# Predicting the Size of California Wildfires and Average Size By County

Nicholas Archambault

7 December 2020

## Introduction

Widespread destruction borne from the spread of summertime wildfires has become an annual scourge for millions of residents of the American west. This summer, skies over California were blanketed for days at a time with apocalyptic orange haze, the toxic sign of wildfire activity surging across the state. The *New York Times* reports that 2020 is the worst in a series of increasingly more devastating fire seasons. As the result of blazes spread across not only California, but Oregon and Washington as well, over two dozen people have died. Over five million acres have burned across the three states, a sixfold increase from the yearly average between 1950 and 2000.

Western wildfires, however, are more than destructive crises – they also represent a symptom of a dangerous long-term environmental trend. As the planet continues to steadily warm, the chaos of 2020's wildfire season will become the yearly norm rather than an anomaly. According to the *Los Angeles Times*, the ten most destructive fires on record have occurred since 1991, fueled by hotter weather, intense drought, and increasingly exurban settlement to set records that are sure to be broken within the coming years.

Fire season precipitates a chain reaction across all levels of a state. Firefighters put their lives on the line to battle the flames, governmental leaders issue evacuation directives, families confront the challenge of obtaining temporary housing. In the aftermath, ecologists seek to understand the root causes of the devastation and work with policymakers to implement more appropriate safety procedures.

A model that can evaluate and predict the size and potential damage of blazes in response to geography and changing climate conditions could save lives and inform the work of those tasked with mitigating wildfires' dangers. This project surveys a subset of nearly two million U.S. wildfires from between 1992 and 2015 and attempts to predict their sizes based on proximity to the nearest city as well as a number of meteorological conditions.

Data were collected from the U.S. Department of Agriculture and the California Irrigation Management Information System at the California Department of Water Resources.

## Data

The following section will provide a brief overview of the data amalgamated for this project, including their sources, important features, and the processes by which they were cleaned into useful forms.

### Fire Data

The primary data source – and the inspiration for this project – was Karen Short's spatial wildfire occurrence data from the archives of the U.S. Department of Agriculture. The data were accessible as a SQLite object easily parsed by the 'RSQLite' package in order for the main table to be extracted. This table contained 40 variables on over 1.88 million wildfires from across the United States between 1992 and 2015.

I eliminated all variables pertaining to the identification of each fire by local and national database indices, then restricted the dataset to fires that began in California and removed all fires without a defined end date. The most challenging aspect of the cleaning was parsing the start and end dates, which were provided on the absolute Julian scale typical of scientific data. I created a function to convert these dates to traditional Gregorian format, essential for combining the fire data with weather data on the days leading up to each fire.

Finally, in order to preemptively avoid web scraping difficulties, I eliminated all fires that began in one year but ended in a new year. After creating a few new variables to mark the month and duration of each fire, the final dataset was left with 16 variables – including the start and end dates, cause, size, and longitude and latitude of each entry – and 91,880 fires.

## Station Data

The ultimate goal of the data cleaning process was to merge the fire data with meteorological data from the days before and of each fire. Meteorological data, however, is organized by each of the 209 statewide weather stations within the California Irrigation Management infrastructure. The first step in the data cleaning process was to identify which station measured the local weather conditions near the location of each of the 91,880 fires.

To access location and identification data for each station, I used the 'cimir' package to query the California Irrigation Management Information System (CIMIS) API. The initial dataset contained 209 unique station entries and nine variables, including the longitude and latitude of each station as well as its date of connection and disconnection to the data repository.

My initial approach to the challenge of assigning a station to each fire was misguided. I attempted to isolate the longitude and latitude coordinates of each fire individually, then cross-reference these coordinates against the longitude and latitude coordinates of all 209 CIMIS stations. A nearest-neighbors search with $k = 1$ took only seconds, yielding the index of the station that measured meteorological conditions for each fire.

I realized, however, that this approach made two erroneous assumptions. First, the nearest-neighbors algorithm uses Euclidean geometry to evaluate distances. The distance between longitudinal and latitudinal coordinates, however, is based on the curvature of the Earth – distinctly non-Euclidean geometry. Second, I realized that many of the stations I had identified as the closest to a particular fire were actually built and connected to the CIMIS network many years after the fire took place. For example, the station measured as closest to a 1997 fire in San Diego County was Borrego Springs. But Borrego Springs did not come online until 2008; thus, I could not retrieve weather data about that fire since the station that my function alleged to have measured it did not arrive until 11 years later. In reality, the closest station at the time of the fire was Escondido, online from 1989 to 1998.

I revised my function to take these changes into account, using the 'sp' and 'geosphere' packages to evaluate non-Euclidean distance ("as the crow flies") and limiting the pool of viable stations for each fire to those whose period of activity included the start and end dates of the fire. The function worked smoothly, and I assigned a station name and number to each row of the fire dataset.

## Weather Data and API Query

With a measuring station assigned to each fire, I could extract weather data from that particular station on the dates leading up to the date of the fire. Wind, temperature, and moisture conditions are crucial factors that can turn a campfire into full-fledged conflagration, so my queries focused on specific variables influencing those larger meteorological trends. After defining a set of 29 parameters to be queried, I pulled the pertinent weather data for each station from the CIMIS API, beginning on 1 January 1992, the start date of my data, and ending on either 31 December 2015 or the end date of the most recent fire measured at that station, whichever came first.

I faced a number of API-related obstacles that prevented me from querying data for each of the 91,880 fires in my dataset. I optimized my code to pull data in batches of no more than 1,750 records at a time, the query limit for the API, but the function consistently stalled and threw API errors. Despite my strict adherence to the batch limit, I received warnings and errors claiming that my API key was invalid or that my network connectivity was unreliable, mere moments after I had pulled batches of thousands of records without a hitch. I generated multiple different API keys for myself and switched to an Ethernet connection in an attempt to resolve these issues, but it still took me two entire days to pull what turned out to be incomplete weather data.

Merging the fire, station and weather data resulted in a dataset, *'total'*, with 47 variables, 29 of which were meteorological, and 69,100 entries, three quarters the length of the California fires dataset with which I started the querying process.

