Predicting wildfire incidence and impact through spatial and weather data
Final Report

Background

Research question

What patterns of land use and weather can explain the incidence and severity of wildfires? We built a predictive model to attempt to answer this question, using a dataset on 1.8 million recorded wildfires in the US (1992-2015) to explore methods of statistical prediction. We cross-check this model with another dataset on fire incidence that covers the state of California, 1951-2000, using older years as a test set.

We focus on predictors that are relevant to disaster risk management, such as changes in land use and drought severity. These predictive methods could help policymakers in arid areas like the American Southwest be warned early so they can take proactive action in the face of increasing risk from climate change.

Policy relevance

Recent extreme wildfire events like those in California and Australia have highlighted the growing importance of fire management. Climate change is making major fire outbreaks in populated areas increasingly likely. It is estimated that in the western US alone, 55% of the increase in fuel aridity (a common metric of fire risk) over the last three decades can be attributed to anthropogenic climate change, contributing to an additional 4.2 million ha of burned area (Abatzoglou & Williams, 2016). On the global scale, climate change is projected to lead to an increase in wildfire risk in untouched tropical ecosystems where few natural fires occur today (Price & Rind, 1994).

Climate change threatens conventional methods of fire risk management like controlled burns by making their consequences less predictable. Also, since a large share of wildfires are ignited by chance events like lightning strikes, they are by nature difficult to pinpoint (Price & Rind, 1994). There is a need, therefore, for better statistical analysis to identify long-term tendencies in wildfire incidence and extent, such as sensitivity to interannual drought variability and the effect of land use changes.

The findings of such analysis could inform systemic changes in agricultural and land use policy. In agricultural policy, changes such as limits on land clearing for agriculture and zoning laws may discourage development in risk-prone areas. In land use planning, better predictions of wildfire risk could inform localities' zoning policies, such as mandating construction with fire-safe building materials or requiring "defensible space" (vegetation-free areas) around structures at risk.

Similar concerns apply to agricultural policy -- more precise geographic estimation of wildfire risk could help authorities decide where to conduct prescribed burns as well as designating at-risk tracts of land as conservation areas.

Regarding emergency response planning, better wildfire modeling could reduce basis risk in disaster insurance and inform where to allocate resources for evacuation readiness. Finally, wildfire risk modeling could inform where states like California get their power -- distributed energy sources, like solar, could both reduce at-risk areas' risk of blackouts during a fire and reduce ignition risk by minimizing the amount of accident-prone transmission lines.

Data description

Data sources

Our primary source of data for this project is a USDA dataset that contains information on every recorded wildfire in the US from 1992 through 2015 - a total of about 1.88 million observations. The dataset (the Fire Occurrence Database, hereafter FOD) contains information on discovery date, ignition point coordinates, and fire extent (in km) for each observation.

Our predictor variables cover a number of relevant climactic, ecological and land use factors. The key climate metric for fire risk is the so-called "fuel moisture content", which measures the latent potential for fire outbreak. On a large scale like ours, fuel moisture content is typically measured as some function of average rainfall, temperature and humidity -- we get these data from the WorldClim database, which contains detailed climate records covering the entire US from 1970 to 2000. All of the climatological variables are available at a resolution of 1/12 degree.

We also incorporate information on human land use from the USGS National Land Cover Database, which classifies areas by their designed use (agricultural, industrial, residential, etc). The NLCD is updated infrequently and only goes back to 2001, so we pick a representative year and assume it is constant over time.

As a supplement to our national analysis, we also consider a preprocessed fire incidence dataset from Mann et. al. (2016) that covers California over the years 1951-2000. This dataset has a slightly different set of predictor variables, covering moisture balance, topography and housing density. The longer time span and different predictors of this dataset allow us to cross-check our main model.

Data processing

Before we can construct a model of wildfire risk, we must define our outcome indicators, our predictor data and our strategy for how to reconcile data from multiple time and spatial scales.

