Wildfire Prediction Using Weather Pattern Analysis

Matthew Presti
Computer Science
University of Colorado
Boulder, CO, USA
mapr2282@colorado.edu

Holly Schwecke
Computer Science
University of Colorado
Denver, CO, USA
hosc2215@colorado.edu

Haoye Tang
Computer Science
University of Colorado
Denver, CO, USA
hata6503@colorado.edu

Abstract

This study explores the complex relationship between weather patterns and wildfire events in California from 2011-2020. Using NOAA weather station data and NASA MODIS satellite wildfire detections, we developed a custom spatial integration methodology and applied two complementary data mining approaches: predictive modeling with XGBoost and unsupervised pattern discovery using UMAP and HDBSCAN clustering.

Our results achieved 80% accuracy in predicting fire intensity based on preceding weather conditions. Contrary to our initial assumptions, wind speed showed surprisingly limited predictive value for fire intensity. Instead, a combination of extended drought periods (days without rain), elevated temperatures, and low humidity emerged as the strongest predictors. No single weather factor dominated—rather, the interaction between multiple variables created conditions conducive to intense wildfires. The methodology developed provides a framework for early warning systems and could inform long-term planning for wildfire management in an era of changing climate conditions.

Introduction

Wildfires represent a growing threat to ecosystems, property, and human lives across the western United States. In recent decades, the frequency, severity, and size of wildfires have increased dramatically, with California experiencing several of its most destructive fire seasons on record. This trend has prompted urgent questions about how to better predict, prepare for, and manage wildfire risk in a changing climate.

Our research addresses three fundamental questions crucial to the advancement of wildfire science and management:

Can we identify specific weather pattern signatures that reliably precede wildfire ignition and intensity? Understanding the precise weather conditions that create high fire risk would enable more targeted early warning systems and resource deployment.

Which weather factors contribute most significantly to extreme fire behavior? While conventional wisdom often emphasizes factors like wind, comprehensive data mining might reveal surprising or counterintuitive relationships.

Can historical weather data help forecast overall fire season severity? The ability to predict particularly dangerous fire seasons months in advance would transform long-term planning and resource allocation.

These questions are increasingly important as climate change alters traditional weather patterns and extends fire seasons. The social and economic costs of wildfires have become staggering. Beyond economic impacts, wildfires threaten irreplaceable natural landscapes, wildlife habitats, and human communities.

While existing research has made progress in wildfire prediction, most approaches struggle to effectively integrate weather and fire data across

both spatial and temporal dimensions. Our project addresses this gap through innovative spatial integration methodologies and complementary data mining approaches. By extracting actionable insights from historical weather-fire relationships, we aim to contribute practical knowledge to the field of data-driven climate models and potentially help mitigate the devastating impacts of wildfires on communities.

Literature Survey

Several researchers have explored the relationship between weather conditions and wildfire activity, employing various data mining and machine learning approaches. Cortez and Morais (2007) applied neural networks and support vector machines to predict forest fire spread based on meteorological data[1]. Jaafari et al. (2019) compared multiple machine learning methods for wildfire susceptibility mapping, finding that ensemble approaches typically outperform individual models [2].

More recently, Jain et al. (2020) combined satellite information with weather data to predict wildfire risk, highlighting how important it is to properly organize and prepare weather pattern data for analysis [3]. Yu et al. (2022) built upon this approach by adding information about drought conditions and plant health. Their study showed better prediction results, especially when forecasting major wildfire events. They found that long-term drought measurements and seasonal vegetation changes were particularly useful indicators for predicting fire behavior [4].

These studies offer helpful approaches, though most struggle to effectively combine weather data and wildfire information across both location and time. Our project seeks to close this gap by developing better methods to integrate data from different geographic areas and by creating testing approaches that account for how wildfire conditions change over time.

Data Sources

NOAA Global Surface Summary of Day (GSOD)

The NOAA GSOD dataset originates from the National Climatic Data Center (NCDC) and USAF Climatology Center. It is available at https://www.kaggle.com/datasets/noaa/noaa-global-surface-summary-of-the-day. This comprehensive dataset covers over 9,000 weather stations worldwide, spanning from 1929 to the present with daily updates. At multiple millions of records and approximately 3.3GB in size, it provides extensive temporal and spatial coverage for our analysis. The dataset includes daily summaries of temperature (mean, max, min), wind (speed, gusts, sustained), precipitation and snow depth, pressure, visibility and dew point, and various weather events including fog, rain, snow, hail, and thunder.