## Final Data Cleaning

To prepare the data for visualization and modeling, I imputed NA values with variable means and removed obvious mistakes in the data, such as soil temperatures of -6999 $^\circ$. Such data cleaning compelled me to carefully scrutinize the units and definitions of each variable. Relative humidity, for example, is a percentage value that cannot be negative. Evapotranspiration is another metric that technically cannot be negative, but sometimes is due to certain rounding and calculation practices. I consulted meteorological definitions in order to ensure that I was not neglecting to remove compromising or mistaken data points.

Next, I engineered a new variable measuring the remoteness of each fire: its proximity, in kilometers, to the nearest city or town. I was interested in understanding whether increased proximity and immediate threat to human life or habitation would result in greater effort or success in preventing a wildfire's spread and duration. I scraped from Wikipedia a list of all cities and towns in California, as well as their counties. After some minor data cleaning, I once more utilized non-Euclidean analysis of distance to identify the city and its county with the closest proximity to each fire in my dataset. This proximity was listed as the new variable 'remoteness'.

In preparation to construct map plots of variables by county, I imported a by-county population dataset, sourced from the World Population Review, as well as a list of county FIPS Code from the USDA.

Finally, I created lagged columns that computed for select meteorological variables the rolling mean for seven and 30 days prior to each fire's occurrence. I reasoned that the backdrop of elevated fire risk conditions due to certain cumulative weather metrics, such as a period of hot, dry weather or an unusually rainy month, may do more to influence a fire's spread than the weather on the very day of the blaze.

The final, complete dataset, *'total.all.csv'*, contained 66 variables and 54,418 entries. Documentation and units for key meteorological variables are listed below:

- size [acres] - size of wildfire
- air.tmp.avg [C] - daily average air temperature
- air.tmp.max [C] - daily maximum air temperature
- air.tmp.min [C] - daily minimum air temperature
- dew.pnt [C] - temperature to which air must be cooled in order completely saturate it with water vapor
- eto [mm] - reference crop evapotranspiration: an estimate of the water used by a 'reference crop', a well-watered, full-cover grass surface, 8-15 cm in height
- prcp [mm] - daily precipitation, all forms
- rel.hum.avg [%] - daily average percentage of water vapor present in air relative to the amount needed for saturation at the same temperature
- rel.hum.max [%] - daily maximum relative humidity
- rel.hum.min [%] - daily minimum relative humidity
- soil.tmp.avg [C] - daily average soil temperature
- soil.tmp.max [C] - daily maximum soil temperature
- soil.tmp.min [C] - daily minimum soil temperature
- sol.rad.avg [W/$m^2$] - daily solar radiation – sunlight – absorbed from space
- sol.rad.net [W/$m^2$] - sunlight absorbed minus sunlight reflected
- vp.avg [kPa] - average daily vapor pressure of water: pressure at which gaseous and liquid states are in equilibrium
- vp.max [kPa] - maximum daily vapor pressure of water

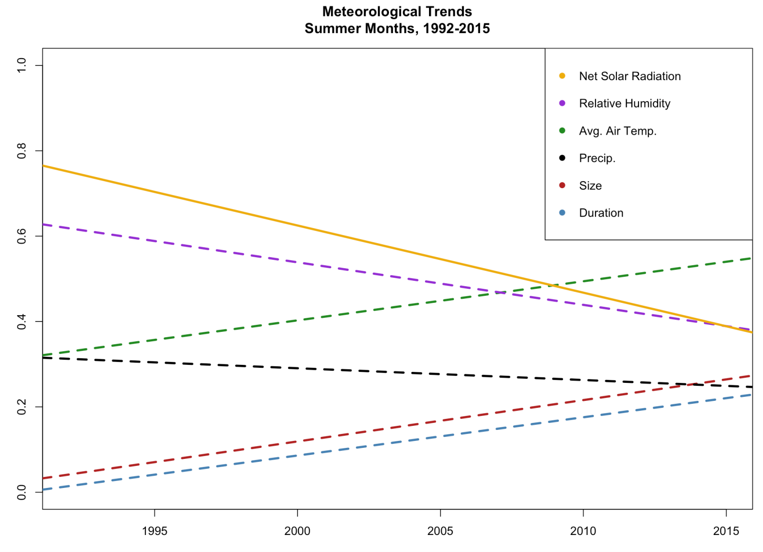
- vp.min [kPa] - minimum daily vapor pressure of water
- wind.ene [m/s] - ENE wind speed
- wind.ese [m/s] - ESE wind speed
- wind.nne [m/s] - NNE wind speed
- wind.nnw [m/s] - NNW wind speed
- wind.ssw [m/s] - SSW wind speed
- wind.sse [m/s] - SSE wind speed
- wind.wnw [m/s] - WNW wind speed
- wind.wsw [m/s] - WSW wind speed
- wind.spd.avg [m/s] - average wind speed
- wind.run [m] - cumulative 'distance' of wind blowing past a given point

Variables including average air temperature, dew point, evapotranspiration, precipitation, average relative humidity, and wind run are also present in columns lagged at seven- and 30-day rolling averages.

## Analysis

In an effort to explore the data visually and uncover latent trends between variables, I identified numerical columns which could be averaged appropriately and aggregated the data by county, year, and month, computing the mean across each variable. The yearly data encompassed 24 years and 51 variables, while the monthly data encompassed all 12 months and 51 variables. From the FIPS dataset, I constructed the full-length FIPS code for each county and appended it to the county data. This code would be essential for plotting maps of variables by county. I also included the population of each county and the number of average acres burned per county, per 1,000 residents. The final county dataset had 55 rows and 55 variables. There are 58 counties in California, and their information was read in alongside the dataset of California cities. Three of the counties are not present in my dataset and because they contain only Census-designated areas and no official cities or towns.

I decided to first examine major trends in the meteorology of California across the years encompassed by the data. The planet is warming, and I wondered whether this trend in warming or other fire-conducive conditions would be evident over the course of just a 24-year span.



