Wildfire Duration Prediction Using Weather and Emission Data

Zachary Holland*, Emmanuel D Leonce†, Dae Won Hwang‡
*nms9dg@virginia.edu †fyb7sx@virginia.edu ‡qsh9fk@virginia.edu
University of Virginia

I. Abstract

This paper examines the factors that influence wildfire duration, with a focus on fuel conditions and weather variables that sustain combustion and spread. We integrate approximately 7.5 million wildfire-day records from the USDA Missoula Fire Lab Emissions Inventory (MFLEI) [1] with nearly 4 billion daily weather observations from gridMET [2] spanning 2003 to 2015. Using a custom spatiotemporal DBSCAN procedure for event identification, we aggregate daily fire detections into event-level records and join them with co-temporal meteorological variables. Feature selection combined Pearson and Spearman correlations, mutual information, ANOVA F-tests, and variance inflation factor analysis to rank predictors. Group-aware validation ensured that all days from a single fire event were kept in the same fold to prevent leakage. Tree-based regressors (Random Forest, XGBoost, Histogram Gradient Boosting) were evaluated, with Random Forest achieving the strongest performance under realistic validation. Results highlight the importance of coarse woody debris fraction, duff fraction, prefire fuel load, large-fuel moisture content, vapor pressure deficit, and geographic location. The work delivers a reproducible, scalable data pipeline and a validated feature set that establish a foundation for future wildfire duration modeling.

II. Introduction

Wildfires have increased in frequency and severity in recent decades, threatening ecosystems, infrastructure, and human lives. The U.S. Forest Service has noted that fires not controlled within the first 24 hours are more likely to escape initial attack and become severe incidents [7]. Burned acreage in the United States has risen from an annual average of approximately 2 to 3 million acres in the 1980s to about 7 million acres between 2010 and 2020 [3], [4], [6].

Predicting the duration of a wildfire is valuable for incident management because it influences crew rotation planning, aircraft scheduling, resource staging, and public safety measures. Longer duration often correlates with greater total emissions, prolonged exposure to smoke, and higher suppression costs. This study investigates whether event-level wildfire duration can be predicted from routinely available emissions data and meteorological variables. We develop a processing pipeline that transforms raw point-based fire detections into structured events, links them to daily weather conditions, and produces engineered features for modeling. Exploratory data analysis revealed seasonal and spatial patterns as well as data quality issues such as systematic zero entries in key fuel variables. All modeling in this work uses group-aware validation keyed to event identifiers to avoid information leakage between training and testing data.

III. Key Variables and Hypothesis

We hypothesize that longer wildfire duration is driven by a combination of fuel availability, fuel dryness, and atmospheric conditions favorable to sustained combustion and spread. Fires with greater prefire fuel loads and lower fuel moisture are expected to burn longer, especially under weather patterns that maintain low relative humidity, high vapor pressure deficit (VPD), elevated temperatures, and sustained winds [9], [11]. Geographic location (latitude, longitude) serves as a proxy for regional climatology, vegetation types, and topography, which collectively influence fuel characteristics and fire behavior.

The variables in our dataset fall into two primary categories: emissions and fuel-related metrics derived from MFLEI, and meteorological variables derived from gridMET. Several engineered features, such as temporal harmonics and rolling averages, were also included to capture seasonal and persistence effects.

A. Emission and Fuel Variables

- **prefire_fuel:** Estimated biomass available before ignition. Higher values provide more energy potential for sustained burning.
- **fuel_moisture_class:** Classifies fuels by moisture content. Lower classes ignite more readily and support faster spread.
- **cwd_frac, duff_frac:** Fractional composition of coarse woody debris and duff layer. Higher values indicate more persistent ground fuel that can smolder for days.
- **area_burned:** Burned area for a given detection; larger areas may reflect more intense or sustained events.
- **covertype / fuelcode:** Vegetation and fuel type classification, influencing burn rate, intensity, and extinction likelihood.
- **ECO, ECO2, ECH4, EPM2_5:** Estimated emissions of carbon monoxide, carbon dioxide, methane, and particulate matter. While not direct drivers, these can proxy for combustion completeness and intensity.

