2021 Wildfire Distribution Risk Model Overview

Risk and Data Analytics
Electric Operations, PG&E
Model Release Date: 10/15/2020
Documents Last Updated: 2/1/2021

TABLE OF CONTENTS

1 Abbreviations, Definitions, and Conventions

2021 - When appended to the name of something, typically a model or other analysis, 2021 refers to the first year for which the model or analysis results will inform planned or executed work. The models documented herein are the 2021 models – they are used to inform 2021 work plans. Note that the 2021 modeling and analysis work was performed during the calendar years 2019 and 2020. When the year is omitted, 2021 may be assumed.

RaDA - The team that created the model is the Risk and Data Analytics team (RaDA). The team, and its models and computer code, have formerly been referred to as Distribution Asset Risk Management (DxARM) or Distribution Risk (DxRisk).

MaxEnt – A Maximum Entropy model applied to spatial range estimation. The name given to a family of models that seek to maximize the information entropy[1] (i.e. instead of the likelihood or some other optimization criteria) of the probability distribution associated with a given set of conditions – in this case, ignition probability, given environmental and asset characteristics. It can also be interpreted as finding the least unique distribution that fits the underlying data.

Maxent - Name of the software used to perform MaxEnt modeling.

WMP – Wildfire mitigation plan. The official expression of PG&E plans as designated by SB-901 to mitigation wildfire risk that includes (non-spatial) MAVF wildfire risk calculations.

Raster data – “Pixelated” spatial data - for example wind speed, elevation, or conductor material – conforming to a well-defined map projection that assigns a geographic coverage area (i.e. a polygon on the surface of the globe) to each data pixel. Gridded weather data in the form of polygons with associated traits is a good example of raster data

Vector data – Tabular data – for example asset IDs and attributes - associated with specific spatial geometries composed of points and lines and conforming to a well-defined map projection that assigns each point to a specific location on the surface of the earth. EDGIS contains vector data on grid assets.

[1] Information entropy is the average level of uncertainty inherent in an outcome derived from a set of variables or covariates

EPSG:32610 – The official projection of PG&E territory topography/geospatial area used for this project - WGS 84 / UTM zone 10N – which has relatively little distortion over California (i.e. the distortion caused by projecting the 3D surface of the earth to 2D map data) and whose units are meters from a fixed spatial location.

Asset attribute data – Data on the characteristics of grid assets. Examples include conductor size, materials, proximity to the coast, or splice counts.

Grid pixels – The subset of raster locations that have grid assets within their boundaries. This modeling was performed using 100m x 100m grid pixels as the corpus of all locations for which input data is needed and predictions will be made.

Fire-season – The period from June 1 to November 30 capturing the typical period of hot and dry weather in PG&E's service territory.

HFTDs - High Fire Threat Districts – Areas within California with elevated (Tier 2) and extreme (Tier 3) fire threat, as developed under CPUC rulemaking R.15-05-006 and adopted by the CPUCs Safety and Enforcement Division. All 2021 modeling was restricted to ignitions and covariates within the HFTDs.

Covariates – The data used as explanatory variables in the formulation of a MaxEnt model. They must be spatially resolved and available for every location for which a prediction will be made such as the number of trees or average precipitation.

Ignition probability – Unless otherwise specified, the odds of at least one ignition within each 100m x 100m grid pixel per-fire season, estimated using MaxEnt as described in this document. Also known as the likelihood of risk event, or LoRE.

Ignition consequences – The spatial data set, based on Technosylva fire simulations under dangerous fire conditions and calibrated to be compatible with PG&E's reported MAVF CoRE values, that multiplies the ignition probability (LoRE) for each grid pixel to produce pixel-level wildfire risk.

Ignition risk – Ignition probability x ignition consequence – balances the severity of an outcome against its likelihood to assess the overall danger associated with potential ignitions at a given locations.

gridMET – A dataset of ~4kM resolution daily meteorological data (derived from satellite imagery), covering the contiguous USA from 1979 to the present.[1]

RTMA – Real-Time Mesoscale Analysis. A NOAA hourly weather raster data product at 2.5kW resolution with hourly timesteps. RTMA has only been available since mid-2015.[2]

MAVF – Stands for "Multi-Attribute Value Function" and refers to the utility-specific risk calculation methodology developed in accordance with principles established by the SMAP Settlement Agreement D.18.12.014

MAVF Consequence Dimensions – The impacts of a risk event such as wildfire or other utility-related events that include damage to equipment, loss of service, and threats to public safety. MAVF captures risk consequences via Reliability, Financial, and Safety dimensions in natural units and converts these into a unitless risk score known as the Multi-Attribute Risk Score (MARS) as discussed in PG&E's 2020 RAMP Report. We are primarily interested in the multi-attribute CoRE values for ignitions in this document.

CoRE – Consequence of risk event used in the MAVF framework. CoRE is multi-attribute ignition consequence for our purposes and will often just be called "consequence" in our documentation.

LoRE – Likelihood of risk event used in the MAVF framework. LoRE is the ignition probability for our purposes.

EORM – Enterprise and Operational Risk Management – the department within PG&E responsible for identifying, quantifying and tracking risk at the enterprise level. EORM implements the company wide MAVF risk calculations.

RAMP – Risk Assessment Mitigation Phase of the General Rate Case proceeding. PG&E filed its 2020 RAMP Report A.20-06-012 in June of 2020.

Red Flag Warnings – Red flag warnings (RFW) are issued by the National Weather Service when extreme fire weather (i.e. hot, dry, and windy) conditions are predicted. Red flag warnings are issued for specific geographies and time ranges. In the context of spatial consequence calculations, we are interested in the spatially differentiated expected count of red flag warnings for all area of the grid.

Technosylva – Fire simulation software whose propagation and consequence outcomes are based on available fuels, topography, and weather data; as well as building structure and population data layers. Technosylva simulation outputs are used as the source of spatially resolved fire severity data that is the primary input into the spatial consequence calculations.

FireSim – Technosylva's fire simulation model

WRRM – The Wildfire Risk Reduction Model developed by Technosylva

FBI – Technosylva's Fire Behavior Index. A scale of 1-5 that captures fire severity as a function of flame length (intensity of burn) and rate of spread. FBI of 3 or greater is expected to require aggressive suppression.

CPZ – Circuit protection zone – the set of all assets protected by a specific protective device. Also referred to as Circuit Segment (CS).

ACSR- aluminum conductor steel-reinforced

Al - aluminum

AWG - American Wire Gauge

CPUC - California Public Utilities Commission

Cu - copper

EDGIS - Electric Distribution Geographic Information System

K - Kelvin

Km - kilometers

kPa - kilopascals

m - meters

mm - millimeters

NED - National Elevation Dataset

PSPS - Public Safety Power Shutoff

s - seconds

SME - subject matter expert(s)

TPI - Topographic Position Index

USGS - United States Geological Survey

ROC-AUC - receiver operator curve – the area under the curve, also referred to as "AUC", is a metric used to evaluate model performance.

2 Executive Summary / Overview

Catastrophic wildfires have become an existential threat to the State of California and pose a significant threat to the safety and economic future of the State's residents as a result of increasing population growth into the wildland urban interface (WUI) and changing climatological conditions. The frequency and severity of these catastrophic wildfire events has increased dramatically over the last 10 years. PG&E recognizes its electrical equipment has been associated with the ignition point for a number of these fires and is working to understand these catastrophic events to maximize planned risk reduction activities. However, PG&E recognizes that the historical methods for understanding and managing wildfire risk need to evolve given the heightened frequency and severity of wildfires. In order to meet this heightened wildfire risk, PG&E has developed a set of models to identify areas of highest potential for ignitions and consequence. PG&E is committed to improving its modeling capabilities as the available information and understanding of wildfires improves.

This document provides a detailed overview of PG&E's current wildfire risk modeling approach: the 2021 Wildfire Distribution Risk Model. This model supersedes the prior iteration of wildfire risk models developed in 2018 (the 2019-2020 Wildfire Risk Model). Key objectives for the 2021 Wildfire Distribution Risk Model are:

1. Provide situational awareness of risk,
2. Enable risk-informed decision making and
3. Enable PG&E to develop line-of-sight on risk reductions from wildfire risk mitigation initiatives.

Recognizing that risk-informed decision making is desired for both asset investment workplans developed on an annual basis and operational decisions, such as PSPS, PG&E has and is developing models specific to the temporal needs of each situation. There are primarily two forms of models that can be used to address wildfire risk. First, planning models support annual workplans and are based on either worst-case conditions such as weather and fuels or cumulative probabilities of failure or ignition. The 2021 Wildfire Distribution Risk Model described herein is a planning model for the Electric Distribution system. Second, operational models, such as those used for PSPS events utilize real-time weather, fuels data, and asset conditions as reflected by maintenance tags or recently completed asset hardening. The Large Fire Probability Model (Distribution) or LFP_D Model, is an example of an operational model. Given the respective application of planning and operational models, planning models are updated on an annual cadence while operational models are updated as frequently as weekly during fire season.

Following the Electric Operations Risk Framework, outlined in section 2 that provides a systematic approach to risk assessment and mitigation, the 2021 Wildfire Distribution Risk Model seeks to quantify the risk of wildfire represented by the probability of electric grid infrastructure caused ignitions combined with the consequences if that ignition propagates to a wildfire. In its entirety, the 2021 Wildfire Risk Model is a set of models that represents failure modes, or risk drivers, underlying ignitions and the consequences of wildfire. These models comprise the components of the wildfire risk formula:

Wildfire Risk = Ignition Probability x Wildfire Consequence

For the first part of this formulation, the "Ignition Probability" portion of the 2021 Wildfire Distribution Risk Model is modeled according to the risk drivers identified in PG&E's 2020 RAMP Report for wildfire risk. From these risk drivers, the 2021 Wildfire Distribution Risk Model developed probabilities for vegetation and equipment failure caused ignitions as they represent 38% and 26% of the grid related ignitions respectively. Within equipment failures, the 2021 Wildfire Distribution Risk Model has developed probabilities for conductor failures. Future modeling efforts will add failure models for other drivers such as 3rd party contact and for other electric grid equipment such as poles and transformers. The modeling framework established with this model will accommodate the future addition of such models.

The predictive power of these risk driver-based models has been greatly improved over the 2018 model in several areas. First, the 2018 model was trained on outages as a proxy for ignitions. Using advanced statistical techniques such as separating the data in to training and test sets and supported by improved data and a more efficient algorithm, a true ignition probability model was developed. Improved data sets have also fueled an improvement in model granularity from the circuit and circuit segment level to 100-meter pixels along the electric distribution lines. Finally, a predictive algorithm called MaxEnt, (a modeling approach often used in biological and environmental application for the identification of species ranges by habitat) was utilized due to its compatibility with available data and modeling objectives. In particular, MaxEnt has the following characteristics: ability to work with spatially explicit inputs and outputs, support for presence/absence probability prediction, ability to work with uncertain location data (i.e. compensating for location uncertainties in historical ignition data), ability to work with relatively few "positives" (i.e. Ignition locations) in a sea of negatives, tendency to converge well with modest amounts of training data, and machine learning heritage which ensures prediction performance is prioritized over "in-sample" training data fit.

The "Wildfire Consequence" portion of the 2021 Wildfire Distribution Risk Model focuses on fire impacts in natural units such as acres burned, number of structures impacted, and variables describing the nature of the fire such as flame length and rate of spread. The key improvement for the 2021 Wildfire Distribution Risk Model is tied to the advanced modeling capabilities of the Technosylva fire simulation tools. In the 2019-2020 Wildfire Risk Model, REAX Engineering provided simulations that relied heavily on the concentration of fuels (based on LANDFIRE 2014 data) to determine the potential for an ignition to propagate to a wildfire. While informative, the Technosylva simulation tool improves on this capability by firstly using an updated ground fuels dataset (LANDFIRE 2016 with fire disturbance updates) and also by modeling what fire science refers to as ladder fuels whereby an ignition will propagate from low fuels, such as grass and brush, to increasingly denser fuels leading to treetops (crowns), as well as updated buildings and population data layers. The result is a more accurate representation of the potential consequences of wildfire in the wildland urban interface and the broader Tier 2 and Tier 3 HFTD areas modeled. Future versions of the consequence model will consider additional areas in the PG&E distribution system.

Bringing the improvements to the both the Ignition Probability and Wildfire Consequence portions of the model together, the 2021 Wildfire Distribution Risk Model now provides an updated and improved measure of wildfire risk. The 2019-2020 Wildfire Risk Model provided a relativistic measure that was instructive for prioritizing circuits and circuit segments, but it did not allow for measuring the degree of risk between those segments. The 2021 Wildfire Distribution Risk Model provides this capability as the risk scores are absolute scaled units. As a result, risk values can now identify how much riskier a location is compared to another, risk can be more accurately compared across wildfire and PG&E's other risk events, and the actual value of risk reduction is now more easily computed.

Even as the predictive power of the 2021 Wildfire Distribution Risk Model has been greatly improved as compared to the 2019-2020 Wildfire Risk Model, PG&E is continuing to develop and refine its risk modeling. The 2021 Wildfire Distribution Risk Model has several limitations; it does not include transmission facilities, has not yet been used to generate risk reduction scenarios matching mitigation plans, and for equipment-involved probability of ignition the model only includes conductors at this time. In 2021, PG&E intends to develop the 2022 Wildfire Distribution Risk Model which will include certain upgrades to the 2021 model and will include data on additional electrical equipment (*e.g.*, poles). In 2021, PG&E is also working to develop a 2022 Wildfire Transmission Risk Model for its transmission facilities that will be similar to the 2021 Wildfire Distribution Risk Model. Finally, PG&E is also working on a pilot Probabilistic Risk Assessment or "PRA" model. The PRA is still conceptual, but, if successfully developed, will integrate all models into a single electric system view of wildfire risk. PG&E is working to develop a reference model of the PRA in 2021 and potentially, depending on the effectiveness of the reference model, to use the PRA for planning in 2022. Improved models will provide more actionable insights that will enable more effective and efficient workplans and allow PG&E to mitigate the risk of wildfire for the State of California and our customers.

3 Document Usage

This document provides a comprehensive overview of all modeling activity that comprises the 2021 Wildfire Distribution Risk Model for a general audience, with more detailed and technical appendices for all major topics. The complete set of topics covered are listed below in Table 1. The next few sections of this document provide a high-level overview of the motivations behind and context for the 2021 risk modeling effort. It is intended to place the modeling work within its strategic and regulatory context and provide a high-level guide to all aspects of modeling performed with discussion of performance and applications. It can be used to understand the vision behind the risk modeling effort and the plans for future developments as well as to gain an executive summary level understanding of the modeling and results.

Following the main body are appendices that cover the Vegetation Probability of Ignition Model and Equipment Probability of Ignition Model in greater detail. These tie to the vegetation and equipment risk drivers identified in the Wildfire risk discussed in PG&E's RAMP Report. These appendices can also be used to understand the data utilized in each model, the relative influence of the different data sets, the precision of the model in predicting ignitions, and areas for future improvement.

The modeling appendices are, in turn, supported by two additional appendices on key methods: Appendix 3: Ignition Probabilities Methods 2021 provides details on the application of the MaxEnt algorithm to provide spatial distribution grid ignition probabilities; and Appendix 4: Ignition Consequence Methods 2021 provides details on the application of the Technosylva simulation data to develop a consequence data set in the MAVF framework, referred to as MAVF CoRE, that is calibrated to the MAVF system level and tranche level scores in the RAMP Report.

In addition to these five written topics, several presentations were developed and used to conduct technical reviews internal and external to the company. These are also available as separate files that can aid in understanding the 2021 modeling, but they were not created to directly support this document.

TABLE 1 - INDEX OF 2021 WILDFIRE RISK MODEL TOPICS (APPENDIX TOPICS ARE CLICKABLE CROSS-REFERENCES)

Section and Topic	Description
2021 Wildfire Distribution Risk Model Overview	The main body of this document - summarizing the context for this work and providing a high-level overview of the approach and results.
Appendix 1: Vegetation-caused Ignition Risk Model 2021	Description of modeling vegetation-caused ignition probabilities and related risk results used to inform 2021 EVM planning and prioritization.
Appendix 2: Conductor-Involved Ignition Risk Model 2021	Description of modeling conductor-involved ignition probabilities and related risk results used to inform System Hardening planning and prioritization.
Appendix 3: Ignition Probabilities Methods 2021	Detailed coverage of the motivation behind and methods used to employ maximum entropy models to make spatial estimates of ignition probabilities (independent of specific applications)
Appendix 4: Ignition Consequence Methods 2021	Detailed coverage of the methods and modeling behind the development of the MAVF-compatible spatial consequence data used in risk calculations based on the Technosylva model.
EVM Risk Model 2021 - Lunch n' Learn	Presentation for a general (internal PG&E) audience on the EVM Risk model.
EVM Risk Model 2021 - Utility Analytics conference	Presentation for a Utility Analytics audience on the EVM risk model
Conductor Risk Model 2021 - Lunch n' Learn	Presentation for a general (internal PG&E) audience on the Conductor Risk model.

4 Objectives, Framework

4.1 Project Objectives

The 2021 Wildfire Distribution Risk model project objectives were to develop a model that:

1. Provides situational awareness of risk
2. Enables risk-informed decision making
3. Enables PG&E to develop line-of-sight on risk reductions from wildfire risk mitigation initiatives

In the pursuit of these objectives, PG&E wildfire risk modeling maturity aimed to progress from relative risk models at the circuit level with system level risk reduction and RSE capabilities to automated quantitative wildfire risk models that include risk reduction and Risk Spend Efficiency (RSE) evaluations ultimately at the asset/structure level. The 2021 Wildfire Risk Model is the second iteration of risk models and is a significant step in improving PG&E's wildfire risk modeling capabilities as measured by the CPUC Utility Wildfire Mitigation Maturity Survey (Maturity Survey).

To accomplish the improvements from the 2019-2020 Wildfire Risk Model to the 2021 Wildfire Distribution Risk Model, a systematic Risk Modeling Framework was used to develop the capabilities identified in the Maturity Survey.

4.2 Framework

The following systematic Risk Modeling Framework has been adopted to develop the capabilities identified in the Maturity Survey. This general framework is shown in Figure 1 - Risk Modeling Framework.

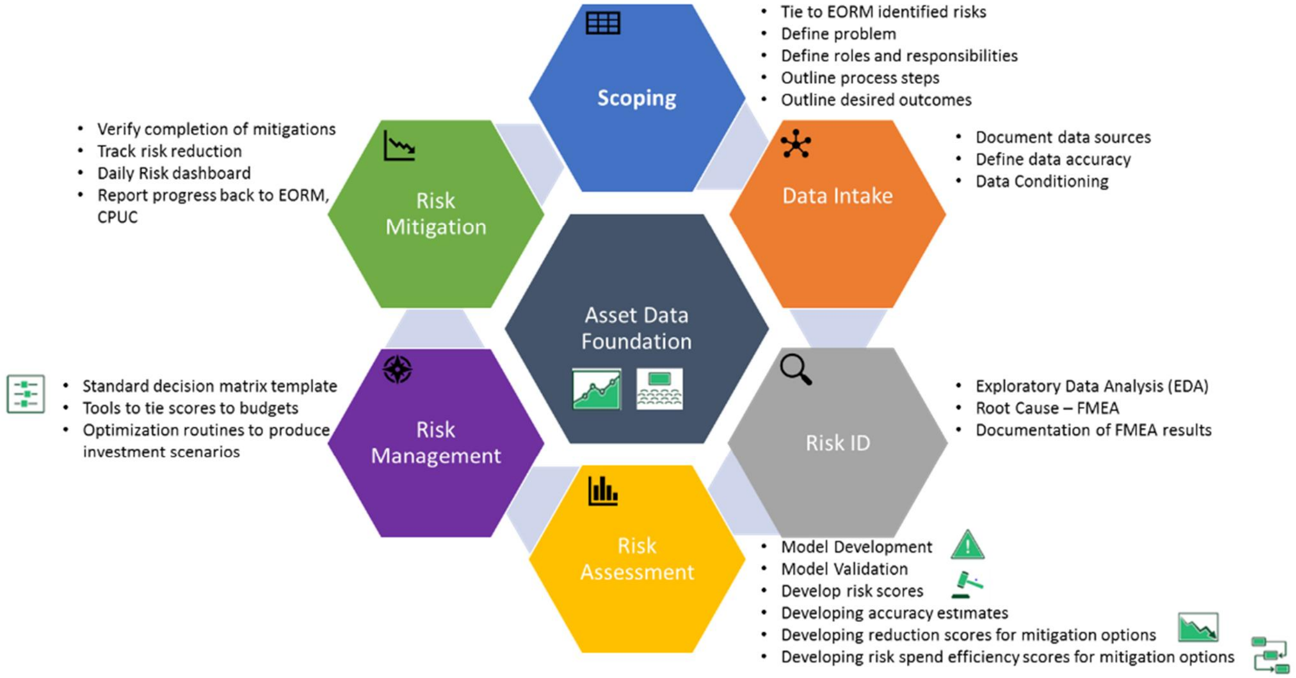

Scoping
• Tie to EORM identified risks
• Define problem
• Define roles and responsibilities
• Outline process steps
• Outline desired outcomes
Data Intake
• Document data sources
• Define data accuracy
• Data Conditioning
Risk ID
• Exploratory Data Analysis (EDA)
• Root Cause – FMEA
• Documentation of FMEA results
Risk Assessment
• Model Development
• Model Validation
• Develop risk scores
• Developing accuracy estimates
• Developing reduction scores for mitigation options
• Developing risk spend efficiency scores for mitigation options
Risk Management
• Standard decision matrix template
• Tools to tie scores to budgets
• Optimization routines to produce investment scenarios
Risk Mitigation
• Verify completion of mitigations
• Track risk reduction
• Daily Risk dashboard
• Report progress back to EORM, CPUC
Asset Data Foundation

FIGURE 1 - RISK MODELING FRAMEWORK

The specific framework steps for the 2021 Wildfire Risk model development are outlined below, beginning with the model Scoping and working through the Data Intake, Risk Identification, Risk Assessment, Risk Management steps to conclude with Risk Mitigation and reporting.

Scoping – defining the problem and desired outcomes. Beginning with the Scoping step, the 2021 Wildfire Distribution Risk Model is tied to the wildfire risk bowtie and risk scores outlined by PG&E's EORM department in our 2020 RAMP Report. Examples include the development of risk scores calibrated to the system MAVF scores and modeling failure modes for the identified wildfire risk drivers. During the scoping step, key desired capabilities were identified tying to the Maturity Survey, such as the improved level of granularity, the ability to aggregate risk scores to different levels such as circuit segments, and the comparability of risk scores to facilitate the development of risk reduction and RSE values.

Data Intake - key data sets are identified and prepared for modeling. For the 2021 Wildfire Distribution Risk Model, vegetation data, ignition data, and asset data were critical data sets that were identified and prepared for modeling usage. As LiDAR data was not fully available at this stage, satellite-derived vegetation characteristics data was provided by one of our project partners, Salo Sciences.

Risk ID - Failure Modes Effect Analysis (FEMA) and Exploratory Data Analysis (EDA) are employed to understand and identify the root cause and characteristics of the problem. From the identified risk drivers in the Wildfire risk bowtie, vegetation-contact and conductor-involved ignitions were the most frequent ignition drivers. Using a previously developed FMEA, EDA was conducted on the identified data sets in the Data Intake step. EDA begins the process of gaining insight from the data before formal modeling begins. This includes understanding the accuracy of the data, patterns including outliers and anomalies, as well as identification of potentially predictive relationships within and between data sets.

Risk Assessment - development of the models and model features. In this step, the model algorithm is selected and trained on the ignition data to provide spatial probabilities of ignition. The Wildfire Consequence Model data was also developed from the Technosylva simulation model. To quantify the predictive power of the model, precision assessments were developed. These metrics informed iterative adjustments that were subsequently made to improve predictive ability. The resulting MAVF risk scores were then calibrated, and validation exercises were held with the Vegetation Management and Distribution Asset Strategy teams that would ultimately use the models to inform their 2021 workplans. At this point the 2021 Wildfire Distribution Risk Model was reviewed and approved by the Wildfire Risk Governance Steering Committee (WRGSC) which is led by PG&E's Chief Risk Officer and made-up of a cross-functional officer team.

Risk Management – insights from models are used to develop work plans. The modeling insights are combined with project factors and variables not incorporated in the models. For example, tree species data was not widely available enough to be fully incorporated in to the EVM Risk model. As a result, the Vegetation Management team applied species data as an overlay to the Vegetation Risk Model to produce the 2021 EVM workplan. With the Distribution Asset Strategy team, model data is combined with information on terrain, customers locations, and customer counts to identify the preferred mitigation alternative. Similar to the risk models, the resulting workplans are also reviewed and approved, as part of this step, by the WRGSC.

Risk Mitigation – monitors and reports the drawdown of risk as work is performed. This is accomplished with model-assigned asset-level risk values as well as validating the model against actual system performance metrics. For example, ignition probability models are validated against actual annual ignitions to capture insights into future improvements. As modeling capabilities improve monitoring the risk drawdown can become a key operational metric.

5 Modeling Methods: Estimating Risk

The 2021 Wildfire Distribution Risk Model formulates risk in probabilistic terms in a manner that is similar to and compatible with the MAVF risk framework established by the SMAP Settlement Agreement. The fundamental concept is that the risk associated with an event, such as a fire ignition, can be expressed as the product of the probability of the event happening and the consequences if it does happen. The MAVF framework calls these the likelihood of risk event (LoRE) and the consequence of risk event (CoRE), respectively. In the 2021 Wildfire Distribution Risk Model, the notation P(ignition) for LoRE ignition probability and C(ignition) for the CoRE consequences of an ignition, is used, as shown below:

Risk = P(ignition) x C(ignition)

The heart of the 2021 risk model is the effort to estimate ignition probabilities and ignition consequences for distribution grid locations in the Tier 2 and Tier 3 High Fire Threat Districts (HFTD). In our documentation we have separated out the more technical discussion of the methods used for estimating P(ignition) and C(ignition) from their application in vegetation-caused and conductors-involved risk models. This overview section discusses all of four of these topics in the sections that follow, but each topic also has a dedicated Appendix that provides significantly more detail on each topic.

5.1 Methods: Ignition Probability

To answer the question of where ignition events are likely to occur, spatially resolved fire season ignition probabilities have been estimated using maximum entropy models (MaxEnt). The MaxEnt model provides relative scores or, if properly calibrated, probabilities for fire-season ignitions per "pixel" of input data. MaxEnt models take the set of locations of ignitions under study and rasterized (i.e., pixelated) data on environmental conditions and asset attributes as explanatory covariates for all locations with grid infrastructure as inputs and output rasterized maps of ignition probabilities.

MaxEnt models have been successfully applied in ecology to the problem of estimating a species' range (*i.e.*, the physical extent of its suitable habitat), given a set of locations where members of that species have been observed and the corresponding environmental conditions at those locations and all candidate locations for the range. In that context, the model assigns a score to every location that captures how similar the conditions at that location are to the locations *where* the species was observed. There is a correspondence between MaxEnt applied to species observations and ranges and ignition locations and at-risk locations –looking for the "range" of grid-caused wildfires - the environmental conditions and asset attributes associated with elevated wildfire probabilities. PG&E has applied MaxEnt methods to event occurrences and their proximate asset and environmental conditions contrasted with the background conditions everywhere else along the distribution grid to identify the locations most likely to experience similar events in the future.

Special topic: Conceptual explanation of how MaxEnt models work

For the 2021 Wildfire Distribution Risk Model, the objective is to identify which environmental conditions and asset attributes (collectively called the model covariates) are more common among ignition locations than they are among all distribution grid locations. For example, tall trees are more common among vegetation-caused ignition locations than they are among typical Distribution grid locations.

Metrics of dryness, HFTD tier assignments, conductor materials and size, and others, can all be checked for such patterns. The ratio of covariate value prevalence at ignition locations to their prevalence across all grid locations is called the *relative occurrence rate*. MaxEnt provides a way of estimating the relative occurrence rate given a fairly modest number of ignition locations. The way it does this is to *fit a statistical distribution of covariate values for ignition locations that is consistent with the values at known ignition locations, but otherwise as similar as possible to the distribution of values found everywhere else along the Distribution grid.*

The similarity criteria described above is enforced using a metric called the *relative information entropy* between the ignition locations and the Distribution grid locations, where the larger that metric is, the more similar the two distributions are. For this reason, the overall approach is referred to as a maximum entropy or MaxEnt estimation of the relative occurrence rate.

When multiplied by the fraction of all grid locations that experience fire-season ignitions annually, the relative occurrence rate is normalized into a distribution that provides the annual probability an ignition will occur for all combinations of values of the covariates. This distribution can be used to look up (aka predict) annual ignition probabilities based on the covariate values found at each Distribution grid location.

5.1.1 Why MaxEnt models?

Different modeling approaches have different strengths and weaknesses and there are always better and worse performing model specifications under any given approach. The selection of a MaxEnt approach was informed by the characteristics of available data and the spatial nature of the planning questions the model needed to address, with multiple specification evaluated to identify models with strong predictive power as well as the explanation of inputs.

Model performance (and selection) should be determined quantitatively based on out-of-sample prediction accuracy for most planning purposes, but performance can only be assed for specific/well specified/narrow questions. In other words, there are no “silver bullet” models that perform well for all questions asked of them. The most important predictions we make using the MaxEnt models is the location-specific fire-season probability of ignitions based on patterns in ignitions from 2015-2018, measured against out-of-sample 2019 data. We have not had the bandwidth to mature all possible alternative model formulations for quantitative comparison to MaxEnt. However, it does perform well compared to earlier models and there are very good reasons we opted to employ it:

(1) MaxEnt, in the lineage of usage we have adopted, is spatially explicit. It is used to answer, “where can I expect X to occur”, which is the most common structure of the questions we’ve been asked to address. It takes spatial data inputs and outputs rasterized (aka pixelated) spatial results that can map directly over grid locations.

(2) Our event data is uncertain around where exactly each ignition occurred and which specific device failed. Ignition locations are recorded in the field and captured where the data acquisition device was, not where the ignition began. Further, there can be GPS signal acquisition challenges in the field that result in location errors. The result is that the outage data associated with ignitions most consistently records their locations, but they are protective device locations, not exact ignition locations. Approaches that require direct assignments of ignition indicators to specific pieces of equipment or precise coordinates of the point of ignition are not viable with currently available data. MaxEnt works with spatially quantized data and is focused on comparing the distribution of conditions at ignition locations to all locations, so some imprecision can be tolerated as relatively small permutations in those distributions.

(3) Unlike traditional “classification” methods, MaxEnt works with “presence only data”, which means that you don’t need accurate labels for all ignitions or all non-ignitions. This is relevant to the locational imprecision noted above – we are assigning ignitions with some spatial uncertainty – but also allows us to side-step the technical modeling issues of the “imbalanced data set” with so few ignitions.

(4) MaxEnt works with relatively small sets of presence data. Ignitions are mercifully rare – good for us all, but bad for statistical power. Any “data hungry” approach whose best performance requires thousands (or more) of data points to fit well, will not work with the reportable ignitions data set.

(5) Under the hood, MaxEnt has similarities with logistic regression, which is a standard choice for “classification” problems like ignition occurrence. However, MaxEnt models are protected from overfitting to training data via regularization and make estimates with presence-only data.

(6) The Maxent software we are using generates derived features from combinations of and breakpoints within the model covariates, accounting for things like covariate interactions, step changes in response, etc., and regularization, eliminating features that don’t improve predictive power. These are not unique to Maxent but are necessary to achieve good out of sample predictive performance.

(7) The Maxent software used performs out of sample prediction testing, and reports modern/standard classification model performance metrics, like the ROC figures, ROC-AUC values, precision, recall, etc. These form the basis of our ability to objectively quantify its performance and compare to other approaches. While not unique to Maxent, these are required capabilities of any approach we would consider.

5.1.2 Data sources and preparation

5.1.2.1 Ignition data

CPUC reportable ignitions were selected as the training "event" data for the 2021 Wildfire Risk Model. The ignitions under study were filtered to have occurred between 2015 and 2018 (2019 data was used to test model predictive power), within HFTD Tier 2 or Tier 3, and to have occurred during the fire season (Jun. 1 – Nov. 30). Ignitions used to train the vegetation-caused ignition probability model had to additionally be labeled as caused by vegetation contact and those used to train the conductor-involved model had to be labeled with conductors as the asset that failed leading to the event. Note that those two sets of ignitions overlap in the case where vegetation damaged a conductor.

Conductor-involved ignitions: There were just under 850 outages (whose locations we use as ignition locations) associated with reportable ignitions that involved conductors from 2015 through 2018. A little under 300 of those occurred in HFTD Tier 2 or Tier 3. Just over 240 of the remaining ignition outages occurred during the fire season. Those events were the ones used to train to 2021 conductor-involved ignition probability model. 60% of those were vegetation-caused; 30% were caused by equipment failures, and the rest had a few miscellaneous causes, including animals and 3rd party contact (mostly car accidents).

Vegetation-caused ignitions: There were just under 470 vegetation-caused ignitions from 2015 through 2018. Right around 260 of those were found in HFTDs Tier 2 and Tier 3 and just over 220 additionally took place during the fire season. Those events were the ones used to train the 2021 vegetation-caused ignition probability model. Over 80% involved conductor damage and more than 75% were labeled as "wire down" events.

5.1.2.2 Explanatory variables (aka covariates)

To have a reportable ignition, fault current needs to be generated, the fault current creates an ignition, and the ignition needs to be viable enough to spread: utility has knowledge of the ignition, the fire travels greater than one linear meter from the ignition point, and the fire propagates beyond utility equipment. Thus, we understand reportable ignitions to be the product of assets interacting with their environment over time.

As visualized in Figure 2, the three categories of data categories of data relevant to modeling that process are: (1) asset attributes (2) spatially varying environmental conditions, determined by location (3) spatio-temporal varying weather conditions, determined by location and time. We are limited by the data available in each category, so we have prepared as many potentially relevant covariates as we can identify and lay our hands on.

(1) For asset attributes, we are interested in attributes that can be changed through mitigation and/or those that are expected to indicate or correlate with degradation. For example, age is expected to correlate with various forms of degradation, whereas, conductors' size and materials determine the susceptibility to structural failure and corrosion, respectively.
(2) For environmental covariates, we are interested in location-specific characteristics that impact vegetation, fuels, and asset health. For example, the coastal indicator is associated with marine layer salinity, a source of corrosion, climatic dryness determines the long-term viability of grass, chaparral, and trees, and terrain determines how sheltered or exposed a location is to wind.
(3) For weather covariates, we are interested in the more proximate environmental causes of failures (like wind and gusts) and factors that influence ignition viability and spread (like ground cover, fuel moisture, and wind). However, we are evaluating these on the timescale of entire fire seasons, so covariates must reflect temporal aggregation, capturing the typical or extreme values of each or some cumulative count or "exposure" to dangerous conditions across the season(s).

There is a detailed discussion of the "pool of variables" in the ignition probability methods appendix (Appendix 3: Ignition Probabilities Methods 2021).

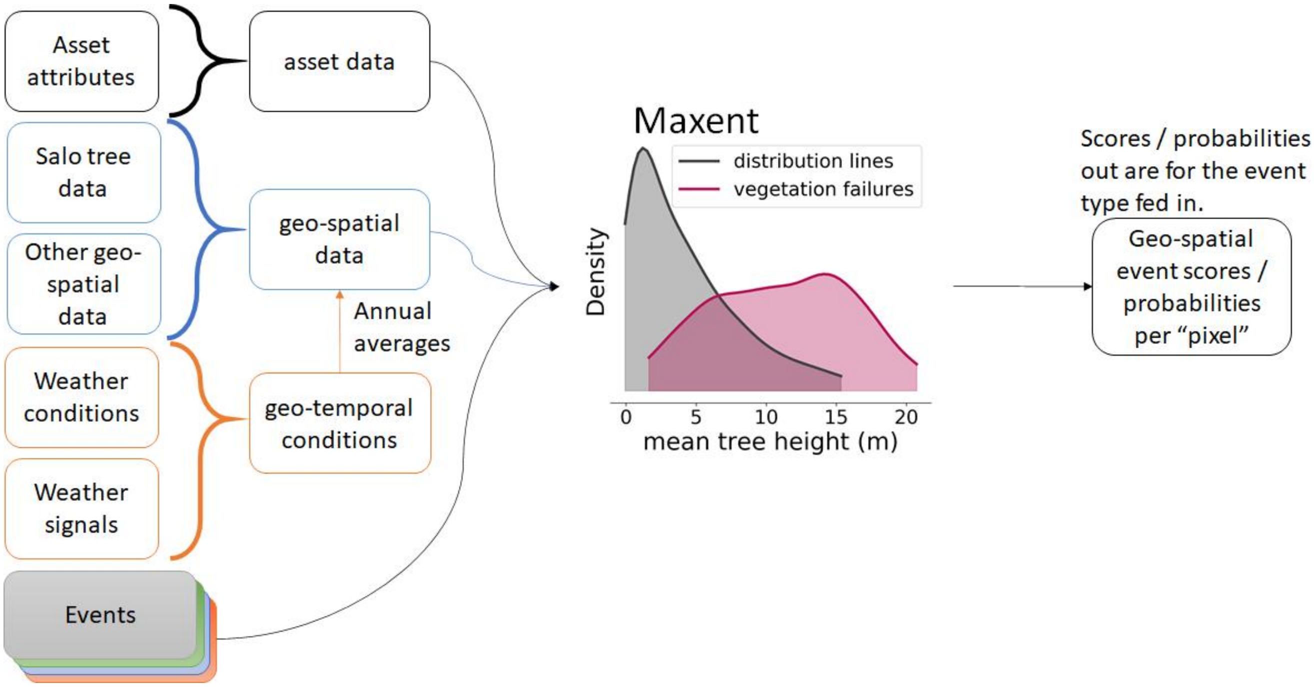


FIGURE 2 - SCHEMATIC OF DATA FLOW THROUGH THE MAXENT MODEL

5.1.2.3 Covariate selection

Modelers us a term "parsimony" to capture the concept that models should be as simple as they can be while still explaining the underlying process. Parsimony is not objectively quantifiable but is not aesthetic either. Without parsimony, a model can overfit the training data, undermining its predictive power and the interpretation of any given covariate can be entangled with the contributions of others like it. We achieved parsimony through two mechanisms: (1) The Maxent modeling software we used "regularizes" model fit by dropping covariates that don't contribute to performance gain when testing out of sample, thus decreasing the risk of overfitting and providing metrics we can evaluate to judge how well it has done. (2) When in possession of multiple covariates that contain similar information or covariates that are directly relevant to mitigation, we have made "editorial" decisions about which covariates to include or exclude while checking that overall performance is not degraded.

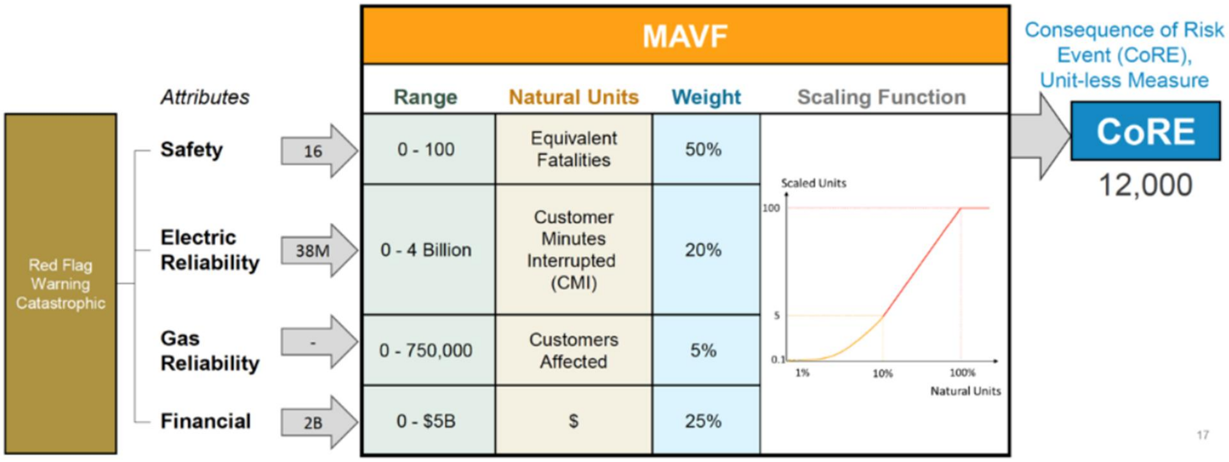
5.2 Methods: Wildfire Consequence

PG&E uses MAVF to calculate the consequence of an event. MAVF is a function for combining consequence impacts of the occurrence of a risk event and creating a single unit-less risk value, known at PG&E as MAVF or MARS. Some of its key features are:

- It formalizes trade-offs between different dimensions of consequence attributes (Safety, Reliability and Financial).
- It captures aversion or indifference over a range of outcomes based on the company's risk management approach.
- It allows comparisons of risk across the company using a common scoring metric.

5.2.1 How MAVF Risk Scoring Works

Figure 3 is the MAVF approved by PG&E's Risk committee for use across company for risk scoring.



Range	Natural Units	Weight
0 - 100	Equivalent Fatalities	50%
0 - 4 Billion	Customer Minutes Interrupted (CMI)	20%
0 - 750,000	Customers Affected	5%
0 - $5B	$	25%

FIGURE 3 - MAVF

The consequence attributes and their respective weights are:

- Financial (25%)
- Safety (50%)
- Electric Reliability (20%)
- and Gas Reliability (5%)

Each outcome in the Consequence model is assigned a score for these categories which is then aggregated to calculate the consequence score.

The consequence values assigned to each simulated fire come from these existing MAVF consequence scores. MAVF divides wildfire risk events into severity categories, modeling each category as a separate set of inputs (think tabulations/counts of historical ignitions that fit into each severity category) and consequence outcomes.

Because the inputs come from multiple sources into the central risk event calculation and then fan back out to the Safety, Reliability, and Financial risk categories, each category is called a risk “bow tie” after what it looks like when diagrammed.

The risk bow tie methodology is a structured way of conceptualizing, representing risk across many types of events. It allows for the risk event to be broken down into the causes, or drivers, of a risk event and the consequences resulting from the risk event. Groupings of drivers or outcomes can be considered as separate tranches and the consequences of the risk event can be calculated for each of these tranches. Tranches segment a system of assets into “like” risk groups because different parts of a system face different hazards, are susceptible to those hazards to different degrees and can result in different consequences given the same event. For instance:

- Material: plastic is not threatened by corrosion compared to metal
- Location: Earthquake in Oakland vs Santa Cruz
- Ambient Conditions: Proximity to vegetation. (combustible material)

A bow tie (Figure 4) quantifies relationships between drivers and outcomes.

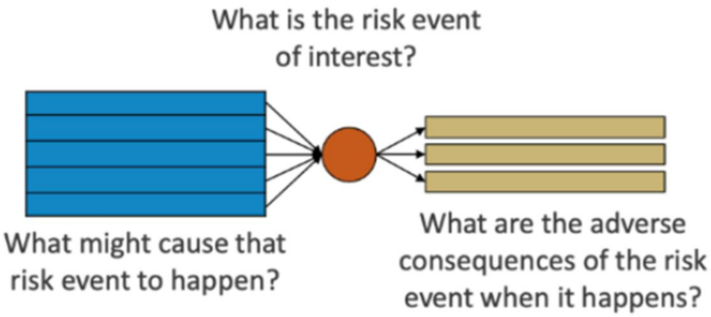

What is the risk event of interest?
What might cause that risk event to happen?
What are the adverse consequences of the risk event when it happens?

FIGURE 4 – BOW TIE STRUCTURE

Under the hood, there are as many bow ties as there are tranches (Figure 5).

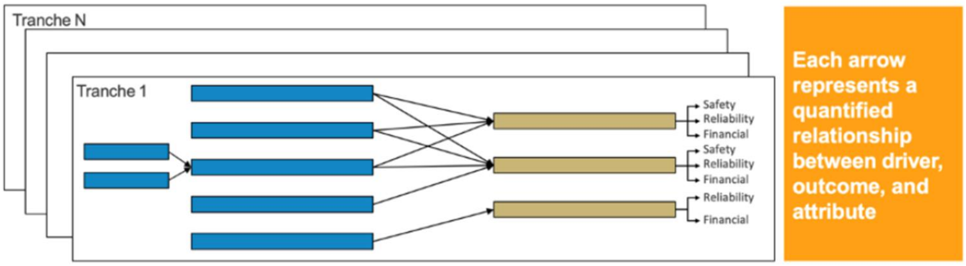


FIGURE 5 - TRANCHES

Figure 6 below provides an example wildfire bow tie.

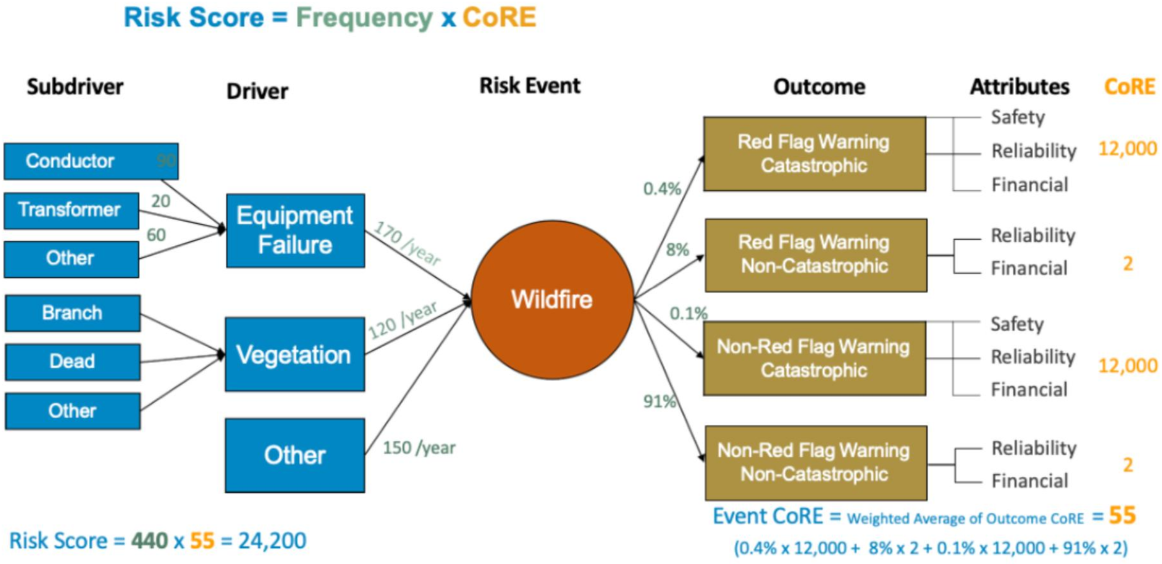

Risk Score = Frequency x CoRE
Subdriver
Driver
Risk Event
Outcome
Attributes
CoRE
Conductor
90
Transformer
20
Other
60
Branch
Dead
Other
Equipment Failure
Vegetation
Other
170 /year
120 /year
150 /year
Wildfire
0.4%
8%
0.1%
91%
Red Flag Warning Catastrophic
Red Flag Warning Non-Catastrophic
Non-Red Flag Warning Catastrophic
Non-Red Flag Warning Non-Catastrophic
Safety
Reliability
Financial
12,000
Reliability
Financial
2
Safety
Reliability
Financial
12,000
Reliability
Financial
2
Event CoRE = Weighted Average of Outcome CoRE = 55
(0.4% x 12,000 + 8% x 2 + 0.1% x 12,000 + 91% x 2)
Risk Score = 440 x 55 = 24,200

FIGURE 6 - EXAMPLE WILDFIRE BOW TIE

5.2.2 Deriving Spatial MAVF CoRE Values

What matters for our purposes is that each bow tie produces CoRE consequence values specific to the categories of events that feed into it and these can become a lookup table for consequence of simulated wildfires as long as they can be mapped into the same categories.

For the 2021 Wildfire Distribution Risk Model effort, which was designed from the ground up to deliver spatially resolved results, the challenge was to map MAVF CoRE values onto a spatial grid. Historically, risk assessments using MAVF scoring have been performed at the enterprise-level without spatially explicit data or models. In other words, the risks are computed in terms of the expected count and severity of "risk events" but not at their specific locations. The purpose of the 2021 Wildfire Distribution Risk Model is to model the spatial variation in risk so that wildfire mitigation efforts can prioritize higher risk assets and locations for mitigation. The development of corresponding spatial MAVF CoRE consequence metrics required mapping the characteristics of every "grid pixel" in the HFTD areas to the categories used to assign ignitions to tranches of consequence already in use in the MAVF framework. These categories include HFTD areas, red flag warning conditions, and fire severity

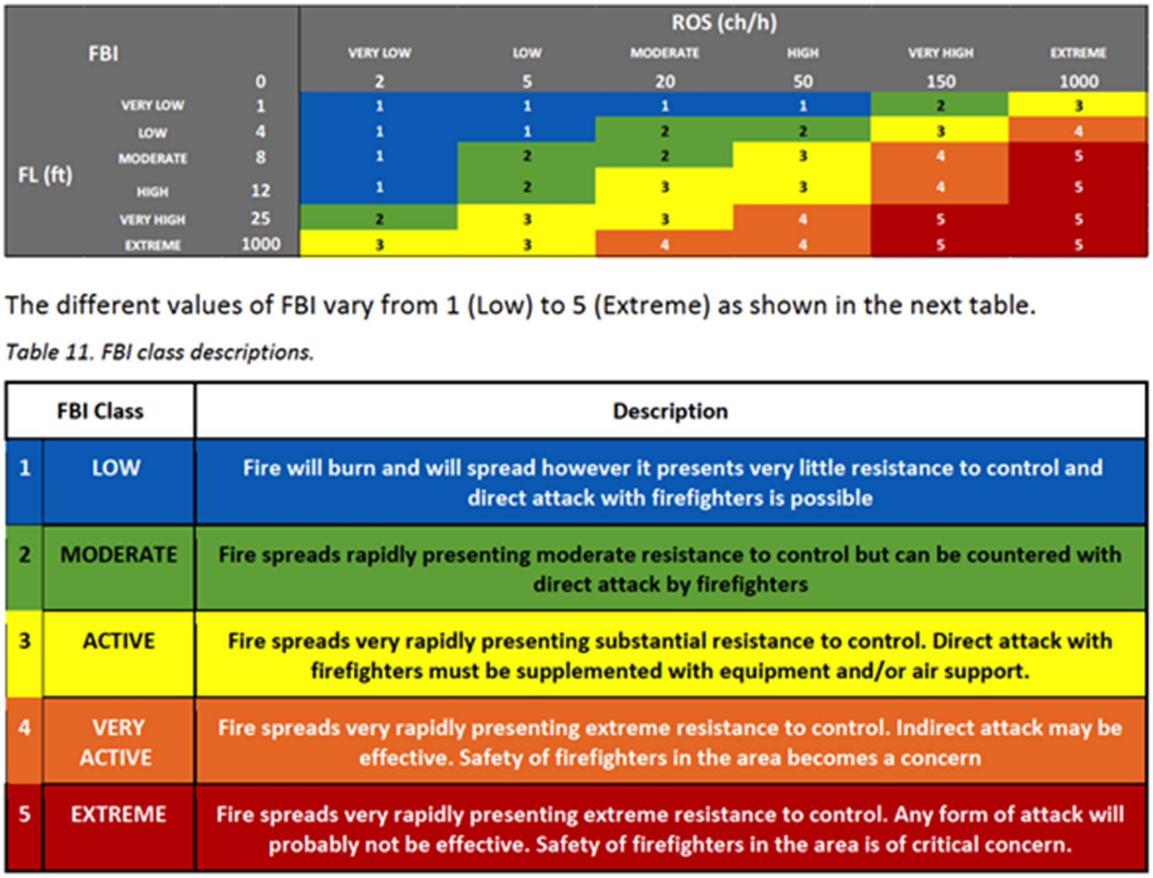
Thus, the spatial consequence values for the 2021 model required spatial estimates of:

1. A simple spatial indicator of whether a given location is within the HFTDs
2. The probability that a location will be under red flag warning at the time of an ignition
3. A spatial breakdown of the likelihood that an ignition would lead to a small, large, destructive, or catastrophic wildfire, given its starting location

Given such estimates, the existing MAVF CoRE values from corresponding bowtie tranches could be applied to each location. The first is very straight forward. We have geo-spatial shape files of the HFTD, so any given location can be assigned an "HFTD indicator". The second was more challenging, but there are also shape files available for every red flag warning called. By stacking those shapes on top of one another, the count of red flag warnings per-fire-season at every location was calculated and rendered into a probability of a red flag warning for any given day.

The fire severity calculation was by far the most complex component of wildfire consequence to estimate (and the most significant in determining the MAVF CoRE values). Technosylva fire simulations under extreme fire weather conditions were used to estimate the likelihood of ignitions growing into fires of Small, Large, Destructive, or Catastrophic extent (these are PG&E specific MAVF wildfire categories), based on Technosylva's fire characteristics, including:

1. The burn area in acres
2. The number of structures within the burn area
3. Technosylva's Fire Behavior Index, assigned on a scale of 1-5 based on the combination of simulated flame length (a metric of burn intensity) and rat of spread (see Figure 7 below for FBI details)



FBI			ROS (ch/h)					
			VERY LOW	LOW	MODERATE	HIGH	VERY HIGH	EXTREME
		0	2	5	20	50	150	1000
FL (ft)	VERY LOW	1	1	1	1	1	2	3
	LOW	4	1	1	2	2	3	4
	MODERATE	8	1	2	2	3	4	5
	HIGH	12	1	2	3	3	4	5
	VERY HIGH	25	2	3	3	4	5	5
	EXTREME	1000	3	3	4	4	5	5

The different values of FBI vary from 1 (Low) to 5 (Extreme) as shown in the next table.

