Dx Risk / Phase 1 / Milestone 1

Prepared by: CDA, Salo, and Presence
For internal PG&E use
2/7/2020

Context – Why are we doing this initiative?

In support of the new EO Risk Paradigm, PG&E is developing a Distribution (Dx) Asset Risk Model (the Model), tuned for Wildfire Risk, which will:

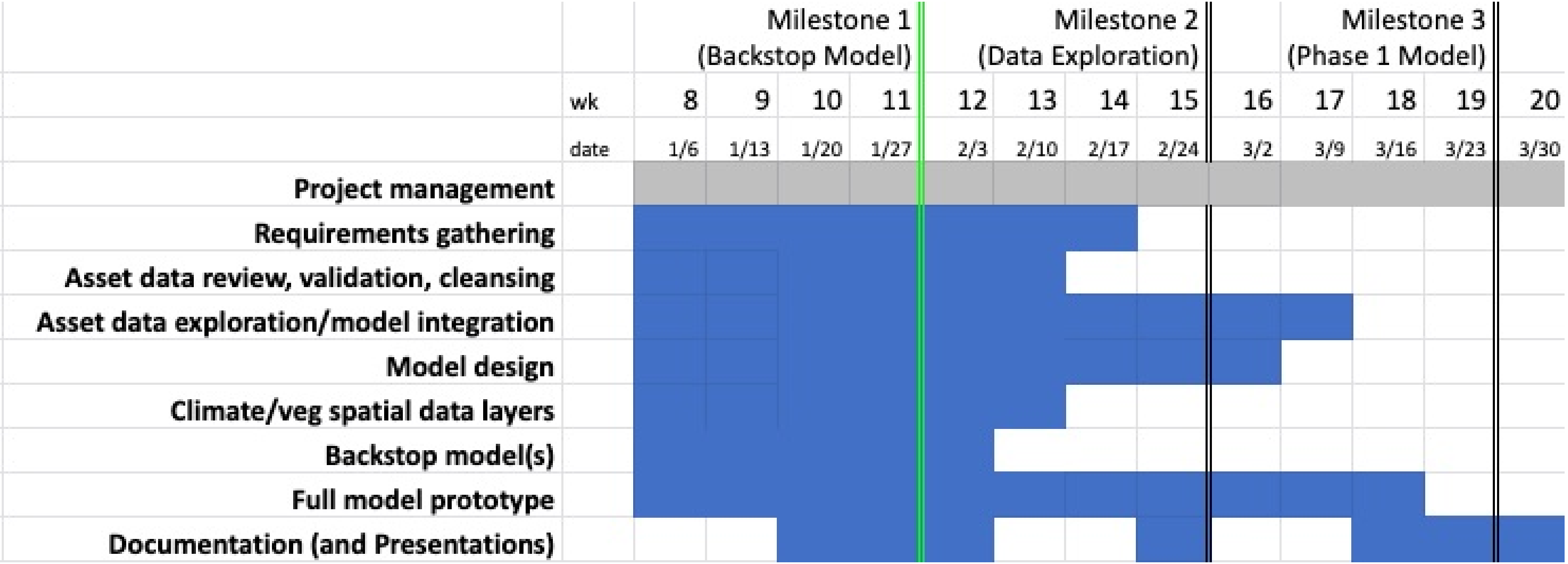
- Provide situational awareness of the current wildfire risk on the Dx system
- Enable risk-informed decision making in the budget planning process
- Allow PG&E to report risk reduction to regulatory entities

Phase 1 key objectives and desired outcomes (end of March 2020)

A Prototype Model has been developed for one or more Dx asset classes such that:

- Statistical experts within PG&E verify that the Model is developed on a solid statistical foundation
- Risk calculation methodology has been approved by EORM
- Prototype results are used to inform the Q1 Dx asset planning budget adjustments.
- The Prototype will only consider Probability of Failure and Wildfire Risk
- MAVF and other components of asset risk will be included in Phase II

Project schedule

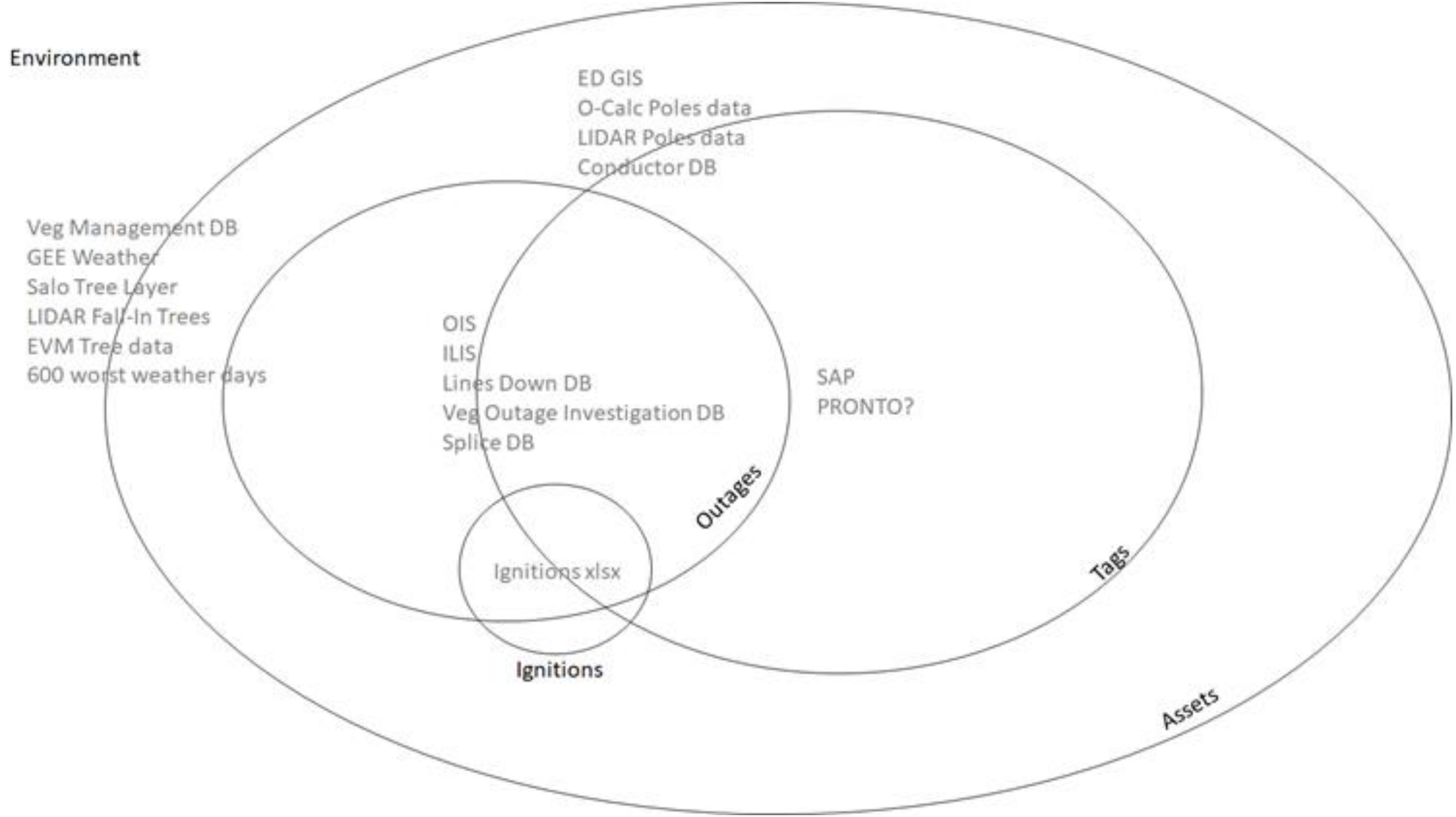


Note that project work commenced in Nov, 2019 - not shown on this chart for clarity