B. Weather and Atmospheric Variables

- **fm100, fm1000:** 100-hour and 1000-hour dead fuel moisture content. Lower values indicate drier large fuels that can sustain combustion over multiple days.
- **vpd:** Vapor pressure deficit; higher values accelerate fuel drying and increase flammability.
- **vs:** Wind speed, which supports flame spread, spotting, and oxygen supply.
- **tmmx, tmmn:** Daily maximum and minimum temperature; warmer conditions reduce fuel moisture and increase ignition likelihood.
- **rmin, rmax:** Daily minimum and maximum relative humidity; lower values exacerbate drying and sustain burning.
- **srad:** Surface solar radiation; higher values promote surface heating and preheating of fuels.
- **etr, pet:** Reference and potential evapotranspiration; higher rates correlate with rapid moisture loss from live and dead fuels.
- **pr:** Daily precipitation; lower or absent precipitation indicates drier antecedent conditions.
- **bi:** Fire danger index; a composite measure of flammability based on weather and fuel data [10].
- **th (wind direction):** Can influence spread direction; not yet integrated into predictive modeling but identified for future work.

C. Supporting Hypothesis with Data

Exploratory data analysis confirmed that many of these variables exhibit seasonal and geographic patterns consistent with known wildfire behavior. For example, high VPD and low fuel moisture coincided with the summer fire season in western states, while duff and coarse woody debris fractions varied strongly with vegetation type and region. Strong correlations between certain fuel variables and event duration in preliminary modeling further support their inclusion in the predictive feature set.

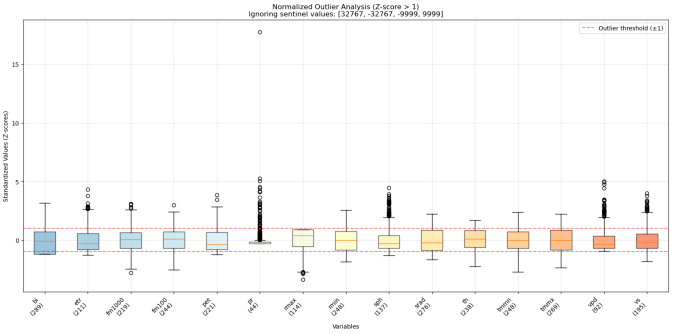

Fig. 1. Boxplots of weather variables show spread and outliers.

IV. Data Strategy and Integration

Our data strategy focused on building a reproducible, scalable pipeline to integrate large-scale wildfire emissions and meteorological records at daily resolution for the period 2003–2015. The goal was to create an event-centric dataset suitable for modeling wildfire duration.

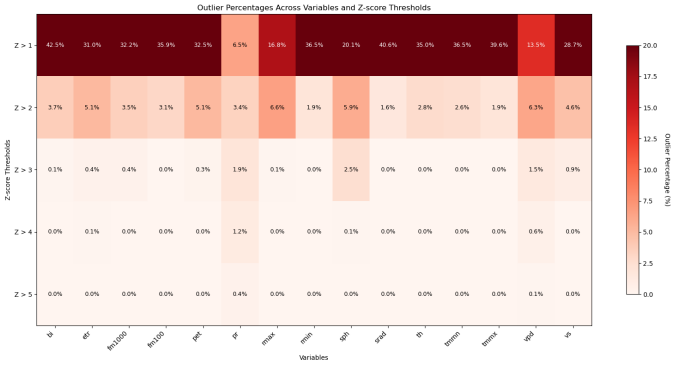

Fig. 2. Outlier percentages across z-score thresholds by variable.