Table 11. FBI class descriptions.

FBI Class		Description
1	LOW	Fire will burn and will spread however it presents very little resistance to control and direct attack with firefighters is possible
2	MODERATE	Fire spreads rapidly presenting moderate resistance to control but can be countered with direct attack by firefighters
3	ACTIVE	Fire spreads very rapidly presenting substantial resistance to control. Direct attack with firefighters must be supplemented with equipment and/or air support.
4	VERY ACTIVE	Fire spreads very rapidly presenting extreme resistance to control. Indirect attack may be effective. Safety of firefighters in the area becomes a concern
5	EXTREME	Fire spreads very rapidly presenting extreme resistance to control. Any form of attack will probably not be effective. Safety of firefighters in the area is of critical concern.

Figure 7: Technosylva's Fire Behavior Index components and description

These characteristics were then used to lookup existing MAVF CoRE values for corresponding tranches and used to compute fire severity assignments for each of the hundreds of simulations conducted per-location. Then the consequence for each simulation outcome could be averaged across all days simulated into averages (and other statistical summaries) of the consequence values for every grid location in the HFTDs areas.

The detailed recipe for using Technosylva simulations and their metrics to create calibrated MAVF CoRE consequence values is:

(1) Assign ignition simulation locations at regular (200m) spacing along all grid locations within HFTDs Tier 2 and Tier 3.
(2) Tabulate the 452 worst historical fire weather days using historical weather data.
(3) For all locations, run a separate 8-hour fire spread simulation for each day of weather data, recording burn area, flame length, impacted structures and FBI on a scale of 1 to 5 for each simulation.
(4) Using pre-existing MAVF consequence scores calculated for all combinations of fire severity (Small, Large, Destructive, Catastrophic), an HFTD indicator, and a red flag warning indicator rendered into a location-specific probability of a red flag warning, assign each simulation output a consequence score.
(5) The rules developed for assigning MAVF fire size to each Technosylva simulation result are:
    a. Small Fire (area < 300 acres)
    b. Large Fire (area > 300 acres)
    c. Destructive Fire (area > 300 acres & 50+ structures impacted)

d. Catastrophic Fire (assigned by ratio of Catastrophic to Destructive fires historically)

(6) Compute statistical extracts of consequence scores for all available simulations at each location – most downstream usage is based on the mean, but variance and others can also be useful.

(7) Assign the resulting mean consequence to each ignition location.

(8) Ensure that simulations can be mapped to all HFTD Tier 2 and Tier 3 grid locations. To do this, simulation output metrics are associated with a 200m x 200m raster pixel with the ignition point in the center, so the results can be assigned spatially to any locations within each pixel.

The details of the spatial consequence modeling methods are found in Appendix 4: Ignition Consequence Methods 2021

5.3 Application: Vegetation-Caused Ignitions for EVM

All vegetation-caused CPUC reportable fire season ignitions from 2015 to 2018 within the HFTD areas were used to model the risk addressed by the EVM program[2]. PG&E withheld 2019 ignition data for use in testing and validating the out of sample predictive power of the model. A MaxEnt model was used to estimate spatial ignition probabilities based on those ignitions. This work was informed by data on vegetation, weather and other environmental conditions. The ignition probabilities were combined with the MAVF CoRE values from the spatial ignition consequence data set to produce 100m x 100m grid-pixel-level risk scores. The pixelated risks were aggregated within each circuit segment (also called Circuit Protection Zone or CPZ) in the HFTD areas to produce the risk summaries provided as inputs used to inform EVM planning and prioritization.

A detailed account of the EVM risk modeling is found in Appendix 1: Vegetation-caused Ignition Risk Model 2021 as well as slides from a presentation on the modeling for a general audience are found in the document named: *EVM Risk Model 2021 - Lunch n' Learn presented 2020_10_21* as well as a separate conference presentation found in the document named: *EVM Risk Model 2021 - Utility Analytics conference presented 2020_10_29*.

Ignition likelihood for vegetation in 2021 was determined based on a probability analysis predicting ignitions in 100m x 100m pixels. The Vegetation Probability of Ignition Model was trained on vegetation ignitions limited to fires season events and CPUC reportable ignitions from 2015 to 2018 and tested using the 2019 ignitions. This data set includes all vegetation related outages that resulted in an ignition. The modeling technique used was a maximum entropy model which provides a way of estimating the relative occurrence rate given a fairly modest number of ignition locations. The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge is the one with the largest entropy, in the context of precisely stated prior data.

100m pixel representation of P(ignition) output from the Vegetation Probability of Ignition Model for the North Bay is shown in Figure 8 below - red is higher, blue is lower, non-HFTD conductors are shown in dark grey.

[2] Note that vegetation-caused conductor-involved ignitions were also modeled by the conductor model.

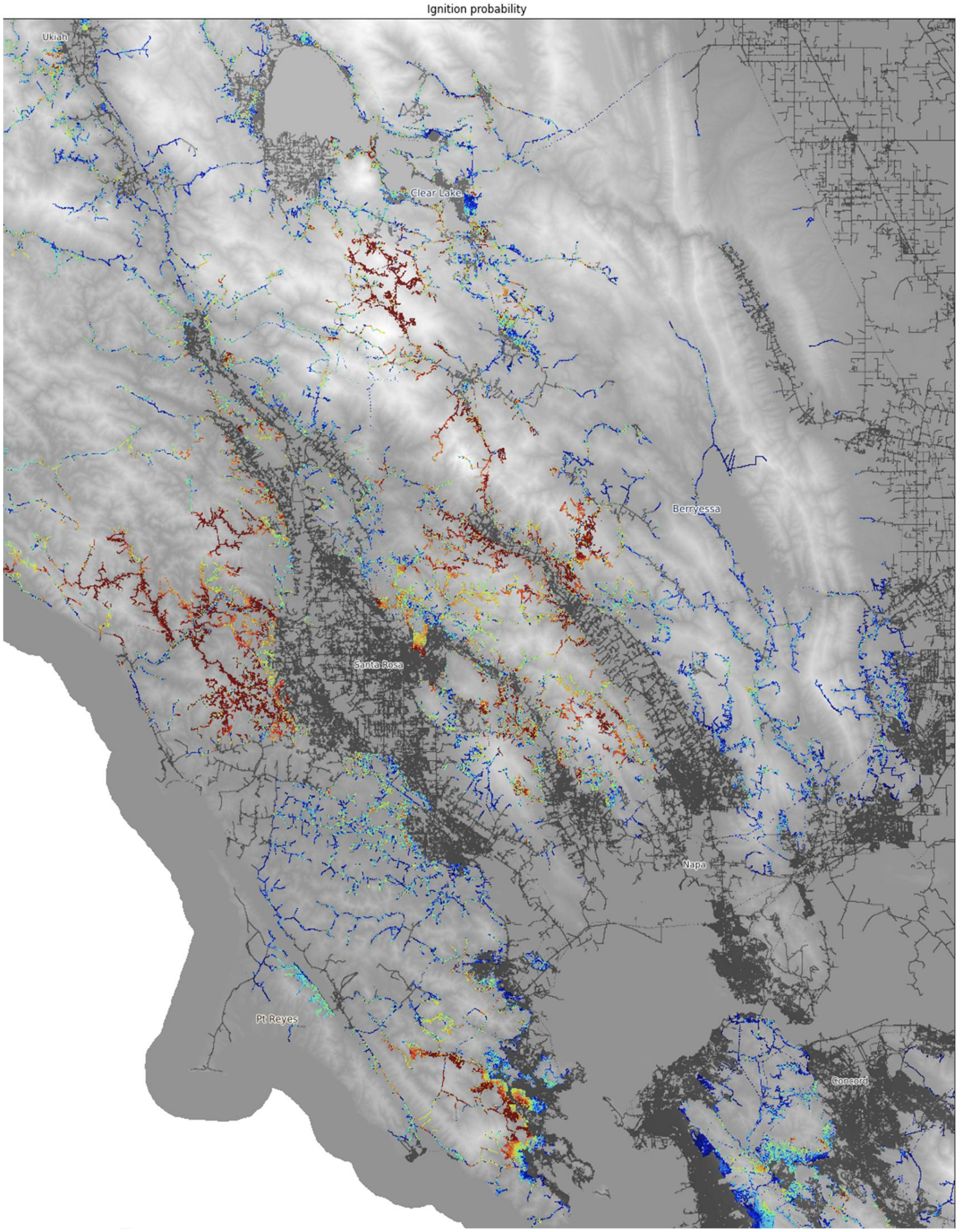


Figure 8 – IGNITION PROBABILITY PER PIXEL FOR THE NORTH BAY - RED IS HIGHER, BLUE IS LOWER COLORED GRID PIXELS ARE WITHIN HFTDS, DARK GRAY GRID PIXELS ARE NOT.

Examples based on the model results rolled up to CPZ summaries are presented below. Interestingly, there are fewer trees (based on the database of known trees maintained by vegetation management) in areas of high consequence: Figure 9 shows scatter plots of per-CPZ data for all 3,000 CPZs analyzed where the y-axis is the count of trees in each CPZ and the x-axis is MAVF CoRE consequence. It shows that higher tree counts tend to be associated with lower consequence values – in other words, there are fewer trees in locations with elevated fire consequences.

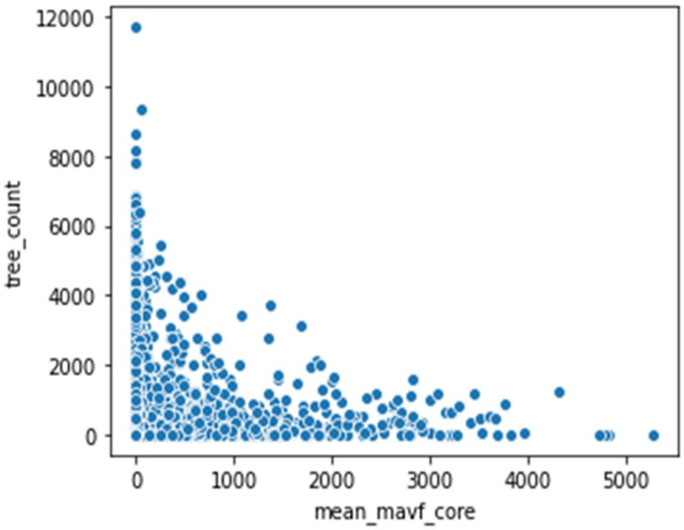


FIGURE 9 - SCATTER PLOT OF CPZS BY CONSEQUENCE AND TREE COUNT

Figure 10 Shows there is not as strong a relationship between P(ignition) and VMD tree density. But the highest P(ignition) values are generally associated with CPZs with fewer trees. This is likely due to the fact that reportable ignitions require dry fuels to grow to reportable proportions (1 m in extent) and areas with fewer trees tend to be hotter and dryer.

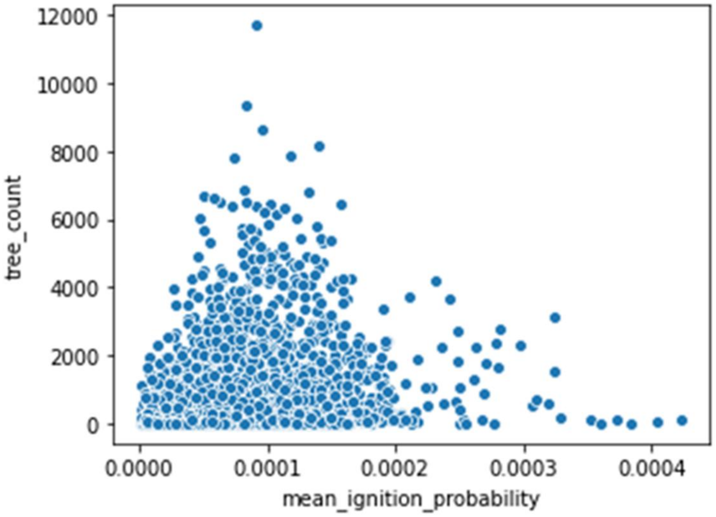


FIGURE 10 - SCATTER PLOT OF CPZS BY P(IGNITION) AND TREE COUNT

Risk is equal to P(ignition) x C(ignition) but we can see that the resulting scores are heavily dominated by Consequence values. Figure 11 plots the Risk score on the y-axis and the two components of that risk calculation (Consequence on the left and P(ignition) on the right) on the x-axes. From these, it can be verified that the Risk score is highly correlated with the Consequence (MAVF CoRE) and less correlated with the P(ignition). This has a lot to do with the fact that the Consequence values range over more orders of magnitude than the P(ignition) values. If you are prioritizing directly by Risk, you are largely prioritizing by Consequence, or the ability for a given location to host a catastrophic wildfire.

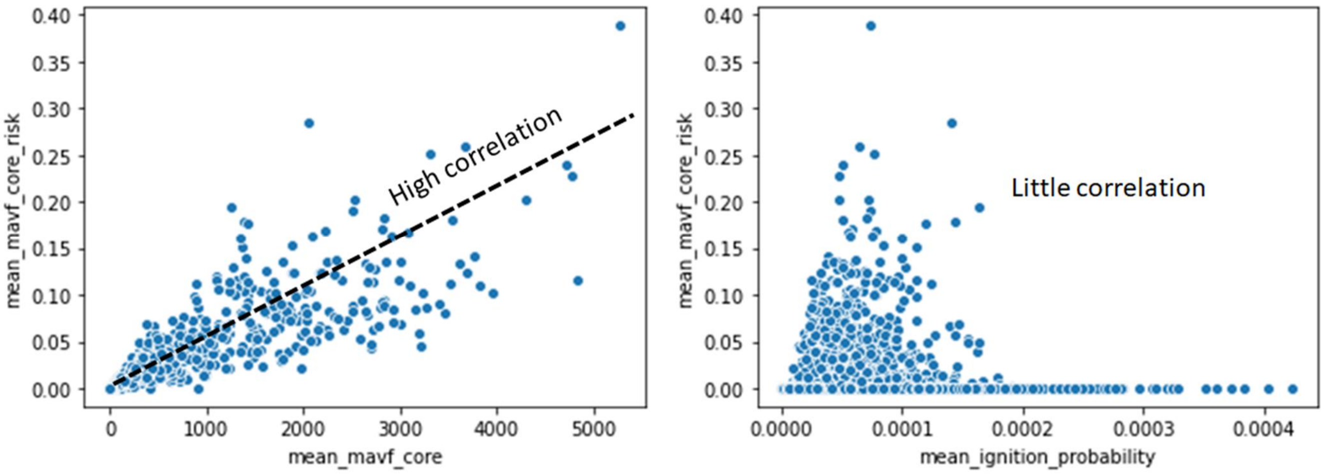


FIGURE 11: SCATTER PLOTS PER-CPZ RISK CORRELATED WITH CONSEQUENCE (LEFT) AND IGNITION PROBABILITY (RIGHT)

Variables in the model included meteorology data, PG&E asset data, and remote sensing data from government and private third parties. A metric called "permutation importance" can be used to quantify how sensitive the model's predicted outputs are to random fluctuations in the given variable's (aka covariate's) input values. The permutation importance of the covariates used in the Vegetation-caused Ignition Probability Model are included below in Table 2. The Pool of covariates section of the MaxEnt ignition probability estimation methods appendix provides detailed information on the meaning and data source of each of the covariates named below.

TABLE 2: VARIABLES IN THE VEGETATION-CAUSED IGNITION PROBABILITY MODEL

Rank	Model Feature	Feature description	Units	Permutation Importance (%)
1	tree-height-max	Satellite derived tree height estimates – highest tree per-raster pixel	m	26.1
2	100-hour-fuels-avg	standard fire modeling metric of fuel dryness for fuels about 1-3" in diameter, mean over season	%	24.1
3	vapor-pressure-deficit-avg	vapor pressure deficit, mean over season	kPa	21.6
4	gusty-summer-day-pct	The percentage of days with sustained hourly wind speeds over 20 mph	%	6
5	HFTD	High Fire Threat District (2 or 3)		4.2
6	precipitation-avg	Seasonal daily average precipitation	mm	3.1
7	Impervious	NLCD imperviousness product - represent urban impervious surfaces as a percentage of developed surface	%	2.8
8	specific-humidity-avg	Seasonal average specific humidity	kg/kg	2.4
9	burn-index-avg	National Fire Danger Rating System (USNFDRS) Burning Index (BI)		2.3
10	wind-max	Annual 99th percentile hourly wind speed at 10m	m/s	1.9
11	temperature-avg	Average of daily maximum temperature in Kelvin	K	1.6
12	windy-summer-day-pct	The percentage of days with sustained hourly wind speeds over 15 mph	%	1
13	local-topography	The topographic position index (TPI) extracted from the USGS national elevation dataset		0.8
14	tree-height-avg	Satellite derived tree height estimates – average per-raster pixel	m	0.8
15	1000-hour-fuels-avg	standard fire modeling metric of fuel dryness for fuels about 3-8" in diameter, mean over season		0.6
16	energy-release-avg	USNFDRS Energy Release Component (ERC)		0.4

Using these variables, a probability of ignition was assigned for each 100m x 100m grid. These probabilities were indexed and calibrated to the total expected ignition frequency.

Updates to this model are planned on an annual basis. In 2021, PG&E aims to incorporate LiDAR informed tree species data so that the predictive power of vegetation caused ignition probabilities will be enhanced to better inform mitigation programs.

5.3.1 EVM model validation

The dataset used to train the model achieved an AUC score of 0.73. The 2019 dataset was used as an out-of-sample test dataset to evaluate the model fit and achieved a score of 0.64 but a randomly withheld test sample from several years achieved a score of 0.72. The minimal reduction in AUC score between the training and testing datasets gives confidence that the model is not overfitting to the training dataset but also raises the possibility that the spatial pattern and other characteristics of 2019 vegetation-caused ignitions deviated slightly from 2015-2018. See the dedicated model validation section, Discussion: Model Validation and Comparison to Previous Work for more discussion of model validation.

5.3.2 EVM model insights and applications

Insights

Vegetation-caused ignitions quite obviously require the presence of fall-in trees close enough and tall enough to contact the overhead circuit. Along similar lines, we expect that all else being equal, dryer and windier conditions will favor both branch failures and fire viability and spread. This explains the sensitivity of the model to tree presence and height data and to metrics of fuel dryness, gustiness, and vapor pressure deficit. However, there are also some counter-intuitive relationships that have emerged from the modeling efforts.

First and foremost, tall trees do not tend to be found under either consistently windy or consistently dry conditions. For the most part, they prefer more benign habitats, but also, their presence lowers local temperatures, increases local humidity and moisture, and lowers local wind speeds. Dryness and wind are major contributors to wildfire risk, but only when they are somewhat anomalous compared to prevailing conditions.

The above relationships contribute to another somewhat counter intuitive result – the areas of highest ignition probability are not the areas of highest ignition consequence. Fires burning in forested areas with mature tall trees are indeed very dangerous, but ignitions start small and often start on the ground. They are more viable as fires that spread to forested areas when they originate under conditions that offer a mix of smaller and larger fuels that are drier and more open to wind than heavily forested areas. Such areas include both man-made and natural transitions from more open or mixed ground cover to more heavily forested areas – and the man-made ones are guaranteed to be proximate to people and structures. Taking a concrete example from the 2021 modeling effort, fire consequences were found to be higher in the Sierra foothills than in higher elevations hosting unbroken forests whereas the ignition probabilities were often found to be higher in within the forests. The recognition of elevated “downhill” consequence is a significant development and improvement compared to earlier modeling.

It is also clear from first principles and confirmed by arborists that not all trees are equally dangerous. Tree species and individual tree health as well as other natural phenomena like pine beetles and plant pathogens like the oak death fungus can all alter the odds and type (branch, trunk, or root) of tree failures. The vegetation management team maintains high quality databases of tree characteristics in general and failed trees in particular based on field observations. However, those data sets are not (yet) comprehensive enough to support full coverage predictions for all grid locations. Field observations combined with remote sensing from lidar and satellite-based surveys hold the promise to improve our understanding and ability to model these relationships in the future.

Applications

The proximate purpose of the vegetation-caused wildfire risk modeling effort was to support risk-informed planning of EVM activities. The 1+ million pixelated spatial risk results were aggregated into representative values for several thousand grid segments called protection zones, with ignition probabilities, consequences, and risk values all reported.

The results were not just delivered in complete form at the end of the modeling effort. Rather, several iterations of the work were produced via ongoing collaborative discussion of the work with the Vegetation Management team. These discussions included regular modeling updates and discussions as well as broader team review of the geography of risk produced by the modeling runs compared to the on the ground working knowledge of risk from division experts. The discussions revealed the distinction between the probability of ignition and ignition consequences – EVM manages the probability of ignition for the most part – and the importance of tree species and health to risk and the choices arborists make in the field. It also revealed the insufficiency of data sets on hand (gathered for other purposes) to support grid-wide predictions based on species or individual tree health. Thus, it was determined that such considerations would be applied by vegetation management experts on top of the model results and that the model team would pursue more comprehensive sources of related data for future use.

5.4 Application: Conductor-Involved Ignitions for System Hardening

All conductor-involved CPUC reportable fire season ignitions from 2015 to 2018 (2019 was held back for testing predictive power) within the HFTDs were used to model the risk addressed by the System Hardening program.[3] A MaxEnt model was used to estimate spatial ignition probabilities based on those ignitions. The ignition probabilities were combined with the MAVF CoRE values from the spatial ignition consequence data set to produce 100m x 100m grid-pixel-level risk scores. This work was informed by data on conductor materials and size, proximity to the coast, and the location of splices. Prior work within PG&E informed our interest in these data fields. The pixelated risks were aggregated within each circuit segment in the HFTD areas to produce the risk summaries provided as inputs used to inform system hardening planning and prioritization.

A detailed account of the Conductors Risk modeling is found in Appendix 2: Conductor-Involved Ignition Risk Model 2021 and slides from a presentation on the modeling for a general audience are found in the document named: *Conductor Risk Model 2021 - Lunch n' Learn presented 2020_10_28*.

Ignition likelihood for equipment in 2021 was determined based on a probability analysis predicting ignitions in 100m x 100m pixels. The Equipment Probability of Ignition Model was trained on conductor failure related ignitions limited to fire season events and CPUC reportable ignitions from 2015 to 2018 and tested using the 2019 ignitions. The modeling technique used was a maximum entropy model which provides a way of estimating the relative occurrence rate given a fairly modest number of ignition locations. The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge is the one with the largest entropy, in the context of precisely stated prior data.

Figure 12 shows that probably and consequence of conductor-involved ignitions are not highly correlated. Locations with elevated likelihood of ignition typically have a small consequence value. This makes mitigation work prioritization more difficult because there are not a clear cluster of locations with high consequence and high probability.

[3] Note that vegetation-caused conductor-involved ignitions were also modeled by the Vegetation Risk Model.

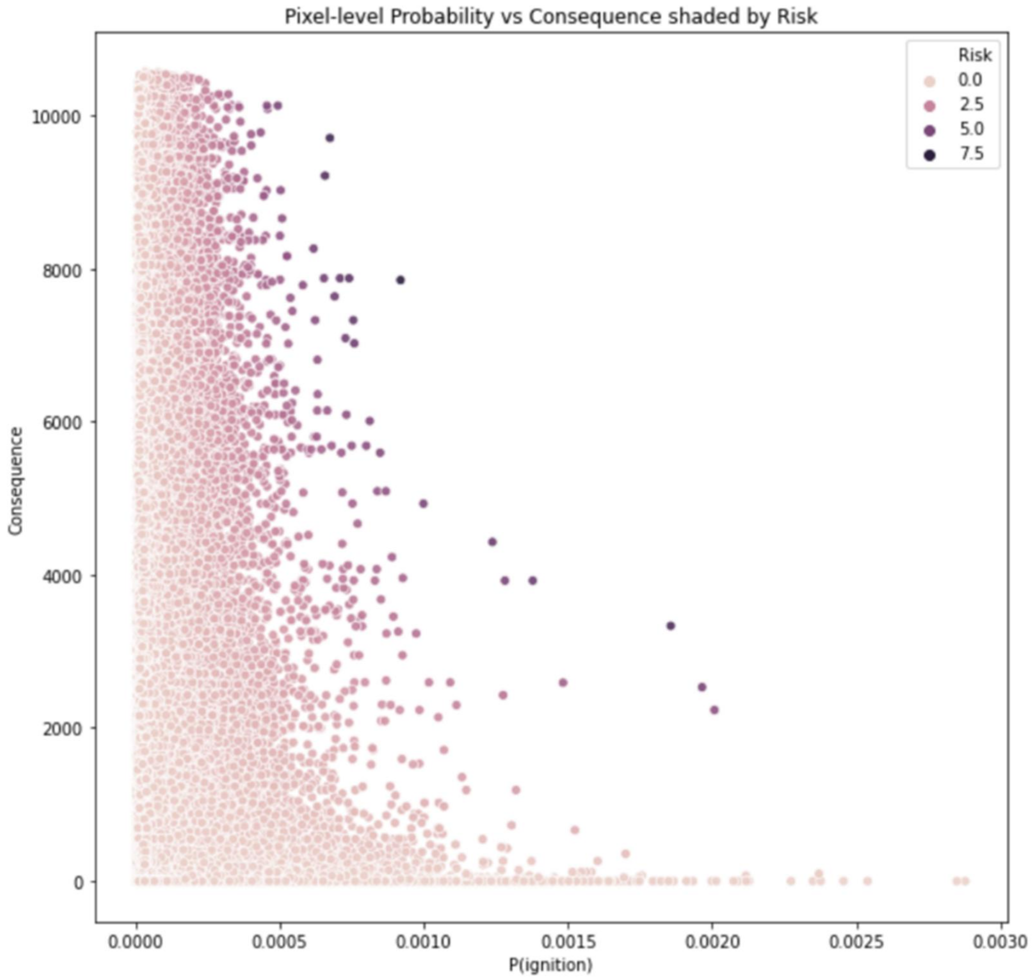


FIGURE 12 - SCATTERPLOT OF THE PIXEL-LEVEL PROBABILITY OF IGNITION ON THE X-AXIS AND MAVF CONSEQUENCE ON THE Y-AXIS WITH SHADING BASED ON THE MAVF RISK VALUES

.

The dominance of consequence can further be demonstrated in the comparison of the ignition probability image and the ignition risk image in the Sonoma area (Figure 13). The locations with a higher likelihood of ignition in the probability image are shown as lower risk areas in the risk image. It may be beneficial to scale the consequence values to gain more influence from the likelihood of an ignition event occurring.

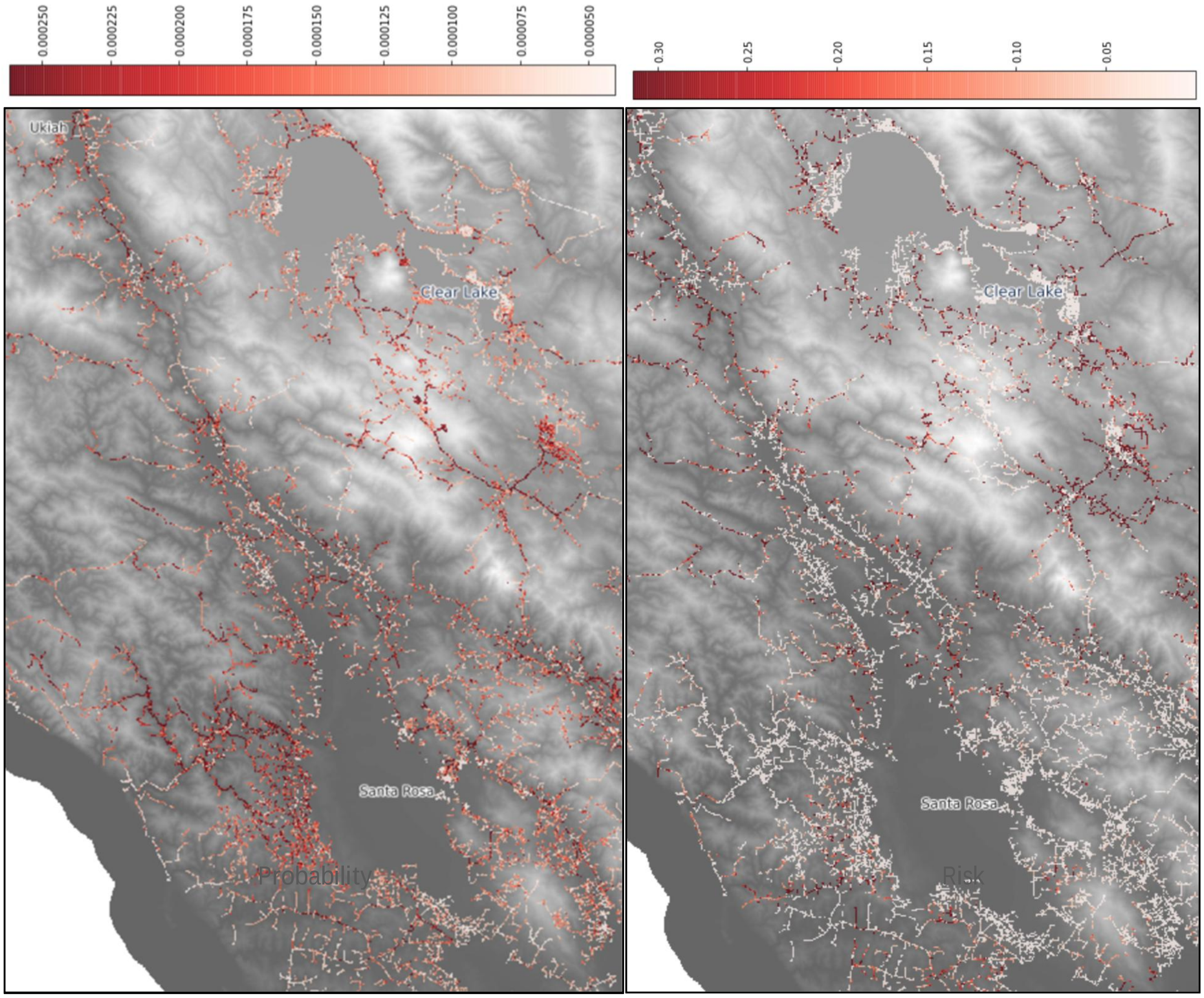


FIGURE 13: COMPARISON OF PROBABILITY AND RISK PIXEL-LEVEL RESULTS IN THE SONOMA AREA. OBSERVE THAT THE AREAS WITH A HIGHER LIKELIHOOD OF IGNITION IN THE PROBABILITY IMAGE (LEFT) ARE SHOWN AS LOWER RISK AREAS IN THE RISK IMAGE (RIGHT). THIS IS THE INFLUENCE FROM THE CONSEQUENCE DATASET.

A range of variables were included in the initial modeling. These included meteorology data, PG&E asset data, and remote sensing data from government and private third parties. A metric called "permutation importance" can be used to quantify how sensitive the model's predicted outputs are to random fluctuations in the given variable's (aka covariate's) input values[4]. The permutation importance of the covariates used in the Conductor-involved Ignition Probability Model are identified below in Table 3. The Pool of covariates section of the MaxEnt methods appendix provides detailed information on the meaning and data source of each of the covariates named below.

TABLE 3: VARIABLES IN THE CONDUCTOR-INVOLVED IGNITION PROBABILITY MODEL

Rank	Model Feature	Feature Description	Units	Permutation Importance
1	Unburnable	non-burnable area	%	30.8
2	precipitation_ave	daily precipitation, mean	mm	29.8
3	conductor_material_acsr	conductor material: ACSR	%	9.7
4	estimated_age	estimated conductor age	years	8.9
5	tree_height_max	max tree height	m	4.3
6	splice_record_exists	Reliability Program splice	%	4.3
7	vapor_pressure deficit_ave	vapor pressure deficit, mean	kPa	4.0
8	conductor_size_2	conductor size: 2	%	3.4
9	conductor_size_4	conductor size: 4	%	1.6
10	100_hour_fuels_ave	100-hour fuel moisture, mean	%	1.1
11	max_temperature_ave	max temperature, mean	K	1.0
12	wind_ave	wind speed, mean	m/s	0.9
13	local_topography	TPI	%	0.2
14	conductor_size_6	conductor size: 6	%	0.1
15	conductor_material_al	conductor material: Al	%	~0
16	conductor_material_cu	conductor material: Cu	%	~0
17	coastal	coastal	%	~0
18	specific_humidity_ave	specific humidity, mean	%	~0

Using these variables, a probability of ignition was assigned for each 100m x 100m grid. These probabilities were indexed and calibrated to the total expected ignition frequency.

Given the amount of time required to develop and test new models, PG&E was only able to include in the Conductor-involved Probability of Ignition Model results in the 2021 Wildfire Distribution Risk assessment – these are the most common and riskiest among the equipment-involved ignitions. Updates to this model are planned on an annual basis. In 2021, we aim to model equipment-caused risk from pole and transformer failures, and to add maintenance tag and asset data in the (renamed) combined Equipment-caused Ignition Probability Model. These additional equipment models will combine with an update to the conductor failure model to improve the predictive power of equipment caused ignition probabilities will be enhanced to better inform mitigation programs.

[4] According to Phillips (2006), "The contribution for each variable is determined by randomly permuting the values of that variable among the training points (both presence and background) and measuring the resulting decrease in training AUC. A large decrease indicates that the model depends heavily on that variable. Values are normalized to give percentages."

5.4.1 Conductor model validation

The dataset used to train the model achieved an AUC score of 0.76. The 2019 dataset was used as an out-of-sample test dataset to evaluate the model fit and achieved a score of 0.74. The minimal reduction in AUC score between the training and testing datasets gives confidence that the model is not overfitting to the training dataset and is able to maintain performance when introduced to new data. See the dedicated model validation section, Discussion: Model Validation and Comparison to Previous Work for more discussion of model validation.

5.4.2 Conductor model insights and applications

First and foremost, the majority of conductor-involved ignitions are due to contract from vegetation. A significant portion of the model fit has to do with either the location of trees or the suitability of local fuels to sustain a fire. The prominence of “unburnable”, which is a very specific data set used widely in the fire modeling community that is derived from the LANDFIRE surface fuel model, may appear to be a completely obvious realization of the fact that sparks on truly unburnable surfaces don’t start fires. However, if unburnable is excluded from the model it is replaced by tree height as the #1 ranked covariate. This underscores the fact that unburnable is capturing locational information about where vegetation-caused outages are possible along with where ignitions are viable.

The contribution to the model fits made by conductor asset attributes, including size, material, and age, are relatively modest individually, but significant collectively. They are consistent with known patterns, like copper corrosion in marine air near the coast, that smaller conductors are structurally weaker, and that older equipment is more likely to suffer from degradation, but all of these are also complicated by our ability to model correlation but not cause and effect. For example, different equipment standards have been in place at different times and equipment that fails often can be expected to see more repair work. Thus there are spatial patterns in equipment attributes governed by when and why equipment was installed where the attributes of the equipment are not the cause of failures. It is also worth noting that vegetation contact with enough force to bring a line down (for example due to trunk failures) will often hit with enough force to bring down even the most robust conductor hardware.

The conductor model results shared a similar outcome with the vegetation-caused model with ignition probabilities and consequence values displaying low correlation. With so many conductor-involved ignitions being caused by vegetation, many of the same explanations for that pattern offered in the vegetation-caused model discussion apply here as well: vegetation moderates wind, dryness, and other conditions of fire spread. However, among the other failure modes, we do observe wind-driven equipment failures. Since wind plays a prominent role in fire spread, future work will focus greater attention on isolating and modeling the causes of such events.

5.5 Discussion: Model Validation and Comparison to Previous Work

As part of the Risk Assessment step in the Risk Modeling Framework, models are reviewed and validated. Validation is conducted on several Quality Assurance (QA) and Quality Control (QC) levels. Two QA methods are employed. First, following good data science and software development practice, data scientists conduct code reviews on each other’s work. Second, model runs include test automation code that checks model outputs to catch erroneous values and quantify performance. The primary ignition probability model performance metrics are based on the accuracy of classification of ignition and non-ignition locations. The standard metrics of classification performance include recall, precision, visualization of the receiver operator curve, and the calculation of the area under that curve.

A number of QC steps are also employed both internal and external to PG&E. Within PG&E, the EORM team reviews the modeling methodology and results to provide feedback and signal their acceptance of the models for use in measuring risk.

Next, PG&E groups that use the risk models to develop mitigation work plans test the model with their subject matter expertise. The PG&E Internal Audit group also has conducted in depth reviews of model methods, results and the application in developing mitigation workplans. Finally, PG&E uses outside expertise to review and validate model methods, code and model results. PG&E is currently contracted Energy and Environmental Economics, Inc. to perform a review and validation of the modeling methodology, code, model results and application to be completed in the Spring 2021.

5.5.1 Performance metrics

Because MaxEnt predictions are based on smoothly varying ignition scores, we can choose a value of the scores that defines the threshold between low and high ignition probability. The lower the threshold, the more ignition locations are predicted (true positive) – the percentage of ignitions predicted is called the recall and the remaining percentage of un-predicted ignitions are called the omission rate. However, a lower threshold will also result is a greater proportion of non-ignition locations getting classified as at elevated risk (false positives). We seek a model that can achieve high recall while minimizing the number of associated false positives.

The receiver-operating curve (ROC) (Figure 14) is a curve with the true positive rate on the y-axis and the false positive rate on the x-axis for all possible omission rates. Each point along the curve represents the tradeoff between making the model omission rate more generous to predict more "true positives" (higher on the y-axis) vs. having that generous omission rate falsely predict ignitions that didn't occur (further right on the x-axis). Any given point along the ROC tells you what fraction of non-ignitions are falsely predicted as ignitions as the "cost" of achieving a given true positive rate for all true positive rates. Along the ROC curve then, predicting only non-ignitions is shown in the lower left corner to predicting only ignitions in the upper right corner. Random guessing will produce a diagonal ROC, whose area would be 0.5. A perfect model would produce an ROC that immediately rises to 100% true positive without any false positives, whose area would be 1. The AUC-ROC is the area under the ROC curve that ranges between 0.5 and 1 and captures how well the model avoid false positives as it captures true positives.

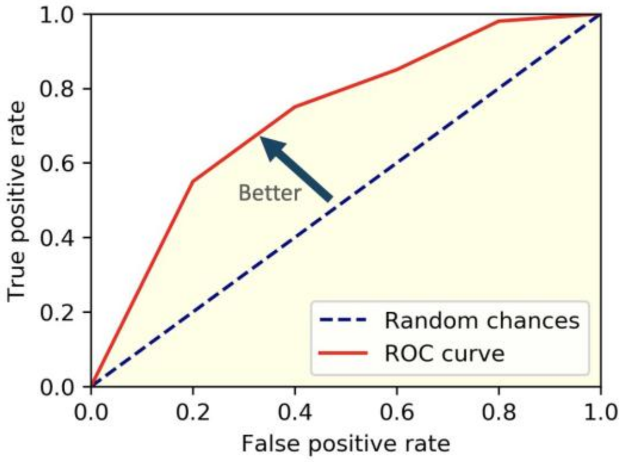


FIGURE 14 - ILLUSTRATION OF A RECEIVER OPERATOR CURVE (ROC) WITH THE ROC-AUC AREA UNDER THE CURVE SHADED

5.5.2 Contribution from each covariate

"Jackknifing" is a modeling term of art for the practice of evaluating a model with and without some element whose impact on results you want to quantify. In our case, we have jackknifed each covariate of our MaxEnt models to compare (a) model performance with all covariates except the one studied and (b) model performance with only the one studied to the performance of the full ensemble of covariates in the "official" model.

This is the best method for answering "how important is X to model performance?" where X is one of the covariates. For this work, the metric of model performance was the regularized training gain[5] - the higher the better.

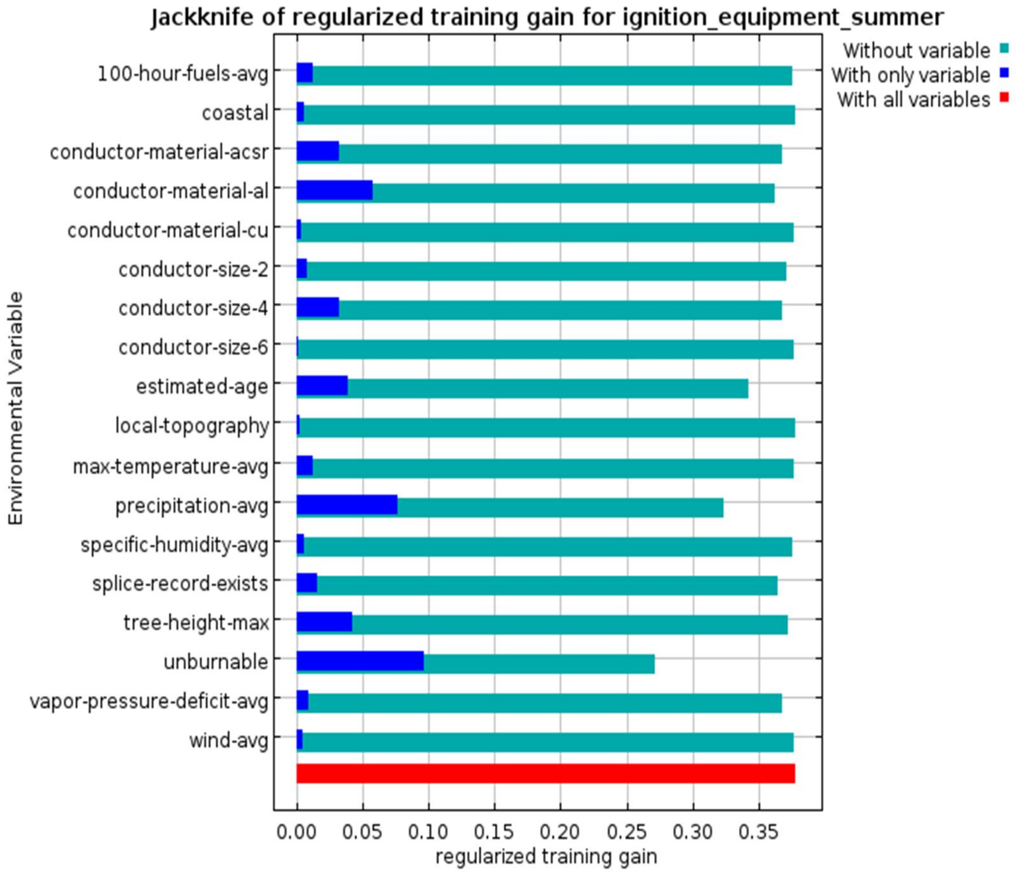


FIGURE 15: CONDUCTOR-INVOLVED IGNITION PROBABILITY MODEL JACKKNIFE RESULTS

[5] "The gain is defined as the average log probability of the presence samples, minus a constant that makes the uniform distribution have zero gain. At the end of the run, the gain indicates how closely the model is concentrated around the presence samples; for example, if the gain is 2, it means that the average likelihood of the presence samples is exp(2) ≈ 7.4 times higher than that of a random background pixel." - https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf

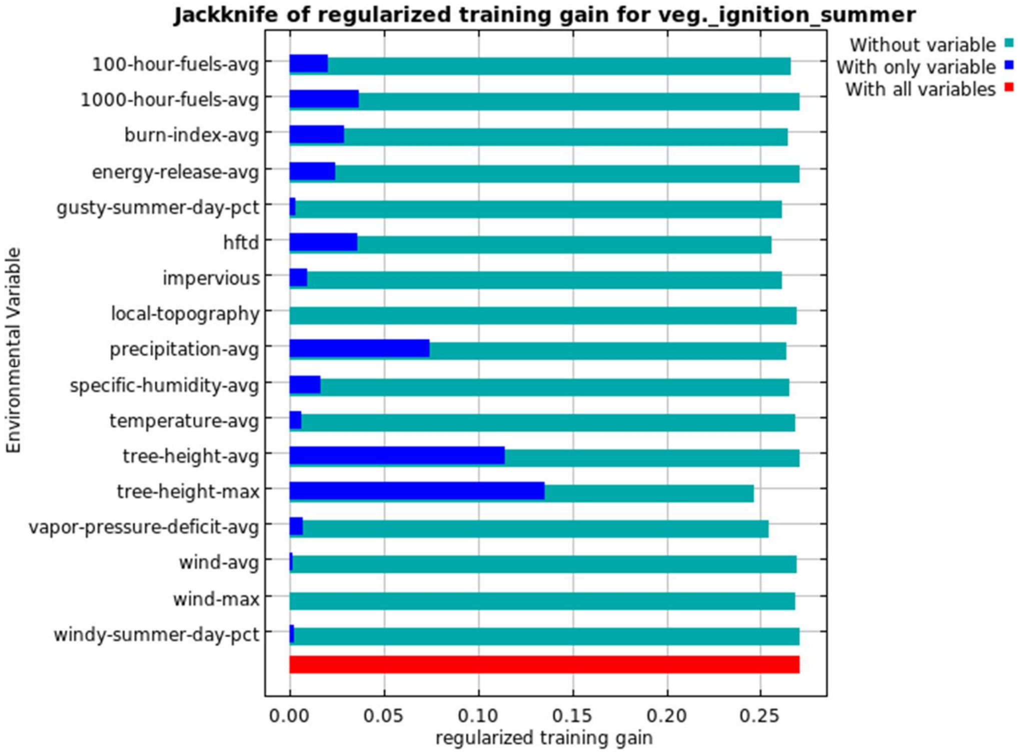


FIGURE 16: VEGETATION-CAUSED IGNITION PROBABILITY MODEL JACKKNIFE RESULTS

5.5.3 Comparison to previous work

The modeling that the 2021 Wildfire Risk Model replaced was produced to support planning for 2019, and is therefore referred to as the 2019 model. That model made estimates of ignition propensities that were not constructed as formal risk calculations. The 2021 model building upon and expands the earlier work with the following improvements:

- Risk score for measurement and prioritization – improved statistical and machine learning methods along with data sets that better describe factors attributed to ignitions and fires
- Wildfire consequence model that better predicts historical destructive wildfires
- Model granularity – base level is 100-meter pixel level that can be aggregated to circuit protection zones or circuits or higher
- Risk score is calibrated to the System MAVF risk scores developed in the RAMP Report

Although the 2019 model did not document a performance metric and appeared to rely purely on in-sample goodness of fit, assessment of true and false positives from that model were nevertheless still possible. The construction of an ROC curve and corresponding AUC measure was calculated using the actual 2019 ignitions compared to predictions. Similar curves and metrics were computed for 2019 ignitions using the 2021 model (recall that it was trained on 2015-2018 data). Comparing the 2019 on the left to the 2021 results on the right, model improvements have resulted in a marked improvement in predictive capability. Without a focus on predictive performance, the 2019 model barely out-performed random chance, whereas the 2021 model tuns in better performance by a wide margin. The interpretation of the ROC curves can be thought of in terms of the fraction of non-ignition locations you would need to harden (x-axis) to ensure that you harden some fraction of ignition locations (y-axis), thus avoiding them. The steeper the curve, the lower the overhead of work done that doesn't avoid ignitions. The circled areas of each figure demonstrate that following the 2019 model to address 50% of locations that will have ignitions would require also addressing roughly 50% of sites that would not have them, whereas the conductor-involved model could be used to address 50% of ignition locations while addressing just over 20% of non-ignition locations and the vegetation-caused ignition model could do the same with addressing around 30% of non-ignition locations. With billions of dollars being applied to hardening the grid, such performance can yield significant gains in the risk performance of dollars spent.

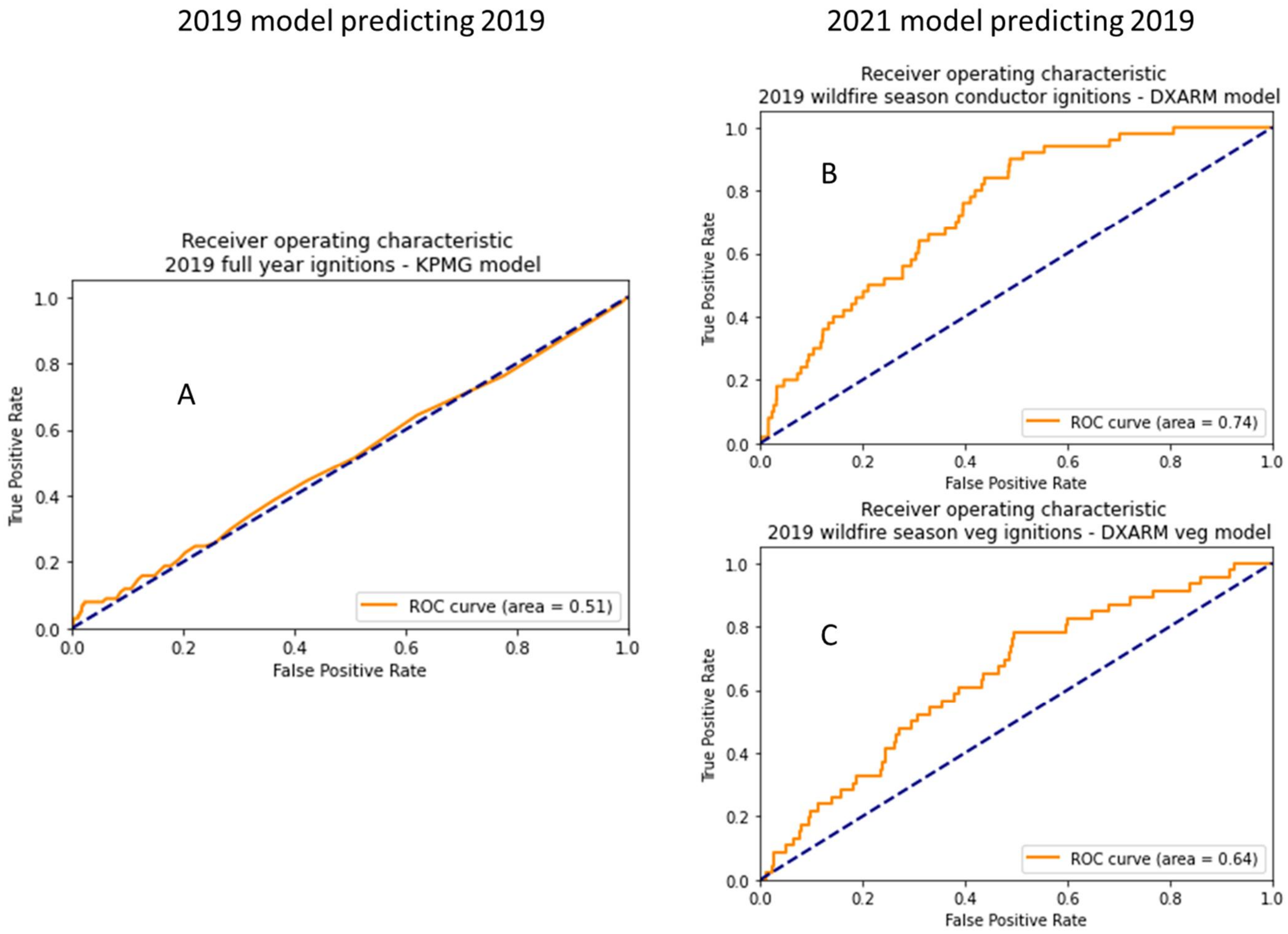


FIGURE 17: RECEIVER OPERATOR CURVES (ROC) FOR OUT-OF-SAMPLE PREDICTION OF 2019 IGNITION LOCATIONS, 2019 MODEL (PANEL A) VS 2021 CONDUCTOR-INVOLVED (PANEL B) AND VEGETATION-CAUSED (PANEL C)

5.5.4 Discussion: Overlapping Risk Between the Two Models

The Vegetation-caused Ignition Risk Model, with detailed documentation found under Appendix 1: Vegetation-caused Ignition Risk Model 2021 and the Conductor-involved Risk Model, with detailed documentation found under Appendix 2: Conductor-Involved Ignition Risk Model 2021 are similar in philosophy and implementation. It is natural to ask how risk results from each model could be combined to capture a more comprehensive profile of risk informed by both vegetation-caused and conductor-involved ignitions. However, it is important to bear in mind that the models were developed separately in the service of EVM planning efforts within Vegetation Management and Asset Hardening planning efforts within Asset Management and their training data sets of ignitions data overlap. The Vegetation-caused Ignition model was trained using ignition events that were known to be caused by vegetation. The Conductor-involved Ignition Risk Model was trained using ignition events that involved conductors. Therefore, the models share the set of ignitions that are both vegetation-caused and involve conductors. Indeed, vegetation-caused ignitions involving conductors - wires down in particular - are the vast majority of ignition events used to train both models. For this reason, simply adding the risk values from both models together without re-calibrating the result would double-count ignition risks (or mitigations applied to prevent them).

The problem of double counting is illustrated by the Venn diagram (Figure 18) depicting the ignition data used for the EMV Risk Model and the Conductor Risk Model. Vegetation-caused ignitions that impacted conductors are found in the inputs to both models, so the P(ignition) and E[ignition] (expected number of ignitions) values for each model cannot be cleanly "added" into a bigger picture view.

