM. Because the study focuses on the Canadian landscape—much of which is situated on a stable craton with minimal significant tectonic or large-scale topographical deformation over the study period—these features are treated as static. Furthermore, as the source data lacks a temporal resolution, a single representative DEM raster was collected for each province/territory. All topographical features are stored as GeoTiff. To characterize fuel types, we utilized the MODIS MCD12Q1 v061 yearly land cover product (Friedl and Sulla-Menashe, 2022). This dataset provides categorical land cover classifications across 17 distinct classes at a 500 m spatial resolution. Given the annual temporal resolution of the source, we compiled individual GeoTIFF rasters for each province/territory, covering every year of the 2001–2020 study period. The density of vegetation is another essential feature in fire dynamics playing a key role in the rate of fire spread. To incorporate a proxy for vegetation density and biomass, LAI data was acquired from the MODIS MOD15A2H v061 product (Myneni et al., 2021). The LAI dataset features an 8-day temporal resolution and a 500 m spatial resolution. All available records spanning from the beginning of 2001 through the end of 2020 were collected for each province/territory and stored in NetCDF format.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

Dynamic meteorological conditions are primary drivers of fire behavior, dictating both the intensity and direction of spread. To capture these variable forces, we extracted eastward and northward wind vectors (u and v components) from the ERA5-Land post-processed daily dataset (Copernicus Climate Change Service, 2024). Additionally, daily skin temperature and dewpoint temperature were acquired from the same source to characterize the thermal gradients and atmospheric moisture dynamics across the study area throughout the fire cycle. Table 2 shows a summary of terrain and meteorological parameters associated with the fire events presented in our dataset.

Table 2. Summary of environmental variables and their sources used for creating WSDS-CAN.

Feature	**Description**	**Temporal acquisition/resolution**	**Spatial resolution**	**Format**	**Data source**
Elevation (m)	Height of terrain	2011-2015/fixed	30 m	GeoTIFF	Copernicus GLO-30 Digital Elevation Model
Slope (degrees)	Steepness of terrain (0-90°)				
Aspect (degrees)	Direction of slope (0-360°)				
Land cover (classes)	Type of terrain	2001-2020/yearly	500 m		MODIS MCD12Q1 v061
Leaf area index (no unit)	Ratio of one-sided leaf area to unit ground area	2001-2020/8-day		NetCDF	MODIS MOD15A2H v061
10m u-component of wind (m/s)	Speed of eastward component of wind at 10 m above ground	2001-2020/daily	9 km		EAR5
10m v-component of wind (m/s)	Speed of northward component of wind at 10 m above ground				
Skin temperature (K)	Temperature of the surface of the Eart				
2m dewpoint temperature (K)	The temperature to which air must be cooled at a 2-meter height to reach 100% relative humidity				

2.1 Preprocessing and Data Preparation

The raw data were aggregated from multiple sources, each characterized by distinct spatial properties and inconsistent Coordinate Reference Systems (CRSs). Consequently, a rigorous preprocessing pipeline was implemented to ensure spatial and temporal alignment across all variables. A CRS defines the mathematical framework used to project two-dimensional values onto the Earth's surface; to maintain consistency with the CNFDB, all environmental parameters were reprojected into the Canada Atlas Lambert projection (EPSG:3978). The fire instances in this catalogue exhibit a significant range of scales, with burned areas spanning from 0.01 km^2 to 2,450 km^2. Given that the dataset is structured to provide a comprehensive historical record—including ignition points, final burned perimeters, and environmental covariates—mapped

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

to individual fire IDs, the preprocessing pipeline must account for these varying spatial extents. Consequently, we generated a customized square bounding box for each fire ID, dynamically scaled to encompass the total extent of the specific event. This bounding box defines the spatial extent used to crop the surrounding landscape, ensuring the capture of environmental covariates that influence fire growth and spread. To embed sufficient contextual information, the dimensions of the box are set to be 30% larger than the maximum diameter of the final burned area geometry—represented as either a polygon or multi-polygon. While this adaptive approach effectively captures the environmental covariates for large-scale fires, the moderate-to-low spatial resolution of certain variables poses a challenge for smaller events. Specifically, for very small perimeters, a strictly proportional buffer may result in a bounding box too small to contain a meaningful number of pixels from coarse-resolution products like ERA5-Land. To mitigate this, we established a minimum threshold of 4 km for the bounding box dimensions, ensuring sufficient spatial context across all data layers.

