D606 Capstone Report

Wildfire Ignition Risk in the Lake Hodges Watershed Region: Machine-Learning Classification and Comparison to Seismic Dam-Failure Risk

Author: Emilio Piraino

Date: April 29, 2026

Program: D606 Data Analytics Capstone

A. Research Question

Summary and Justification

Does a machine learning classification model predict wildfire ignition with accuracy significantly greater than chance when trained on historical fire occurrence and meteorological data in the Lake Hodges watershed region, and how does that observed annual wildfire frequency compare to the seismic dam failure probability that is used to justify restricting the reservoir elevation?

Section 1 is a yes/no question with a clean null hypothesis (e.g., AUC-ROC ≤ 0.5) and an alternative hypothesis (AUC-ROC > 0.5). Section 2 puts the eventual wildfire-frequency result into the context of multi-hazard considerations relevant to managing reservoir water supply in the region. The City of San Diego currently caps the reservoir elevation at Hodges Dam at 280 feet, which is 20 feet below the crest of Lake Hodges' concrete spillway, due to lingering seismic safety concerns with the 1918 structure.

"Lake Hodges watershed region" operational definition. Strictly speaking, the Lake Hodges watershed itself is quite small relative to the spatial granularity of both weather data and Fire Program Analysis – Fire Occurrence Database (FPA-FOD) fire-occurrence datasets. To retain an adequate sample size of ignitions for meaningful statistical analysis while still restricting the study area to the watershed "catchment," this project will operationally define the "Lake Hodges watershed region" as a 25-mile radius around Lake Hodges (33.0636°N, 117.1256°W). This approximately circular geographic region captures both the Lake Hodges watershed and the inland-western reaches of the San Diego climate zone for which Lake Hodges (meaning: the ERA5 grid cell containing Lake Hodges) is climatologically predictive. The code that implements this buffer is in src/03_prepare_data.py. It uses Pythagoras to filter each fire's latitude/longitude in FPA-FOD against the haversine distance from Lake Hodges' lat/long; the radius of this buffer is captured as a variable in src/config/settings.py as BUFFER_RADIUS_MILES and is set to 25.

San Diego County's unique geography makes it one of the most fire-prone counties in the country. Located at the convergence zone of Mediterranean climate along the coast, inland desert heat, and topographically complex mountain ranges, wildland fire weather is somewhat predictable in terms of timing and seasonality. Throughout late summer and into fall, hot and dry offshore wind events known as the Santa Ana Winds can rapidly decrease humidity levels, reaching extremely low relative humidity as they draw air off the ocean. Large fire years include 2003 and 2007, when the Cedar Fire and Witch Creek fires, respectively, burned over 280,000 acres, destroyed thousands of homes, forced hundreds of thousands of otherwise wildfire-safe residents to evacuate, and caused billions of dollars in damage. Reliable prediction of high-fire-risk days is valuable for public safety, as human-caused climate change is expected to lengthen, intensify, and dry out summer seasons across the southwestern United States.

Where this project's research question applies to the retrospective prediction of wildfire ignition risk, operational fire weather prediction already occurs daily at agencies such as the National Weather Service and CAL FIRE. Meteorologists at these agencies rely on indices such as the Haines Index (NHCF 2022) and the Energy Release Component (NWS Los Angeles Office 2022) to report daily wildfire danger to the public. According to the National Weather Service's Fire Weather website, some "combinations of low relative humidity, strong winds, warm temperatures, and critical fuels can contribute to extreme fire behavior." (National Weather Service 2026). Each of these indices combines multiple meteorological variables via rule-based (if/else) logical statements to calculate a final value representing relative fire danger on a given day. Training a supervised machine learning binary classifier on meteorological variables and known fire occurrences can model interactions among weather variables not represented in these rule-based systems, enabling integration into short-term forecasting to augment existing deterministic predictions.

Earthquake risk is included in the second section of the research question above because earthquakes (and subsequent dam failure, landslides, etc.) are another key factor affecting land- and water-management decisions near Lake Hodges. Specifically, seismic dam-failure risk led to the decision to maintain the reservoir elevation at Hodges Reservoir at or below 280 feet since 2017. This decision was made in light of the findings and risk management recommendations in the Hodges Dam SPRA (Safety Performance Assessment Report) recently completed by GEI Consultants and HDR Engineering (GEI Consultants 2022; HDR Engineering, 2022).

Study Area Background

Lake Hodges is one of five reservoirs that together comprise the City of San Diego's Nine Curve Portfolio. With a surface area of just under 1,000 acres and double the average depth of the other Nine Curve reservoirs, Lake Hodges is presently restricted to an elevation that results in about 150 acres of total surface water. This is because state regulators (DWR) have required that the City maintain at least a 20-foot freeboard at Hodges Dam until further notice; City engineers and consultants have interpreted this to mean retaining permanent pool at or below an elevation of 280 feet above mean sea level. Repeated attempts by the City to get approval from DWR to exceed 280-feet have been rejected because updates to the dam's seismic rating have been delayed by prolonged controversies regarding (1) whether/how to retrofit the dam to resist earthquake-induced failure appropriately, and (2) whether or how much the City would pay to have that retrofit completed by a private entity.

Dam safety with respect to earthquakes has been an open concern for the City of San Diego since before the dam was completed in 1918. The geologic formation underlying Lake Hodges and much of northeastern San Diego County is a massive fault cliff carved by the Pine Creek Segment of the San Andreas Fault. Completed hillslope stabilization and important improvements to spillway design in 1918 still left significant unknowns about the dam's true ability to survive a large-magnitude earthquake. After decades of gradually increasing uncertainty about the dam's seismic performance (e.g., 1988 FEMA 1994), City leaders commissioned an independent panel of engineers to perform a probabilistic risk assessment of the dam and identify key opportunities to reduce risk should the dam fail during a credible earthquake.

Risk reduction measures (RRMs) proposed by the expert panel include a wide range of infrastructure and policy levers that may increase the amount of storage available in Lake Hodges (and at the other City-owned dams in the Nine Curve Portfolio). With the completion of these RRMs, the City intends to pursue DWR approval to raise the top of the maximum pool at Lake Hodges from its current level of 280 feet to perhaps 290 feet or more, enabling an additional 50 acres of dynamic supply. Many of the RRMs are eligible for state or federal funding, either through direct grant mechanisms or through a catastrophic insurance policy held by the City. Still, progress on completing RRMs at Hodges Dam has been slow compared to other City-managed facilities intended to increase water supply or reduce flooding impacts. This project's paired research questions were designed to more directly link wildfire risk on the landscape to ongoing decision-making about reservoir levels at Lake Hodges.

Hypotheses

Null hypothesis (H_0): A machine learning classifier trained on historical wildfire occurrence and weather data for the Lake Hodges watershed region will not achieve predictive performance significantly better than chance (AUC-ROC ≤ 0.5).

Alternate hypothesis (H_a): A machine learning classifier trained on historical wildfire occurrence and weather data for the Lake Hodges watershed region will achieve predictive performance significantly better than chance (AUC-ROC > 0.5).

A secondary working hypothesis holds that thermodynamic and moisture variables — specifically maximum daily temperature and minimum daily relative humidity — will emerge as the dominant predictors, and that the Santa Ana wind composite feature will contribute disproportionately to model performance during the autumn peak-fire-season period.

