Predicting annual vegetation-driven ignition probabilities using a spatial model—a maximum entropy analysis

Convergence Data Analytics, Salo Sciences, Presence Product Group

Introduction

In support of risk-based Electric Operations planning, PG&E is developing a Distribution (Dx) Asset Risk Model, which is 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 $risk = ignition\ probability \cdot consequences$. Both the likelihood and the consequences of an ignition are conditioned, to a degree, on the environmental conditions experienced by distribution assets prior to and during an ignition. 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 all hazard trees. However, the degree to which multiple environmental conditions interact to determine the probability of failure is not well characterized at a systems level.

In this report, we describe a novel approach for characterizing ignition probabilities across PG&E's network of distribution assets as conditioned by multiple environmental patterns: wind and gust speeds, temperature, vegetation structure, and topography. To maintain an aggressive project timeline—while working to gain the credentials to access internal data—we developed this model using public data sources. These data included ignitions reported to the CPUC from 2014-2018, Dx grid locations derived from the spatial data set published to support the Integration Capacity Analysis Map, weather station data and satellite-derived vegetation data. We focused on predicting ignitions caused by vegetation contact, which were the source of 641 of 2,233 ignitions between 2014 and 2018; the most common category of reported ignitions. Since we didn't know *a priori* how interacting environmental conditions at a systems level lead to failures, it was a challenge to define a model to characterize the roles of multiple covarying environmental patterns. In this analysis, we focused on evaluating the relative importance of annual environmental patterns. By aggregating data to an annual time scale, this report serves as an exploratory analysis of how long-term environmental patterns condition ignition probabilities. We used this to identify where to focus efforts to characterize how fine-scale temporal (i.e. hourly, daily) environmental patterns precede ignitions, highlighted in *Next Steps*.

Methods

Model form

We used a spatially-explicit maximum entropy model, MaxEnt[1,2], to link statewide maps of environmental patterns to the locations of PG&E distribution assets and ignition events in order to predict the probability of an ignition for each asset as conditioned by environmental conditions (Fig. 1). The

[1] Elith *et al.* 2010

[2] Software available for download at this link

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

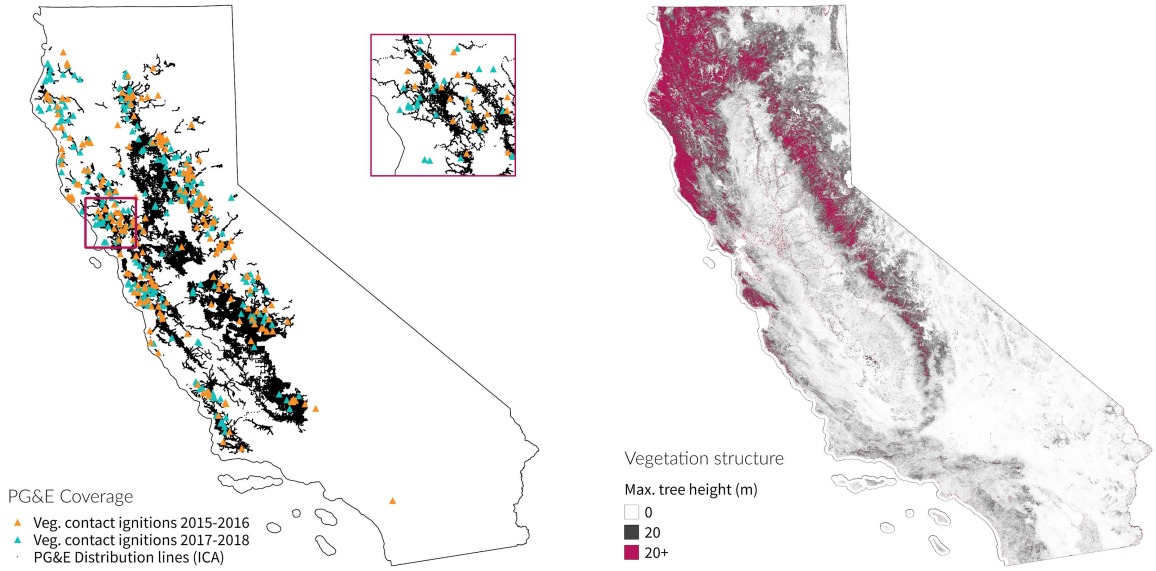


Figure 1. (left) The locations of PG&E distribution lines and ignition locations. Ignition data provided by PG&E to the CPUC, distribution data provided through PG&E and accessed via the Interconnection Capacity Analysis (ICA)[3]. (right) Maximum tree heights mapped across the state of California. Vegetation 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.

of environmental conditions experienced by all similar assets. This is formulated mathematically as:

$$Pr(y = 1|z) = f_1(z) \cdot Pr(y = 1) \div f(z) \quad (1)$$

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.

Input data

We used the PG&E ignitions database to generate the ignition locations ($y = 1$). This database contains the locations and source of all PG&E-related ignitions from 2013 to 2019. We subset these ignitions using the "Contact From Object" and "Date" fields to generate a list of ignitions caused by vegetation contact over the years 2015-2018. To characterize z, we generated and evaluated 10 statewide

[3] Data available at this link

environmental covariates using data from 2015-2018. These included mean wind speeds, maximum wind speeds, mean gust speeds, maximum gust speeds, mean temperatures, maximum temperatures, mean surrounding tree height, maximum surrounding tree height, local topographic position, and landscape topographic position (Fig. 2)[4]. The "Input covariate data" section of Appendix 1 describes the covariates in more detail. Using MaxEnt, we computed feature transformations $f(z)$ using the *hinge* and *product* feature options. The hinge transformation allows piecewise linear functions to be fit to each environmental pattern, and product transformations fit features based on interactions between covariates (e.g. max wind speed · mean temperature). The spatial extent of $f(z)$ was constrained to just the locations of distribution grid assets, based on PG&E's publicly-accessible Interconnection Capacity Analysis (ICA) locations. We used the "LineDetail" dataset to extract distribution line data, converted these data to point locations using the center of each line, and removed assets where we lacked environmental data, resulting in 934,202 distribution point locations. To characterize the spatial distributions of ignition probabilities, and develop an early sense for the risks posed by ignitions, we subset the distribution point locations according to the CPUC High Fire Threat Districts (HFTDs)[5].

Model outputs

For $Pr(y = 1|z)$, MaxEnt generates multiple conditional probability estimates: relative occurrence rates, omission rates, and probability scores[6]. These scores are assigned to every location on the map where environmental data are available.

- The **relative occurrence rate** ranks the relative probability of ignitions under local environmental conditions, which is useful for calculating rank-ordered probability scores.
- These relative rates are then transformed into **cumulative probability scores**—or **omission rates**—which are scores from 0-100. Thresholding this score will approximate the percent of ignitions omitted from prediction (e.g., setting a threshold at >20 will produce a binary map of likely/unlikely ignition locations that should omit 20% of ignition locations).
- These relative scores do not contain information on how frequently ignitions tend to occur, however. Translating relative to **absolute probabilities** was performed by computing a mean-centered logistic regression. This step depends on a model parameter, $\boldsymbol{\tau}$, which scales probabilities according to the frequency that ignitions occur across the grid.

In our case, ignitions are rare—there have been approximately 100 ignitions per year across the nearly 1 million asset locations. We empirically estimated $\boldsymbol{\tau}$ as 0.0016, or a 0.16% probability that an asset experiencing the conditions that lead to failure will result in an ignition. Absolute probability scores are highly sensitive to this uncertain estimate—which is explored further in *Next Steps*.

Evaluating model performance

We evaluated model performance by splitting the ignitions data into two groups: ignitions from 2015-2016 and from 2017-2018. We trained the model to calculate ignition probabilities using the 2015-2016 data (210 locations) and evaluate its performance on the ignitions from 2017-2018 (266 locations). We report two model performance metrics: recall scores and the area under the receiver

[4] These data are described in *Appendix 1*, Table 1
[5] HFTD data available at this link.
[6] Explained in detail by Merow *et al*. 2013.

operator curve (AUC). To compute recall scores, we set a threshold of >5 on the omission rate maps to approximate a 95% confidence interval for predicting, in a binary sense, locations where vegetation

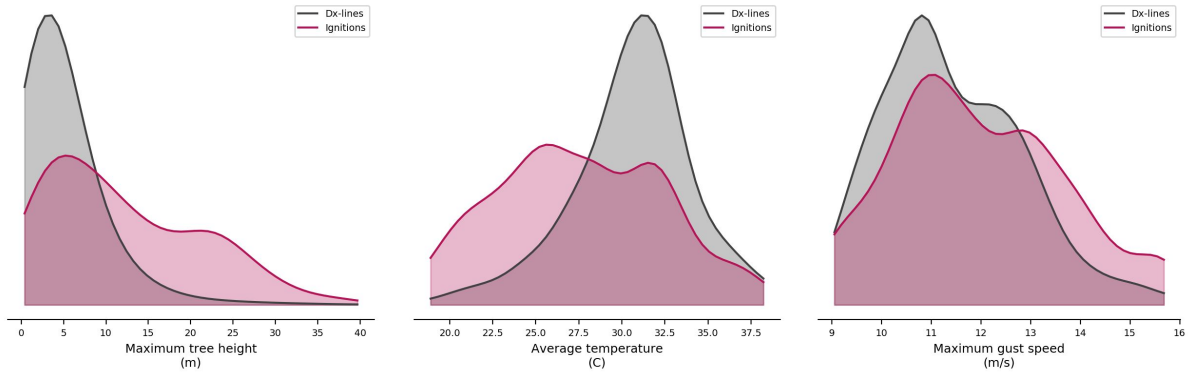


Figure 2. Normalized density distribution plots comparing the environmental conditions between all distribution lines (grey) and all ignition locations (red) for maximum tree height, average temperature, and maximum gust speed (from left to right). Ignition locations were more likely to occur in areas with taller trees nearby and in areas with lower average temperatures compared to the full population of distribution lines. Maximum gust speeds at ignition locations were similar to gust speeds at all distribution lines, with some sites experiencing disproportionately high maximum gust speeds.

contact could lead to ignitions (Fig. 3; black dots) and where ignitions were unlikely (Fig. 3; white dots). For rhetorical simplicity, we refer to these thresholds as labeling assets as at-risk and not at-risk of vegetation contact, respectively. 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 map of omission rates accurately constrains the extent where ignitions could occur from contact.

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

Table 1. Confusion matrix for the four prediction outcomes from a binary prediction of ignition likelihood.

Our second model performance metric, AUC, estimates separability. In our case, AUC scores represent how well at-risk assets can be discriminated from assets where contact ignitions are unlikely to occur[7]. AUC is scored from 0.5 to 1 based on the true positive and false positive rates, where a score of 0.5 indicates two populations that cannot be distinguished, and a score of 1 indicates perfect separability. 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. Here, we report recall and AUC scores for the training and testing datasets to evaluate the degree to which the model is overfit to the environmental conditions in the training data. We also include the results of a sensitivity analysis, where a

[7] Read more about AUC at this link.

leave-one-out jackknife procedure was applied to each environmental covariate in order to evaluate how model performance changed with the inclusion or exclusion of each covariate.