(Dashed lines indicate trends contributing to or indicative of worsening fire conditions)

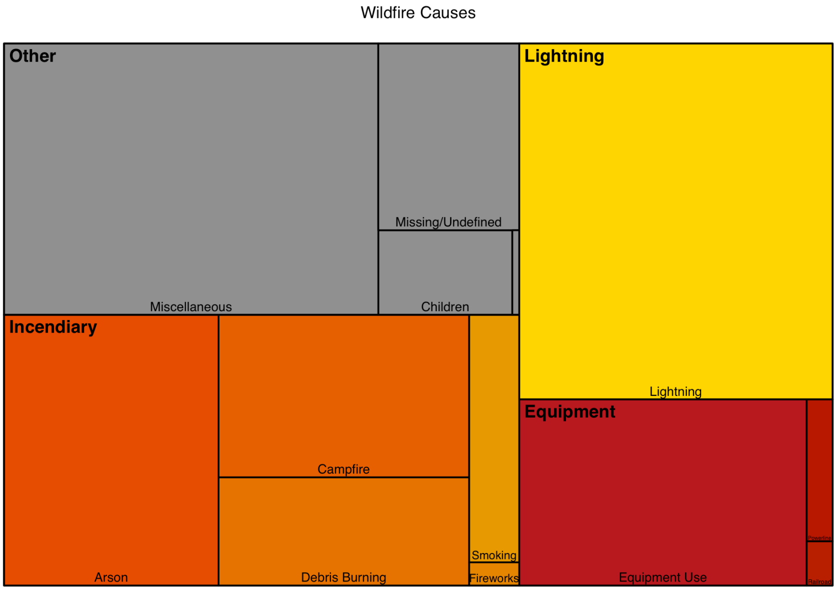
For each year from 1992 to 2015, I collected all data entries for the summer months, June through August, as measured by a single station in Fresno. I averaged these entries for six standardized meteorological indicators across each year, then fit a linear regression to those averages and plotted the results to observe overall trends in the movement of such averages.

The figure shows the direction in which the average value of each metric in the summer months is trending. Dashed lines indicate those trends corresponding to worsening fire conditions. All but one of the indicators are shifting in directions that make larger, hotter, and longer fires more likely, corroborating the previously-mentioned work done by the *Los Angeles Times* and *New York Times* showing that the devastation of the 2020 fire season is merely a continuation of an ongoing trend.

Fires start more easily when air temperature is high and relative humidity is low. As shown on the plot, relative humidity and precipitation during California fire season are both decreasing, while average summer month air temperature has risen over the past 24 years. The result, as shown by the red and blue lines, is that fires are burning larger and longer.

I then turned my attention to understanding the most typical causes of fires. The fire dataset included over dozen distinct causes to which wildfires were attributed. According to the *Los Angeles Times*, around 84% of wildfires across the country are started by human-related activities, including arson, power equipment, and discarded cigarette butts. The only true 'natural' cause of fire is lightning strikes.

I grouped fire causes into four rough groups: lightning; equipment, including vehicles, power lines, and railroads; incendiary incidents, including debris burning, arson, and campfires; and 'Other', a catch-all for children's accidents, burning buildings, spontaneous combustion of flammable materials, and blazes brought about by a combination of factors.

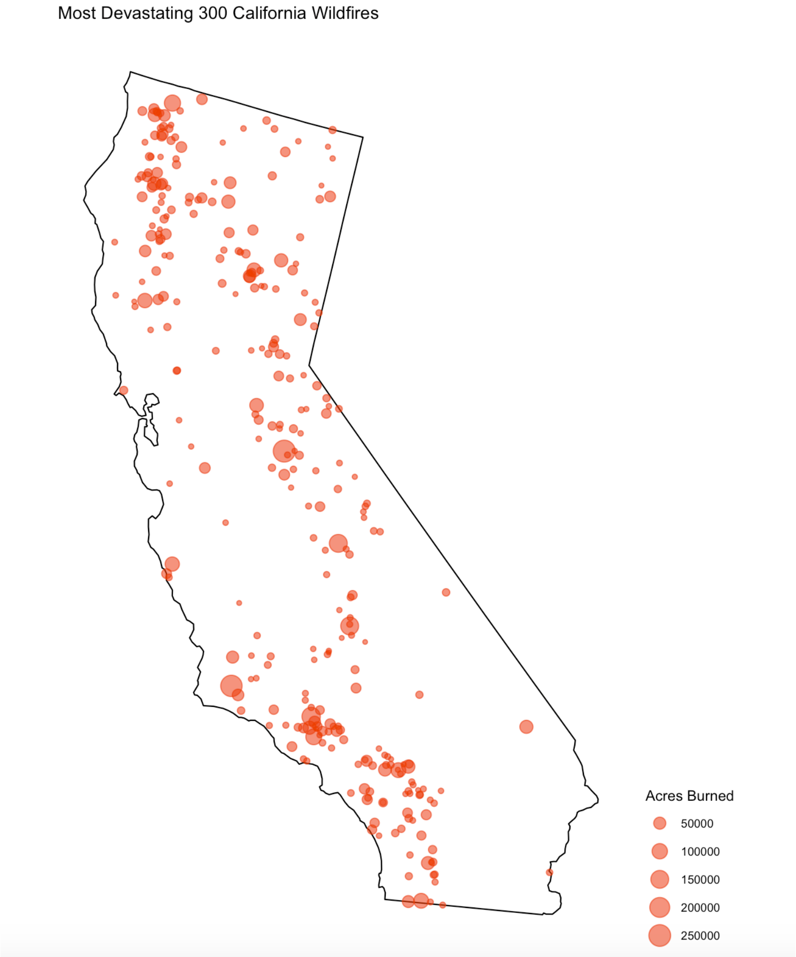


The figure falls in line with the proportions calculated by the *Los Angeles Times*, with approximately three quarters of all California wildfires in the dataset attributable to human negligence or activity. This number is likely a low estimate, considering the large proportion of causes categorized as 'Missing' or simply 'Miscellaneous'.

Wildfires are part of the natural life cycle of a healthy forest; periodic, naturally-arising fires and 'controlled burns' conducted by government personnel help clear the underbrush of dead and decaying biomass, allowing for re-fertilization of the forest floor and reinvigorated plant growth. My original intention was to analyze the fire dataset in two groups: one comprising controlled burns conducted by the U.S. Forest Service or other entities, and one limited to spontaneously-arising incidents. My intuition was that the weather conditions pre-empting these very different types of wildfires would be starkly different. However, no variable within the original USDA fire database serves as an indicator that a burn was government-mandated. As such, it is likely that the majority of fires brought about by miscellaneous causes are actually components of intentional, beneficial ecological management.