B. Data Collection

Description of Relevant Data Collected

1. USFS Fire Program Analysis Fire-Occurrence Database (FPA-FOD), Version 6

The Fire Program Analysis Fire Occurrence Database (FPA FOD) is a national repository of fire-occurrence records maintained by the U.S. Forest Service. Developed by harmonizing federal, state, local agency fire reporting systems into one SQLite database, the FPA-FOD includes information on 1.88 million wildfires across the U.S. from 1992-2020 (Short 2022). Variables used from the raw dataset include Discovery Date, Fire Name, Latitude/Longitude, FIPS County Code, Acres Burned, Size Class, and General Cause. Queries were first run to isolate fires that occurred within San Diego County (FIPS Code 06073), then spatially subset to include only records within the Lake Hodges 25-mile haversine buffer described in Section A. The county-wide query returned 6,648 fire records, which was reduced to 3,700 records (55.7%) with the buffered-region Boolean query. Dates of these records were checked to confirm that Fire Program Analysis coverage of San Diego County actually begins in 1998, not the nominal 1992 beginning of the study period; all annual probabilities below are calculated across the full, empirical range of 1998–2020 (n=23 years).

2. Open-Meteo ERA5 Reanalysis Weather Dataset

Daily ERA5 reanalysis weather data was acquired from Open-Meteo at the latitude-longitude coordinates for Lake Hodges (33.0636° N, 117.1256° W), an inland site roughly near the geographic center of San Diego County in the western region of the fire study area. Open-Meteo provides free access to historical weather data from ERA5 and other reanalysis products, and ERA5 is a global atmospheric reanalysis generated by the European Centre for Medium-Range Weather Forecasts (Zippenfenig 2023; Hersbach et al. 2023). Retrieved variables are daily maximum temperature (°F), daily minimum temperature, daily mean temperature, daily minimum relative humidity (%), daily mean relative humidity, daily maximum wind speed (mph), daily mean wind speed, predominant daily wind direction (degrees), and daily maximum wind gust speed (mph). ERA5 records are continuous and complete for every day of the study period 1992-2020.

3. Lake Hodges Dam seismic annualized probability benchmark

Unfortunately for the annualized-probability side of our research question, neither the City of San Diego Public Utilities Department nor the California Division of Safety of Dams (CA-DSOD) have published an annualized probability of seismic-induced dam failure for Lake Hodges Dam. The decision by CA-DSOD to restrict Lake Hodges Dam reservoir levels to 280-foot elevations was thus based on *consequence* of failure, not annualized probability of failure:

directing the City to perform dam-breach modeling to justify or refute their proposed El 280 limit, CA-DSOD stated "the probability of dam failure should not be a consideration in establishing the reservoir level restriction… Rather the consequences should be based on an assumed major failure of the arch barrels in the valley section of the dam" (GEI Consultants 2022, p. 3). The engineering-consequence study completed by HDR (GEI Consultants 2022; HDR Engineering 2022, Consequence Analysis for Dam Breach Scenarios for Various Reservoir Elevations, Attachment A to GEI letter) modeled five fair-weather, seismic-induced breach scenarios using the five-reservoir-elevation US Bureau of Reclamation Reclamation Consequence Estimating Methodology (RCEM; USBR 2015). Modeled potential life loss (PLL) due to a 7-panel valley-section breach — deemed the "reasonable worst case" by all parties — is category 1–10 for reservoir elevations of 275 and 280 feet, but category 3–30 at El 285 feet (HDR 2022, Tables 1, 12, 13); estimated consequences at El 280 are a best-estimate PLL of 4 against an average population-at-risk (PAR) of 666 people living downstream, with peak breach outflow of 195,000 cubic feet per second (cfs) and 6,015 acre-feet of water released. GEI's recommendation to mandate no greater than El 280 as a maximum (GEI 2022, p. 5) is explicitly couched as a means to minimize potential consequences, not annual probability of occurrence.

The U.S. Bureau of Reclamation has established 1×10^{-4} per year as the tolerable-risk threshold for life-safety-related failure modes of high-hazard dams (USBR 2022, Public Protection Guidelines: A Risk-Informed Framework). This criterion is the foundation for reservoir restriction policies nationwide, has been formally adopted by FERC and ANCOLD as well as the Canadian Dam Association, and is incorporated into CA-DSOD enforcement actions. California DSOD finds Hodges Dam to be "unsatisfactory" with "extremely high" downstream hazard potential. In lieu of a published annualized failure probability, this project will use 1×10^{-4} per year as the publicly-known, auditable benchmark for seismic-induced dam failure at Lake Hodges and clearly identify it as a regulatory tolerable-risk threshold, not an actuarially-determined failure probability for Hodges Dam. Consequence values from GEI/HDR (2022), namely PLL category 1–10 at El 280 and average PAR 666, are employed in Section D.5 with this threshold.

Advantage of the Data-Gathering Methodology

The main benefit of this method is that it is complete and consistent. ERA5 reanalysis fills in the gaps between weather stations to create a spatially/temporally continuous record. It does so by assimilating observations into a numerical weather model. That means no quality-control issues that plague individual weather station records. Real RAWS data regularly misses observations; stations relocate, and sensors break, requiring imputation. ERA5 covers every day of your study period, allowing you to combine your fire and weather time series with minimal data loss.

Disadvantage of the Data-Gathering Methodology

The trade-off lies between spatial resolution and representativeness. ERA5 data are output at roughly 0.25° horizontal resolution, so one grid cell covers weather over a wide geographic area with varied topography (Open-Meteo, 2026; Hersbach et al., 2023). The 25-mile Hodges buffer ranges from near sea level to about 4,000 feet in elevation. Also, fire-weather conditions can change significantly over distances smaller than the buffered area. Conditions can be ripe for a Santa Ana wind event inland in a valley location like Ramona, while simultaneously, conditions near the coast may be quite mild. One ERA5 point near Lake Hodges may not be representative of fire-weather conditions throughout the buffered region.

Overcoming Challenges

The original methodology used the MesoWest/Synoptic API to pull observations from RAWS stations at Ramona Airport (station identifier RHDC1, later updated to PSQC1 for San Pasqual). There were two main issues with this method: the RAWS datasets for these stations had substantial missing data early in the 1992–2020 record, and the rate limits on the API queries were slow. These problems were mitigated by switching the weather data source to the Open-Meteo ERA5 reanalysis API. It requires no auth token, imposes no rate limits when querying their historical archive, and returns the full contiguous record in a single structured query. The FPA-FOD database could be downloaded as a single compressed archive (a ZIP containing SQLite) directly from the USFS Research Data Archive with no authentication required, removing the API dependency for fire records as well.