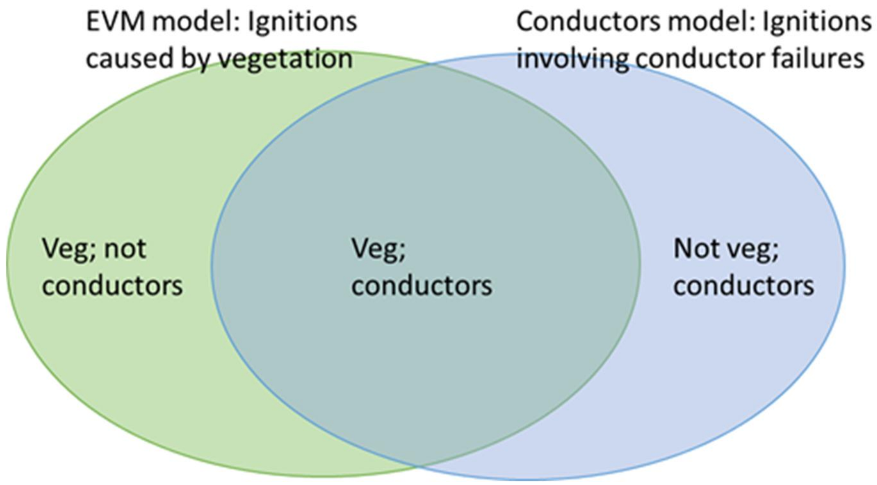

EVM model: Ignitions caused by vegetation
Conductors model: Ignitions involving conductor failures
Veg; not conductors
Veg; conductors
Not veg; conductors

FIGURE 18 - VENN DIAGRAM COMPARING EVM AND CONDUCTOR MODELS

The solution to the double-counting described above is to break the ignitions into the three categories revealed by the Venn diagram and model each separately, using the "sum" of those results to guide planning and risk mitigation across each category. Exactly that type of model is under development under the name "Composite Model". See the *Future Work* section for more details.

6 Maturity Survey Areas of Improvement

Key improvements with the 2021 Wildfire Risk Model have advanced capabilities in three CPUC Utility Wildfire Mitigation Maturity Survey areas and laid the groundwork for two others.

6.1 Key Improvements

Ignition Risk Estimation. Ignition probabilities are now based on a quantitative model for two of the failure modes identified in the RAMP wildfire risk bowtie. These ignition models are developed at the 100-meter pixel granularity that can be aggregated to circuit segment and circuit level views. As additional models are added to represent more failure modes the modeling of ignitions will be become more complete. Precision metrics to measure the predictive power of the model have also been developed to track this improvement over time.

Estimation of Wildfire Consequences on Communities. Wildfire consequence is now produced at the 100-meter pixel granularity based on acreage and structures. Based on an 8-hour Technosylva simulation, the consequence scoring is developed with fire spread acreage, structure, rate of spread and flame length.

Estimation of Wildfire and PSPS Risk Reduction Impact of Initiatives. The quantitative wildfire risk values produced with the ignition probability and wildfire consequence models now allow for calculation of the risk reduction provided by mitigation alternatives. Prior risk model risk scores only provided a relative ranking but lacked quantitative ability to measure risk. By applying both Subject Matter Expert and data informed effectiveness values for each mitigation alternative, risk reduction for mitigation alternatives can be estimated. As data is collected on the performance of wildfire mitigations, such as system hardening, risk reduction values will be determined based on quantitative models.

6.2 Future Improvements

Risk-Based Grid Hardening and Cost Efficiency. Building on the quantitative risk values, future model improvements will enable the development of RSE values for mitigation alternatives.

Portfolio-Wide Initiative Allocation Methodology. MAVF calibrated risk scores provide the framework for the development of risk scores representing the portfolio of risks. In the future, a grid location or circuit segment would have a wildfire risk score but also a public safety risk score or a reliability risk score. These values will enable a portfolio-wide allocation of risk mitigation plans tuned to most efficiently reduce risk.

7 Future Work

The risk modelling team has already initiated work on the 2022 Wildfire Risk Models. The key area for improvement in the 2022 model is to enable the combination and comparison of risk scores for individual risk drivers. For this reason, the 2022 model is often referred to as the Composite Model as it will assemble the vegetation and conductor equipment failure risk models into a risk model framework which will allow for a granular composite risk score. This risk score will allow for the identification of high-risk circuits segments but also insights in to whether vegetation or equipment risk is the predominant driver. Coupled with improved abilities to measure risk reduction for mitigation alternatives and the 2022 model will allow comparison of the effectiveness of vegetation mitigations with equipment mitigations.

Planned additional improvements include:

- additional equipment failures models for poles and pole mounted transformers
- inclusion of LiDAR data and tree species data
- inclusion of inspection and repairs tag data
- inclusion of PSPS consequence dimensions to assess holistic mitigation effectiveness
- re-introduction of a new egress model to complement the wildfire consequence data set

The aim with these improvements is to further mature PG&E’s wildfire risk modeling capabilities and the effectiveness of mitigation workplans. As these aims are realized, risk models will provide ever more actionable insights that will enable PG&E to more effectively target and deploy wildfire mitigations for the benefit of the State of California and its residents.

For the two models documented herein, the composite solution would be to model the three areas of the Venn diagram (see Figure 18 above) separately and composite their results as needed into representations suitable for EVM or asset hardening. Figure 19 below illustrates that separation.

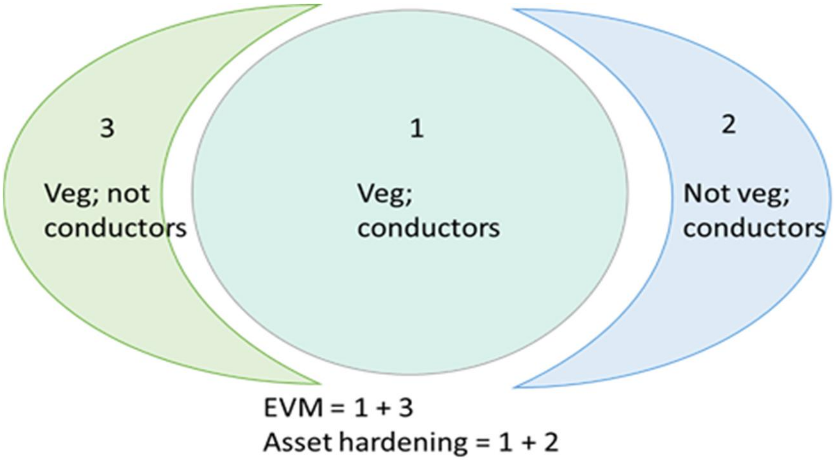


FIGURE 19 – COMPOSITE MODEL SUBSET CATEGORIES FOR VEGETATION-CAUSED AND CONDUCTOR-INVOLVED IGNITIONS

8 Project Team

PG&E's 2021 Wildfire Risk Model is a collaborative effort bringing together internal PG&E risk teams and external experts in risk and data science lead by the Risk and Data Analytics team within PG&E's Electric Operations Asset Management group. The team consists of the following external and internal groups and individuals.

Convergence Data Analytics (lead contractor) – Led by Sam Borgeson, PhD, UC Berkeley. Convergence Data Analytics was founded out of the Stanford Sustainable Systems Lab and specializes in bridging data science and analytics insights from academia into the utility industry. Sam has a background in software development, consulting, and data-driven modeling and has provided technical leadership and oversight for the model development. Their company website is: http://www.convergenceda.com/

Individual consultant members of the CDA team with expertise in asset failure modeling, meteorological and ecosystem data preparation, event classification and arrival process modeling and data engineering.

Laurel Dunn - PhD, UC Berkeley, is a Data Scientist and Engineering specializing in Electric Power Systems and risk and decision analysis.

Eric Waller – PhD, UC Berkeley, is an expert on plant distribution modeling with specialization in Plant Biogeography and Fire Ecology.

Philip Price, PhD, University of Kentucky and former staff scientist at Lawrence Berkeley National Lab, is an expert in statistical modeling including Bayesian modeling, statistical sampling and validation methods.

Salo Sciences – Chris Anderson, PhD Stanford, and David Marvin, PhD Michigan, and their team bring expertise in remote sensing of forests, application of machine learning to predicting utility-caused ignition. His team provided the early technical implementation of the Maximum Entropy algorithm. Dr. Marvin leads model development for the California Forest Observatory. Their company website is: https://salo.ai

Presence Product Group – Cooper Marcus, MBA Kellogg School of Management, is a Product Manager with expertise in software project management and execution. He provides project management, development of custom tooling for data engineering and data science, and technical support for the project's team of data scientists and analysts. https://presencepg.com/

Duffy Gillman – Information Sciences, Software Architect

Berty Pribilovics – Full stack Software Engineer

PG&E Strategic Data Science Team – Led by JP Dolphin, Director and the following team members:

Doron Bergman, PhD, UC Santa Barbara – Expert Data Scientist

Carolyn Meldgin, PhD, University of Illinois at Urbana-Champaign – Expert Data Scientist

PG&E Risk Management and Safety Team – Led by Benson Wong, Senior Manager and the following team members:

Srini Vanukuri – Expert Data Scientist

PG&E Risk and Data Analytics Team – Led by Jon Eric Thalman, PE, Senior Manager and the following team members:

Andrea Brown, PE – Senior Data Scientist

Manuj Sharma – Principal Product Manager

Carlos Rengifo – Expert Data Analyst

Appendix 1: Vegetation-caused Ignition Risk Model 2021

9 Executive Summary

The PG&E Risk and Data Analytics (RaDA) team developed this Vegetation Risk Model to estimate the risk posed by fire season (Jun. 1 – Nov. 30) ignitions caused by vegetation interacting with overhead conductor segments of the distribution grid within the high fire threat district (HFTD) Tier 2 & 3 areas. The model predictions inform the Vegetation Management team of the locations to be at elevated risk from vegetation-caused ignitions in a typical, or planning, year.

The Vegetation Risk Model (also referred to simply as "the model" in this document) is based on the definition of Risk = (probability of ignition) x (consequence of ignition) at all locations along the grid. More specifically, it is based on the conditional probability that a vegetation-caused CPUC-reportable ignition will occur within a fire season period within a given 100m by 100m pixel location containing overhead conductors in HFTD tiers 2 and 3 multiplied by a spatial rendering of EORM's MAVF CoRE values from tranches related to wildfire. The ignition probabilities were estimated using a maximum entropy algorithm (described in detail in Appendix 3: Ignition Probabilities Methods 2021) and the consequence values were derived using a spatially gridded set of fire simulations to determine expected fire severity (described in detail in Appendix 4: Ignition Consequence Methods 2021). Model features (also referred to as variables or covariates) include high-spatial-resolution environmental and meteorological data. The key features driving predicted probabilities and risk are tree height, vegetation dryness, and atmospheric dryness.

As is typical of risk models, risk is calculated by multiplying probability by consequence. To model the probability of ignition and the consequence separately, both calculations look at a variety of environmental conditions and asset data that may or may not influence these two quantities. A data driven approach can expose which environmental conditions have actual influence. The conditions under which consequence is high need not be the same for which ignition probability is high. This applies to both time and location considerations.

Risk results were calculated at each 100m by 100m pixel and aggregated to larger segments of the distribution system, called circuit protection zones (CPZs). A spreadsheet of results for each CPZ, and geospatial files of the results, were developed and provided to the Vegetation Management (VM) team to enable the use of probability, consequence, and risk results to inform the planning of EVM mitigations that are intended to reduce the risk of vegetation-caused wildfires.

The conditions for ignition spread, including hot dry weather, small dry fuels mixed in with more substantial fuels to "ladder into[6]", and significant winds are partially at odds with the conditions found within forested areas, which tend to be more moist, lower temperature, and less windy than more open areas. This tension manifests in the results, were the densest stands of tall trees tend to feature elevated ignition probabilities, but do not feature the most extreme fire spread and consequence values. More open areas with mixed vegetation types tend to be the locations with significant ignition probabilities and significant consequences due to spread.

In a nutshell and very broadly speaking, ignitions near but not deep within forested areas are found to be the "riskiest" However, the consequence data tends to dominate over the ignition probabilities in the risk value calculations (risk correlates more strongly with consequence), as consequence values range over more orders of magnitude than the probability of ignition values. If you are prioritizing directly by risk, you are largely prioritizing by consequence, suggesting that a region with attributes that have high potential for catastrophic results (i.e. large or intense spread after an ignition) is sufficient cause for prioritization of work.

[6] In fire science ladder refers to the ability of a fire to progress from one fuel source to another

Another potentially counter-intuitive finding is that prevailing wind speeds are not found to be strongly predictive of annualized ignition probabilities. Our ignition probability model pools data from several fire seasons into a single ignitions data set to discern the difference in environmental conditions and asset characteristics associated with ignition locations. Due to these data requirements, the environmental conditions are shared by all ignitions – they are not drawn from the weather at the time of the ignitions, but rather from the typical climate in the location of the ignitions. The variability they capture is spatial, not temporal. With that context, we observe that winds tend to be stronger near the coast, where other contributors to fire risk, like heat and dryness, tend to be muted. We also observe that vegetation in windy areas tends to be shorter and stouter or at least more robust to wind. Finally, we note that forested areas are a form of "surface roughness" that disrupts and slows winds. In other words, tall/dense stands of trees tend to be negatively correlated with wind.

The primary method used to evaluate the predictive algorithm's goodness-of-fit was the receiver operator curve – the area under the curve (ROC-AUC) metric. The ROC curve measures the ratio of true positives rate to the false positive rate. The area under the ROC curve of 0.5 represents random chance and 1.0 represents perfect prediction. The full ignition dataset used to train the model achieved an AUC score of 0.737. Using a random split of the ignition dataset into separate train (75%) and test (25%) subsets (selected at random from all ignition inputs), the model achieved an AUC of 0.727 when evaluated against the training data, and an AUC of 0.716 when evaluated against the testing data, close enough to affirm good out-of-sample predictive power. Yet another out-of-sample evaluation, of the model trained on the entirety of the 2015-2018 dataset of ignitions, against some ignition data from 2019, yielded an AUC of 0.64.

10 Introduction

10.1 Background

Wildfires are an increasing problem in California with larger fires occurring more frequently (Figure 20). Before PG&E began conducting Public Safety Power Shutoff (PSPS) events to prevent vegetation from causing ignitions that could lead to a catastrophic wildfire, there were 476 vegetation-caused ignitions from 2015 through 2018 within our service territory. 405 of those occurred during the June through November period commonly recognized as the period most likely for ignitions to grow into wildfires and 222 occurred within HFTDs, nearly all of which damaged conductors. Because vegetation caused ignitions have historically been a significant contributor to ignitions within PG&E territory, PG&E's current commitment to reduce the likelihood (probability) and consequence of catastrophic wildfires makes it imperative that we understand the risks associated with vegetation-caused ignitions and use that information to inform the prioritization of mitigation through EVM.

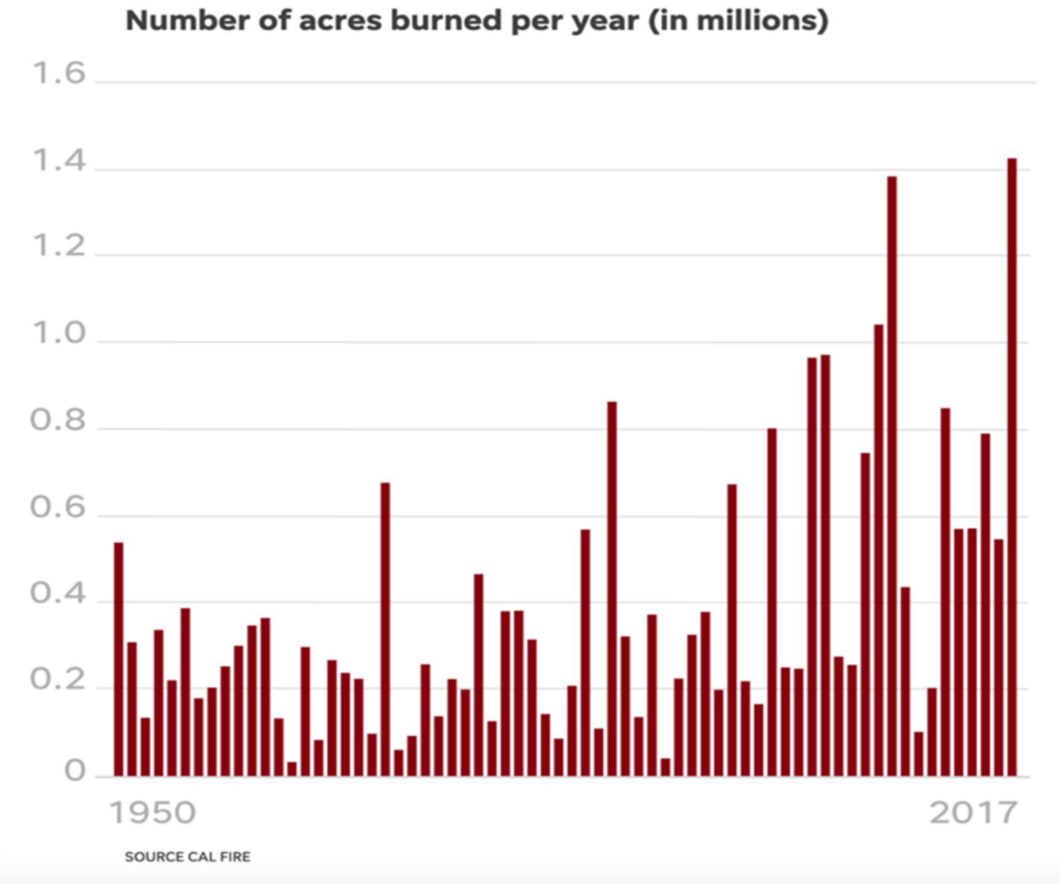


FIGURE 20 - TOTAL ACRES BURNED IN CALIFORNIA PER YEAR FROM 1950 TO 2017

PG&E conducts a vegetation management program to manage the risk of nearby vegetation contributing to electric grid-caused catastrophic wildfires within the service territory. For example, routine vegetation management work requires clearance of 4 feet around power lines in high fire-threat areas, with a recommended minimum clearance of 12 feet or more at time of prune to ensure compliance year-round. Hazardous vegetation such as dead or dying trees that pose a potential risk to the lines are also removed.

In response to the increasing wildfire threat the state of California faces, PG&E started an Enhanced Vegetation Management (EVM) program in 2019 that applies more stringent criteria. For example, overhanging limbs and branches directly above the lines are removed. Additionally, the condition of all trees tall enough to strike the lines, is also evaluated for is assessed to determine which require mitigation work[7].

10.2 Project

PG&E's system has approximately 25,000 miles of electric distribution grid lines in the high fire threat district (HFTD). The EVM program can work approximately 1,800 miles of distribution grid lines per year (based on approximate mileage worked

[7] see https://www.pge.com/en_US/safety/emergency-preparedness/natural-disaster/wildfires/vegetation-management.page

in 2019). To enable PG&E to not only reduce wildfire risk, but to reduce it quickly, the EVM program needs to differentiate between locations with comparatively higher and lower risk within HFTD areas. To this end, PG&E created a vegetation-caused ignition risk model to support the EVM team's process of risk-based planning and prioritization of their work.

As is typical is risk modeling projects, risk equals the probability of an event times the consequence of that event. The need to differentiate risk by location necessitates a modeling approach that captures how local conditions determine the probability of and consequences from an event.

Ignition events were modeled, and specifically, those ignitions that met the CPUC reporting requirement. A model of consequences is employed, described in Appendix 4: Ignition Consequence Methods 2021and focus here on describing a model for the probability of a vegetation-caused ignition.

A maximum entropy (MaxEnt) model is used for the probability, described in Appendix 3: Ignition Probabilities Methods 2021 and combine it with the consequence. This document also includes a discussion of the resultant risk profile from vegetation, and the manner in which it can inform EVM work.

As stated above, the model described in this document was built to support planning across one to many years of EVM effort. Our results are expected risk per-fire-season, inclusive of all weather conditions experienced during each fire season.

For risk calculations, the model first makes localized predictions of fire-season ignition probabilities, given environmental conditions and asset attributes. We make these estimates using a spatial model that works with 100m x 100m "grid pixels". Then the ignition probabilities are multiplied through by fire consequence data[8], based on worst-fire-weather simulation outcomes for the same 100m pixel locations. Thus, the model computes risk-per-pixel values across all modeled grid pixels - every 100m square containing distribution grid infrastructure within the HFTDs in this case.

The model then assigns the computed "pixel risk" across all trees in the vegetation management database (VMD) found within each pixel. The risk associated with each tree can then be aggregated to provide risk data at the circuit protection zone (CPZ) level.

10.3 Document Usage

This appendix is intended for both technical and non-technical audiences to explain the purpose of the model, summarize the methodology used to develop the model, and highlight the performance of the model.

Additional sections are referenced, and should be consulted to gain a full understanding of the model and process, the context in which this work was performed, and closely related work – these include:

- Section 2– Summarizes the regulatory, management, and analytical context in which the work described in these documents was performed, introduces the team and related work.
- Appendix 3: Ignition Probabilities Methods 2021 - Details MaxEnt modeling and Circuit Protection Zone (CPZ) aggregation.
- Appendix 4: Ignition Consequence Methods 2021– Details modeling of wildfire consequence.
    - Spatial Wildfire Consequence 2021 - Lunch n' Learn presented 2020_10_16 (PG&E EORM, 2020) – PG&E internal presentation.
    - Spatial Wildfire Consequence 2021 - Lunch n' Learn presented 2020_10_16 (PG&E EORM, 2020) – PG&E internal presentation.

[8] Spatial fire consequence data is based on MAVF framework tranche categories and values mapped to specific locations using historical data on Red Flag Warning Locations and fire spread simulations at points separated by 200m all along the distribution grid in HFTD tiers 2 and 3 – see "Methods – Spatial Consequence 2021" for details.

- EVM Risk Model 2021 – Lunch n' Learn presented 2020_10_21 (PG&E Risk and Data Analytics, 2020) – PG&E internal presentation related to this document.
- EVM Risk Model 2021 – Utility Analytics conference presented 2020_10_29 (PG&E Risk and Data Analytics, 2020) – Utility Analytics conference presentation related to this document.
- Appendix 2: Conductor-Involved Ignition Risk Model 2021 – Details a model similar to that described herein, but focused on wildfire risk related to conductor failure.
    - Conductor Risk Model 2021 – Lunch n' Learn presented 2020_10_28 (PG&E Risk and Data Analytics, 2020) – PG&E internal presentation

10.4 Applications

The predicted risk values for each CPZ can be used to inform EVM planning decisions to reduce wildfire risk. The VM team can use the risk predictions to identify those CPZs with elevated risk of an annual ignition event involving vegetation when developing wildfire mitigation workplans.

This tool is also intended to provide a method for quantifying the risk reduction achieved through planned and executed EVM work.

11 Vegetation Ignition Probability Model

This section summarizes the most important characteristics of the vegetation-caused ignition risk modeling performed by highlighting data and methods adopted through an iterative and consultative process of engagement between RaDA and Vegetation Management staff.

11.1 Modeling Framework

The Vegetation Model is a classification model that identifies the likelihood of at least one ignition per year at a given 100m by 100m pixel location. More specifically, the Vegetation Model models the conditional probability that a reportable ignition will occur within a given year within a given pixel location.

The model was fit using a presence-only maximum entropy (MaxEnt) algorithm. The algorithm is based on the assumption that the most unique characteristics of locations that have experienced ignitions in the past will predict the locations that experience ignitions in the future. This algorithm is explained in more detail in Section *21.2*.

The model was trained using reportable ignitions, as defined by the California Public Utilities Commission (CPUC) and outlined in Section *11.2*. The ignitions were filtered for ignition events associated with an vegetation interacting with grid assets, as identified during the follow-up investigation. The ignitions were also filtered by date, using only ignition events that occurred during the fire season (Jun. 1 – Nov. 30). Ignitions data ranged from 2015 to 2019, and 2019 ignitions data were withheld from the model training dataset to use as a test dataset.

11.2 CPUC-Reportable Ignitions

CPUC reporting requirements are limited to reportable fire events that meet the following criteria:

- A self-propagating fire of material other than electrical and/or communication facilities,
- The resulting fire traveled greater than one linear meter from the ignition point, and
- The utility has knowledge that the fire occurred

Fires that caused damage to utility facilities and whose ignition is not associated with utility facilities are excluded from this reporting requirement (CPUC, 2014)

11.3 Model Assumptions

Risk that can be addressed by the EVM program (computed per-pixel) is assumed to exist only in those pixels where the vegetation management database (VMD) has recorded trees. To be clear, the probability and consequence components of risk are computed for all grid pixels within the HFTDs but can only be assigned to trees that are known to the VM team.

- Some fraction of risk is assumed non-addressable through EVM. This could be due to imperfect information (trees that look healthy but aren't), practical constraints (landowners refusing to allow tree work), fluke events (trees not deemed close enough or tall enough to hit the lines contacting them anyway during storm events - we hear reports of branches blown uphill for example), or future growth that could lead to impact over the risk planning horizon.
- Tree work is assumed to change the probability of an ignition (by reducing the likelihood that the tree contacts and harms a Distribution line), but not change the consequence of ignitions, so risk reduction from EVM can be expressed in terms of a lower P(ignition).

11.4 Model Features

Model covariates includes to following, which are described in detail in the “Pool of Covariates” section of Appendix 3: Ignition Probabilities Methods 2021:

- local-topography – whether the local areas is uphill or downhill from its surroundings
- precipitation-avg – annual average precipitation
- specific-humidity-avg – annual average humidity
- temperature-avg – annual average temperature
- tree-height-avg – average of remote sensed tree height within each 100m x 100m pixel
- tree-height-max - maximum of remote-sensed tree height within each 100m x 100m pixel
- vapor-pressure-deficit-avg – annual average VPD, where VPD is a metric of dryness
- gusty-summer-day-pct – the percentage of all summer days with hourly gusts over 20 mph
- windy-summer-day-pct – the percentage of all summer days with sustained hourly wind speeds over 15 mph
- wind-avg - hourly average wind speed at 10m, averaged from 2016 to 2018
- wind-max - annual 99th percentile hourly wind speed at 10m, assessed over 2016 to 2018
- 100-hour-fuels-avg - the GRIDMET variable use is known as fm-100, and is a standard fire modeling metric of fuel dryness for fuels about 1-3" in diameter - intermediate sized fuels, averaged for 2014-2016
- 1000-hour-fuels-avg – GRIDMET fm-1000, as defined above, but for 3-8" in diameter, averaged for 2014-2016
- burn-index-avg - the US, the National Fire Danger Rating System (USNFDRS) Burning Index (BI), averaged for 2014-2016
- energy-release-avg - USNFDRS Energy Release Component (ERC), averaged for 2014-2016
- hftd – categorical variable identifying tiers 2 and 3
- impervious - NLCD imperviousness products represent urban impervious surfaces as a percentage of developed surface over every 30-meter pixel in the United States, scaled to 100m
- Outage data is drawn from ILIS and vegetation-caused outage reports; ignition data is drawn from 2015-2018 CPUC reportable ignitions. 2019 data used for out of sample performance testing.
- Only ILIS entries and ignitions flagged as “vegetation caused” are used; and
- Only fire season (Jun. 1 – Nov. 30)outages and ignitions are used.
- Weather and environmental covariates are drawn from public data sources, like gridMET, RTMA, the National Elevation Database - NED, the National Land Cover Database – NLCD.
- Tree cover and height drawn from Salo Science proprietary data derived from satellite imagery using computer vision algorithms.

- Vegetation-caused ignitions from 2015-2018 within HFTDs used as “presence data” inputs for MatEnt model.
- Rasterized weather, environmental, and tree data for just HFTD grid pixels used as covariates for MaxEnt model.
- Model outputs calibrated via tau calibration (see Methods – Ignition Probability Modeling 2021) into probability of ignition for every HFTD grid pixel.
- Ignition probability pixel values paired with corresponding pixel values of spatial wildfire consequence and multiplied to produce per-pixel risk.
- Risk values mapped spatially to trees in the VMD.
- Risk values and supporting metadata rolled-up to CPZ-level summaries, delivered as official model results.

11.4.1 EVM 2019 Trees

A list of trees worked in 2019 by the EVM program from the Oracle EVMGIS database is utilized as a reference scenario of EVM work, but not as a source of tree locations (see VMD 2019 candidate trees for that source). The tree data was filtered as follows:

- The tree work date (TW_WORK_DATE) was in 2019;
- The tree work was completed; and
- The tree diameter at breast height was at least 4 inches.
- The tree data is not from LiDAR, but rather inspections

The tree diameter at breast height criterion was provided by VM staff as a criterion to distinguish between tree work where branches or the entire tree is removed, versus brush clearing. Brush is lower vegetation that does not pose a risk for interacting with live wire or causing a pole to fall, but rather pose a danger by providing ignitable fuel for a fire.

This data represents the entirety of the trees that have undergone EVM work in 2019. There are 133,666 tree records in our data extract - all of them record latitude/longitude coordinates. VM confirms that the number of records is a good approximation of the number of trees worked per year.

11.4.2 VMD 2019 candidate trees

The data set is the source of tree presence/absence and species used for the Vegetation Risk modeling work. The risks are calculated for all grid pixels in the HFTDs but are applied to the EVM 2019 trees and summarized per-CPZ for the purposes of quantifying “work” per CPZ for prioritization and planning.

Vegetation work at PG&E is informed by inspections that cover the entirety of every circuit on the grid every year. The inspections are recorded in the VMD SQL-SERVER database. Each inspection records a number of trees, some tagged as requiring work, others tagged as not.

We take the set of trees tagged in the VMD as requiring work during 2019 to represent the set of trees that are candidates for having EVM work performed on them (at the time of the modeling effort, 2020 inspections were not completed).

Much like the EVM data pull, we consider only trees where the diameter of the tree at breast height (DBH) is at least 4 inches. This is done to ignore brush clearing instances. Further, we consider only maintenance inspections since we were informed that they are the only inspections that are comprehensive. These considerations are made to cover the entire tree population presenting risks to the grid.

Tree inspections can happen several times a year in the same location. We pick the first inspection of 2019 for each location to get the set of trees closest to what it was at the start of 2019.

- 3.9M VMD 2019 candidate trees, aka trees inspected in 2019

- 2.3M of those are within HFTD tiers 2 and 3
- 804k of those are trees considered EVM regional priority species

11.5 Model Evaluation

The primary method used to evaluate the model's goodness-of-fit, and therefore the suitability for using as a prediction tool, was the ROC-AUC (Receiver-Operator Curve – Area Under the Curve) metric.

The receiver-operator curve based on the prediction of the training data for the model is depicted in Figure 21 below. A few performance metrics like the area under the ROC curve, precision and recall are listed beneath it.

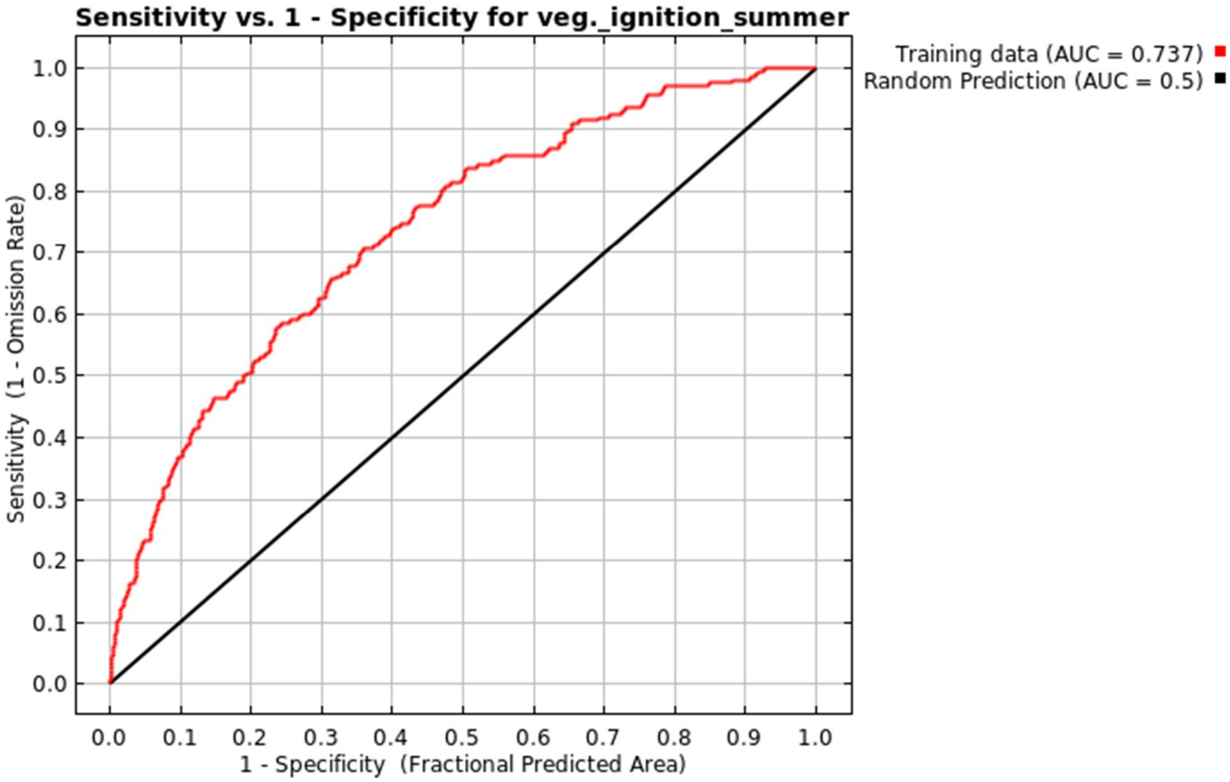


FIGURE 21 - ROC CURVE FOR THE IGNITION PROBABILITY MODEL RESULTS. DEPICTS THE TRADEOFF BETWEEN "TRUE POSITIVE" FRACTION ON THE Y-AXIS (THE FRACTION OF ALL IGNITIONS IN PREDICTED BY THE MODEL) VS. THE FALSE POSITIVE FRACTION (THE FRACTION OF NON-IGNITIONS PRECITED

ROC-AUC: 0.737 (the area under the ROC curve 0.5 is random guessing; 1.0 is perfect prediction)

precision: 0.00038 (the fraction of all predicted high ignition risk locations at 0.95 recall that experienced an ignition)

recall: 0.960 (the fraction of all ignition identified by the model when using an omission rate of 0.95)

Note that the threshold where the logistic is interpreted as a 1 is theoretically tuned to recall 0.95 of the events. Thus, with this model's recall, we are looking for a number close to 0.95 to confirm that the expected behavior is seen empirically.

A separate run that trained on 75% of the ignitions data, selected at random and tested on the other 25% achieved an in-sample AUC of 0.727 and test sample AUC of 0.716, so we do not believe that the model is significantly over-fitting the training data and is suitable for use in a predictive setting.

11.5.1 Permutation Importance (of Covariates)

The permutation importance (sensitivity of the model to the values in the underlying training data) is defined to be *the decrease in a model score (a chosen performance metric) when a single feature value is randomly increased or decreased*. The contribution for each variable is determined by randomly permuting the values of each variable among the training points (both presence and background) and measuring the resulting decrease in training gain – a metric of model performance. This technique benefits from being model-agnostic and can be calculated many times with different permutations of the feature. We can use permutation importance to identify which of the variables studied has the greatest predictive effect upon ignition probability. As seen below, tree-height-max, 100-hour-fuels-avg, and vapor-pressure-deficit-avg are the covariates whose values the model predictions are most sensitive to. In other words, the model finds the presence of tall trees in dry areas to be most predictive of vegetation-caused ignitions.

The complete description of all input variables is described in the "Pool of covariates" section in Appendix 3: Ignition Probabilities Methods 2021. The permutation importance of each model parameter is displayed below:

TABLE 4 - PERMUTATION IMPORTANCE OF MODEL FEATURES

Rank	Model Feature	Permutation Importance
1	tree-height-max	26.1
2	100-hour-fuels-avg	24.1
3	vapor-pressure-deficit-avg	21.6
4	gusty-summer-day-pct	6
5	hftd	4.2
6	precipitation-avg	3.1
7	impervious	2.8
8	specific-humidity-avg	2.4
9	burn-index-avg	2.3
10	wind-max	1.9
11	temperature-avg	1.6
12	windy-summer-day-pct	1
13	local-topography	0.8
14	tree-height-avg	0.8
15	1000-hour-fuels-avg	0.6
16	energy-release-avg	0.4
17	wind-avg	0.2

11.5.2 Jackknifed Model Results

A model's "jackknifed" performance is computed by systematically excluding each covariate (one at a time) and re-running the model to determine the decrease in model performance absent the missing covariate - the greater the decline in performance, the more important that covariate is to the overall model.

This technique can also establish how well each covariate explains the results (i.e. vegetation caused ignitions) when it is the *only covariate in the model*. Figure 22 below depicts the jackknifed performance (training gain is a metric of model performance – the higher the better) of the ignition probability model, with the turquoise bars representing the model performance when each covariate is left out and the dark blue bars representing the model performance when each covariate is the only explanatory variable. Both of these categories can be compared to the model performance of the full set of covariates at the bottom in red. From this we can see that "tree-height-max" is the best single covariate as well as the one whose absence most degrades the model performance. With all other variables intact, the loss of vapor pressure deficit degrades model performance by the second greatest amount, despite not performing that well by itself. One can also verify that HFTD and impervious ground cover (i.e. non-flammable) are particularly important to the model's final performance.

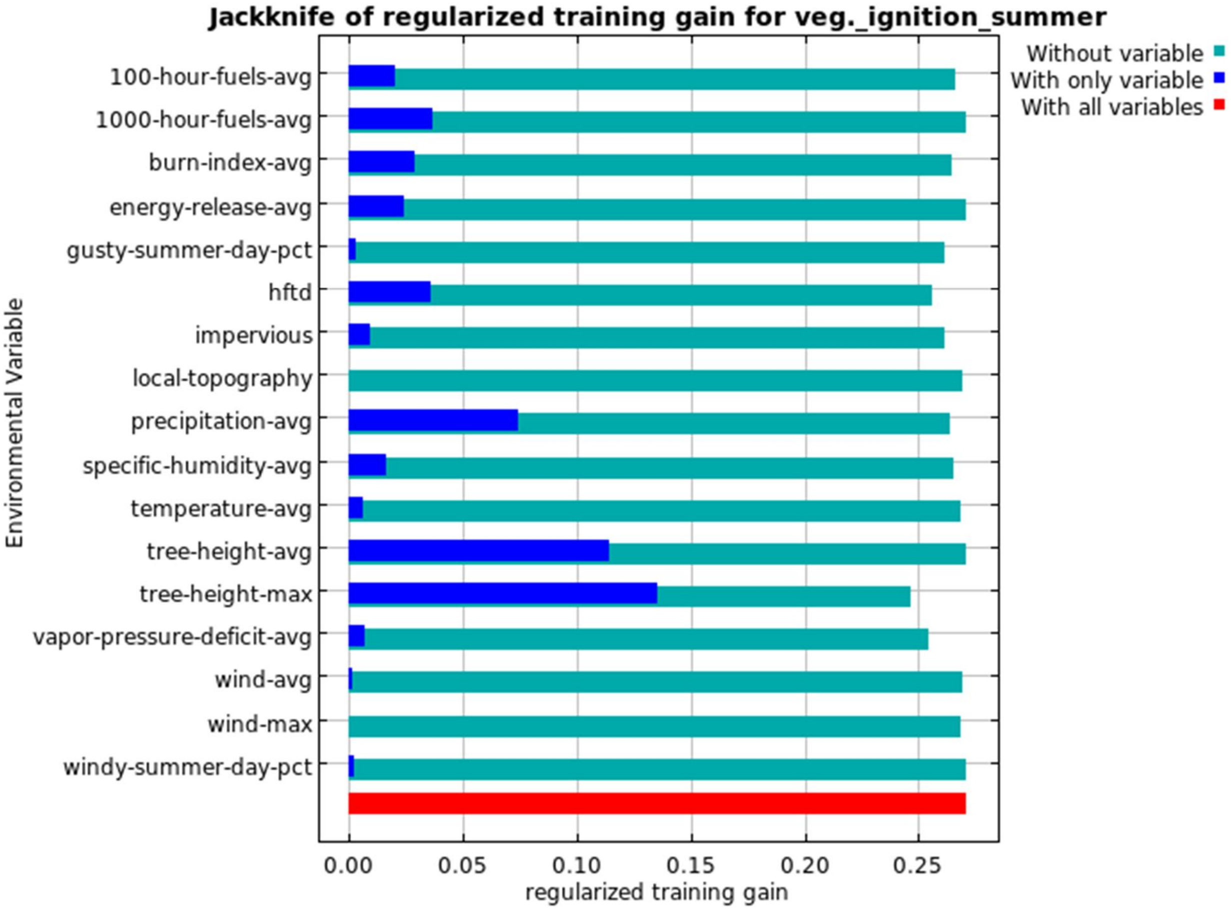


FIGURE 22 - JACKKNIFE OF REGULARIZED TRAINING GAIN FOR VEGETATION IGNITION MODEL

12 Methodology

For risk determination purposes, we identify probabilities of ignition for each 100m x 100m grid pixel in the HFTDs and multiply them by spatially resolved consequence data for each pixel to determine pixel-specific risk, then aggregate risk to CPZs to determine the CPZ- specific ignition risks across the HFTD area.

12.1 Circuit Protection Zones (or Circuit Segments)

Circuit Protection Zones (CPZ) were selected as the appropriate segmentation of the grid to report risk results because they are the most granular scale at which outages are reliably captured by the system protective devices – and outages are an important factor for model training. Furthermore, the predecessor 2018 model utilized CPZs as the smallest aggregation of risk, so for comparison purposes, the 2021 model utilized the same approach of aggregating to CPZs.

For the Vegetation Risk Model 2021 documented herein, a mid-2018 CPZ vintage was requested by stakeholders. This vintage selected facilitated comparison to the predecessor 2018 model but made more challenging comparison to the Conductor Risk Model 2021, whose stakeholders selected a more modern CPZ vintage.

More information about CPZs, their vintages, and their challenges and limitations is in Section 31.

12.2 Estimating Probability(ignition) using MaxEnt

Our goal is to produce high spatial resolution predictions of risk. To achieve this, we employ a Maximum Entropy (MaxEnt) modeling approach to train models to predict the probability of ignitions over the time frame of a single wildfire season in 100m x 100m pixels on a map grid.

Our model is an instance of supervised machine learning and therefore requires learning from examples of actual events that occurred. We train our model on 4 wildfire seasons of ignitions, 2015 through 2018 inclusive, these are the first 4 years for which reportable ignitions were tracked and this approach saves the 2019 data for out of sample validation. Wildfire season is defined as Jun. 1 – Nov. 30, inclusive. Incidents accumulate over time, but what matters for risk assessment is the rate at which veg-caused failures (ignitions caused by vegetation) occur, so our model is calibrated to predict annual counts of fire-season veg ignition events.

MaxEnt modeling is useful for our modeling purposes since it assigns similar probabilities of events to different locations experiencing similar conditions. To predict veg-caused failures in the distribution grid, we make the reasonable assumption that such failures likely occur in locations with conditions that are similar to those where past failures occurred.

Comprehensive details on the modeling approach (but not the exact set of covariates we used, which are documented herein) can be found in Appendix 3: Ignition Probabilities Methods 2021.

12.3 Consequence

Wildfire consequence estimates the resulting damage if an ignition event occurs at a specific location. For this model, the multi-attribute value function (MAVF) consequence of risk event (CoRE) dataset was used, a consequence dataset provided by the Enterprise and Operational Risk Management (EORM) team at PG&E that combines safety, financial, and reliability types of damages. More information about MAVF CoRE consequence is included in Appendix 4: Ignition Consequence Methods 2021.

12.4 Risk

As is typical in risk modeling, we define risk as:

*Risk(event) = P(event) * C(event)*

where P(event) is the probability of the event occurring and C(event) is the consequence of the event occurring.

Since we are working with event probabilities derived from MaxEnt models for every 100m pixel along the grid, we use consequence data with the same spatial resolution to produce “risk per-pixel” values.

Then we can produce an aggregate risk across a given Circuit Protection Zone (CPZ) by summing or averaging the risk values across the pixels within that CPZ (see Methods – Ignition Probability Modeling 2021 (PG&E Risk and Data Analytics, 2020) for more details on CPZs).

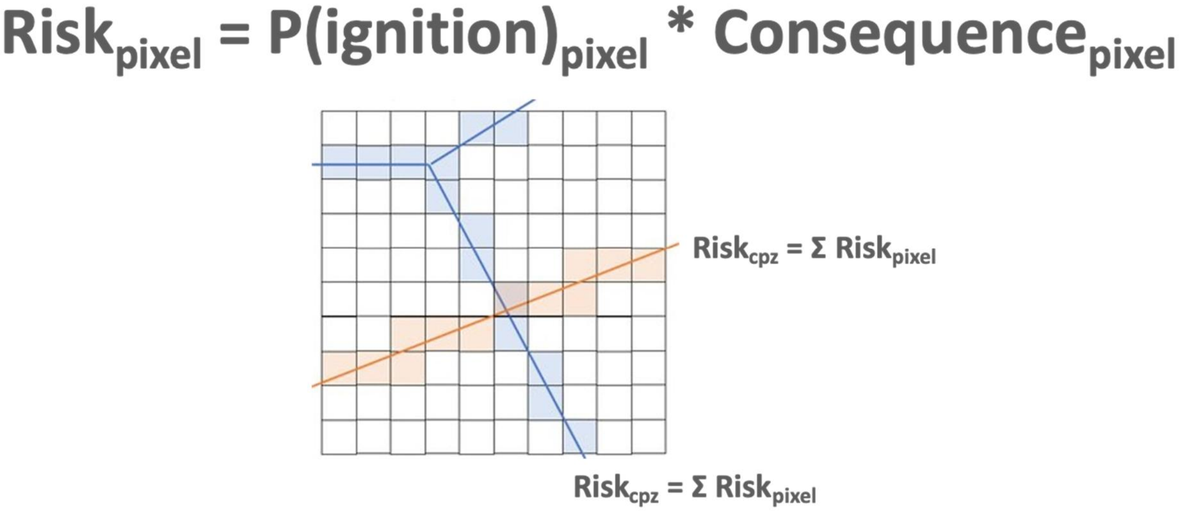
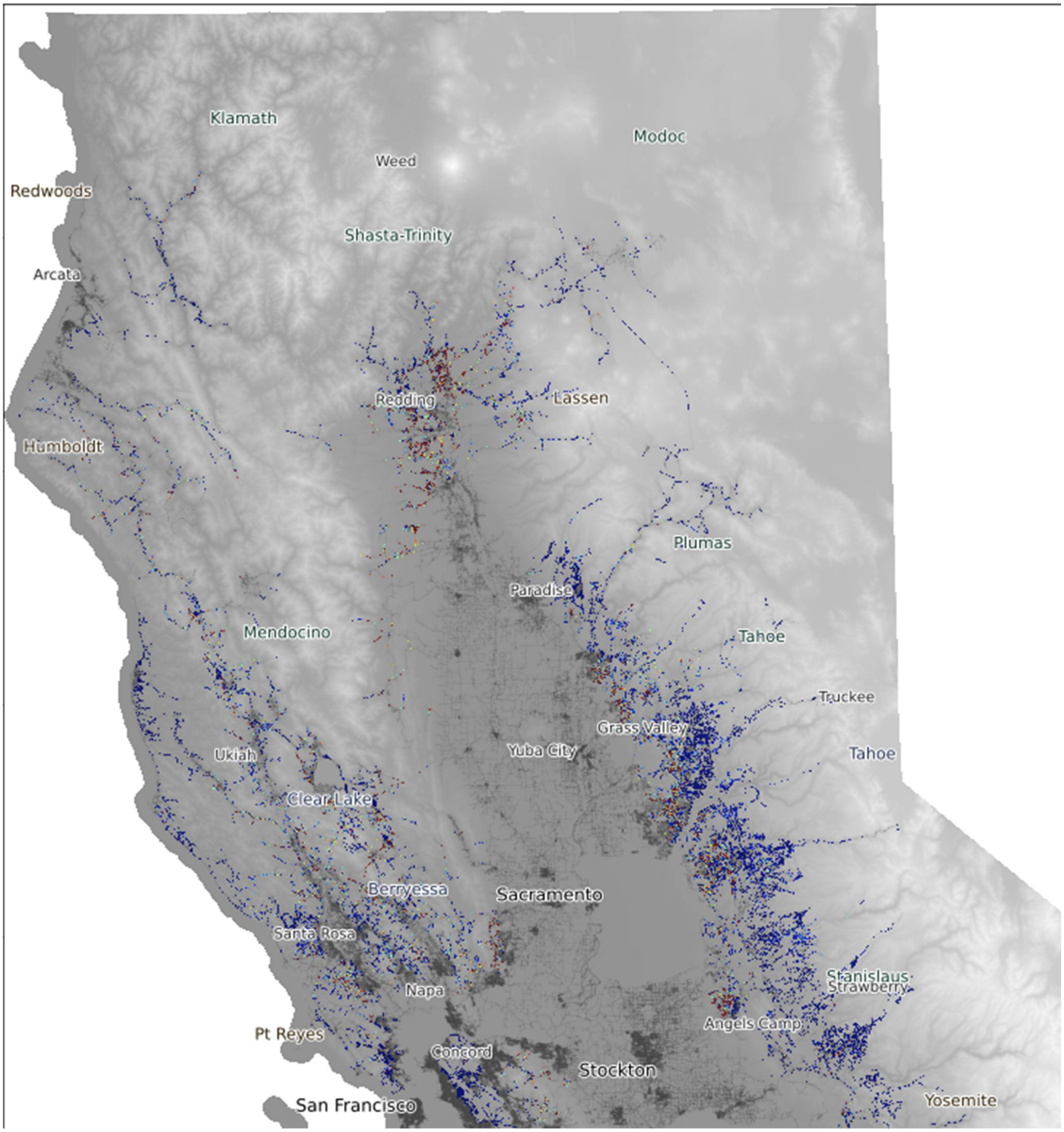

$Risk_{pixel} = P(ignition)_{pixel} * Consequence_{pixel}$
$Risk_{cpz} = \Sigma\ Risk_{pixel}$
$Risk_{cpz} = \Sigma\ Risk_{pixel}$

FIGURE 23 - A DIAGRAM OF HOW PIXELS ARE AGGREGATED TO CPZS

13 Risk Results

Risk results are mapped below, with the results broken into two areas north and south of San Francisco spanning two pages to allow a higher "zoom level" of CPZ level risk scores.

Ignition risk



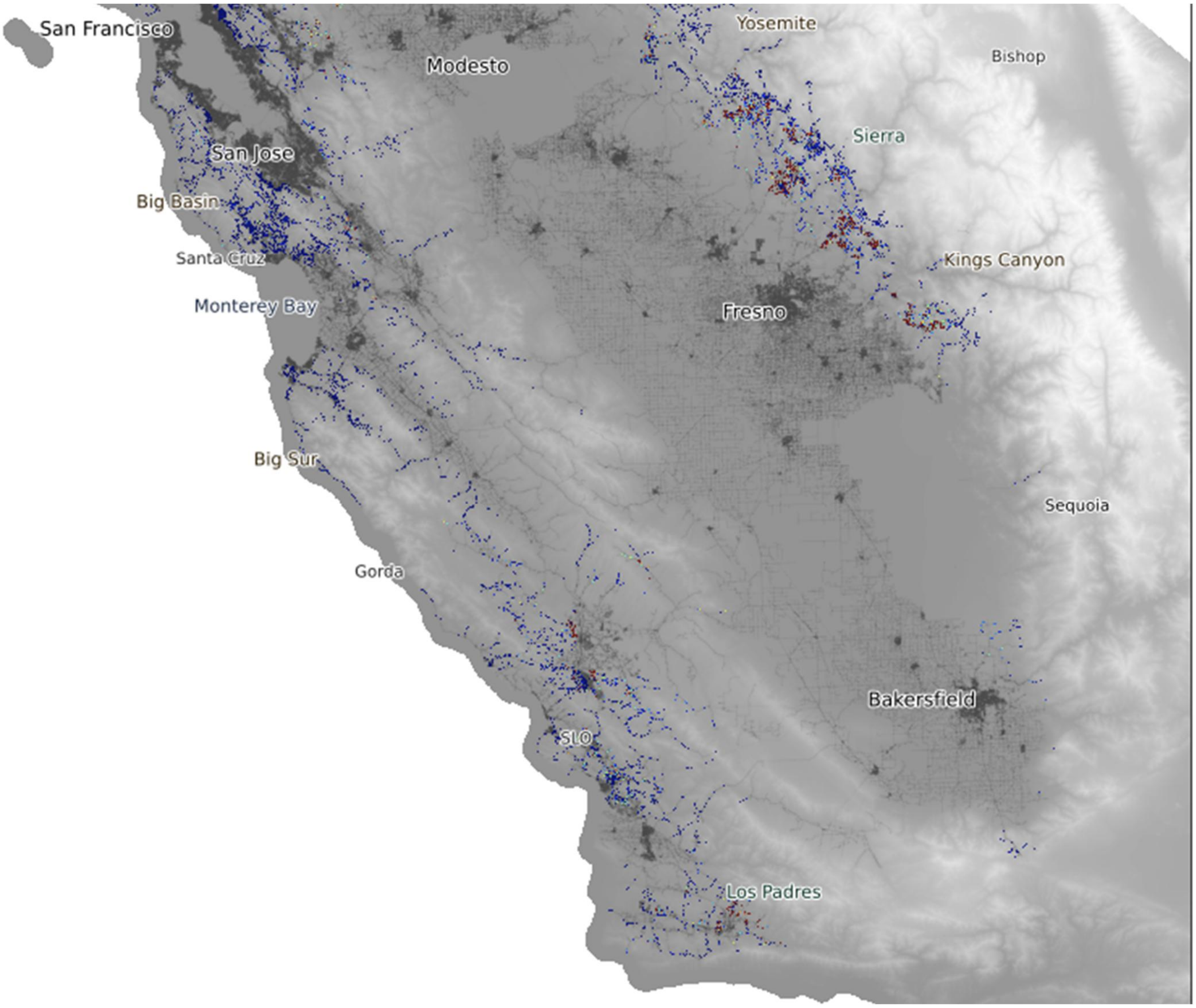


FIGURE 24 - PER-CPZ RISK - RED IS HIGHER, BLUE IS LOWER

13.1 Maxent Model Results

Maxent is the name of the software used to run our MaxEnt models. Maxent computes on a pixel-by-pixel basis. The blue "grid pixels" in Figure 25 below are the grid locations (including both primary and secondary conductors) that were modeled in Maxent. Only grid locations are modeled, because modeling wildfire risk from PG&E assets makes no sense where there are no such assets. Tiers 2 and 3 are filled with orange and red shading respectively, and non-HFTD areas are grey. The darker grey grid pixels indicate locations of the grid, but because they are outside of HFTD 2 and 3, they were not modeled in this analysis.