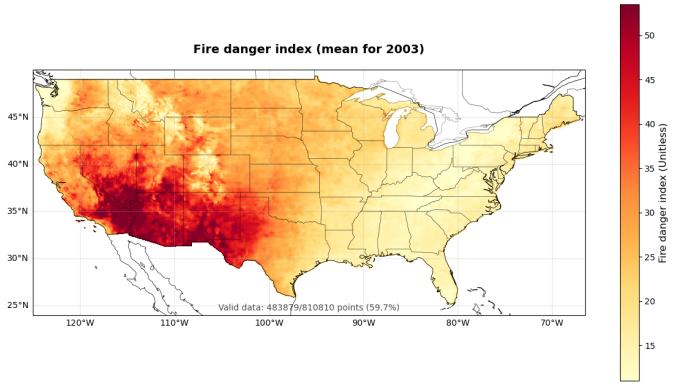

Fig. 3. Fire Danger Index map (2003): regional variability in risk.

A. Weather Processing: NetCDF to Parquet/BigQuery

Daily gridMET weather variables were provided as netCDF (.nc) files segmented by year and variable, totaling around 150 files and nearly 4 billion data points. These files store data as masked arrays, which preserve spatial and temporal

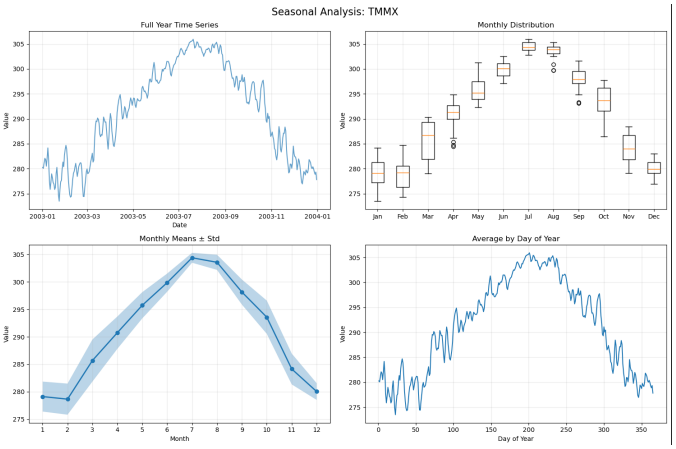

Fig. 4. Seasonal structure in `TMAX`: time series, monthly, and annual cycles.

dimensions but are not directly compatible with most tabular processing tools. Using `xarray` in combination with `dask` and `netCDF4`, we converted each file to Parquet format, retaining the $(\text{lat}, \text{lon}, \text{date})$ index structure. To support different analysis patterns, we produced both:

- *Long* format tables (variable, value) for flexible aggregation and filtering.
- *Wide* format tables (columns for each variable) to streamline repeated joins with the emissions dataset.

Quality control routines flagged out-of-range values, contiguous clusters of NAs, and inconsistencies in temporal coverage for later imputation (see Sec. V-C).

B. Fire Event Identification with DBSCAN

The MFLEI dataset contains roughly 7.5 million wildfire-day snapshots, each representing a single detection. To model at the event level, we needed to group related detections into coherent wildfire events. We used a custom spatiotemporal DBSCAN clustering process, tuned to the resolution of MFLEI coordinates and detection dates. The DBSCAN output assigned a persistent `fire_id` to each record, which served two critical purposes:

1) Enabled aggregation of daily detections into event-level features such as duration, spread rate, and total emissions.
2) Supported group-aware validation by ensuring that all observations from the same event stayed within the same fold.

DBSCAN was used only for event construction, not as a predictive algorithm.

C. Joining Emissions and Weather Data

Following event construction, we performed a spatial–temporal join to associate each MFLEI event-day record with co-located meteorological variables from gridMET. The weather data were stored as daily 4 km-resolution rasters in netCDF format and converted to Parquet tables in both long and wide formats.

The join process was executed in two stages:

1) **Temporal match:** Event-day timestamps were matched exactly to the gridMET daily index (UTC-based).
2) **Spatial match:** Fire detections were mapped to the centroid of the nearest gridMET cell using a k-d tree built on the (lat, lon) grid. For detections falling outside the contiguous gridMET domain (rare), we performed a fallback nearest-neighbor search within a 0.25° great-circle radius.