## Maps

I used the 'usmap' library to spatially analyze fire incidents across the state and by county. Understanding the geographic distribution of wildfires in the state is not only a fascinating indicator of the interplay between varying geographical and meteorological conditions, it is crucial for officials evaluating where the most urgent fire prevention action is needed.



The above bubble plot shows the spatial distribution of the 300 largest wildfires from the dataset in an effort to understand which areas of the state suffer from the worst blazes.

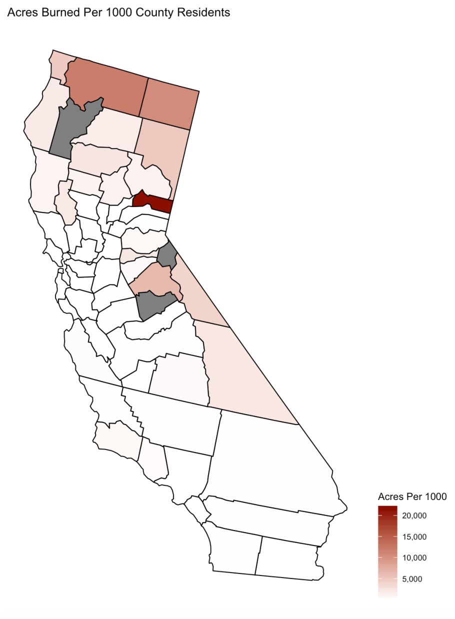
The southeastern area is known for its dryness and proximity to neighboring deserts, and tinderbox conditions arise from the combination of multiple atmospheric phenomena. The Sundowner winds blow dry air from north to south during the prime of fire season, causing humidity to plummet, while the Santa Ana winds transport desert air west from Utah and Nevada over the Sierra Nevada mountains and into the lower pressure basin of southern and central California. These winds are primarily responsible for large blazes in the Central Valley and the Los Angeles and San Diego areas.

In the north, meanwhile, the temperature is cooler and more humid, but vast sprawls of woodland provide ample kindling for fires to spread far and fast.

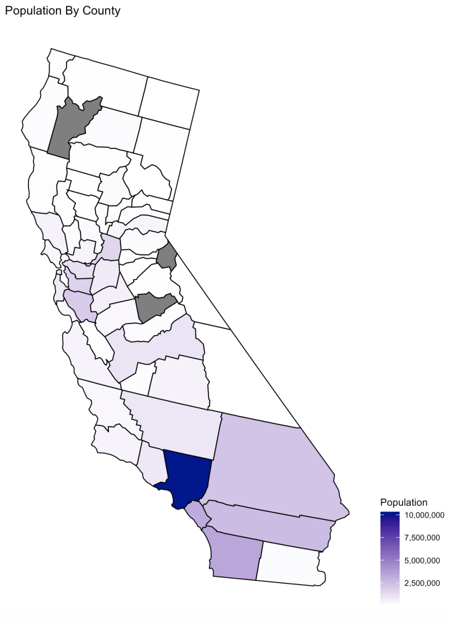
### By County

Having aggregated the data by county, I created plots to examine disparities between which counties were hit hardest by wildfires. I first graphed county population and the number of acres burned per 1,000 residents on adjacent figures.

It is clear that Los Angeles County and the southern part of the state are home to a high concentration of the total population. As such, it makes sense that acres burned per 1,000 residents is lower in the south. In the heavily forested north, home to far fewer people, fires are more widespread and difficult to contain. The smaller, more remote communities in the northern part



Acres Burned Per 1000 County Residents

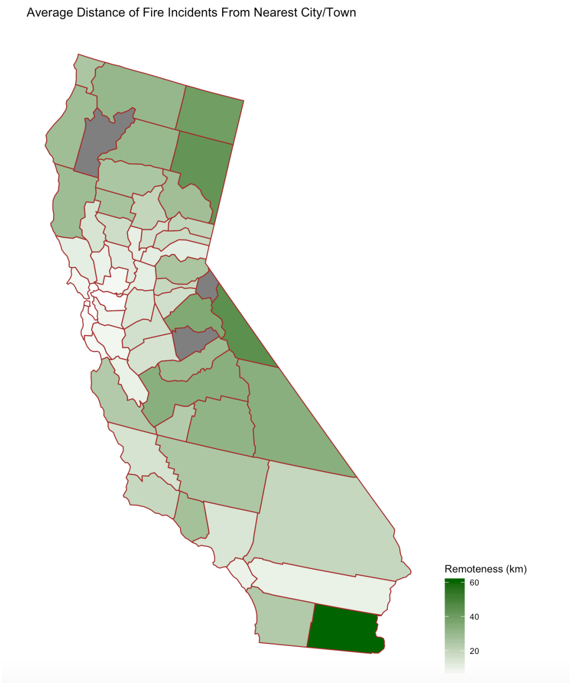


Population By County

of the state likely do not benefit from the fire-containment resources, firefighting force, and less flammable landscape composition in the way the south does.

The three counties for which my dataset contains no data are shaded gray.