The bounding box is centered on the centroid of the burned area geometry, with a fixed internal grid resolution of 128×128 pixels. As a result, the effective pixel size for each instance is determined by the ratio of the total box dimensions to the 128-pixel grid width. Figure 2 presents the preparation of the multi-layer dataset, and the details are elaborated in the following sections.

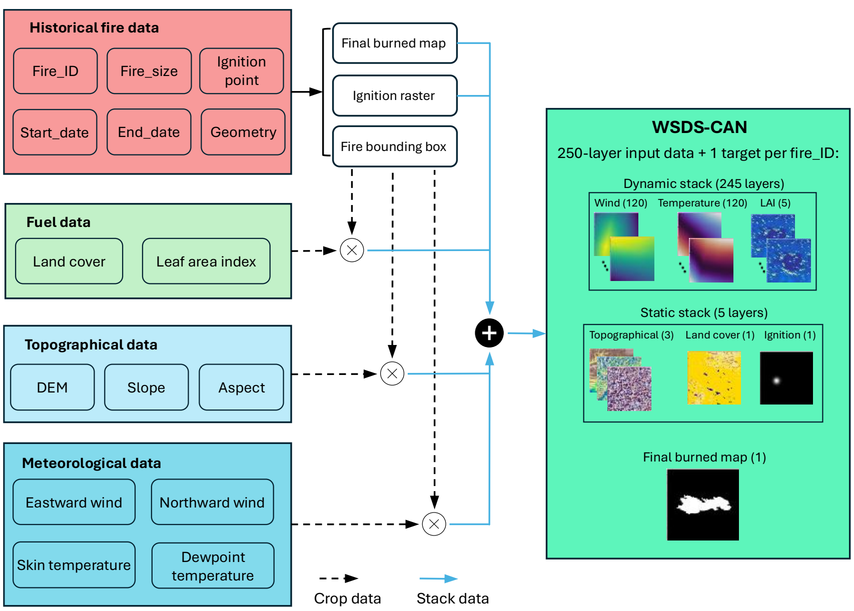


Figure 2. Flowchart of multi-layer data preparation by preprocessing raw data from multiple databases.

2.1.1 Ignition

In CNFDB, the ignition points of fire events are reported as latitude and longitude pairs. These coordinates are provided by Canadian fire management agencies—including provinces, territories, and parks—and are subject to localization errors such as coordinate rounding, manual reporting inaccuracies, or discrepancies between the detected ignition and the initial fire

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

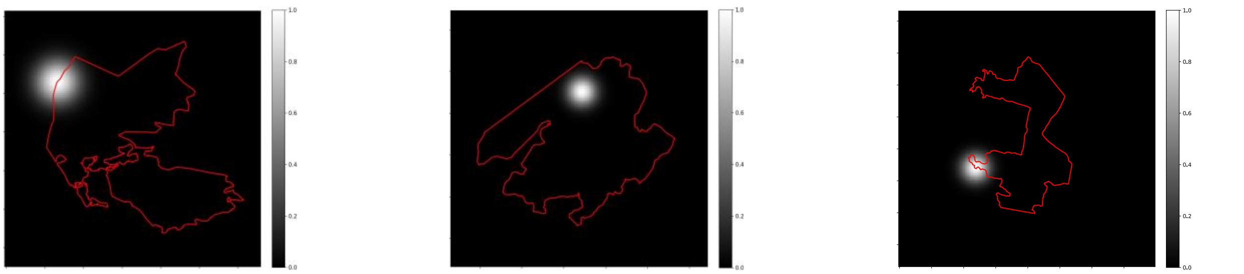
Open Access Earth System Science Data Discussions

report. To account for these uncertainties, we generate a Gaussian blob centered on the reported ignition point, with a variance proportional to the fire's total burned area as described by Equation (1).

$$\sigma \propto \frac{r_m}{pixel\ size} \tag{1}$$

Where, σ is the standard deviation of the Gaussian blob, r_m is radius of the smallest enclosing circle for the fire's burned extent, and the pixel size denotes the spatial resolution of a single pixel in the bounding box. Therefore, the ignition raster for each fire instance is modelled as a circular, normalized Gaussian distribution centered on the reported ignition coordinates, with a spatial extent proportional to the total burned area. The pixel values represent the probability of a location being the true origin; values are maximized at the centroid and decay as the distance from the center increases. Because the source coordinates were extracted directly from the CNFDB (already provided in EPSG:3978), no additional CRS reprojections were required for this layer.