HFTD 2&3 conductor locations

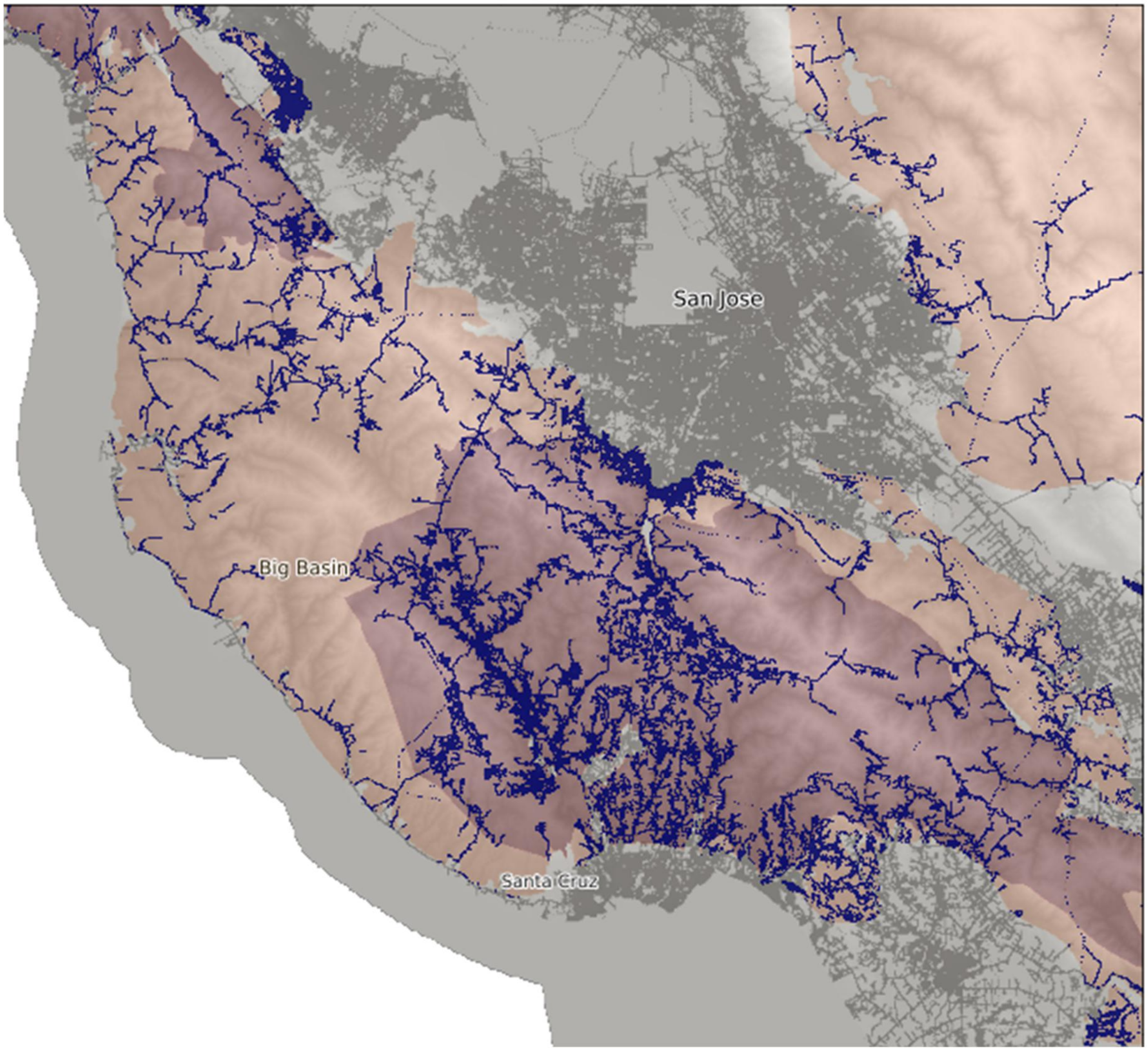


FIGURE 25 - CONDUCTOR PIXELS IN BLUE, HFTD 3 IN DEEPER RED, HFTD 2 IN LIGHTER ORANGE, FOR THE COASTAL AREA INCLUDING SANTA CRUZ AND THE LOWER SAN FRANCISCO PENINSULA

The figures below present two example Maxent raster covariates that are combined with other data in the modeling: per-pixel maximum tree height (Figure 26) and average wind speed (Figure 27) for the southern San Francisco Peninsula, including Big Basin and Santa Cruz. Note that all covariates are masked to HFTD areas (meaning non-HFTD areas are filtered out) because this analysis considered only those areas.

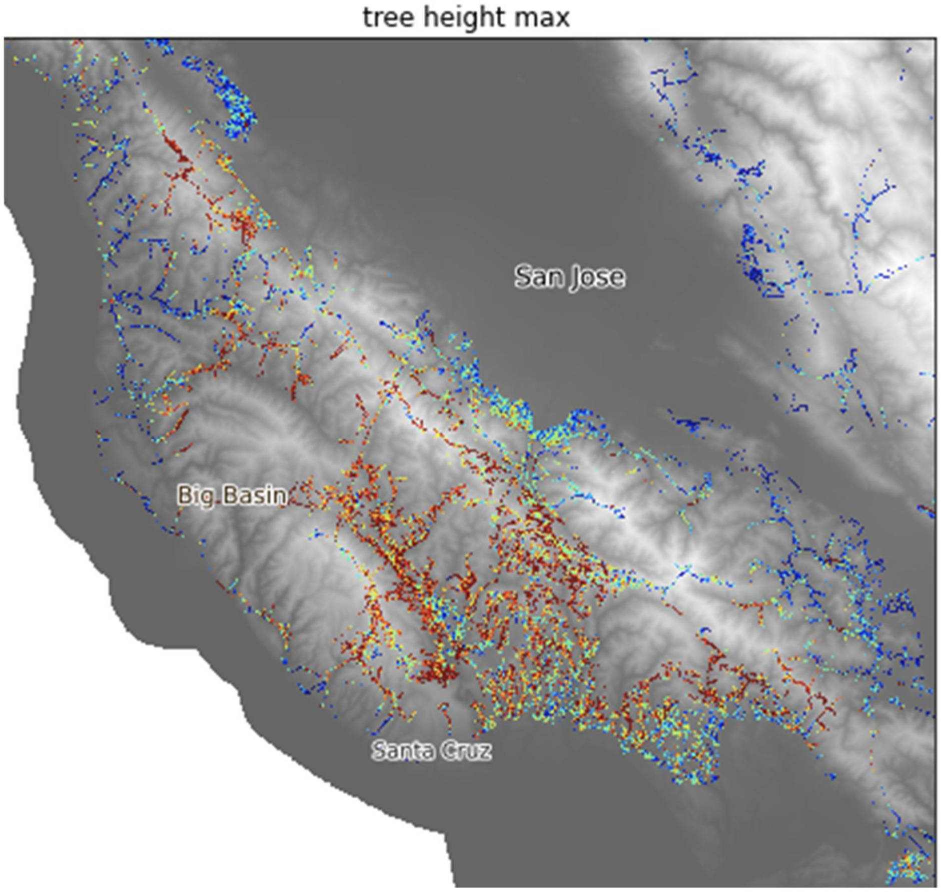


FIGURE 26 - PER-PIXEL MAXIMUM TREE HEIGHT - RED IS HIGHER, BLUE IS LOWER

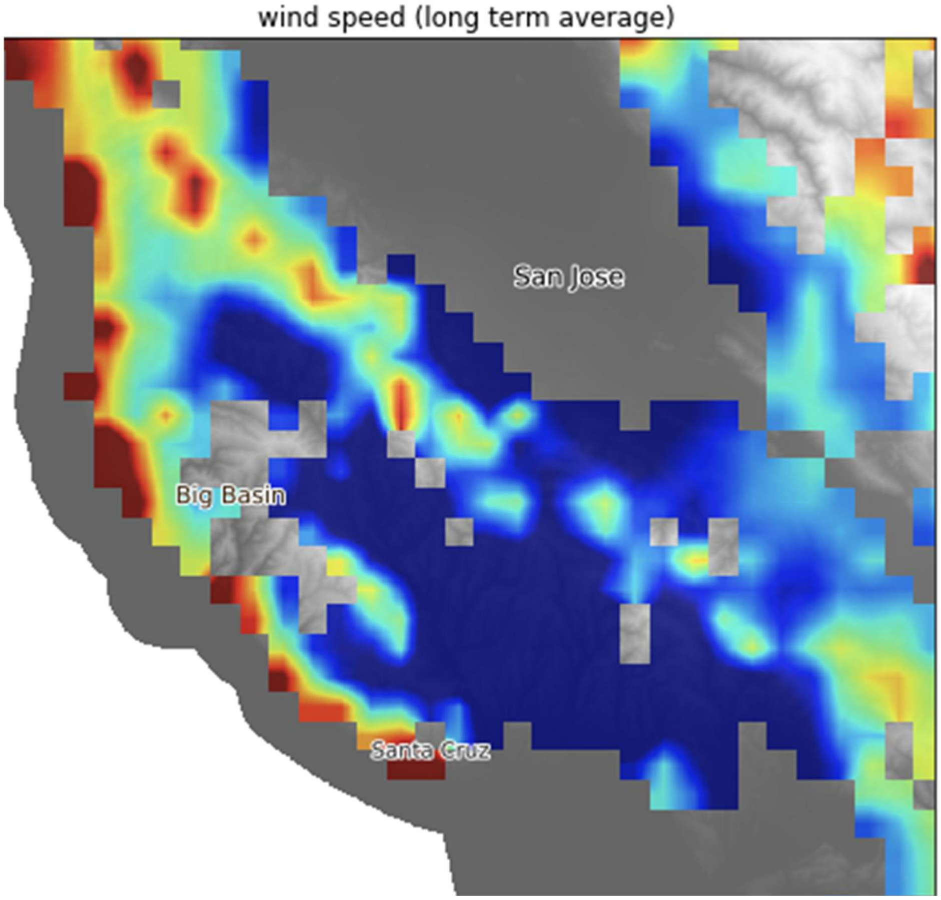


FIGURE 27 - PER POLAR-ORBITING METEOROLOGY SATELLITES (POMS) CELL LONG TERM AVERAGE WIND SPEED - RED IS HIGHER, BLUE IS LOWER

The resulting model output, the probability of vegetation-caused fire season ignitions for the same area, is depicted in Figure 28 below with highest ignition probabilities colored red and lowest colored blue.

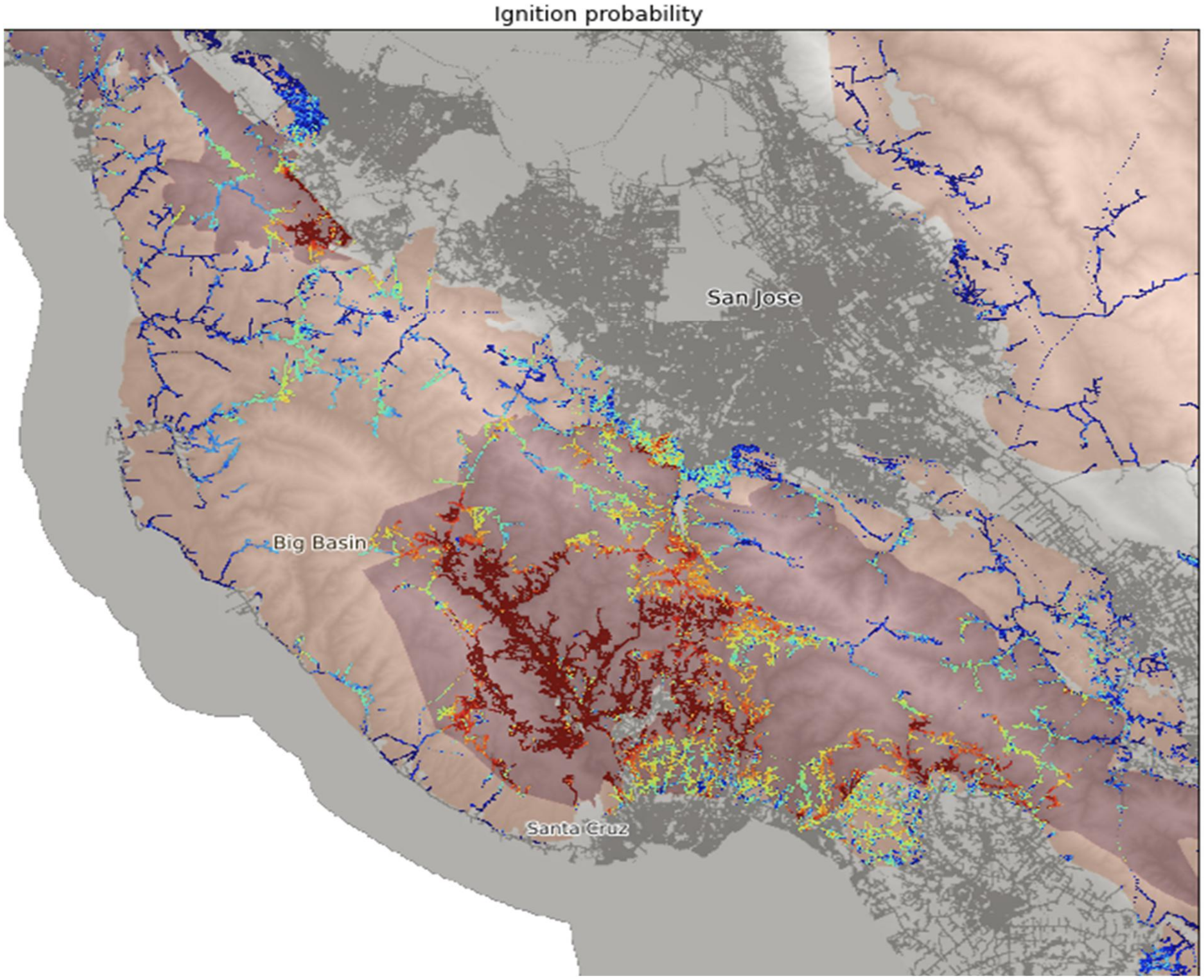


FIGURE 28 -IGNITION PROBABILITY RESULTS PER PIXEL, RED IS HIGHER, BLUE IS LOWER. GREY LINES ARE GRID-PIXELS OUTSIDE HFTDS, TRANSLUCENT ORANGE AND RED AREAS ARE HFDT TIERS 2 AND 3 RESPECTIVELY.

100m pixel representation of P(ignition) output from Maxent model for North Bay is shown in Figure 29 below - red is higher, blue is lower, non-HFTD conductors are shown in dark grey.

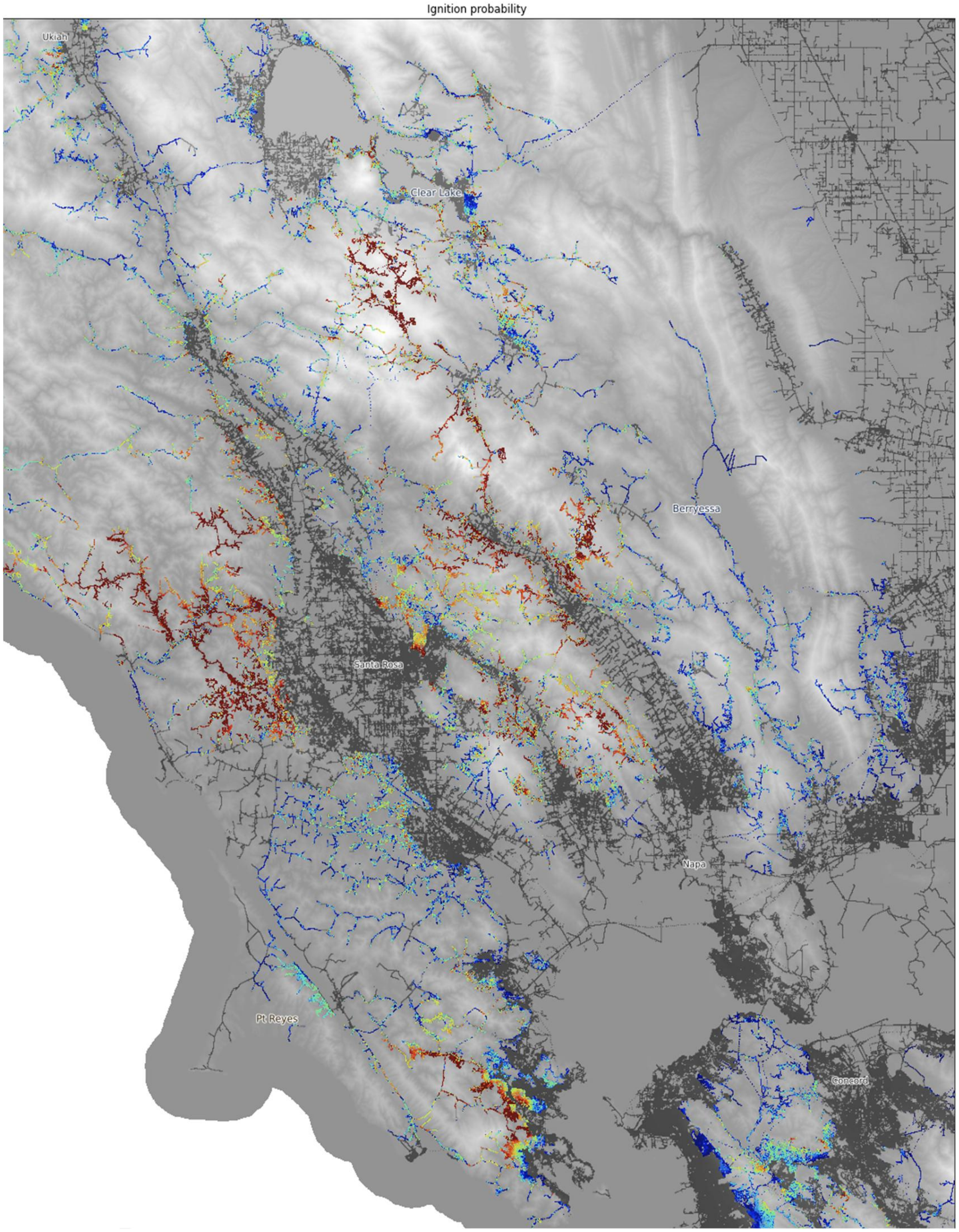


FIGURE 29 – IGNITION PROBABILITY PER PIXEL FOR THE NORTH BAY - RED IS HIGHER, BLUE IS LOWER COLORED GRID PIXELS ARE WITHIN HFTDS, DARK GRAY GRID PIXELS ARE NOT.

13.2 CPZ-level Aggregate Tree and Risk Metrics

The “official results” from this modeling effort were rolled-up to summaries at the Circuit Protection Zones (CPZ) level. See section 31 for details of the CPZs to which we aggregate model results.

Because the VMD 2019 trees data set is a list of trees, the tree locations can be used to read out their P(ignition) and consequence values and then those values along with tree attributes, including their risk scores can be rolled up, or aggregated, to CPZ-level summaries. Those summaries cover approximately 3,000 CPZs in HFTDs 2 and 3 and can be used for exploring tree risks and prioritizing EVM work.

This model covers approximately all CPZs in HFTD 2 and 3, as they existed in approximately mid-2018. This mid-2018 CPZ vintage was requested by stakeholders to facilitate comparison to CPZ-specific model results from the predecessor 2018 EVM risk model.

Figure 30 below provides a normalized histogram and density curve of the count of trees within each CPZ. Counts are widely varied, with most CPZs including few trees, but some CPZs include thousands of trees.

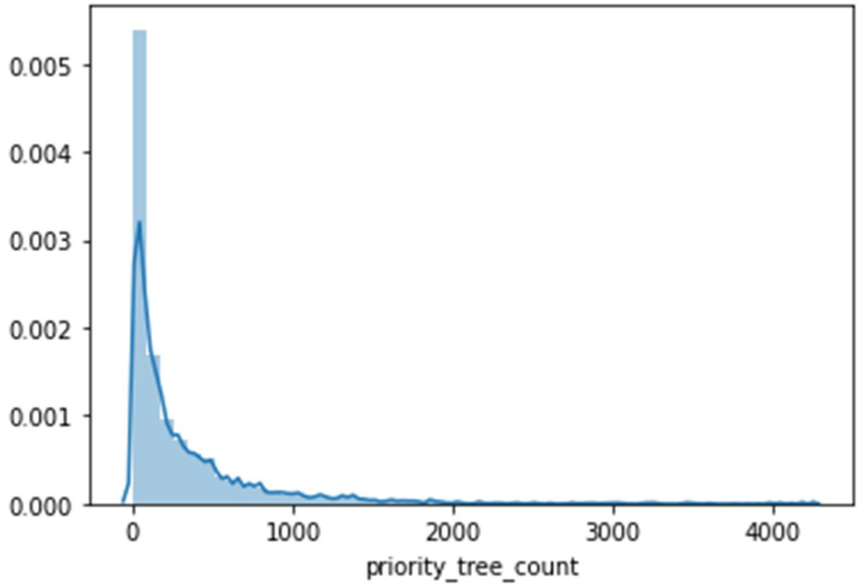


FIGURE 30 - COUNT OF TREES PER CPZ: Y-AXIS IS THE PROBABILITY DENSITY (THE AREA UNDER THE CURVE IS 1) AND THE X-AXIS IS THE NUMBER OF PRIORITY TREES FOUND WITHIN EACH CPZ

There are fewer trees in areas of high consequence: Figure 31 below shows scatter plots of per-CPZ data for all 3,000 CPZs analyzed where the y-axis is the count of trees in each CPZ and the x-axis is MAVF core consequence. It shows that higher tree counts tend to be associated with lower consequence values – in other words, there are fewer trees in locations with elevated fire consequences. This is as true for fire burn area as it is. For metrics like structures burned that related to settlement patterns (and presumably tree removal). This could be due to the fact that fire spread more readily in grasslands and chaparral than in forests - fire fighters talk about fuel ladders that allow fires to burn up into forest canopies (but the ladder doesn’t start there). Here, and in the plots to follow, we are looking at Risk, P(ignition), and Consequences all averaged per pixel, so that different CPZs can be compared in a fair way, despite having a variety of sizes.

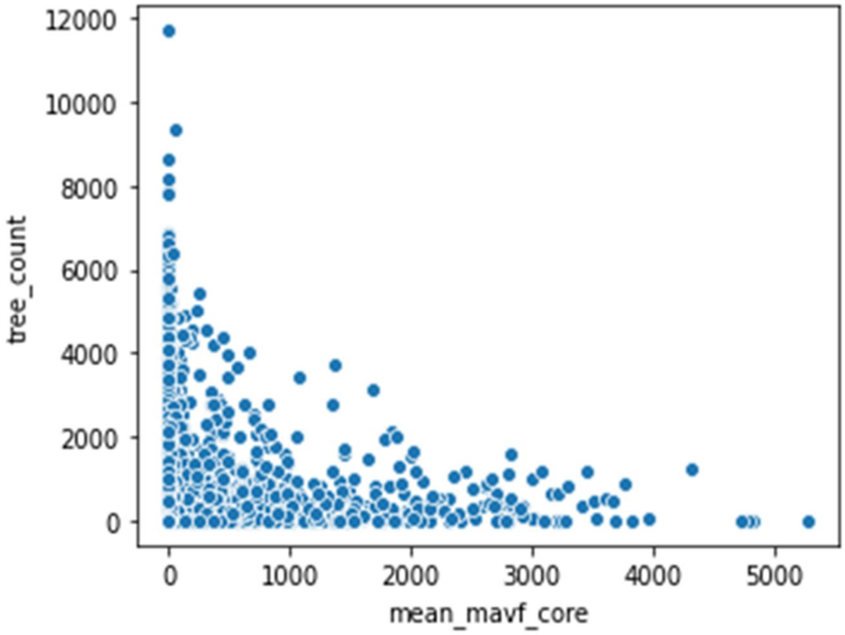


FIGURE 31 - SCATTER PLOT OF CPZS BY CONSEQUENCE AND TREE COUNT

Figure 32 Shows there is not as strong a relationship between P(ignition) and VMD tree density. But the highest P(ignition) values are generally associated with CPZs with fewer trees. This is likely due to the fact that reportable ignitions require dry fuels to grow to reportable proportions (1 m in extent) and areas with fewer trees tend to be hotter and dryer.

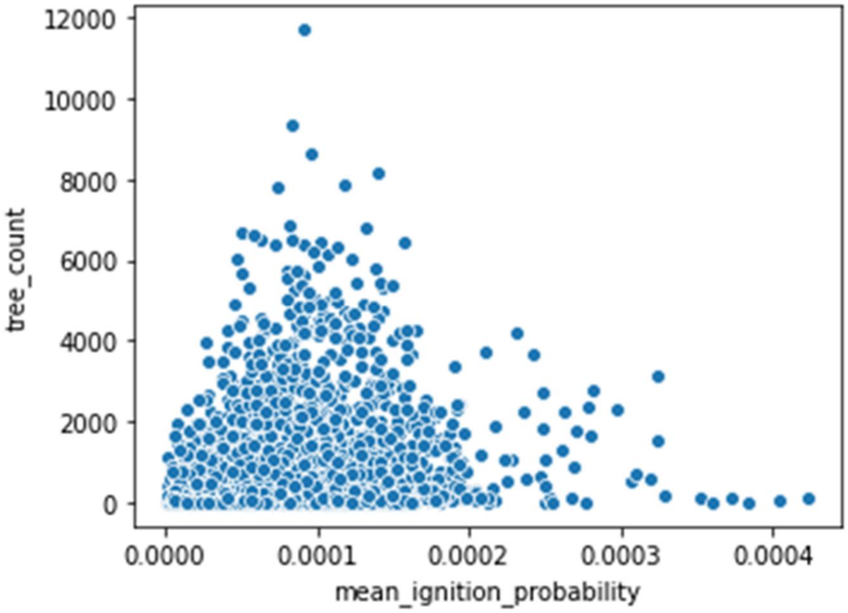


FIGURE 32 - SCATTER PLOT OF CPZS BY PROBABILITY(IGNITION) AND TREE COUNT

Risk scores are dominated by consequence values. Figure 33 below plot the Risk on the y-axis and the two components of that risk calculation (Consequence on the left and P(ignition) on the right) on the x-axes. From these, it can be verified that the Risk score is highly correlated with the Consequence and not particularly correlated with the P(ignition), although the areas of highest ignition probabilities are all considered lower Consequence. This has a lot to do with the fact that the Consequence values range over more orders of magnitude than the P(ignition) values. If you are prioritizing directly by Risk, you are largely prioritizing by Consequence, or the ability for a given location to host a catastrophic wildfire.

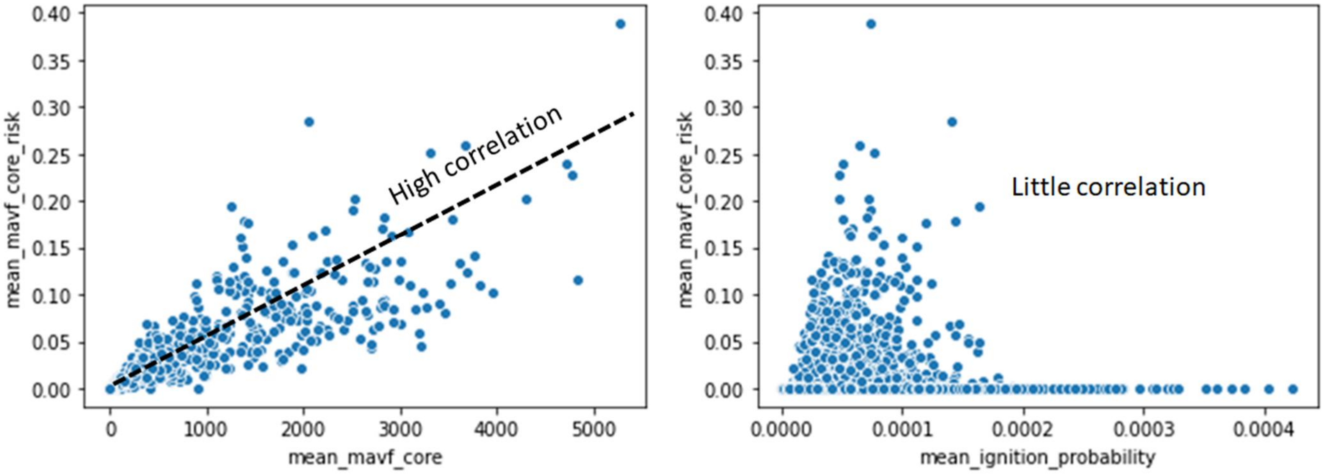


FIGURE 33 - SCATTER PLOTS PER-CPZ RISK CORRELATED WITH CONSEQUENCE (LEFT) AND IGNITION PROBABILITY (RIGHT)

Figure 34 below illustrates the locations of relatively high and low consequence, by CPZ: It is a map of the Sierra foothills, depicting the CPZ per-pixel average values of the MAVF CoRE consequence metric. Consequence tends to be highest further west/downslope in grassy and scrubby areas at the western edge of the Central Valley.

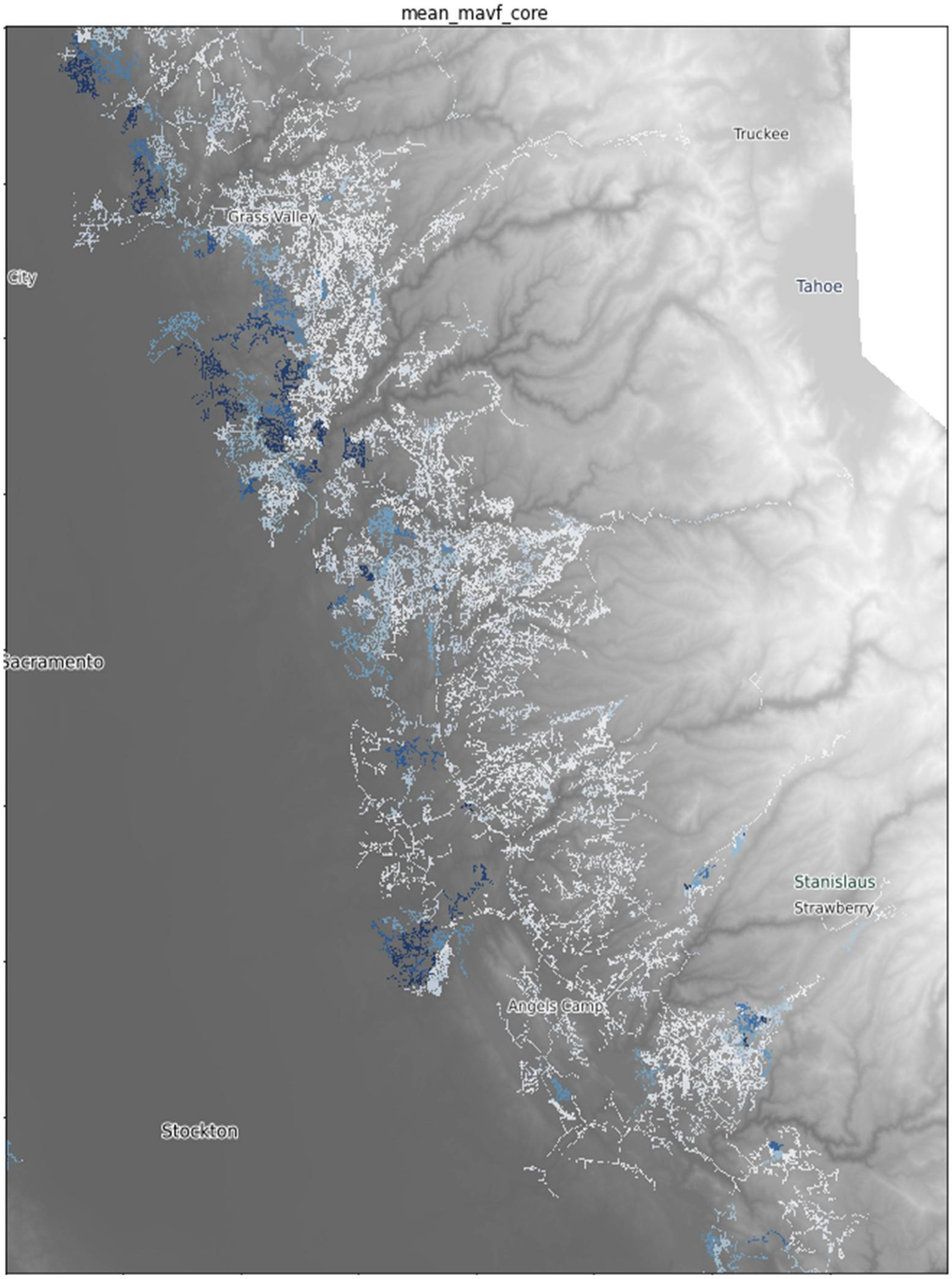


FIGURE 34 - CONSEQUENCE BY CPZ - DEEPER BLUE INDICATE HIGHER VALUES, WHITE INDICATE LOWER VALUES

Figure 35 depicts P(ignition) and Figure 36 depicts tree count. Both metrics tend to increase to the east of the Central Valley, but they are not perfectly aligned.

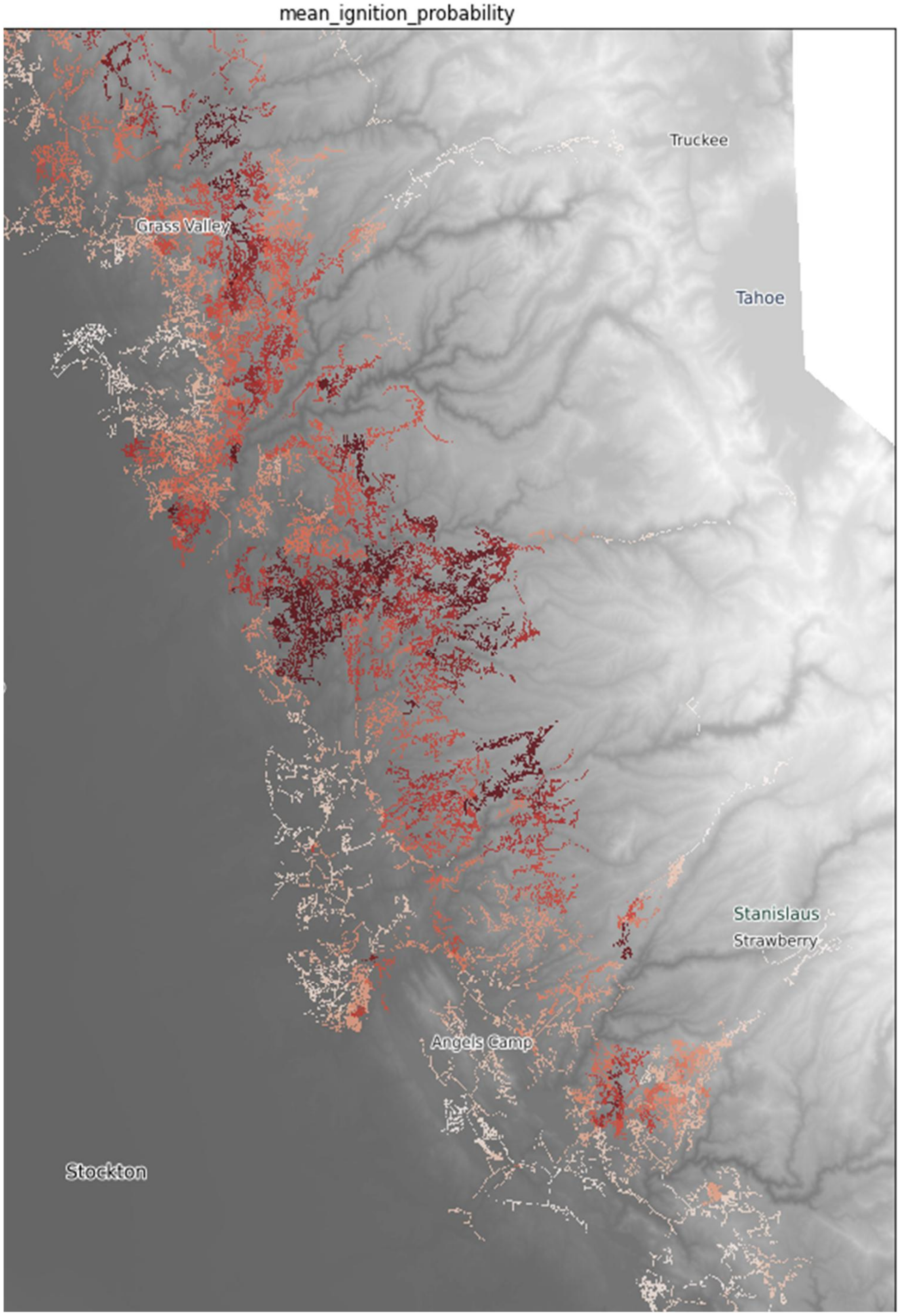


FIGURE 35 - MEAN PROBABILITY(IGNITION) - DEEPER RED INDICATE HIGHER VALUES, WHITE INDICATE LOWER VALUES

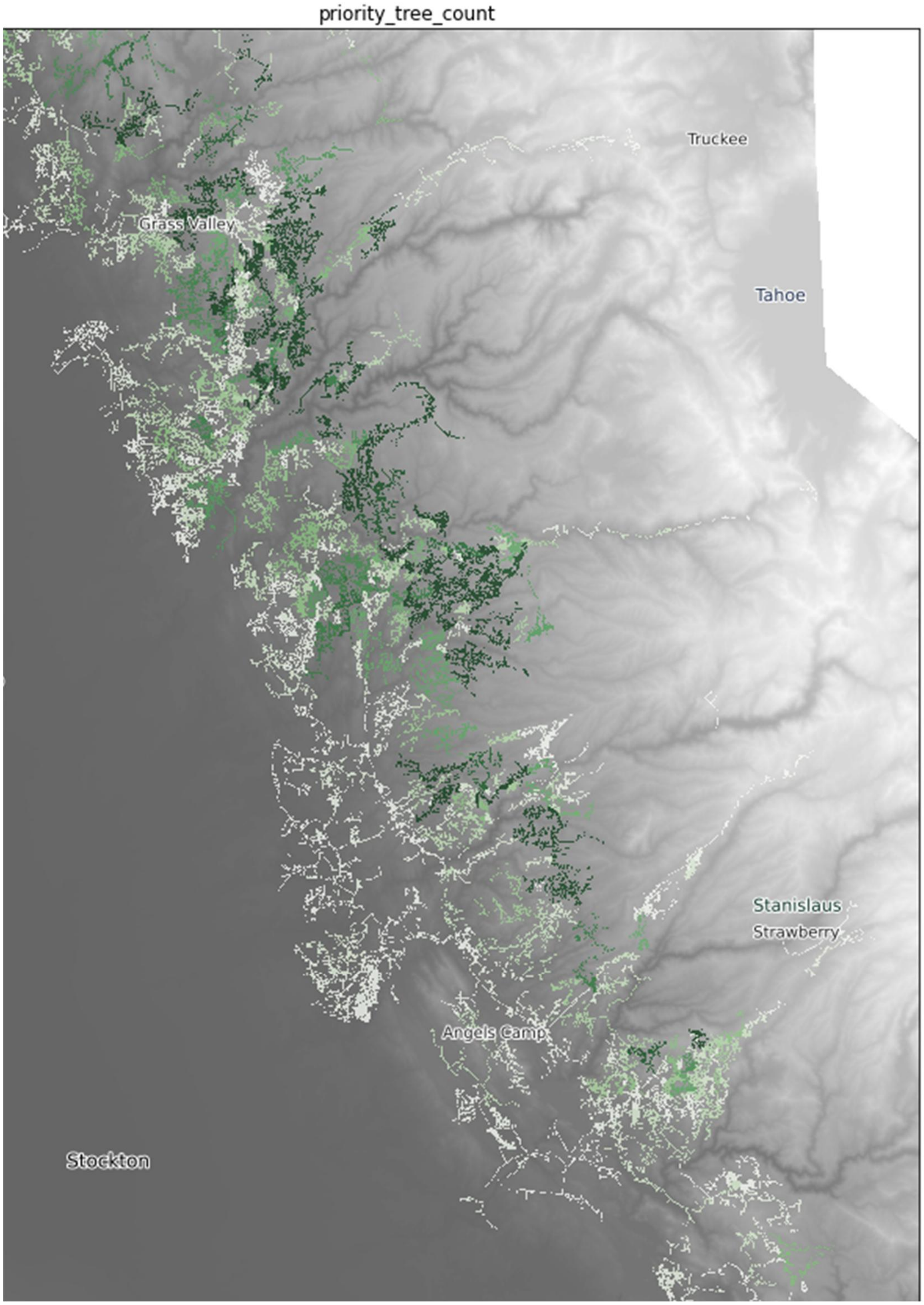


FIGURE 36 TREE COUNT - DEEPER GREEN INDICATE HIGHER COUNTS, WHITE INDICATE LOWER COUNTS

14 Validation

Risk results were reviewed by SMEs within PG&E with expertise in veg management and wildfire risk analysis. Feedback from these SMEs supported the adoption of the Technosylva-derived consequence dataset (Reax consequence data was used in the 2018 risk model) and updates to improve the delineation of CPZs within the HFTD areas. The pixel level results were imported into the Google Earth platform to enable a desktop analysis review of the pixel-level results along a CPZ. A field assessment to verify model results along two high risk CPZs was conducted.

A governance committee was established to verify that this 2021 model was a continuous improvement when compared to the previous 2018 model used for veg management planning. The committee approved the 2021 model.

As documented in the modeling sections, model performance from in-sample and out of sample runs was tabulated. The ROC plot from that test is reproduced below. The fact that the curves for Test data and Training data align so well and the AUC values are so close supports the interpretation that the model predicts well out of sample.

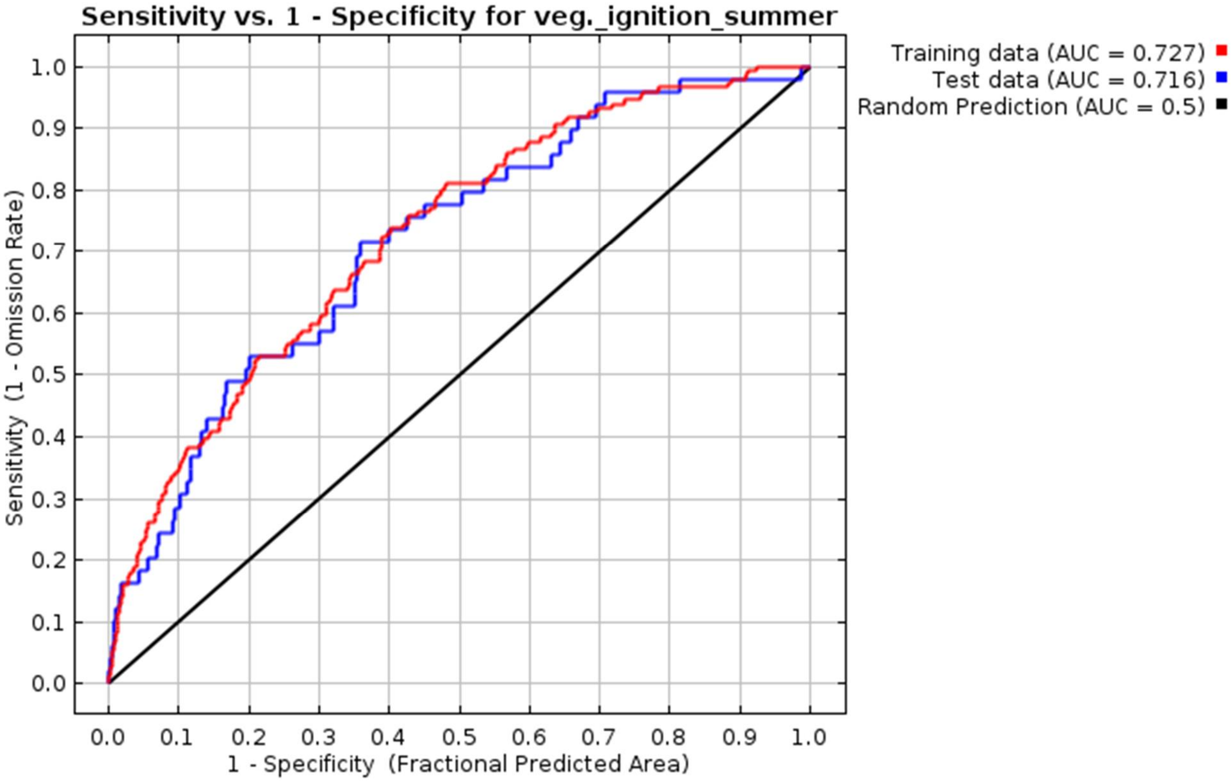


FIGURE 37 IN AND OUT OF SAMPLE MAXENT MODEL PERFORMANCE

15 Model Limitations

- There is additional tree data derived from LiDAR surveys of trees in HFTD. Incorporation of LiDAR tree survey data was not feasible on the timeline of this 2021 model. In the coming months we hope to incorporate LiDAR data in the 2022 iteration of the work described in this report. LiDAR data offers a complete picture of all the trees present in the vicinity of the electric grid assets within HFTD, in contrast to the VMD tree data which is known to be incomplete. Detailed and complete information about the trees that could possibly interact with the grid, is expected to improve the predictive performance of our model.

- Worst case fire simulation consequence data is not technically the correct consequence data to apply to all ignitions because ignitions take place over a wide range of mostly low-risk weather conditions, but it is conservative in that it considers all ignitions under dangerous fire conditions.
- In reality Vegetation Management has multiple categories of EVM work – especially branch work vs. tree removal based on avoiding different failure modes – branch failures, trunk failures, and root failures. We have modeled all vegetation-caused ignitions together, but in the future, modeling each cause category separately would allow Vegetation Management to better prioritize each specific category of work they do.
- No VM activity is permanent. Trees grow. Therefore, both the costs and benefits of VM work require careful consideration of the planning time frame and work repetition cadence involved. EVM work is maintained in subsequent years under the maintenance program.
  - Questions: Should maintenance costs be part of the risk-spend calculations? If not, should the risks be modeled as increasing over time as the trees grow back?
  - For this work we are only looking at a single year of EVM and the associated near-term (~1 year) risk reduction where regrowth is not an important consideration.

15.1 Overhead Branches vs. All EVM

EVM covers several areas of activity, but species selection as conceived by the VM team is primarily focused on trimming overhead branches, and thus on the mitigation of branch-failure-caused outages and ignitions. However, some trees whose branches need to be worked are removed instead if the branch work would effectively destroy the tree. Unfortunately, the EVM DB records the outcome, not the motivation, so only approximate estimates of branch-motivated effort are available. Rough estimates are that up to 20% of "removal" outcomes are due to branch concerns. For our initial modeling pass, we have not differentiated types of EVM and tree failure types and therefore have implicitly assumed that all categories of EVM work benefit from species selection in the same manner.

16 FAQ

What is the relationship between tree count/density and ignition risk? Why doesn't tree count play a more explicit role in the model?

To model the probabilities of vegetation-cased ignitions, and the mitigation caused by EVM, you must model trees. The question becomes what data sets on trees are available. This modeling effort relied on two different sources of tree data: (1) satellite-sensed data on tree canopy heights (and presence/absence) without the ability to resolve individual trees and (2) Individual tree records in vegetation management's VM database (aka the VMD). VM is clear that their database is not a comprehensive survey of all trees and they caution against using that data as ground truth for tree counts, so neither source provides an accurate count. We built predictive models using only the remote-sensed tree data and compared them to models that also incorporated VMD-based metrics of tree density. We found evidence that the VMD-derived tree density metric was leading to over-fitting on the training sample, with significant degradation of performance of out-of-sample prediction. Over -fitting is a situation where the trained model corresponds too closely or exactly to a particular set of data and may therefore fail to predict future observations reliably. The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e. the noise) as if that variation represented underlying model structure. With indications that the model was over-fitting and knowing that the counts were imperfect to begin with, we opted to exclude them from the "official" 2020 ignition probability model.

Further, when both VMD species outage scores and counts are included (experimentally) in the model, it is the species outage scores and not the counts that have higher explanatory power. As we've learned from VM, species matter quite a bit to the probability of outages and ignitions. Resources such as the Pacific North-West tree failure database (Pacific North

West Tree Failure Database, n.d.) and the Western Tree Failure Database (Western Tree Failure Database, n.d.) record tree failure data and demonstrate the different propensity of failure various tree species have. Additionally, significant research has been on-going regarding the mechanical properties of trees, and how it may differ depending on tree species (primer on trees interacting with wind, n.d.; Smiley, Matheny, & Lilly, 2012; Matheny & Clark, 2009) . All this leads us to believe that tree density is not as significant as tree species in predicting ignitions.

Finally, recall that we are predicting ignitions, not outages and not locations where fires spread to. Our other strong predictors, VPD and 100-hour fuel moisture, relate to dryness. Here is where you can get some counter-intuitive results - dense tree settlement requires moisture, shade and leaf litter are better at retaining moisture, and evapotranspiration increases humidity. Also, dense tree settlement tends to diffuse wind, lowering its impact compared to more open spaces. Taken all together, it *is harder* to get a reportable ignition within dense tree cover than in areas with dryer conditions. Those areas will tend to be more open, so the relationship between trees and ignitions is not simply a function of tree density.

The pattern of P(ignition) presents evidence that there is tension between the count of VMD trees and the dryness/exposure to wind required to produce a reportable ignition. The CPZs with highest P(ignition) do not tend to be those with the highest count of VMD trees.

Separate from the ignition probabilities, trees and other land cover are also included as inputs in the simulation models of fire spread used to compute MAVF core consequences. There is a general pattern of fires spreading more quickly in more “open” land cover where flame lengths (a metric of fire intensity) tend to be longer, so fires can “ladder” up from grass and scrub to denser fuels. Again, there is a trade-off between tree density and the early viability and spread of the fires.

Why doesn’t wind play a more prominent explanatory role in the model?

The model we've been asked to build assesses spatial differences in ignition likelihoods rather than temporal ones and treats all ignitions equally, with downstream consequences for each ignition calculated using fire simulations whose input weather is drawn from more than 400 of the worst fire weather days in the last 30 years. Wind covariates do help to explain the location of ignitions, but ignitions from wind event days make up a small minority of all ignitions and therefore wind event weather data does not play a prominent role in our model (unlike the PSPS model, which is entirely focused on short term weather).

The model described in this document was built to support planning across one-to-many years of effort. Our results are expected risk per-fire-season, inclusive of all-weather conditions experienced during each fire season. The expectation value realized by the model’s ignition probabilities are scaled to reproduce the average annual count of vegetation-caused fire-season ignitions used to train the model -- and those probabilities are differentiated spatially, with weather covariates aggregated temporally (i.e. averages, mins, and maxes) across fire seasons. In other words, entire seasons of weather data are pooled together to answer the question *“where is the probability of ignitions relatively high and relatively low, over planning timeframes?”*.

If we had been asked to build a model that predicts the likelihood of outages given a specific weather pattern, both wind speed and direction would play a significant role in the predictions. However, when modeling all ignitions over longer periods of time, prevailing wind speeds and directions play a different role. The EVM model is built on the assumption that past events predict future outcomes. As long as there are a similar number of wind events in similar locations over time, the model is already accounting for wind impacts on annual ignitions. However, the majority of ignitions are not caused by wind at all and 95% of outages do not occur during NE wind event days.

In that context, a consistent number of dangerous wind events can be presumed to occur each season in any given location, so it turns out that tree characteristics - especially height and proximity to conductors - are better predictors of the locations of high outage and ignition probabilities than seasonal wind summaries. This is largely due to the self-evident fact that trees incapable of contacting lines does not cause trouble, but can only be modeled because we have utilized a state of the art remote sensed tree height data layer that makes it possible to include such specific information in our model. Prevailing wind conditions do also help to explain where events occur, but the tree data is a stronger predictor. One way to interpret our results is that they assume more outages are occurring during dangerous wind conditions than other days, but that all else being equal, you can rely on a similar number of such events in the future as have occurred in the past.