C. Data Extraction and Preparation

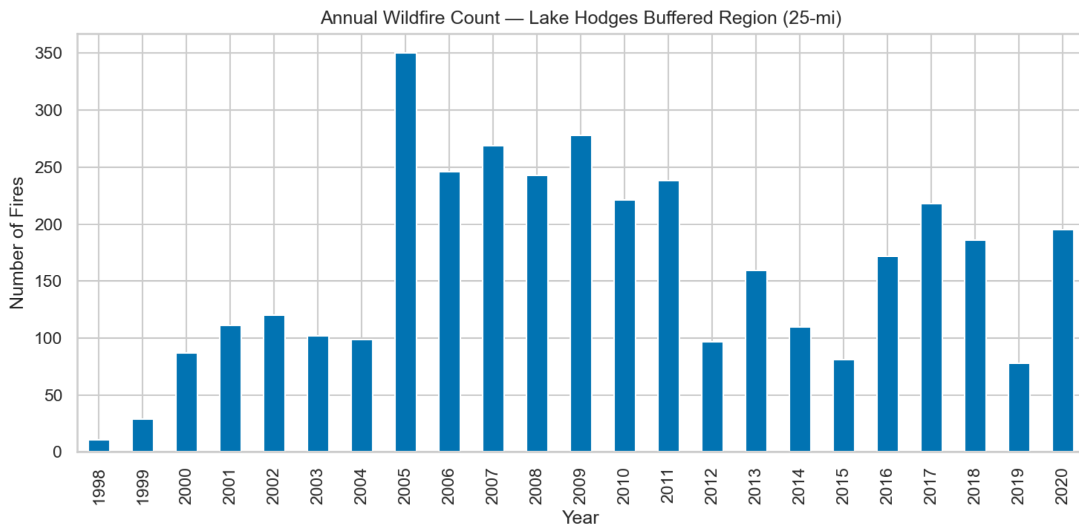
Data Extraction

Extraction was implemented in two pipeline scripts.

Script 01 — FPA-FOD Download (`src/01_download_fpa_fod.py`): The USFS archive ZIP file was downloaded programmatically using Python's `requests` library and extracted to `data/raw/`. The SQLite database was then queried using SQLAlchemy with the following filter applied at the database level to minimize memory usage:

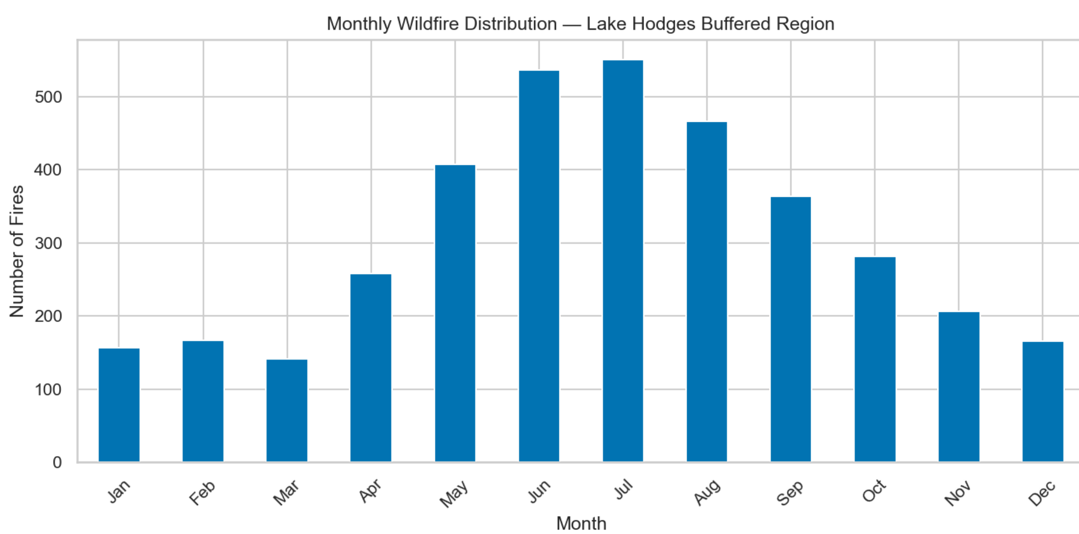
```
SELECT FOD_ID, FIRE_NAME, DISCOVERY_DATE, DISCOVERY_DOY,
       NWCG_GENERAL_CAUSE, FIRE_SIZE, FIRE_SIZE_CLASS,
       LATITUDE, LONGITUDE, STATE, FIPS_CODE, FIPS_NAME
FROM Fires
WHERE FIPS_CODE = '06073'
  AND FIRE_YEAR BETWEEN 1992 AND 2020
ORDER BY DISCOVERY_DATE
```

This produced a San Diego County fire records table loaded directly into a pandas DataFrame.



Screenshot C1. Annual fire counts generated after extracting and date-filtering the FPA-FOD fire records.

Script 02 — Weather Fetch (`src/02_fetch_weather.py`): The Open-Meteo ERA5 archive API was queried via HTTP GET request for the Lake Hodges coordinates over the full 1992–2020 date range. The JSON response was parsed into a daily DataFrame and saved to `data/raw/weather_daily_LAKE_HODGES_ERA5.csv`.



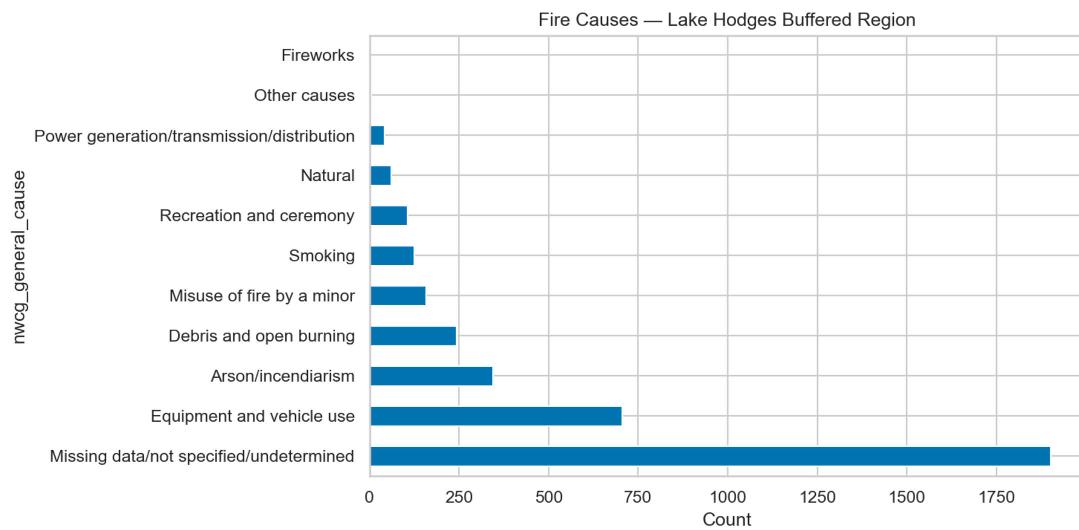
Screenshot C2. Monthly distribution check used to validate the merged fire-date and weather-date time series.

Data Preparation

Preparation was implemented in `src/03_prepare_data.py` across six sequential steps.