Figure 3 illustrates the ignition raster for a random selection of fire IDs from the dataset. Representing the ignition point as a Gaussian blob provides significant advantages for ML models by offering stronger spatial signals regarding the fire's origin. A single-pixel point (a 'hot pixel') in a sparse raster of zeros can be easily overlooked during the convolution and pooling layers of a model, whereas a multi-pixel Gaussian distribution ensures that the initialization feature is preserved and prioritized during training.



Fire ID: 17LA-LAKE2_2017	Fire ID: 06SE-ZANGEZA	Fire ID: 15WY-LRIR02
Fire size: 52Km^2	Fire size: 10 Km^2	Fire size: 1.84 Km^2

Figure 3. ignition raster of three different file samples shown against their final burned extent. The ignition raster is saved as Gaussian blobs (white ball) only.

2.1.2 Land cover

To characterize the fuel types in the fire's vicinity, we utilized the MODIS MCD12Q1 v061 (Friedl and Sulla-Menashe, 2022) land cover product. Since the repository contains one GeoTIFF per province/territory for each year, we implemented a selection logic based on the CNFDB metadata: for each fire ID, the administrative region was identified via the `SRC_AGENCY` column, and the appropriate annual raster was retrieved using the `YEAR` column. The raw data, originally provided in a Sinusoidal projection, were reprojected to EPSG:3978 prior to being cropped by the fire's bounding box. This categorical dataset comprises 17 vegetation classes, with all no data or fill-value pixels assigned a value of zero.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

2.1.3 DEM, Slope, Aspect

To capture the topographical characteristics of the terrain surrounding each fire, DEM raster was collected from the Copernicus GLO-30 dataset for each province/territory. Due to the high spatial resolution of this product, the data volume for an entire province often exceeded the transfer limits of the source server. To address this, we partitioned the data acquisition into multiple smaller queries; these individual segments were subsequently tiled and stitched together using the `rasterio.merge` utility. For each fire ID, the corresponding regional DEM was selected based on the `SRC_AGENCY` attribute. The native WGS 84 (EPSG:4326) coordinate system of the raw DEM data was reprojected to EPSG:3978 before being cropped to the specific spatial extent of the fire's bounding box.

Slope represents the angular measurement of terrain steepness, ranging from 0° to 90°. In vegetated regions, steeper terrain is typically prone to more rapid fire spread due to convective heating and closer proximity of the fuel bed to the rising flames. Aspect indicates the compass orientation of the slope, ranging from 0° to 360°, where 0° signifies a north-facing slope and 180° signifies a south-facing slope. In the Canadian landscape, south-facing slopes often present a higher fire risk due to increased solar radiation exposure, which results in drier fuels. For each fire ID, Slope and Aspect were derived from the cropped DEM raster using the `richdem` library. The no data value for all topographical layers was standardized to -9999.

2.1.4 Leaf Area Index

Apart from type of vegetation, the density of the biomass is essential as it defines the combustion potential. LAI (m^2/m^2) represents the one-sided green leaf area per unit ground surface area serving as a primary indicator of canopy fuel load. It can be used as a proxy to both fuel density and the overall combustibility of the vegetation. Since LAI features an 8-day temporal resolution, it is not possible to obtain unique data for every day of a fire's duration. To account for the influence of LAI on fire development, we partitioned the period between the ignition and extinction dates into five equal-length intervals. Using the `Pandas` library, we then identified the LAI frame closest to the center of each temporal bin.