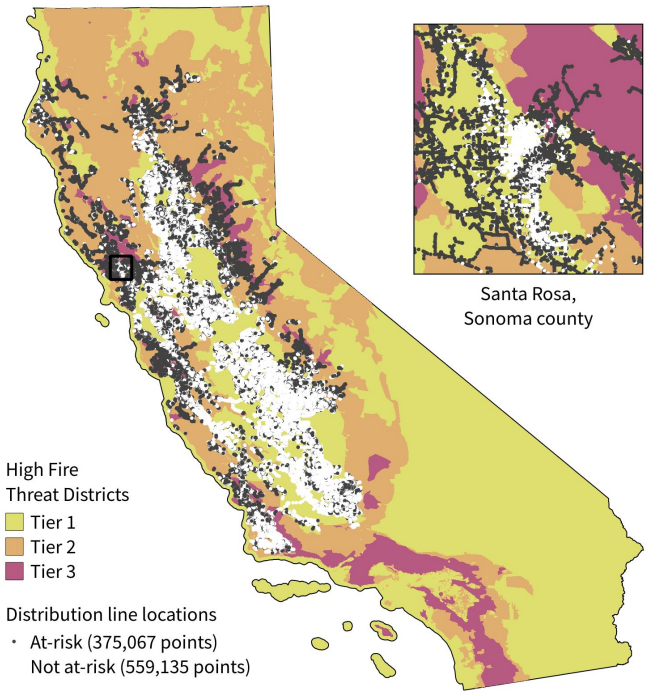


Figure 3. Map of the distribution line locations, colored by assets that were predicted as at-risk (black) and not at-risk (white). This delineation between assets was calculated by setting a 5% threshold on the omission rate predictions to estimate where assets are or are not likely exposed to vegetation contact.

Results

We trained a MaxEnt model to predict relative ignition probabilities using 210 ignition locations from 2015-2016 based on 10 environmental covariates. These relative probabilities were scaled to omission rates then thresholded to approximate a 95% confidence interval for predicting which assets are/are not at risk of vegetation contact ignitions. Based on the training data, we calculated AUC scores of 0.765—or, a 76.5% chance the model can distinguish between assets where ignitions will or will not occur. Using the omission rate threshold, we calculated recall scores of 0.799—80% of the ignitions are contained within the areas defined by the top 95% of failure probabilities. When we applied the model to predict the 266 out of sample ignitions from 2017-2018, we computed an AUC score of 0.755 and a recall score of 0.781. The small differences in AUC and recall scores between training and testing data suggest that the model is not strongly overfit to the environmental conditions in the training data.

The recall and AUC scores from the training data tell complementary stories. Setting a 95% threshold on the omission rates was an attempt to, in a binary sense, identify areas where vegetation contact ignitions are possible from areas where they're unlikely. If the model were able to perfectly discriminate between these areas, we would expect recall scores to be 0.95. The computed recall score of 0.799 on the training data indicates that the current model predictions omit 20% of ignitions from the thresholded likely/unlikely ignition locations. Multiply the computed recall scores by the 95% confidence threshold, and the product is 0.76—close to the model AUC score of 0.765. As AUC scores estimate how

well at-risk assets can be discriminated from not at-risk assets, we interpret our model results to suggest we have about 76% confidence in the model identifying where ignitions are and are not likely to occur.

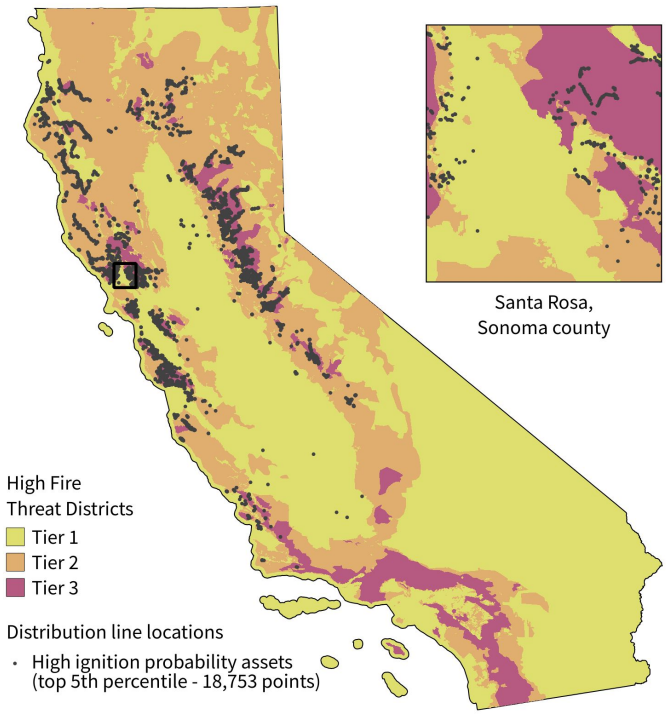


Figure 4. Map of distribution line locations predicted as most likely to lead to a vegetation contact-driven ignition. This subset of assets was delineated by selecting points where the ignition probability was in the top 5th percentile—above 0.17%.

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.

After evaluating which assets were at-risk, we performed mean-centered logistic regression to calculate absolute probability scores to rank ignition probability by asset. To coarsely evaluate the accuracy of these predictions, we computed the sum of all asset probability scores, which we interpret to represent the number of predicted ignitions for each time period, then compared those to the total number of observed ignitions. 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.

Next, we subset the distribution assets into groups according to HFTD tiers and evaluated the ignition probability scores for each asset in each tier (Fig. 5). We calculated the mean ignition probability of Tier 1 assets as 0.03%, the mean ignition probability for Tier 2 assets as 0.08%, and the mean ignition probability for Tier 3 assets as 0.16%. To identify a subset of the assets most likely to lead to ignitions, we computed the 95th percentile of ignition probabilities probabilities for assets classified as at-risk. The 95th percentile of ignition probabilities was 0.176%, and 18,753 assets are above this threshold (Fig. 4). The low per-asset probability scores are a function of the relative rarity of ignitions (~100 per year) compared to the number of distribution assets (~1 million).

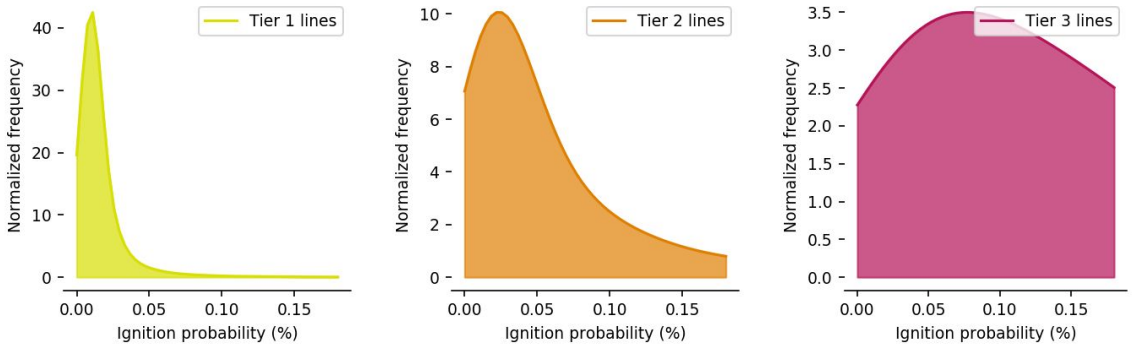


Figure 5. Normalized frequency distributions of predicted ignition probabilities across all distribution point locations for Tier 1, Tier 2 and Tier 3 HFTDs (left, center and right). Our model predicts that the majority of distribution grid assets have a low probability of failure, but that the distribution of ignition probabilities is proportionately higher in areas with higher fire threats.

Next steps

We predicted ignition probabilities from vegetation contact across 934,202 locations along distribution lines using 210 ignition records and statewide maps of vegetation structure, wind and gust speeds, temperature patterns and topography. Using an omission rate threshold, we classified assets into at-risk and not-at-risk lines, predicting 40% of assets were at-risk. Based on the model performance metrics, we interpret our results to suggest we have about 76% confidence in the model discriminating between these two classes. The prediction probabilities appeared well-calibrated for the 2015-2016 training dataset (229 predicted, 210 observed) but less-so for the 2017-2018 testing dataset (200 predicted, 266 observed). When aggregating these results by HFTD, we found ignition probabilities greatly increased from Tier 1 areas (0.03% mean) to Tier 2 areas (0.8%) to Tier 3 areas (0.16%). By arbitrarily setting a threshold at the top 95th percentile of ignition probabilities, we identified 18,753 assets as high priority locations for vegetation management. Based on these results, we have identified the following opportunities for the next steps in predicting ignition probabilities.

- Include temporally-explicit data or evaluate time-sensitive modeling approaches. Based on the discrepancy between observed and predicted ignitions in the 2017-2018 data (and on the results of a sensitivity analysis[8]), we found annual weather pattern data alone were not strong predictors of ignition probabilities. We expect this is because ignition events tend to occur during anomalous weather events, but we only evaluated the effects of annual weather patterns. Establishing metrics that capture anomalous weather conditions, or setting up temporally-explicit models to identify conditions at the time of failure, may help overcome this effect.
- Include data on pole and conductor health. Asset health may play a role in predicting ignition probabilities, and with access to more data on the endogenous condition of each asset we expect to improve our predictions of ignition events.
- Include additional model evaluation metrics. AUC and recall scores are effective for evaluating whether assets can be distinguished as at-risk or not-at-risk. But our probability estimates are difficult to validate. In our current modeling approach we spread the low probability scores across a large number of assets and evaluated whether the sum of those probabilities approximated the

[8] See *Appendix 1* Table 2

observed ignitions. But this is a very non-specific approach; it's not clear that the locations where ignitions occurred were high probability assets. We also don't yet have a sense for whether the predicted probabilities overestimate probabilities for assets that did not cause ignitions, or the extent to which we are dealing with a potentially zero-inflated problem (i.e., a large number of assets that experience vegetation contact and wire-down but did not lead to an ignition).

- Re-evaluate the $\boldsymbol{\tau}$ parameter, which scales relative to absolute ignition probabilities. We tested multiple $\boldsymbol{\tau}$ settings to scale the absolute probability scores, and the sum of all probability scores across all assets should approximately yield the total number of ignitions[9]. $\boldsymbol{\tau}$ was scaled using the total number of distribution grid assets, including areas that were not considered at-risk. However, the sum of ignition probabilities from only at-risk assets was much lower—184.87 predicted ignitions for 2015-2016. It's not yet clear whether $\boldsymbol{\tau}$ should be re-scaled to adjust the probability scores so that only the sum of probabilities from at-risk assets should be considered, or if summing across all assets to include probabilities from all assets captures the uncertainty of characterizing at-risk/not at-risk assets.
- Address the large discrepancy between the predicted and observed ignitions in the 2017-2018 test dataset. This is the result of differences between the input environmental covariates from different years. In this case, only weather data were updated: vegetation and topography were constant between years. Since the weather covariates captured aggregate environmental conditions (mean and maximum conditions during each time periods), this result suggests that the mean and max. weather conditions in 2017-2018 were likely to lead to fewer ignitions than the conditions in 2015-2016—which was not validated by the observed ignitions.

The next iteration of our predictive model will build on this Milestone 1 modeling effort to establish a comprehensive modeling framework that accounts for prevailing environmental conditions, time of failure conditions, and hierarchically considers the causal chain that leads to ignitions. In this analysis we directly modeled the probability that an ignition would occur. In the next iteration, with access to more data, we intend to represent multiple grid failure processes—from contact to outages to lines down to ignitions. Between these linkages are interventions that maintenance, repair, and mitigation efforts can weaken and break, thus reducing the probability of a wildfire ignition.