What have we learned about seasonal wind and outages/ignitions for our modeling?

(1) Prevailing wind speeds (average daily maximum gust speed, for example) are poor predictors of events. Consistent wind "harvests" limbs over time and impact the strength and morphology of trees, a fact that complicates the relationship between prevailing wind speed and vegetation contact. Also, consistent winds are also fairly common along the coasts, and therefore correlated with cooler temperatures and higher humidity.

(2) Dangerous wind conditions are dangerous in part due to their novelty. If elevated wind speeds from a particular direction are unusual for a given location, they are more likely to contact trees with limbs that have not yet been tested. We find that the percentage of days with high wind gusts has predictive power on top of prevailing speeds, but the relationship appears to predict more incidents in areas where high gusts are less common.

(3) In exploratory work, the locations of *outages* differ significantly according to wind direction. If we isolate NE or NW wind days, we see a very different pattern in where outages occur, and this is consistent with the interpretation that "untested" trees are more likely to drop limbs and that topological interactions with wind can produce different outcomes depending on wind direction.

(4) Ignitions that start under hot/dry/wind conditions are more dangerous than others. Areas of high consequence are determined by fire simulations using worst historical fire conditions, especially NE wind events. So that portion of the risk is captured outside of the ignition probability modeling that we've been discussing but is very much integrated into the risk scores we report.

(5) We have just kicked off work on the next generation of our modeling - the so-called "composite model" that will be able to model subsets of events, like events during NE wind days, separately and do the proper bookkeeping to combine them into total counts. That model will allow for closer inspection of the causes and conditions associated with the 5% of outages that occur during NE wind events and, if needed, will allow for them to be weighted with greater importance than others in contributing to overall risk. However, there are significant data, software, and methodological problems that need to be solved before the composite model is fully operational.

17 References and Data Sources

CPUC. (2014, February 5). *Decision Adopting Regulations to Reduce the Fire Hazards Associated with Overhead Electric Utility Facilities and Aerial Communications Facilities.* Retrieved from https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M087/K892/87892306.PDF

Matheny, N., & Clark, J. R. (2009). Tree risk assessment: What we know (and what we don't). *Arborist News 18(1)*, 12-19.

Ohring, M. (1995). Failure and Reliability of Electronic Materials and Devices. *Engineering Materials Science, https://www.sciencedirect.com/science/article/pii/B9780125249959500398*, 747-788.

Pacific North West Tree Failure Database. (n.d.). Retrieved from https://www.arcgis.com/apps/opsdashboard/index.html#/23d3d47df6ee46d1b1f34e2910e467dc

PG&E Digital Catalyst. (2019). *STAR: The System Tool for Asset Risk - Primary Overhead Conductors.* Retrieved from PG&E Wiki: https://wiki.comp.pge.com/display/SW/Primary+Overhead+Conductors#PrimaryOverheadConductors-EDGISConductorDataset

PG&E Electric Operations. (2020). *2021 Asset Management Plans.*

PG&E EORM. (2020). *Methods - Spatial Wildfire Consequence 2021.*

PG&E EORM. (2020). *Spatial Wildfire Consequence 2021 - Lunch n' Learn presented 2020_10_16.*

PG&E Risk and Data Analytics. (2020). *Conductor Risk Model 2021 - Lunch n' Learn presented 2020_10_28.*

PG&E Risk and Data Analytics. (2020). *EVM Risk Model 2021 - Lunch n' Learn presented 2020_10_21.*

PG&E Risk and Data Analytics. (2020). *EVM Risk Model 2021 – Utility Analytics conference presented 2020_10_29.*

PG&E Risk and Data Analytics. (2020). *Methods - Ignition Probability Modeling 2021.*

primer on trees interacting with wind. (n.d.). Retrieved from https://ucanr.edu/sites/treefail/files/204521.pdf

Smiley, E. T., Matheny, N., & Lilly, S. (2012). Qualitative Tree Risk Assessment. *Arborist News 21(1)*, 1-20.

The Nature Conservancy. (n.d.). *Topographic Position and Landforms Analysis.* Retrieved from http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf

University of California, Merced. (n.d.). *GRIDMET: University of Idaho Gridded Surface Meteorlogical Dataset.* Retrieved from Earth Engine Data Catalog: https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET#description

USGS. (2016). *LANDFIRE dataset.* Retrieved from https://landfire.cr.usgs.gov/distmeta/servlet/gov.usgs.edc.MetaBuilder?TYPE=HTML&DATASET=FBK.

Western Tree Failure Database. (n.d.). Retrieved from https://ucanr.edu/sites/treefail/CTFRP_Statistics/50_or_more_753/

Appendix 2: Conductor-Involved Ignition Risk Model 2021

18 Executive Summary

The PG&E Risk and Data Analytics (RaDA) team developed this Conductor-involved Risk Model to estimate the risk posed by fire season (Jun. 1 – Nov. 30) ignitions involving primary overhead conductor segments of the distribution grid within the high fire-threat district (HFTD) Tier 2 & 3 areas. The model predictions inform the Asset Strategy – Grid Design team of the locations the model has identified to be at elevated risk of an ignition event involving a conductor failure occurring in a given planning year.

The Conductor Risk Model (the "model") uses a maximum entropy algorithm to predict the conditional probability that a CPUC-reportable ignition will occur within a fire season period within a given 100m by 100m pixel location containing primary overhead conductors. Model features (also referred to as variables or covariates) include high-spatial-resolution environmental data, meteorological data, and characteristics of the primary overhead conductors. The key features driving predicted probabilities are percentage un-burnable ground cover, average daily precipitation, conductor age, and conductor material.

The primary method used to evaluate the predictive algorithm's goodness-of-fit was the receiver operator curve – the area under the curve (ROC-AUC) metric. The area under the ROC curve of 0.5 represents random chance and 1.0 represents perfect prediction. The dataset used to train the model achieved an AUC score of 0.76 and the 2019 out-of-sample test dataset achieved an AUC score of 0.74.

As is typical of risk models, risk is calculated by multiplying probability by consequence. Risk results were calculated at each 100m by 100m pixel and aggregated to larger segments of the distribution system, called circuit protection zones (CPZs). A spreadsheet of results for each CPZ, and geospatial files of the results, were developed and provided to the Asset Strategy team to enable the use of probability, consequence, and risk results to inform the planning of system hardening mitigations that are intended to reduce the risk of conductor-involved wildfires.

19 Introduction

19.1 Background

Wildfires are an increasing problem in California with larger fires occurring more frequently (Figure 38). Before PG&E began conducting Public Safety Power Shutoff (PSPS) events to prevent conductor contacts or failures from causing ignitions that could lead to a catastrophic wildfire, there were, on average, 60 conductor involved ignitions per year within the HFTD areas. Roughly 85% of those occurred during the June through November period commonly recognized as the period most likely for ignitions to grow into wildfires. Because conductor failure – whether caused by internal weakness, high wind loading, or a vegetation contact – has historically been a significant contributor to ignitions within PG&E territory, PG&E's current commitment to reduce the likelihood (probability) and consequence of catastrophic wildfires makes it imperative that we understand the risk of conductor involved ignitions when considering prioritization of wildfire mitigation projects.

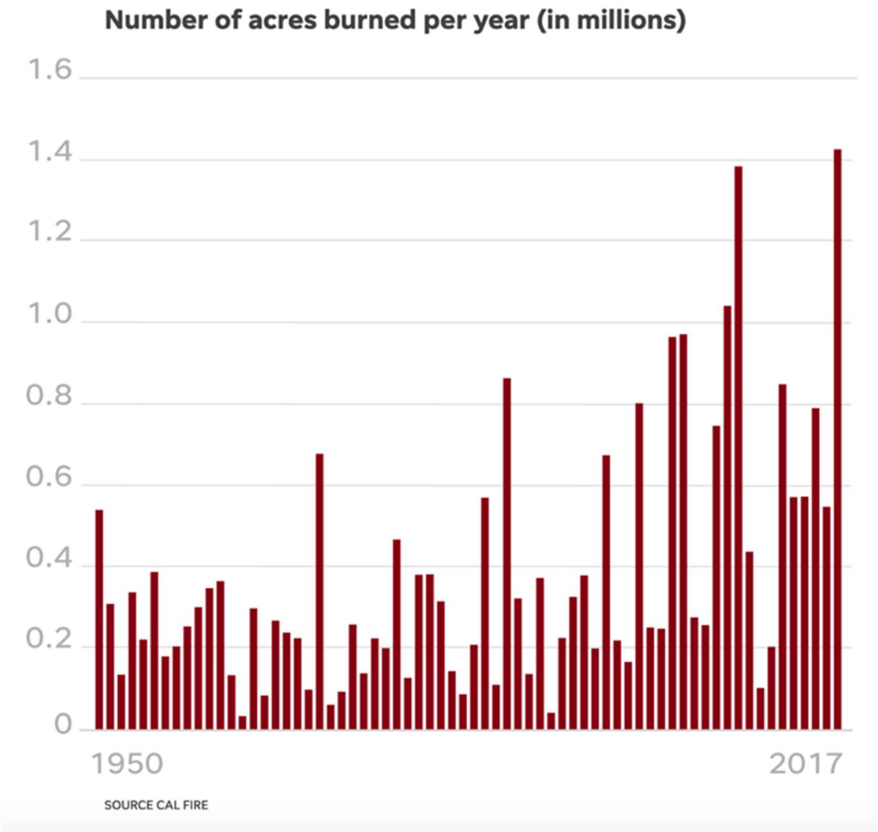
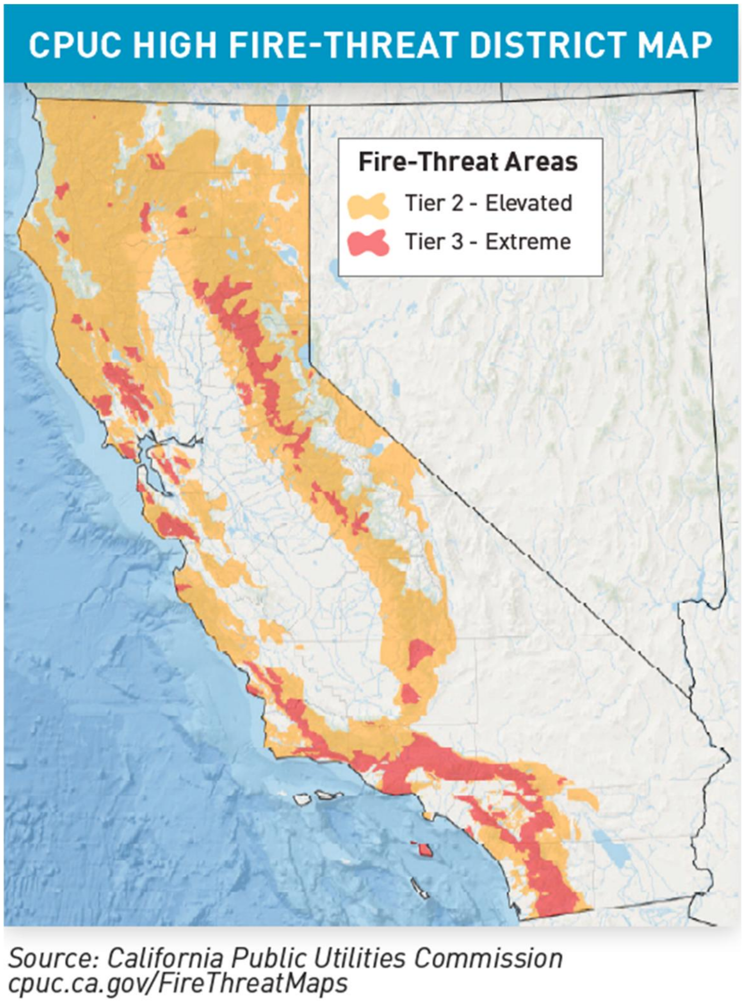


FIGURE 38. NUMBER OF ACRES BURNED PER YEAR IN CALIFORNIA FROM 1950 TO 2017

19.2 Project

The PG&E Risk and Data Analytics (RaDA) team developed the Conductor Model to estimate the probability of a fire season ignition occurring across segments of the overhead distribution grid within the high fire-threat district (HFTD) Tier 2 & 3 areas (Error! Reference source not found.37). To identify a risk output, the RaDA team combined the probability of ignition with the consequence of wildfire to calculate wildfire risk for each 100m by 100m pixel associated with primary overhead conductors across PG&E's distribution system. The pixelated risk values were aggregated to larger segments of the distribution system, called circuit protection zones (CPZs).



Source: California Public Utilities Commission
cpuc.ca.gov/FireThreatMaps

FIGURE 39 MAP OF THE CPUC HIGH FIRE-THREAT DISTRICTS (HFTD)

19.3 Usage

This section is intended for both technical and non-technical audiences to explain the purpose of the model, summarize the methodology used to develop the model, highlight the performance of the model, and discuss the risk results.

19.4 Applications

The predicted risk values of conductor failure and risk of ignition for each CPZ can be used to inform system hardening mitigation planning decisions to reduce wildfire risk. The Asset Strategy – Grid Design team can use the risk predictions to identify those CPZs with elevated risk of an annual ignition event involving a conductor failure when developing wildfire mitigation workplans.

This tool is also intended to provide a method for quantifying the risk reduction achieved through the wildfire mitigation work related to minimizing conductor failures that can lead to ignitions.

19.5 Model Limitations

While the model has improved performance in predicting ignition locations when compared to the previous model, the model does not:

- Prescribe the type of mitigation that should be performed to reduce risk along a CPZ.
- Predict the likelihood of conductor failure. This model predicts the likelihood of an ignition where the ignitions are filtered specifically for conductor-involved events.
- Indicate the root cause of conductor failures. While the features in the model help *predict* where conductor-related ignitions are expected to occur and *may* give limited insight into causes of conductor failures, it does *not* indicate that the feature is *causing* conductor-failures.
- Provide the tradeoff risk for alternative mitigation types. Currently, the training dataset used for the vegetation risk model has some overlap with the training dataset used for this model. This means that the risk values cannot be compared between the models.

20 Conductor Ignition Probability Model

20.1 Model Framework

The Conductor Model is a classification model that identifies the likelihood of at least one ignition per year at a given 100m by 100m pixel location. More specifically, the Conductor Model models the conditional probability that a reportable ignition will occur within a given year within a given pixel location.

The model was fit using a presence-only maximum entropy (MaxEnt) algorithm. The algorithm is based on the assumption that the most unique characteristics of locations that have experienced ignitions in the past will predict the locations that experience ignitions in the future. This algorithm is explained in more detail in Section 21.2.

The model was trained using reportable ignitions, as defined by the California Public Utilities Commission (CPUC) and outlined below. The ignitions were filtered for ignition events associated with an overhead conductor failing, as identified during the follow-up investigation. The ignitions were also filtered by date, using only ignition events that occurred during the fire season (Jun. 1 – Nov. 30). Ignitions data ranged from 2015 to 2019, and 2019 ignitions data were withheld from the model training dataset to use as a test dataset. Available full-season ignitions data ranged from 2015 to 2019, and 2019 ignitions data were withheld from the model training dataset to use as a test dataset.

20.1.1 CPUC-Reportable Ignitions

CPUC reporting requirements are limited to reportable fire events that meet the following criteria:

- A self-propagating fire of material other than electrical and/or communication facilities,
- The resulting fire traveled greater than one linear meter from the ignition point, and
- The utility has knowledge that the fire occurred

Fires that caused damage to utility facilities and whose ignition is not associated with utility facilities are excluded from this reporting requirement (CPUC, 2014)

20.2 Model Assumptions

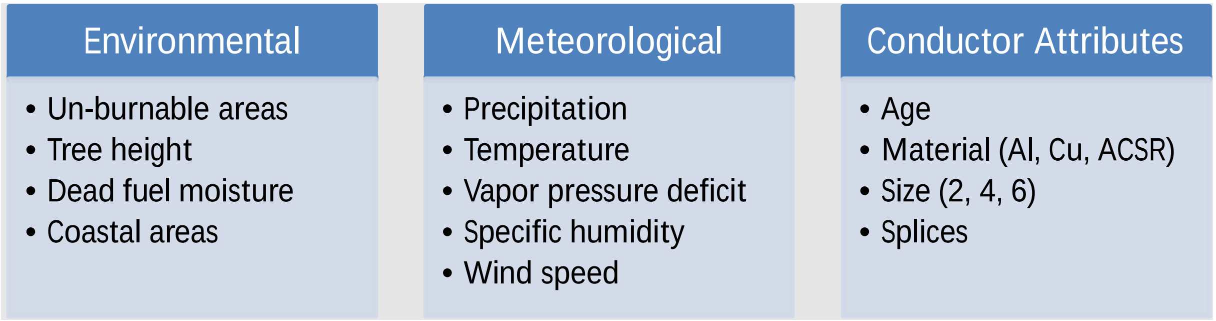
The following assumptions were made when developing this model:

1. CPUC-reportable ignitions do not consider the resulting fire area or volume other than meeting the minimum size threshold for the CPUC reportable ignition.

2. The model does not consider the mitigation effect of PSPS events that started occurring in 2019 and may overpredict the annual number of ignitions as a result.
3. The model drew upon conductor attribute data of recent vintage. Those attributes were used for all modeling, despite the fact that some ignitions studied occurred five years prior to the attribute snapshot.

20.3 Model Features

A combination of high-resolution environmental data, meteorological data, and conductor characteristics related to elevated failure rates were used as feature inputs to the model (Figure 40). Each feature dataset is discussed in more detail in the following subsections, and the importance of the features are discussed in the following Model Evaluation section (20.4). Data sets were selected based on availability. As highlighted in the Future Improvements section (7.0) future models will continue to work to incorporate more data sources to improve predictive power.



Environmental
- Un-burnable areas
- Tree height
- Dead fuel moisture
- Coastal areas

Meteorological
- Precipitation
- Temperature
- Vapor pressure deficit
- Specific humidity
- Wind speed

Conductor Attributes
- Age
- Material (Al, Cu, ACSR)
- Size (2, 4, 6)
- Splices

FIGURE 40 - CATEGORIES OF DATASETS USED AS INPUTS, OR FEATURES, IN THE MODEL

20.3.1 Environmental Data

20.3.1.1 Un-burnable

The “un-burnable” feature is a land surface descriptor similar to imperviousness that includes surfaces that typically don’t ignite when a spark occurs. The feature was derived from the designated non-burnable land use types within the 2016 LANDFIRE surface fuel model (USGS, 2016). The model feature is the portion of the 100m x 100m pixel identified as un-burnable. The LANDFIRE model‘s designated non-burnable spatial layers over-layed to create the composite spatial feature include:

- urban
- perennial snow/ice
- agriculture
- water, and
- barren.

20.3.1.2 Tree height

Tree height data were obtained from a third-party vendor, Salo Sciences, and the “tree-height-max" feature was developed by calculating the maximum tree height, in meters, for each 100m x 100m pixel area along the distribution grid, according to the processed satellite data provided by Salo. The satellite imagery was collected in November 2019.

The mean tree height was tested as a feature in the model and removed because it did not notably contribute to the prediction power.

20.3.1.3 Dead fuel moisture

The dead fuel moisture data were obtained from GRIDMET, and the “100-hour-fuels" feature was included in the model. The GRIDMET dataset used is a standard fire modeling metric of fuel dryness.

The 1000-hour dead fuel moisture data were also tested as a feature in the model and removed because it did not contribute to the prediction power.

20.3.1.4 Coastal

Coastal areas were identified using a binary feature in the model. Conductor geometries are tagged with a coastal indicator field in EDGIS. Coastal marine layer weather conditions are a known factor in driving conductor corrosion.

20.3.1.5 Topography

The relative topography of the area was also used as a feature in the model. The topographic position index (TPI) was extracted from a United States Geological Survey (USGS) national elevation dataset (NED) at 100-meter resolution. The TPI compares the cell elevation to the mean elevation for the local neighboring area (positive values are above the mean and negative values are below the mean) (The Nature Conservancy)

20.3.2 Meteorological Data

Gridded Surface Meteorological dataset (GRIDMET) (University of California, Merced) meteorological data was retrieved via Google Earth Engine at a resolution of 2.5 arc minutes, or roughly 4-kilometer resolution. This data is adapted, subsequently up-sampled to the 100 meter100m x 100m grid pixel units the model utilizes. The dataset included daily meteorological measurements for the full 3-year time period from 2014 to 2016. The maximum entropy algorithm is purely a spatial model and does not include a temporal variable. For this reason, the daily values were averaged over the full time period to use as features representing the local climate of a location in the model. Since maximum entropy utilizes the spatial variation between locations to identify areas with similar characteristics, the time period gives sufficient coverage to identify areas with similar climate characteristics. Future improvements to the model include the migration to internal PG&E meteorological data.

20.3.2.1 Precipitation

The average daily precipitation was calculated from the GRIDMET dataset. The daily total precipitation, in millimeters (mm), was averaged from 2014 to 2016.

20.3.2.2 Vapor Pressure Deficit

The average vapor pressure deficit was calculated from the GRIDMET dataset. The daily average, in kPa, was averaged from 2014 to 2016.

20.3.2.3 Specific Humidity

The average specific humidity was calculated from the GRIDMET dataset. The daily average, in %, was averaged from 2014 to 2016.

20.3.2.4 Temperature

The average maximum temperature was calculated from the GRIDMET dataset. The daily maximum, in units of Kelvin, was averaged from 2014 to 2016.

20.3.2.5 Wind

The hourly average wind velocity and gust velocity at a height of 10-m was calculated from the Real-Time Mesoscale Analysis (RTMA) dataset at a resolution of 2.5-km. The daily mean values of each, in meters-per-second, were averaged from 2016 to 2018 to create the wind_ave feature.

The average daily maximum values were also calculated and tested as a feature in the model, but it was removed because it did not notably contribute to the prediction power.

20.3.3 Conductor Attributes

The characteristics of conductors are particularly important model features because they help explain why a conductor breaks and can therefore help quantify the effectiveness of wildfire mitigation work.

For example, the reduction in the ignition probability when a segment of smaller diameter size 6 copper conductors is replaced with larger diameter size 2 aluminum conductor steel reinforced (ACSR) conductors can be quantified by the model, holding all other environmental variables constant. In this mitigation scenario, the conductor material, conductor size, and conductor age attributes would all change in the model. The conductor attribute features included in the model were identified by outages subject matter experts (SMEs) to be associated with elevated conductor failure rates (Figure 41).

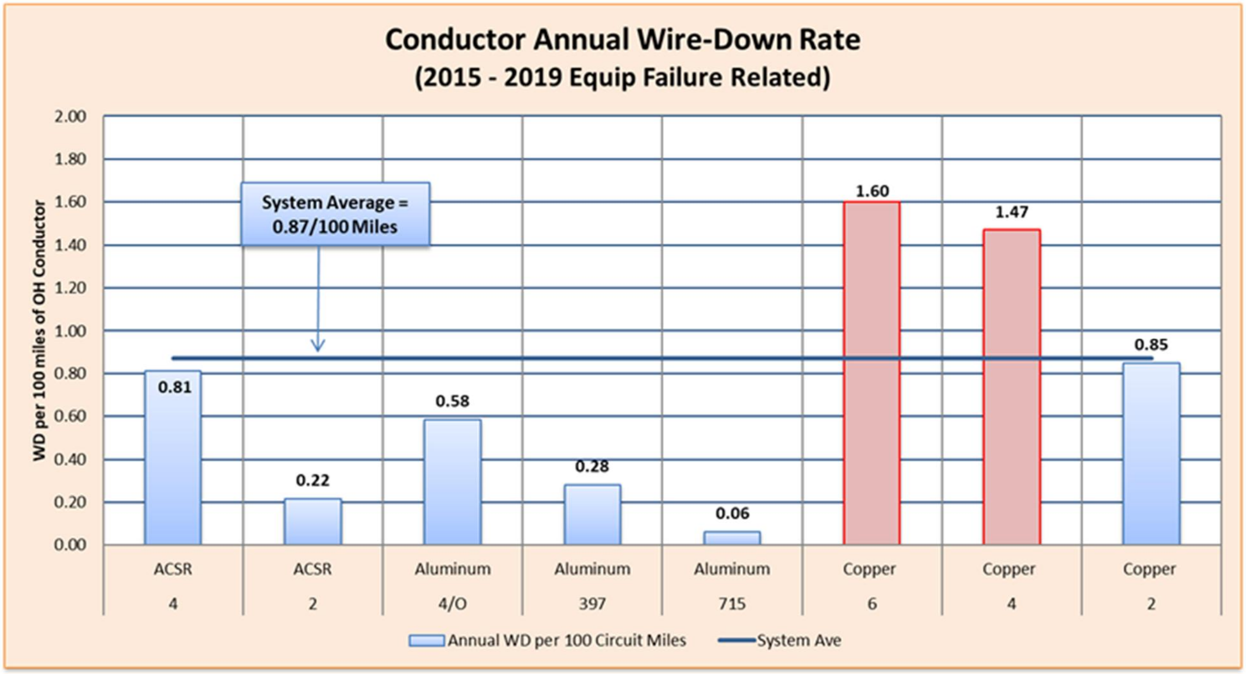


FIGURE 41 - 2015 - 2019 CONDUCTOR ANNUAL WIRE-DOWN RATES AS SHOWN IN THE 2021 ASSET MANAGEMENT PLANS (PG&E ELECTRIC OPERATIONS, 2020)

20.3.3.1 Conductor Age

The estimated conductor age (the “estimated-age”) was calculated as the number of years since the installation year, as listed in EDGIS. If the installation date was missing or invalid, then the estimated age in the STAR model dataset was used (as extracted from the primary conductor dataset in the Foundry platform). The installation date was determined to be invalid if:

1. It fell within the 1986 to 1990 time period, an unreliable default value in the dataset,
2. It was greater than the current date, or
3. It was less than 1901.

The STAR model estimated the conductor age using the average age of the poles associated with the conductor or, if pole age could not be calculated, the average age of the conductors in the service territory (PG&E Digital Catalyst, 2019).

The response chart shown in Figure 42 is an output from the model and shows that conductors *less than* 15-years-old or *more than* 100-years-old increase the model ignition probability. This corresponds to the concept of a "bathtub curve" (Figure 43) for equipment failure rates, where young equipment tends to fail at higher rates due to defectiveness and older equipment tends to fail at higher rates due to wear- out. However, conductor age warrants additional investigation to identify the conductor age as a driver of, rather than correlated to, increased ignition probability. For example, the newest conductors may have been replacements made after the 2015-2018 ignitions, possibly as a result of fire, rather than the other way around.

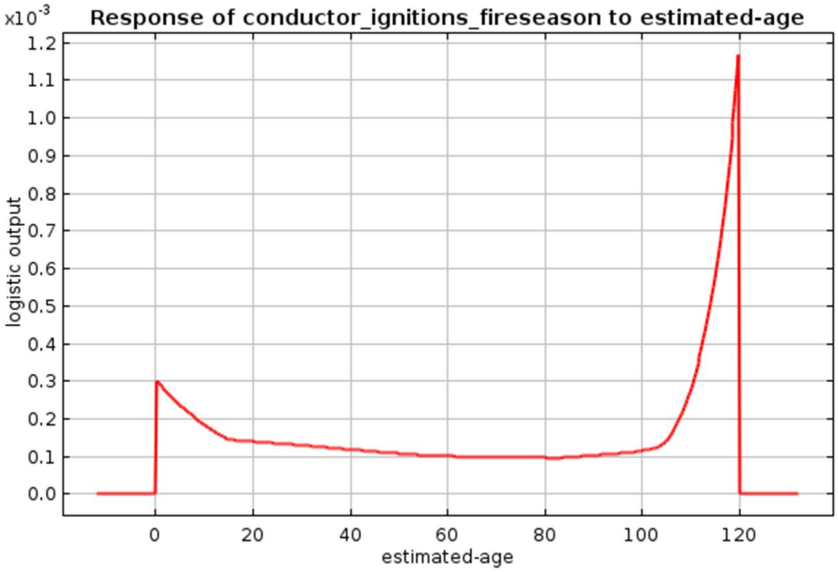


FIGURE 42 - RESPONSE CHART FROM THE MODEL OUTPUT FOR CONDUCTOR AGE, IF ONLY THAT FEATURE WERE INCLUDED IN THE MODEL

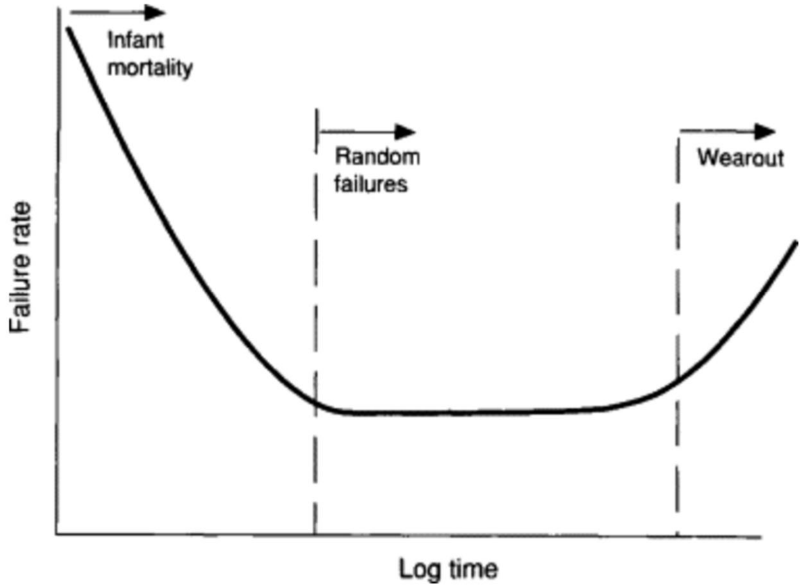


FIGURE 43 - EXAMPLE BATHTUB CURVE FOR ELECTRICAL EQUIPMENT FAILURE RATES OVER TIME (OHRING, 1995).

20.3.3.2 Conductor Material

The type of conductor material was split into one-hot encoded or dummy variables, which identified conductor materials aluminum (Al), copper (Cu), and ACSR (“conductor-material-al", “conductor-material-cu", and “conductor-material-acsr", respectively) as binary model features. As shown in Figure 41, SMEs have identified that the Cu conductor material is correlated to elevated wire-down rates compared to ACSR and Al. The response chart for the ACSR feature shown in Figure 44 supports that conductors *not* of ACSR material increased the ignition probability. However, the response charts also show that the Al feature slightly increased the ignition probability and the Cu feature had minimal effect on the ignition probability. See model limitation sections.

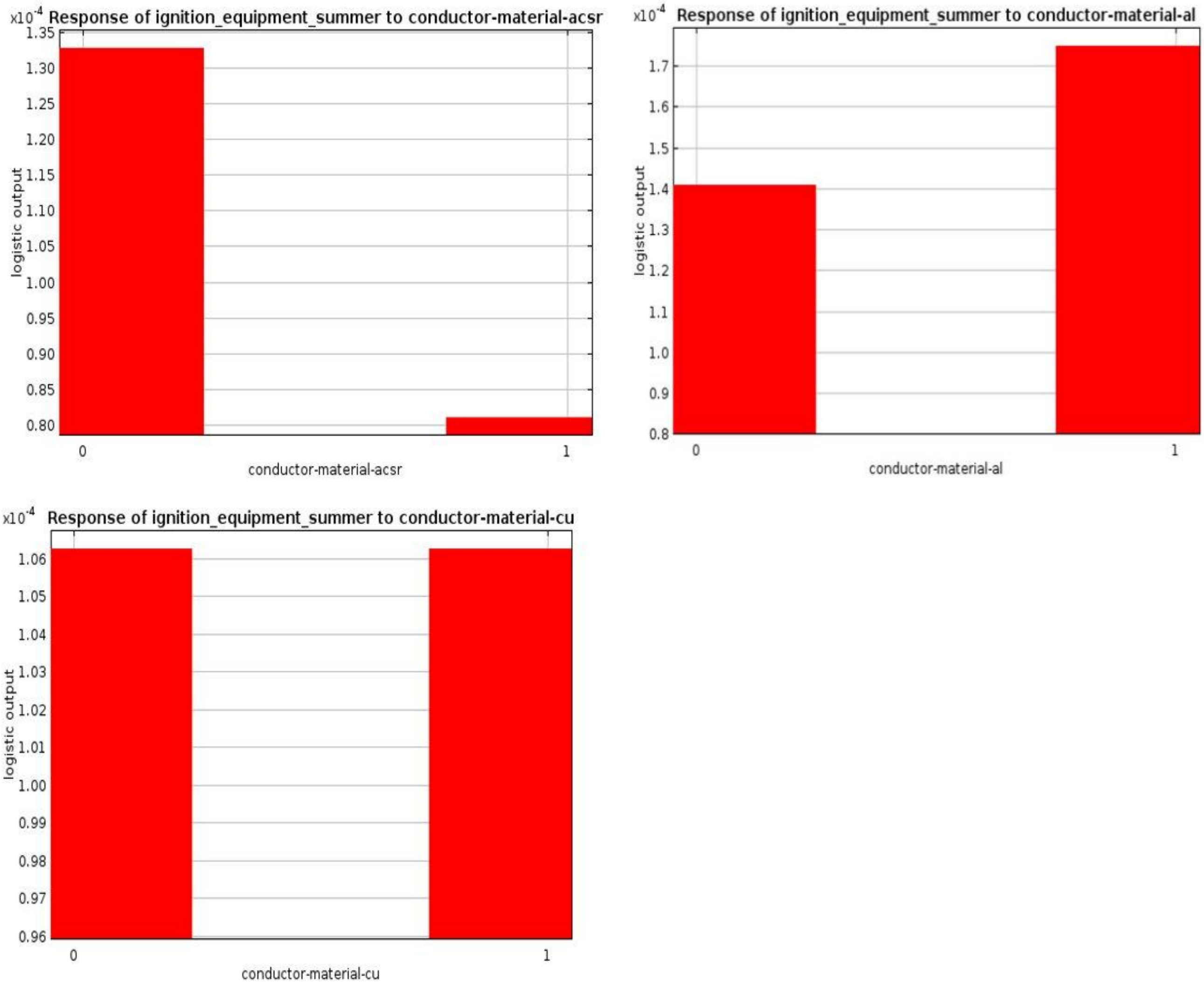


FIGURE 44 - RESPONSE CURVES FROM THE MODEL OUTPUT FOR ACSR, AL, AND CU, IF ONLY THAT CONDUCTOR MATERIAL FEATURE WERE INCLUDED IN THE MODEL

20.3.3.3 Conductor Size

The conductor size dataset was split into one-hot encoded dummy variables, which identified conductor size 2, 4, and 6 (“conductor-size-2", “conductor-size-4", and “conductor-size-6", respectively) as binary model features. Conductor sizes are defined using the American Wire Gauge (AWG) standardized wire gauge system (Figure 45) and smaller cross-sectional areas are associated with larger size numbers (in other words, the smaller the gauge number, the thicker the wire).

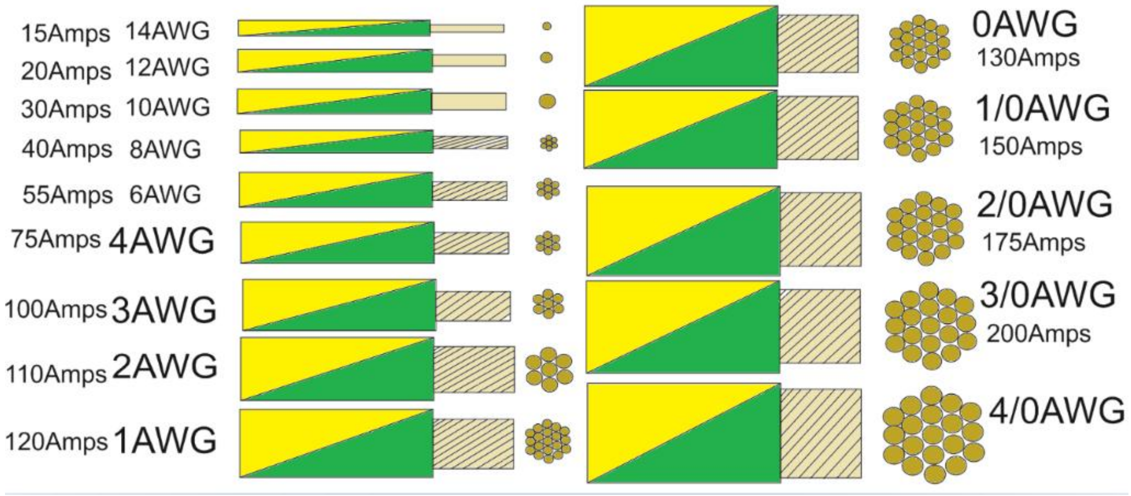


FIGURE 45 - WIRE SIZES AND TYPICAL MAXIMUM AMPERAGE – SOURCE HTTPS://COMMONS.WIKIMEDIA.ORG/WIKI/FILE:WIRE_SIZE_GROUNDING_CONDUCTORS.PNG

As shown in Figure 41, SMEs have identified that smaller diameter conductor sizes (size 6) are correlated to elevated wire-down rates compared the larger diameter sizes (size 2 and size 4). The response charts shown in Figure 46 support that conductors with smaller diameters (i.e. conductor size 6, conductors *not* of size 2, and conductors *not* of size 4) increased the ignition probability.

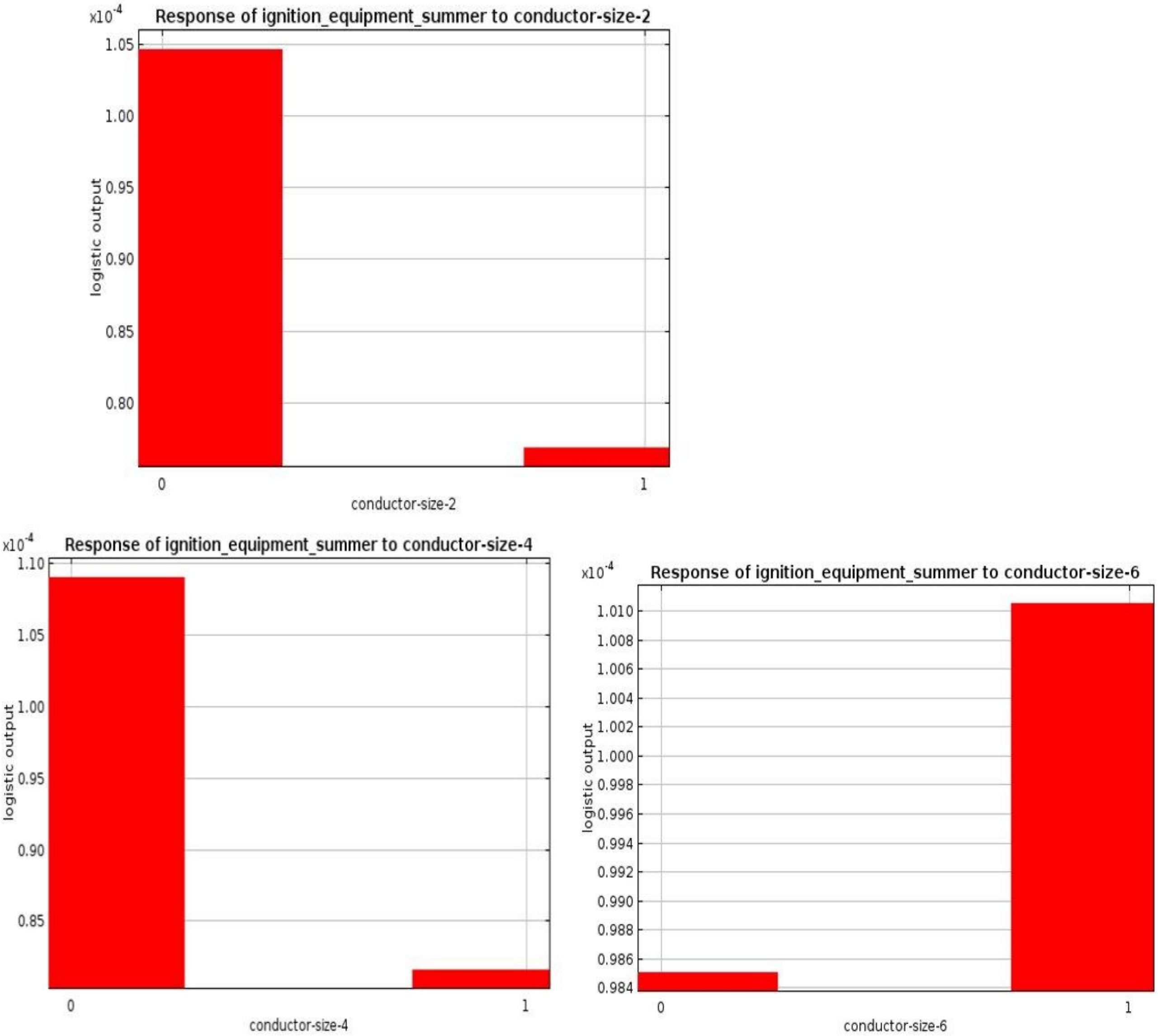


FIGURE 46 - RESPONSE CURVES FROM THE MODEL OUTPUT FOR CONDUCTOR SIZE 2, 4, AND 6, IF ONLY THAT CONDUCTOR SIZE FEATURE WERE INCLUDED IN THE MODEL

20.3.3.4 Splices

Splices were identified from the splices database table (Emili Scaief, 2020). In order to prevent splice locations from introducing bias to the model, only the Reliability Program splice records were used, which only included spans with more than three per phase. Other splice record programs focused specifically on fire areas or outage events, which cannot be used as an input to a model predicting where these events occur. The splices database only includes the latitude and longitude of the splice and the circuit name. Splices were mapped to conductors using a spatial join and then validated using the circuit name. Nearly half of the 50,000 splice records could not be spatially mapped to a conductor and more than

3,500 splice records were removed from the dataset because they did not pass the circuit name data validation step when mapped to a conductor, meaning the circuit name in the splices table did not match the circuit name in EDGIS.

A binary feature was created (called "splice-record-exists") based on whether a splice record existed for a conductor. As shown in the response chart in Figure 47, conductors *without* a splice record were associated with increased ignition probability. For example, the mapping issues described above could have biased the data. Also, the vintage of the splice data is 2020, so lines that were replaced between 2015 and 2020 would register as not having splices even if they historically had them. Alternately, the splices feature could be helping the model differentiate 3-phase spans from 1- and 2-phase spans, where ignitions are more likely to occur on 1- and 2-phase spans. In other words, the model is not identifying splices as a driver of increased ignition probability, as would be expected. This is not surprising given the small coverage of recorded splices across the distribution grid. Obtaining a different source of data for location and number of splices per conductor may improve how the model uses splices to predict ignition probability.

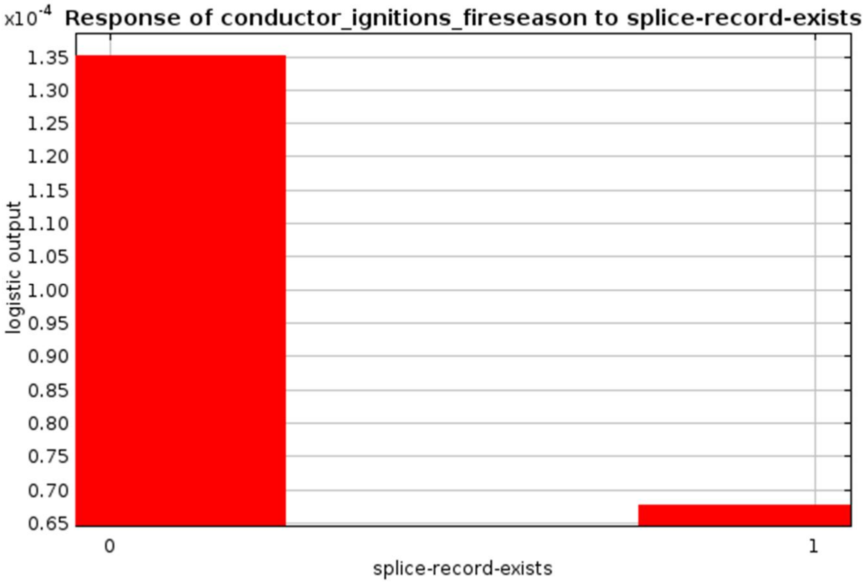


FIGURE 47 - RESPONSE CHART FROM THE MODEL OUTPUT FOR SPLICE RECORD IF ONLY THAT FEATURE WERE INCLUDED IN THE MODEL

The total splice count was also tested as a feature in the model and removed because it did not contribute to the prediction power.

20.3.4 Feature Correlation

The correlation between features is displayed as a heatmap in Figure 48. Strong correlation is observed between conductor_material_cu and conductor_size_6 as well as conductor_material_acsr and conductor_material_cu. However, these features are important to include in the model in order to consider how the mitigation of replacing these conductors affects the likelihood of ignition. As shown in the 'Permutation Importance' scores in Table 5, the model automatically deprioritized the conductor_material_cu feature, and the correlation of this feature to the other conductor attributes is planned to be addressed as the modeling matures in future iterations.

There are also strong correlations observed between the following meteorological variables, which are all characteristics of the dryness of the climate where a conductor is located:

- specific humidity (specific_humidity_ave),

- 100-hour dead fuel moisture content (100_hour_fuels_ave),
- vapor pressure deficit (vapor_pressure_deficit_ave), and
- a coastal indicator (coastal).

The Maxent program automatically selects the best features to use when optimizing the model for regularized training gain, and the coastal indicator and specific humidity features were deprioritized in the model, as demonstrated by the 'Permutation Importance' scores in Table 5.

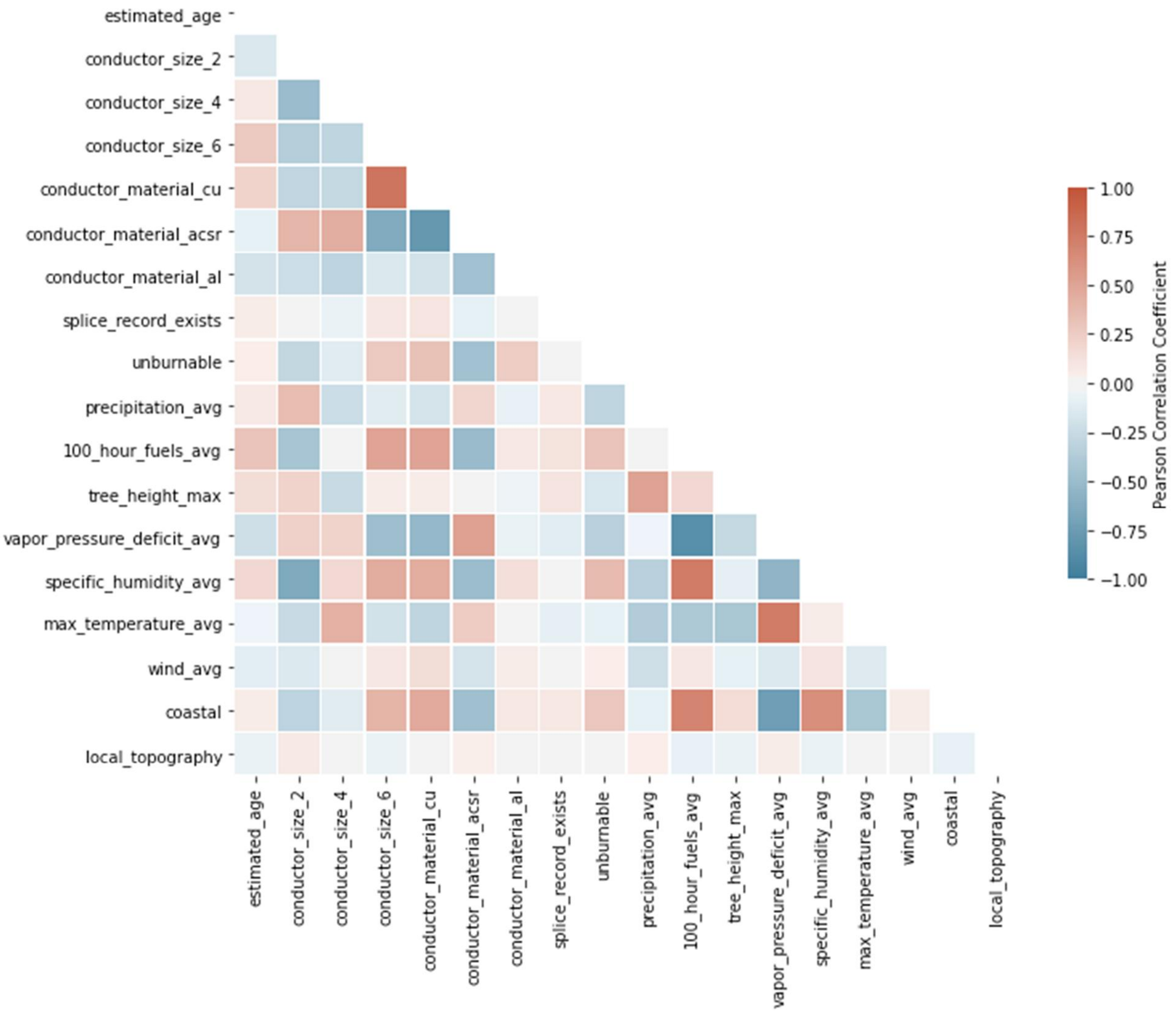


FIGURE 48 - HEATMAP OF PEARSON CORRELATION COEFFICIENT BETWEEN EACH FEATURE

20.4 Model Evaluation

20.4.1 Feature Importance

Feature importance scores are output by the model and included in Table 5. The features are listed in the table by rank according to their permutation importance score. A jackknife chart demonstrating regularized training gain for each feature is shown as Figure 49. As demonstrated in the table and figure, the unburnable feature and mean daily precipitation feature were the primary drivers for ignition prediction. The unburnable feature gives the model information about where a spark is not likely to lead to an ignition due to the type of ground surface (e.g. a paved surface). The precipitation feature shows correlation to the maximum tree height feature and may be an indicator of where trees are located that can fall into a conductor, but this would require further investigation to confirm.

Secondary drivers of prediction were the conductor attributes: conductor material, estimated conductor age, conductor size, and splices as well as the maximum tree height and mean daily vapor pressure deficit. As discussed in Section 20.3.3, the conductor characteristics give information about the vulnerability of a conductor to failure. The tree height gives the model information about where vegetation can fall into a conductor. The vapor pressure deficit gives the model information about how dry the climate is surrounding a conductor.

TABLE 5. RANKED FEATURE IMPORTANCE SCORES

Rank	Model Feature	Feature Description	Units	Permutation Importance
1	unburnable	non-burnable area	%	30.8
2	precipitation_ave	daily precipitation, mean	mm	29.8
3	conductor_material_acsr	conductor material: ACSR	%	9.7
4	estimated_age	estimated conductor age	years	8.9
5	tree_height_max	max tree height	m	4.3
6	splice_record_exists	Reliability Program splice	%	4.3
7	vapor_pressure deficit_ave	vapor pressure deficit, mean	kPa	4.0
8	conductor_size_2	conductor size: 2	%	3.4
9	conductor_size_4	conductor size: 4	%	1.6
10	100_hour_fuels_ave	100-hour fuel moisture, mean	%	1.1
11	max_temperature_ave	max temperature, mean	K	1.0
12	wind_ave	wind speed, mean	m/s	0.9
13	local_topography	TPI	%	0.2
14	conductor_size_6	conductor size: 6	%	0.1
15	conductor_material_al	conductor material: Al	%	~0
16	conductor_material_cu	conductor material: Cu	%	~0
17	coastal	coastal	%	~0
18	specific_humidity_ave	specific humidity, mean	%	~0

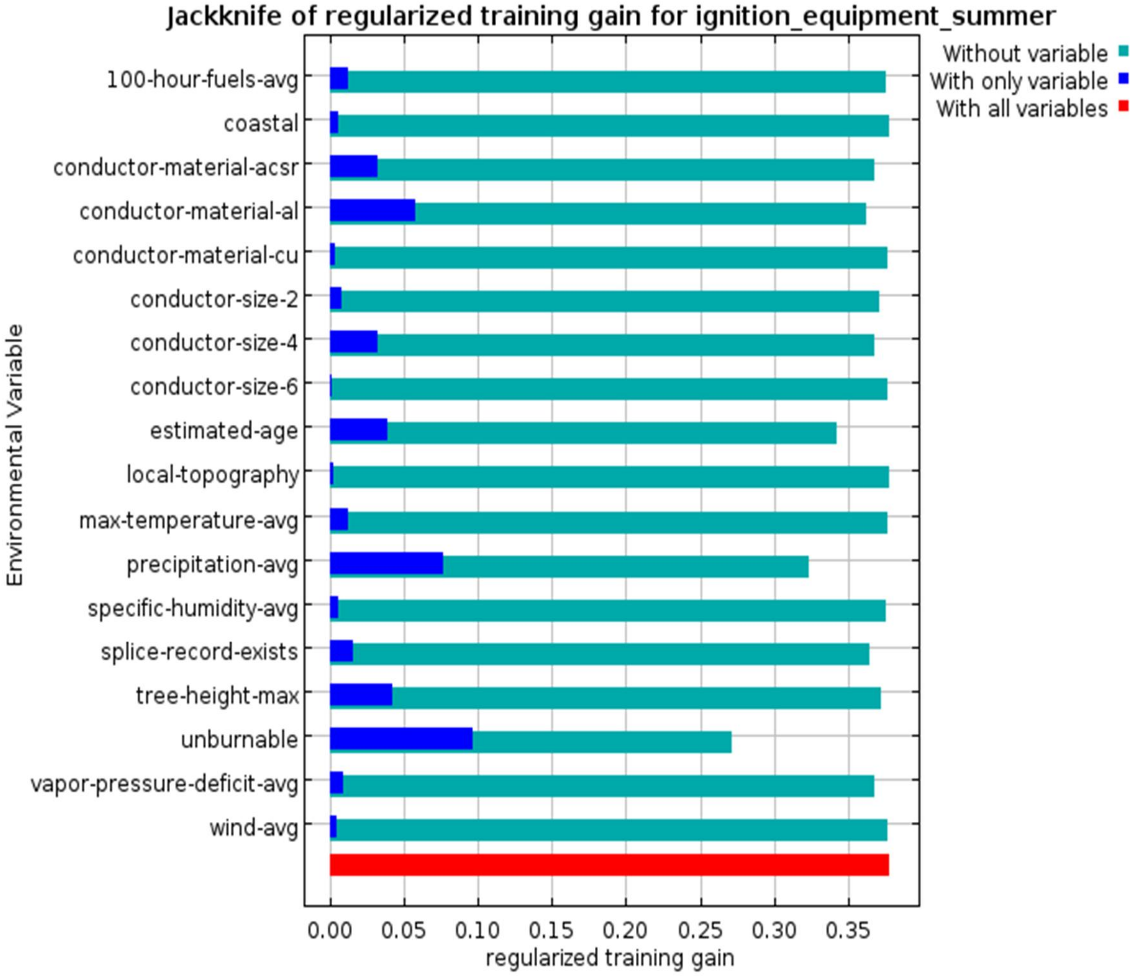


FIGURE 49 - MODEL OUTPUT OF A JACKKNIFE CHART OF REGULARIZED TRAINING GAIN BY FEATURE

20.4.2 Metrics

The primary method used to evaluate the goodness-of-fit was the ROC-AUC metric, which is described in detail in Appendix 3: Ignition Probabilities Methods 2021.