While this approach mitigates the temporal discrepancy, challenges remain for short-duration events. For fires lasting fewer than five days, we specifically selected the LAI frames nearest to the start and end dates to capture the vegetation contrast immediately before and after the event; in these cases, only two unique LAI frames were generated. Once the appropriate frames were selected, they were reprojected from their native Sinusoidal projection to EPSG:3978 and cropped to the fire's bounding box. The no data values were standardized as `numpy.nan`. Figure 4 displays five sequential LAI frames alongside the corresponding final burned area map for four randomly selected fire IDs. By comparing the temporal sequence to the final burn perimeter, the progression of the fire can be observed; the expanding regions of deep blue pixels, which represent the lowest LAI values, correlate with the spatial growth of the fire as captured in the final frame.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

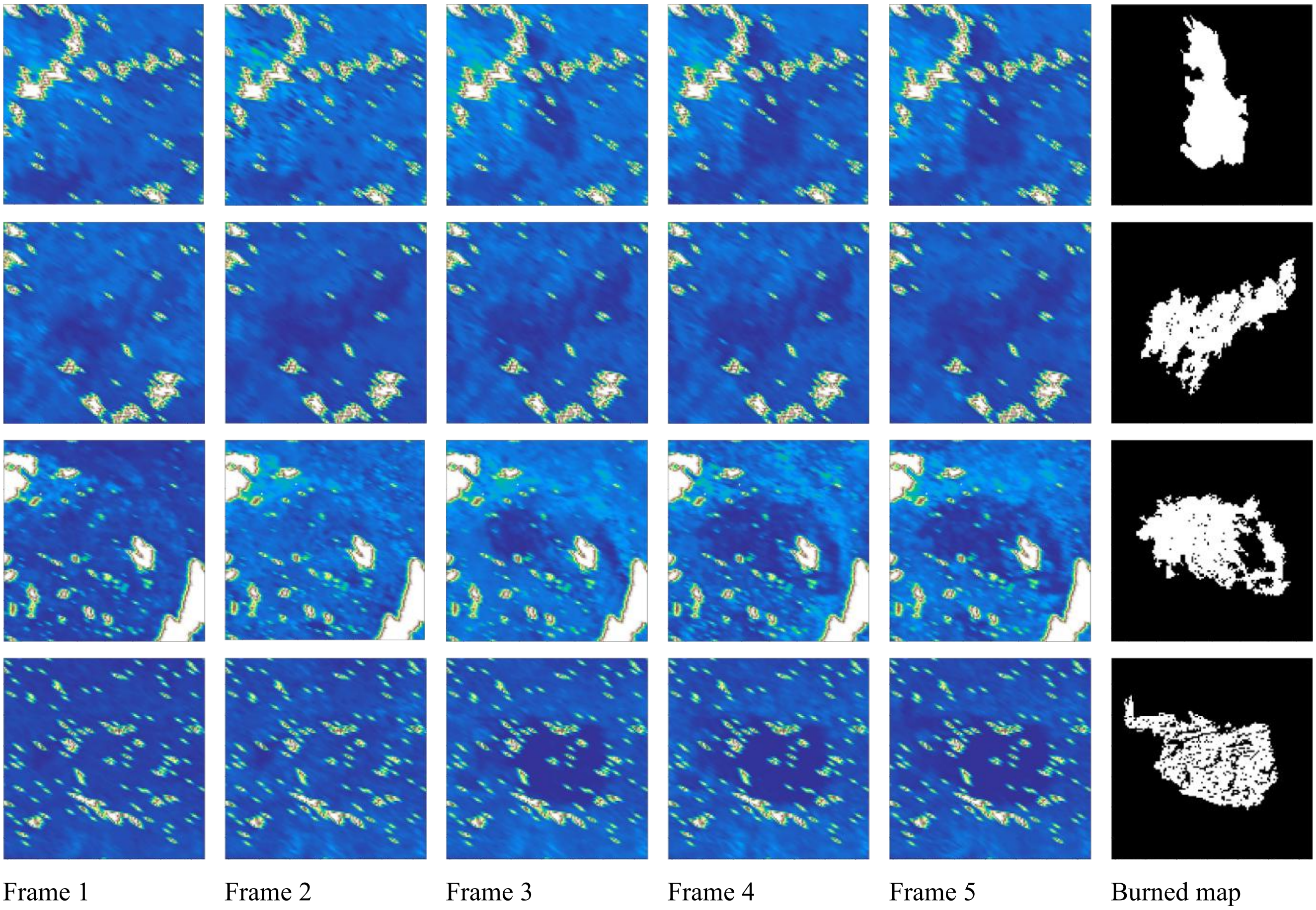


Figure 4. LAI frames of four random fire events. Fire IDs from the first to the last rows are 12PN-GILBERT, 15LA-JEAN, 15WY-PHILION, 16SR-LEXI, respectively.