[9] i.e. this evaluation criterion is a bit circular.

Appendix 1: Modeling details

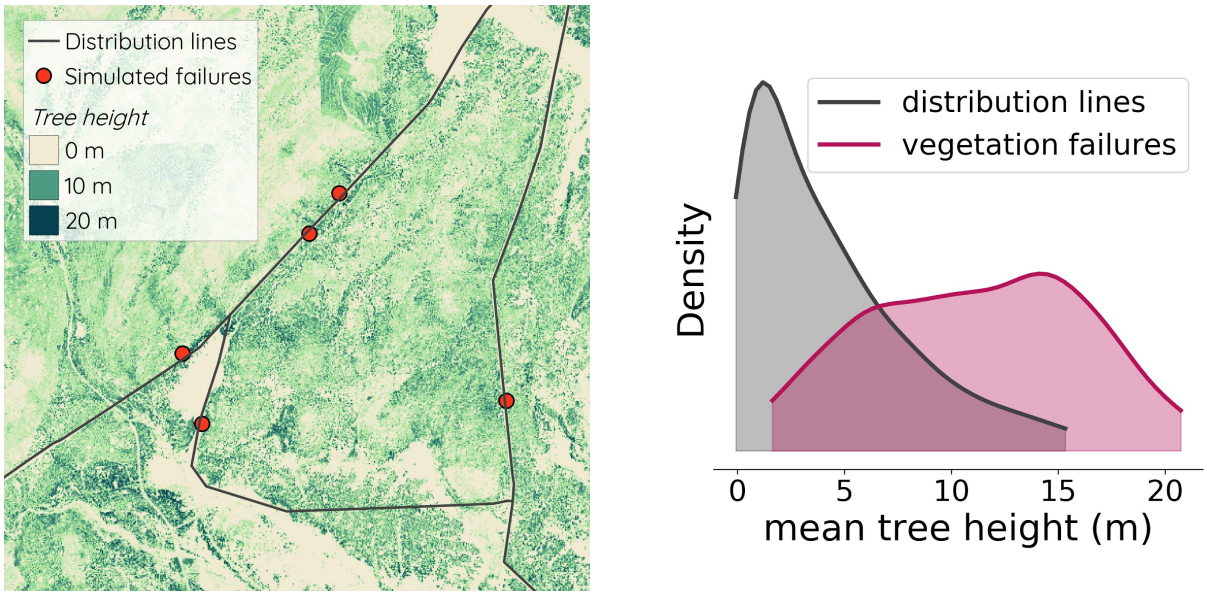


Figure 6. (left) Map showing simulated ignition points (red dots) across a set of distribution lines (black lines). (right) Simulated statistical distributions of tree height measurements from the point locations of ignitions (red) and all distribution assets (black).

The MaxEnt approach to computing conditional probabilities

We used the MaxEnt software package[10] to compute the conditional probability of occurrence of a failure event along distribution lines. Originally developed to predict the geographic distributions of plant and animal species, MaxEnt assesses the relative probabilities that discrete events occur across a landscape. In this case, we used point locations where vegetation contact ignitions occurred to predict the probability of contact failures 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 encroached by vegetation. This range of variation can be described according to a statistical distribution: the majority of lines contain no vegetation encroachment, a small portion with some small trees overhanging lines, and a much smaller proportion with very tall trees overhanging. However, if you were to examine the subset of lines where vegetation contact failures occurred, you would likely find that the statistical distributions shifted, finding a higher proportion of tall trees along assets that failed. Comparing these distributions, you could infer that assets fail more often due to vegetation contact when lines are surrounded by tall trees. You would calculate high probabilities when surrounded by tall trees, moderate probabilities when surrounded by small trees, and zero probability when there are no trees.

This spatially-explicit contact failure modeling approach relies on three key datasets: i) the geographic locations of all operating assets exposed to contact, ii) records of where and when certain assets failed, and iii) a set of environmental predictor data that we expect to determine the probability of failure. In our model formulation, these can be re-framed as i) the landscape of analysis (L), ii) locations

[10] Elith *et al*. 2010, *A statistical explanation of MaxEnt for ecologists*, documented here, available here

where failures have been observed ($y = 1$), and iii) 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:

$$Pr(y = 1|z) = f_1(z) \cdot Pr(y = 1) \div f(z) \quad (1)$$

The null hypothesis is that failures are likely to occur in proportion to the distribution of environmental conditions across all assets ($f(z)$)—failures are equally likely to occur everywhere. We update this assumption using the distribution of conditions at sites where failures have occurred ($f_1(z)$)—using the ratio of these two distributions to calculate the relative probability of failure. The only quantity that is lacking from eq. 1 is the second term, $Pr(y = 1)$, the prevalence of failures (the proportion of failed assets) across the landscape. With information on where failures did not occur, we can compute a prevalence scaler to move from relative probability to probability of failure.

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 failure in one place or another. Because prevalence data are not typically available for calculating the conditional probability of failure, a work-around was implemented (i.e., 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. Knowing τ would solve the non-identifiability of prevalence; without that, MaxEnt arbitrarily sets $\tau = 0.5$. This log transformation is monotone (order preserving) with the raw output. Here, we calculated the prevalence score by computing the average rate of failure as $count(ignitions) \cdot count(assets)$, or $210 \div 934{,}202$, as 0.022%. Unfortunately, due to some technical issues, we rescaled this number by 7 based on the proportion of area PG&E's distribution grid covers relative to the state of California. As a result, the final τ value was calculated as 0.16%.

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 smooth, nonlinear functions to map probability calculations in a continuous fashion.

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[11]
	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[12]
	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)[13]
	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

Table 1. The input environmental covariates to predict ignition probabilities, the units of measurement, the scale of observations, and the data source.

[11] RTMA description available at this link
[12] MODIS LST description available at this link
[13] TPI description available at this link

All input covariates, excluding the vegetation data, were computed in Google Earth Engine[14]. For vegetation contact failures, we expected a higher proportion of nearby trees to be more likely to lead to contact, and taller trees near assets are more likely to lead to contact. However, it was not clear whether vegetation contact failures would be driven by average nearby vegetation patterns or by tall "outlier" trees.

We expected higher wind speeds and higher temperatures to increase the probability of failure. But it was not clear *a priori* whether sustained (i.e., mean) wind/heat pressure is more important than extreme events (i.e. local extremes / high frequency heat or wind events). We expected these patterns to be moderated by the topographic position of an asset (e.g. whether an asset is on a ridgetop or in a valley).

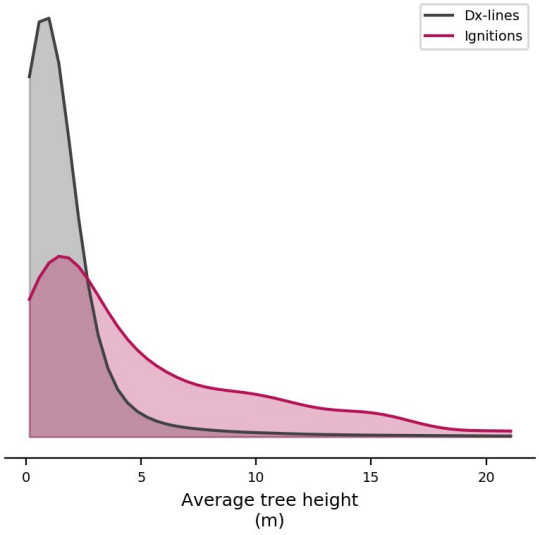
Feature importance scores

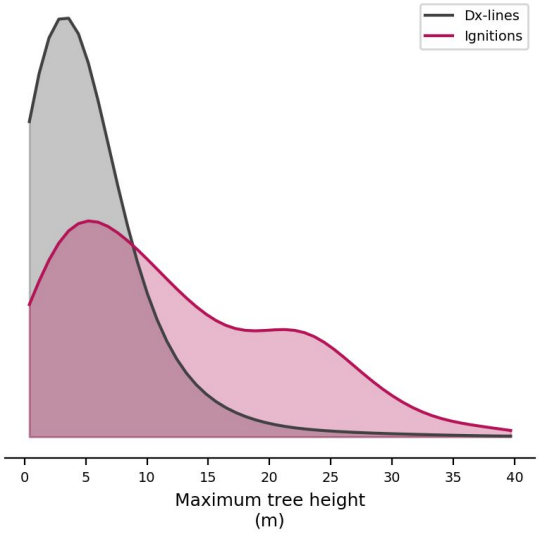
Variable	Percent contribution	Permutation importance
Max. tree height	60.6	63.5
Avg. tree height	23.2	0
Local topographic position	3.6	0
Average gust speed	3.3	16
Landscape topographic position	3.3	0
Max. temperature	1.8	2.4
Max. gust speed	1.6	0
Avg. temperature	1.4	9.6
Avg. wind speed	0.9	3.6
Max wind speed	0.3	4.9

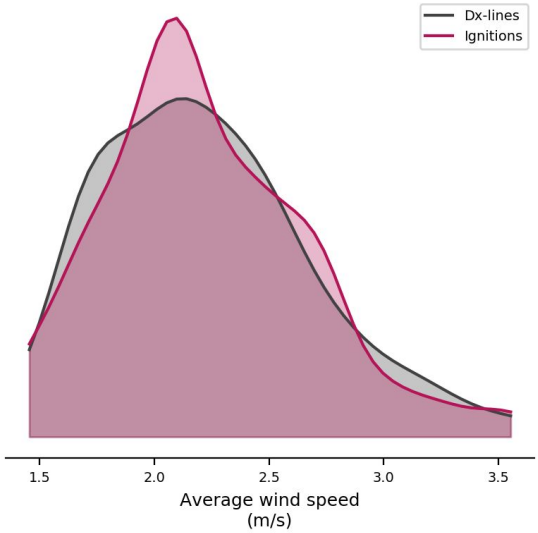
Table 2. Results from a jackknife sensitivity analysis, which quantified the relative importance of each covariate by computing changes in model performance when each covariate was excluded from analysis. Percent contribution calculates how model 'gain' changes in response to random permutations in feature coefficients (which are fit from the input covariates). Permutation importance scores are calculated by randomly altering the values of a single covariate and recomputing model performance. These models are rescaled to percentages based on how model performance changed across all covariate permutations.