The dataset used to train the model achieved an AUC score of 0.76 (Figure 50). As displayed in Figure 51, the 2019 dataset was used as a test dataset to evaluate the model fit and achieved a score of 0.74. The minimal reduction in AUC score between the training and testing datasets gives confidence that the model is not overfitting to the training dataset and is able to maintain performance when introduced to new data.

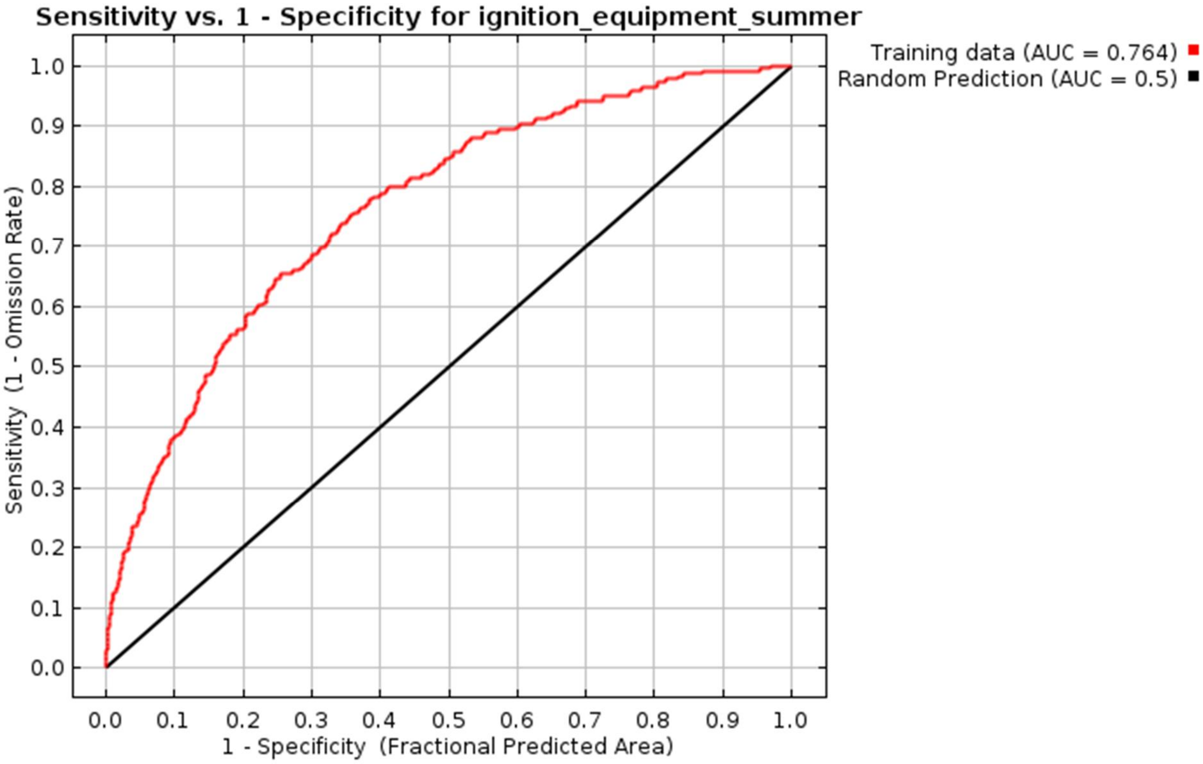


FIGURE 50 - THE ROC CURVE FOR THE PIXEL TRAINING DATASET

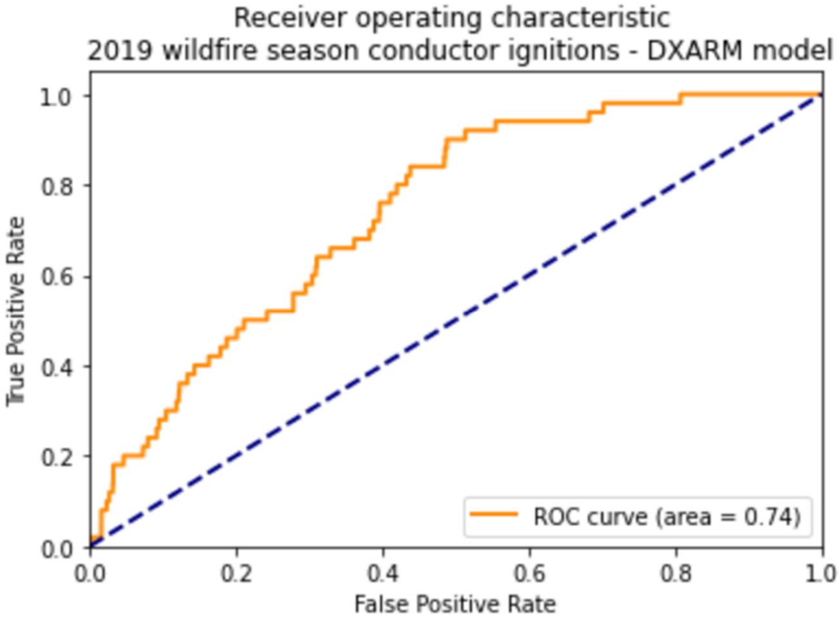


FIGURE 51 - THE ROC CURVE FOR THE TEST DATASET (2019 CPUC-REPORTABLE IGNITIONS)

21 Methodology

The following sections describe the methodology used to predict the likelihood of an ignition, the consequence dataset used, and the calculation of risk.

21.1 Circuit Protection Zones

Considering the topology of the grid was important when segmenting the distribution system into units of work for the mitigation planning process. CPZs were selected as the appropriate segmentation of the grid to report risk results because they are the most granular scale at which outages are reliably captured by the system protective devices. Furthermore, the predecessor 2018 model utilized CPZs as the smallest aggregation of risk, so for comparison purposes, the 2021 model utilized the same CPZ aggregation technique.

For the Conductor Risk Model 2021 documented herein, a mid-2020 CPZ vintage was requested by stakeholders. This vintage selection caused results of this model to fairly closely align with the current (as of late-2020) as-designed configuration of the grid, but caused the 2021 model results to be challenging to compare to the predecessor 2018 model results, because CPZs had changed so much in the interim.

More information about CPZs, vintages, and related challenges and limitations is in section 31.

21.2 Estimating Probability using MaxEnt

This model was fit using a presence-only maximum entropy (MaxEnt) algorithm, which is a supervised machine learning algorithm that requires learning from historical events (i.e. “training” the model). The model was trained on four wildfire seasons of ignitions, 2015 through 2018 inclusive. Wildfire season was defined as Jun. 1 through Nov. 30. The algorithm essentially assigns similar probabilities of events to locations with similar conditions (or features). In other words, to predict conductor-involved ignitions along the distribution grid, the model reasonably assumes that such events are likely to occur in locations with conditions that are similar to those where past events occurred.

Comprehensive details on the modeling approach can be found in Appendix 3: Ignition Probabilities Methods 2021.

21.3 Consequence

Wildfire consequence estimates the resulting damage if an ignition event occurs at a specific location. For this model, the multi-attribute value function (MAVF) consequence of risk event (CoRE) dataset was used, a consequence dataset provided by the Enterprise and Operational Risk Management (EORM) team at PG&E that combines safety, financial, and reliability types of damages. More information about MAVF CoRE consequence is included in Appendix 4: Ignition Consequence Methods 2021.

21.4 Risk

Risk is calculated by combining the likelihood of an event occurring with the consequence of that event (i.e. probability * consequence). For the current model, risk is calculated at a scale of 100m x 100m pixels associated with primary overhead conductors across the distribution system. Then, the pixel risk is aggregated for each CPZ by calculating the mean and sum of the pixel risk values (The advantage of the mean aggregation technique it is less influenced by the number of pixels in, or the size of, the CPZ. The advantage of the summed aggregation technique is that it captures all risk within a CPZ.)

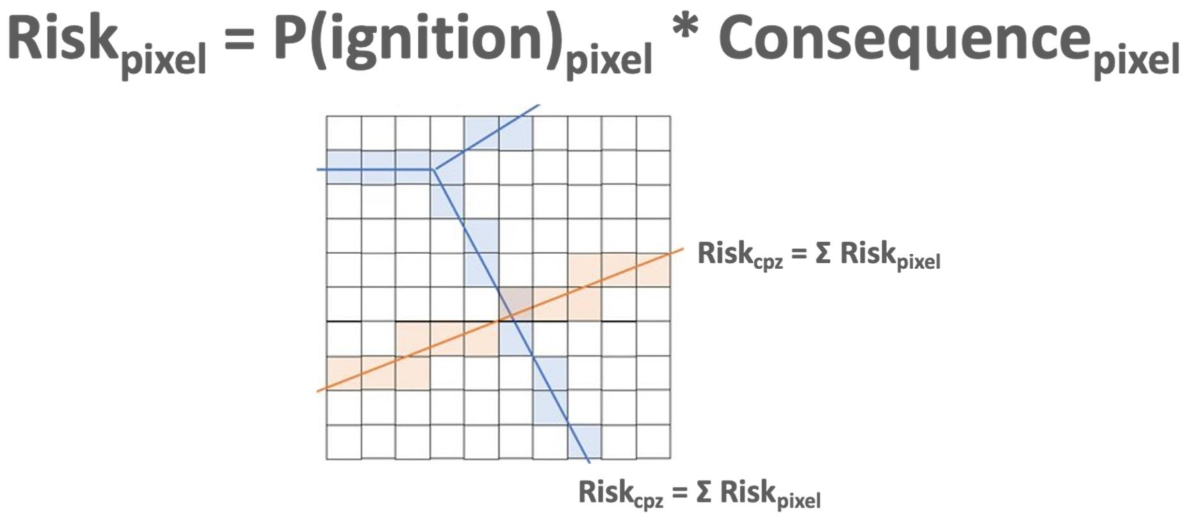

$\text{Risk}_{\text{pixel}} = \text{P(ignition)}_{\text{pixel}} * \text{Consequence}_{\text{pixel}}$
$\text{Risk}_{\text{cpz}} = \Sigma\ \text{Risk}_{\text{pixel}}$
$\text{Risk}_{\text{cpz}} = \Sigma\ \text{Risk}_{\text{pixel}}$

FIGURE 52 A DIAGRAM OF HOW PIXELS ARE AGGREGATED TO CPZS

22 Risk Results

22.1 Pixel-level Results

Risk results were calculated at the 100m x 100m pixel-level and are summarized in *Table 6* below. The probability, consequence, and risk values are all unitless and follow a lognormal distribution with many low values and a few very high values. As listed in the table, the ignition probability values sum to 60, meaning that the model expects 60 conductor-related reportable ignitions to occur in a given year within the HFTD Tier 2 & 3 areas. This value is an external calibration input to the model that was calculated using the mean number of ignitions for the 4-years of training data. The (arbitrary units) consequence values sum to over 730 million, and future iterations of the model will scale the consequence values so that the consequence dataset is calibrated to the per-event risk values for the relevant tranche type of ignitions modeled (i.e. conductor involved) reported through documents like the WMP. This will allow for the consequence and risk values to align between different wildfire risk models and higher-level reported enterprise risk values.

TABLE 6. SUMMARY OF PIXEL-LEVEL PROBABILITY, CONSEQUENCE, AND RISK VALUES

Descriptive Statistics	Probability	Consequence	Risk
mean	0.992 x 10-4	1,198.79	0.0858
standard deviation	0.970 x 10-4	2,229.97	0.1975
minimum	0.004 x 10-4	0.06	2 x 10-8
25% quartile	0.461 x 10-4	0.07	6 x 10-6
50% quartile	0.718 x 10-4	23.93	0.0022
75% quartile	1.190 x 10-4	1,188.99	0.0828
maximum	28.778 x 10-4	10,554.31	7.2126
sum	60	730,438,270	52,317

Figure 53 shows that probably and consequence are not positively correlated. Locations with elevated likelihood of ignition typically have a small consequence value. This makes mitigation work prioritization more difficult because there are not a clear cluster of locations with high consequence and high probability. The figure also shows that the highest probability values *do not* have an elevated risk value while the highest consequence values *do* have an elevated risk value (risk is shown in the figure with shading). This is due to the scale of the consequence values (0.06 to 10,554) compared to the scale of the probability values ($0.004*10^{-4}$ to $28.8*10^{-4}$), which cause the consequence values to dominate the risk result. The dominance of consequence can further be demonstrated in the comparison of the ignition probability image and the ignition risk image in the Sonoma area (Figure 54). The locations with a higher likelihood of ignition in the probability image are shown as lower risk areas in the risk image. It may be beneficial to scale the consequence values to gain more influence from the likelihood of an ignition event occurring.

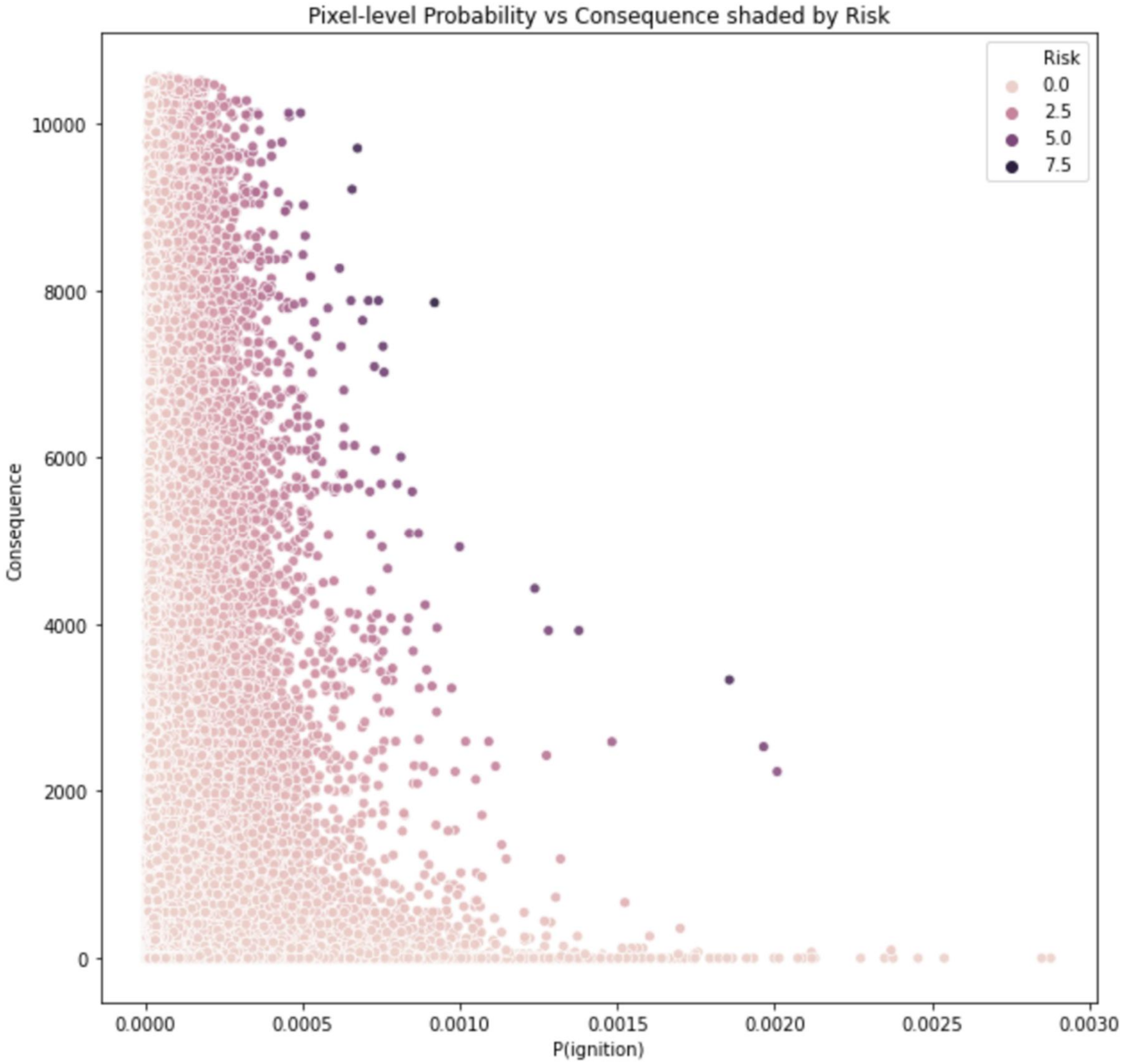
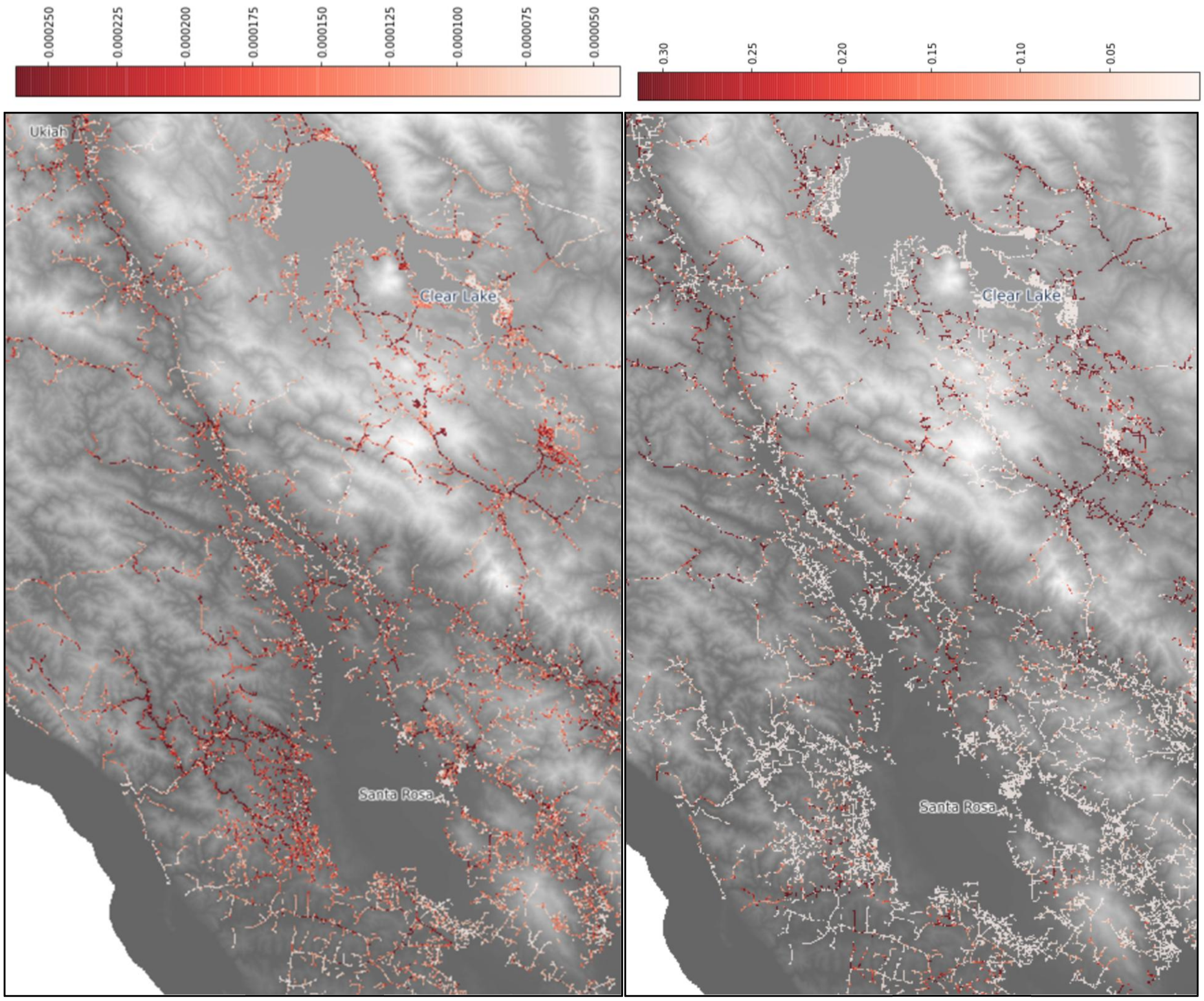


Figure 53 - Scatterplot of the pixel-level probability of ignition on the x-axis and MAVF consequence on the y-axis with shading based on the MAVF risk values



Probability

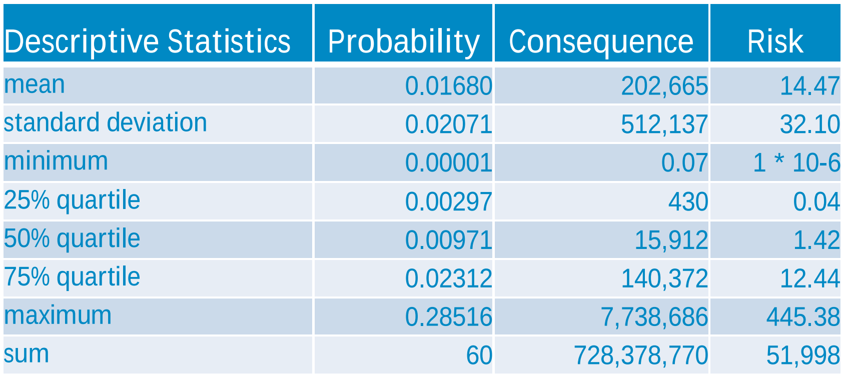
Risk

FIGURE 54 - COMPARISON OF PROBABILITY AND RISK PIXEL-LEVEL RESULTS IN THE SONOMA AREA. OBSERVE THAT THE AREAS WITH A HIGHER LIKELIHOOD OF IGNITION IN THE PROBABILITY IMAGE (LEFT) ARE SHOWN AS LOWER RISK AREAS IN THE RISK IMAGE (RIGHT). THIS IS THE INFLUENCE FROM THE CONSEQUENCE DATASET.

22.2 CPZ-level Results

The pixel values were aggregated to CPZs and results delineated by CPZ are summarized in *Table 7* below. A spreadsheet of results for each CPZ was developed to enable stakeholders to utilize probability and risk results in the planning process.

TABLE 7. SUMMARY OF CPZ-LEVEL PROBABILITY, CONSEQUENCE, AND RISK VALUES



Descriptive Statistics	Probability	Consequence	Risk
mean	0.01680	202,665	14.47
standard deviation	0.02071	512,137	32.10
minimum	0.00001	0.07	1 * 10-6
25% quartile	0.00297	430	0.04
50% quartile	0.00971	15,912	1.42
75% quartile	0.02312	140,372	12.44
maximum	0.28516	7,738,686	445.38
sum	60	728,378,770	51,998

23 Validation

Risk results were reviewed by SMEs within PG&E with expertise in asset management and wildfire risk analysis to verify that locations the model identified as high risk were locations that SMEs intuitively believed to be at high wildfire risk. Feedback from these SMEs resulted in the selection of a new consequence dataset and updates to improve the delineation of CPZs within the HFTD areas. The pixel-level results were imported into the Google Earth platform to enable a desktop analysis review of the pixel-level results along a CPZ. A field assessment to verify model results along two high risk CPZs was also conducted.

A governance committee was established to verify that this 2021 model was a continuous improvement when compared to the previous 2018 model used for system hardening planning. The committee approved the 2021 model.

24 Future Improvements

In order to get past some of the limitations the model currently has and to improve the model performance, the following potential improvements have been identified:

1. Develop a composite probability model that can assign risk reduction impacts to various types of mitigation work.
2. Investigate whether the splice locations collected from the overhead inspections improve how the model uses the splices feature.
3. Back-test the model to see if its predictions correlate to the locations of actual line breaks and line hits (damages and hazards) over the past few years of PSPS and pre-PSPS, irrespective of ignitions.
4. Transition to PG&E internal meteorology data that covers the entire modeled time period, including testing and validation years.
5. Research and test methods for handling risk reduction due to mitigation work completed within the model algorithm.
6. Include locations where system hardening mitigation work, like covered conductors, has been completed.
7. Consider methods for improved handling of the imbalanced ignitions dataset. Potential methods include:
    a. Utilize an algorithm that predicts the probability of an ignition when an outage event occurs (i.e. P(ignition | outage)). This allows the use of outage data, which is much less imbalanced than the ignitions dataset.
    b. Experimenting more with regularization and alternative model metrics.
8. Consider how to handle the reduced ignitions due to PSPS events. Potential methods include:
    a. Consider PSPS as a type of mitigation
    b. Incorporate the identified failures during post-PSPS as outage failures in a model that predicts ignitions given an outage.
9. Test an algorithm that can handle time-series data and, if the imbalance problem can be remediated, model at finer time scales (e.g. quarters or months).
10. Test an algorithm that can better handle missing data, such as a tree-based algorithm

25 References and Data Sources

CPUC. (2014, February 5). *Decision Adopting Regulations to Reduce the Fire Hazards Associated with Overhead Electric Utility Facilities and Aerial Communications Facilities.* Retrieved from https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M087/K892/87892306.PDF

Matheny, N., & Clark, J. R. (2009). Tree risk assessment: What we know (and what we don't). *Arborist News 18(1)*, 12-19.

Ohring, M. (1995). Failure and Reliability of Electronic Materials and Devices. *Engineering Materials Science, https://www.sciencedirect.com/science/article/pii/B9780125249959500398*, 747-788.

Pacific North West Tree Failure Database. (n.d.). Retrieved from https://www.arcgis.com/apps/opsdashboard/index.html#/23d3d47df6ee46d1b1f34e2910e467dc

PG&E Digital Catalyst. (2019). *STAR: The System Tool for Asset Risk - Primary Overhead Conductors.* Retrieved from PG&E Wiki: https://wiki.comp.pge.com/display/SW/Primary+Overhead+Conductors#PrimaryOverheadConductors-EDGISConductorDataset

PG&E Electric Operations. (2020). *2021 Asset Management Plans.*

PG&E EORM. (2020). *Methods - Spatial Wildfire Consequence 2021.*

PG&E EORM. (2020). *Spatial Wildfire Consequence 2021 - Lunch n' Learn presented 2020_10_16.*

PG&E Risk and Data Analytics. (2020). *Conductor Risk Model 2021 - Lunch n' Learn presented 2020_10_28.*

PG&E Risk and Data Analytics. (2020). *EVM Risk Model 2021 - Lunch n' Learn presented 2020_10_21.*

PG&E Risk and Data Analytics. (2020). *EVM Risk Model 2021 – Utility Analytics conference presented 2020_10_29.*

PG&E Risk and Data Analytics. (2020). *Methods - Ignition Probability Modeling 2021.*

primer on trees interacting with wind. (n.d.). Retrieved from https://ucanr.edu/sites/treefail/files/204521.pdf

Smiley, E. T., Matheny, N., & Lilly, S. (2012). Qualitative Tree Risk Assessment. *Arborist News 21(1)*, 1-20.

The Nature Conservancy. (n.d.). *Topographic Position and Landforms Analysis.* Retrieved from http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf

University of California, Merced. (n.d.). *GRIDMET: University of Idaho Gridded Surface Meteorlogical Dataset.* Retrieved from Earth Engine Data Catalog: https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET#description

USGS. (2016). *LANDFIRE dataset.* Retrieved from https://landfire.cr.usgs.gov/distmeta/servlet/gov.usgs.edc.MetaBuilder?TYPE=HTML&DATASET=FBK.

Western Tree Failure Database. (n.d.). Retrieved from https://ucanr.edu/sites/treefail/CTFRP_Statistics/50_or_more_753/

Appendix 3: Ignition Probabilities Methods 2021

26 Executive Summary / Overview

26.1 Section Usage

This section was written to explain the technical motivation for, and methods involved in applying the principle of maximum entropy to spatial probabilities of grid-caused fire-season wildfire ignitions. This approach was used to estimate ignition probabilities for both the 2021 Vegetation-caused risk model and the 2021 Conductor-involved risk model. Instead of repeating the technical motivation and methods of training such models in the documentation specific to each model, this section has been separated out to explain the maximum entropy approach in technical detail. It should be read as a technical companion to the risk modeling sections but is not required to read and understand them.

The other component of wildfire risk is the expected consequence of an ignition, given that it is ignited under dangerous fire conditions. The technical motivation and methods associated with the derivation of ignition consequence, also known as the MAVF CoRE values, are detailed in a separate methods section that is also intended to serve as an optional technical companion to the 2021 EVM and Conductor risk models.

27 Introduction

In support of risk-based Electric Operations planning, PG&E has developed distribution asset risk models, designed to quantify wildfire risks from the distribution system at planning and situational awareness timescales, support risk-based decision making, and enable reporting of risk reduction activities to regulators and the public. To do this, PG&E characterizes wildfire risk as *risk=ignition probability x wildfire consequence*. Both the likelihood and the consequences of an ignition are conditioned, to a degree, on the environmental factors (for example wind and gust speeds, temperature, vegetation structure, and topography) experienced by distribution assets and their age and other physical characteristics. To-date, multiple teams within PG&E have characterized the roles of specific environmental conditions that precede ignitions. For example, the meteorological team has developed fire weather indices, and the vegetation management team has identified the locations of hazard trees. This work seeks to build on that understanding to rigorously quantify the degree to which multiple environmental and asset covariates interact to determine the probability of ignitions at both fine and system levels.

To answer the question of *where* ignition events are likely to occur, we have estimated fire season ignition probabilities using maximum entropy models (MaxEnt) pioneered in the modeling of ecological ranges of species. These models are trained on ignition (or outage) locations and gridded spatial (raster) environmental and asset attribute data. The data can draw from a specific time period, but the model itself is dedicated to spatial, not temporal, patterns. The Maxent model provides relative scores or, if properly calibrated, probabilities for fire-season ignitions per “pixel” of input data.

The principle of maximum entropy (MaxEnt) refers to the expectation that all things being equal, systems will tend to be found with the macroscopic properties that can arise from the greatest number of underlying micro-scale configurations - a system’s information entropy is closely related to the number of micro-scale configurations. That is to say that, information entropy is higher when a system’s macro-state is consistent with a higher number of micro-scale states or configurations. In other words, the characteristics of high entropy systems can be captured by relatively few statistical variables. For the wildfire risk model application, the aim is to identify the simplest characterization of the envelope of environmental and asset attributes within which the maximum number of ignitions is found. Just as some models, i.e. regression, can be estimated by selecting parameters that offer the maximum likelihood solution, MaxEnt models are solved by tuning their parameters to maximize information entropy. In essence, MaxEnt applies a mathematical analog to Ocham’s razor: the least unique solution is the most likely one.

For the Wildfire Risk Model, the objective is to identify which environmental conditions and asset attributes (collectively called the model covariates) are more common among ignition locations than they are among all distribution grid locations. For example, tall trees are more common among vegetation caused ignition locations than they are among typical Distribution grid locations. Metrics of dryness, HFTD tier assignments, conductor materials and size, and others, can all be checked for such patterns. The ratio of covariate value prevalence at ignition locations to their prevalence across all grid locations is called the relative occurrence rate. MaxEnt provides a way of estimating the relative occurrence rate given a fairly modest number of ignition locations. The way it does this is to fit a statistical distribution of covariate values for ignition locations that is consistent with the values at known ignition locations, but otherwise as similar as possible to the distribution of values found everywhere else along the Distribution grid. The similarity criteria are enforced using a metric called the relative information entropy between the ignition locations and the Distribution grid locations, where the larger that metric is, the more similar the two distributions are. For this reason, the overall approach is referred to as a maximum entropy or MaxEnt estimation of the relative occurrence rate. When multiplied by the fraction of all grid locations that experience ignitions annually, the relative occurrence rate is normalized into an estimate of the annual probability an ignition will occur for all values of the covariates. This can be used to look up (aka predict) annual ignition probabilities based on the covariate values found at each Distribution grid location.

MaxEnt models have been successfully applied in ecology to the problem of estimating a species' range (i.e. the physical extent of its suitable habitat), given a set of locations where members of that species have been observed and the corresponding environmental conditions at those locations and all candidate locations for the range. In that context, the model assigns a score to every location that captures how similar the conditions at that location are to the locations where the species was observed. The correspondence between MaxEnt applied to species observations and ranges and outage/ignition locations and at-risk locations is fairly obvious – we are looking for the "range" of grid-caused wildfires - the environmental conditions and asset attributes associated with elevated wildfire probabilities. We have applied MaxEnt methods to event occurrences and their proximate asset and environmental conditions contrasted with the background conditions everywhere else along the distribution grid to identify the locations most likely to experience similar events in the future.

28 Ignition Probability Model

The objective of this work is to specify a model for characterizing the probability of an ignition, given what is known about the environment and the condition of the grid. These probabilities are to be coupled with fire spread models to describe the consequences of an ignition should it occur. The primary application for the model will be to examine how different grid hardening measures can mitigate wildfire risks that are inherent to operating the grid.

In light of these requirements, relevant models must provide several capabilities:

- Data-driven estimates describing the probability of an ignition;
- Temporally and spatially explicit estimates of when and where ignitions are likely to occur;
- Characterization of causal factors that give rise to ignition;
- Model structure suitable for inferring how mitigation measures reduce the probability that an ignition will occur.

The usage of MaxEnt addresses these criteria in the following ways:

1. MaxEnt provides spatially explicit estimates of ignition probabilities
2. Those estimates are based on a regularized fit to environmental and asset derived covariate data, meaning the contribution of each factors can be evaluated individually
3. The strength of parameter fit to each covariate allows prediction for outcomes with changed asset attributes and environmental conditions

Figure 55 below illustrates the data inputs and outputs of the MaxEnt estimation. Note: the software package used to train MaxEnt models is called Maxent (lowercase "e"), and the use of a lowercase "e" in referring to the model should invoke the specific software implementation we have adopted.

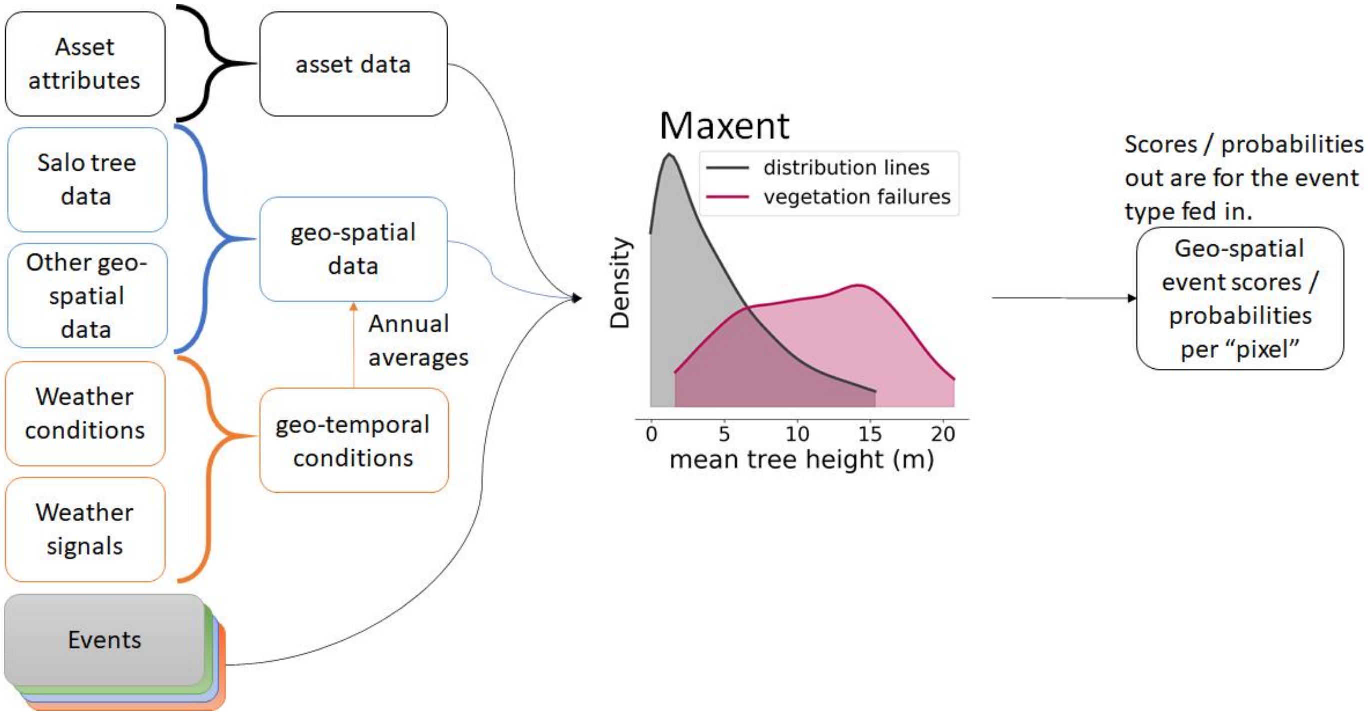


FIGURE 55 - SCHEMATIC OF DATA FLOW THROUGH THE MAXENT MODEL

28.1 What makes this a hard problem

The model described here constitutes the current state of the art in probabilistic risk modeling for wildfire mitigation. The current report provides a mathematical framework for quantifying these risks. It is well documented in the literature, however, that accurate parameterization of rare events such as wildfire ignition is a notoriously challenging problem to solve. Here, we summarize why the characteristics of the problem and the requirements of such models make this problem challenging to solve.

28.1.1 Sparse/Imbalanced Data

The tools that are most used to fit and evaluate machine learning models are predicated on the assumption that the data are sufficient to train a model. These assumptions may break down when data are inaccurate or when observations are sparse. If observations are too sparse to capture the full range of outcomes that could occur, it is possible to arrive at an interpretation of the data that is blind to these outcomes. As a corollary, the prediction of ignition vs. non-ignitions must recognize that non-ignitions are the norm on most days in most locations. In other words, you'd be right almost all the time if you predicted "no ignition" yet doing so would be entirely useless in addressing the task at hand. Any viable approach must recognize that ignitions for an imbalanced data set, where the positive class is both rare and extremely important to predict. MaxEnt models are trained based on the comparison between covariate values associated with presence data vs. covariates associated with the background of all data. As such, they can be described as presence-only models that do not require or assume balanced data sets.

28.1.2 Zero Inflation

Problems where there are substantially more observations of non-events than there are observations of events suggests that most of the data describe conditions that could be completely irrelevant to characterizing the sensitivities that are of interest. Zero inflation describes a property of certain random variables where it is not always possible for a particular outcome to occur. For example, wildfire ignitions are not viable under wet conditions or on paved parking lots. A model trained on zero-inflated data can yield parameter estimates more heavily influenced by conditions of little interest than by those conditions that are indeed problematic, thereby masking sensitivities to conditions that pose the most severe risks.

To address the possibility of zero inflation, our models seek to filter out locations and conditions under which ignitions are impossible for all practical purposes. Our models are restricted to grid pixel locations within HFTDs and the ignition data is restricted to fire season only. We further condition our estimates using covariates that describe the degree to which land surface is burnable and the presence and height of nearby trees.

28.1.3 Characterizing Probabilities

Characterizing probabilities is an inherently statistical undertaking. Elevated risk offers no guarantee that a particular outcome will occur, and there are undoubtedly instances in the data where ignitions *did not* occur despite relatively higher risk. The current work describes a mathematical basis for estimating the probability that a particular outcome will occur, and documents methods for training, tuning, and testing the model using event data from the past. However, it is worthwhile to note that the true probability distributions that give rise to the data can never be exactly known and the assessment of the accuracy of spatially localized predicted probabilities to high precision would require ignition observation for all grid locations - more ignition data that accrues in decades of grid operations.

28.2 Model Framework

We used a spatially-explicit maximum entropy model, MaxEnt[9], to link statewide maps of environmental patterns to the locations of PG&E distribution assets and ignition events (Figure 56) in order to predict the probability of an ignition for each asset as conditioned by environmental conditions (Figure 57). Vegetation data and other environmental covariates were spatially matched to the distribution and ignition data and used to predict the probability of vegetation contact-driven ignitions across the distribution grid. Vegetation data were provided by Salo Sciences.

[9] Elith *et al*. 2010; estimation software available at at this link,

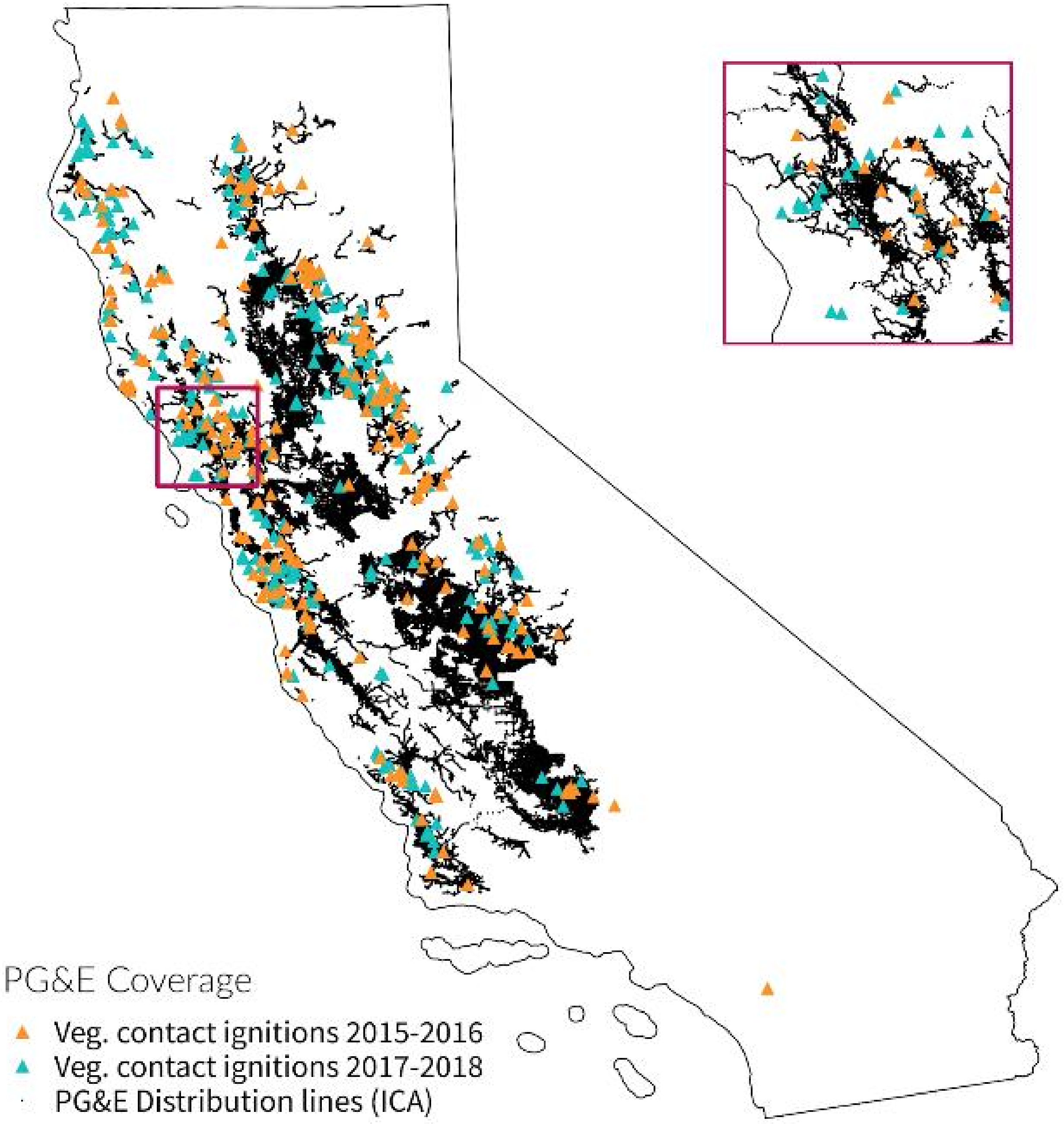

PG&E Coverage
▲ Veg. contact ignitions 2015-2016
▲ Veg. contact ignitions 2017-2018
· PG&E Distribution lines (ICA)

Figure 56 - The locations of PG&E distribution lines and ignition locations. Ignition data provided by PG&E to the CPUC, distribution data provided through PG&E

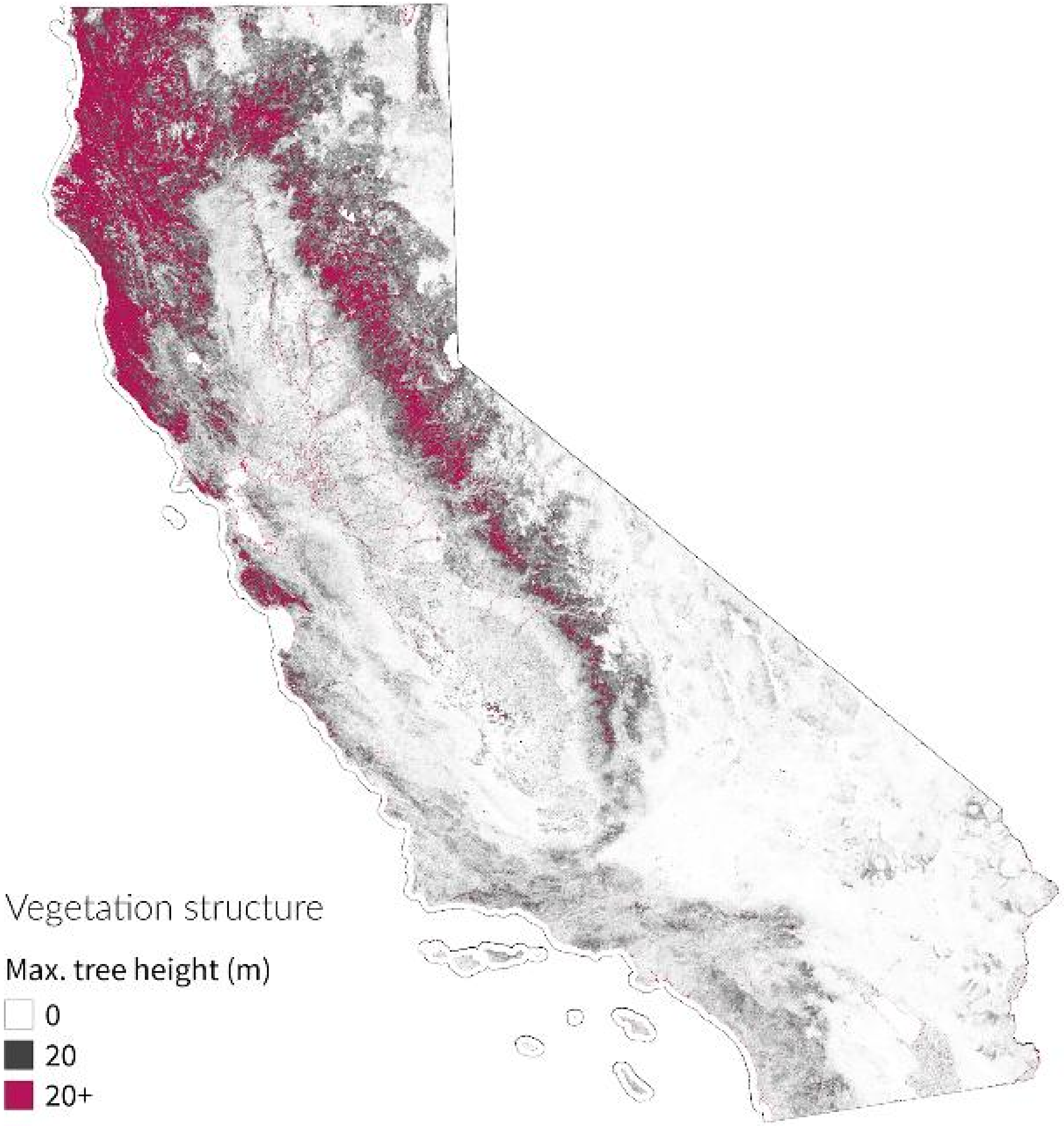


FIGURE 57 - MAXIMUM TREE HEIGHT

The guiding principle behind this approach is that, for environmentally driven failures, the probability of failure can be calculated by comparing the range of environmental conditions at failure sites to the range of environmental conditions experienced by all similar assets. This is formulated mathematically as:

$$\Pr(y=1|z)=f_1(z)\ \Pr(y=1)\ /\ f(z) \qquad (1)$$

Where *y* represents an ignition event, *y* = 1 are locations where an ignition occurred, *z* is a vector of environmental covariates, *f*(*z*) is a probability distribution of non-linear feature transformations derived from the vector of covariates across all distribution assets, $f_1(z)$ is a probability distribution of features derived from the covariates at ignition locations, and $Pr(y = 1|z)$ is the probability that an ignition occurs at a point on the landscape as conditioned by environmental conditions. In our framework, this formulates that we can calculate the probability of an environmentally-driven ignition ($Pr(y = 1|z)$) as the conditional probabilities of environmental conditions experienced at ignition locations ($f_1(z) \cdot Pr(y = 1)$) divided by the conditional probabilities of environmental conditions experienced by all distribution assets ($f(z)$). The math behind the MaxEnt model is similar to logistic regression with linear, piecewise, and interaction covariate terms, aiming to maximize the information entropy of the predicted ignition probabilities.

In plain English, MaxEnt assesses the relative probabilities that events occur across a landscape. In this case, we use point locations where ignitions occur to predict the probability of ignitions across all distribution assets. The guiding principle behind this approach is that the probability of ignition can be calculated by comparing the range of environmental conditions at failure sites to the range of environmental conditions experienced by all assets.

Distribution lines are subject to a range of environmental conditions. Wind speeds and temperatures vary across lines. Some are exposed on hilltops. Others are near vegetation. This range of variation can be described according to a statistical distribution: the majority of lines are not in proximity to vegetation, a small portion with some small trees overhanging lines, and a much smaller proportion with very tall trees overhanging. However, when we examine the subset of lines where ignitions from various causes have occurred, we find that the statistical distributions are shifted. For example, there is a higher proportion of tall trees near sites of vegetation-caused ignitions. Comparing these distributions, you can infer that assets fail more often due to vegetation contact when lines are surrounded by tall trees. For that particular cause, the model will calculate high probabilities for areas surrounded by tall trees, moderate probabilities for areas surrounded by small trees, and zero probability when there are no trees.

This spatially-explicit ignition modeling approach relies on three key datasets: 1) the geographic locations of all operating assets exposed to the conditions that cause ignitions, 2) records of where and when ignitions occurred, and 3) a set of environmental predictor data that we expect to determine the probability of ignition. In our model formulation, these can be re-framed as 1) the landscape of analysis (L), 2) locations where failures have been observed ($y = 1$), and 3) a vector of environmental covariates (z).

If we define $f(z)$ to be the background probability density of covariates across L, $f_1(z)$ to be the probability density of covariates across L where failures occurred, and $f(z)$ where failures did not occur. The quantity that we wish to estimate is the probability of failure, conditioned on environment: $Pr(y = 1|z)$. Strictly presence-only (i.e., failure-only) data only allow us to model $f_1(z)$, which on its own cannot approximate probability of presence. Presence/background data allows us to model both $f_1(z)$ and $f(z)$, and this gets to within a constant of $Pr(y = 1|z)$, as Bayes' rule gives equation (1).