## Remoteness

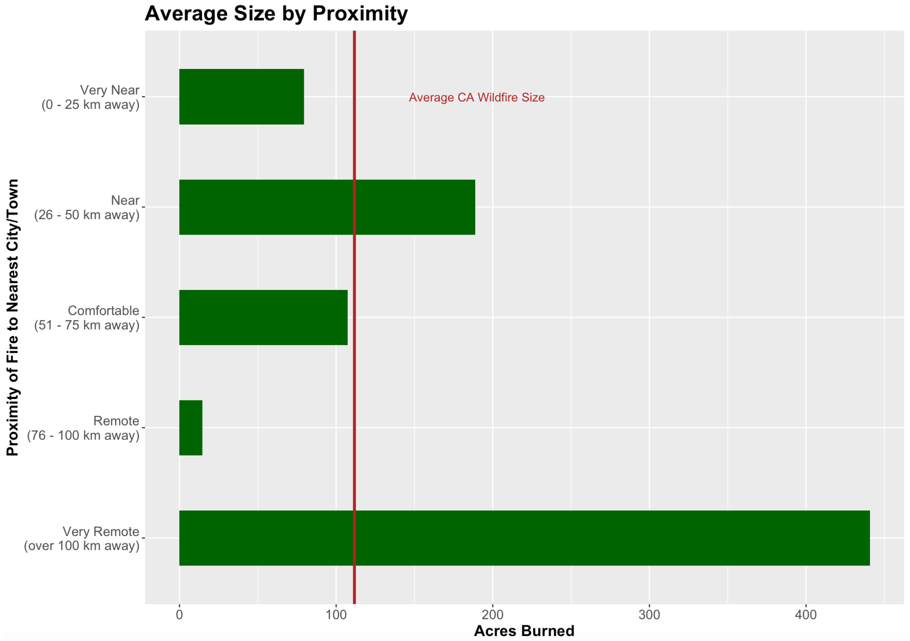


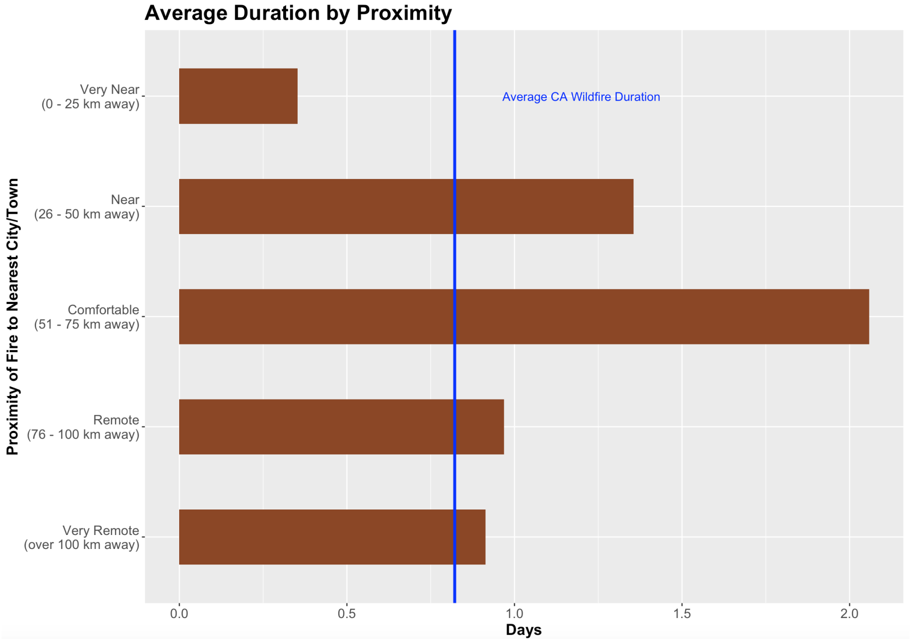
Average Distance of Fire Incidents From Nearest City/Town

Building off the idea that northern, more remote counties may lack the resources to fight fires as well as they could, I examined the 'remoteness' variable previously created to understand average fire distance from the nearest city or town within each county.

Counties outside the central Bay Area are generally more remote. In counties in the north and east, near the Sierra Nevada range, fires are generally more distant from local cities and towns. This plot supports the notion that northern and more remote counties may experience larger and more destructive fires not only because of their heavy forest coverage, but also because resources are less accessible.

I continued to evaluate remoteness by grouping counties into five tiers based on the average remoteness of their fires.





The distributions of average fire size and duration across these tiers does not follow the pattern one might expect. Fires that are over 100 km away from the nearest community tend to burn the largest footprint – this makes sense, given that the such fires are not immediate threats to human life and, as such, attempts by resources to control them may be less urgent. However, remote fires 76-100 km away from a community burn a smaller average footprint than nearby fires. One would be wrong to assume that a clear and direct association exists between fires that are closer to cities and fires that are prevented from burning broad swaths of land.

The plot of duration shows a similarly contradictory pattern. Fires that are remote and 'very remote' tend to burn, on average, for half the time of some closer fires.

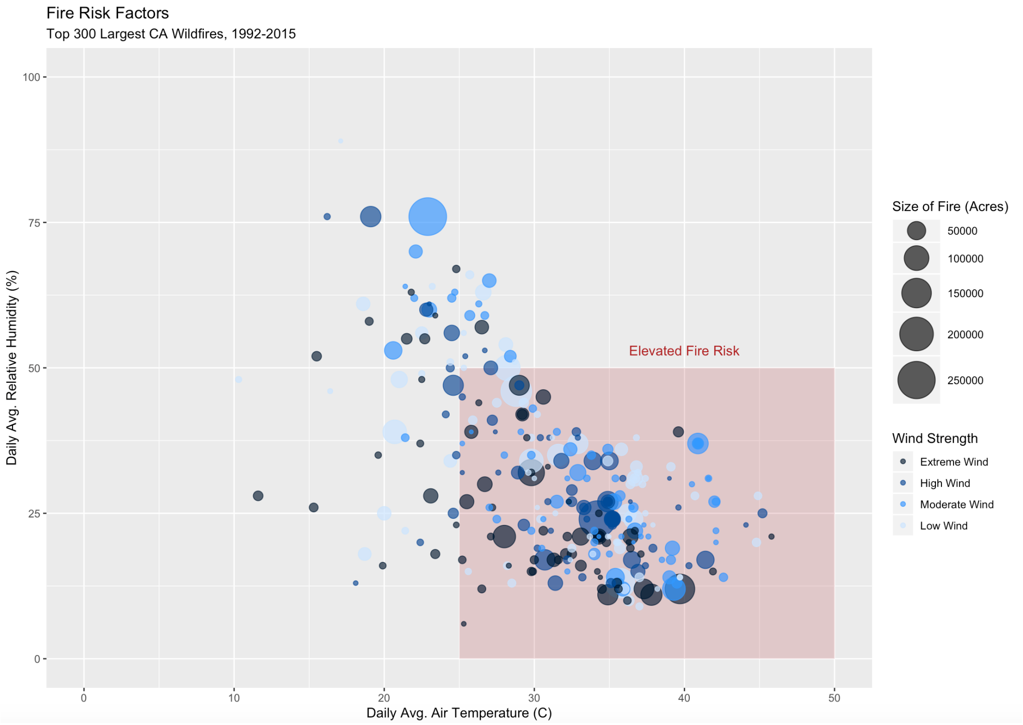
These results seem to contradict the previous notion that more remote locales have difficulty accessing resources in time to address wildfires. If that were the case, one would observe more of a consistent association between remoteness and both greater size and duration.