2.1.5 Wind and Temperature

Wind and temperature are fundamental meteorological drivers that shape fire spread patterns and influence burn severity. To capture these dynamics, we collected daily eastward (u-component) and northward (v-component) wind vectors at a 10-meter height, alongside daily skin temperature and 2-meter dewpoint temperature. These parameters were retrieved from the ERA5-Land post-processed daily statistics (Copernicus Climate Change Service, 2024) for each province/territory. The raw data is provided at a 9 km spatial resolution in a Sinusoidal projection. To compensate for the coarse resolution of these meteorological parameters, we first evaluated the data coverage within each fire's bounding box. If the box dimensions were insufficient to capture a meaningful spatial gradient—defined as failing to encompass at least one ERA5-Land cell and its immediate neighbours—the box was systematically expanded to encapsulate the necessary atmospheric context. Once the spatial extent was finalized to include these broader meteorological features, it was used to crop the raw data. Furthermore, we applied bilinear interpolation during the resampling process to smooth the sharp grid boundaries and provide a more

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

continuous representation of the wind and temperature fields. Following these considerations, daily meteorological data for each fire ID is generated from the ignition date through the extinction date, covering a duration of up to 60 days. These data are stored as multi-layer 128×128-pixel GeoTIFF tensors. Consequently, the dynamic meteorological features for a single fire ID can be summarized as a series of $60 \times 128 \times 128$ tensors for the u-component of wind, the v-component of wind, skin temperature, and 2-meter dewpoint temperature. The no data values within these meteorological layers are standardized as `numpy.nan`.

3 Machine Learning Application

This section demonstrates the utility of the WSDS-CAN dataset for the task of wildfire spread prediction. For this application, we implemented the neural network architecture and the training strategy proposed in our prior work (Keshmiri and Wahid, 2026). The dataset was partitioned into training, validation, and test sets using a 70:15:15 ratio. Regarding the static input layers, Elevation, Slope, and Ignition were retained as single-channel raster. Conversely, the Aspect layer was decomposed into its sine and cosine components—forming two separate channels—to resolve the inherent angular cyclicity of compass headings. The categorical land cover data is expanded into an eight-channel feature set via one-hot encoding to facilitate numerical processing. To ensure input dimension consistency across the dataset, the temporal depth of the daily dynamic features—including eastward and northward wind components, skin temperature, and dewpoint temperature—is fixed at 60 channels. Fire instances with durations shorter than 60 days are standardized using zero-padding. While a straightforward approach involves feeding all input channels simultaneously to generate a single prediction of the final burned area, a more robust and informative strategy involves a hybrid input method. In this approach, static channels and LAI are provided as initial context, while the daily meteorological layers are fed into the model iteratively. This simulates the temporal evolution and spread of the fire on a day-by-day basis.

To address the class imbalance between "Fire" and "No Fire" pixels, we utilized a spatially weighted binary cross-entropy with logits (BCEWithLogitsLoss) as the objective function. A custom weight map was developed to prioritize four distinct zones derived from the ignition location and the ground-truth burned area. The highest weight ($w = 50$) is assigned to ignition pixels with values exceeding 0.4, ensuring the model identifies the fire's origin. The main interior of the burned area is assigned the second-highest weight ($w = 15$) to emphasize the burned extent against the background. To suppress excessive false positives and penalize boundary violations, a three-pixel-wide "containment ring" surrounding the fire perimeter is weighted significantly ($w = 10$). Finally, all the background ("No Fire") pixels are assigned a unite weight. ($w = 1$). Figure 5 illustrates the generated daily burned maps overlaid on the final burned area for a specific fire instance in Manitoba (Fire ID: MB_MB2018NE116_2018). These predictions are produced using the iterative simulation approach, where static features and LAI are provided as initial context, and meteorological data are fed into the model sequentially to represent daily spread.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