We obtained a CSV version that includes latitude and longitude coordinates for each weather station, which significantly simplified our spatial integration approach and improved the precision of our weather-to-wildfire data mapping.

NASA MODIS Satellite Wildfire Data

The NASA MODIS satellite wildfire dataset is sourced from NASA MODIS satellite data, available via Kaggle at https://www.kaggle.com/datasets/avkashchauhan/california-wildfire-dataset-from-2000-2021. This dataset covers the period from 2000 through March 25th, 2022, focusing on a geographical range that encompasses the the state of California. This dataset provides comprehensive geographic wildfire data including fire detection points, dates, and confidence levels, allowing for detailed spatial-temporal analysis of wildfire patterns.

Data Processing Methodology

Our data processing workflow consisted of several stages designed to transform raw weather and wildfire data into a format suitable for pattern mining and predictive modeling. This multi-stage approach allowed us to effectively handle the spatial and temporal dimensions of the data.

Data Acquisition and Preparation

Both datasets required careful preprocessing before they could be integrated. For the NOAA GSOD dataset, we filtered to focus on weather stations in California, where our wildfire data was concentrated. We performed data cleaning to handle missing values, standardize units, and ensure consistent formatting. For the NASA MODIS wildfire data, we filtered to the study period (2011-2020) and extracted relevant attributes such as acquisition date, coordinates, and fire radiative power (FRP).

Wildfire Event Clustering

A critical innovation in our approach was the development of a clustering algorithm to transform individual satellite fire detections into coherent wildfire events. The MODIS data provided point-in-time fire detections, but these needed to be aggregated into actual fire events for meaningful analysis.

Our clustering approach grouped fire detections using both spatial and temporal proximity:

Spatial clustering: Fire detections within 10km of each other were considered part of the same potential event.

Temporal clustering: Only detections from the same month were grouped together to avoid creating artificially extended events.

This process converted approximately 7.8 million individual fire detections into a more manageable and meaningful set of distinct wildfire events, each with attributes such as:

Start and end dates: The temporal boundaries of the fire event.

Duration in days: How long the fire persisted according to satellite detections.

Geographic location: Latitude/longitude coordinates of the fire's most intense point.

Fire intensity metrics: Maximum, average, and total Fire Radiative Power (FRP).

Detection count: Total number of satellite detection points associated with the event.

The implementation used a spatial clustering algorithm that processed the data in several stages. First, we segmented the data by year and month to ensure temporal cohesion within potential events. For each temporal segment, we implemented an iterative clustering approach where an unassigned detection point would serve as the seed for a new event. We then calculated the Euclidean distance between this seed point and all other unassigned points in the same temporal window.

For computational efficiency, we approximated the geographic distance using a simplified conversion (1 degree ≈ 111km) with a cosine correction for longitude distances based on latitude. Points within the specified distance threshold (10km) were assigned to the same event cluster. This process continued until all points were assigned to an event cluster, resulting in a set of spatially and temporally coherent wildfire events. Each event received a unique identifier, allowing subsequent analysis to operate at the event level rather than individual detections.

Spatial Integration of Weather and Wildfire Data

One of the key challenges in our project was integrating point-based weather station data with wildfire event locations. We implemented a buffer-based spatial integration method that associated each wildfire event with the nearest weather stations:

Proximity identification: For each wildfire event, we identified the closest weather station(s) within a 50km radius.

Temporal data extraction: This approach allowed us to extract relevant weather data preceding each fire event.

Station prioritization: In cases where multiple stations were available, we prioritized the nearest station with complete data.

The implementation utilized a KDTree spatial indexing structure from SciPy to efficiently identify stations within a specified radius of each fire event. This data structure significantly improved computational efficiency compared to a brute-force approach, especially given the large number of weather stations (>9,000) and fire events.

We converted the 50km search radius to approximate degrees (dividing by 111km per degree) for compatibility with the coordinate system. For each wildfire event location, the algorithm queried the KDTree to find all weather stations within the specified radius. The IDs of relevant stations were stored in a set data structure to eliminate duplicates, creating an optimized list of stations relevant to our analysis. This spatial filtering step substantially reduced the amount of weather data to process while ensuring all relevant data for wildfire events was retained.

Feature Engineering

Our feature engineering approach focused on extracting meaningful weather patterns that preceded wildfire events. For each wildfire event, we collected weather data from the nearest station for the 30 days leading up to the fire start date. This temporal lookback window allowed us to capture the cumulative weather conditions that might have contributed to fire ignition and spread. We engineered a diverse set of features across several categories:

Temperature Features

Average temperature: Mean temperature over the 30-day period (avg_temp_30d).