28.2.1 Scaling probability scores

MaxEnt first estimates the ratio $f_1(z)/f(z)$, which is referred to as the "raw" output. This is the core calculation, giving insight about what features are important and estimating the relative likelihood of ignition in one place or another. Because prevalence data are not typically available for calculating the conditional probability of occurrence in MaxEnt models (i.e. the true population of a species whose range is being studied), MaxEnt estimates occurrence probabilities using what is called the "logistic" output. This treats the log of the output: $\eta(z) = log(f_1(z)/f(z))$ as a logit score, and calibrates the intercept so the implied probability of presence at sites with "typical" failure conditions (where $\eta(z)$ = the average value of $\eta(z)$ under f_1) is a parameter τ, as in

$$\Pr(\mathrm{y} = 1|\mathbf{z}) = \tau e^{\eta(\mathbf{z})-r}/(1-\tau + \tau e^{\eta(\mathbf{z})-r}) \qquad (2)$$

Knowing τ solves the non-identifiability of prevalence; without that, MaxEnt arbitrarily sets $\tau = 0.5$. This log transformation is monotone (order preserving) with the raw output. For our work, the true number of ignitions is well tracked, so we calculated the prevalence score by computing the average rate of failure such that the expected count of ignitions per-fire-season based on Pr(y=1|z) match the annual average observed in the underlying model training data. We refer to this normalization step as τ calibration.

28.2.2 Feature selection

MaxEnt fits models to features—an expanded set of transformations of the original covariates. Fitted functions are defined over many features, resulting in more features than covariates. There are six feature classes: linear, product, quadratic, hinge, threshold and categorical. Products include all possible pairwise combinations of covariates, fitting simple interactions. Threshold features allow "steps" in the fitted function. Hinge features allow changes in the gradient of the response. Multiple threshold or hinge features can be fit for one covariate, generating complex functions. Hinge features alone create a model akin to a GAM: an additive model with nonlinear fitted functions of varying complexity but without sudden steps from thresholds. In this analysis, we included just product and hinge features. This combination captures interaction terms between covariates (e.g., between wind speeds and tree height) and fits nonlinear functions in a piecewise but continuous fashion. The features generated through interaction and hinge functions are then evaluated during the model fitting process, which seeks to maximize regularized gain, so only the features that improve the model fit without over-fitting to training data are kept in the final model formulation.

28.2.3 Inferring Causality

Ignitions result from complex interactions between weather, ecosystems, and grid assets. These dynamics are not perfectly understood and often are monitored only indirectly. The number of splices on a particular span, for example, does not tell us about the condition that those splices are in. While we may find correlations to suggest that certain attributes of the system contribute to elevated risk, it may not be possible to infer causal relationships. This poses a challenge to quantifying risk reduction associated with mitigation measures.

The current work describes an approach that uses expert judgement to inject assumptions about how mitigation measures will alter correlations observed in the data. To develop robust statistical methods for doing "causal inference" typically requires randomized controlled trials or other managed interventions rather than the "natural experiment" of past outcomes this work is based on.

28.3 Event data "presence" observations

The pool of all ignition data is the starting point for "presence" data used to train MaxEnt. These models perform better if they are trained on data that is as specific as possible to the conditions under which you want to predict the outcomes. The ignitions are typically filtered by cause (e.g. vegetation-caused vs. equipment failure), equipment involved (e.g. conductors vs. poles), date range (e.g. 2015-2018), fire season (Jun. 1 – Nov. 30) and occurrence in HFTD tiers 2 and 3.

Filtering criteria are a result of the combination of model client needs, model limitations, and data source limitations. Client needs are things like "We need to know the risk posed by conductor failures that occur during fire season" – in which case, training data would be filtered to only those events involving conductor failures that occurred during fire season. Model limitations include things like "The model is not currently able to correctly interpret the change in outages and ignitions that occurred in 2019 and 2020 due to Public Safety Power Shutoff (PSPS) events" – in which case, training data would be filtered to exclude events from 2019 and 2020. Data source limitations are things like "Consequence data is available only for HFTD 2 and 3 locations (thus, no Risk can be calculated for non-HFTD 2 and 3 locations)" – in which case, training data would be filtered to exclude non-HFTD 2 and 3 locations.

The locations of the resulting set of filtered ignitions (constituting a specific sub-category of all ignitions) are used to lookup the values of environmental, weather, and asset data that share the same grid pixels. The distributions of those covariate values constitute the presence distribution which will be compared to the background distribution of values associated with all grid pixels. Model predictions will produce higher probabilities under conditions unique to the presence distribution when compared to the background.

28.4 Pool of covariates

The following table summarizes the pool of raster covariates developed to date for use in MaxEnt model runs.

Covariate	Category	Source	Spatial resolution	Units	Descriptions
100-hour fuels	Meterological data	gridMET	~4km	%	Unless otherwise noted, all GRIDMET data aggregated from 2014 to 2016. The dead fuel moisture data were obtained from GRIDMET, and the "100-hour-fuels" feature was included in the model. The exact GRIDMET variable use is known as fm-100, and is a standard fire modeling metric of fuel dryness for fuels about 1-3" in diameter - intermediate sized fuels.

1000-hour fuels	Meterological data	gridMET	~4km	%	fm-1000, as defined above, but for 3-8" in diameter.
burn index	Meterological data	gridMET	~4km		the US, the National Fire Danger Rating System (USNFDRS) Burning Index (BI)
energy release	Meterological data	gridMET	~4km		USNFDRS Energy Release Component (ERC)
precipitation average	Meterological data	gridMET	~4km	mm	Daily precipitation average
specific humidity	Meterological data	gridMET	~4km	kg/kg	Specific humidity
vapor pressure deficit avg	Meterological data	gridMET	~4km	kPa	Measure how much water is in the air compared to how much it could hold at the given temperature. VPD drives evapotranspiration and is the mechanism for fuels drying out during fire season.
temperature max average	Meterological data	gridMET	~4km	K	Average of daily maximum temperature in Kelvin (recall that it is sensed via satellite)
wind avg	Meterological data	RTMA	~2.5km	m/s	Hourly average wind speed at 10m, averaged from 2016 to 2018
wind max	Meterological data	RTMA	~2.5km	m/s	Annual 99th percentile hourly wind speed at 10m assessed over 2016 to 2018
windy summer day pct	Meterological data	RTMA	~2.5km		The percentage of days with sustained hourly wind speeds over 15 mph
gusty summer day pct	Meterological data	RTMA	~2.5km		The percentage of days with sustained hourly wind speeds over 20 mph
tree height max	Tree data	Salo Sciences	100m		Tree height data were obtained from a third-party vendor, Salo, and the “tree-height-max" feature was developed by calculating the maximum tree height, in meters, for each 100m x 100m pixel area along the distribution grid, according to the processed satellite data provided by Salo. The satellite imagery was collected in November 2019.
tree height average	Tree data	Salo Sciences	100m		Same as above but taking the pixel average height.
impervious	Surface condition	NLCD	100m	%	NLCD imperviousness products represent urban impervious surfaces as a percentage of developed surface over every 30-meter pixel in the United States, scaled to 100m.

unburnable	Surface condition	LANDFIRE 2016 Surface Fuels Model	100m	%	The “un-burnable” feature is a land surface descriptor similar to imperviousness that includes surfaces that typically don’t ignite when a spark occurs. The feature was derived from several land use types within the 2016 LANDFIRE surface fuel model (USGS, 2016) and is the percentage of the 100m x 100m pixel identified as un-burnable. The land use types considered “un-burnable” in the composite spatial layer include: urban, snow/ice, agriculture, water, and barren.
local topography	Surface condition	NED National Elevation Database	100m		The relative topography of the area was also used as a feature in the model. The topographic position index (TPI) was extracted from a USGS national elevation dataset (NED) at 100-meter resolution. The TPI compares the cell elevation to the mean elevation for the local neighboring area (positive values are above the mean and negative values are below the mean) (The Nature Conservancy).
hftd	HFTD	CPUC	100m		Categorical variable that is 1 for non-HFTD locations, 2 for Tier 2 and 3 for Tier 3.
Age	Asset data	EDGIS Conductors	100m		The estimated conductor age (the “estimated-age”) was calculated as the number of years since the installation year, as listed in ED-GIS. If the installation date was missing or invalid, then the estimated age in the STAR model dataset was used
Materials	Asset data	EDGIS Conductors	100m		The type of conductor material was split into one-hot encoded dummy variables, which identified conductor materials aluminum (Al), copper (Cu), and ACSR (“conductor-material-al", “conductor-material-cu", and “conductor-material-acsr", respectively) as binary model features.
Size	Asset data	EDGIS Conductors	100m		The conductor size dataset was split into one-hot encoded dummy variables, which identified conductor size 2, 4, and 6 (“conductor-size-2", “conductor-size-4", and “conductor-size-6", respectively) as binary model features. Lower numbers correspond with larger diameters.
Splice count	Asset data	EDGIS Conductors	100m		Splices were identified from the splices database table (Emili Scaief, 2020). In order to prevent splice locations from introducing bias to the model, only the Reliability Program splice records were used, which only included spans with more than three per

					phase.
Coastal indicator	Asset data	EDGIS Conductors	100m		Coastal areas were identified using a binary feature in the model. Coastal areas within PG&E service territory were mapped internally in PG&E and conductors are tagged with a coastal indicator field in ED-GIS.

29 Assessing Results

As described above, the Maxent software we use performs feature selection on covariates under a regularization protocol designed to avoid over-fitting (the condition where a model is so optimized to fit the training data that it loses predictive power). It also computes model sensitivity to random perturbations in both input data (permutation importance) and the resulting λ fit parameters (percent contribution) and quantifies the model gain via jackknifing (Figure 58) for all leave-one-out and include-only-one permutations of model covariates, capturing the individual explanatory power and the unique explanatory power of each covariate. These metrics are computed for all “official” model runs.

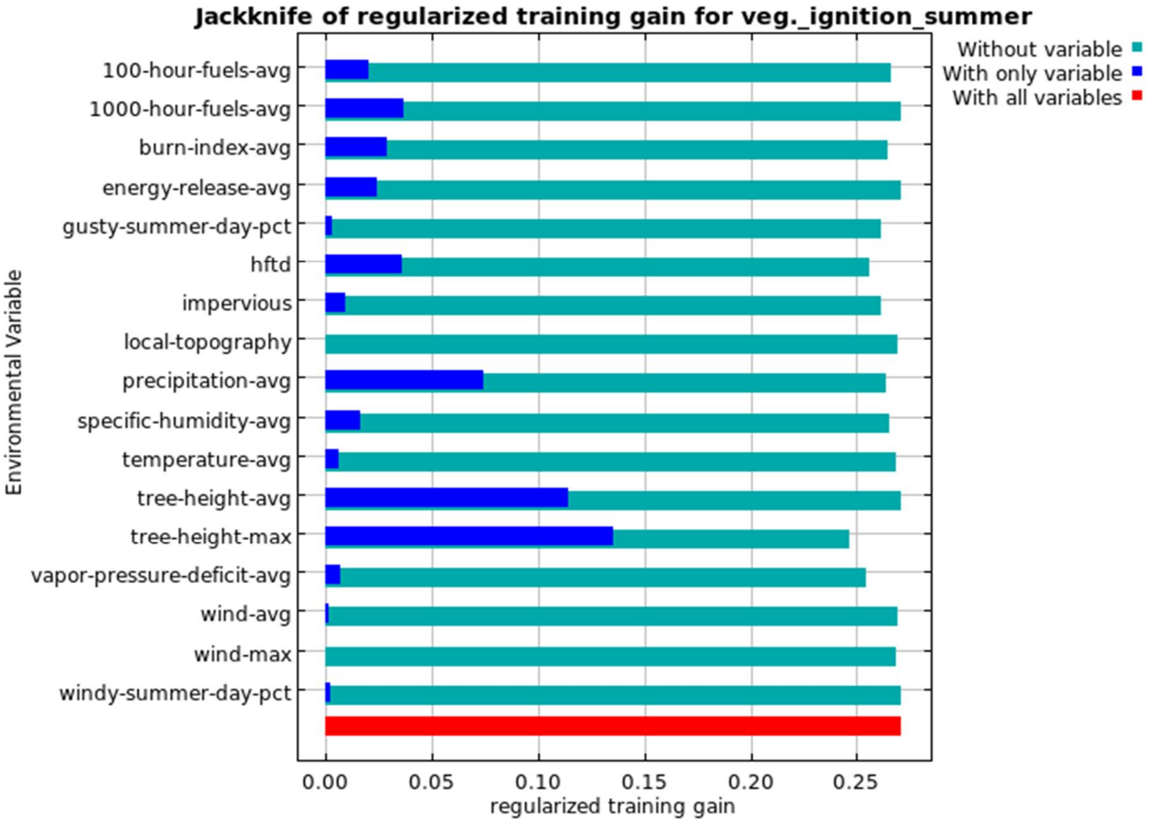


FIGURE 58 - JACKKNIFE TABLE

We also report two model performance metrics: recall scores and the area under the receiver operator curve (AUC). To compute recall scores, we set a threshold of >5 on the omission rate that maps approximately to a 95% confidence interval for predicting, in a binary sense, locations where ignitions are likely and where ignitions are unlikely. We computed recall scores as the number of ignitions in the test data located within at-risk areas (true positives) divided by the total number of ignitions in the test data (i.e., TP / (TP + FN)). If this score is near 95%, then the predicted ignitions at the omission rates accurately constrains the extent where ignitions could occur.

TABLE 1. CONFUSION MATRIX FOR THE FOUR PREDICTION OUTCOMES FROM A BINARY PREDICTION OF IGNITION LIKELIHOOD.

		Predicted to be at-risk	
		True	False
Ignition observed	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

Our second model performance metric, AUC, estimates separability. In concrete terms, the AUC is the ROC-AUC, or area under the receiver-operating curve (ROC) – see Error! Reference source not found.59. The ROC is a curve with the true positive rate on the y-axis and the false positive rate on the x-axis. Each point along the curve represents the tradeoff between making the model omission rate more generous to predict more “true positives” (higher on the y-axis) vs. having that generous omission rate falsely predict ignitions that didn’t occur (further right on the x-axis). Any given point along the ROC tells you what fraction of non-ignitions are falsely predicted as ignitions as the “cost” of achieving a given true positive rate for all true positive rates. Along the ROC curve then, predicting only non-ignitions is shown in the lower left corner to predicting only ignitions in the upper right corner. Random guessing will produce a diagonal ROC, whose area would be 0.5. A perfect model would produce an ROC that immediately rises to 100% true positive without any false positives, whose area would be 1. The AUC-ROC is this a metric between 0.5 and 1 that captures how well the model avoid false positives as it captures true positives.

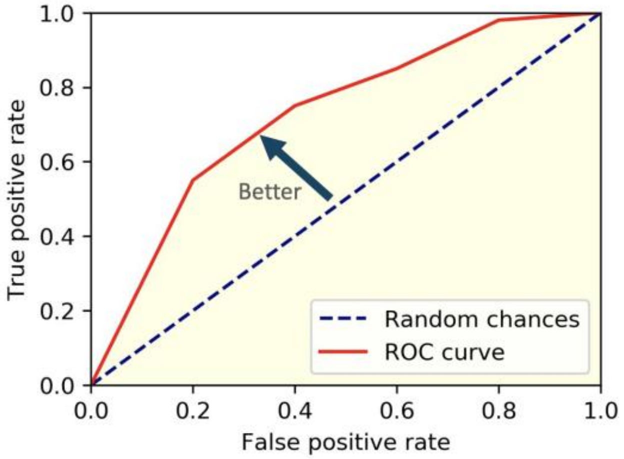


FIGURE 59 ILLUSTRATION OF A RECEIVER OPERATOR CURVE (ROC) WITH THE ROC-AUC AREA UNDER THE CURVE SHADED

In our case, an AUC score of 0.7 can be interpreted as a 70% chance that the model will be able to distinguish between where ignitions are and are not likely to occur.

Finally, we evaluate model predictive performance by splitting the ignitions data into two groups, training on one set and testing predictions on the other. Since the model is not trained on the test sample, its prediction performance on that sample are more accurate metrics of model performance in a planning context. One logical formulation for a test sample is to use one or more entire fire seasons. Our initial modeling was based on a training set of ignitions from 2015-2016 and a test set from 2017-2018. However more recently we've been using 2015-2018 to train the model and 2019 to test.

However, there is significant variability in ignition activity from year to year, so a model trained on several years' worth of data can be expected to predict for a typical year, not necessarily the test year. To isolate potential over-fitting from the question of year-over-year variability in ignitions, a purely random test sample can be taken instead. The AUC for those test predictions can be more unambiguously interpreted as a metric of how well the model will predict out of sample in general.

30 Model Limitations

MaxEnt models are structured around resolving spatial differences in the likelihood of an event occurring, like an ignition or a siting of a specific species. However, for rare events like wildfire ignitions, they require data spanning a significant passage of time to achieve statistical significance. This, in turn, means that MaxEnt results are not particularly well informed by temporally varying conditions. They tend to require pre-filtering of data to a specific set of conditions/location relevant to a given question. For example, studying just north easterly wind events would require filtering all event and time varying input data to the subset of days when such conditions occurred and then fitting a MaxEnt model using those inputs. However, it is not always possible to slice the data so finely and still maintain enough statistical power to get a predictive fit.

MaxEnt models are also tethered to providing rasterized predictions. One has to be careful when attributing the risk within a given raster pixel across several assets that are located within that pixel. Other model types are better tuned to modeling specific assets directly. However, it should be noted that outage and ignition data are not very well resolved spatially. The former is reported with the location of the protective device that triggered, the later have locations captured in the field, but often at some distance from the ignition. The result is that there are limits to how specifically either can be assigned to specific assets.

MaxEnt models are also “presence only” models. They do not require or take into account fully labeled data (i.e. where the non-ignitions are also labeled). This causes complications related to estimating ignition probabilities in that they need to be calibrated against expected ignition counts. Classification models of other types would more directly estimate probabilities. However, it is important to consider whether all ignition are actually observed and the extent to which spatial uncertainty prevents clean binary labeling at the asset level.

31 Circuit Protection Zones (CPZ)

Circuit Segments referred to as Circuit Protection Zones (CPZ) were selected as the appropriate segmentation of the grid to report risk results because they are the most granular scale at which outages are reliably captured by the system protective devices – and outages are an important factor for model training. Furthermore, the predecessor 2018 model utilized CPZs as the smallest aggregation of risk, so for comparison purposes, the 2021 model utilized a similar approach of aggregating assets and risk to CPZs.

Circuit Protection Zones (CPZs - also referred to as distribution shutoff zones and circuit segments, and related to, but different from, protection zones and zones of protection) are a term used at PG&E - but the term had not previously been clearly defined. Furthermore, no canonical source data that lists CPZs was found to be available - thus the RaDA team created a data processing pipeline that allocates distribution grid assets to CPZs, lists and assigns identifiers to CPZs, adds various metadata to the list of CPZs, and characterizes the spatial extent of each CPZ by mapping the grid assets that belong to each CPZ.

CPZs are defined here as the smallest non-overlapping sections of the distribution grid that can be de-energized by circuit breakers and line reclosers (including trip savers and fuse savers) that are typically in the closed position at the time of aggregation (see the Vintage section below). While there are numerous other types of protection and de-energization devices, including interrupters, sectionalizers etc, we do not “split” circuits into CPZs at those devices - we split only at circuit breakers and line reclosers.

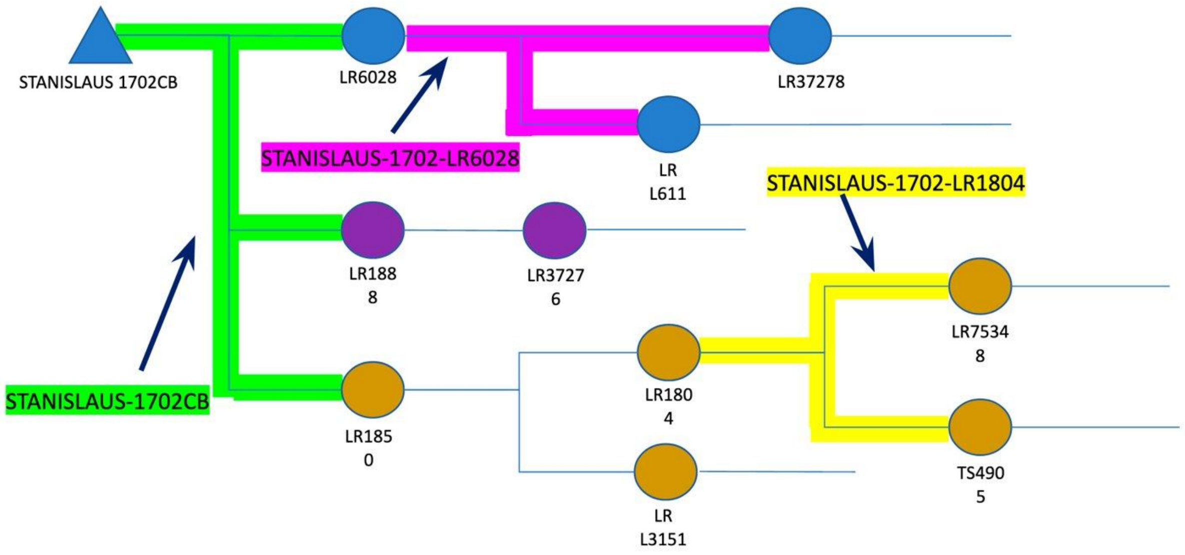


FIGURE 60 CPZ NAMING - LR = LINE RECLOSER, CB = CIRCUIT BREAKER

As of mid-2020, there were approximately 10,800 circuit breakers and line reclosers with an associated CPZ in the distribution grid. In HFTD there were approximately 3,800 circuit breakers and line reclosers with an associated CPZ.

CPZs are determined for the “normal” or “as built” configuration of the grid. Automated and manual switching may cause the grid to reconfigure temporarily or persistently, however we ignore those configurations in favor of the default configuration intended by circuit designers and as defined in our source data (trace tables mapped to asset data in EDGIS). The as-built configuration is the designed structure of the grid and is the most appropriate way to think about the grid for planning purposes over many years.

CPZs are referred to by circuit protection zone identifiers (CPZ ID). In this analysis, CPZ IDs are formed by concatenating the circuit or feeder name with the equipment identifiers of the protecting devices at the upstream point, or “start”, of each circuit protection zone. Each CPZ terminates at “downstream” protection device(s) - or, for zones that terminate only at service points (eg. houses, businesses, etc), they terminate at those service points. CPZs that start at the “start” of a circuit are named with the protection device that controls the circuit.

31.1 Vintage

CPZs are defined by the “normal” or “as built” configuration of the grid at a specific point in time. The point in time is referred to as the “vintage” of the CPZ aggregation. RaDA has created two CPZ vintages – one each for each of the EVM Risk Model 2021 and the Conductor Risk Model 2021.

An additional vintage is expected to be created for 2022 models, unless a replacement for CPZs is implemented by that time.

31.2 Challenges and Limitations

CPZs are aggregations of the grid as defined by certain protective devices. Both the grid extent, and the location of these protective devices, change over time. In particular, hundreds of new protective devices have been installed in HFTD areas in 2019 and 2020, motivated primarily by the desire to be able to shut off power to smaller sections of the grid during PSPS events.

These changes produce changes in the shapes, CPZ IDs, and asset “contents” of CPZs that make year-over-year model result comparisons challenging. For example, the Conductor Risk Model 2021 results included ~800 CPZs that had no direct equivalent in the predecessor 2018 model results. Virtually all of the distribution assets in these ~800 “new” CPZs existed previously, but the addition of new protective devices caused new CPZs to come into existence when aggregation of assets to CPZs was performed to create a new CPZ vintage for the Conductor Risk Model 2021.

CPZs are thus a sub-optimal unit of grid and risk aggregation if year-over-year comparisons are desired, as is typically the case when considering wildfire risk models. However, no suitable replacement has been proposed and accepted by the many stakeholders of this work.

32 References and Data Sources

CPUC. (2014, February 5). *Decision Adopting Regulations to Reduce the Fire Hazards Associated with Overhead Electric Utility Facilities and Aerial Communications Facilities.* Retrieved from https://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M087/K892/87892306.PDF

Matheny, N., & Clark, J. R. (2009). Tree risk assessment: What we know (and what we don't). *Arborist News 18(1)*, 12-19.

Ohring, M. (1995). Failure and Reliability of Electronic Materials and Devices. *Engineering Materials Science, https://www.sciencedirect.com/science/article/pii/B9780125249959500398*, 747-788.

Pacific North West Tree Failure Database. (n.d.). Retrieved from https://www.arcgis.com/apps/opsdashboard/index.html#/23d3d47df6ee46d1b1f34e2910e467dc

PG&E Digital Catalyst. (2019). *STAR: The System Tool for Asset Risk - Primary Overhead Conductors.* Retrieved from PG&E Wiki: https://wiki.comp.pge.com/display/SW/Primary+Overhead+Conductors#PrimaryOverheadConductors-EDGISConductorDataset

PG&E Electric Operations. (2020). *2021 Asset Management Plans.*

PG&E EORM. (2020). *Methods - Spatial Wildfire Consequence 2021.*

PG&E EORM. (2020). *Spatial Wildfire Consequence 2021 - Lunch n' Learn presented 2020_10_16.*

PG&E Risk and Data Analytics. (2020). *Conductor Risk Model 2021 - Lunch n' Learn presented 2020_10_28.*

PG&E Risk and Data Analytics. (2020). *EVM Risk Model 2021 - Lunch n' Learn presented 2020_10_21.*

PG&E Risk and Data Analytics. (2020). *EVM Risk Model 2021 – Utility Analytics conference presented 2020_10_29.*

PG&E Risk and Data Analytics. (2020). *Methods - Ignition Probability Modeling 2021.*

primer on trees interacting with wind. (n.d.). Retrieved from https://ucanr.edu/sites/treefail/files/204521.pdf

Smiley, E. T., Matheny, N., & Lilly, S. (2012). Qualitative Tree Risk Assessment. *Arborist News 21(1)*, 1-20.

The Nature Conservancy. (n.d.). *Topographic Position and Landforms Analysis.* Retrieved from http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf

University of California, Merced. (n.d.). *GRIDMET: University of Idaho Gridded Surface Meteorlogical Dataset.* Retrieved from Earth Engine Data Catalog: https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET#description

USGS. (2016). *LANDFIRE dataset.* Retrieved from https://landfire.cr.usgs.gov/distmeta/servlet/gov.usgs.edc.MetaBuilder?TYPE=HTML&DATASET=FBK.

Western Tree Failure Database. (n.d.). Retrieved from https://ucanr.edu/sites/treefail/CTFRP_Statistics/50_or_more_753/

32.1 Example ignition vs. background covariate distributions

The following distributions (Figure 61) were created during the initial phase of development of the model of vegetation-caused ignitions. They are included here to illustrate the differences between environmental conditions at ignition locations vs. the entire grid. The greater the difference between ignition locations (pink) and all distribution grid locations (grey), the more definitive the MaxEnt predictions can be.

First are distributions related to the per-pixel average vs. maximum tree height, as estimated by Salo Sciences using computer vision algorithms on satellite imagery. It is easy to verify that however it is quantified, the presence of taller trees is much more common at locations of vegetation caused ignitions.

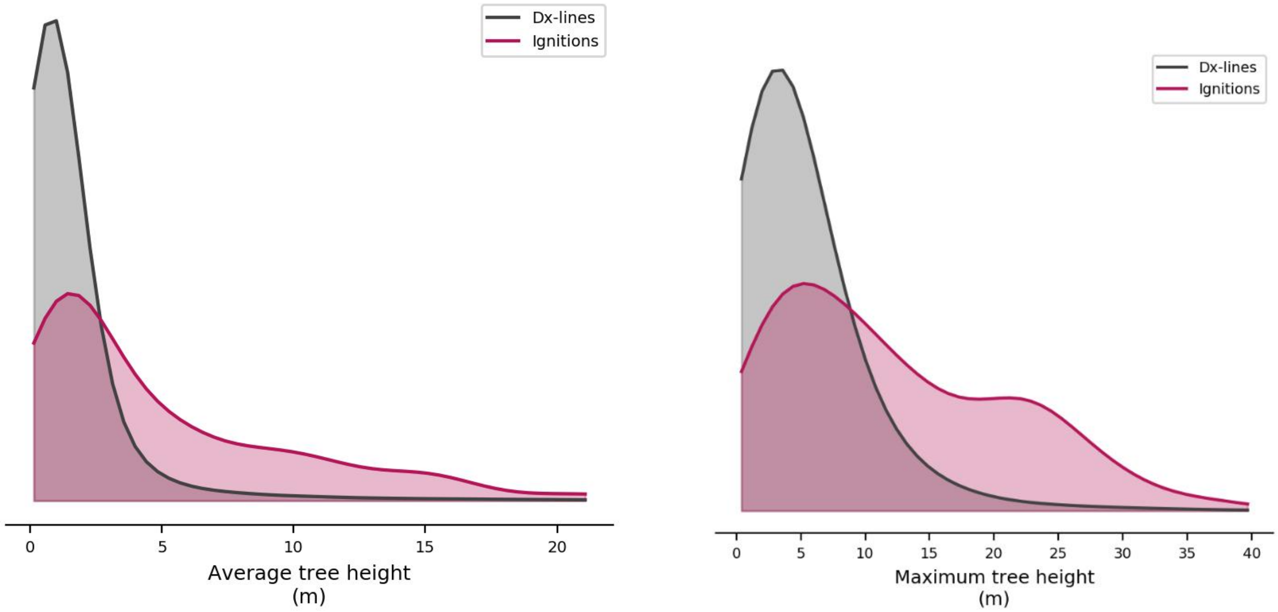


FIGURE 61 - TREE HEIGHT DISTRIBUTIONS

Next are distributions of multi-year average wind speed metrics based on GRIDMET (Figure 62) – daily average and daily maximum wind speeds. These appear to run counter to the intuition that dangerous fires are propelled by high winds and that vegetation failures are caused by high winds, but what they actually illustrate is important to understanding what is being modeled and how in our spatially-explicit model. First, it should be said that these figures are showing minimal differences between annual average wind conditions (for both daily average or maximum values) across ignition locations and the background. There is more discussion of wind and its role in Appendix 1: Vegetation-caused Ignition Risk Model 2021, but as an illustration of how to makes sense of a potentially count-intuitive result:

(1) Note that we are not modeling dangerous ignitions. We are modeling (and predicting) all ignitions – vegetation cause in this case. The majority of vegetation-caused ignitions are not associated with extreme winds (even though the most dangerous ones are).
(2) We are not modeling the moment of ignition – we are modeling the typical fire season. Our ignition data spans multiple years and the corresponding covariates must do the same. All of the wind speed variability in those distributions is spatial, not temporal.
(3) Recall that coastal areas, with relatively low ignitions risk due to higher moisture and humidity, are consistently windy.
(4) Note that consistently windy environments stunt and shape trees or event select for certain species that can handle the wind.
(5) Note that forested areas tend to break the wind, lowering wind speeds compared to more open spaces.
(6) For there to be trees with weak limb and trunks that haven't failed (yet) at a given site, that site has to have experienced only moderate winds up to the time of their eventual failure. In many cases, it will be because the winds are unusual for a given location that they "harvest" limbs.
(7) Other modeling approaches focused on question of when assets fail rather than where they fail do find more prominent role for wind in explaining some ignitions and the find that windy days have elevated probabilities of both outages and ignitions. MaxEnt models span entire fire-seasons and they expect the typical number of extreme wind days will occur in each location as they have in the past. Metrics related to wind gusts do have additional explanatory power (see below).

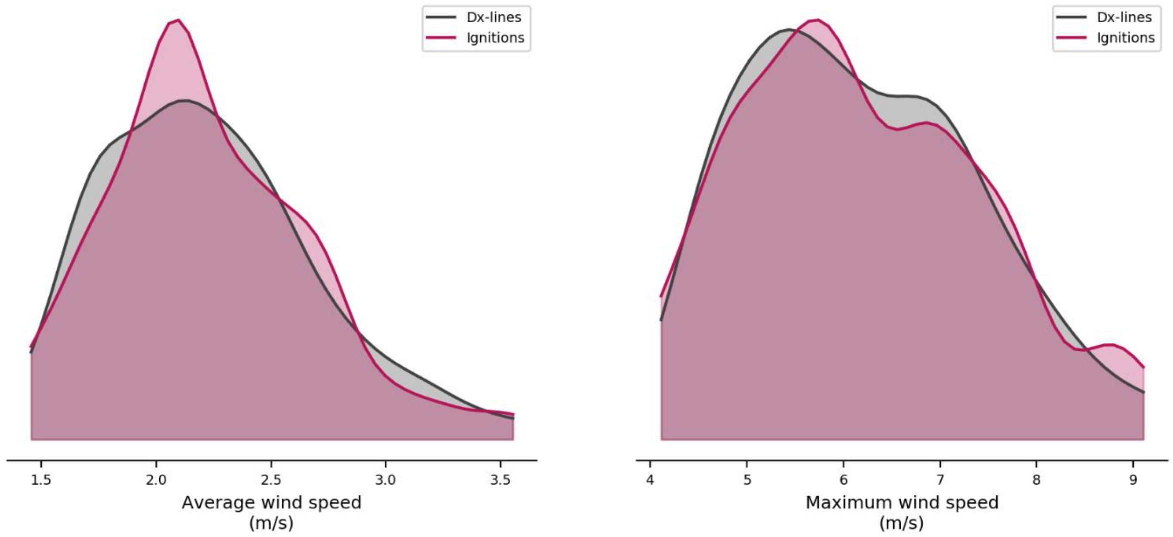


FIGURE 62 - WIND SPEED DISTRIBUTIONS

Finally, we look at gust speeds (left) and average temperatures (Figure 63). The gust speed metric presented comes from RTMA data and is the average of the daily maximum hourly wind gust speed. Here we can see that sites with ignitions do have a longer tail of gust speeds that background locations, a fact consistent with high gusts being a prominent mechanism that breaks branches. And finally, average temperatures at vegetation-caused ignition locations are cooler in larger numbers than background locations. At first glance, this might be counter intuitive due to the relationship between high temperatures, dryness, and ignitions. However, it is important to recall that the hottest locations with grid infrastructure lack trees and are therefore highly unlikely to experience vegetation-caused outages. The prevalence of lower temperatures among vegetation-caused ignition locations is most likely due to the fact that areas with trees tend to be cooler and moister – both because trees prefer such habitat, and their shade, leaf litter, and evapotranspiration all serve to make their surrounding cooler and more humid than they would otherwise be.

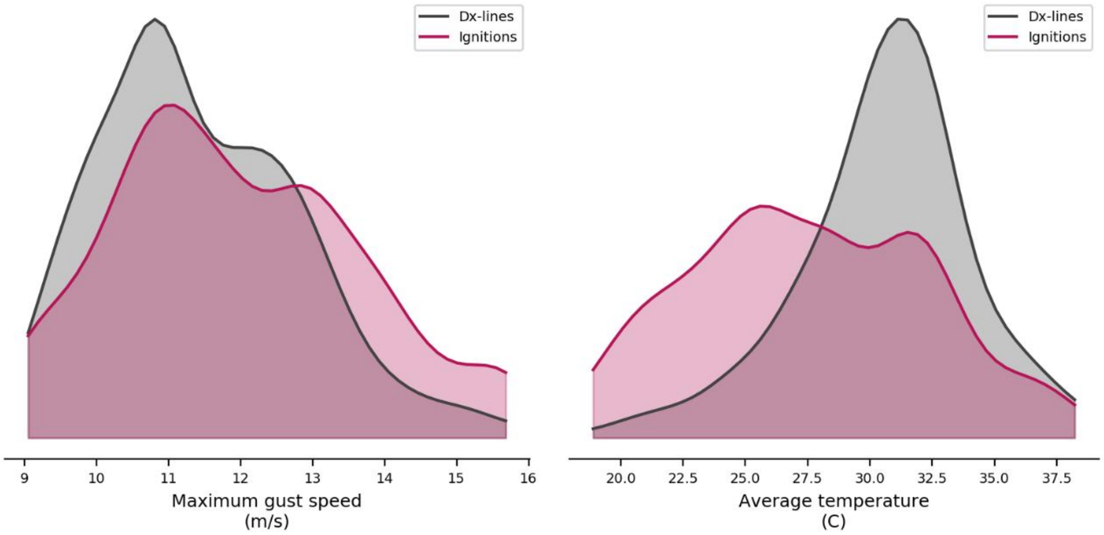
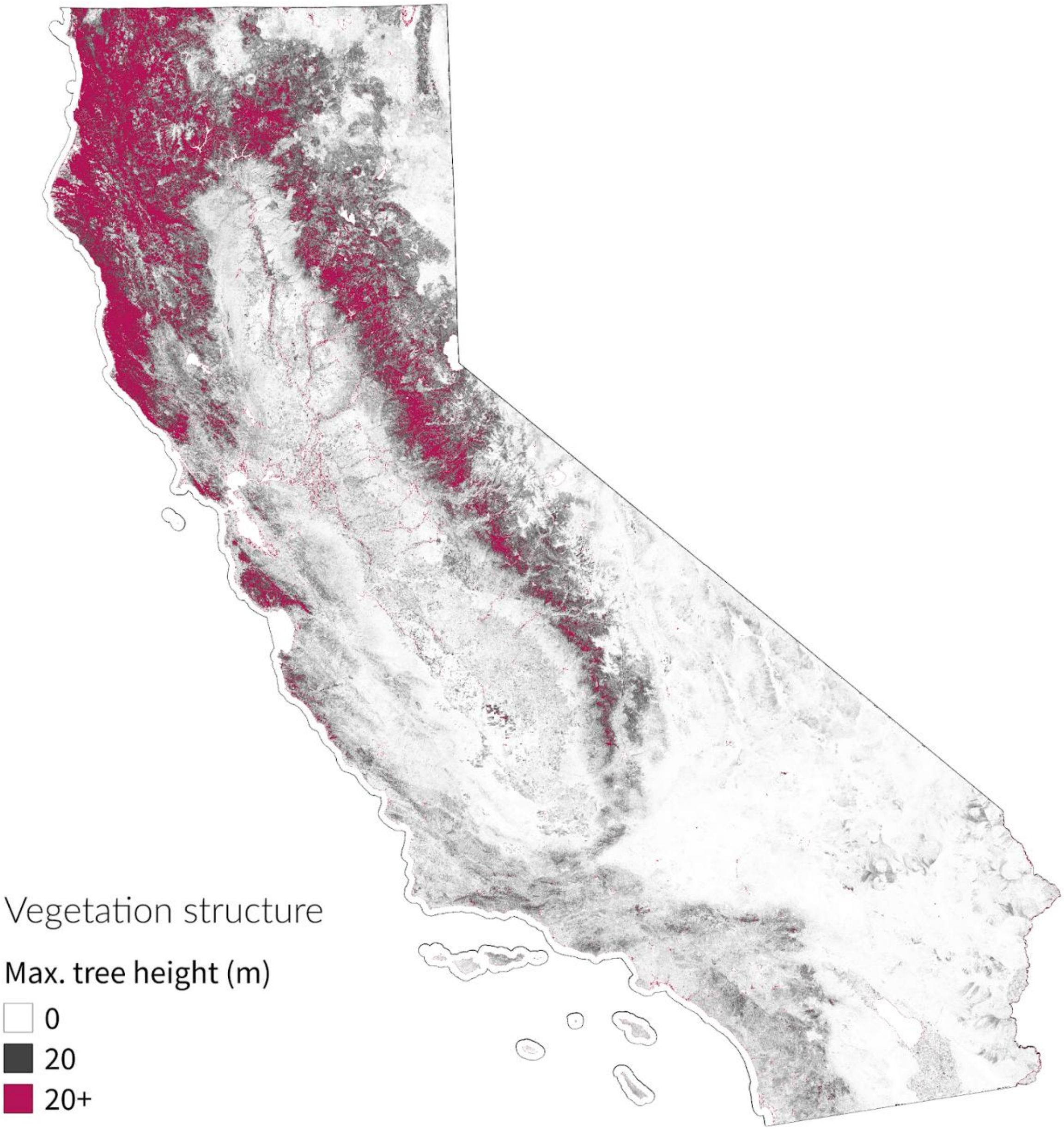
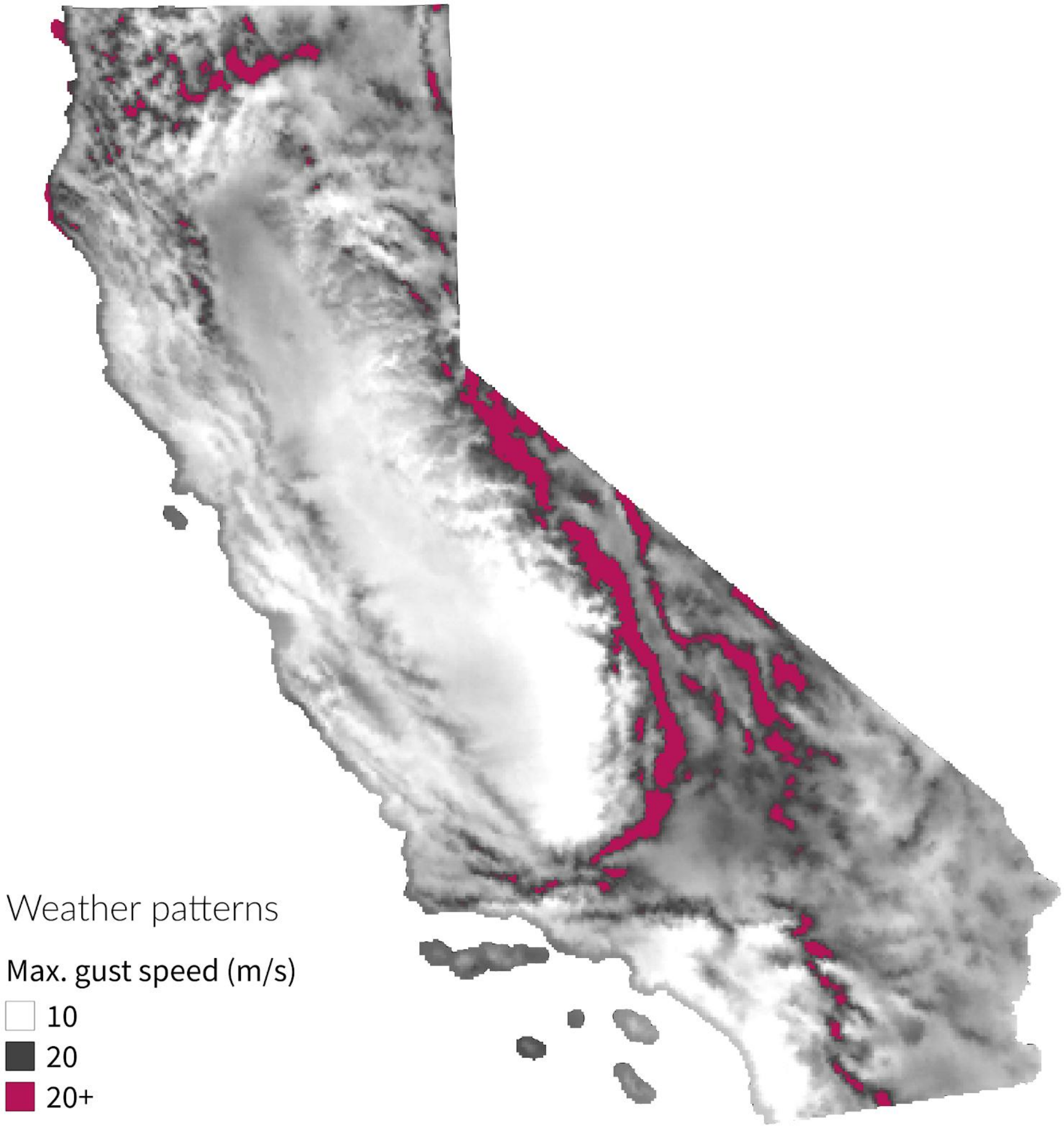
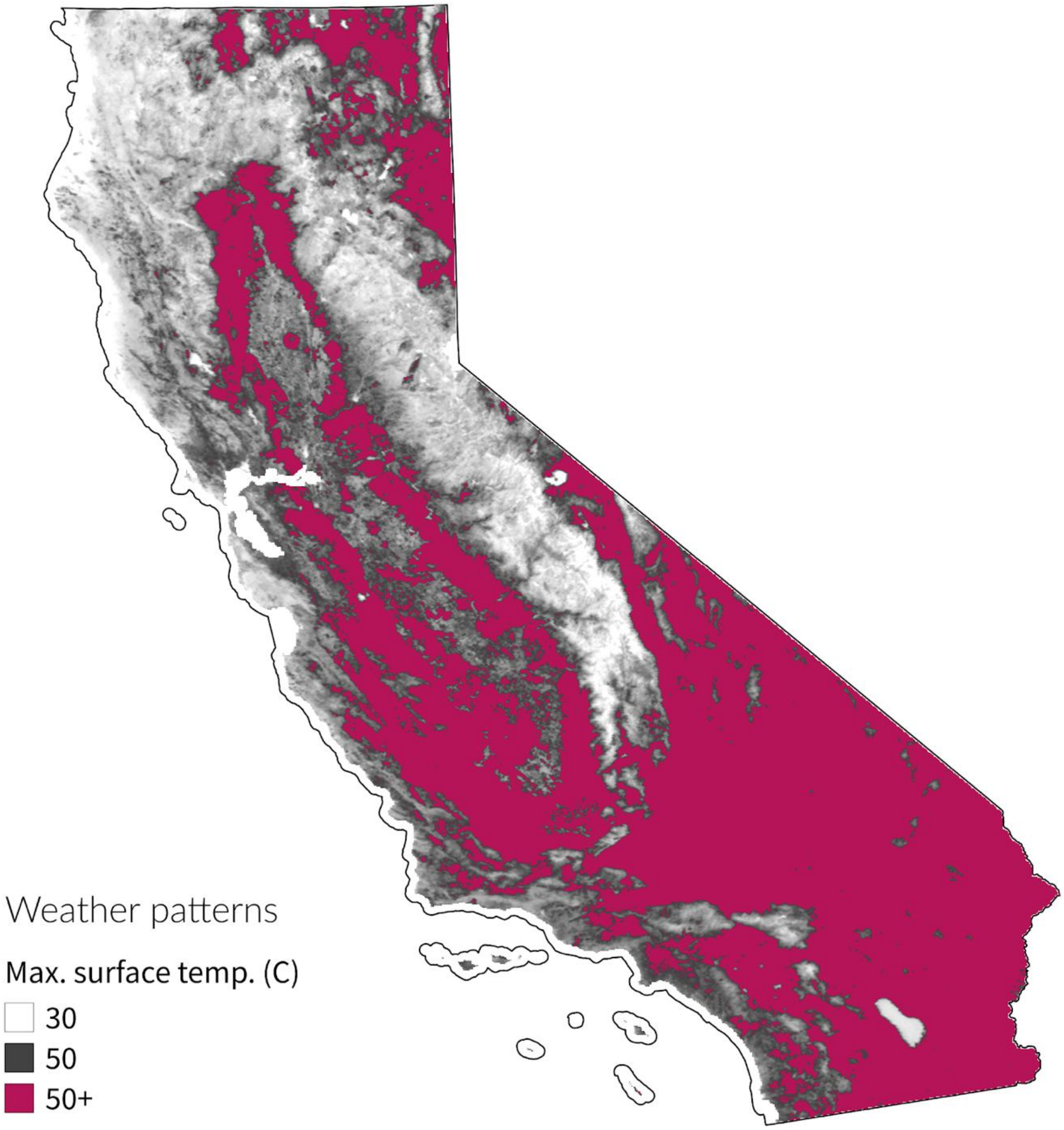


FIGURE 63 - GUST AND TEMP DISTRIBUTIONS

32.2 Example spatial covariates




Weather patterns
Max. gust speed (m/s)
□ 10
■ 20
🟥 20+


Weather patterns
Max. surface temp. (C)
□ 30
■ 50
🟥 50+

Appendix 4: Ignition Consequence Methods 2021

33 Executive Summary / Overview

This appendix explains the methods, assumptions, inputs, and outputs of the calculation of ignition consequence for the 2021 risk models. Recall that risk = ignition probability x ignition consequence, or alternately in the terminology of the MAVF framework, risk = LoRE x CoRE, where LoRE and CoRE are the likelihood and consequence of a risk event, respectively. It is intended as a technical methods companion to Appendix 1: Vegetation-caused Ignition Risk Model 2021and Appendix 2: Conductor-Involved Ignition Risk Model 2021 and is therefore intended for readers who want to understand the modeling process in more technical detail. The most important concept of this appendix is that the 2021 models require a version of consequence data that expresses spatially varying patterns in expected wildfire consequence. This appendix describes the derivation of “spatial consequence” values that are consistent with the risk methodology (and value totals) found in PG&E’s existing wildfire mitigation plans (WMP).

Additional sections and documents are referenced, and should be consulted to gain a full understanding of the model and process, the context in which this work was performed, and closely related work.

34 Introduction

PG&E is developing a Wildfire Consequence Model, which will enable mapping of Wildfire Consequence to 100m x 100m PG&E grid pixels.

Catastrophic wildland fires have become a major threat throughout the state of California and pose significant threat to the safety and economic future of the state. PG&E recognizes our electrical equipment has been the ignition point for a number of these fires and is working to understand these catastrophic events to maximize planned risk reduction activities. Enterprise risks are calculated, reported, and managed through the MAVF framework developed at the CPUC. Historically enterprise-level MAVF calculations have been performed without spatially explicit data or models. In other words, the risks are computed in terms of the expected count and severity of “risk events” but not their locations. The frequency and severity of these catastrophic fire events has increased dramatically over the last 10 years. The historical methods for quantifying fire risk need to evolve to manage the increasing population in the wildland urban interface and California’s warming and drying climate.

The purpose of the 2021 risk models is to model the spatial variation in risk so that wildfire mitigation efforts can prioritize higher risk assets. This appendix explains the development of the new spatially explicit MAVF CoRE consequence metrics (generally referred to simply as consequence herein and in related documentation) that are consistent with the enterprise-wide risk numbers reported in the most recent Wildfire Mitigation Plan (WMP).

35 Methods

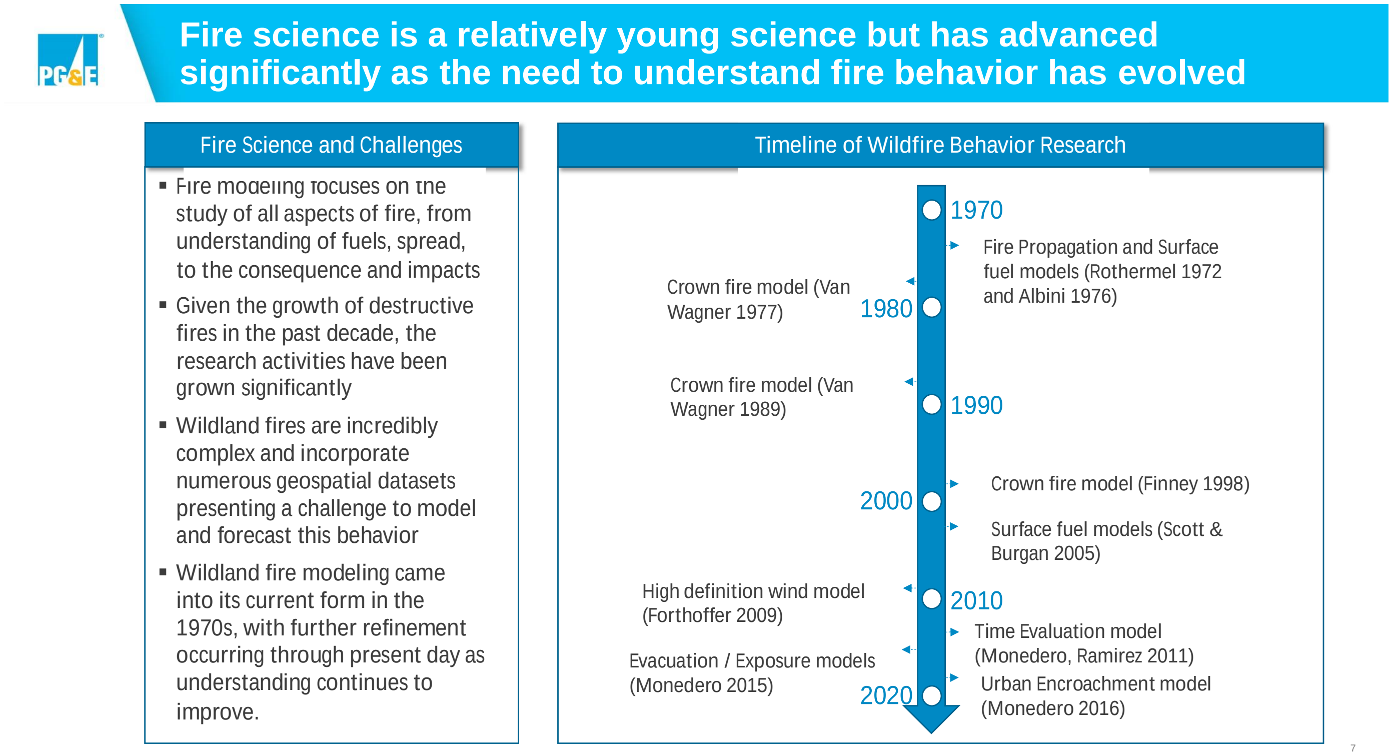
The PG&E Meteorology team worked with fire simulation vendor Technosylva to create fire spread simulation every 200m along the PG&E network within HFTD regions. Under this agreement, fire simulations using the worst historical days of fire weather (compiled from in-house historical weather data by the meteorology team) are performed for every location. By leveraging the Technosylva fire spread modeling outputs and historical Red Flag Warning shapefiles, the Wildfire consequence model was developed to distinguish the consequence within the HFTD tranche at 100m x 100m grid level.