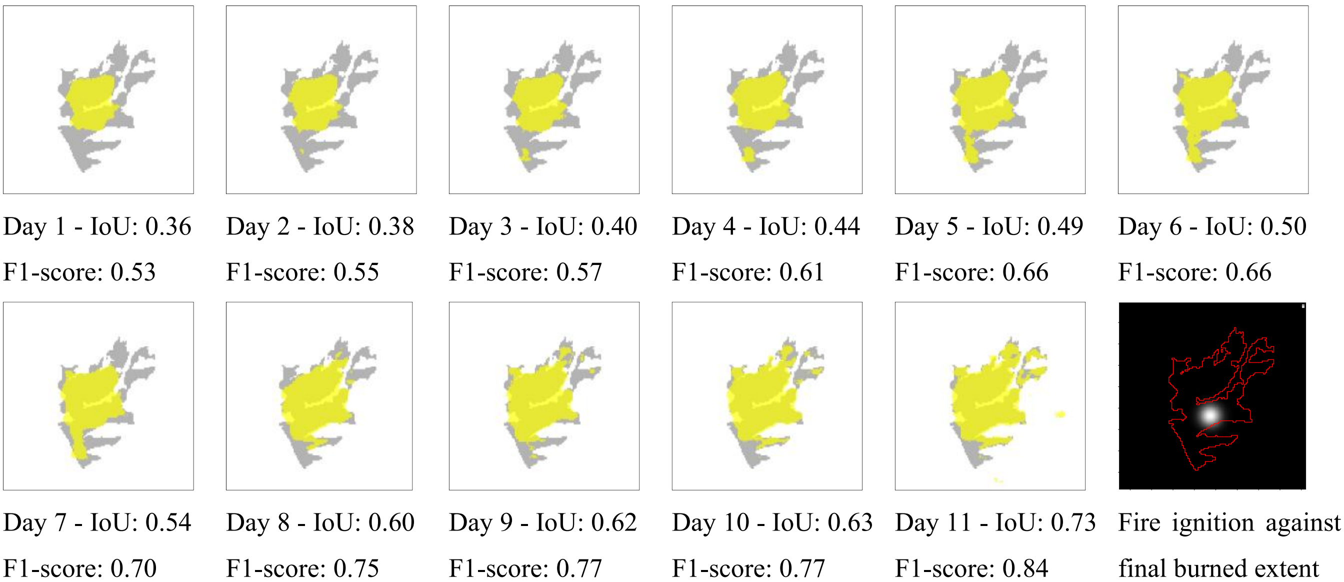


Figure 5. Predicted burned masks of a random fire sample generated by a neural network model trained on WSDS-CAN dataset. Yellow is the predicted mask; grey is the final burned mask. The final frame illustrates the Gaussian-modelled ignition point overlaid against the total burned extent.

In this experiment, as new meteorological data is sequentially introduced to the model, updated burned area predictions are generated. Successive iterations demonstrate that the model is capable of capturing and following the overall spatial patterns of the burned extent. The consistent upward trend in evaluation metrics, specifically the F1-score and IoU, validates that the model effectively interprets the relevance of dynamic inputs in relation to the target final burned area. To further demonstrate the utility of the WSDS-CAN dataset, we present the performance metrics for the fire spread prediction experiment across four distinct neural network architectures. All models were trained using the same training, validation, and test partitions, utilizing BCEWithLogitsLoss as the objective function; the comparative results are summarized in Table 3.

Table 3. Comparison of average performance metrics across all test samples for various neural network architectures.

Network architecture	Precision	Recall	F1-score	IoU	AUC
SE-ResUNet	0.65	0.84	0.66	0.51	0.97
DeepLabv3	0.60	0.71	0.64	0.50	0.97
UNet++	0.55	0.80	0.64	0.51	0.96
UNet	0.50	0.82	0.62	0.48	0.94

4 Conclusions

The Wildfire Spread Dataset for Canada (WSDS-CAN) provides a comprehensive, high-resolution spatiotemporal framework for the advancement of wildfire spread modelling in Canadian boreal forests. By integrating multi-source satellite observations, including MODIS land cover and LAI, with high-resolution Copernicus GLO-30 DEM terrain features and ERA5-Land meteorological drivers, this dataset captures the complex interplay between fuel, topography, and weather. Spanning two decades (2001–2020), WSDS-CAN offers the historical depth and spatial consistency necessary to train sophisticated deep learning architectures.