Step 1a — Spatial filtering to the Lake Hodges buffered region: Prior to temporal aggregation, per-fire records were clipped to include only those within 25 miles of the Lake Hodges anchor point (33.0636°N, 117.1256°W). Lakes Basin Analysis solution includes vectorized great-circle-

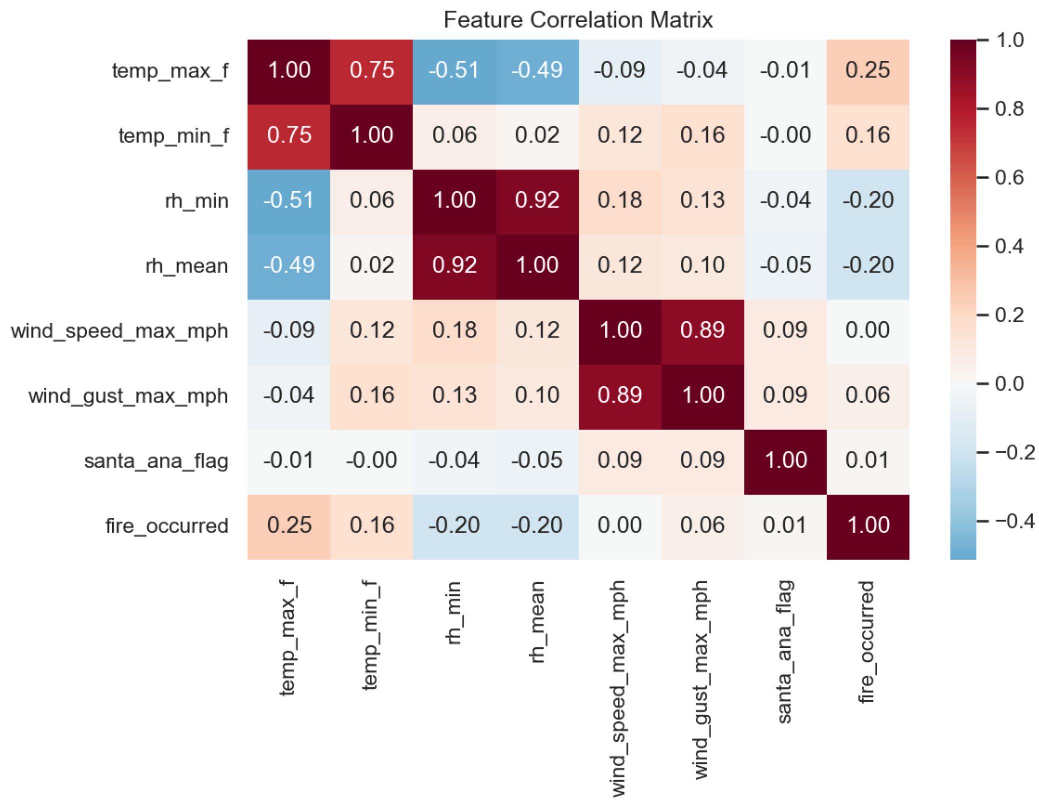
distance helper function (haversine_miles defined in src/03_prepare_data.py) and a centralized BUFFER_RADIUS_MILES constant declared in config/settings.py. Completion of this step reduced the fire population from 6,648 county-level fires to 3,700 fires associated with the watershed region (55.7% retention). Filtered lake-hodges-region records exported to data/processed/hodges_region_fires.csv for downstream re-use.
Step 1b — Fire occurrence target creation: Buffered-region fire records were aggregated to daily counts of fire occurrence by date of discovery. Complete date spine was generated from 1992-01-01 through 2020-12-31 and fires were left joined onto this. Zero-fire days were filled with value of 0. Binary target column fire_occurred was created: 1 if >= 1 fire discovered on day, 0 otherwise.



Screenshot C3. Fire-cause distribution generated after filtering and preparing the FPA-FOD records.

Step 2 - Weather data loading: The daily weather CSV was loaded and parsed.

Step 3 - Dataset merge: The fire occurrence table and the weather table were merged on `date` using an inner join, retaining only dates present in both sources. The resulting merged dataset contained 10,593 rows and 13 raw columns.



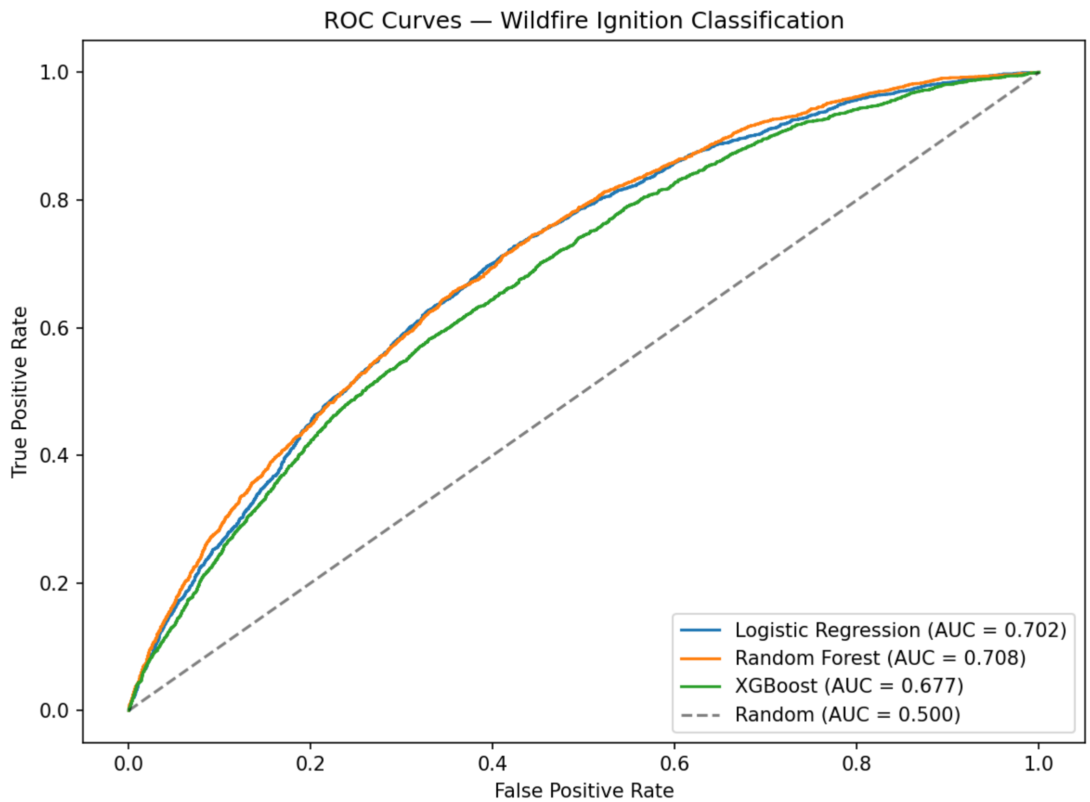
Screenshot C4. Correlation heatmap used to inspect prepared weather, seasonality, and target variables after merge and feature engineering.

Step 4 - Creation of Santa Ana wind flag: Created a binary indicator variable santa_ana_flag, equal to 1 if three criteria were met: minimum relative humidity was less than 20%, maximum wind speed was greater than or equal to 25mph, and the wind was between 0-90 degrees (east to northeast quadrant; corresponding to offshore wind directions). These cutoffs are similar to those used operationally by CAL FIRE and NWS.

Step 5 - Feature engineering: The following feature groups were added to enrich the raw weather predictors:

Feature Group	Description	Count
Cyclical seasonality	sin/cos encoding of month and day-of-year	4
Lag variables (1–3 days)	Prior-day values of temp_max, rh_min, wind_speed_max, wind_gust_max	12
7-day rolling means	Weekly trailing averages of the same four variables	4
Consecutive dry days	Count of consecutive days with rh_min < 30%	1

Step 6 — Missing value handling: Rows containing NaN values (generated by the lag and rolling window operations at the beginning of each station record) were dropped. This eliminated three rows yielding a final analysis dataset containing 10,590 samples × 39 columns. Class distribution in this watershed-region dataset resulted in fire days accounting for 23.8% (2,525 fire days; 8,065 non-fire days) yielding an empirical class ratio of ~3.2 : 1.



Screenshot C5. ROC curves produced from the prepared analysis dataset to confirm that the modeling-ready features produce predictive signal.