Maximum temperature: Highest temperature recorded in the 30-day period (max_temp_30d).

Average daily maximum: Mean of daily maximum temperatures over the 30-day period (avg_max_temp_30d).

Precipitation Features

Total precipitation: Sum of all rainfall over the 30-day period (total_precip_30d).

Drought duration: Number of days without significant rainfall (< 0.01 inches) (days_without_rain).

Wind Features

Average wind speed: Mean wind speed over the 30-day period (avg_wind_speed). **Maximum wind speed**: Highest wind speed recorded in the period (max_wind_speed).

Humidity Features

Average dew point: Mean dew point temperature over the period (avg_dewpoint).

Derived Features

Temperature-precipitation ratio: A custom drought metric calculated as the ratio of average temperature to total precipitation (with a small constant added to avoid division by zero).

The feature engineering implementation involved several technical steps. First, we used a KDTree spatial data structure to efficiently find the nearest weather station for each wildfire event. For each fire event, we calculated the geographic distance to the nearest station and stored this as part of our feature set.

The temporal filtering required precise handling of datetime objects. We identified the fire's start date and filtered the station's weather data to include only records from the 30-day window preceding the fire start. Records from stations with insufficient data coverage (fewer than 5 days of data in the 30-day window) were excluded to ensure robust feature calculations.

For each qualifying fire-station pair, we calculated statistical aggregations (means, maximums, sums) of weather variables within the temporal window. The derived drought metric (temperature-precipitation ratio) required special handling to avoid division by zero errors by adding a small constant (0.1) to the precipitation sum. We also implemented threshold-based counting, such as for calculating days without significant rainfall (precipitation below 0.01 inches).

The process produced a feature vector for each wildfire event that represented the comprehensive weather conditions preceding the fire, creating a rich dataset for our subsequent pattern mining and prediction tasks.

Data Mining and Analysis Techniques

We employed two complementary data mining approaches to extract insights from our integrated dataset: (1) an XGBoost-based predictive modeling approach focused on fire intensity prediction, and (2) a dimension reduction and clustering approach aimed at identifying distinct weather patterns associated with wildfire events.

Predictive Modeling with XGBoost

For our primary analysis approach, we utilized gradient boosting via XGBoost, a powerful ensemble technique well-suited for this type of predictive task. The XGBoost model was trained to predict fire intensity (measured by Fire Radiative Power) based on the engineered weather features.

Our implementation included:

Data Preparation: Target variable was fire intensity categorized as "high" or "low" based on Fire Radiative Power (FRP). Features included the full set of engineered weather features. We used an 80% training, 20% testing split with temporal preservation (earlier years for training, later years for testing).

Model Training: We implemented an XGBoost regressor model with optimized hyperparameters. The model was trained using a standard 80/20 train/test split.

Feature Importance Analysis: We employed SHAP (SHapley Additive exPlanations) values to interpret feature contributions and conducted a decadal analysis to identify consistent predictors across years.

Model Evaluation: We assessed performance using classification metrics including accuracy,

precision, recall, and F1-score, complemented by confusion matrix analysis.

Dimension Reduction and Clustering with UMAP and HDBSCAN

As a complementary approach, we implemented unsupervised learning techniques to discover natural patterns in the weather data:

Dimension Reduction: We applied UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction of the weather data, preserving the high-dimensional structure while projecting to 2D for visualization.

Clustering with HDBSCAN: We utilized HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to identify distinct weather patterns. This density-based approach is capable of finding clusters of varying shapes and sizes.

Cluster Analysis: We cross-referenced clusters with wildfire events to identify high fire-rate clusters, calculated fire rates for each cluster (number of fire events / number of observations), and characterized clusters with radar charts to visualize feature profiles.

Comparative Analysis: We compared high fire-rate clusters with baseline clusters to identify distinguishing weather patterns and extracted insights about weather conditions most associated with wildfire events.

Results and Findings

Our data mining approaches yielded several key insights about the relationship between weather patterns and wildfire activity. We present these findings organized by analysis method.

XGBoost Predictive Modeling Results

The XGBoost model achieved strong predictive performance in classifying fire intensity based on preceding weather conditions:

Overall Model Performance: Classification accuracy reached 80%, with balanced precision and recall for both low and high-intensity fire classes. The model showed consistent performance across the decade-long study period (2011-2020).