It is worth noting that the levels of remoteness tiers were chosen arbitrarily. Adjustment of the divides between these levels could result in different distributions or observed trends.

## Bubble Plot

Finally, I constructed a bubble plot to illuminate a more focused understanding of key meteorological parameters responsible for facilitating the spread of fires: temperature, relative humidity, and wind. Fire conditions are heightened when air temperature rises but relative humidity – or the proportion of water vapor saturation within the air – falls. Once fires begin, they are most efficiently spread by gusts of wind that carry embers and a hot mix of air and ash across forests and communities.

This figure plots daily average relative humidity against daily average air temperature for the 300 largest fires in the dataset. The size of the bubble corresponds to the size of the fire, while the color corresponds to wind strength on the day of the fire. I created this variable by breaking the 'wind run' variable into four distinct levels. Wind run is the measured 'distance' of cumulative wind that passes a given point over a particular timeframe. In this case, 'low wind' is designated by less than 130 km of daily wind run, while 'extreme wind' is indicative of wind run greater than 200 km.



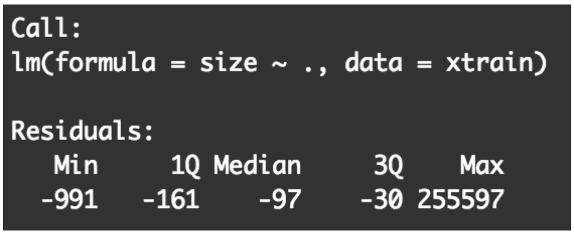
The red shaded square indicates conditions that lead to elevated fire risk: air temperatures over 25 ° C and relative humidity under 50%. It is clear that most larger fires fall into the bottom right quadrant of the plot, meaning they took place on days exceeding these meteorological thresholds. Of these fires, most larger ones are indicated by darker colored bubbles, meaning their spreads were fueled by gusting winds.

# Modeling and Results

Armed with the insights gleaned from visualizing the relationships between variables, I began to test different types and parameters of linear regression models to predict wildfire size with as low error as possible.

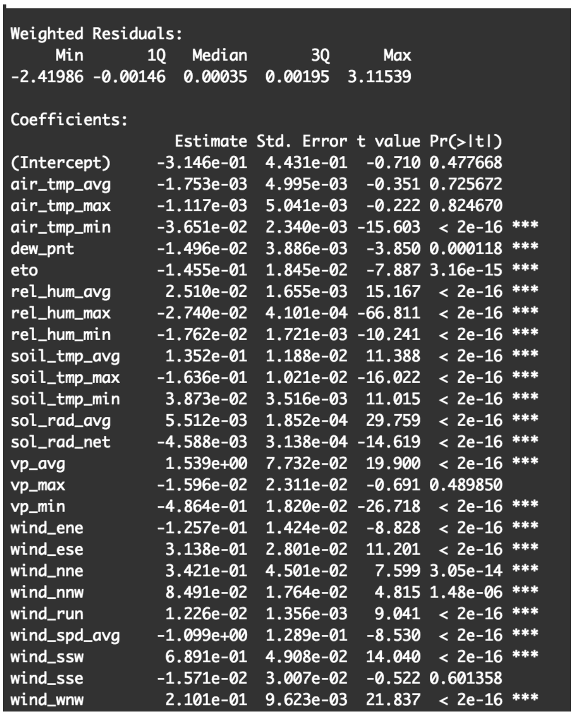
I first implemented a model of wildfire size predicted by all other variables in order to understand how a baseline model would perform. Predictors for this model included all daily meteorological variables, as well as seven- and 30-day lagged variables.

The initial model's poor fit is evident in examination of the coefficients of key variables. Air temperature, for example, has a negative coefficient, indicating that wildfire size tends to diminish



```
Call:
lm(formula = size ~ ., data = xtrain)

Residuals:
   Min     1Q Median     3Q    Max 
  -991   -161    -97    -30 255597 
```



```
Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-2.41986 -0.00146  0.00035  0.00195  3.11539 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -3.146e-01  4.431e-01  -0.710 0.477668    
air_tmp_avg    -1.753e-03  4.995e-03  -0.351 0.725672    
air_tmp_max    -1.117e-03  5.041e-03  -0.222 0.824670    
air_tmp_min    -3.651e-02  2.340e-03 -15.603  < 2e-16 ***
dew_pnt        -1.496e-02  3.886e-03  -3.850 0.000118 ***
eto            -1.455e-01  1.845e-02  -7.887 3.16e-15 ***
rel_hum_avg     2.510e-02  1.655e-03  15.167  < 2e-16 ***
rel_hum_max    -2.740e-02  4.101e-04 -66.811  < 2e-16 ***
rel_hum_min    -1.762e-02  1.721e-03 -10.241  < 2e-16 ***
soil_tmp_avg    1.352e-01  1.188e-02  11.388  < 2e-16 ***
soil_tmp_max   -1.636e-01  1.021e-02 -16.022  < 2e-16 ***
soil_tmp_min    3.873e-02  3.516e-03  11.015  < 2e-16 ***
sol_rad_avg     5.512e-03  1.852e-04  29.759  < 2e-16 ***
sol_rad_net    -4.588e-03  3.138e-04 -14.619  < 2e-16 ***
vp_avg          1.539e+00  7.732e-02  19.900  < 2e-16 ***
vp_max         -1.596e-02  2.311e-02  -0.691 0.489850    
vp_min         -4.864e-01  1.820e-02 -26.718  < 2e-16 ***
wind_ene       -1.257e-01  1.424e-02  -8.828  < 2e-16 ***
wind_ese        3.138e-01  2.801e-02  11.201  < 2e-16 ***
wind_nne        3.421e-01  4.501e-02   7.599 3.05e-14 ***
wind_nnw        8.491e-02  1.764e-02   4.815 1.48e-06 ***
wind_run        1.226e-02  1.356e-03   9.041  < 2e-16 ***
wind_spd_avg   -1.099e+00  1.289e-01  -8.530  < 2e-16 ***
wind_ssw        6.891e-01  4.908e-02  14.040  < 2e-16 ***
wind_sse       -1.571e-02  3.007e-02  -0.522 0.601358    
wind_wnw        2.101e-01  9.623e-03  21.837  < 2e-16 ***
```

as temperatures rise. This runs contrary to both intuitive assumptions and empirical ecological data.