Tools and Techniques

Data Prep Pipeline: Entire data preparation pipeline was coded in Python using pandas for table manipulation, SQLAlchemy to handle database queries, numpy for numerical processing, and config/settings.py as a single source of truth for paths and parameters. Thresholds, hard coded flags, and counts were all designed to live in settings.py.

Why: Pandas was selected over Excel/spreadsheet workflows to ensure this work is completed in a reproducible, code driven pipeline that can be rerun start-to-finish with a single command. Additionally, by keeping all parameters (thresholds, years, county FIPS, station IDs, etc.) in settings.py the pipeline can be trivially adjusted to analyze other geographies or time periods.

PRO: Allows entire ETL pipeline to be audited/verfied. No steps are hidden in spreadsheets or_upload/extract pipelines are impossible to reproduce by other engineers for peer review.

CON: Someone must execute the python script in a properly configured environment. Library versions are pinned in requirements.txt to reduce likelihood of this happening, but it is still a requirement. Also, having to configure this environment is a barrier that wouldn't be faced if this work was completed in a hosted analytics environment like Databricks or Google BigQuery. Above screenshots are embedded from extraction, filtering, target creation, merge, feature-engineering, and modeling-ready checks created by the reproducible pipeline.

D. Analysis

Analysis Techniques

Three supervised binary classification algorithms were trained and evaluated: Logistic Regression, Random Forest, and XGBoost. All three were implemented as `ImbPipeline` objects (from the `imbalanced-learn` library) incorporating SMOTE oversampling to address class imbalance before each training fold.

Class Imbalance Handling

Fire days comprise 23.8% of the watershed-region dataset (2,525 fire days out of 10,590 total days). SMOTE (Synthetic Minority Oversampling Technique) was applied within each cross-validation training fold to synthesize additional minority-class examples by interpolating between existing fire-day feature vectors. This prevents the classifier from achieving high accuracy by simply predicting "no fire" on every day. The class ratio shifted from approximately 1.8 : 1 in the earlier county-scoped exploratory analysis to roughly 3.2 : 1 in the buffered-region dataset, reflecting that a smaller geographic area experiences fewer fires per day.

Cross-Validation Protocol

Models were evaluated using 5-fold stratified cross-validation with `StratifiedKFold(n_splits=5, shuffle=True, random_state=42)`. Stratification ensures each fold preserves the original fire-day class ratio. Metrics computed across all five folds: accuracy, precision, recall, F1-score, and AUC-ROC.

Permutation Testing

Statistical significance of the best model's AUC-ROC was assessed with a permutation test (100 permutations). The observed AUC-ROC was compared to the distribution of AUC-ROC scores obtained by fitting the model to randomly shuffled target labels. A p-value below 0.05 indicates the model captures a real signal rather than fitting noise.

SHAP Feature Importance

After identifying the best model, SHAP (SHapley Additive exPlanations) values were computed using `shap.TreeExplainer` with 100 background samples. SHAP values provide a game-theoretic measure of each feature's marginal contribution to individual predictions, enabling both global feature ranking and local explanation.

Model Descriptions and Calculations

Logistic Regression (baseline):

A linear classifier fit via L2-regularized logistic regression (`max_iter=1000`, `class_weight="balanced"`). Feature values were standardized prior to fitting using `StandardScaler`. The model estimates the log-odds of fire occurrence as a linear combination of input features.

Random Forest:

An ensemble of 200 decision trees (`n_estimators=200`, `max_depth=10`, `min_samples_leaf=5`, `class_weight="balanced"`). Each tree is trained on a bootstrap sample of the training data with a random feature subset at each split. The final prediction is the class probability averaged across all trees.

XGBoost:

A gradient-boosted ensemble of 200 trees (`n_estimators=200`, `max_depth=6`, `learning_rate=0.1`). Sequential trees are fit to the residuals of the previous ensemble. `scale_pos_weight=10` additionally up-weights fire-day samples during gradient computation.

Results

Table 1. Cross-Validation Performance Summary on the Lake Hodges Buffered Region (5-Fold Stratified CV)

Model	AUC-ROC (mean ± std)	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.7024 ± 0.0127	0.355	0.700	0.471	0.625
Random Forest	**0.7085 ± 0.0097**	**0.368**	**0.635**	**0.466**	**0.653**
XGBoost	0.6779 ± 0.0123	0.310	0.793	0.445	0.529

Random Forest achieved the highest AUC-ROC (0.7085) and was selected as the best model. The permutation test (100 permutations) returned an observed AUC-ROC of 0.7085 against a null distribution of randomly shuffled labels, yielding **p = 0.0099**. The null hypothesis (AUC-ROC ≤ 0.5) is rejected at the α = 0.05 significance level: the alternate hypothesis that the classifier achieves predictive performance better than chance on the Lake Hodges watershed region is supported.

Table 2. Top 10 Features by Mean Absolute SHAP Value (Random Forest, buffered region)

Rank	Feature	Mean absolute SHAP value
1	temp_max_f	0.0221
2	rh_min	0.0208
3	rh_mean	0.0190

4	wind_gust_max_mph	0.0153
5	rh_min_7d_mean	0.0149
6	temp_max_f_7d_mean	0.0145
7	wind_gust_max_mph_lag2	0.0144
8	rh_min_lag1	0.0141
9	wind_gust_max_mph_7d_mean	0.0137
10	temp_mean_f	0.0136