Feature Importance Analysis: A central finding of our research is that no single dominant feature emerged as the primary predictor of fire intensity. Instead, a combination of features collectively contributed to fire risk prediction. As shown in Figure 1, top features (in approximate order of importance) included: days without rain (drought duration), average maximum temperature over 30 days, dew point (humidity indicator), average temperature, and temperature-precipitation ratio (derived drought measure). Notably, wind speed metrics (both average and maximum) showed surprisingly limited predictive value for fire intensity.

This absence of a dominant predictor underscores the complex, multifaceted nature of wildfire behavior. Fire intensity appears to be an emergent property arising from the interaction of multiple weather variables rather than being driven by any single factor. This complexity highlights why previous fire prediction efforts that focused too narrowly on individual weather elements have possibly fallen short.

The model's feature importance distribution was relatively flat compared to many predictive models in other domains, where one or two features typically dominate. In our case, the top five features contributed relatively similar importance weights, suggesting that accurate fire prediction requires monitoring a constellation of

weather conditions rather than relying on simplified metrics.

Relationship Between Weather Variables and Fire Intensity: Our analysis revealed nuanced relationships between weather variables and fire intensity. As illustrated in Figure 2, there is a clear positive correlation between temperature metrics and fire intensity. Strong negative correlation between precipitation/humidity measures and fire intensity was also observed. Days without rain emerged as particularly important, with longer dry periods strongly associated with more intense fires. The custom temperature-precipitation ratio showed good predictive power, suggesting the importance of the balance between heat and moisture.

These correlations, however, are not straightforward linear relationships. Our model captured complex interaction effects between variables. For instance, the impact of temperature on fire intensity was magnified after extended periods without rain, suggesting a compounding effect when multiple risk factors align. This interaction effect helps explain why fire behavior can sometimes seem unpredictable when assessed through the lens of individual weather measurements.

Wind Speed and Fire Intensity: Contrary to preliminary expectations, wind speed did not show a strong correlation with fire intensity in our model, as demonstrated in Figure 3. This finding held true for both average wind speed and maximum wind speed measurements. While wind may play a critical role in fire spread, our analysis suggests it is not as significant a factor in predicting fire intensity (FRP) compared to temperature and drought-related variables.

This counter-intuitive finding challenges conventional wisdom about wildfire behavior, which often emphasizes wind as a primary driver. Our results suggest that the role of wind may be more context-dependent than previously thought, potentially becoming significant only when other conditions (drought, temperature) have already created favorable burning conditions. This finding has important implications for fire risk assessment systems that might currently overweight wind factors.

UMAP and HDBSCAN Clustering Results

The unsupervised learning approach revealed distinct weather patterns with varying associations to wildfire activity:

Cluster Distribution: UMAP successfully reduced the dimensionality of the weather data while preserving meaningful structure, as shown in Figure 4. HDBSCAN identified approximately 55 distinct weather pattern clusters. Two dominant clusters (55 and 31) contained the majority of weather observations (334,137 and 325,216 observations respectively).

The identification of these distinct weather patterns suggests that California's climate can be effectively categorized into meaningful regimes, some of which create significantly higher fire risk. This structured approach to weather pattern classification offers a valuable framework for long-term climate and fire risk planning.

Fire Rate Analysis: Clusters 35 and 49 showed notably higher fire rates compared to baseline clusters. These high fire-rate clusters were relatively small (30 and 46 observations respectively). The fire rates for these clusters were 0.733333 and 0.608696, much higher than the average fire rate across all clusters.

What's particularly notable is that these high-risk weather patterns represent relatively rare conditions, occurring in less than 0.1% of overall weather observations. This rarity helps

explain why catastrophic fire conditions can seem to emerge suddenly—they represent uncommon but distinctive weather patterns that dramatically elevate fire risk. The identification of these specific high-risk patterns could significantly improve targeted early warning systems.

Cluster Characterization: Radar chart analysis revealed the distinctive profiles of high fire-rate clusters, as illustrated in Figure 5. Cluster 35 was characterized by elevated temperature, low precipitation, low humidity (dew point), and moderate wind speeds. Cluster 49 showed similar drought conditions but with some variations in the exact pattern of temperature and humidity.

The radar chart visualization helps quantify exactly how these high-risk weather patterns differ from typical conditions. Rather than showing extreme values on any single dimension, these high-risk clusters represent a specific constellation of moderate-to-high risk factors across multiple dimensions. This multifaceted profile again underscores why simplistic, single-variable approaches to fire risk assessment often fall short.