Where we are now

- Requirements gathering and data research
  - 17+ meetings over 2 months with key stakeholders
  - Comprehensive catalog of relevant data sets, written documents defining the modeling problem(s) and related approaches and tradeoffs
- Infrastructure - cloud-based data science environment
  - jupyter/python/geopandas/rasterio
  - AWS SageMaker environment - same platform as ARAD
  - Team members have access to: PG&E private data, collaboration tools, source code repositories, etc
- Infrastructure - modeling in software
  - Pipeline for gathering and formatting geo-spatial data
  - Pipeline for augmenting any location (by lat/lon) with geo-spatial information
    - Ignition sites, Dx grid, etc.
  - Software system to prepare data for, configure, execute, and post-process MaxEnt modeling runs
- Backstop model

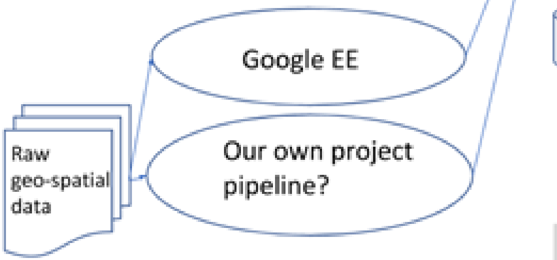
Asset and event data set relationships



Geo-data processing pipeline

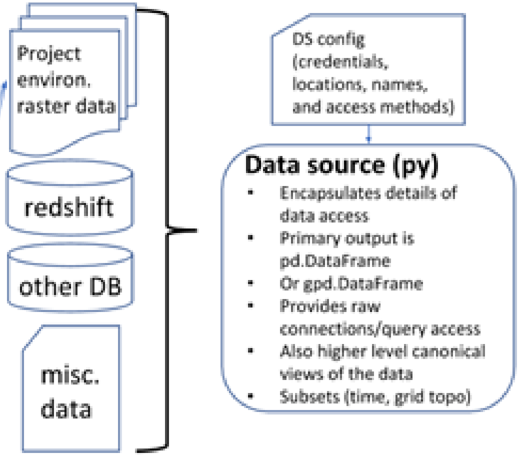
Batch environmental data preparation

1. Choose data source(s)
2. Isolate dates
    1. climate trends
    2. weather at event
3. Isolate specific fields (bands)
4. Summarize with max/min/mean, etc.
5. Render to project env data raster(s)
6. GEE (1a) and project-specific (2) pipelines




```
# pseudo code for GEE implementation

# official GEE name/account local data
data_set_name = 'NOAA/NWS/RTMA'
bands = ['WIND','LST', etc.]
date_range = ['2016-01-01','2019-01-01']
summary_fn = ['min','max','mean','5pct','95pct']
raw_data = load_bands(data_set_name, bands, date_range)
# one band per band x summary_fn
out_data = summarize_data(raw_data, summarize_fn)
# Export.image.toAsset, to tifs in AWS, if possible
save_out(out_data, [resolution, location, etc.])
```

Data storage and access




```
from dx_risk import data_source as ds

ds.init('dx_risk_data_config.yaml')

# low level interface based on connection config
conn = ds.get_connection('redshift')
tags = pd.read_sql( 'select * from SAP.tag_table',
                    con = conn)

# where is the data?
data_dir = ds.RASTER_DATA_DIR
dblb = ds.get_raster('dry_bulb_file_name.tif')

# high level interface based on config
tags2 = ds.get('inspection_tags')
dblb2 = ds.get('raster.dry_bulb')
```

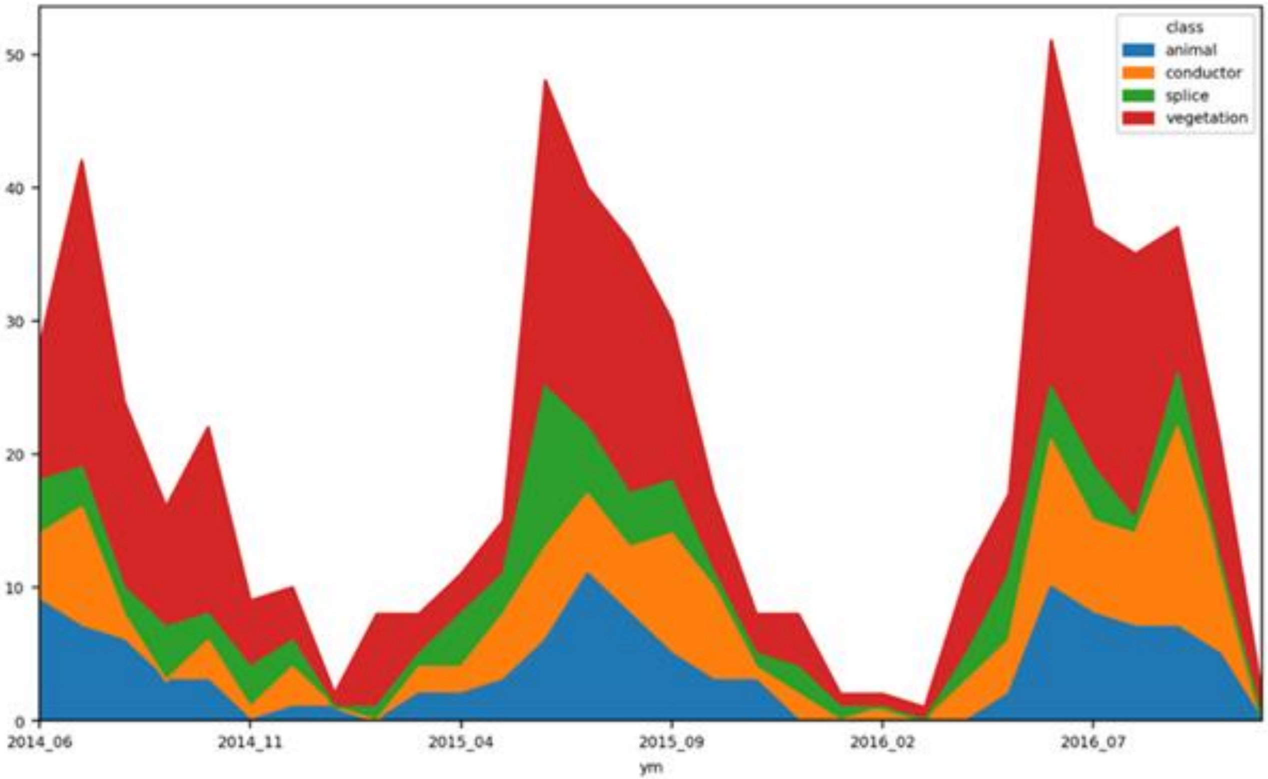
Asset/event augmentation

1. Select and load time/space points
    1. Input (g)pd.DataFrame
    2. GEE input: csv
2. Select env data to augment with
3. Bulk lookup in raster(s) by lat/lon
4. Output augmented data
    1. gpd.DataFrame
    2. GEE: csv ready for gpd import

```
from dx_risk import data_source as ds
From dx_risk import geo_tagger

ds.init('dx_risk_data_config.yaml')

# geo_tagger is passed whatever config it
# needs to find raster data and
# interpret named augmentation fields
gt = geo_tagger( ds.RASTER_DATA_DIR,
                 ds.RASTER_CONFIG)

ign = ds.get('ignitions') # gpd.DataFrame
ign_plus = gt.augment( ign,
                       ['dry_bulb.mean',
                        'VPD.mean',
                        etc.        ] )

# ign_plus becomes what Chris calls
# "vector data" in the Maxent framework

# All of this code is upstream from
# his ipynb example
```