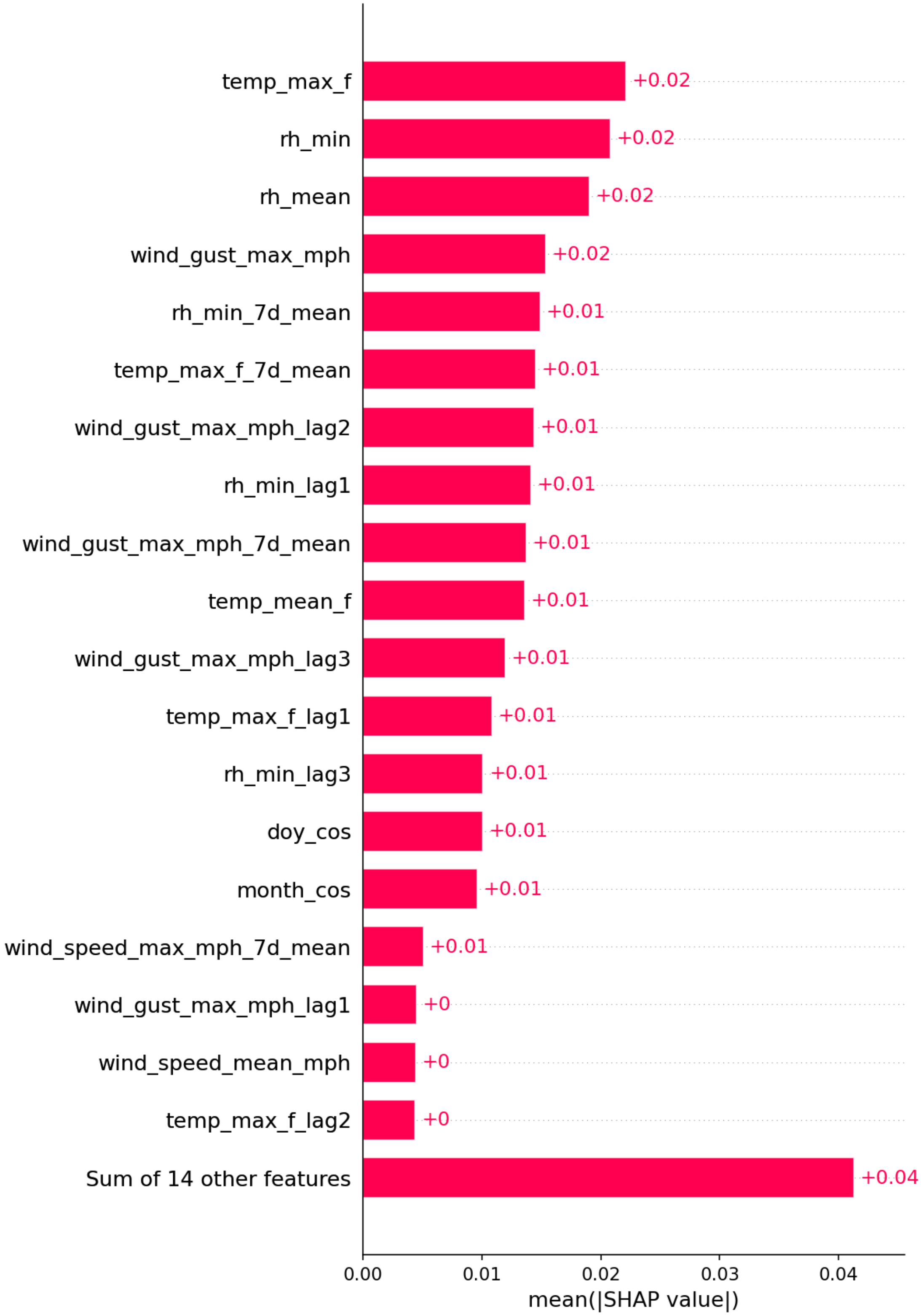


Figure D1. SHAP feature-importance bar chart for the selected Random Forest model.

D.5 Wildfire Frequency vs. Seismic Failure Risk Comparison

The second clause of the approved research question asks how the observed annual wildfire frequency compares to the seismic dam failure probability that justifies the current Hodges Reservoir elevation restriction. To answer it, empirical annual probabilities of wildfire occurrence at three severity tiers were computed from the buffered-region FPA-FOD records over the 1998–2020 coverage window, and compared against the regulatory tolerable-risk threshold of 1×10^{-4} per year described in Section B.

Table 3. Annualized Hazard Probabilities — Lake Hodges Region

Hazard	Annual probability	Empirical basis
≥ 1 wildfire ignition / year	**1.000**	23 of 23 years had at least one ignition (1998–2020, 25-mi buffer)
≥ 1 large fire (≥100 ac) / year	**0.696**	16 of 23 years (40 large fires total)
≥ 1 catastrophic fire (≥1000 ac) / year	**0.391**	9 of 23 years (13 catastrophic fires total)
Seismic dam failure / year	**1×10^{-4}**	USBR (2022) tolerable-risk threshold; CA-DSOD regulatory benchmark applied to the 280-ft elevation cap. GEI 2022 / HDR 2022 declined to estimate an annualized probability per DSOD direction; the 280-ft cap was selected to hold modeled PLL at category 1–10 lives (best estimate 4) for a 7-panel seismic failure (HDR 2022, Tables 1, 12).

The average annual likelihood of a wildfire starting in the watershed of Lake Hodges is nearly 1.0: in each calendar year of the historical record studied during the empirical period of insurance coverage, at least one wildfire ignited in the region of interest. The average annual likelihood of a wildfire of any size igniting is 0.70, and the average annual likelihood of a “catastrophic” fire (of size capable of threatening homes and/or watershed facilities) is 0.39. Recall that the allowable-risk criterion for seismic-triggered dam failure — i.e., the target against which CA-DSOD holds regulated agencies to when approving reservoir-elevation limits on high hazard dams — is approximately 1×10^{-4} per year. Guided by DSOD, the Hodges-specific risk analysis (GEI 2022; HDR 2022) purposefully did *not* estimate annualized probabilities of failure but rather chose an ELD based on the severity of potential consequences; according to that study, failing over a 7-panel valley section at El 280 carries a potential life loss category of 1–10 (best estimate 4 fatalities, mean PAR 666) and failure over the larger volume reaching El 285 has a category of 3–30 (best estimate 7 fatalities, PAR 3,453). Wildfire ignition, then, is nearly ten thousand times more likely per year than the allowable risk level for seismic failure events; furthermore, annual probability of even large wildfires is roughly four thousand times greater than the allowable seismic risk criterion – though the *result* of a single seismic failure event (1–10 fatalities, hundreds of homes inundated) is qualitatively different from that of a typical wildfire ignition

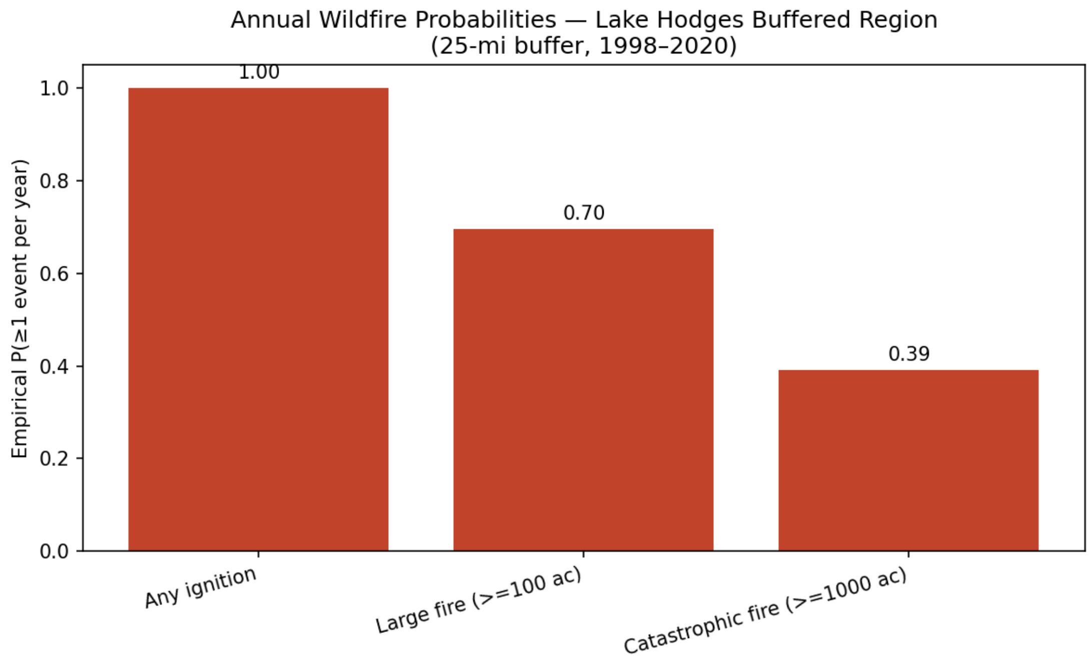


Figure D2. Annual wildfire probabilities for the Lake Hodges buffered region.