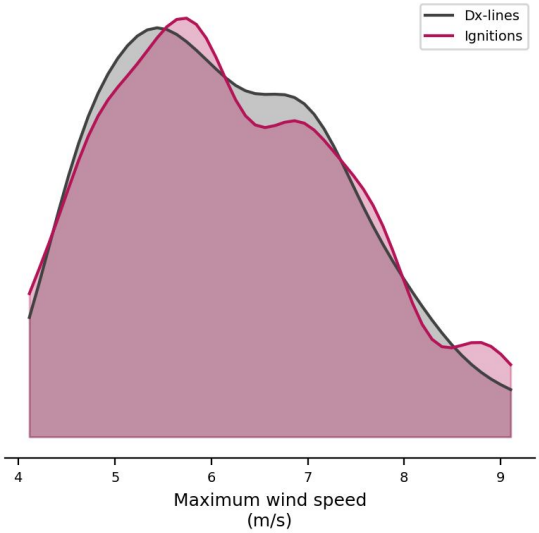
[14] https://code.earthengine.google.com/

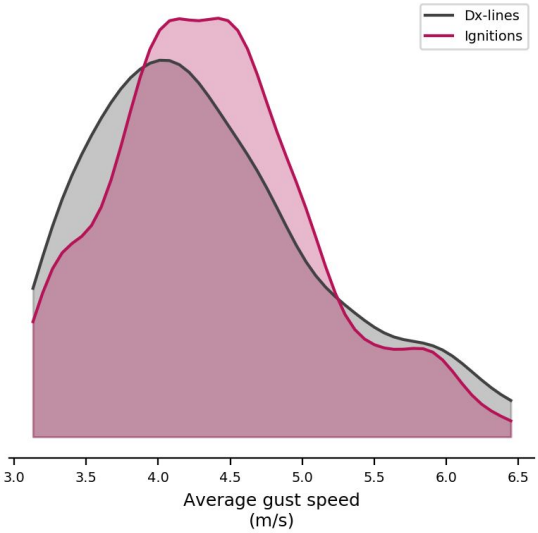
Appendix 2: Environmental covariate distributions

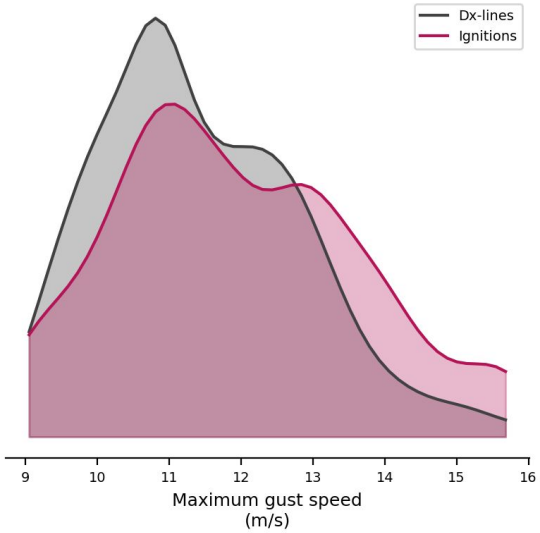


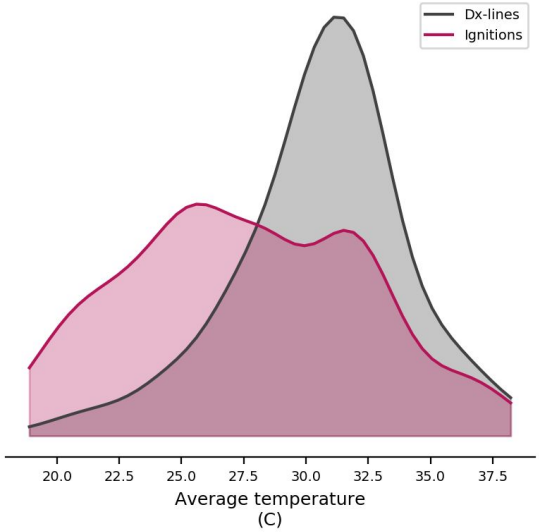


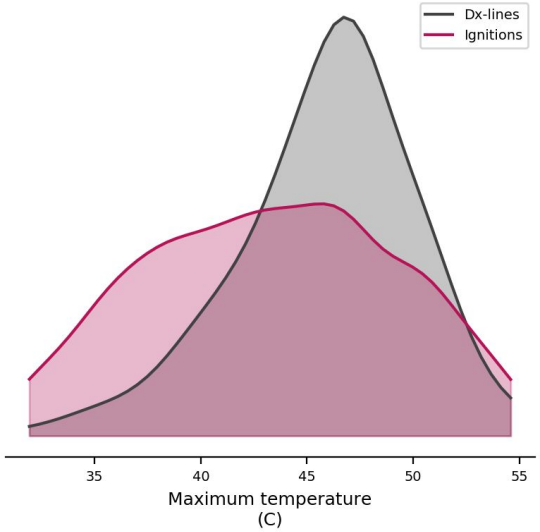


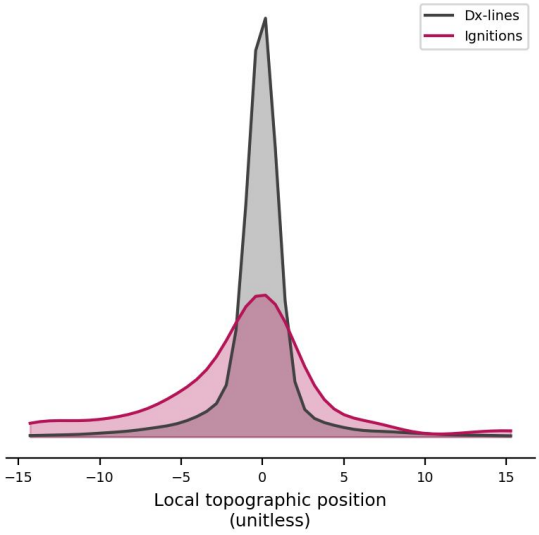


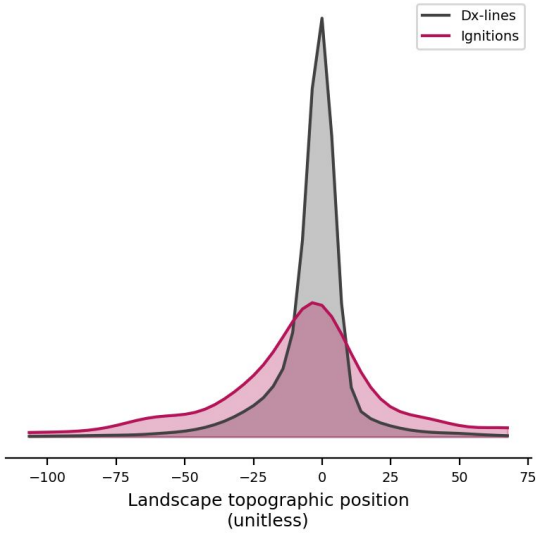

Dx-lines
Ignitions
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Average gust speed
(m/s)


Dx-lines
Ignitions
9
10
11
12
13
14
15
16
Maximum gust speed
(m/s)

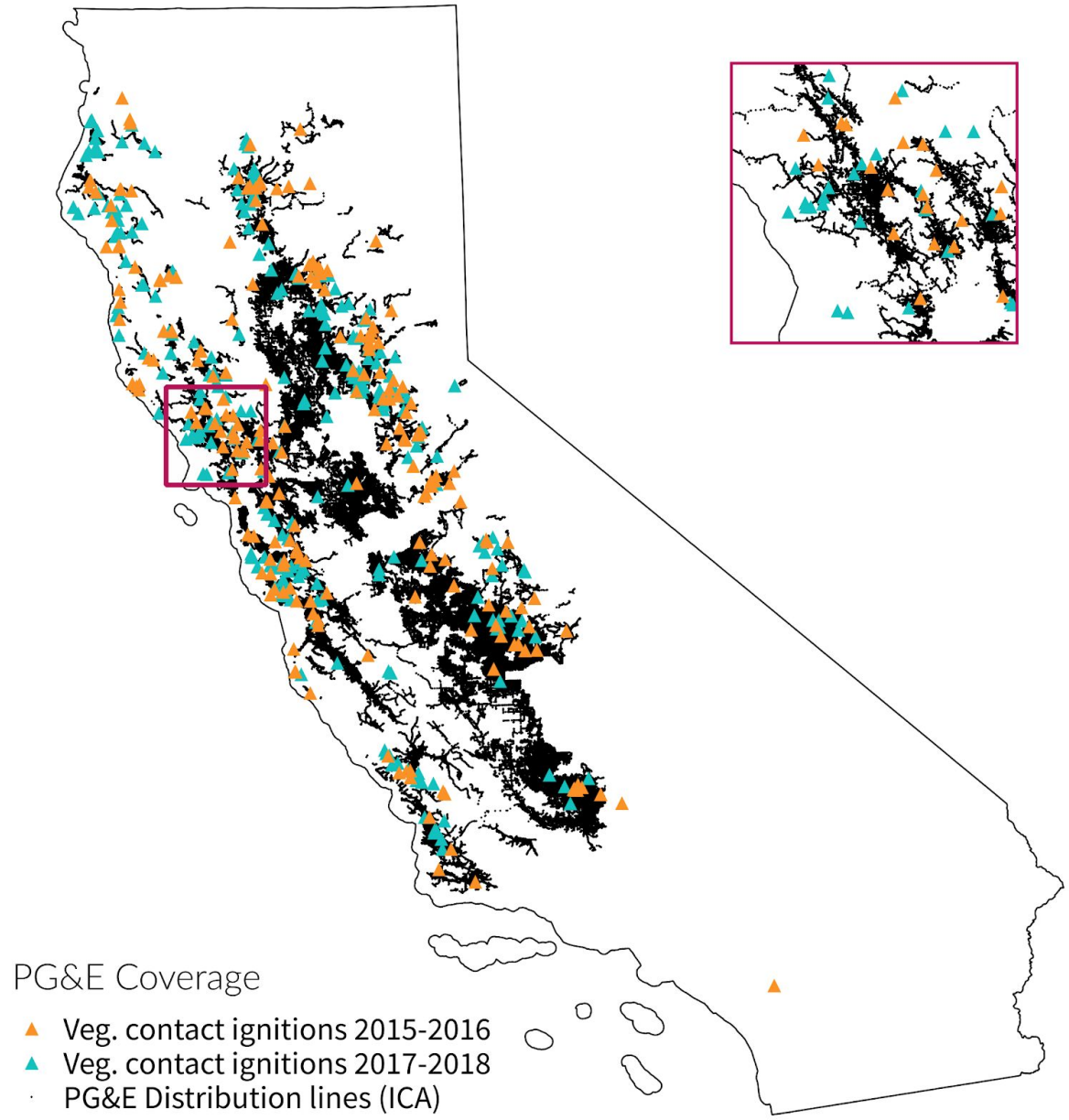

Dx-lines
Ignitions
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Average temperature
(C)