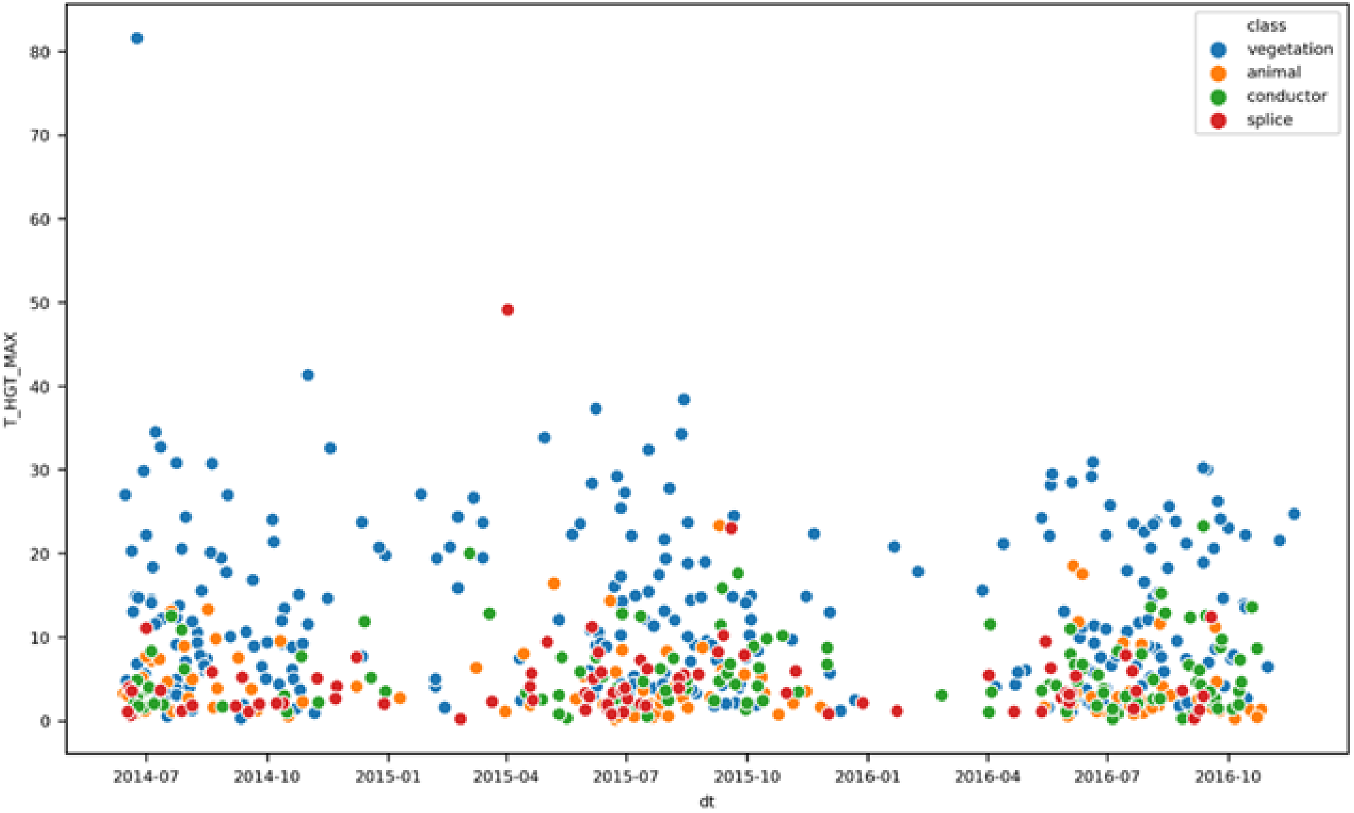
Vegetation caused ignitions

Model progress - Seasonal frequency of ignitions by ignition cause class



Model progress - Ignitions by tree height, date, class

Max tree height (m) per 100m "pixel"



date

Key modeling considerations

1. Ignitions are rare events
2. Multiple points of failure precede an ignition
3. The drivers of failures include both endogenous and exogenous processes
4. Failures can result from instantaneous and cumulative processes
5. Multiple forms of uncertainty in available data
    a. Relational topology unclear (e.g., hard to link outages to wire-downs to ignitions)
    b. Spatial uncertainty high (recorded positions are often imprecise)
6. Physical models are robust and easy to interpret, but only describe a few processes
7. Statistical models can identify novel failure patterns, but are easily biased in predicting rare events
8. Needs to be sensitive to management activities
9. Needs to improve over time as new data comes in from the field