To scale this process to billions of row–row comparisons, we leveraged Google BigQuery's geospatial functions (`ST_DISTANCE`, `ST_CLOSESTPOINT`) and partitioned tables by date to maximize pruning. The pipeline first generated an intermediate table mapping every `fire_id, date` pair to the corresponding weather grid cell ID, then joined this to the wide-format gridMET table for feature enrichment.

The result was a unified dataset where each event-day row contains:

- MFLEI-derived fuel and emissions metrics.
- Collocated daily weather variables (temperature, humidity, wind speed/direction, solar radiation, precipitation, fire danger indices, etc.).
- Spatial metadata (lat/lon, eco-region).

This integration step produced the final modeling-ready table, with all features aligned at the daily temporal resolution and event ID granularity.

D. Feature Engineering

We engineered additional predictors to capture seasonal patterns, temporal persistence, and categorical effects:

- Temporal harmonics (`day_of_year_sin/cos`) to encode seasonality.
- Short-term rolling means and lag features for key variables (e.g., fuel moisture, VPD).
- Encoded categorical variables such as `fuelcode` and `covertype` using frequency or target encoding.
- Retained raw coordinates (latitude, longitude), which mutual information analysis showed to be informative proxies for regional climate and vegetation.

E. Reproducibility and Infrastructure

All processing steps were version-controlled and logged to ensure reproducibility:

- NetCDF-to-Parquet conversion jobs stored in Google Cloud Storage, followed by ingestion to BigQuery.
- Parameterized DBSCAN job outputting an event manifest (`fire_id` and parameters used).
- Joins and feature engineering tracked with schema versioning and execution timestamps.
- All datasets, queries, and processing scripts stored in a shared repository for handoff to future teams.

V. Learning Setup

A. Target Definition

The target variable, *duration*, was defined as the number of days from the first to the last active detection within a clustered `fire_id`. Each model row represented an event-level instance. In this study, we modeled total duration based on conditions observed during the event as a whole.

B. Validation Protocol

We employed group-aware K-fold cross-validation, ensuring that all observations from a given `fire_id` were contained in a single fold. This avoided information leakage that would otherwise occur if days from the same event appeared in both training and validation sets.

C. Imputation Strategy

Given frequent zero entries in fuel and emissions variables that likely represented missing values, we adopted a multi-tier imputation approach:

1) **Hot-deck imputation** for isolated gaps (about 1% of imputed values).
2) **k-NN imputation** within local spatial and temporal neighborhoods (about 40% of imputed values).
3) **Seasonal or yearly climatology substitution** for structured gaps (about 59% of imputed values).

All imputations were tracked with provenance tags to allow for sensitivity analysis.

D. Models Considered

We evaluated three primary tree-based regressors:

- Random Forest
- XGBoost
- Histogram Gradient Boosting

These models were selected for their ability to handle mixed data types, non-linear relationships, and moderate multicollinearity. k-Nearest Neighbors regression was included as a simple baseline. Hyperparameters were kept modest, as the focus was on establishing baseline learnability and identifying important predictors rather than exhaustive tuning.

VI. Feature Selection and Modeling Results

Feature selection used Pearson/Spearman correlations, Mutual Information (MI), ANOVA F-test, and VIF to control redundancy. These lenses provide complementary perspectives: linear association, non-linear dependence, discriminative power, and multicollinearity. Statistical summaries are reported in Tables I–V.

A. Top Features by Correlation (Pearson & Spearman)

TABLE I
Top Features by Pearson and Spearman Correlation

Feature	Pearson	Spearman	Avg
cwd_frac	0.252	0.238	0.245
duff_frac	0.245	0.238	0.241
longitude	0.135	0.221	0.178
prefire_fuel	0.188	0.161	0.175
fm1000_value	0.123	0.192	0.157
fm100_value	0.084	0.166	0.125
day_of_year_sin	0.088	0.152	0.120
covertype	0.117	0.103	0.110
bi_value	0.081	0.134	0.107
rmax_value	0.078	0.135	0.107