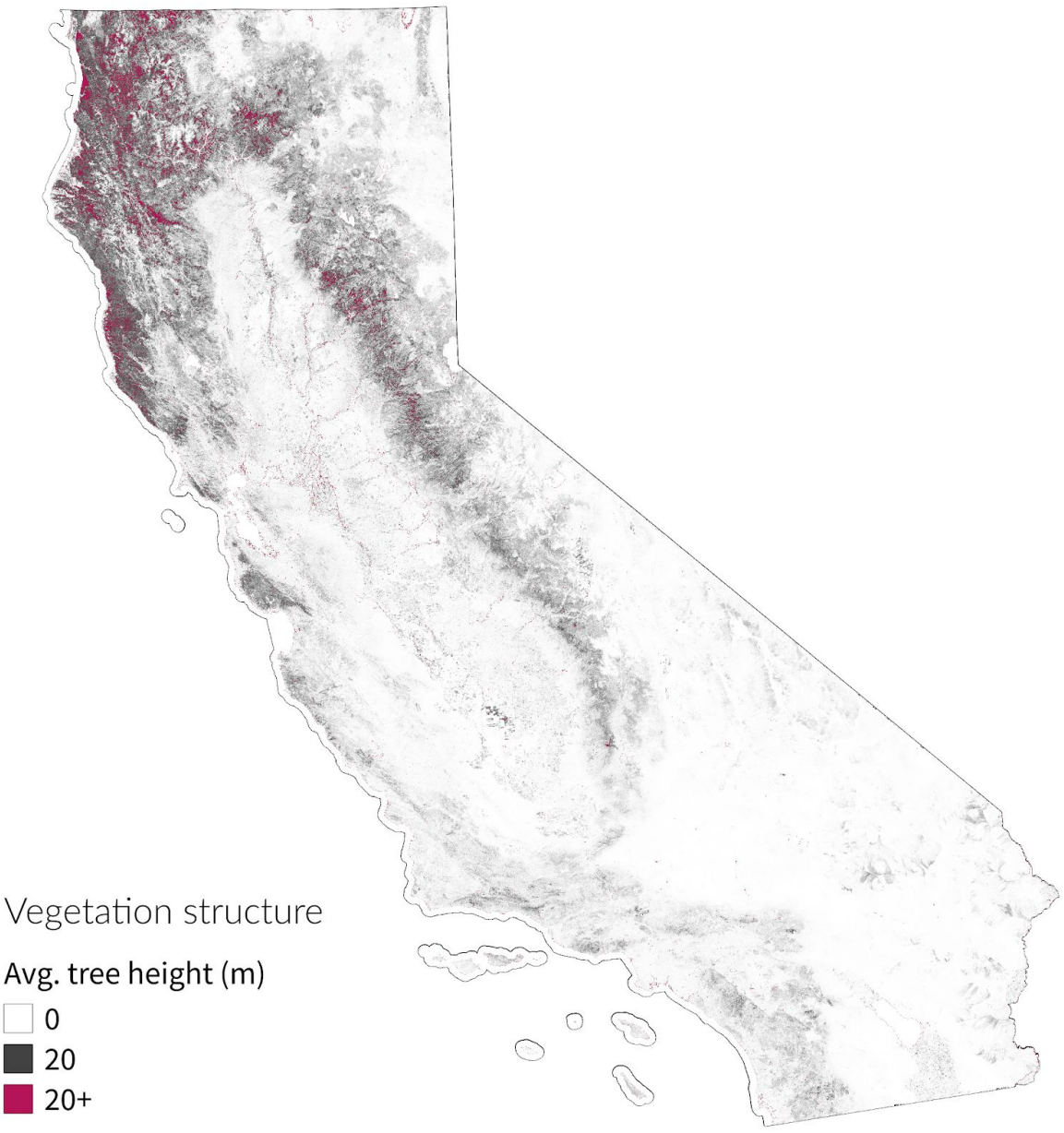

Dx-lines
Ignitions
35
40
45
50
55
Maximum temperature
(C)

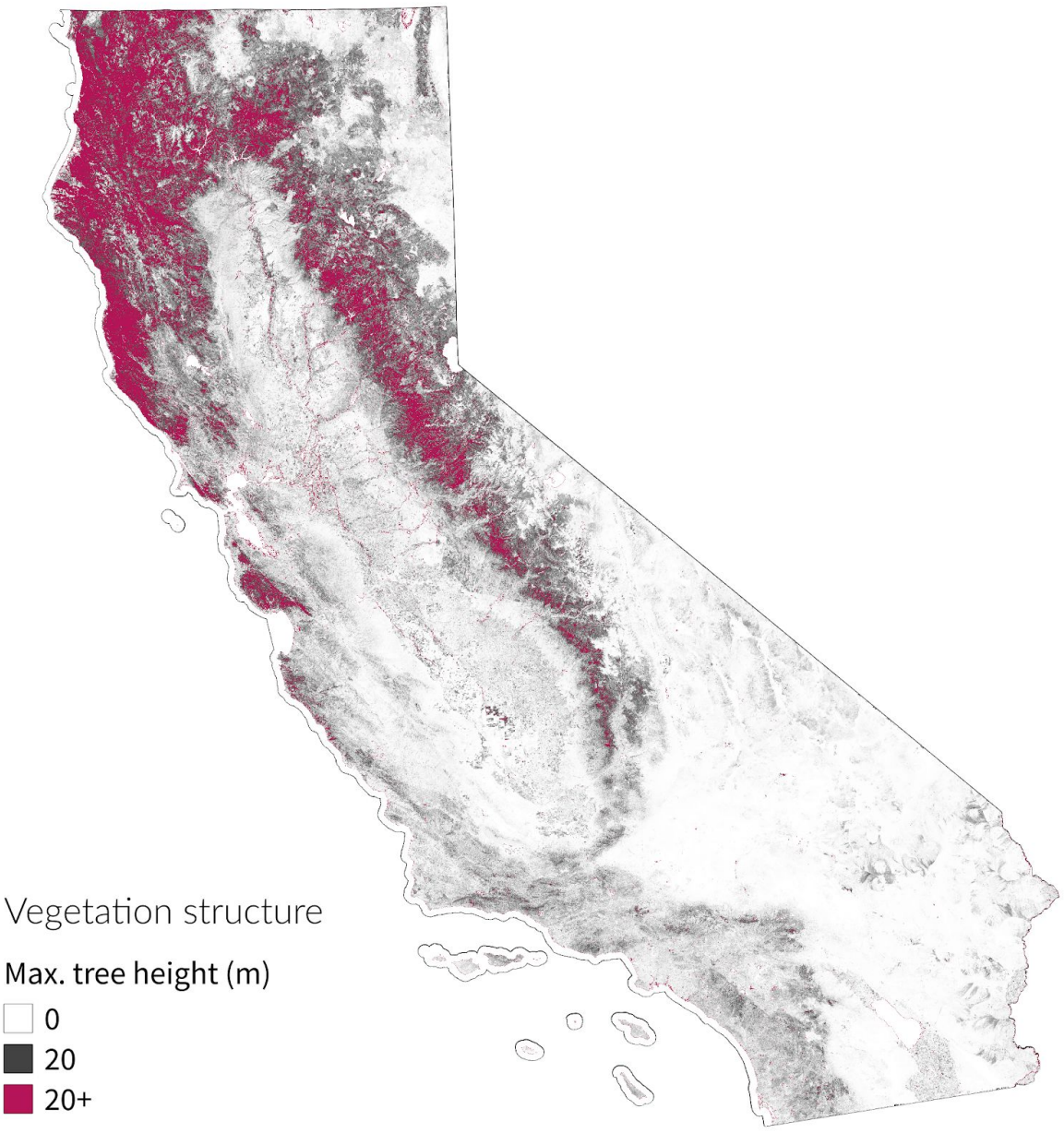

Dx-lines
Ignitions
−15
−10
−5
0
5
10
15
Local topographic position
(unitless)

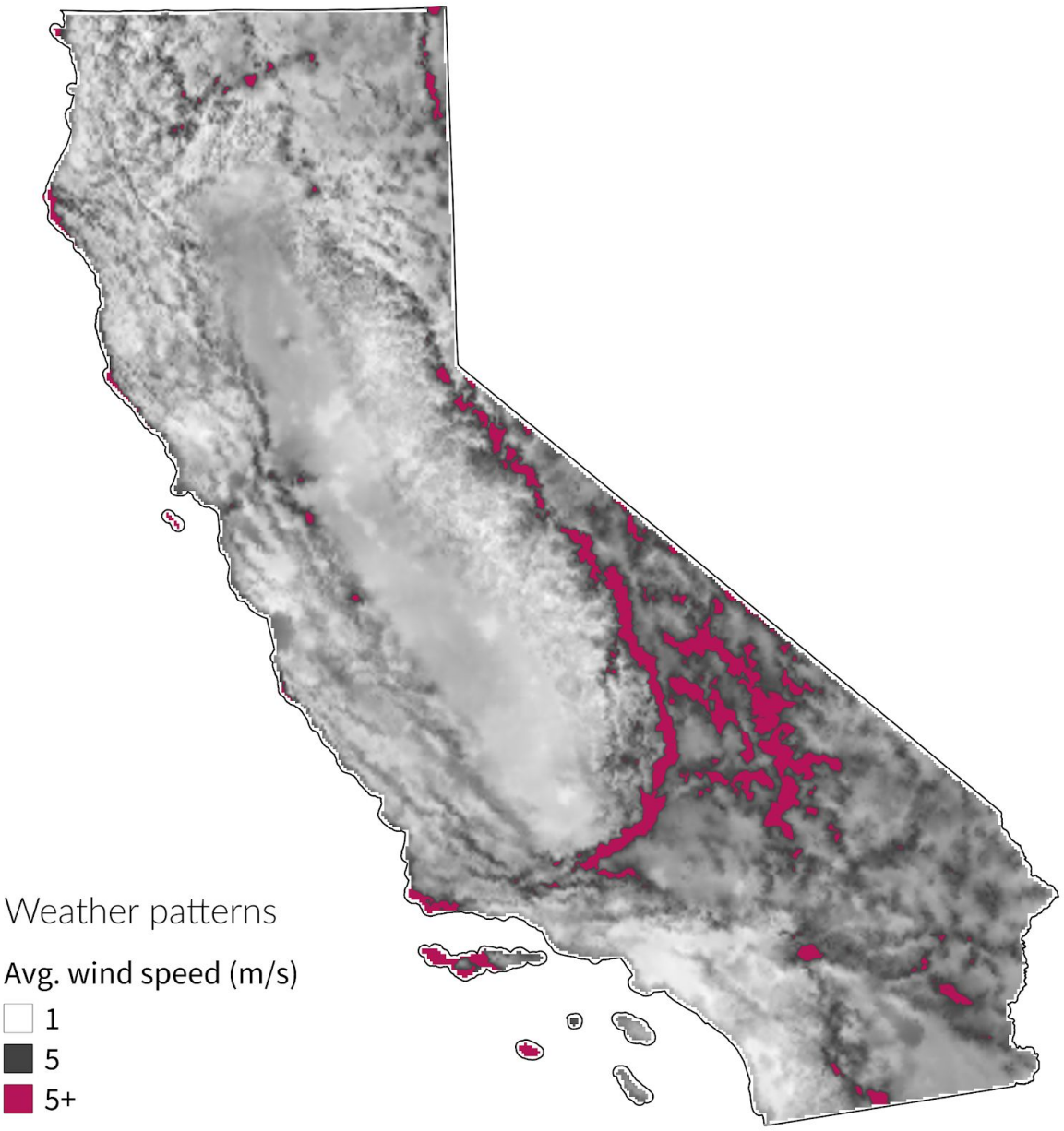

Dx-lines
Ignitions
−100
−75
−50
−25
0
25
50
75
Landscape topographic position
(unitless)