Spatial model: structure and inputs

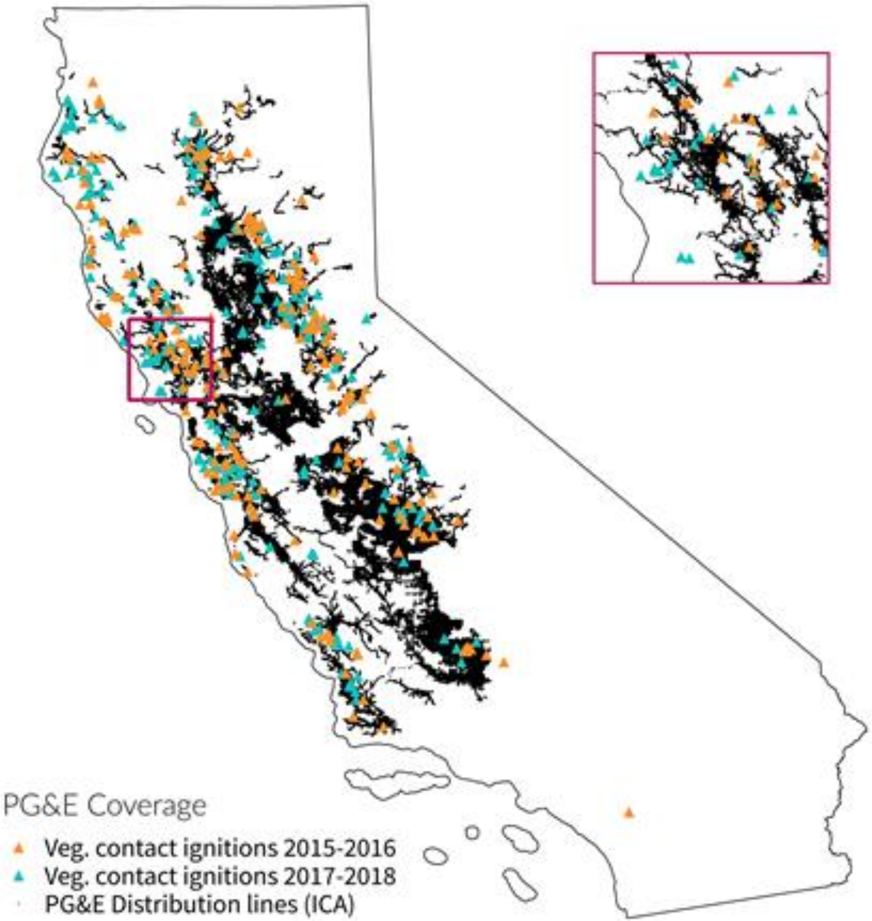
Ignitions 2014-2018

Dx locations from Integration

Capacity Assessment spatial files

Ignitions from PG&E internal data

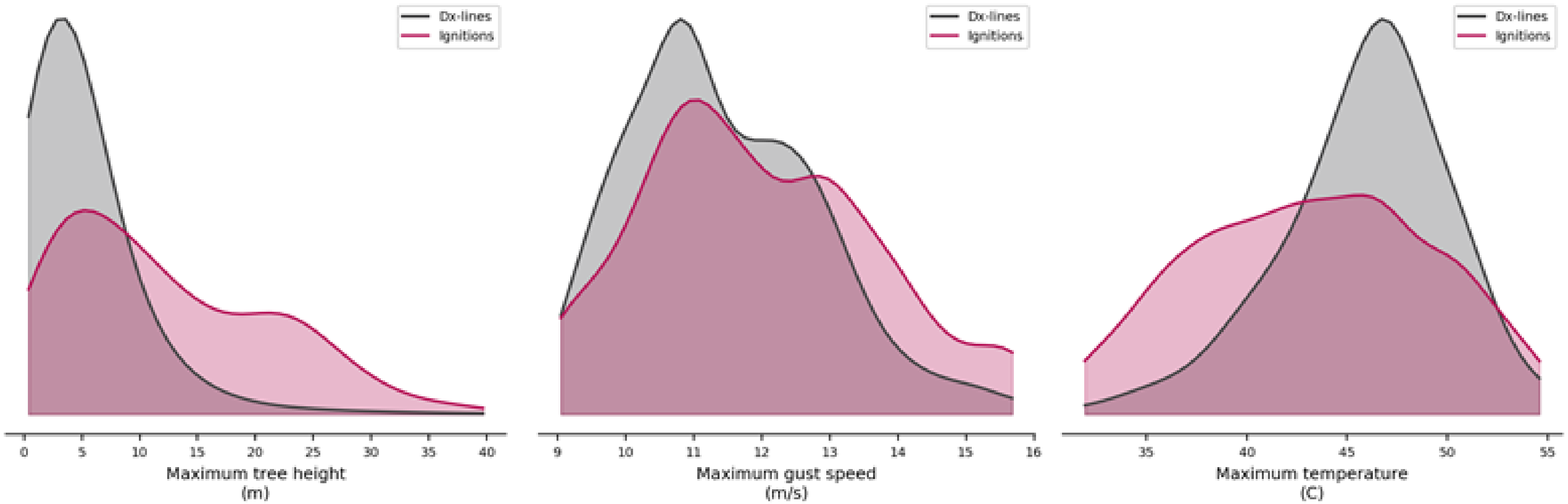
Restricted to vegetation contact



High level overview of modeling approach

MaxEnt, or maximum (information) entropy, models were developed to derive probable ranges of species given the set of locations where they have been sited.

In or case, MaxEnt models discriminate between environmental conditions at the sites of ignitions and a set of “background” locations without reported ignitions, where our background is the full Dx grid.



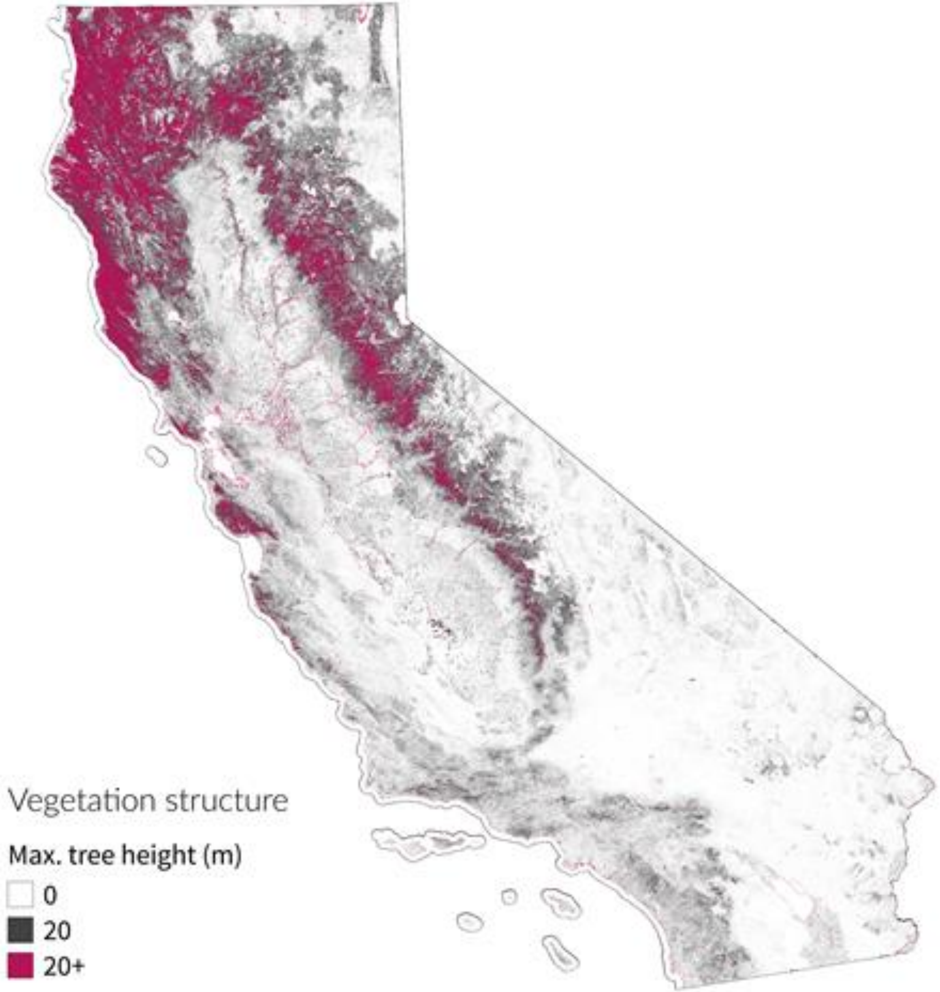
Salo tree height data

Forest Net outputs

100m resolution

Full state coverage

Regular updates possible



Input covariate data

Class	Covariate	Unit	Spatial scale	Notes
Vegetation	Mean tree height	(m)	100 m*	Mean tree height of area around asset
	Tallest nearby trees	(m)	100 m*	Calculated as maximum tree height in area around an asset
Wind	Mean wind speed	(m/s)	2,500 m	From RTMA
	Local wind speed maximum	(m/s)	2,500 m	Calculated as the 99th percentile of local wind speeds
Gust	Mean gust speed	(m/s)	2,500 m	From RTMA
	Local gust speed maximum	(m/s)	2,500 m	Calculated as the 99th percentile of local gust speeds
Temperature	Mean temperature	(°C)	1,000 m	From MODIS LST
	Local temperature maximum	(°C)	1,000 m	Calculated as the 99th percentile of local temperatures
Topography	Local topographic position	unitless	100 m*	From the topographic position index (TPI)
	Landscape topographic position	unitless	1,000 m*	Calculating TPI at fine and large scales allows distinguishing multiple landforms (i.e. difference in local and landscape topography)

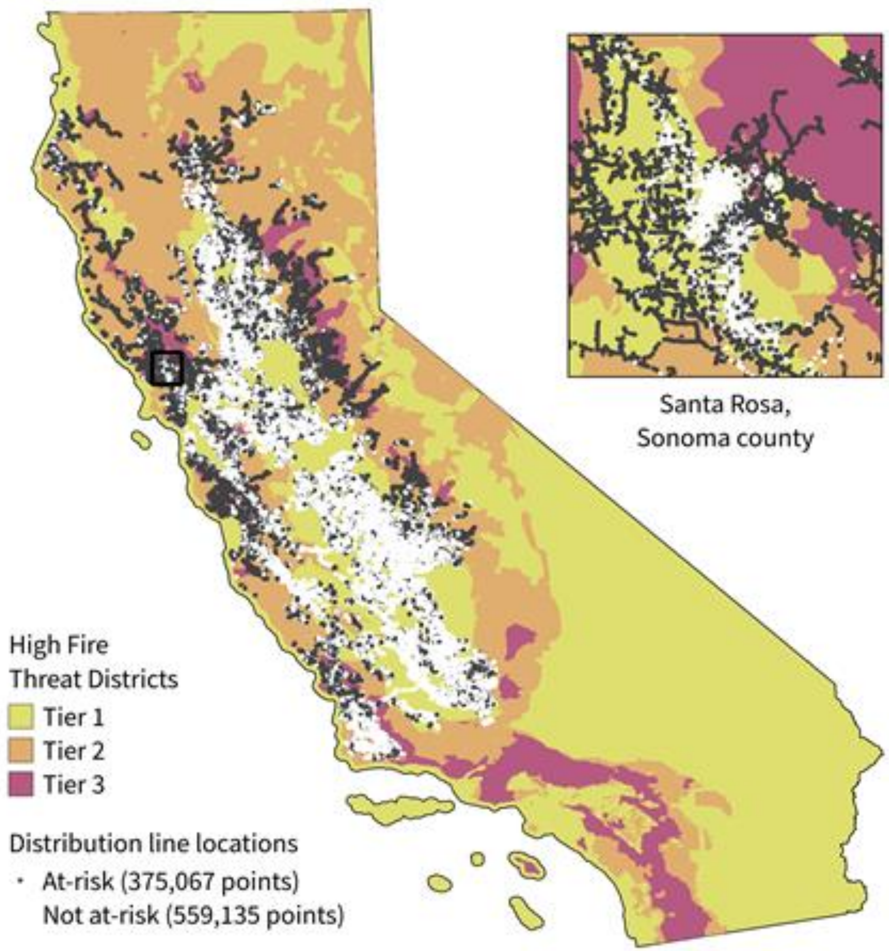
*can be calculated at finer spatial scales