B. Top Features by Mutual Information

TABLE II
Top Features by Mutual Information

Feature	MI Score
longitude	0.6834
latitude	0.5423
day_of_year_sin	0.2773
day_of_year_cos	0.2747
srad_value	0.2678
prefire_fuel	0.2254
duff_frac	0.2146
cwd_frac	0.2145
fuelcode	0.1811
rmax_value	0.1708

C. Top Features by ANOVA F-test

TABLE III
Top Features by ANOVA F-test

Feature	F-score
cwd_frac	489,947.16
duff_frac	461,037.61
prefire_fuel	266,910.04
longitude	133,868.33
fm1000_value	111,089.27
covertype	100,803.28
day_of_year_sin	56,019.62
burn_source	54,274.63
fm100_value	51,751.55
bi_value	47,77642

D. Top Features by VIF (Low Multicollinearity)

TABLE IV
Top Features by VIF (Low Multicollinearity)

Feature	VIF
pr_value	1.31
th_value	1.10
prefire_fuel	2.86
longitude	4.52
season	3.60

E. Overall Combined Ranking

TABLE V
Overall Combined Feature Ranking

Rank	Feature
1	longitude
2	prefire_fuel
3	cwd_frac
4	duff_frac
5	day_of_year_sin
6	fm1000_value
7	srad_value
8	day_of_year_cos
9	bi_value
10	fm100_value

F. Elbow and Selection Rationale

The 2nd-derivative elbow suggested only two features, but this was deemed too aggressive given sharp early gains and the known robustness of tree ensembles to moderate feature counts. We retained 27 features as a practical balance, enabling permutation-based importance to adjudicate which features materially change predictions.

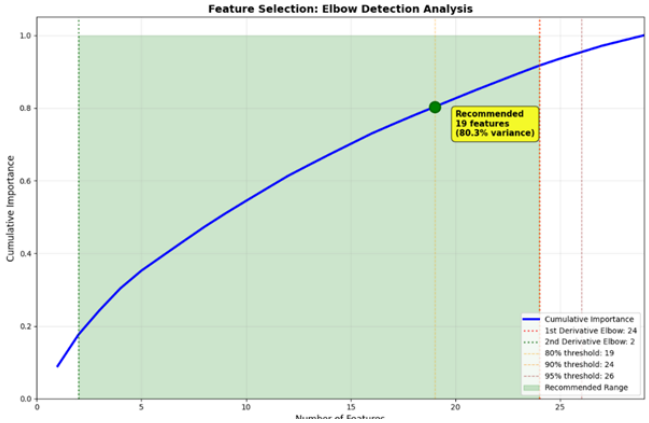

Fig. 5. Elbow detection for feature count selection. The second-derivative elbow suggested two features; we retained a larger set (27) based on first-derivative stability and ensemble tolerance.

TABLE VI
MODEL PERFORMANCE OVERVIEW

Model	R^2	RMSE	MAE
RandomForestRegressor	0.982	1.584	0.242
XGBRegressor	0.708	6.466	3.453
HistGradientBoosting	0.646	7.125	4.051

G. Model Performance

We report three standard metrics:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}, \qquad \text{MAE} = \frac{1}{n}\sum_i |y_i - \hat{y}_i|, \tag{1}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{2}$$

Early row-level splits overstated performance by mixing days from the same event; group-aware splits corrected this leakage and still confirmed learnability, with Random Forests leading among tabular baselines.

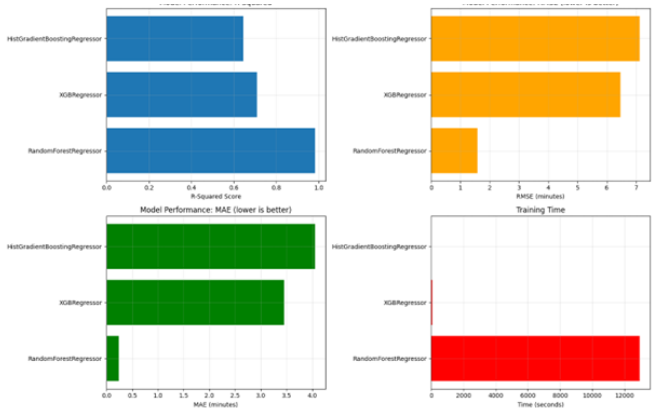

Fig. 6. Performance summary on the full dataset. Random Forest achieved the strongest R^2 but required group-aware splits to avoid fire-level leakage.