To ensure spatial and physical consistency, all data layers were reprojected from their native formats (such as Sinusoidal and WGS 84) to the EPSG:3978 coordinate system, which is optimized for the Canadian landmass. Furthermore, the implementation of an adaptive bounding box strategy ensured that coarse-resolution meteorological data were captured with sufficient spatial context; by systematically expanding the fire's extent to encompass neighbouring ERA5 cells, we preserved meaningful atmospheric gradients essential for model learning.

The utility of the dataset was demonstrated through a fire spread prediction experiment using an iterative simulation approach. This methodology effectively translates daily dynamic features into predictable growth patterns, as evidenced by the consistent upward trend in IoU and F1-score metrics during temporal progression. Furthermore, a comparative analysis of four neural network architectures namely: SE-ResUNet, DeepLabv3, UNet++, and UNet, validated the robustness of the data. Ultimately, WSDS-CAN bridges the gap between raw geospatial data and actionable predictive modelling, supporting efforts to mitigate the environmental and socioeconomic impacts of wildfires in the face of a changing climate.

Data Availability

The dataset can be accessed from https://doi.org/10.5281/zenodo.20264141 (Keshmiri and Wahid, 2026).

Author contributions

Conceptualization, H.K. and K.A.W.; methodology, H.K. and K.A.W.; software, H.K.; validation, H.K. and K.A.W.; formal analysis, H.K.; investigation, H.K.; resources, H.K.; data curation, H.K.; writing—original draft preparation, H.K.; writing—review and editing, H.K. and K.A.W.; visualization, H.K.; supervision, K.A.W.; project administration, K.A.W.; funding acquisition, K.A.W.

Competing interests

The authors declare no conflicts of interest.

Disclaimer

Copernicus Publications adds a standard disclaimer: "Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper.

While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher."
Please feel free to add disclaimer text at your choice, if applicable.

Acknowledgements

During the preparation of this manuscript, the authors used Google Gemini 3 Flash, a 2026 LLM model, for the purposes of proofreading, text and image refinement. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Financial support

This word was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant number RGPIN-2018-06274.

Review statement

The review statement will be added by Copernicus Publications listing the handling editor as well as all contributing referees according to their status anonymous or identified.

References

Andela, N., Morton, D. C., Giglio, L., Paugam, R., Chen, Y., Hantson, S., Van Der Werf, G. R., and Randerson, J. T.: The Global Fire Atlas of individual fire size, duration, speed and direction, Earth Syst. Sci. Data, 11, 529–552, https://doi.org/10.5194/essd-11-529-2019, 2019.

Artés, T., Oom, D., de Rigo, D., Durrant, T. H., Maianti, P., Libertà, G., and San-Miguel-Ayanz, J.: A global wildfire dataset for the analysis of fire regimes and fire behaviour, Sci. Data, 6, 296, https://doi.org/10.1038/s41597-019-0312-2, 2019.

Balch, J. K., St. Denis, L. A., Mahood, A. L., Mietkiewicz, N. P., Williams, T. M., McGlinchy, J., and Cook, M. C.: Fired (Fire events delineation): An open, flexible algorithm and database of us fire events derived from the modis burned area product (2001–2019), Remote Sens. (Basel)., 12, 1–18, https://doi.org/10.3390/rs12213498, 2020.

Barber, Q. E., Jain, P., Whitman, E., Thompson, D. K., Guindon, L., Parks, S. A., Wang, X., Hethcoat, M. G., and Parisien, M.-A.: The Canadian Fire Spread Dataset, Sci. Data, 11, 764, https://doi.org/10.1038/s41597-024-03436-4, 2024.

Copernicus Climate Change Service, Climate Data Store, (2024): ERA5-land post-processed daily-statistics from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.e9c9c792, Accessed on 25 October 2025.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

European Space Agency (2024). Copernicus Global Digital Elevation Model. Distributed by OpenTopography. https://doi.org/10.5069/G9028PQB. Accessed on 28 October 2025.

Friedl, M. and Sulla-Menashe, D.: MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061, NASA Land Processes Distributed Active Archive Center [data set], https://doi.org/10.5067/MODIS/MCD12Q1.061, 2022.