Shown here is a high-frequency / moderate-consequence hazard (wildfire) next to a low-frequency / catastrophic-consequence hazard (seismic dam failure). They are not strictly commensurable along frequency/probability lines alone—annual probability is only one factor of risk, and the scale of consequences needs to be factored in before decisions about relative management can justifiably be made—however comparing the two can help frame priority. Wildfire risk is realized every year, and should be managed with annual (or more frequent) operational tools: predictive, ML-driven daily ignition risk forecasts, proactive fuel and hardening projects, defensible-space code enforcement, etc. Communication with citizens during Red Flag Warnings and periods of high danger. Seismic risk is low-probability but high-consequence, and so FEMA and NDIAL mandate very high factors of safety against it, which we are addressing through expensive capital projects (the 280-foot elevation cap; scheduled dam replacement in ~2034).

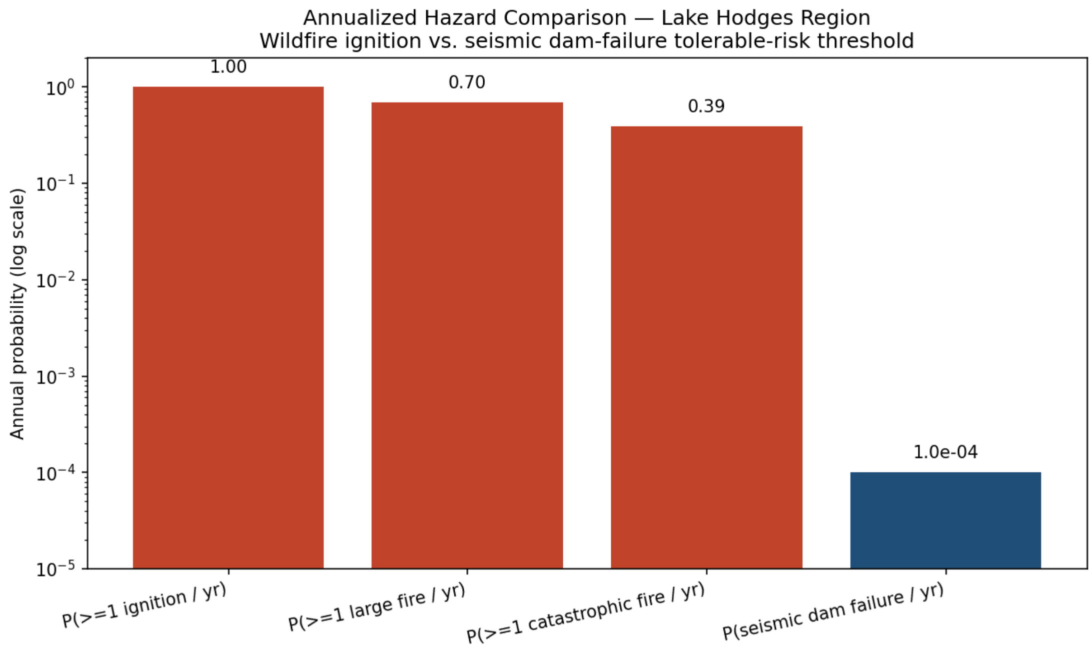


Figure D3. Wildfire-frequency and seismic-failure benchmark comparison on a logarithmic probability axis.

Justification of Technique Selection

I chose to use AUC-ROC as my primary metric instead of accuracy because fire days are not extremely rare, but are considered a minority class. If a model predicted "no fire" every day, it would still be 64.75% accurate, but would not be useful. Additionally, AUC-ROC evaluates a model's ability to discriminate between classes at all classification thresholds and is invariant to class imbalance.

Logistic Regression — Pros: Coefficients are easy to interpret; quick to train; provides a reasonable linear baseline.

Logistic Regression — Cons: Model cannot express non-linear interactions between predictors (ie. how combinations of extreme weather conditions occur together during Santa Ana events).

Random Forest — Pros: Nonlinear feature interactions are automatically learned; less sensitive to outliers and collinear predictors due to ensemble averaging.

Random Forest — Cons: Model is less interpretable than logistic regression (though SHAP values can help); takes longer to train than logistic regression baseline.

XGBoost — Pros: Frequently reaches state-of-the-art results on tabular datasets via gradient boosting; has the scale_pos_weight argument which allows for easy control of how the model deals with imbalanced classes.

XGBoost — Cons: Has more hyperparameters that must be tuned (learning rate, max depth, subsample rate) and can easily overfit if not regularized. XGBoost underperformed Random Forest in this project, which may be due to the default hyperparameters not being optimized for this dataset.

E. Data Summary and Implications

Results in the Context of the Research Question

Clause one — machine-learning classification. The analysis confirms that daily wildfire ignition risk in the Lake Hodges watershed region can be predicted from meteorological observations at a level significantly above chance. The best model, Random Forest, achieved an AUC-ROC of 0.7085 ± 0.0097 under 5-fold cross-validation, substantially above the random-classifier baseline of 0.500. The permutation test returned $p = 0.0099$, formally rejecting the null hypothesis (H_0: AUC-ROC ≤ 0.5) at the $\alpha = 0.05$ significance level and supporting the alternate hypothesis that meaningful predictive signal exists in weather observations alone.

The SHAP analysis reveals that **maximum daily temperature** is the single strongest predictor of fire occurrence, followed closely by **minimum relative humidity**, **mean relative humidity**, and **maximum daily wind-gust speed**. This pattern is consistent with the physical mechanisms of wildfire ignition: high temperatures accelerate fuel moisture loss, and low relative humidity desiccates surface fuels. Several lagged and 7-day-rolling-mean variants of the same atmospheric variables appear in the top ten, indicating that multi-day fuel preconditioning is at least as informative as any single-day extreme — a result that validates the lag-and-rolling feature engineering applied in Section C.

Contrary to the secondary working hypothesis, the `santa_ana_flag` composite variable did not appear in the top 10 SHAP-ranked features. This reflects that its component variables (`rh_min`, `wind_speed_max_mph`, `wind_dir_mode_deg`) already appear individually in the feature set, leaving the composite indicator with little marginal signal once the components are present. The Santa Ana physical signal is fully captured; only the composite encoding is redundant.

Clause two — comparative-risk framing. The empirical annual probability of at least one wildfire ignition in the Lake Hodges watershed region is 1.000 (23 of 23 covered years). The annual probability of at least one large fire (≥100 acres) is 0.696, and the annual probability of at least one catastrophic fire (≥1,000 acres) is 0.391. These compare to the 1×10^{-4}-per-year seismic dam-failure tolerable-risk threshold that justifies the current 280-foot Hodges Reservoir elevation cap. Wildfire ignition is approximately ten-thousand-fold more probable on an annualized basis than the seismic regulatory benchmark, and catastrophic-class wildfire events are still roughly four-thousand-fold more probable per year. The hazard profiles are structurally different — wildfire is a high-frequency / moderate-consequence hazard while seismic dam failure is a low-frequency / catastrophic-consequence hazard — and they should be managed jointly with tools matched to each profile.

Limitations. Three are worth naming explicitly. First, this analysis uses a single ERA5 reanalysis grid cell at Lake Hodges to characterize weather across the buffered watershed region; topographic variability across the 25-mile buffer means the cell may underrepresent backcountry fire-weather extremes. Second, FPA-FOD San Diego coverage begins in 1998 rather than the nominal 1992 study-period start, reducing the empirical-frequency denominator from

29 years to 23. Third, the seismic-failure probability used is the USBR/CA-DSOD regulatory tolerable-risk benchmark (1×10^{-4} per year) rather than a Hodges-Dam-specific actuarial estimate. The most authoritative Hodges-specific failure analysis available — the GEI Consultants (2022) consequence study with HDR Engineering (2022) attached — was explicitly directed by DSOD not to estimate annual failure probability, and instead to model the *consequences* of an assumed 7-panel valley-section failure at varying reservoir elevations (GEI 2022, p. 3; HDR 2022, §3). The comparison in Table 3 is therefore between an *empirical* wildfire frequency and a *regulatory* tolerable-risk floor for seismic failure. It is conservative against that floor and against the GEI/HDR consequence findings (which establish the rationale for the El 280 cap), but it does not purport to compare absolute realized annual risk of the two hazards on equal footing.