Backstop model: results

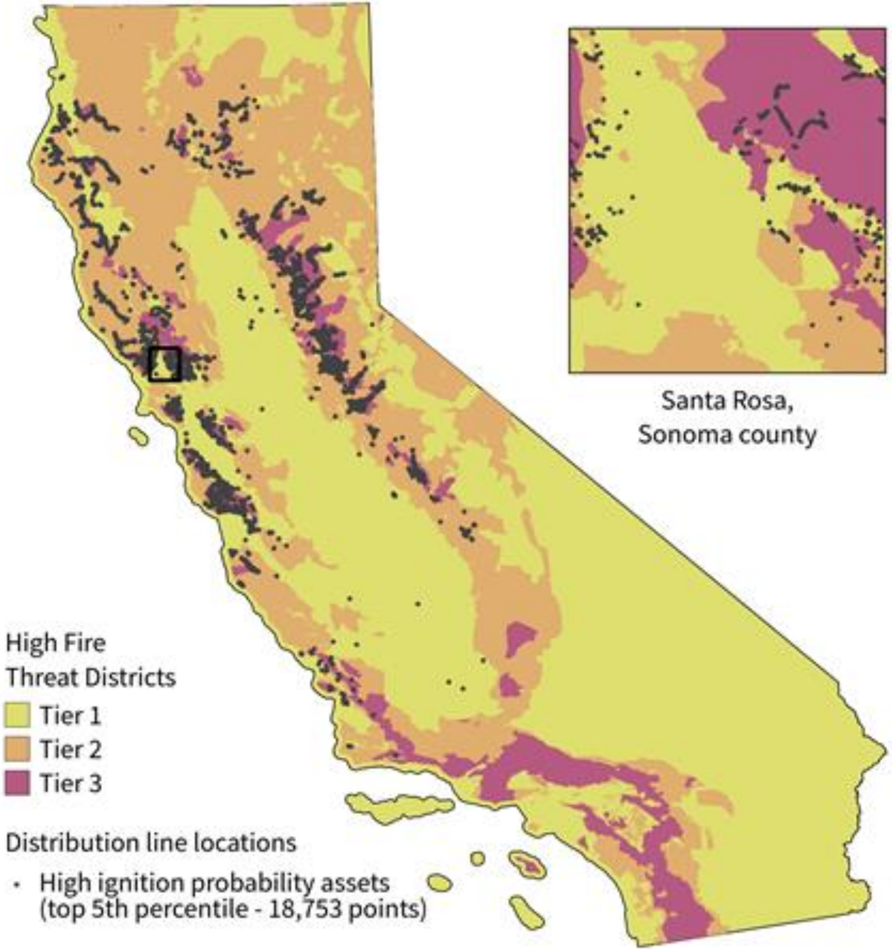
Low and high risk assets against HFTDs

At risk threshold set to produce

5% omission rate



Top %5 of of predicted Ignition probability

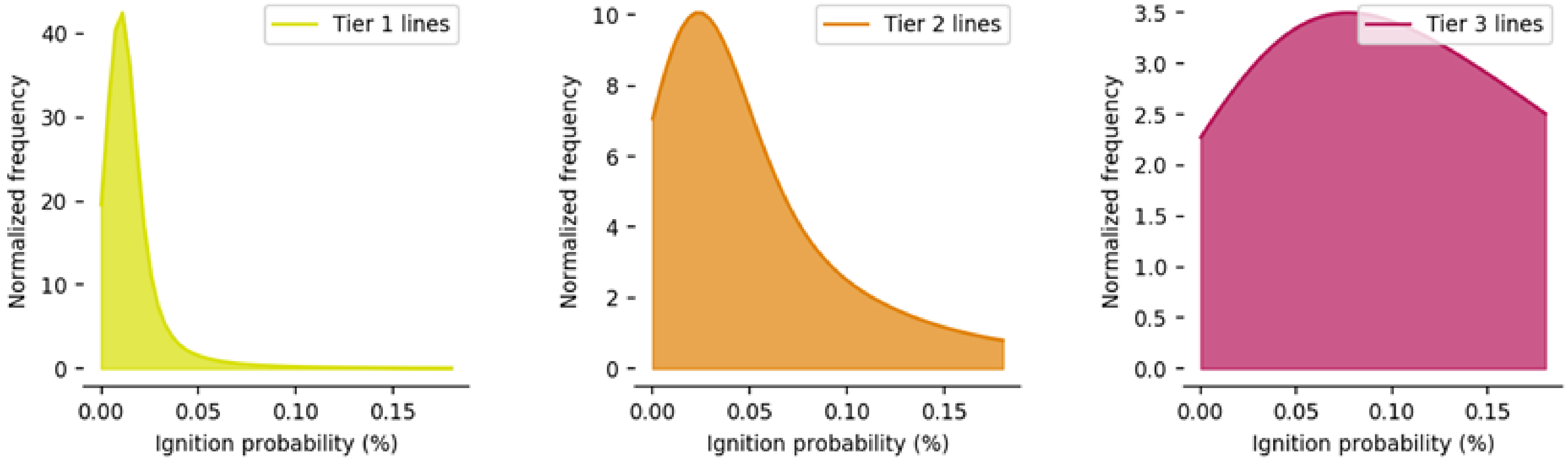


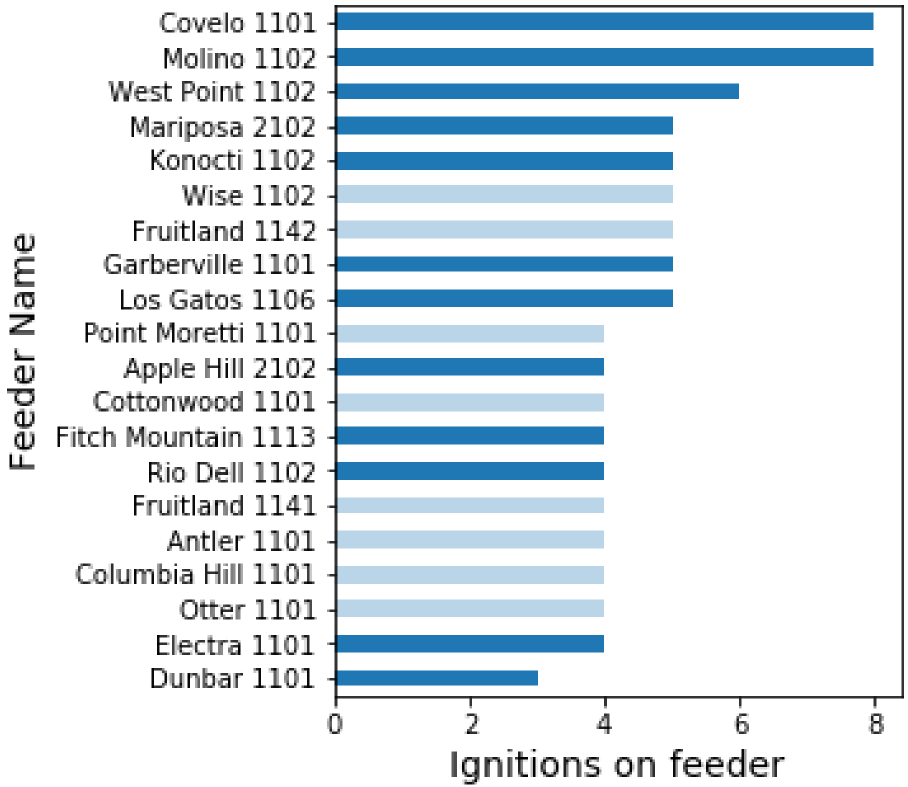
Model performance

	Training (2015-2016)	Testing (2017-2018)
AUC (probability the model can distinguish high risk territory from low risk territory - 0.5 is random chance; 1.0 is never wrong)	0.765	0.755
Recall at 95% omission TP / (TP + FN) (Fraction of ignitions found within the high risk territory)	0.799	0.781
Predicted ignition count	229.1	200.0
Observed ignition count	210	266