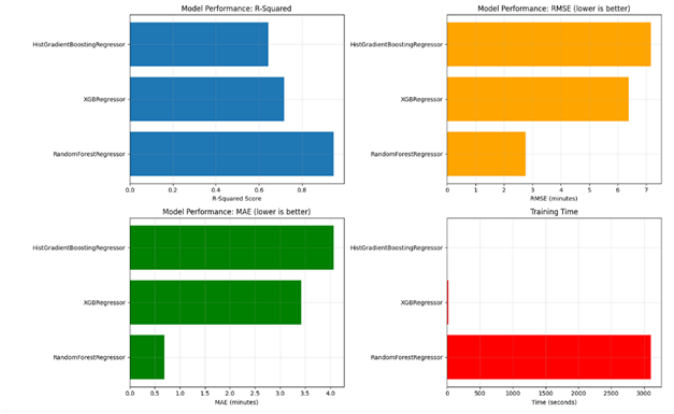

Fig. 7. Performance on a 25% sample. While metrics dipped relative to the full dataset, the ranking of top predictors remained stable.

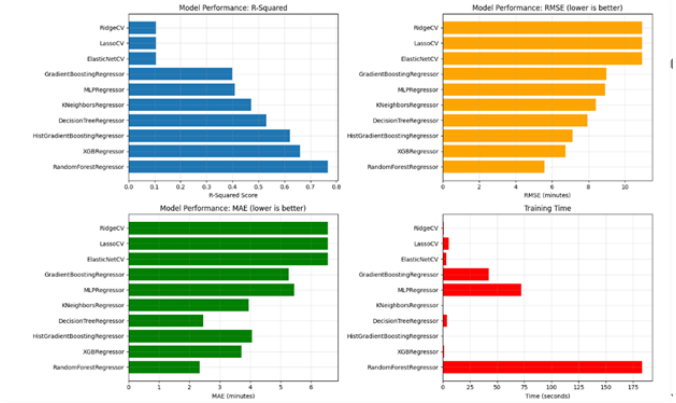

Fig. 8. Screening results across 40+ regressors (LazyPredict). Tree ensembles (RF, XGB, HistGB) consistently outperformed others, motivating their use as baselines.

H. Feature Importance

Built-in (RF/XGB) and permutation importances consistently elevate `duff_frac`, `cwd_frac`, `prefire_fuel`, `fm1000_value`, and location terms (longitude/latitude), matching the statistical rankings in Tables I–V. Agreement between model-agnostic and model-based views supports robustness.

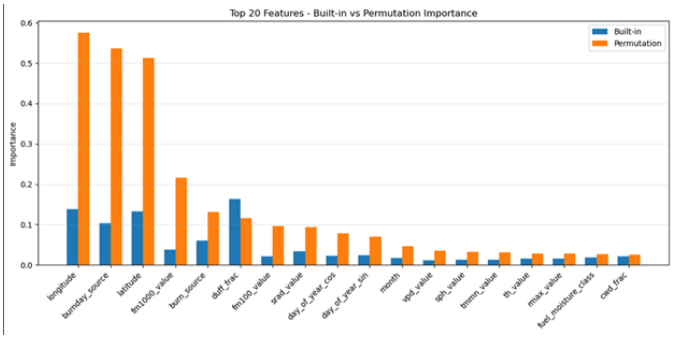

Fig. 9. Built-in feature importance from tree ensembles highlighting *duff_frac*, *cwd_frac*, *longitude*, *prefire_fuel*, and *fm1000_value*.