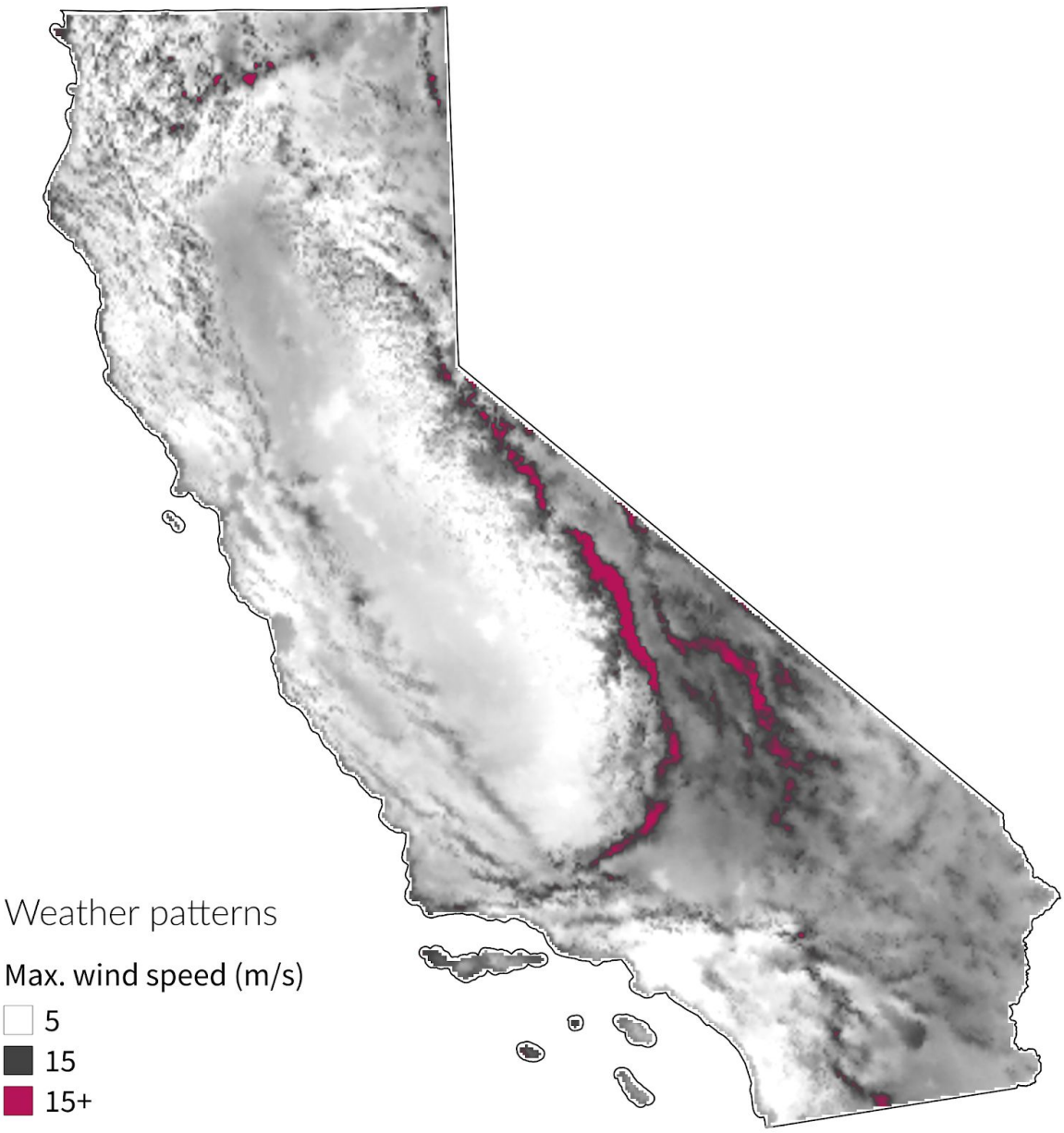
Appendix 3: Maps

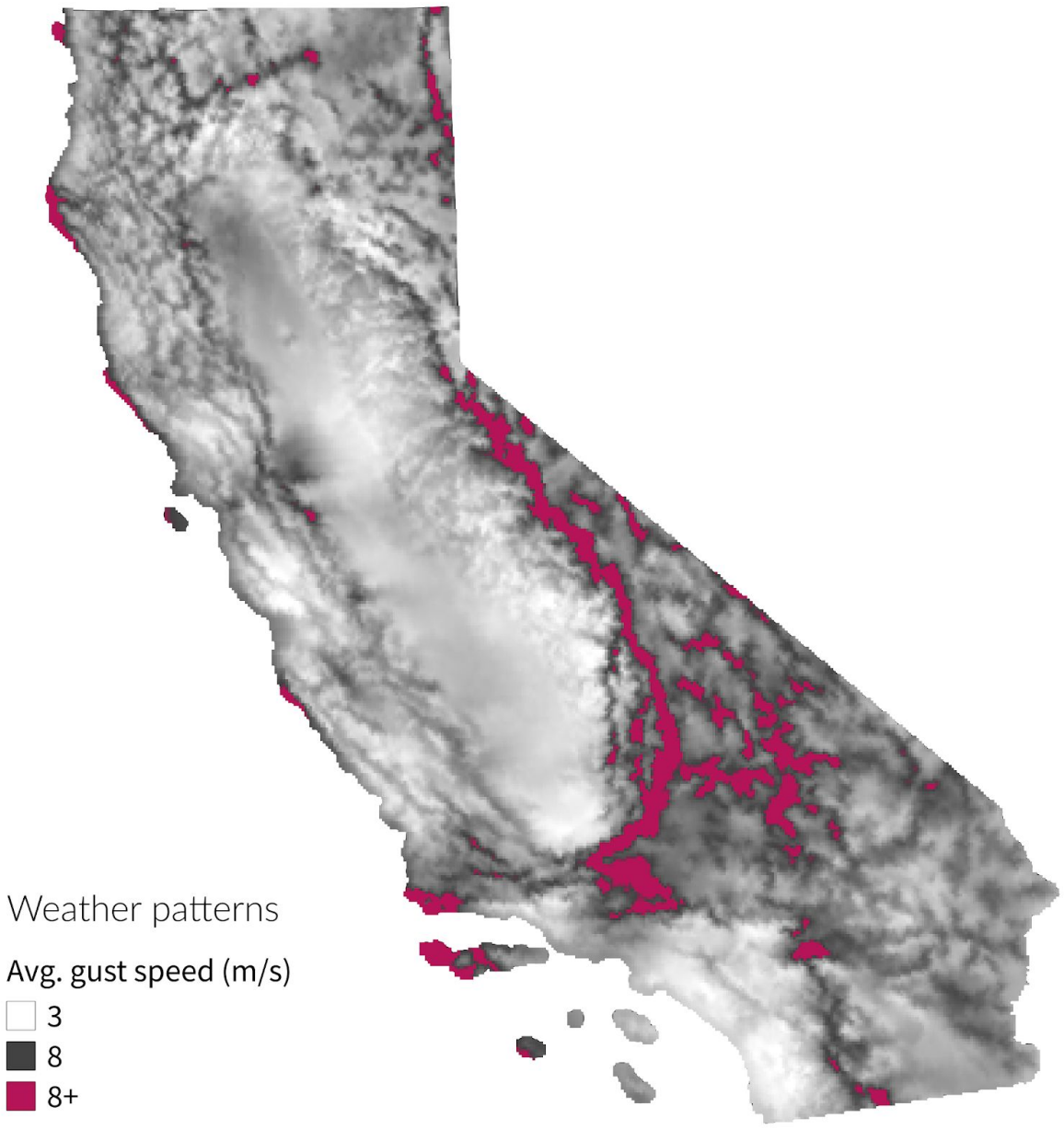


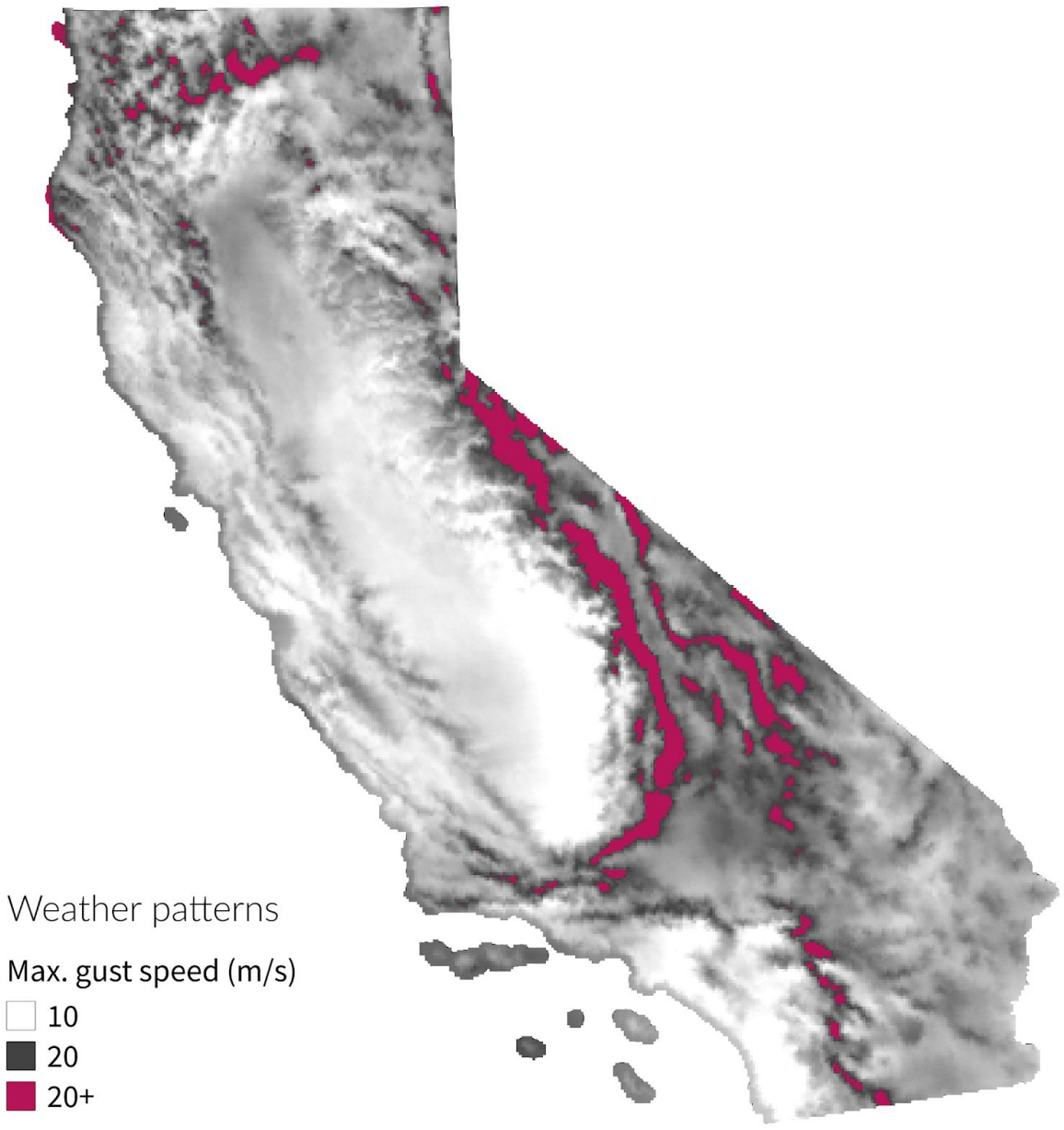

Vegetation structure
Avg. tree height (m)
0
20
20+

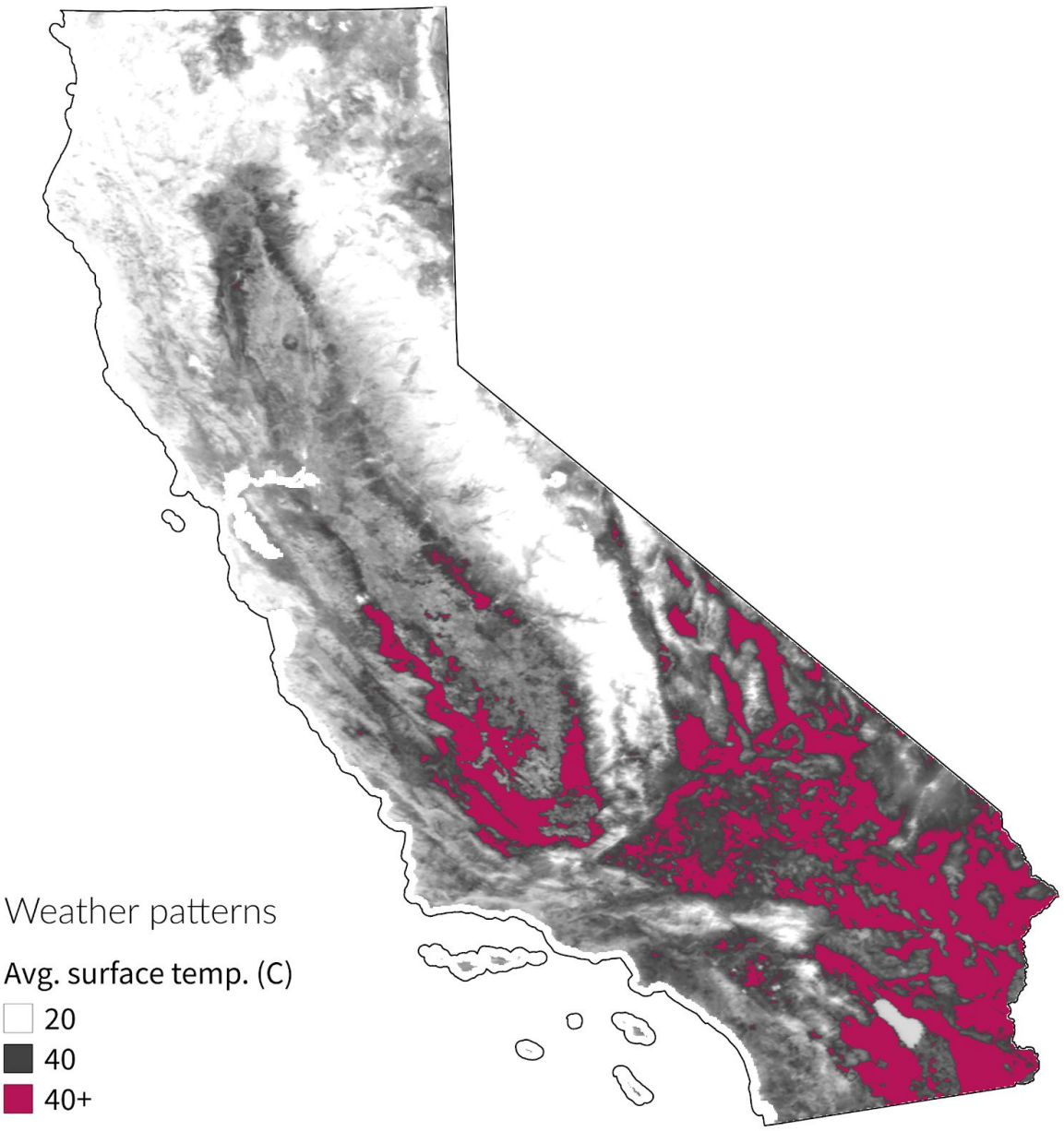

Vegetation structure
Max. tree height (m)
0
20
20+

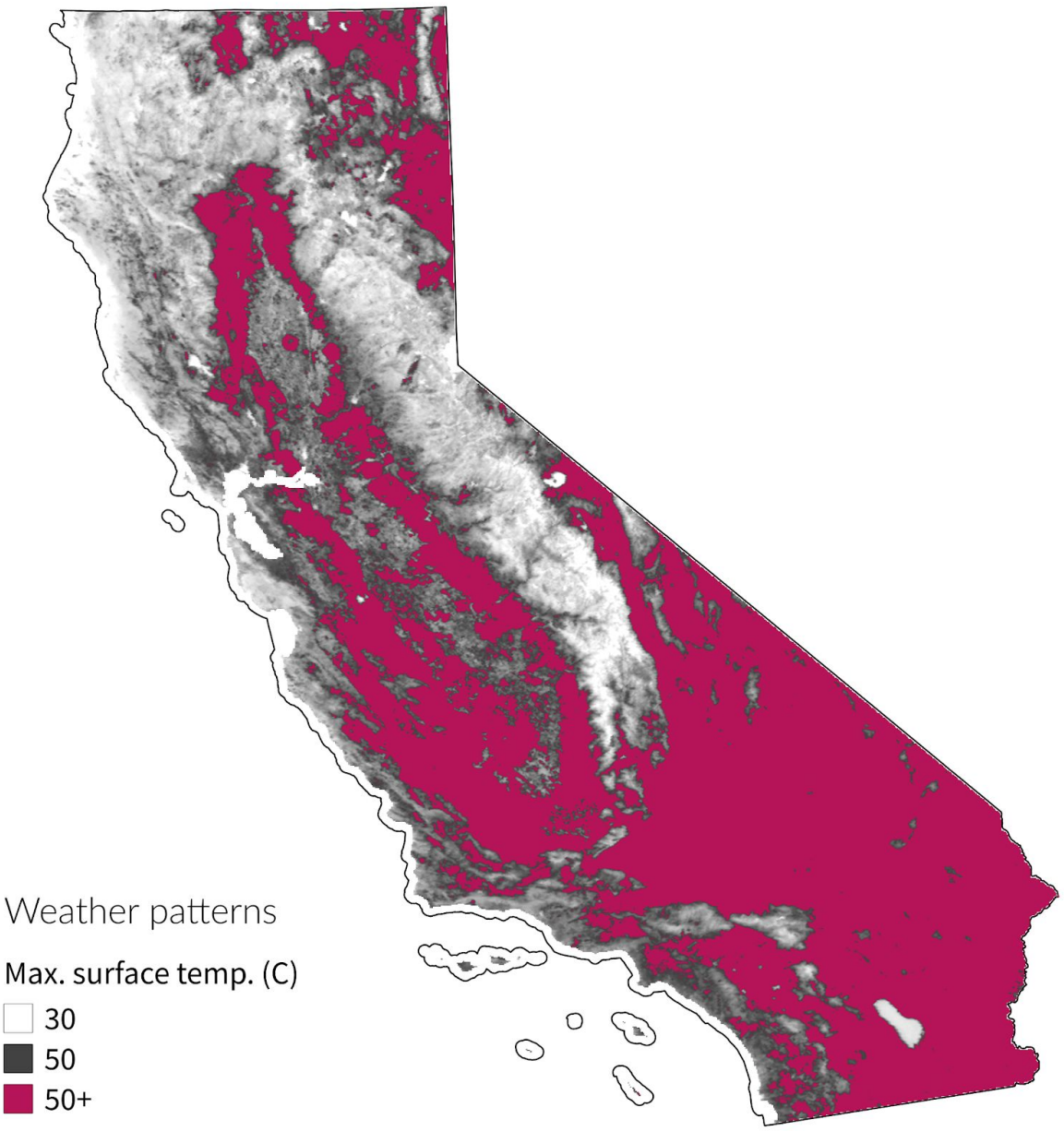

Weather patterns
Avg. wind speed (m/s)
1
5
5+

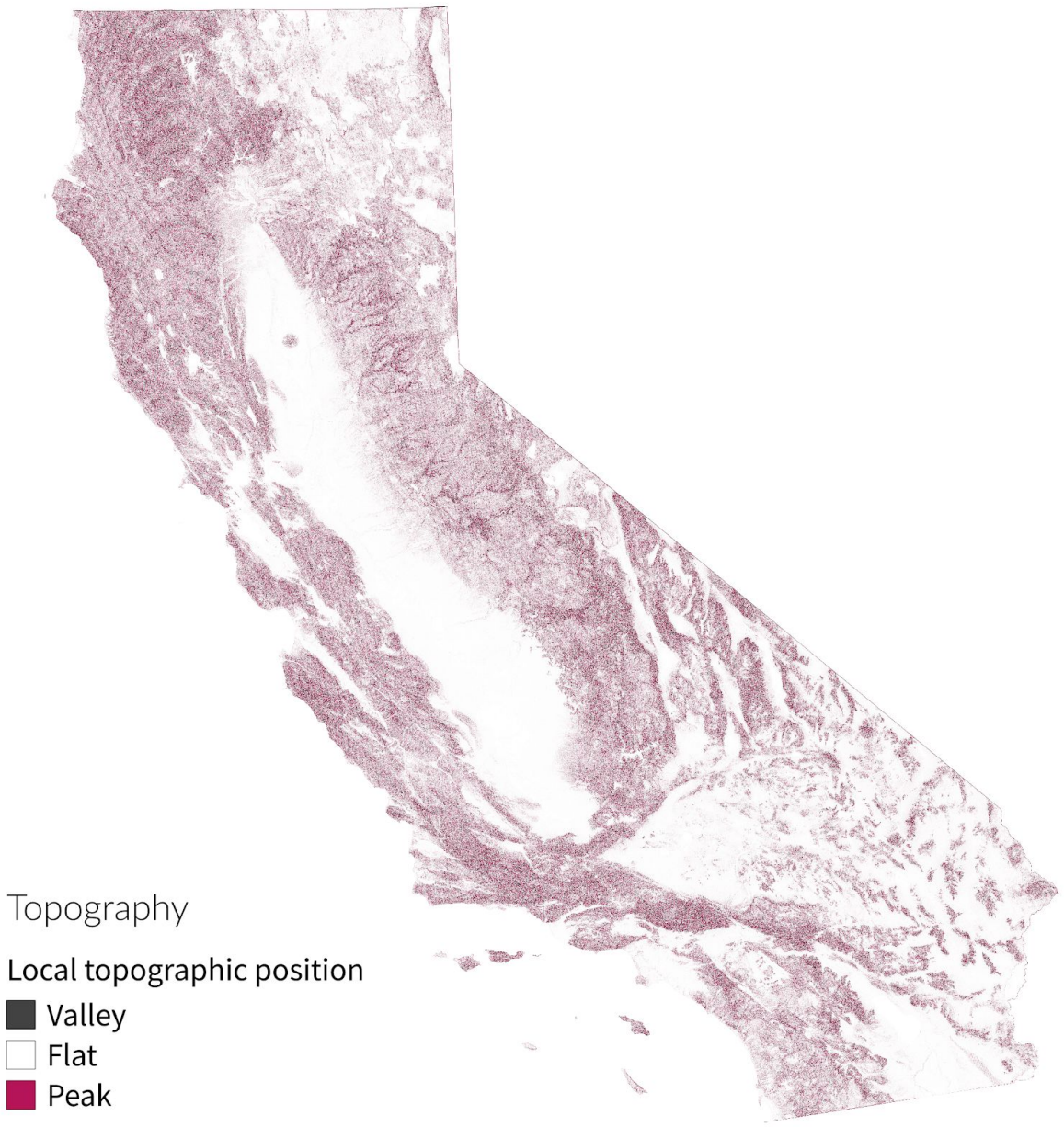