Giglio, L., Schroeder, W., and Justice, C. O.: The collection 6 MODIS active fire detection algorithm and fire products, Remote Sens. Environ., 178, 31–41, https://doi.org/10.1016/j.rse.2016.02.054, 2016.

Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L., and Justice, C. O.: The Collection 6 MODIS burned area mapping algorithm and product, Remote Sens. Environ., 217, 72–85, https://doi.org/10.1016/j.rse.2018.08.005, 2018.

Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., and Chen, Y.-F.: 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, https://doi.org/10.1109/TGRS.2022.3192974, 2022.

Kelley, D. I., Burton, C., Di Giuseppe, F., Jones, M. W., Barbosa, M. L. F., Brambleby, E., McNorton, J. R., Liu, Z., Bradley, A. S. I., Blackford, K., Burke, E., Ciavarella, A., Di Tomaso, E., Eden, J., Ferreira, I. J. M., Fiedler, L., Hartley, A. J., Keeping, T. R., Lampe, S., Lombardi, A., Mataveli, G., Qu, Y., Silva, P. S., Spuler, F. R., Steinmann, C. B., Torres-Vázquez, M. Á., Veiga, R., van Wees, D., Wessel, J. B., Wright, E., Bilbao, B., Bourbonnais, M., Gao, C., Di Bella, C. M., Dintwe, K., Donovan, V. M., Harris, S., Kukavskaya, E. A., N'Dri, A. B., Santín, C., Selaya, G., Sjöström, J., Abatzoglou, J. T., Andela, N., Carmenta, R., Chuvieco, E., Giglio, L., Hamilton, D. S., Hantson, S., Meier, S., Parrington, M., Sadegh, M., San-Miguel-Ayanz, J., Sedano, F., Turco, M., van der Werf, G. R., Veraverbeke, S., Anderson, L. O., Clarke, H., Fernandes, P. M., and Kolden, C. A.: State of Wildfires 2024–2025, Earth Syst. Sci. Data, 17, 5377–5488, https://doi.org/10.5194/essd-17-5377-2025, 2025.

Keshmiri, H. and Wahid, K. A.: A UAV–Satellite Hybrid Pipeline for Wildfire Detection and Dynamic Perimeter Prediction, Drones, 10, https://doi.org/10.3390/drones10040263, 2026.

Keshmiri, H. and Wahid, K. A.: WSDS-CAN: Wildfire Spread Prediction Dataset for Canadian Boreal Forests [Data set]. Zenodo. https://doi.org/10.5281/zenodo.20264141, 2026.

Kirchmeier-Young, M. C., Malinina, E., Barber, Q. E., Garcia Perdomo, K., Curasi, S. R., Liang, Y., Jain, P., Gillett, N. P., Parisien, M. A., Cannon, A. J., Lima, A. R., Arora, V. K., Boulanger, Y., Melton, J. R., Van Vliet, L., and Zhang, X.: Human driven climate change increased the likelihood of the 2023 record area burned in Canada, NPJ Clim. Atmos. Sci., 7, https://doi.org/10.1038/s41612-024-00841-9, 2024.

Laurent, P., Mouillot, F., Yue, C., Ciais, P., Moreno, M. V., and Nogueira, J. M. P.: FRY, a global database of fire patch functional traits derived from space-borne burned area products, Sci. Data, 5, 180132, https://doi.org/10.1038/sdata.2018.132, 2018.

Myneni, R., Knyazikhin, Y., and Park, T.: MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V061, NASA Land Processes Distributed Active Archive Center [data set], https://doi.org/10.5067/MODIS/MOD15A2H.061, 2021.

https://doi.org/10.5194/essd-2026-378
Preprint. Discussion started: 22 July 2026

National Fire Database Fire Polygon Data: https://cwfis.cfs.nrcan.gc.ca/downloads/nfdb/fire_poly/current_version/NFDB_poly.zip, Accessed on 31 October 2025.

Skakun, R., Castilla, G., Metsaranta, J., Whitman, E., Rodrigue, S., Little, J., Groenewegen, K., and Coyle, M.: Extending the National Burned Area Composite Time Series of Wildfires in Canada, Remote Sens. (Basel)., 14, https://doi.org/10.3390/rs14133050, 2022.