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

Asset ignition probabilities by HFTD Tier





Data sources and how used

Ignitions data (bars) [limited to veg]

Our predictions (filtered 100 highest risk feeders)

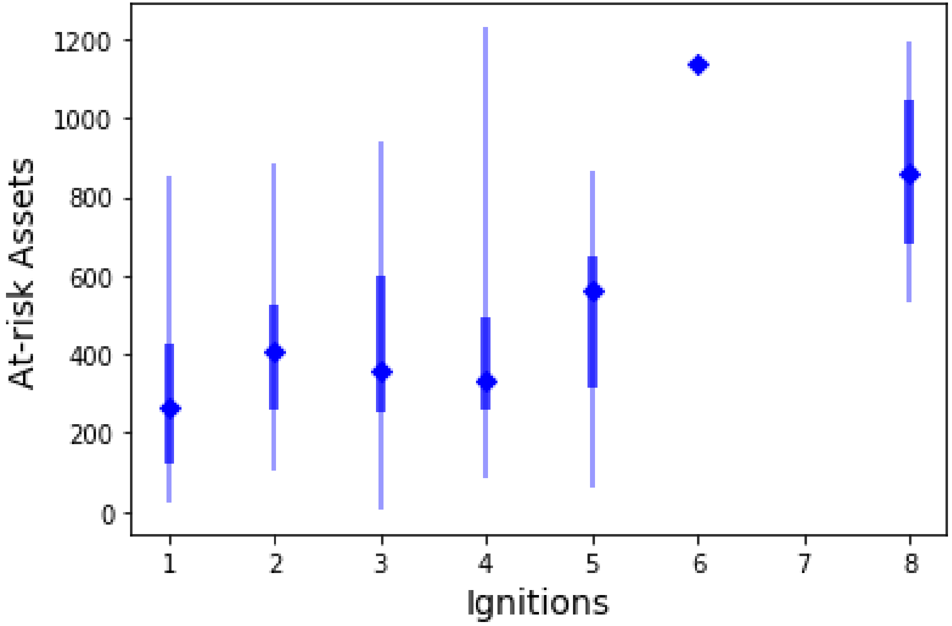
ICA data (filtered for inclusion)

Description

Number of ignitions per feeder. Dark blue indicates feeders among the 100 feeders with the highest risk score.

Comments and Caveats

Limited to feeders included in both ignitions and ICA datasets.



Data sources and how used

Our predictions (vertical axis)

Ignition data (counts along x axis)

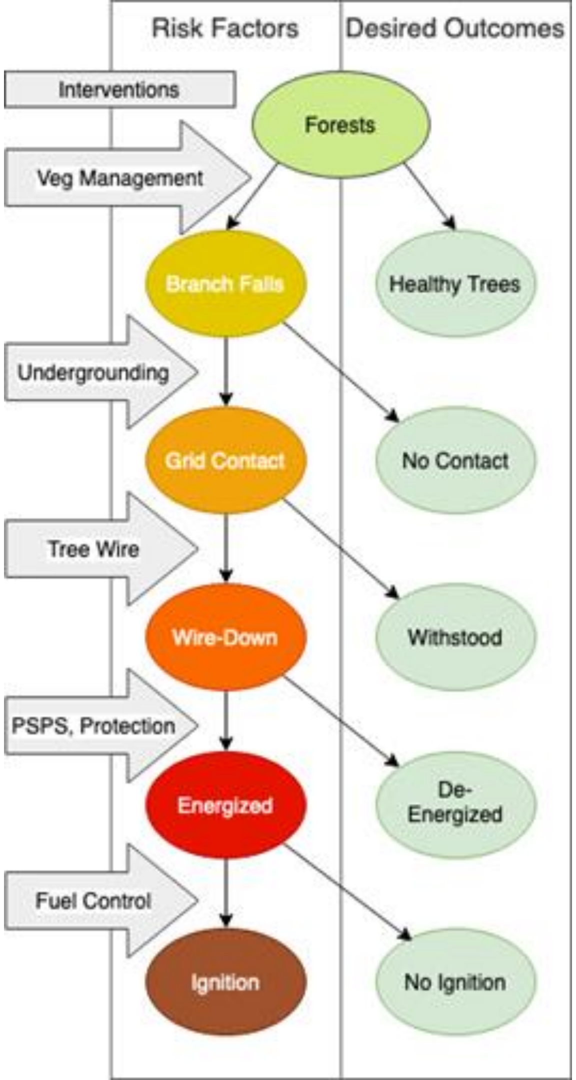
Description

Bar chart showing risk distribution by feeder (y-axis) grouped by the number of ignitions that actually occurred on that feeder (x-axis)

Comments and Caveats

Only one feeder had 6 ignitions, and none had 7.

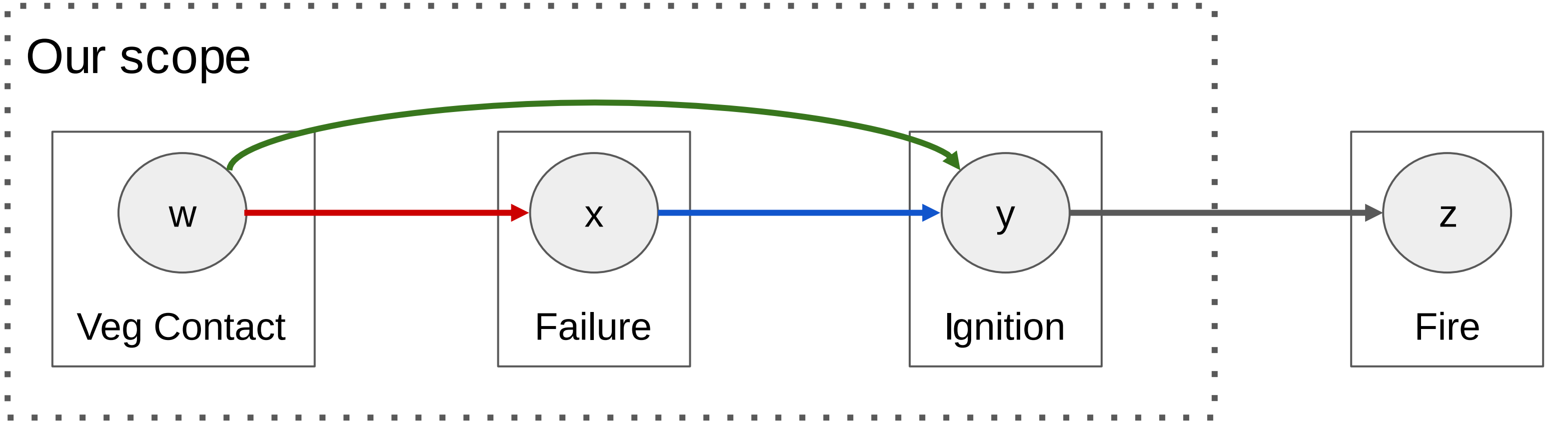
Our next steps


Risk Factors
Desired Outcomes
Interventions
Forests
Veg Management
Branch Falls
Healthy Trees
Undergrounding
Grid Contact
No Contact
Tree Wire
Wire-Down
Withstood
PSPS, Protection
Energized
De-
Energized
Fuel Control
Ignition
No Ignition

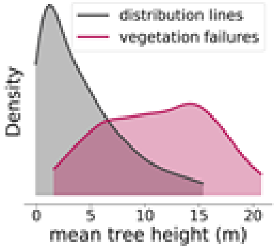
What is the output of our model?