Comparative Evaluation: When comparing with the XGBoost results, the clustering approach identified similar weather factors associated with fires. However, the clustering approach was less effective for predictive purposes compared to the directed XGBoost approach. The greatest value of the clustering approach was in visualizing and characterizing distinct weather patterns.

The complementary nature of these two approaches—supervised and unsupervised learning—strengthens confidence in our findings. When independent methodologies point to similar conclusions about the relationship between weather patterns and fire risk, it suggests the identified patterns are robust rather than artifacts of a particular analytical approach.

Methodological Assessment: The clustering analysis provided complementary insights but was deemed less suitable than predictive modeling for the specific task of fire intensity prediction. High fire-rate clusters had relatively few observations, limiting statistical power. Nevertheless, the cluster analysis helped confirm key relationships identified in the XGBoost approach.

Applications and Implications

Our findings have several practical applications for wildfire management, risk assessment, and climate model development.

Scalable and Modular Analysis Framework

Year-Specific Analysis: Our framework is designed to be highly modular, allowing for analysis of specific years or ranges of years independently. This modularity enables both historical analysis and ongoing operational implementation as new data becomes available.

Configurable Parameters: The system incorporates controllable parameters—including the temporal lookback window and various model hyperparameters—that can be adjusted based on specific use cases or geographical regions. This flexibility allows the approach to be tailored to different fire regimes or adapted as climate conditions evolve.

Extensible Architecture: The pipeline's component-based design facilitates future enhancements, such as integration of additional data sources or implementation of alternative modeling approaches. Each module (data acquisition, spatial integration, feature

engineering, and prediction) can be independently optimized or upgraded.

Short-term Fire Risk Assessment

The predictive model developed in this study could be integrated into existing fire risk assessment systems to improve early warning capabilities:

30-Day Early Warning System: By monitoring the specific weather patterns identified in our study, fire management agencies could receive advanced warning of potentially high-intensity fire conditions. The 30-day lookback pattern suggests that effective early warning is possible by tracking cumulative weather conditions.

Targeted Resource Allocation: During periods when multiple fire risk factors align (extended drought, high temperatures, low humidity), fire management agencies could strategically position resources in high-risk areas. This preemptive approach could reduce response times and improve containment of new fire starts.

Risk Communication: Our findings on the specific weather conditions most strongly associated with intense fires could improve public communications about fire danger. More precise messaging could encourage appropriate precautionary behaviors during high-risk periods.

Computational Efficiency: Through our custom spatial integration methodology, the system achieves significant computational optimization that enables processing of large-scale datasets. This efficiency makes the approach viable for operational deployment in resource-constrained environments.

Hyperparameter Optimization Potential: While our current implementation demonstrates strong predictive performance, there remains considerable potential for further optimization through systematic exploration of model hyperparameters and feature engineering approaches. This represents a promising direction for future research and operational refinement.

Integration with Climate Models

Our findings on weather-fire relationships could be integrated with broader climate modeling efforts:

Expert Model Component: The predictive model could serve as an expert component within larger Earth system and climate models. This would improve the representation of wildfire dynamics in climate projections.

Scenario Analysis: Using our model, various climate change scenarios could be assessed for their implications for future fire regimes. This could inform both mitigation and adaptation planning.

Research Extension: Our methodology provides a foundation for expanded research incorporating additional variables such as vegetation type, land use, and human activity patterns. The integration framework developed for weather and fire data could be adapted for other environmental data sources.

Conclusion and Future Work

This study has successfully mined the relationship between weather patterns and wildfire activity in California, yielding valuable insights about the specific weather conditions that precede high-intensity fire events. Our multi-method approach combining predictive modeling with exploratory clustering revealed consistent findings about the importance of extended drought periods, elevated temperatures, and low humidity in creating conditions conducive to intense wildfires.

The most significant contribution of this work is the identification of specific weather pattern signatures that reliably precede wildfire events, providing actionable insights for fire management agencies. By demonstrating that no single weather factor dominates fire prediction, but rather that a combination of interacting factors drives fire intensity, our work highlights the complexity of wildfire-weather relationships and the importance of comprehensive monitoring approaches.

Limitations

Several limitations to our current approach suggest directions for future work:

Temporal Range: Our decade-long analysis period (2011-2020) may not capture longer-term climate cycles relevant to fire regimes.

Data Resolution: The point-based nature of weather station data means some spatial interpolation was necessary, potentially missing localized weather patterns. Daily weather summaries may miss sub-daily extreme conditions that influence fire behavior.