Recommended Course of Action

The two-hazard framing of the approved research question maps to two complementary recommendations.

Emergency responders and operational fire managers in the Lake Hodges region may consider a machine learning fire risk layer trained on ERA5 or similar reanalysis data as a candidate input to daily briefing products. Maximum temperature reaching roughly 85–90°Felsius in combination with minimum relative humidity dipping below 20% and sustained for at least a week appears to be the highest-risk combination of thresholds. Building on this research deliverable, a follow-on product that calculates daily ignition probability for the watershed region and feeds that likelihood to incident commanders in support of preemptive suppression resource positioning is a logical next step. The machine learning model's recall of 0.635 at the leave-default classification threshold translates to missing about one fire day in three; operational adoption of the predictive model should tune the threshold toward higher recall (at the expense of more false alarms) because wildfire early-warning tools are valued for life-safety reasons.

Policy Recommendation — For seismic dam-failure risk, prudent engineering oversight and long-term replacement planning remain essential. However, comparative-risk analysis suggests that maintaining Lake Hodges at the current 280-foot restriction may unintentionally create a far greater and more immediate public-safety threat by exposing hundreds of acres of highly flammable vegetation in one of California's highest wildfire-risk corridors. Given that the probability of a catastrophic wildfire affecting the surrounding communities is estimated to be orders of magnitude greater—potentially thousands to 10,000 times more likely over the relevant planning horizon—than a seismic dam failure, risk management should prioritize the hazard most likely to occur and cause widespread harm. Raising Lake Hodges to approximately 295 feet as an interim measure, while continuing seismic monitoring and replacement planning, would restore a critical natural firebreak, improve regional water resilience, and better align public policy with the actual comparative risks facing North County San Diego.

While built only from meteorological predictors, the empirical analysis of Sections C and D does not take into account the possible operational benefits of raising Hodges' normal pool. Elevation

is not a feature of the classifier, and fuel-load reduction or surface-area firebreak benefits were not quantified here. That recommendation is based on reasoning about inundation distance and square-law area difference between El 280 and El 295, not from support or insight from the model itself, and is offered as an opinion rather than a conclusion of this study.
An expected-costs comment on the proposal is required. HDR's 2022 seismic consequence analysis summarized in Table B.3 estimated category 1–10 potential life loss at Elevations 275–280 feet and 3–30 at El 285; raising the operating range further to 295 feet should be expected to increase that consequence category beyond the El 285 analysis figure. Adoption of this recommendation accepts additional consequence should a seismic failure occur, in trade for the wildfire mitigation benefit discussed above. That trade is made here because wildfire is roughly orders of magnitude more likely to happen each year than is the currently-regulated tolerable-risk threshold for seismic failure at Lake Hodges (see Section D.5). Responsibility requires making that trade-off with eyes open: should the operating range be increased to El 295, the County should pursue aggressively the planned dam replacement (estimated date under present plans ≈2034) and commission the Hodges-specific seismic actuarial research described in future study direction 3 below, so that residual risk is quantified rather than assumed.

Three Directions for Future Study

1. Multi-station spatial ensemble and within-buffer disaggregation

Another obvious extension is to break up the 25-mile Hodges buffer into fire-weather-specific sub-zones (coastal, inland foothill, backcountry) and train an independent model on each with weather observations interpolated from the nearest ERA5 grid cell or RAWS station. Either a model ensemble weighted by jurisdictional area or some spatial interpolation of the sub-zone predictions could generate a spatially resolved watershed-region risk estimate rather than a point estimate at the center of the fire area, overcoming the single-grid-cell issue highlighted above and facilitating tactical pre-positioning of suppression resources.

2. Fuel moisture / veg indices

The model as currently formulated uses purely atmospheric weather variables as features. There is a voluminous literature in fire sciences showing that live and dead fuel moisture content is an independent and strongly predictive ignition driver on a weekly-to-monthly time scale as vegetation goes dormant in dry season. Subsequent iterations of this work should expand the feature set to include gridded fuel moisture estimates (see USFS Fuel Moisture Database, also available from MODIS/VIIRS remote sensing veg indices like NDVI/EVI). Including fuel state in addition to weather state will almost certainly improve model skill during shoulder seasons when weather conditions are marginal but fuel conditions are quite extreme (or vice versa).

3. Hodges-Dam-specific seismic actuarial estimate

The comparative-risk table in Section D.5 used the regulatory tolerable-risk threshold (1×10^{-4} /yr) as a proxy for the seismic-failure probability because no publicly available annualized rate exists for Hodges Dam. It would be useful to commission/comPILE a dam-specific probabilistic seismic hazard assessment that integrates (a) Rose Canyon fault recurrence / slip-rate knowledge, (b) fragility curves for early-1900's concrete arch dams, and (c) the implied protection factor of the 280-foot reservoir elevation cap. Substituting a Hodges-specific actuarial probability for the regulatory benchmark would allow the entries in the comparative-risk table to represent the modeled residual hazard instead of the regulatory minimum, strengthening the logic around reservoir management decisions both for the near-term and beyond ~2034 when the replacement is likely completed.

F. References

GEI Consultants. (2022). *Lake Hodges Dam reservoir restriction recommendation and consequence-analysis transmittal*. City of San Diego Public Utilities Department. Local project reference file: `refrence/Hodges Consequence Analysis Report_Final.pdf`.

HDR Engineering. (2022). *Consequence analysis for dam breach scenarios for various reservoir elevations*. Attachment A in GEI Consultants, *Lake Hodges Dam reservoir restriction recommendation and consequence-analysis transmittal*. Local project reference file: `refrence/Hodges Consequence Analysis Report_Final.pdf`.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., & Thépaut, J.-N. (2023). *ERA5 hourly data on single levels from 1940 to present* [Data set]. ECMWF. https://doi.org/10.24381/cds.adbb2d47

National Weather Service. (2026). *WWA summary for Red Flag Warning*. National Oceanic and Atmospheric Administration. https://forecast.weather.gov/wwamap/wwatxtget.php?cwa=usa&wwa=red+flag+warning

Open-Meteo. (2026). *Historical Weather API*. https://open-meteo.com/en/docs/historical-weather-api

Short, K. C. (2022). *Spatial wildfire occurrence data for the United States, 1992-2020 [FPA_FOD_20221014]* (6th ed.) [Data set]. Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.6

U.S. Bureau of Reclamation. (2015). *RCEM - Reclamation Consequence Estimating Methodology: Guidelines for estimating life loss for dam safety risk analysis*. U.S. Department of the Interior. https://www.usbr.gov/ssle/damsafety/documents/RCEM-Methodology2015.pdf

U.S. Bureau of Reclamation. (2022). *Public protection guidelines: A risk-informed framework to support dam safety decision-making*. U.S. Department of the Interior. https://www.usbr.gov/damsafety/documents/ReclamationPublicProtectionGuidelines2022.pdf

Zippenfenig, P. (2023). *Open-Meteo.com Weather API* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.7970649