Estimated risk reduction due to decisions related to:

1. **Veg management**
2. **Grid hardening**
3. **Protection**



Integrating multiple functional forms



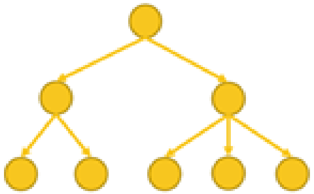
$$log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

MaxEnt

- Calculates p(failure) over long time period to characterize failure probabilities under aggregate environmental conditions
- Serves as a prior estimate of failure probability to predict short-term failure probability

Fragility curves

- Evaluate long-term endogenous p(failure)
- Serves as a physically-based prior estimate of failure probability to predict short-term failure probability



Decision / regression trees

- Uses prior estimates of exogenous/endogenous failures as features to split trees
- Can experiment to identify the best time scales for analysis

Discussion

Using the thresholded predictions, we identified 375,067 assets that were at-risk, or 40.1% of the

934,202 conductor locations (Fig. 3). Of those assets, 284,250 of the HFTD Tier 1 assets were classified

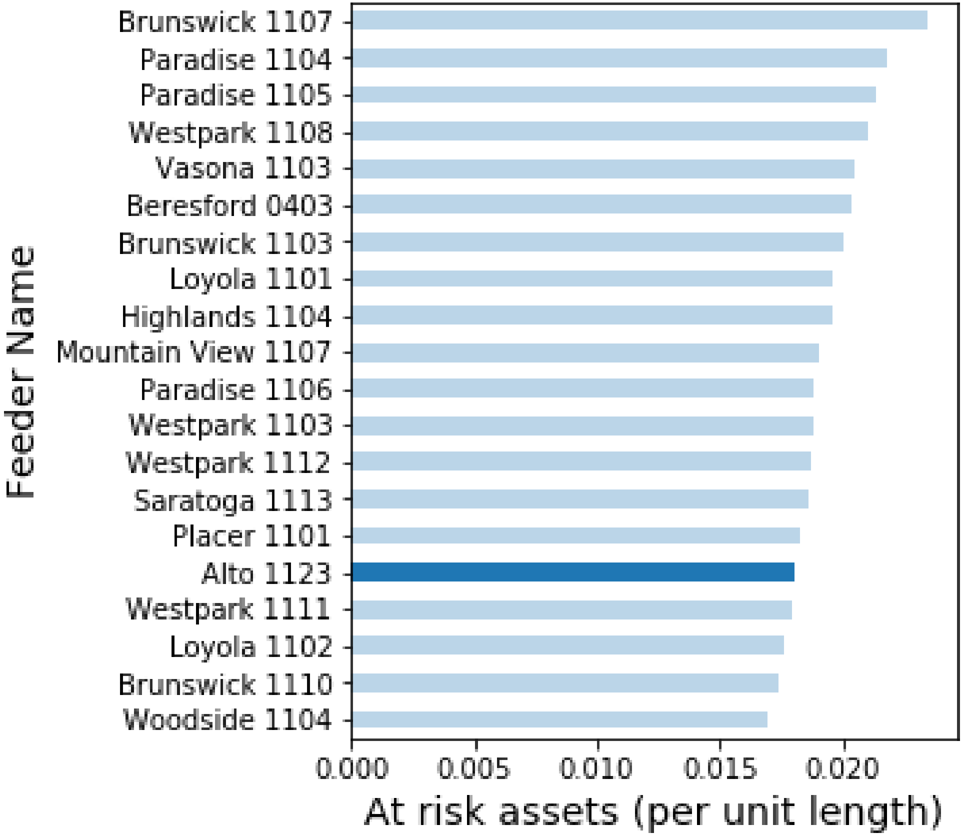
as at-risk (34.4% of 825,511 assets), 61,013 of the Tier 2 assets were classified as at-risk (79.1%), and

29,804 of the Tier 3 assets were classified as at-risk (94.4%) based on the 2015-2016 predictions.

For the 2015-2016 data, the sum of all predicted ignition probabilities was 229.1,

compared to 210 observations during that period. For the 2017-2018 data, the sum of all predicted

ignition probabilities was 200.0 for 2018, compared to 266 observations during that period.



Data sources and how used

Our predictions (width of bar)

ICA data (normalize by length)

Ignition data (color)

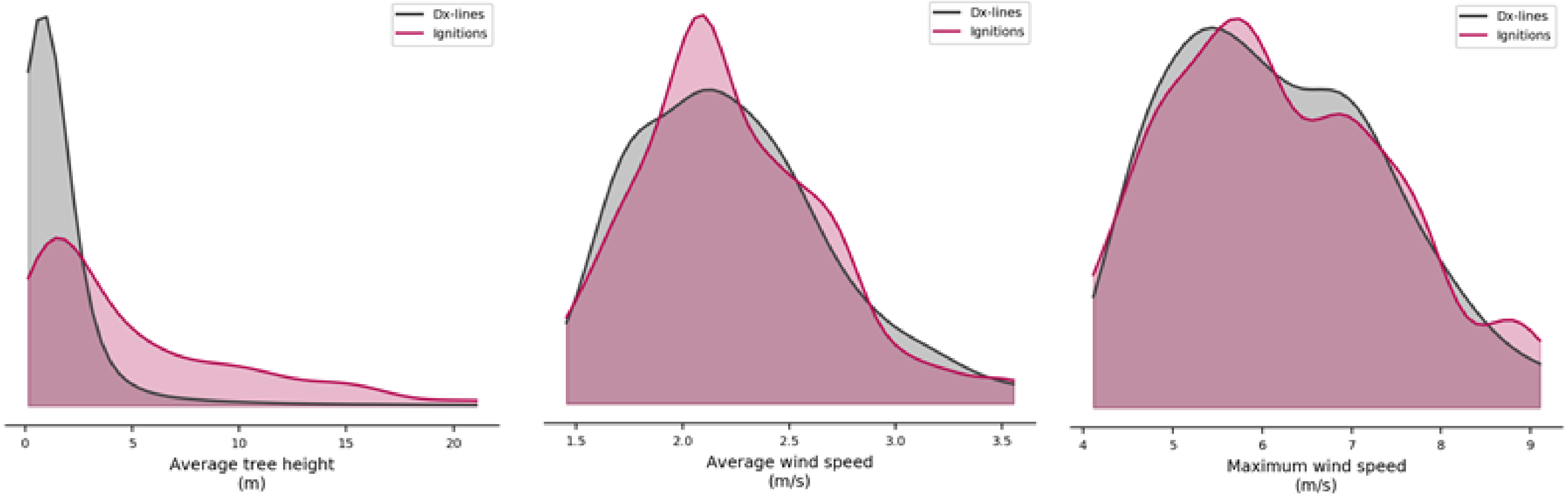
Description

Number of at risk assets normalized by length of distribution circuit.