Additional Variables: Our analysis focused primarily on weather variables and did not incorporate vegetation condition, human activity, or landscape factors that also influence fire behavior.

References

[1] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," in New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, 2007, pp. 512-523.

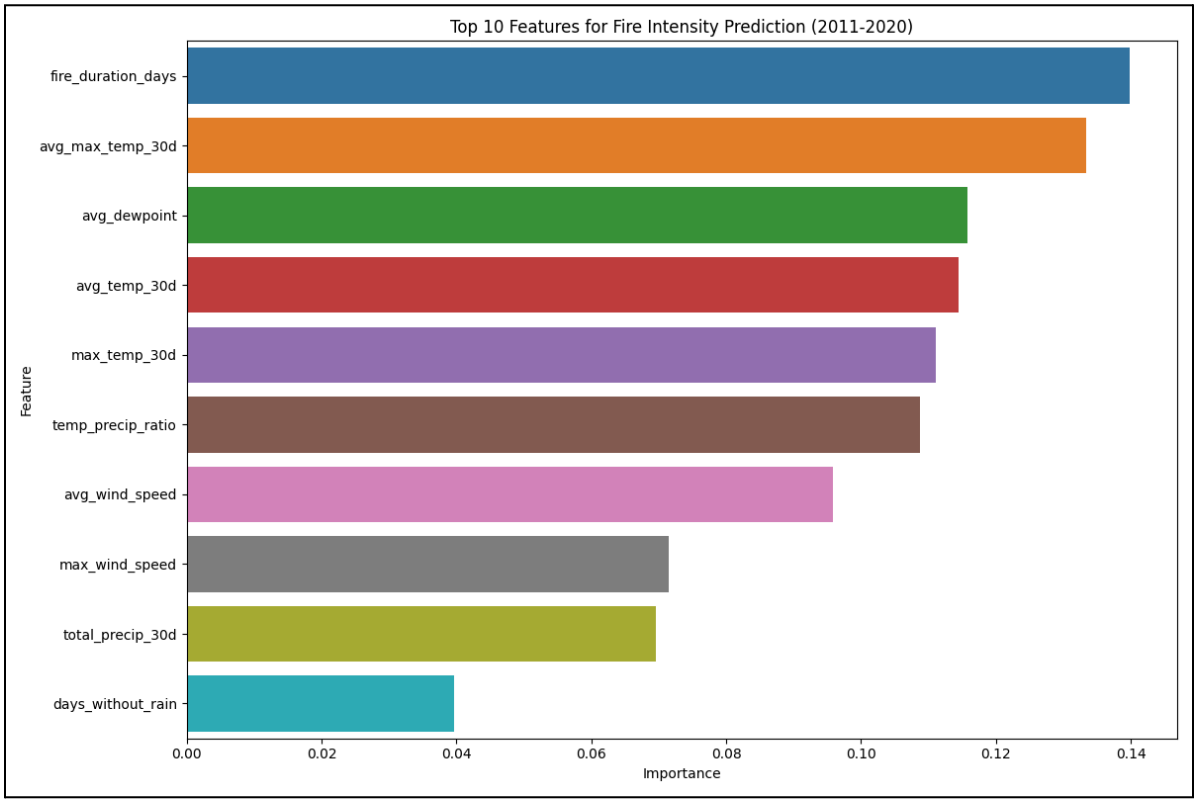
[2] A. Jaafari, E. K. Zenner, and B. T. Pham, "Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers," Ecological Informatics, vol. 43, pp. 200-211, 2018.

[3] P. Jain, S. C. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan, "A review of machine learning applications in wildfire science and management," Environmental Reviews, vol. 28, no. 4, pp. 478-505, 2020.

[4] Y. Yu, J. J. Sharma, A. C. Fouilloux, R. Strandberg, and H. F. Zhou, "Machine learning for weather-related wildfire prediction: A methodological review," Environmental Research Letters, vol. 17, no. 3, 2022.