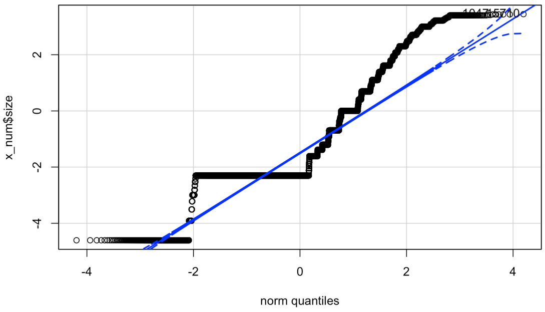
The model performed so poorly due to the extremely skewed distribution of the response variable, wildfire size. Of roughly 54,000 data points, the sizes of over half are under a single acre. The range of sizes, however, extends past 250,000 acres.

This model has little predictive power – the RMSE for predictions on a randomized test dataset was unacceptably high. Additionally, the significance of this model's predictors at the 0.05 level cannot be trusted due to the absence of normalized response variable residuals. A log transformation was insufficient to render the residuals normalized, and despite implementing a Box-Cox transform to correct the 'size' variable toward normality as best I could, this variable remains so extremely skewed that I believe a different model type is needed to evaluate it.

Typical practice is to exclude major outliers from the data before running regression, where an outlier is defined as any point 1.5 times the inter-quartile range outside the inter-quartile range. I had initially refrained from taking this step, since restricting my data to values within the traditionally acceptable range would eliminate over half the data. The IQR of the size variable, for example, is just 0.9. Eliminating typical outliers would cut my data from over 54,000 points to just over 17,000.

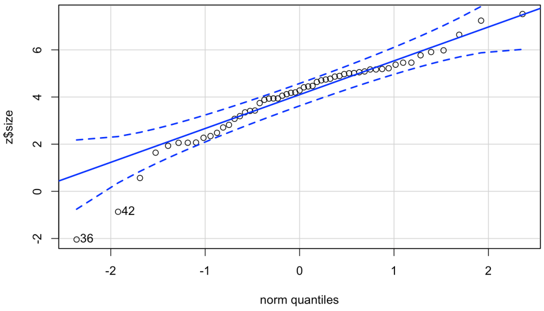
I had already determined that the transformed model yielded minimal information, and I was curious as to whether 'cherry picking' the data that would make the model look good might lead to an improved fit. However, even with size limited to values between 0.1 and 2.3 acres and every other column's values restricted to the appropriate range, size was not close to a normal distribution.

I then refined my set of predictors to exclude all highly correlated min and max values, as well as all lagged variables. The new predictors included daily air temperature, dew point, evapotranspiration, relative humidity, solar radiation, vapor pressure, wind run, and remoteness. Once more, despite curtailing outliers, limiting multicollinear predictors and implementing a Box-Cox transform, the size variable was far from normally distributed and no meaningful information could be gleaned from the model.

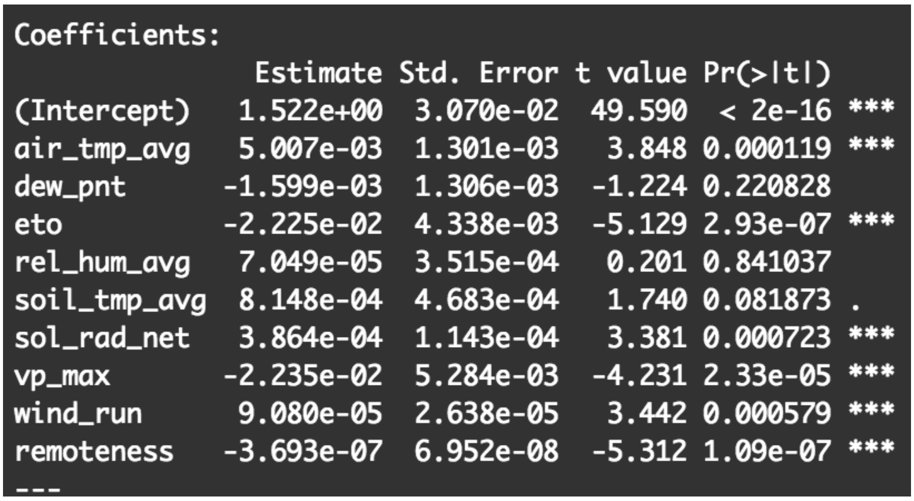


### By County

I concluded that a different type of model would be necessary to correct for the skewness of the response variable and make accurate predictions of individual wildfire sizes. Unable to meet my initial goal, I shifted my focus to predicting average wildfire size by county. I split the list of 55 county averages into a randomized training set of 44 rows and a test set of 11 rows, and I performed univariate and multivariate normality tests of the data. This time, a log transformation worked beautifully in normalizing the residuals for average wildfire size.



After limiting the input to just nine predictors in order to control for multicollinearity, I could be confident in the model's output and the significance of its predictors. I performed both-ways, backward, and forward step-wise regression; forward step-wise regression performed the best, eliminating a need to implement principal component analysis.