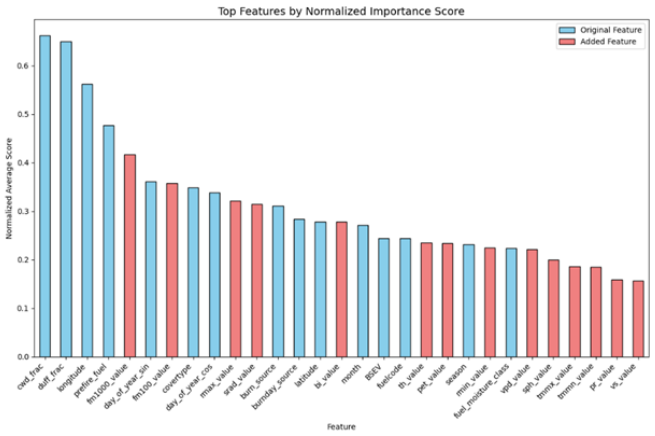

Fig. 10. Normalized, model-agnostic importance (e.g., permutation-based) confirming robustness of the top predictors across models and reducing single-model bias.

VII. Operational Use, Uncertainty, and Research Status

The methods and results presented here are part of a preliminary research effort intended to explore the feasibility of predicting wildfire duration from integrated emissions and meteorological datasets. While our pipeline demonstrates strong in-sample performance, cross-validation diagnostics revealed clear signs of overfitting, particularly for the Random Forest model, indicating that predictive accuracy is likely overstated when applied to unseen events.

For operational wildfire management, a more robust model would be required before deployment. Based on the unified MFLEI–gridMET dataset, a validated and generalizable duration prediction model could directly support:

- **Staging timelines:** Allowing incident management teams to anticipate the duration of required ground and aerial resources.
- **Mutual-aid requests:** Facilitating earlier interagency coordination to reduce delays during escalation.
- **Air-resource allocation:** Informing how long water- or retardant-dropping aircraft will be needed, enabling efficient flight-hour and fuel budgeting.

We recommend that any operational system output not only a point estimate but also a confidence interval or prediction interval, using methods such as quantile regression forests, conformal prediction, or bootstrapped ensembles. Incorporating live data feeds—near-real-time weather, updated fuel moisture indices, and fire detections, would enable rolling reforecasts, similar to operational numerical weather prediction systems.

VIII. Limitations and Recommendations for Future Work

Although the integrated dataset and baseline models establish that wildfire duration is learnable from emissions and meteorological predictors, there are important limitations:

- **Overfitting in current models:** The top-performing models achieved very high R^2 on group-aware folds but exhibited instability across feature subsets. This suggests that the current feature set may contain variables with limited generalization value. A key priority for the next research team should be rigorous feature selection and dimensionality reduction to improve robustness.
- **Wind direction and terrain:** Current predictors do not incorporate wind vectors aligned with spread direction or terrain attributes such as slope and aspect. Integrating these will require geospatial alignment of daily fire perimeters or rasterized burn footprints with topography and wind fields.
- **Generalization across regions and years:** While group-aware validation prevents leakage within the dataset, true generalization requires holding out entire eco-regions or years to simulate forecasting in novel climate regimes.
- **Temporal modeling:** Our predictors are aggregated at the event level. Building multi-day sequential inputs could enable spatiotemporal models (e.g., ConvLSTM, temporal transformers) that better capture persistence and escalation effects.
- **Uncertainty quantification:** Operational systems must provide uncertainty estimates to reflect rapidly changing fire and weather conditions.

IX. Conclusion

We developed and validated an end-to-end, event-centric pipeline for predicting wildfire duration from large-scale emissions and meteorological data. The workflow integrates:

1) Spatiotemporal clustering of MFLEI detections into event IDs via DBSCAN.
2) Conversion and integration of daily gridMET weather rasters into BigQuery tables.
3) Event-level feature engineering combining raw and derived meteorological and fuel metrics.