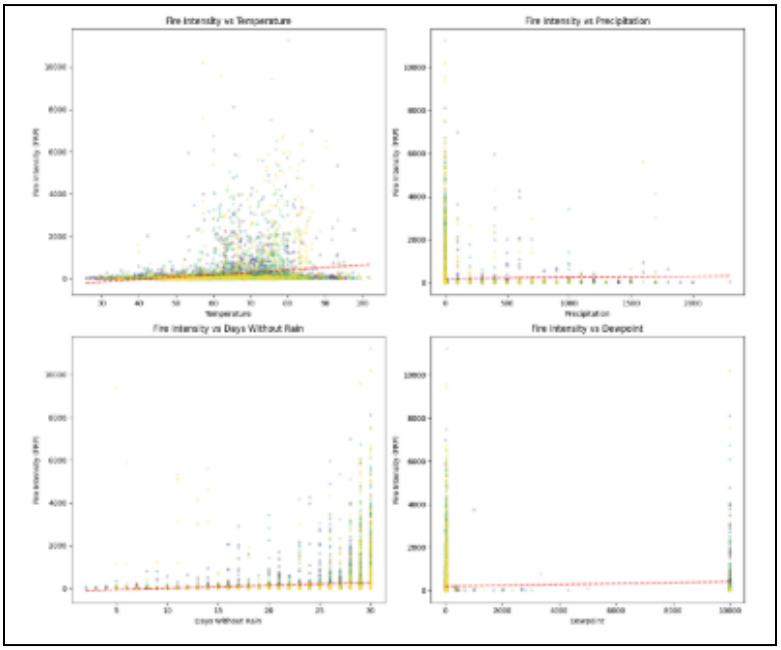
Visualizations:

Figure 1:



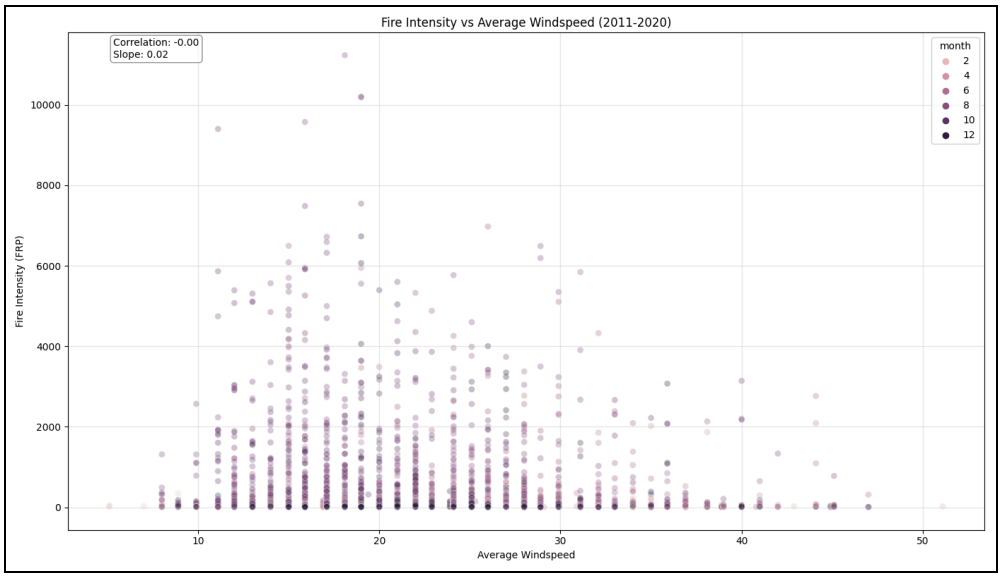
Top 10 Features for Fire Intensity Prediction (2011-2020) using XGboost

Figure 2:



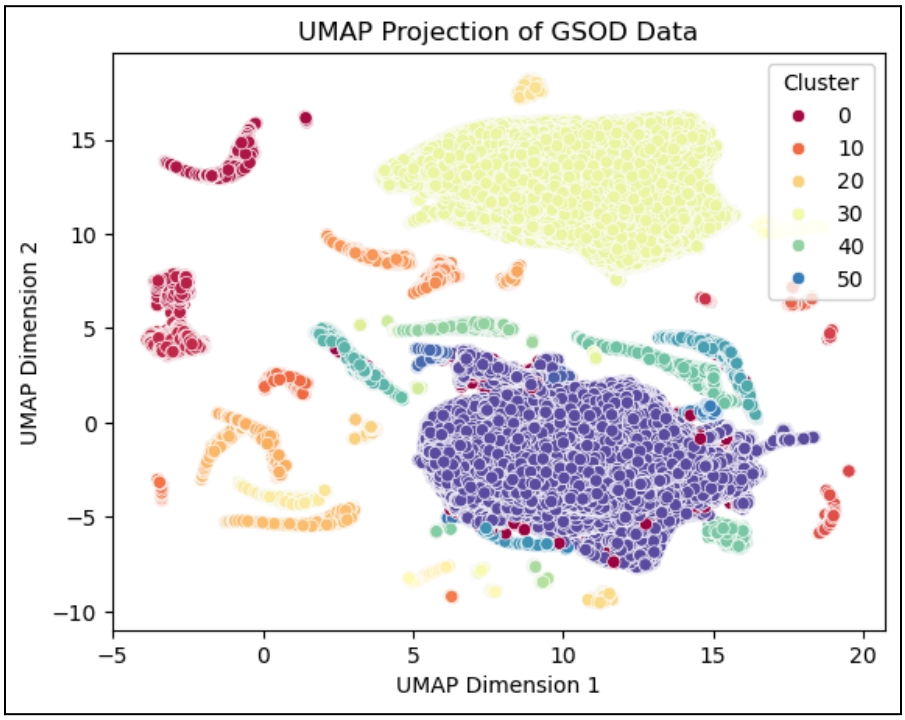
Fire Intensity vs Weather Variables Scatter plot showing relationships between fire intensity and temperature, precipitation, days without rain, and dewpoint.