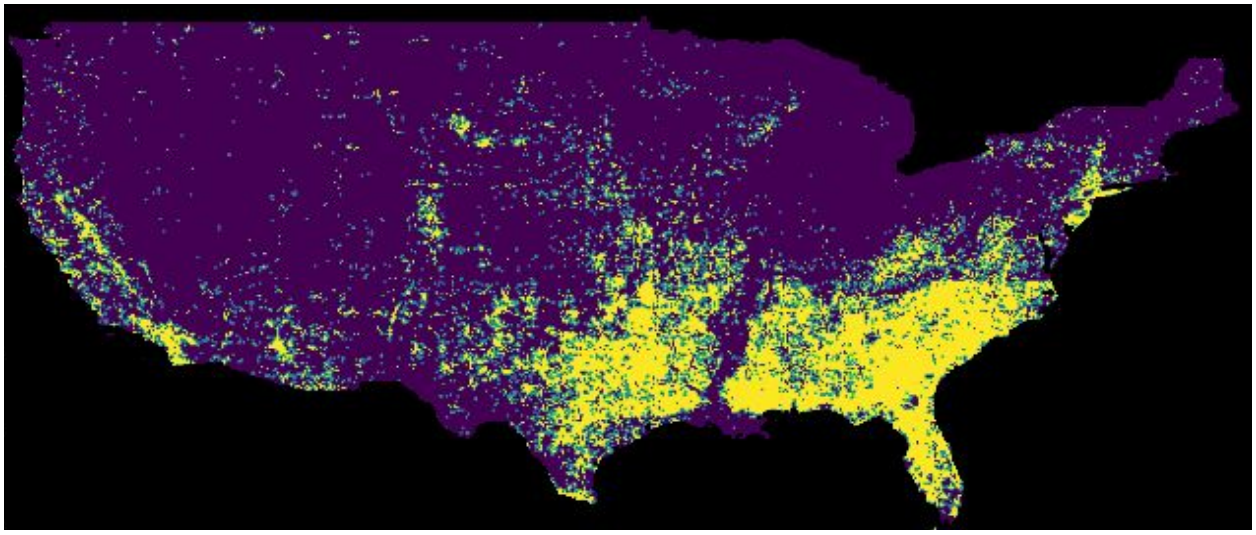
Our climatology data has a resolution of 1/12 degree, so we aggregate everything else up to that grid level. For our outcome variable, fire incidence, we simply count the number of fires in each grid cell,

based on their ignition coordinates. For land cover, which is a categorical variable, we take the mode (most common category) in each grid cell.

Since climatology is expressed as monthly averages, we also aggregate fire incidence by the month in which it occurred.

Summary statistics

After preprocessing, we find that about 30% of pixels in the US have at least one fire recorded. Below is an illustration of all fires that occured in January as an illustration:



Analysis

Model selection

The choice of model for this setting depends on whether or not we believe our data is a comprehensive record of all wildfires over the sample period. If it is not, we must treat it as unbalanced -- meaning that we only have a sample of the presence of wildfires, and no data on their absence. Conventional prediction methods like logistic regression do not perform well in such settings. Instead, we should apply a method that is well-suited to presence-only data, such as maximum entropy estimation. In this setting, we would apply maximum entropy as an "unconditional" method of estimation rather than a "conditional" one, since we only have data for places where y = 1 (i.e. where there is a fire outbreak). We could then apply Bayes' rule along with some assumptions about the underlying sampling methodology to get some estimate of p(y=1|x) from a model of p(x|y=1).

If we do believe our data is comprehensive, however, then a conditional model is preferable. In that case, we could apply methods like logistic regression. However, balance (i.e. significantly more 0's than 1's) is still a potential threat to validity in this setting.

Given time constraints, we were only able to test conditional models on this data, specifically logistic regression and neural networks. However, as discussed in the conclusion, unconditional models may be a promising solution to some of the prediction issues we encountered in this setting.

Analysis 1 – National data

For our first analysis, we test a logistic regression model for wildfire classification, using the national FOD data.

Model
We transform the fire count data into a binary indicator – 1 for any fires over the sample period, 0 for none. These indicators are **per month**. Our predictor variables include climatology (temperature, precipitation and wind speed) and land use classification (a dummy variable for each category of land use, 16 in total).

Since our data are spatial, we must account for spatial autocorrelation in the outcome and predictor variables. We do this by adding spatial "lags" for each variable, where the *k*th spatial lag is defined as the average of that variable over the grid cells at a taxicab distance *k* from the center. We test both 1st and through the 5th spatial lags in this analysis.

We train our model on a random 80% of the national data and test it on the remaining 20%. On the neural network, we had 70% used for training, 10% used for intermediate validation, and the remaining 20% for testing.