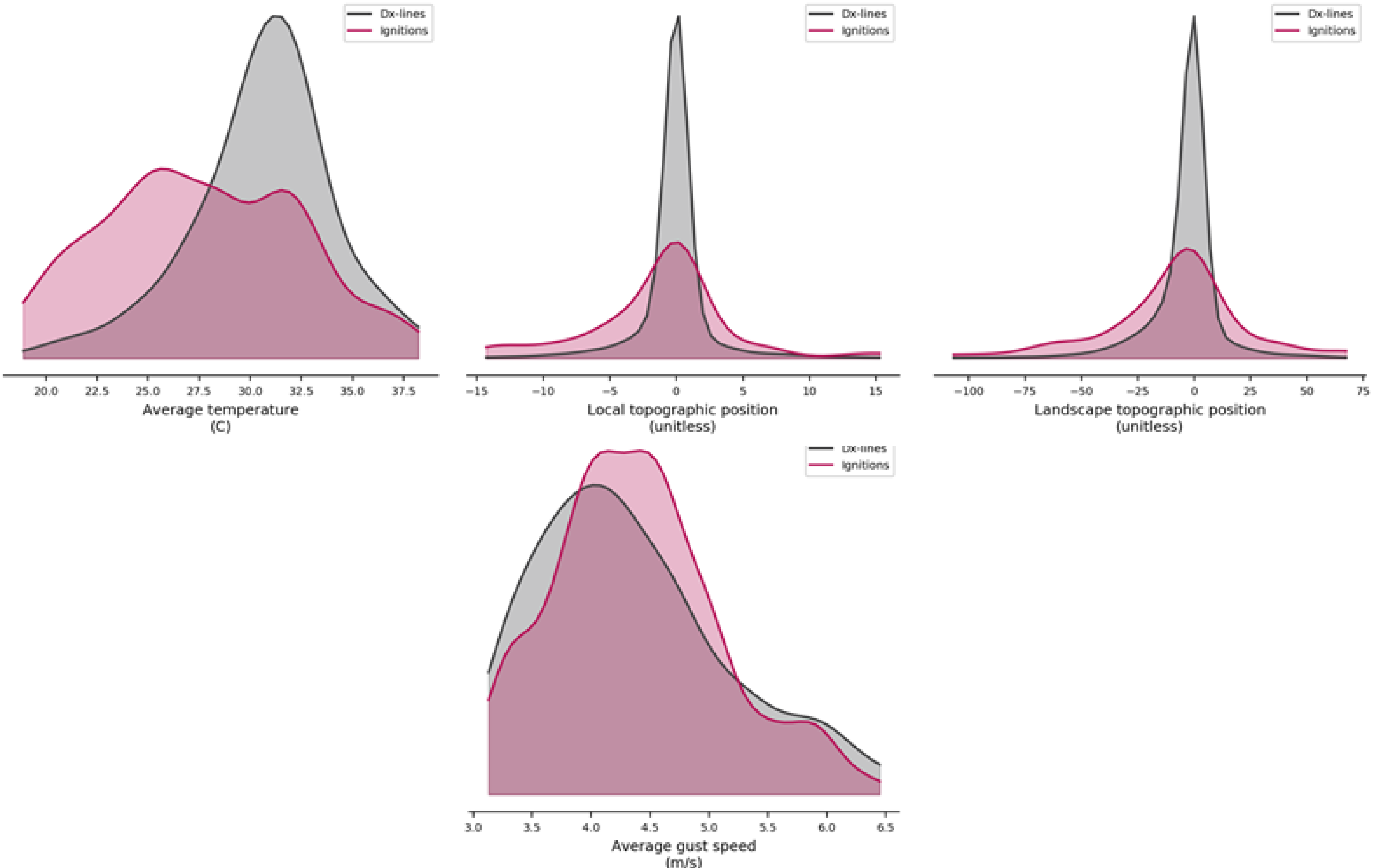
Colors indicates feeders that also rank highly in ignitions per unit length!

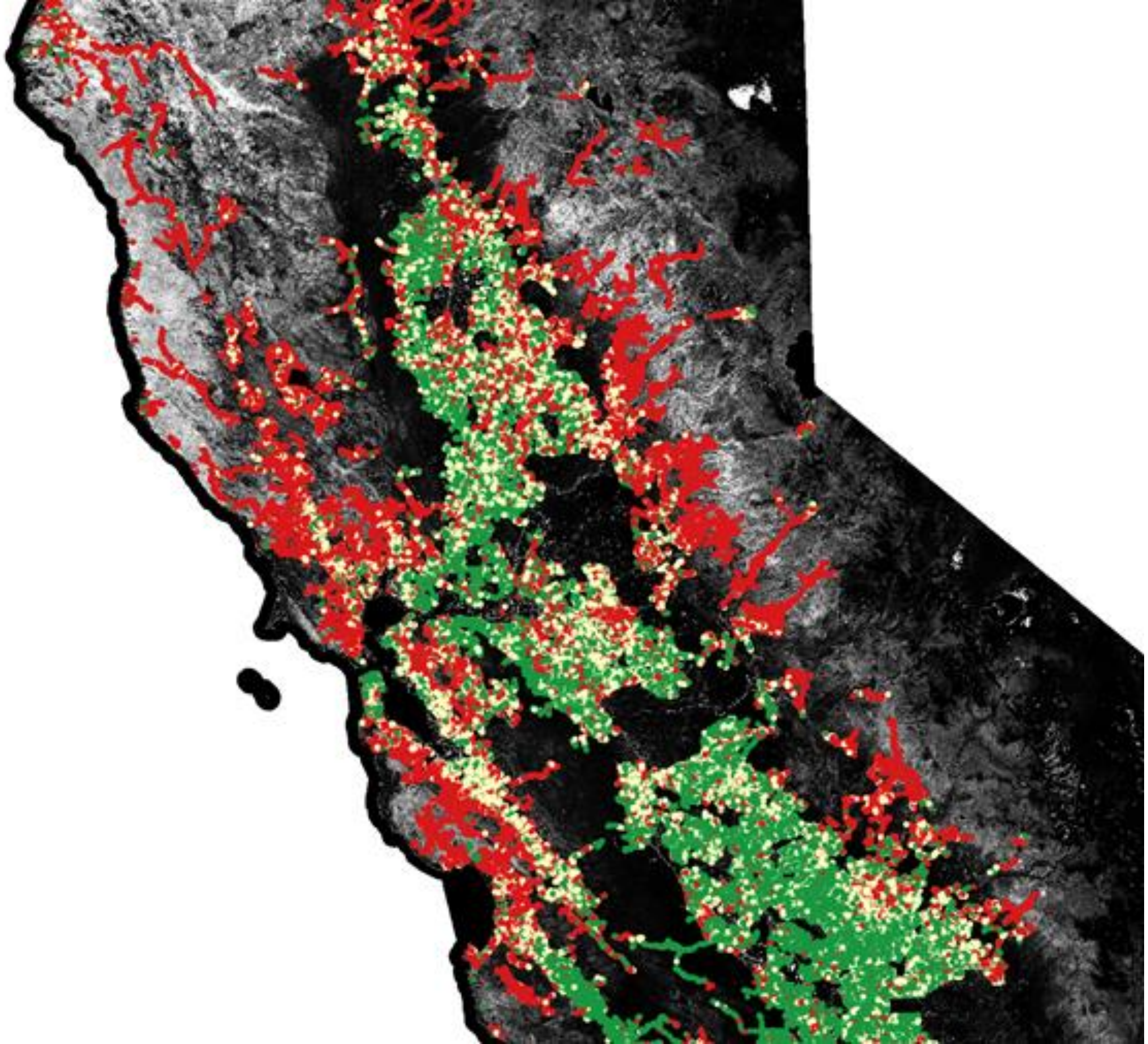
Comments and Caveats

Paradise and Grass Valley at risk



Dx-lines
Ignitions
0
5
10
15
20
Average tree height
(m)
Dx-lines
Ignitions
1.5
2.0
2.5
3.0
3.5
Average wind speed
(m/s)
Dx-lines
Ignitions
4
5
6
7
8
9
Maximum wind speed
(m/s)


Dx-lines
Ignitions
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Average temperature
(C)
Dx-lines
Ignitions
−15
−10
−5
0
5
10
15
Local topographic position
(unitless)
Dx-lines
Ignitions
−100
−75
−50
−25
0
25
50
75
Landscape topographic position
(unitless)
Ignitions
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Average gust speed
(m/s)



What our Phase 1 Model should do

1. Appropriately weight failure events to ensure high recall
2. Represent the multiple failure processes that lead to ignitions
3. Include predictive features that capture steady state and dynamic conditions over multiple timescales
4. Establish simple heuristics for cross-referencing datasets that can improve over time
5. Flag outliers to evaluate whether an event represents a novel failure process or is highly uncertain
6. Include results from physical models in a statistical modeling framework