Figure 3:



Fire Intensity vs Average Windspeed (2011-2020) Scatter plot showing the relationship between fire intensity (FRP) and average wind speed.

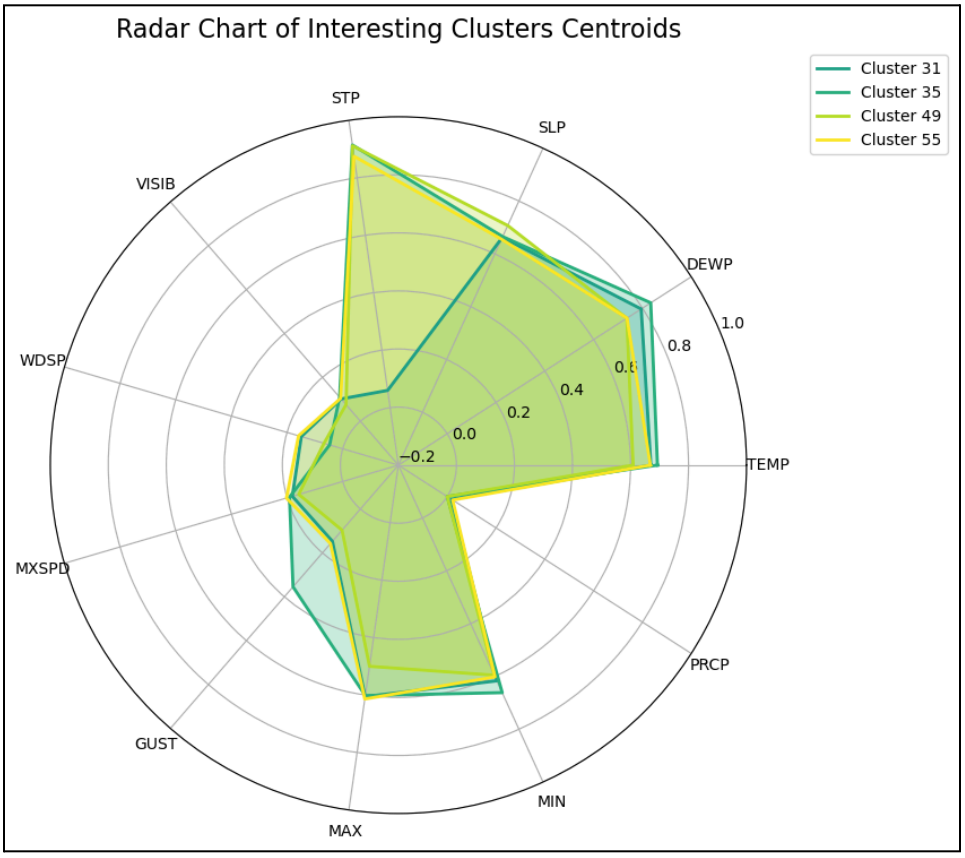
Figure 4



cluster	
55	334137
31	325216
3	10062
2	7500
45	6659

UMAP Projection of GSOD Data Two-dimensional UMAP projection of weather data showing the clustering of approximately 55 weather pattern types, with cluster size annotations indicating the dominance of clusters 55 and 31.

Figure 5:



	fire_rate	cluster_size
35	0.733333	30
49	0.608696	46
19	0.391892	74
30	0.333333	90
4	0.179856	139

Radar Chart of Interesting Clusters Centroids Radar chart comparing feature profiles of high fire-rate clusters (35 and 49) to baseline clusters.