Results
The model fit statistics for our logistic regression model are as follows:

- **Accuracy: 85%**
- **Precision: 77%**
- **Recall: 69%**

For reference, the null accuracy (i.e. predicting all values to be 0's) of this data is 70%. This suggests that our model offers some novel information, but its predictive power is limited.

The coefficients on land use classification are also interesting in their own right, as they offer some indication of which types of land are the greatest risk factors, including human settlement.

Variable name	Description	Coefficient
fires_L1_%	First lag of fire count	2.191
tavg_L1_avg	First lag of average temp	0.117
prec_L1_avg	First lag of precipitation	0.001
srad_L1_avg	First lag of solar radiation	0.000
wind_L1_avg	First lag of wind speed	-0.119
vapr_L1_avg	First lag of vapor pressure	-0.361
LC11_L1_%	Open water	-0.099
LC12_L1_%	Perennial ice/snow	0.000
LC21_L1_%	Developed open space	0.045
LC22_L1_%	Developed low intensity	0.020
LC23_L1_%	Developed med intensity	0.013
LC24_L1_%	Developed high intensity	0.002
LC31_L1_%	Barren land	-0.037
LC41_L1_%	Deciduous forest	0.108
LC42_L1_%	Evergreen forest	0.303
LC43_L1_%	Mixed forest	0.076
LC52_L1_%	Shrub/scrub	-0.308
LC71_L1_%	Grassland	-0.059
LC81_L1_%	Pasture / hay	0.067
LC82_L1_%	Cultivated crops	-0.321
LC90_L1_%	Woody wetlands	0.113

We can see that residential development and pasturing (variables highlighted in red) are positively associated with wildfire risk, whereas barren or minimally cultivated land (variables highlighted in blue) are negatively associated with wildfire risk.

Robustness

The results above do not change significantly between the models with 1 and 5 spatial lags. They also do not change significantly when variables are standardized or a neural net classification model is used instead of logistic regression. Even rebalancing the data (i.e. randomly subsetting such that there is an equal share of 1's and 0's in the outcome variable) does not improve model performance.

Analysis 2 - California data

Model

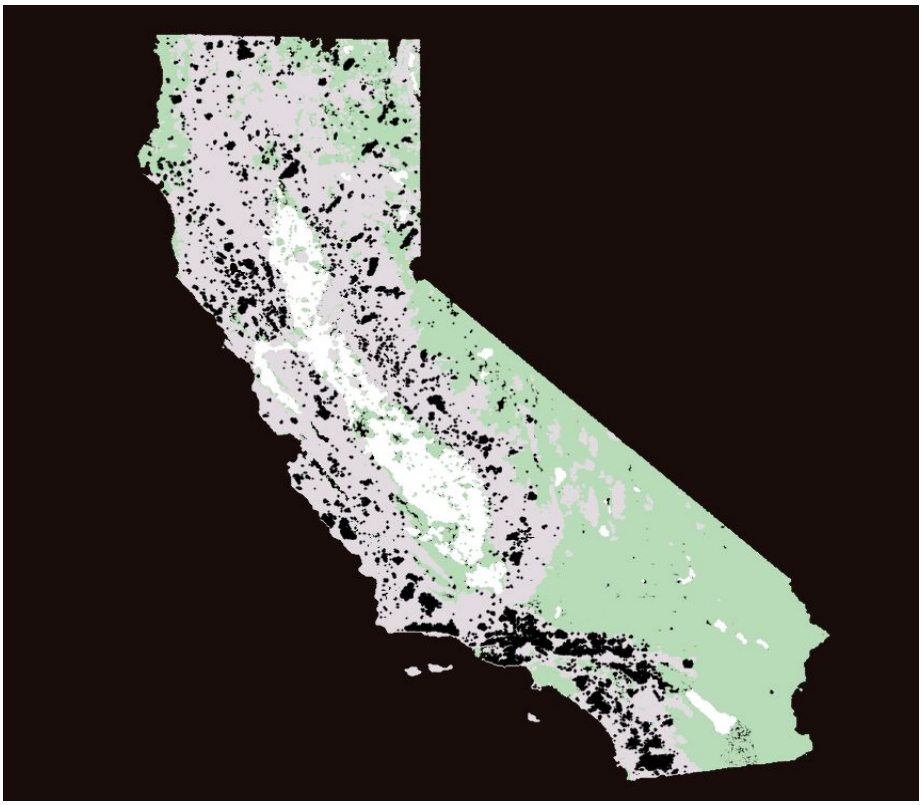
As a supplement to our national model, we try a similar approach on the Mann et. al. (2016) California dataset. As described above, this dataset includes a somewhat different set of predictors, and is based on wildfire data from the state of California, but the basic model formulation is the same.