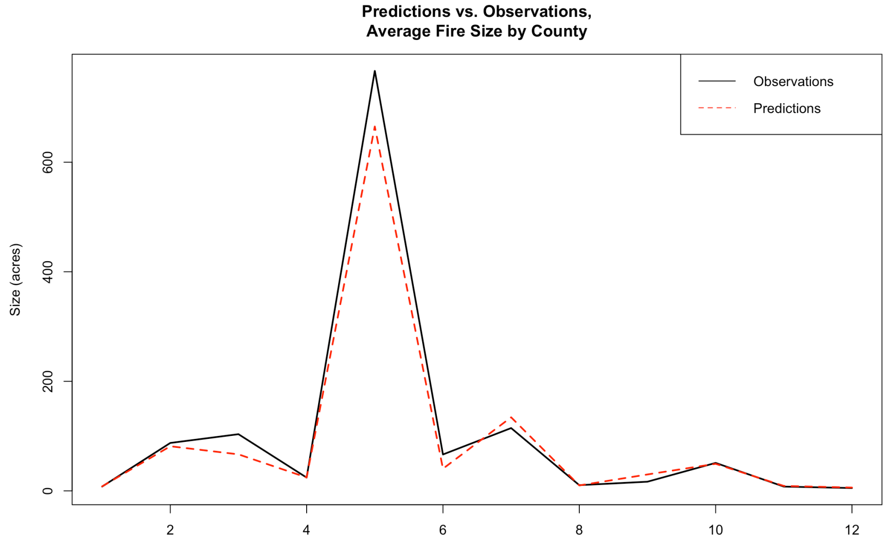
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.522e+00  3.070e-02  49.590  < 2e-16 ***
air_tmp_avg  5.007e-03  1.301e-03   3.848 0.000119 ***
dew_pnt     -1.599e-03  1.306e-03  -1.224 0.220828    
eto         -2.225e-02  4.338e-03  -5.129 2.93e-07 ***
rel_hum_avg  7.049e-05  3.515e-04   0.201 0.841037    
soil_tmp_avg 8.148e-04  4.683e-04   1.740 0.081873 .  
sol_rad_net  3.864e-04  1.143e-04   3.381 0.000723 ***
vp_max      -2.235e-02  5.284e-03  -4.231 2.33e-05 ***
wind_run     9.080e-05  2.638e-05   3.442 0.000579 ***
remoteness  -3.693e-07  6.952e-08  -5.312 1.09e-07 ***
---
```

All but three logged predictors of log size are significant at the 0.05 level. Air temperature and wind run exhibit positive coefficients, as does net solar radiation. These are three crucial ingredients for accelerating wildfire spread, and I expected their coefficients to indeed be positive based on prior visualization of their relationships with fire size.

The negative coefficients for dew point and evapotranspiration, two metrics evaluating moisture content in the air and ground, also align with expectations. The greatest surprise is a positive coefficient for relative humidity. I expected a negative relationship between relative humidity and size, since heavier, moister air seems intuitively like it would limit wildfire spread. One possible explanation is that greater relative humidity often comes with more consistent or stronger gusts of wind, which could counterbalance the influence of moisture-laden air in order to carry fires across a larger geographical footprint.

The negative coefficient for remoteness further refutes my initial assumption that remote fires are more difficult to combat, or that they are often not as urgently addressed since they do not pose immediate threats to communities. The model indicates that larger fires are generally less remote, implying that more remote fires are contained more rapidly, before they are allowed to grow. This perhaps indicates that remote fires are higher priorities for government officials and firefighters since they typically occur in heavily vegetated areas where an ample supply of combustible biomass can fuel rapid growth.

The predicted average results for 11 randomized test counties were substantially better than the predictions for the sizes of individual fires. These predictions for average fire size by county can be found in my submitted zip file.



Modeling average wildfire size per county does not provide the same granular detail that a could be gleaned from an accurate model for the size of individual blazes, but it can nonetheless be useful for ecologists and policymakers seeking to obtain a holistic familiarity with the dynamic interplay of moisture, wind, and fire spread in climatically diverse regions of California.

## Conclusion

In this project, I scraped California weather station data and combined it with population statistics, engineered variables, and a subset of a database of 1.88 million wildfires in an attempt to predict wildfire size by a combination of meteorological conditions.

After a challenging data scraping and cleaning process, all prospective linear models predicting the size of individuals failed to meet crucial assumptions for normalized residuals, despite model variations incorporating hand-picked variable selection and the implementation of weighted models and Box-Cox transforms. The size data represented within this dataset is subject to so much inherent skewness that it is counterproductive and uninformative to fit it with even the most precisely transformed linear model. I believe a different distribution, such as a Poisson, or a non-parametric machine learning technique, such as a decision tree, would be better suited to generate accurate predictions based on such data.

Failing to meet my initial goal, I re-calibrated my model to predict average fire size by California county. The revised model met assumptions of normality, and its coefficients confirmed that the most influential factors for accelerating the spread of wildfires are air temperature and wind. Building upon this type of model in complexity and accuracy could provide officials with general understanding of the specific meteorological conditions of areas across California, allowing fire containment and safety protocols to be adjusted accordingly.

The tumultuous set of results obtained through exploration of wildfire data in this project leads me to conclude that prediction of individual wildfire sizes is a task too nuanced for a linear model to successfully undertake. Next steps to improve upon the efforts made in this project would involve a non-parametric approach or, perhaps, classification rather than regression. Fires are grouped into classes based on their sizes:

- class A: less than 0.25 acres
- class B: 0.25 - 10 acres
- class C: 10 - 100 acres
- class D: 100 - 300 acres
- class E: 300 - 1,000 acres
- class F: 1,000 - 5,000 acres
- class G: greater than 5,000 acres

I would be interested in exploring whether a classification-based approach to this problem would lead to better results. At the very least, such an approach would allow the model more latitude, rather than compelling it to identify the a singular size value for the wildfire based on evolving, erratic weather conditions.

The inability of this model to predict specific size values speaks to the inherent unpredictability of individual wildfires. With this project, I sought to understand whether there are fundamental markers or particular combination of conditions that indicate, with any certainty, the likelihood of a fire to spread to devastating proportions. This project has resoundingly rejected that notion, forcing me to conclude that wildfire spread is largely non-deterministic and subject to the volatility of unique circumstances. My visualizations revealed that some of the largest California fires spread on days of relatively little wind, moderate temperatures, and substantial humidity – subpar fire conditions.

The best meteorological models examine large-scale trends in atmospheric circulation and surface heating on far greater timescales than individual days. Perhaps an accurate model of wildfire spread is only achievable based on analysis of patterns with higher levels of stability and longevity than daily weather conditions.