Preliminary results highlight duff consumption (`duff_frac`), coarse woody debris fraction (`cwd_frac`), prefire fuel load, large-fuel moisture indices (`fm1000`, `fm100`), and geographic coordinates as consistently strong predictors. However, observed overfitting underscores the need for refined feature selection, expanded validation strategies, and the inclusion of additional physical predictors before operational deployment. Future work should focus on these areas, alongside the development of rolling forecast capabilities that update as new observations arrive.

Beyond its immediate findings, this study's primary contribution is the creation of a reproducible, versioned, and extensible data-processing and modeling framework for wildfire duration analysis. By standardizing the ingestion, integration, and validation steps, the pipeline lowers the barrier for subsequent teams to test new algorithms, incorporate emerging datasets, and evaluate additional response variables. As such, the value of this work lies not only in its current predictive performance but in its role as a foundation for iterative improvement, collaborative extension, and eventual transition toward operational wildfire decision-support systems.

REFERENCES

[1] S. Urbanski, W. Hao, and S. Baker, "US wildfire emissions estimates using the Missoula Fire Lab Emissions Inventory (MFLEI): Data set and documentation," *USFS*, 2017. [Online]. Available: https://www.fs.usda.gov/rmrs/ (search: MFLEI). [Accessed: Aug. 8, 2025].

[2] J. T. Abatzoglou, "Development of gridded surface meteorological data for ecological applications and modeling," *Int. J. Climatol.*, vol. 33, no. 1, pp. 121–131, 2013. [Online]. Available: https://www.climatologylab.org/gridmet.html. [Accessed: Aug. 8, 2025].

[3] U.S. EPA, "Climate Change Indicators: Wildfires," 2023. [Online]. Available: https://www.epa.gov/climate-indicators/climate-change-indicators-wildfires. [Accessed: Aug. 8, 2025].

[4] Z. Hausfather, "How global warming has increased US wildfires," 2018. [Online]. Available: https://www.carbonbrief.org/factcheck-how-global-warming-has-increased-us-wildfires/. [Accessed: Aug. 8, 2025].

[5] A. Shadrin, R. Solovyev, and A. Kustov, "Wildfire spreading prediction using multimodal data and deep neural network approach," *Scientific Reports*, vol. 14, no. 1, 2024. https://doi.org/10.1038/s41598-024-52821-x

[6] National Interagency Fire Center, "Wildland Fire Statistics," 2024. [Online]. Available: https://www.nifc.gov/fire-information/statistics. [Accessed: Aug. 8, 2025].

[7] U.S. Forest Service, "Initial Attack Success Rates and Large Fire Cost Drivers," 2016. [Online]. Available: https://www.fs.usda.gov/. [Accessed: Aug. 8, 2025].

[8] National Wildfire Coordinating Group, "Incident Command System Operations Section Chief, ICS-420-1," 2022. [Online]. Available: https://www.nwcg.gov/. [Accessed: Aug. 8, 2025].

[9] W. M. Jolly, M. A. Cochrane, P. H. Freeborn, et al., "Climate-induced variations in global wildfire danger from 1979 to 2013," *Nature Communications*, vol. 6, 7537, 2015. https://doi.org/10.1038/ncomms8537

[10] National Wildfire Coordinating Group, "WIMS User's Guide, Appendix E: NFDRS Technical Reference," 2011. [Online]. Available: https://wildfireweb-prod-media-bucket.s3.us-gov-west-1.amazonaws.com/s3fs-public/2022-11/Appx_E_NFDRS_Technical_Reference.pdf. [Accessed: Aug. 8, 2025].

[11] J. T. Abatzoglou and A. P. Williams, "Impact of anthropogenic climate change on wildfire across western US forests," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. 11770–11775, 2016.

[12] Federal Emergency Management Agency, "ICS-100: Introduction to the Incident Command System," 2022. [Online]. Available: https://training.fema.gov/is/courseoverview.aspx?code=IS-100.c. [Accessed: Aug. 8, 2025].

[13] National Wildfire Coordinating Group, "Interagency Air Operations Guide," PMS 505, 2021. [Online]. Available: https://www.nwcg.gov/publications/505. [Accessed: Aug. 8, 2025].