We separate the data into three periods: 1951 - 1975, 1976 - 2000, and 2001 - 2025, where the latter is based on forecasted climate data from CMIP5 rather than empirical observations. Our approach is to train the model on 1976 - 2000, test it on 1951 - 1975, and use the 2001 - 2025 period as a forward-looking projection. As an additional cross-check, we can also compare the forecasted burned area in 2001 - 2015 to the actual fire incidence data from the FOD dataset used in Analysis 1.

Results

- **Accuracy: 60%**
- **Precision: 11%**
- **Recall: 86%**

Null accuracy for this dataset is 92%. This model performs worse on accuracy than the national model (likely driven by the lower base rate), but much better on recall, which is arguably a more relevant metric in this setting.



The black area on this map is the actual burned area 1951-1975, and the grey is the predicted burned area from our model for the same period.

We can also project into the future, where it is apparent that our modeled burned area will increase in the coming years.

Robustness

In addition to training on different subsets of years as described above, we can also compare our model to the FOD data (not used for estimation) as a cross-check:

Orange dots are FOD fire ignition points (scaled by intensity) 2001 - 2015, and the grey area is our predicted burned area over the same period. We can see that there is generally high coincidence between the two datasets with the exception of some fires in the southeastern portion of the state (which, as a desert climate, might be less amenable to this kind of analysis).

The results of our Analysis 2 model do not change significantly when using decision trees or random forests.

Conclusion

Summary of analysis

The mixed performance of our models suggests that the data is not well-separable by any function. In particular, all of our analyses yield a high degree of false positives, even after rebalancing the training data, trying different predictors / datasets and trying non-linear prediction models like neutral nets.

In this setting, false positives may be less of a concern than false negatives, given the high downside risk of failing to adequately prepare for a wildfire. The recall metric (disregarding false positives) is fairly good in some specifications. Additionally, the coefficients on land use in the regression model may have interesting policy implications in their own right.

Despite these limited successes, it is worth considering why wildfires appear so resistant to easy prediction, despite the wealth of data available. This diagnosis can point the way toward better models in the future.

Potential explanations

Balance
One potential explanation for the limited success of the model is class imbalance in the data, given that most (~70%) grid cells never had a fire over the sample period. However, we have enough observations to safely rebalance the data, and model performance does not improve after rebalancing. Thus, balance is likely not a sufficient explanation.

Nonlinearity
Another issue might be that fuel moisture content – the ultimate driver of wildfire risk – might not be a simple additive function of our climatological variables. Indeed, models to directly estimate fuel moisture content often have a complicated functional form (cf. Ceccato et. al. 2003).

Were this the case, we would expect a neural network classifier to perform better than logistic regression, because neural networks consider higher-order polynomials and interaction terms in their model search. However, neural networks do not appear to offer an improvement in this case. Moreover, other papers (ex. Mann et. al. 2016) have been successful in predicting long-term fire tendency with a simple functional form. Therefore this also does not appear to be a sufficient explanation.

Appropriate predictors
A related explanation is that we may have chosen the wrong climate variables – long-term tendencies in temperature, precipitation and wind speed may not be appropriate for predicting fire risk. Mann et. al. (2016) addresses this by using evapotranspiration and climatic water deficit instead, as well as incorporating topography. However, our attempt to replicate our methodology using their variables (Analysis 2) does not yield appreciably better results. We do get somewhat better recall with the Mann et. al. dataset, but also lower accuracy than the national model. Thus this is a partial explanation at best.

Time scale
Our model considers wildfire risk on a climatological scale – that is, average tendencies on the span of decades. However, it might be the case that most variability in wildfire risk is driven not by long-term spatial differences, but by interannual or seasonal variability. If so, our model is simply on the wrong scale.

There is some merit to this idea – interannual and decadal climate variability in the US were much greater than long-term trend variability over this period, and most practical wildfire risk models used by forestry services are on a much shorter time scale than ours. Again, however, other analyses like Mann et. al. (2016) have had success predicting wildfire risk on a climatological scale. This suggests that our efforts are not totally off base, and that we just have not found the right combination of data and modeling approach.