The basic recipe for using these simulations to create calibrated MAVF CoRE consequence values is:

(1) Assign ignition simulation locations at regular (200m) spacing along all grid locations within HFTDs 2 and 3.
(2) Tabulate the 452 worst historical fire weather days using historical weather data.
(3) For all locations, run a separate 8-hour fire spread simulation for each day of weather data, recording burn area, flame length, impacted structures and FBI on a scale of 1 to 5 for each simulation.
(4) Using pre-existing MAVF "bow tie" consequence scores calculated for all combinations of fire severity (Small, Large, Destructive, Catastrophic), an HFTD indicator, and a red flag warning indicator rendered into a location-specific probability of a red flag warning, assign each simulation output a consequence score.
(5) The rule for assigning fire size are: Small Fire (area < 300 acres), Large Fire (area > 300 acres), Destructive Fire (area > 300 acres & 50+ Structures destroyed), Catastrophic Fire (assigned by ratio of Catastrophic to Destructive fires historically)
(6) Compute statistical extracts of consequence scores for all available simulations at each location – most downstream usage is based on the mean, but variance and others can also be useful.
(7) Assign the resulting mean consequence to each ignition location.
(8) Ensure that simulations can be mapped to all HFTD 2 and 3 grid locations. To do this, simulation output metrics are associated with a 200m x 200m raster pixel with the ignition point in the center, so the results can be assigned spatially to any locations within each pixel.
(9) Save a GeoTiff with derived consequence for every 100m x 100m grid pixel with a value assigned through the above steps.

36 Fire simulation

Figure 64 provides a summary of the history of fire simulation models from the 1970s. The basic formula is that fires spread according to available fuels, local topography, and winds, but the field is still young in many ways. Fire modelers are racing to learn how to model complex interactions driving modern wildfires. Areas of research and improvement include the rapid and long distance spread of fire through wind-driven embers, fires that are self-propelled and accelerated by convection driven winds, the fuel effects as drought- and beetle-killed trees (estimated at more than 150 million in the state) begin to decompose and fall over, and the impacts of an ever hotter and dryer climate as predictions stretch into the future.



Fire science is a relatively young science but has advanced significantly as the need to understand fire behavior has evolved

Fire Science and Challenges

- Fire modeling focuses on the study of all aspects of fire, from understanding of fuels, spread, to the consequence and impacts
- Given the growth of destructive fires in the past decade, the research activities have been grown significantly
- Wildland fires are incredibly complex and incorporate numerous geospatial datasets presenting a challenge to model and forecast this behavior
- Wildland fire modeling came into its current form in the 1970s, with further refinement occurring through present day as understanding continues to improve.

Timeline of Wildfire Behavior Research

1970

Fire Propagation and Surface fuel models (Rothermel 1972 and Albini 1976)

Crown fire model (Van Wagner 1977)

1980

Crown fire model (Van Wagner 1989)

1990

Crown fire model (Finney 1998)

2000

Surface fuel models (Scott & Burgan 2005)

High definition wind model (Forthoffer 2009)

2010

Time Evaluation model (Monedero, Ramirez 2011)

Evacuation / Exposure models (Monedero 2015)

2020

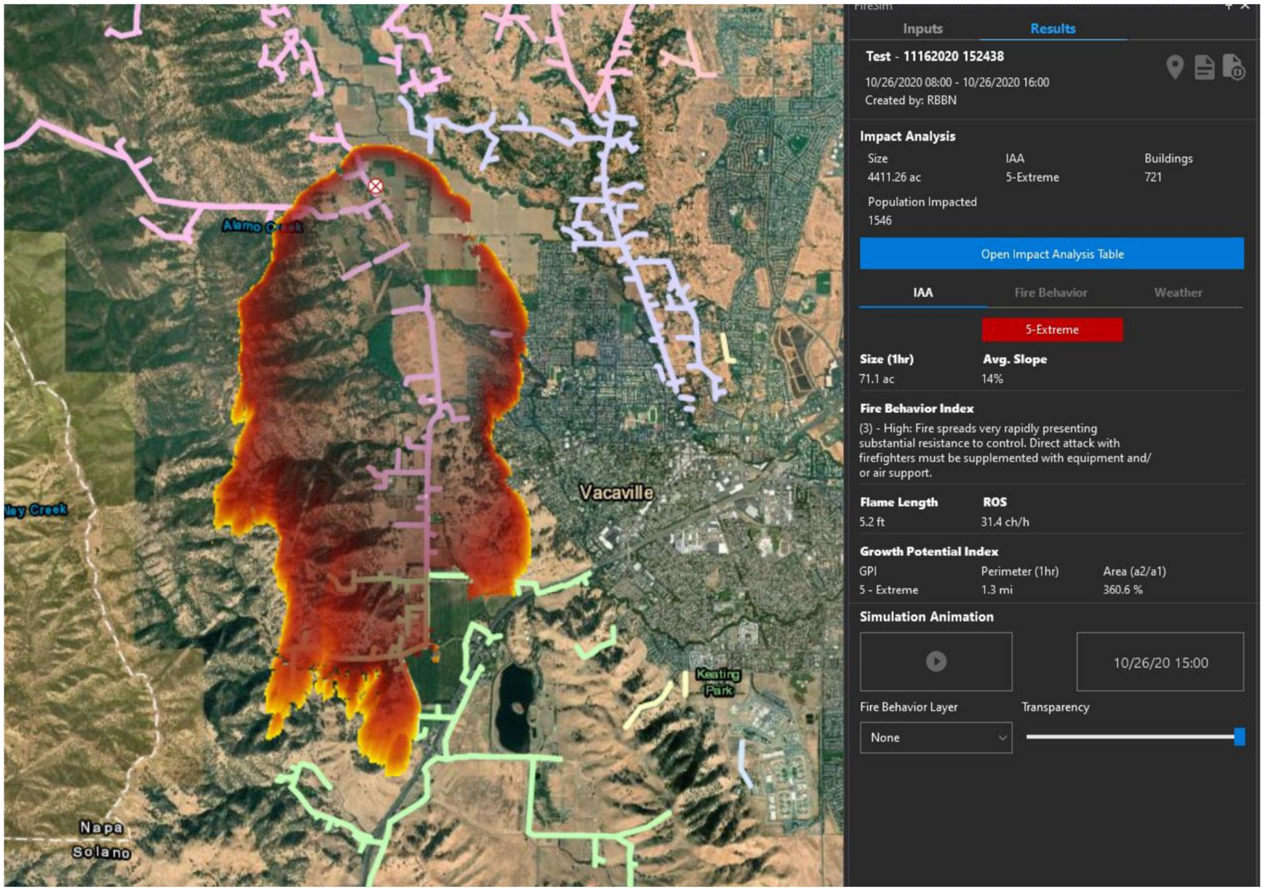
Urban Encroachment model (Monedero 2016)

7

Figure 64 - SUMMARY OF THE HISTORY OF WILDFIRE MODELING TECHNIQUES

Technosylva provides a product named, FireSim to offer real time insights into potential catastrophic fires and the Wildfire Risk Reduction Model (WRRM) which supports asset management across the organization in developing risk reduction programs.

When fires occur, either caused by asset equipment failure, or more commonly through other ignition sources, it is important to quickly understand where the fire is going and what it will impact. FireSim provides an on-demand capability to create a spread prediction and obtain detailed information on potential impacts based on topology, meteorology, and fuels type and condition from updated satellite imagery. Impact analysis includes population, buildings and company assets. Figure 65 below illustrates simulation outputs from FireSim.



Test - 11162020 152438

10/26/2020 08:00 - 10/26/2020 16:00
Created by: RBBN

Impact Analysis

Size	IAA	Buildings
4411.26 ac	5-Extreme	721

Population Impacted
1546

5-Extreme

Size (1hr)	Avg. Slope
71.1 ac	14%

Fire Behavior Index

(3) - High: Fire spreads very rapidly presenting substantial resistance to control. Direct attack with firefighters must be supplemented with equipment and/ or air support.

Flame Length	ROS
5.2 ft	31.4 ch/h

Growth Potential Index

GPI	Perimeter (1hr)	Area (a2/a1)
5 - Extreme	1.3 mi	360.6 %

Simulation Animation

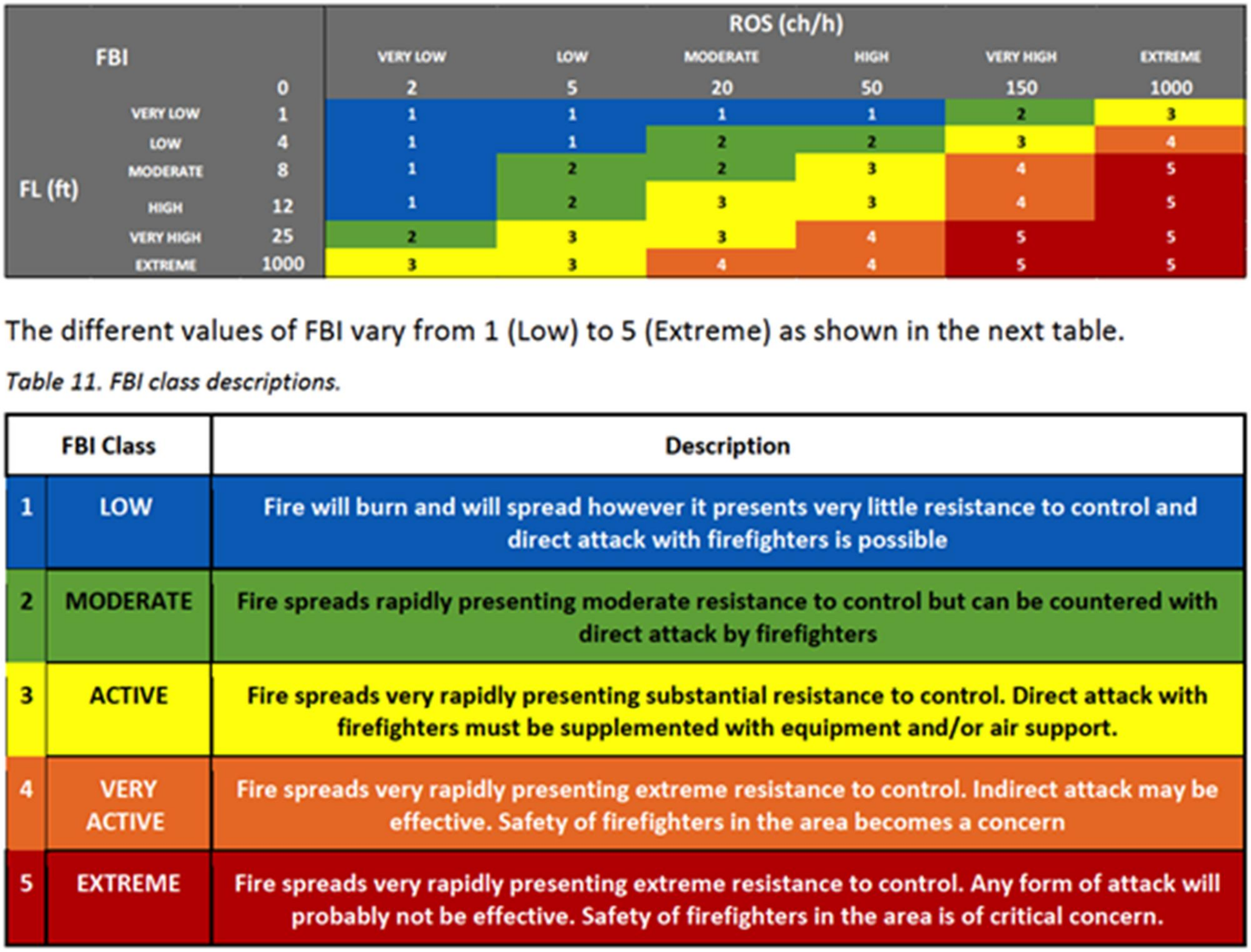
10/26/20 15:00

Fire Behavior Layer: None

Transparency

FIGURE 65 - EXAMPLE VISUALIZATIONS OF TECHNOSYLVA FIRESIM RESULTS, MODELING THE EXPECTED EXTENT OF A MODELED FIRE AND THE INTENSITY OF BURN

The primary simulation output relevant to this work is the analysis of fires modeled from ignition locations at regularly spaced points along the entire set of grid infrastructure within HFTDs 2 and 3. These 8-hour simulations, conducted using weather data from 452 worst historical fire weather days at each location provide key consequence metrics summarizing burn area, structures impacted, and fire behavior index (FBI) for each simulation run. The criteria for FBI assignment, based on flame length (a metric of burn intensity) and rate of spread (ROS) and a description of each FBI class are illustrated in Figure 66 below.



FBI			ROS (ch/h)					
			VERY LOW	LOW	MODERATE	HIGH	VERY HIGH	EXTREME
		0	2	5	20	50	150	1000
FL (ft)	VERY LOW	1	1	1	1	1	2	3
	LOW	4	1	1	2	2	3	4
	MODERATE	8	1	2	2	3	4	5
	HIGH	12	1	2	3	3	4	5
	VERY HIGH	25	2	3	3	4	5	5
	EXTREME	1000	3	3	4	4	5	5

The different values of FBI vary from 1 (Low) to 5 (Extreme) as shown in the next table.

Table 11. FBI class descriptions.

FBI Class		Description
1	LOW	Fire will burn and will spread however it presents very little resistance to control and direct attack with firefighters is possible
2	MODERATE	Fire spreads rapidly presenting moderate resistance to control but can be countered with direct attack by firefighters
3	ACTIVE	Fire spreads very rapidly presenting substantial resistance to control. Direct attack with firefighters must be supplemented with equipment and/or air support.
4	VERY ACTIVE	Fire spreads very rapidly presenting extreme resistance to control. Indirect attack may be effective. Safety of firefighters in the area becomes a concern
5	EXTREME	Fire spreads very rapidly presenting extreme resistance to control. Any form of attack will probably not be effective. Safety of firefighters in the area is of critical concern.

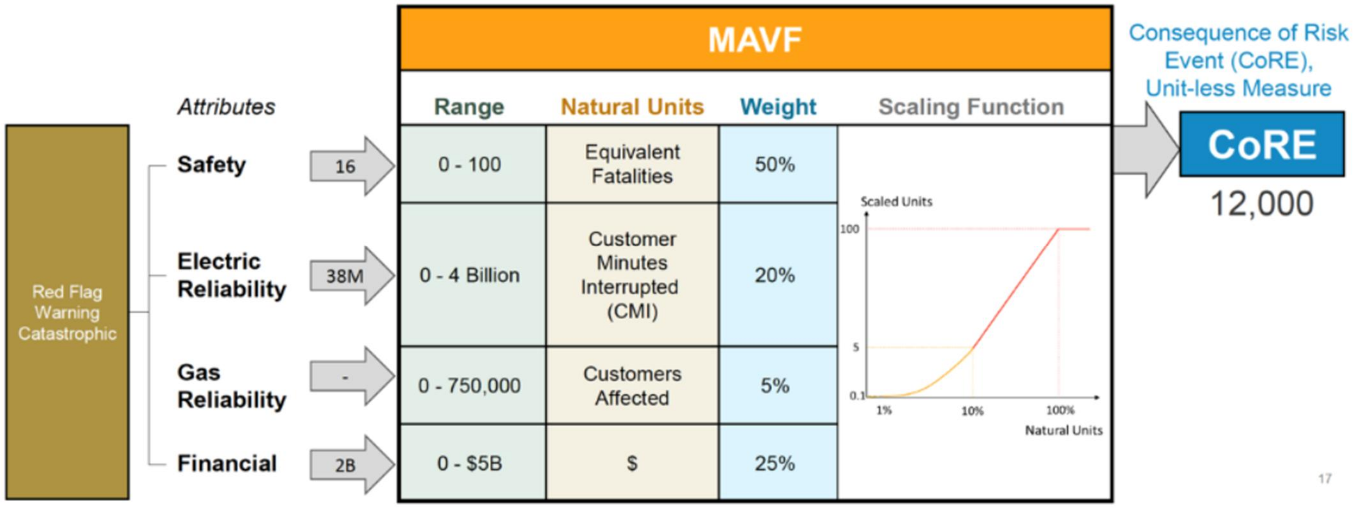
FIGURE 66 - FIRE BEHAVIOR INDEX COMPONENTS AND DESCRIPTION

37 Existing MAVF “bow tie” Consequences

PG&E uses a Multi Attribute Variable Function (MAVF) to calculate the consequence of an event. MAVF is a tool for combining potential consequences of the occurrence of a risk event and creating a single quantification of risk values. Some of its key features are:

- It formalizes trade-offs between different dimensions of consequence attributes (Safety, Reliability and Financial).
- It captures aversion or indifference over a range of outcomes based on the company’s risk management preference.
- It allows comparisons of risk across the company using a common metric.

Figure 67 is the MAVF approved by PG&E’s Risk committee for use across company for Risk scoring.



Red Flag Warning Catastrophic

MAVF

Attributes		Range	Natural Units	Weight	Scaling Function
Safety	16	0 - 100	Equivalent Fatalities	50%	
Electric Reliability	38M	0 - 4 Billion	Customer Minutes Interrupted (CMI)	20%	
Gas Reliability	-	0 - 750,000	Customers Affected	5%	
Financial	2B	0 - $5B	$	25%	

Consequence of Risk Event (CoRE), Unit-less Measure

CoRE

12,000

FIGURE 67 - MAVF

The consequence attributes and their respective weights are:

- Financial (25%)
- Safety (50%)
- Electric Reliability (20%)

Each outcome in the Consequence model is assigned a score for these 3 categories which is then aggregated to calculate the consequence score.

The consequence values assigned to each simulated fire come from these existing MAVF consequence scores. The main idea is that MAVF divides wildfire risk events into severity categories, modeling each category as a separate set of inputs (think tabulations/counts of historical ignitions that fit into each severity category) and consequence outcomes. Because the inputs come from multiple sources into the central risk event calculation and then fan back out to the Safety, Reliability, and Financial risk categories, each category is called a “bow tie” after what it looks like when diagrammed.

The bow tie methodology is a structured way of conceptualizing, representing risk across many types of events. It breaks down the causes of a risk event into separate tranches and calculates the adverse consequences of the risk event for each of these tranches. Tranches segment a system of assets into “like” risk groups because different parts of a system face different hazards, are susceptible to those hazards to different degrees and can result in different consequences given the same event. For instance:

- Material: plastic is not threatened by corrosion compared to metal
- Location: Earthquake in Oakland vs Santa Cruz
- Ambient Conditions: Proximity to vegetation. (combustible material)

A bow tie (Figure 68) quantifies relationships between drivers and outcomes.

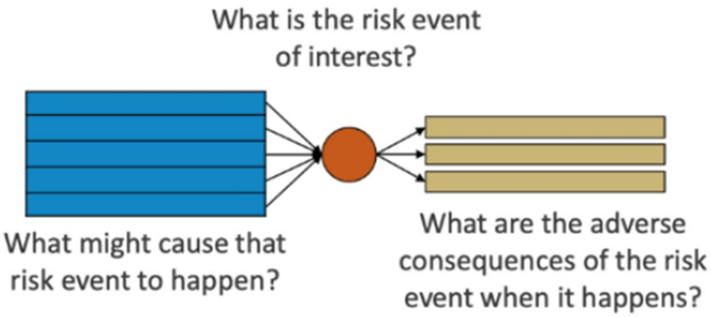

What is the risk event of interest?
What might cause that risk event to happen?
What are the adverse consequences of the risk event when it happens?

FIGURE 68 - BOW TIE STRUCTURE

Under the hood, there are as many bow ties as there are tranches (Figure 69).

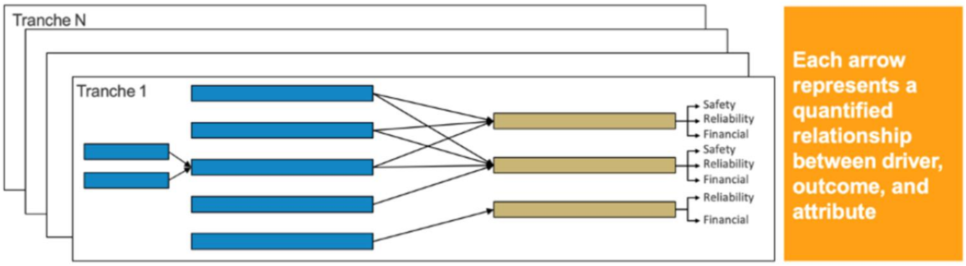


FIGURE 69 - TRANCHES

Figure 70 below provides an example wildfire bow tie.

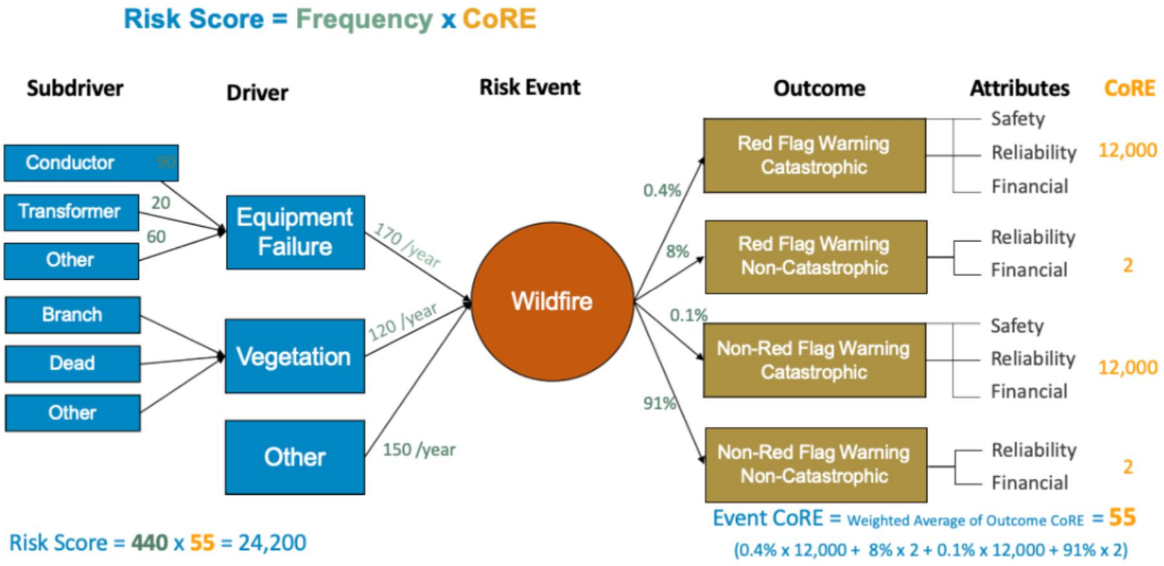
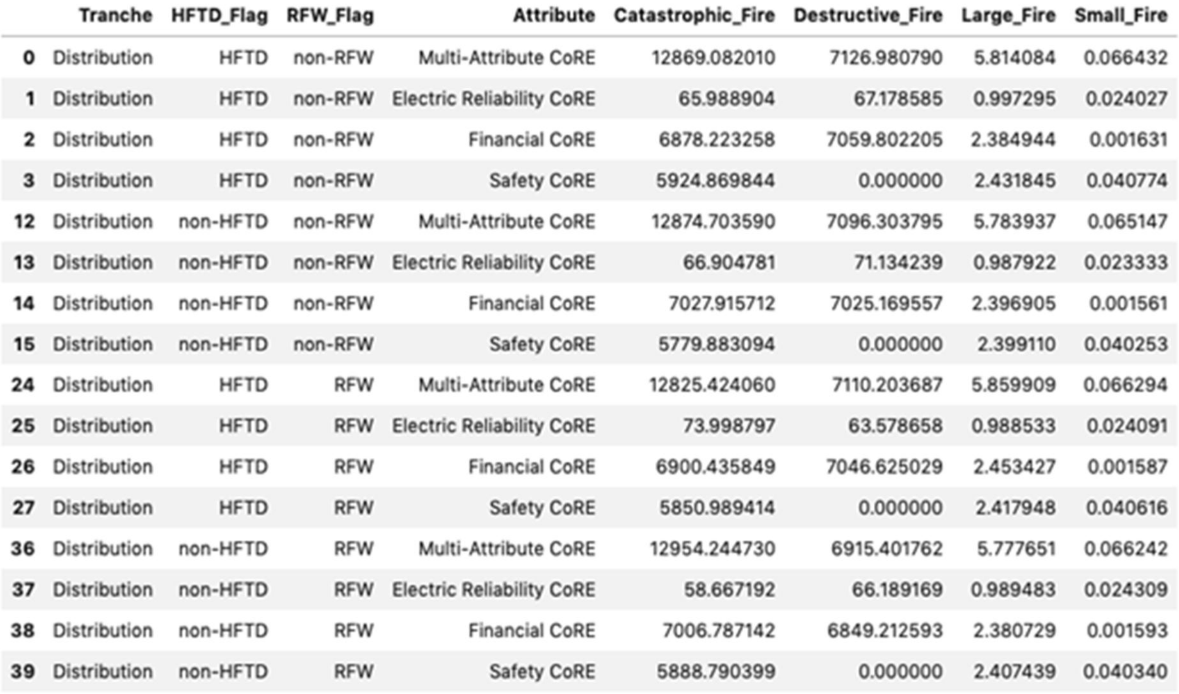

Risk Score = Frequency x CoRE
Subdriver
Driver
Risk Event
Outcome
Attributes
CoRE
Conductor
90
Transformer
20
Other
60
Branch
Dead
Other
Equipment Failure
Vegetation
Other
170 /year
120 /year
150 /year
Wildfire
0.4%
8%
0.1%
91%
Red Flag Warning Catastrophic
Red Flag Warning Non-Catastrophic
Non-Red Flag Warning Catastrophic
Non-Red Flag Warning Non-Catastrophic
Safety
Reliability
Financial
12,000
Reliability
Financial
2
Safety
Reliability
Financial
12,000
Reliability
Financial
2
Event CoRE = Weighted Average of Outcome CoRE = 55
(0.4% x 12,000 + 8% x 2 + 0.1% x 12,000 + 91% x 2)
Risk Score = 440 x 55 = 24,200

FIGURE 70 - EXAMPLE WILDFIRE BOW TIE

What matters for our purposes is that each bow tie produces CoRE consequence values specific to the categories of events that feed into it and these can become a lookup table for consequence of simulated wildfires as long as they can be mapped into the same categories. Table 8 below provides the actual MAVF CoRE values for wildfire-relevant bow tie and consequence categories.

TABLE 8 - MAVF CORE VALUES



	Tranche	HFTD_Flag	RFW_Flag	Attribute	Catastrophic_Fire	Destructive_Fire	Large_Fire	Small_Fire
0	Distribution	HFTD	non-RFW	Multi-Attribute CoRE	12869.082010	7126.980790	5.814084	0.066432
1	Distribution	HFTD	non-RFW	Electric Reliability CoRE	65.988904	67.178585	0.997295	0.024027
2	Distribution	HFTD	non-RFW	Financial CoRE	6878.223258	7059.802205	2.384944	0.001631
3	Distribution	HFTD	non-RFW	Safety CoRE	5924.869844	0.000000	2.431845	0.040774
12	Distribution	non-HFTD	non-RFW	Multi-Attribute CoRE	12874.703590	7096.303795	5.783937	0.065147
13	Distribution	non-HFTD	non-RFW	Electric Reliability CoRE	66.904781	71.134239	0.987922	0.023333
14	Distribution	non-HFTD	non-RFW	Financial CoRE	7027.915712	7025.169557	2.396905	0.001561
15	Distribution	non-HFTD	non-RFW	Safety CoRE	5779.883094	0.000000	2.399110	0.040253
24	Distribution	HFTD	RFW	Multi-Attribute CoRE	12825.424060	7110.203687	5.859909	0.066294
25	Distribution	HFTD	RFW	Electric Reliability CoRE	73.998797	63.578658	0.988533	0.024091
26	Distribution	HFTD	RFW	Financial CoRE	6900.435849	7046.625029	2.453427	0.001587
27	Distribution	HFTD	RFW	Safety CoRE	5850.989414	0.000000	2.417948	0.040616
36	Distribution	non-HFTD	RFW	Multi-Attribute CoRE	12954.244730	6915.401762	5.777651	0.066242
37	Distribution	non-HFTD	RFW	Electric Reliability CoRE	58.667192	66.189169	0.989483	0.024309
38	Distribution	non-HFTD	RFW	Financial CoRE	7006.787142	6849.212593	2.380729	0.001593
39	Distribution	non-HFTD	RFW	Safety CoRE	5888.790399	0.000000	2.407439	0.040340

38 Mapping Consequence Values onto Spatial Locations

To derive spatial consequence values for each simulated fire, each location is classified by its HFTD, red flag warning, and fire severity attributes. This classification identifies   which row of data to use for the consequence score data for each simulated fire.

Specifically, FireSim outputs are used to distribute the Historical MAVF Consequence values at the Tranche Level to the 100m pixel level. Instead of uniformly distributing the consequence value within the Tranche, FireSim outputs are used to allocate the consequence values at each location.

HFTD is a pure function of location that is either True or False, so that assignment can just be looked up spatially using the official CPUC HFTD shape file.

Red flag warnings (RFW) are time and location specific, so the scores used for them are a weighted average of the probability that each location will be under a Red Flag warning on any given day during the fire season. These are calculated using Red Flag Warning shape files from 2015-2019, with probabilities rendered to 100m x 100m spatial pixels.

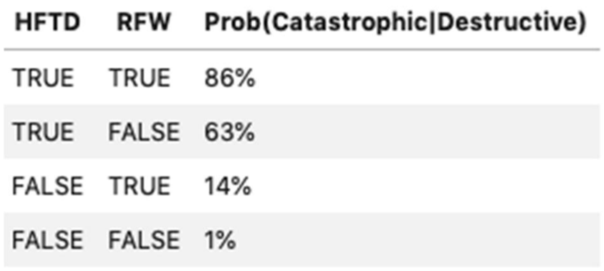
Fire severity is the most complicated to assign but is still deterministically based on simulated fire metrics for each simulation. The rules for fire severity assignment for each simulation are:

- Small Fire (< 300 acres)

- Large Fire (> 300 acres)
- Destructive Fire (> 300 acres & 50+ Structures destroyed OR FBI >= 3)
- Catastrophic Fire (Destructive & at least a serious injury)

As serious injuries are not an output from the Technosylva data, Catastrophic Fire probability cannot be determined directly. Instead, the probabilities in Table 9are derived from CalFire dataset, consistent with the numbers used in the wildfire bow tie model.

TABLE 9 - CATASTROPHIC FIRE CONDITIONAL PROBABILITY

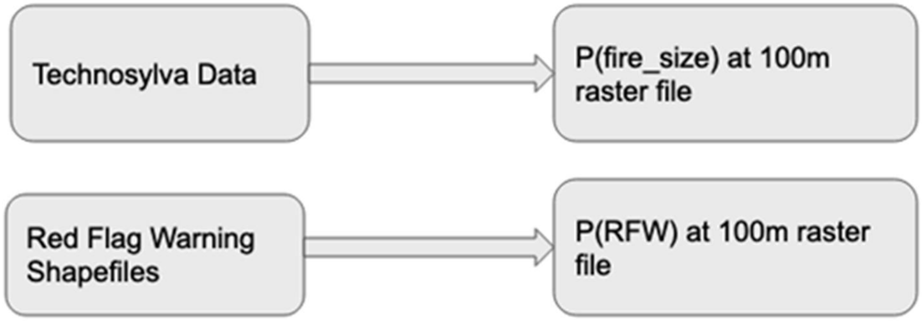


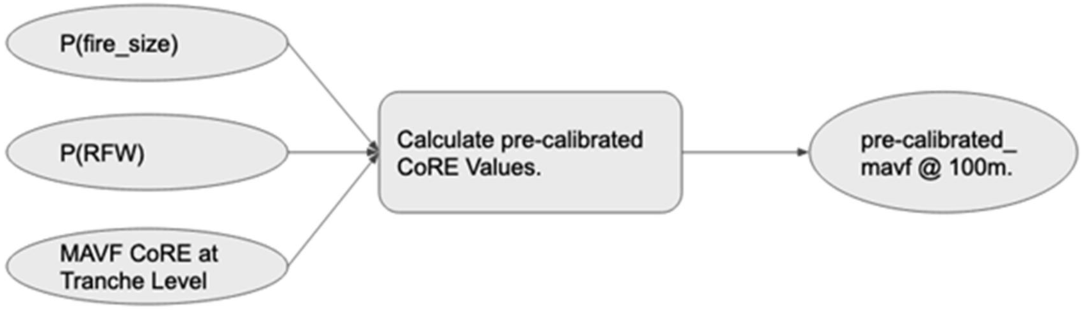
HFTD	RFW	Prob(Catastrophic|Destructive)
TRUE	TRUE	86%
TRUE	FALSE	63%
FALSE	TRUE	14%
FALSE	FALSE	1%

In other words, 86% of destructive fires within the HFTDs under RFW conditions are expected to be catastrophic, whereas just 1% of those outside the HFTDs and not under RFW conditions are.

Once each simulated outcome has an HFTD assignment, a RFW probability, and a severity probability, the bow tie consequence values (for all 3 consequence categories and their sum) for each category are assigned with the appropriate probability weights. Then the probability weighted average consequence is computed for all weather days simulated for each ignition location, yielding the spatial MAVF consequence scores. For reasons discussed in the next section, these scores require a final calibration step to tie together with the total risk values reported in the most recent WMP. The results prior to the final calibration are called the pre-calibrated consequence data.

The key modeling steps of this calculation are illustrated below, with fire size and RFW probabilities derived from their relevant input data and fire size, RFW, and HFTD (not visualized) combined to lookup bow tie consequence values, averaged into pre-calibrated values across weather days and rendered to 100m x 100m raster data:


Technosylva Data
P(fire_size) at 100m raster file
Red Flag Warning Shapefiles
P(RFW) at 100m raster file


P(fire_size)
P(RFW)
MAVF CoRE at Tranche Level
Calculate pre-calibrated CoRE Values.
pre-calibrated_ mavf @ 100m.

The two maps below illustrate mean burn area (Figure 71) and FBI (Figure 72) values from the Technosylva fire simulations for the North Bay. These are key inputs into the fire severity classification calculation.

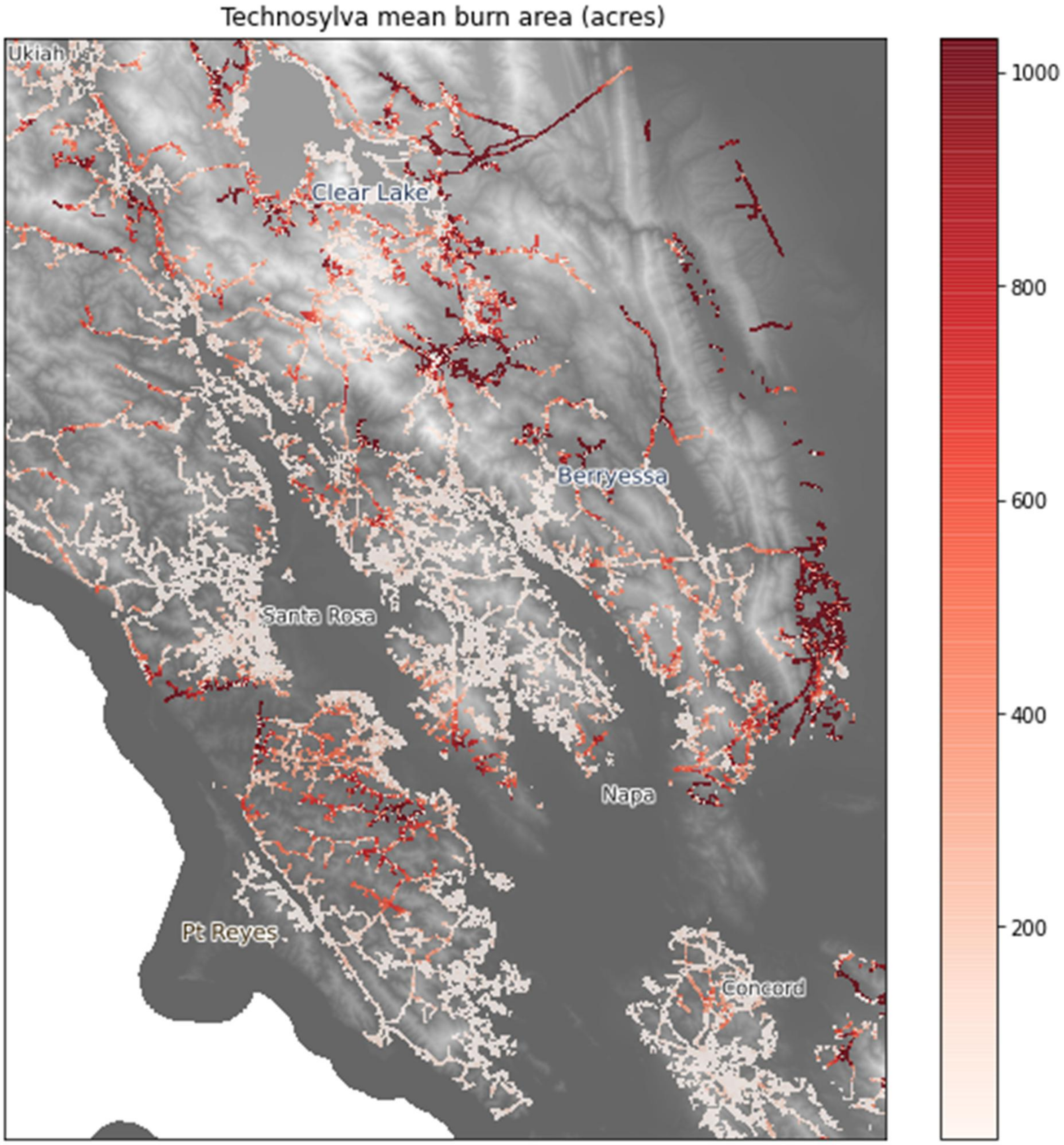


FIGURE 71 - TECHNOSYLVA MEAN BURN AREA IN ACRES

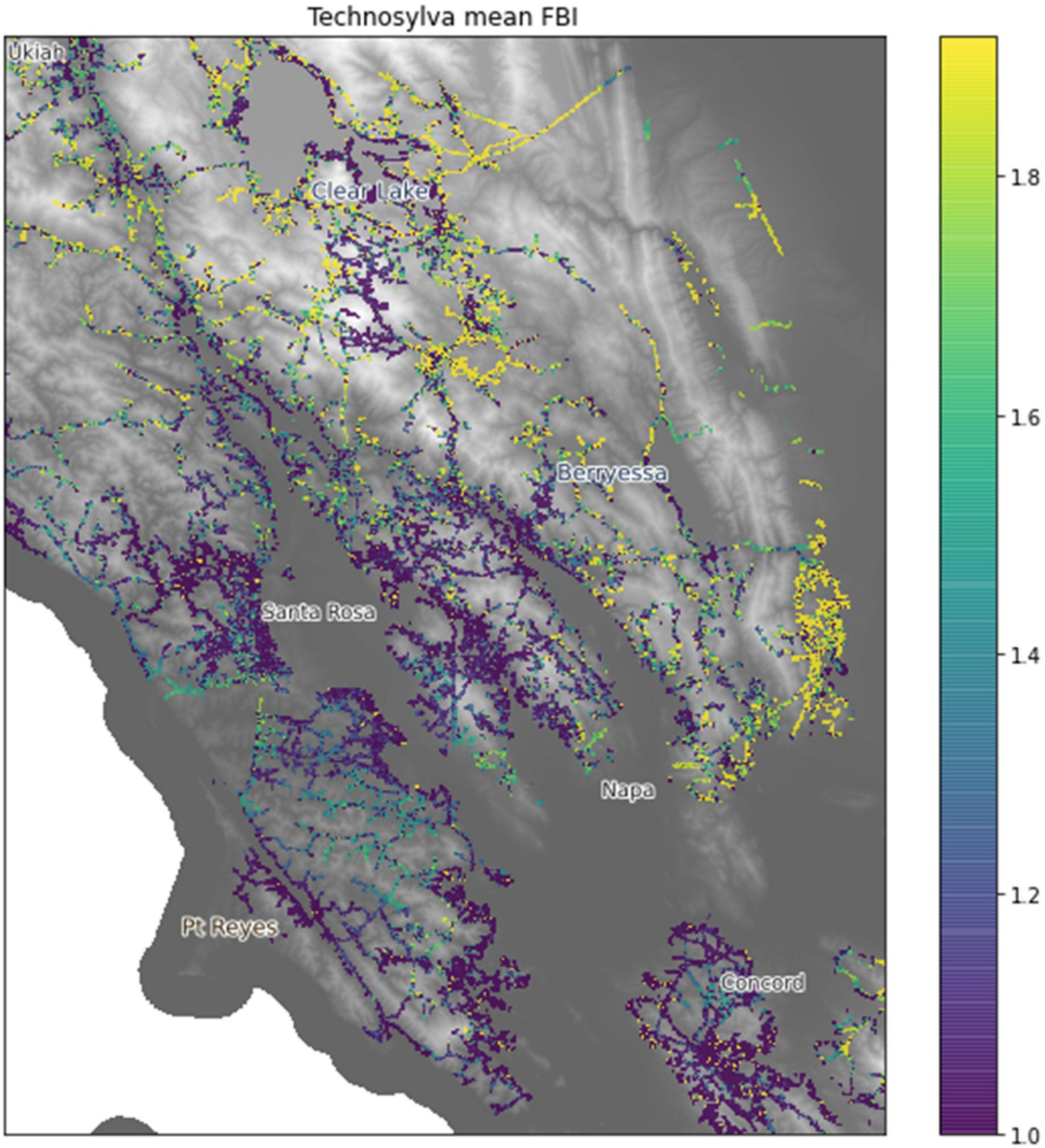


FIGURE 72 - TECHNOSYLVA MEAN FBI

Figure 73 below maps the spatial consequence values for the North Bay, with HFTDs highlighted in translucent orange and red and all grid pixels in light grey. The influence of burn area and FBI on the final result can be verified via cross comparison of the maps.

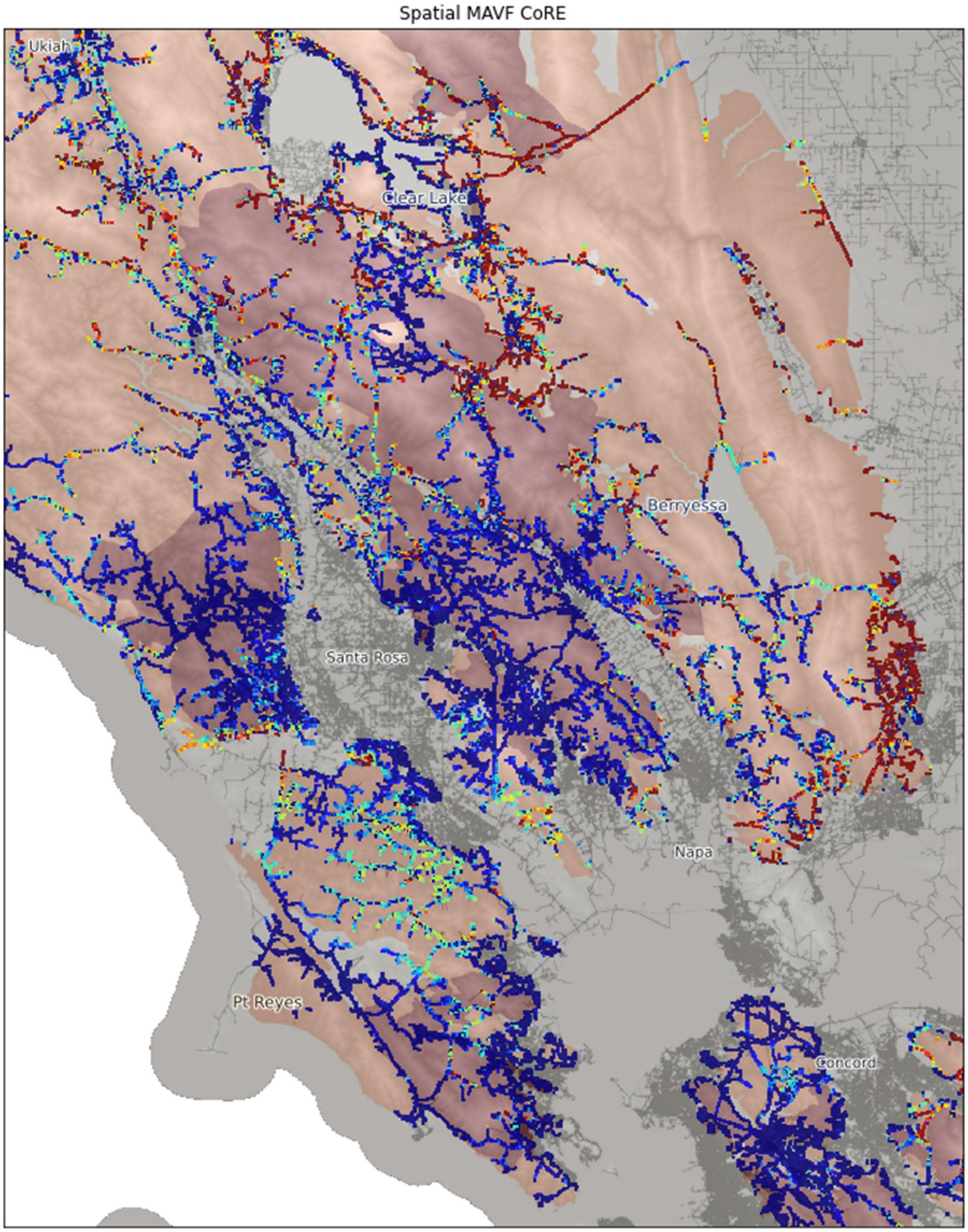


FIGURE 73 - PER-PIXEL CONSEQUENCE VALUES, RED IS HIGHER, BLUE IS LOWER

39 Consequence calibration

The derived consequence scores are consistent with the bow tie MAVF CoRE values, assumptions and methods, but risk is ultimately a function of the number of ignitions. To ensure that our work is consistent with the risk values computed using the standard bow tie models, the consequence data needs to be calibrated. As we are calculating the Technosylva fire spread modeling using worst weather conditions, pre-calibrated CoRE overestimates the consequence and data does not match with wildfire bow tie model results, which is using last 5 years historical ignitions data. We can calibrate the CoRE by matching uncalibrated Risk per event to the weighted average CoRE from the wildfire bow tie model. By applying uniform calibration factor across all pixels, this method preserves the relative consequence difference between 100m pixels.

Specifically, to calibrate MAVF CoRE consequence data we follow these steps:

(1) For each ignition cause and impacted equipment type, lookup the total risk reported in the WMP. This is the risk associated with all ignitions from that combination of characteristics.

(2) Note that the risk is the product of the per-event risk and the count of events in that category.

(3) By dividing the total risk by the count of events in its category associated with the WMP calculations, we obtain *a per-event risk for each cause/equipment type* risk category from the WMP. These values were 101.7 for vegetation cause ignitions and 59.7 for conductor involved ignitions for the 2021 risk models.

(4) Compute uncalibrated risk using the MaxEnt ignition probabilities as LoRE x pre-calibrated CoRE, sum across all grid pixels to compute total risk, and divide by the expected count of ignitions predicted by the ignition probabilities (aka the sum(LoRE)) to get the uncalibrated risk per-event.

(5) Take the ratio of the WMP risk per-event and the uncalibrated risk per-event (Calibration factor = Calibrated Risk per Event / Uncalibrated Risk per Event) and use it to multiply the CoRE values (Calibrated CoRE = CoRE * Calibration factor). Note that this multiplicative scaling does not change the rank order of any results – it simply re-scales the values so they add to the WMP values.

(6) The total risk associated with calibrated CoRE will now equal the risk associated with the WMP for the same number of risk events.

40 Validation

Figure 74 below illustrates the consequence scores for major named fires when simulated via Technosylva to consequence scores from Reax, the simulation software used for prior wildfire consequence assessment. For this comparison the score for the most destructive simulation for each location was plotted. Technosylva, along the x-axis, more consistently associates elevated consequence with these destructive real-world high-risk fires.

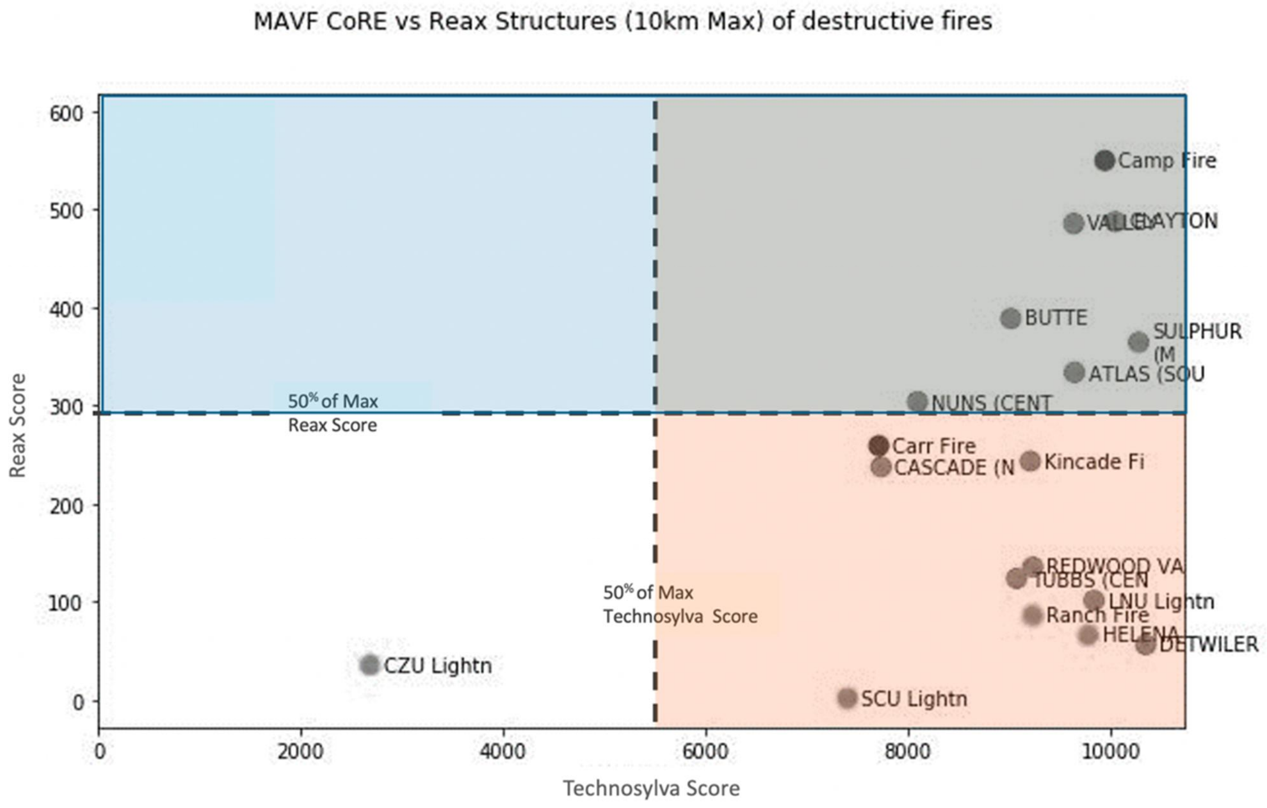


FIGURE 74 - COMPARISON OF REAX AND TECHNOSYLVA SCORES FOR NOTABLE WILDFIRES

Reax Score

- Previous risk models used the REAX wildfire consequence model.
- Relies on fuels as a main parameter to determine wildfire spread, however fuels data for the Reax runs performed pre-dates significant update from the fire modeling community.
- Uses census tract population to compute structures impacted, which can distort the locations of highest damage.
- REAX scores just a portion of destructive historical fires high

Technosylva Consequence Score

- Uses the Technosylva model which models ladder effect of fire moving from grass to scrub to treetops.
- Includes more current fuels data and more accurate structures data.
- Consequence scores most historical catastrophic fires high.

41 Model Limitations

- Technosylva fire spread simulations are done only within in HFTD Tiers 2 and 3 and only using high risk (worst) weather conditions.

- Simulations model the first 8 hours of a fire as a proxy for the destructive potential of the fire. The Technosylva simulations used in this analysis were for 8 hours.
- Destructive Fire Probability is sensitive to the parameter values we chose for FBI, Acres and Buildings.
- Fire simulation models are not capable of modeling the most active and destructive wildfires we experience in California. Work on wind driven ember transport, positive convective wind feedback loops, and the very significant standing fuels from drought and beetle damage is ongoing.
- Fires are (fortunately) too rare to empirically validate predictions with high statistical confidence.
- MAVF tranches and the function itself have several free parameters whose values reasonable people might disagree on.
- Wildfire risk appears to be an emergent outcome of climate change characterized by non-linear response to conditions due to threshold crossing and feedbacks. This makes it difficult to model or calibrate based on empirical data (i.e. from the past).
- Firefighting has a significant impact on how large fires grow and how destructive they are. Fire simulations do not account for suppression activity and can therefore make unrealistic spread predictions.