Weather patterns
Max. wind speed (m/s)
5
15
15+

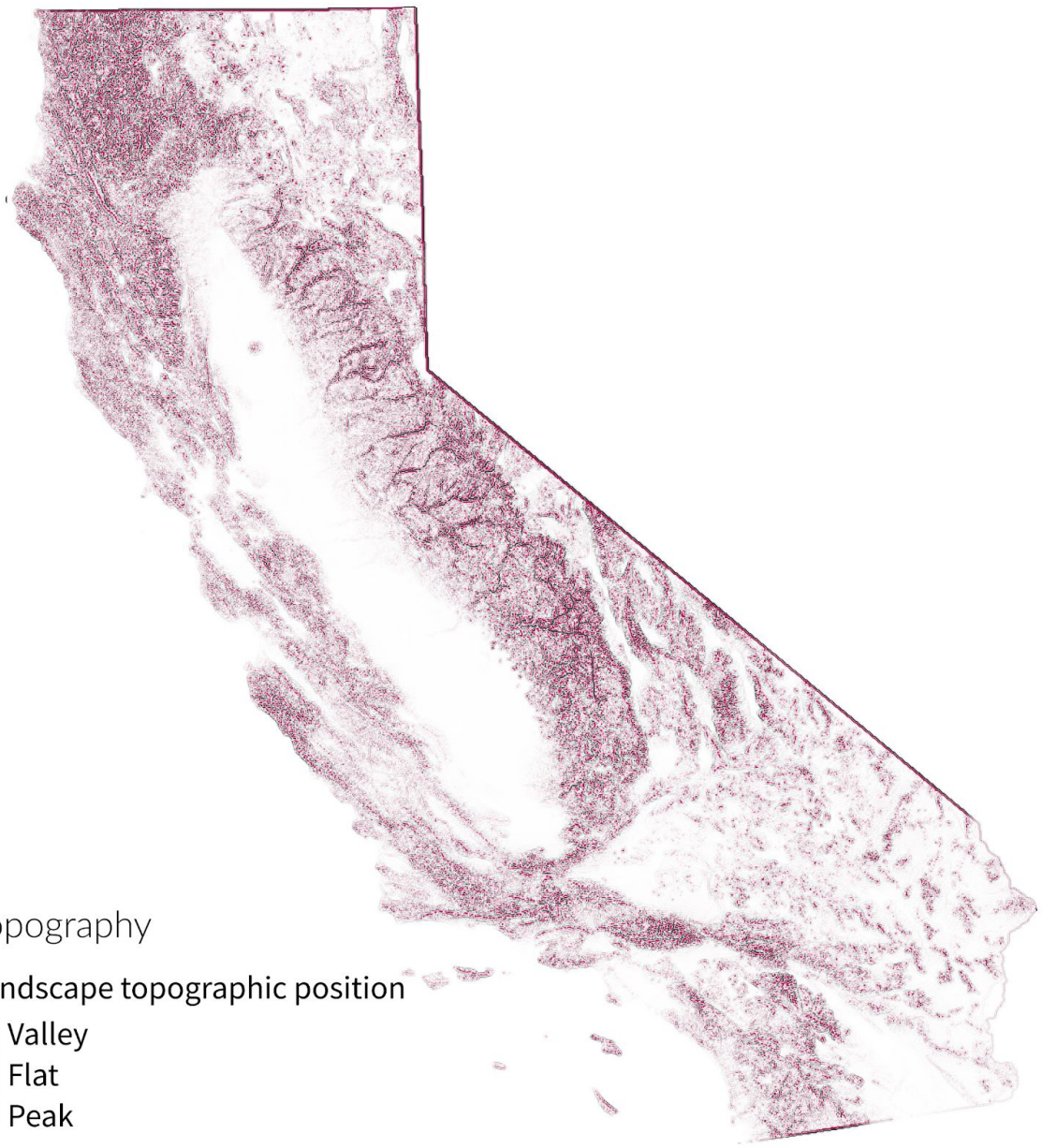

Weather patterns
Avg. gust speed (m/s)
3
8
8+

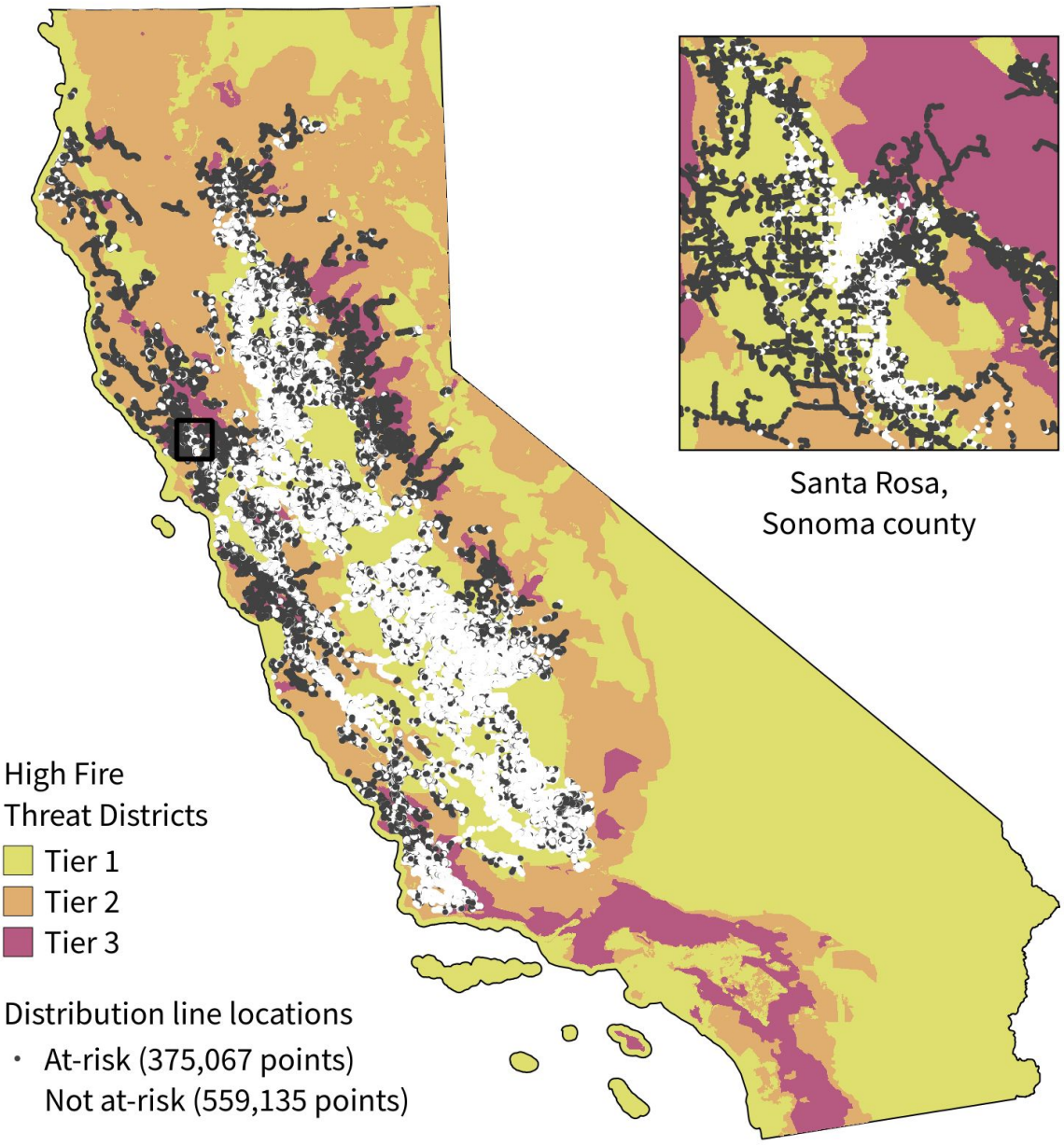

Weather patterns
Max. gust speed (m/s)
10
20
20+

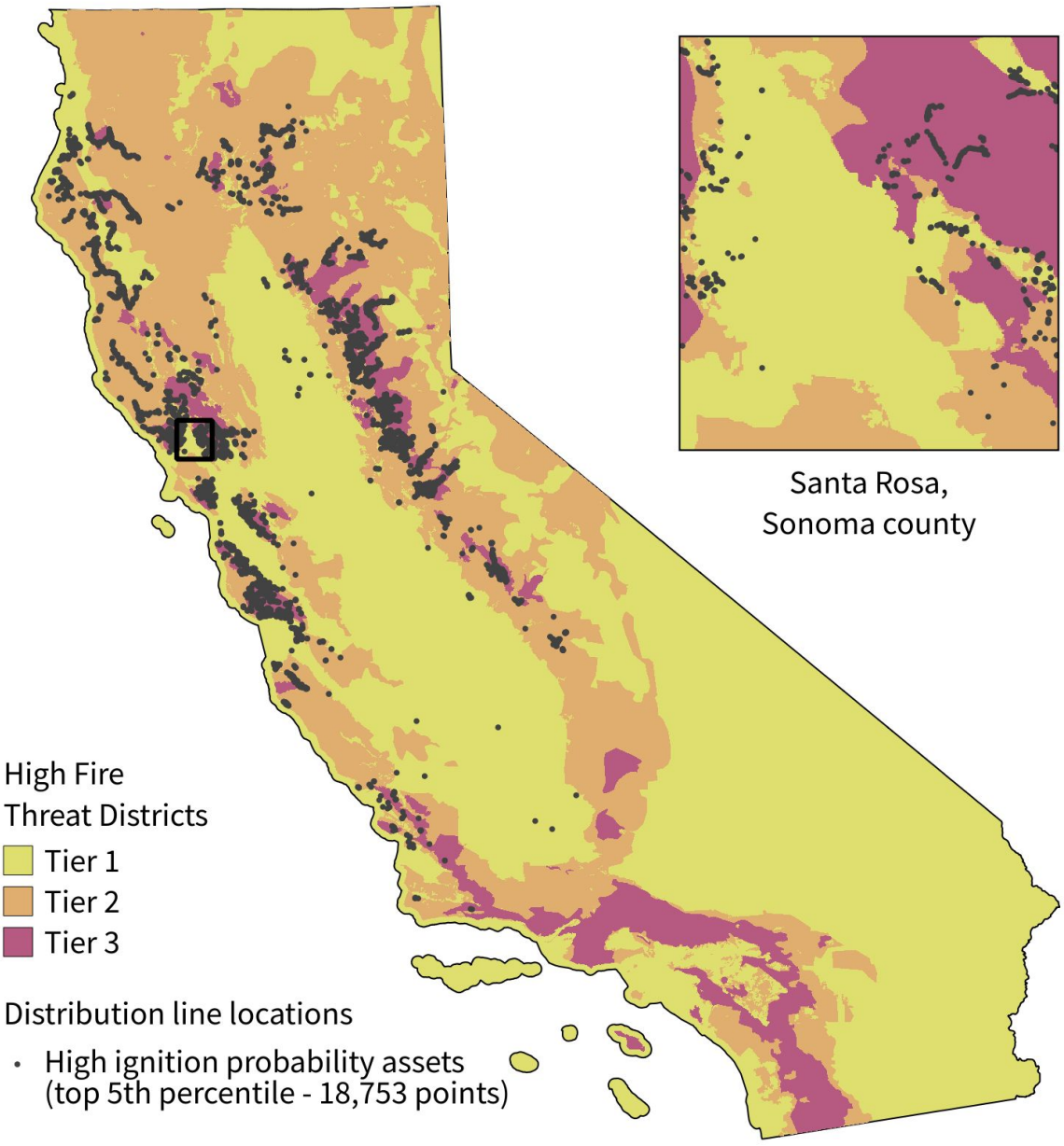

Weather patterns
Avg. surface temp. (C)
20
40
40+


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


Topography
Local topographic position
Valley
Flat
Peak


Topography
Landscape topographic position
Valley
Flat
Peak


Santa Rosa,
Sonoma county
High Fire
Threat Districts
Tier 1
Tier 2
Tier 3
Distribution line locations
At-risk (375,067 points)
Not at-risk (559,135 points)


Santa Rosa,
Sonoma county
High Fire
Threat Districts
Tier 1
Tier 2
Tier 3
Distribution line locations
High ignition probability assets
(top 5th percentile - 18,753 points)