Comprehensiveness of sample
Finally, there is the issue we raised in the "Model Selection" discussion - can our data truly be considered a comprehensive record of wildfires over the sample period? If not, than the imbalance in our data is much greater than measured, motivating the use of "presence-only" methods, as discussed previously.

Recognizing this, Mann et. al. (2016) use a "zero-inflated" negative binomial model, which amounts to estimating two separate regressions - one for the presence of any fire, one for the conditional count of fires - and using the results of the former to adjust the latter. The better performance of their model suggests there is some merit to this explanation.

Policy implications

Effects of built environment
One clear policy implication from this analysis is that human activity appears to be a risk multiplier for fire risk, as captured by the positive coefficients on residential and pasture land. In contrast, open / uncultivated space appears to reduce fire risk.

Together, these findings motivate "defensible space" policies, which require properties in areas of high fire risk to maintain a minimum perimeter of cleared land. Some states already have such policies, but they could be expanded, particularly as the risk of wildfire outbreak is projected to become greater in the coming decades.

Potential effects of climate change
On that note, our findings highlight two important facts about the effects of climate change: First, as demonstrated by our California analysis, climate change will almost certainly increase the extent of wildfire outbreaks in the highest-risk states. Second, the extent of that increase is still highly contested. For instance, previous estimates of the increase in burned area in California by the end of the century range from +15% and +50% (Mann et. al. (2016)). The difficulties we encountered in our own analysis illustrate how the scientific community is still a long way from being able to present projections of the increase in wildfire risk that are precise enough to be actionable for local-level policymakers.

Importance of interannual variability
Last, as mentioned in the "potential explanations" section, there may be more value in predicting wildfires on the interannual or decadal scale than the long-term climatological scale, even if one is concerned with adaptation to climate change. Not only is there more climate variability on this scale, but the methods for prediction are better understood. By taking action to become more resilient to extreme weather events *today,* local governments may be better positioned to adapt to climate change in the future - the principle behind Columbia University's ACToday project.

Direction for further research

Unconditional models
Based on our diagnosis above, the most promising direction for future research appears to be models built for presence-only data, such as the zero-inflated negative binomial. Unconditional / generative probability distribution estimation methods like maximum entropy have also had promising applications in ecology (ex. Phillips et. al. (2006)).

Higher-frequency data
Additionally, it would be informative to estimate a similar model using higher-frequency (ex. yearly or monthly) data, as observation suggests there is more variability at this scale.

Restricting to major fires
Last, it might be helpful to take the opposite approach and narrow the scope of our model rather than broadening it -- specifically, by restricting our analysis to only very large fires. Major events might be more distinctive in our data. This would limit the generalizability of our model but could potentially improve its performance.

Appendix

Member contributions

- Max Mauerman co-wrote this report, contributed to the presentation and contributed analysis ideas and literature review.
- Pei Yin Teo conducted analyses on the California dataset, contributed to the pre-processing for the land cover data and contributed to the presentation.
- Mughil Pari preprocessed the climatology and FOD data and ran the logistic regressions, as well as tested the neural networks. He also contributed to the requisite topics on the presentation.
- Priyanka Sethy built the presentation and contributed to the report.

References

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences*, *113*(42), 11770–11775. https://doi.org/10.1073/pnas.1607171113

Bryant, B. P., & Westerling, A. L. (2014). Scenarios for future wildfire risk in California: Links between changing demography, land use, climate, and wildfire: SCENARIOS FOR FUTURE WILDFIRE RISK IN CALIFORNIA. *Environmetrics*, *25*(6), 454–471. https://doi.org/10.1002/env.2280

Ceccato, P., Leblon, B., Chuvieco, E., Flasse, S., & Carlson, J. D. (2003). Estimation of Live Fuel Moisture Content. In *Wildland Fire Danger Estimation and Mapping: Vol. Volume 4* (pp. 63–90). WORLD SCIENTIFIC. https://doi.org/10.1142/9789812791177_0003

Defensible Space. (n.d.). *Los Padres ForestWatch*. Retrieved March 7, 2020, from https://lpfw.org/fire/defensible-space/

Dixon, L., Tsang, F., & Fitts, G. (n.d.). *California's Fourth Climate Change Assessment*. 105.

Fire-Safe Homes. (n.d.). *Los Padres ForestWatch*. Retrieved March 7, 2020, from https://lpfw.org/fire/fire-safe-homes/

FLASSE, S. P., & CECCATO, P. (1996). A contextual algorithm for AVHRR fire detection. *International Journal of Remote Sensing*, *17*(2), 419–424. https://doi.org/10.1080/01431169608949018

Flasse, S., Trigg, S., Ceccato, P., Perryman, A., Hudak, A., Thompson, M., Brockett, B., Drame, M., Ntabeni, T., Frost, P., Landmann, T., & Roux, J. le. (2004). 8. Remote Sensing Of Vegetation Fires And Its Contribution To A Fire Management Information System. *USDA Forest Service / UNL Faculty Publications*. https://digitalcommons.unl.edu/usdafsfacpub/189

Gutiérrez-Vélez, V. H., Uriarte, M., DeFries, R., Pinedo-Vásquez, M., Fernandes, K., Ceccato, P., Baethgen, W., & Padoch, C. (2014). Land cover change interacts with drought severity to change fire regimes in Western Amazonia. *Ecological Applications*, *24*(6), 1323–1340. https://doi.org/10.1890/13-2101.1

Homer, C., Dewitz, J., Jin, S., Xian, G., Costello, C., Danielson, P., Gass, L., Funk, M., Wickham, J., Stehman, S., Auch, R., & Riitters, K. (2020). Conterminous United States land cover change patterns 2001–2016 from the 2016 National Land Cover Database. *ISPRS Journal of Photogrammetry and Remote Sensing*, *162*, 184–199. https://doi.org/10.1016/j.isprsjprs.2020.02.019

Is Your Home In a Fire Hazard Severity Zone? (n.d.). Retrieved March 7, 2020, from https://www.arcgis.com/home/item.html?id=5e96315793d445419b6c96f89ce5d153

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163. JSTOR.

Land Use Planning to Reduce Wildfire Risk. (2016, January 19). Headwaters Economics. https://headwaterseconomics.org/wildfire/solutions/lessons-five-cities/

Mann, M. L., Batllori, E., Moritz, M. A., Waller, E. K., Berck, P., Flint, A. L., Flint, L. E., & Dolfi, E. (2016). Incorporating Anthropogenic Influences into Fire Probability Models: Effects of Human Activity and Climate Change on Fire Activity in California. *PLOS ONE*, *11*(4), e0153589. https://doi.org/10.1371/journal.pone.0153589

National Land Cover Database. (n.d.). Retrieved March 7, 2020, from https://www.usgs.gov/centers/eros/science/national-land-cover-database?qt-science_center_objects=0#qt-science_center_objects

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Price, C., & Rind, D. (1994). The Impact of a 2 × CO2 Climate on Lightning-Caused Fires. *Journal of Climate*, 7(10), 1484–1494. https://doi.org/10.1175/1520-0442(1994)007<1484:TIOACC>2.0.CO;2

Roberts, D. (2019, October 28). *Wildfires and blackouts mean Californians need solar panels and microgrids*. Vox. https://www.vox.com/energy-and-environment/2019/10/28/20926446/california-grid-distributed-energy

Short, K. C. (n.d.). *Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508] (4th Edition)* [Data set]. https://doi.org/10.2737/RDS-2013-0009.4

Some Wonder if Electric Microgrids Could Light the Way in California. (n.d.). Retrieved March 7, 2020, from https://pew.org/32fF7F3

Stevens-Rumann, C. S., Kemp, K. B., Higuera, P. E., Harvey, B. J., Rother, M. T., Donato, D. C., Morgan, P., & Veblen, T. T. (2018). Evidence for declining forest resilience to wildfires under climate change. *Ecology Letters*, *21*(2), 243–252. https://doi.org/10.1111/ele.12889

WorldClim Version2 | WorldClim—Global Climate Data. (n.d.). Retrieved March 7, 2020, from http://worldclim.org/version2

Yebra, M., Chuvieco, E., & Riaño, D. (2008). Estimation of live fuel moisture content from MODIS images for fire risk assessment. *Agricultural and Forest Meteorology*, *148*(4), 523–536. https://doi.org/10.1016/j.agrformet.2007.12.005