Fire Susceptibility in the western United States with ConvLSTM

By Caroline Juang (csj2116)

EAEE E4000: Machine Learning for Environmental Engineering and Sciences
Professor Pierre Gentine
Fall 2021

This final project was collaborated on/ discussed with Jianing Fang (jf3423), but was carried out by Caroline Juang.

Files included in this project (separate attachments)

1. **Preprocessing.pdf** – Python notebook file with data preparation.
2. **Training.pdf** – Python notebook file with data preprocessing, model training, and validation/visualization.

Introduction

Fire activity, in both fire frequency and in forest area burned, is increasing in the western United States (US) due to a combination of anthropogenic and natural climate change (Abatzoglou & Williams, 2016). Though fire is a naturally-occurring phenomenon in western US forests, rapid increases in burned area in 2020 and 2021 resulted in record-breaking fire years in a four-decade record (Aspegren, 2021). With continued warming and drying and abundant fuel left to burn in western US forests, there will be continued increases in burned area (Abatzoglou et al., 2021); suggesting needed fire management to prevent possible risks to human health and infrastructure and extreme disturbances to forest ecosystems. Improved modeling and forecasting methods are required to understand which regions may be the most vulnerable to this future fire load, and which environmental and climate factors strongly related to fire activity could be controlled.

Empirical methods of modeling western US forest fire are popular because of the strong correlations between fire and weather and climate variables, especially between vapor-pressure deficit (VPD) and forest-fire area burned (Williams & Abatzoglou, 2016). However, modeling fires is challenging because there is a nonstationarity in fire-climate relationships—fire behavior in the past as a function of its past drivers does not necessarily mean the fires will behave the same way with the same drivers in the future. For example, anthropogenic climate change is resulting in forested areas in Yellowstone National Park to burn more frequently, from every 100-300 years to every <30 years, which can rapidly transform the structure of forests and make them more vulnerable to new fire activity (Westerling et al., 2011). Human-led fire management and suppression can modify which areas fires are allowed to burn, or ignite new fires, modifying where and when fires could potentially burn (Higuera et al., 2015). Empirically-driven statistical modeling studies generally use linear regression to tie forest-fire area burned to potential drivers, (e.g. Littell et al., 2018). Machine learning is increasingly being used for fire modeling, done with decision tree-based techniques (e.g. Parisien & Moritz, 2009; Tehrany et al, 2019). These techniques capture the spatial patterns of fires and fire, but without a temporal aspect, they do not capture the nonstationarity that occurs as fires or non-climate factors modify the environment.

In this final project, I aim to combine the spatial and temporal information of past fire events to improve our understanding of forest-fire burned area and its drivers, even with the nonstationarity of fire. Using a Convolutional Long-Short Term Memory (ConvLSTM) framework, I used five drivers that span climate and non-climate variables that include physical environment and human-related drivers to generate monthly predictions of whether an area will or will not burn.

Methods

Forest Fire and Drivers

The area of study in this project is the western US, which is the continental US west of 103 degrees longitude, and the time period 1984-2019. This area of study is part of my Ph.D. research, and for my research I have been developing a database of western US fires by combining the Monitoring Trends in Burn Severity (MTBS) data product based on Landsat data (Eidenshink et al., 2007) and smaller government agency databases. The final database used contains more than 18,000 fires in forest and non-forest regions, shown in Figure 1 (Juang et al., in review). The database was transformed into a monthly gridded dataset of area burned in gridcells at a resolution of 12 km. All other datasets were also transformed into 12 km-resolution gridded data products. Each gridcell contains area burned in km^2.

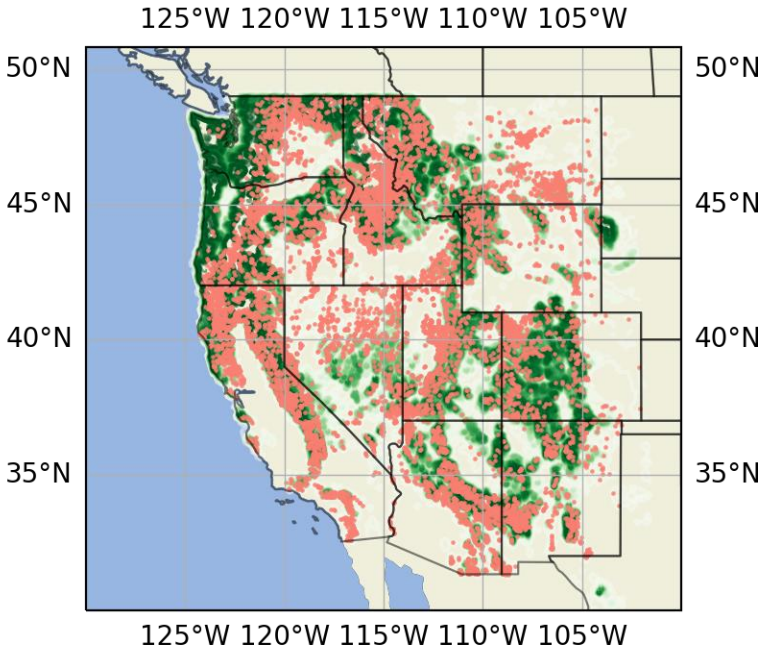


Figure 1: Locations of fires (1984-2019) in the Juang et al. (under review) database, in red. The size of the points are not to scale with fire size. In green: the fractional forest area for regions of the western US.

In this study, I used five variables—past area burned, average wind speed, VPD anomaly (VPDz), maximum temperature anomaly (Tmaxz), and the number of months since a fire last burned (return months)—to create a simple model relating drivers of forest-fire burned area to the complex range of climate, environment, and human factors (Table 1). These particular variables were selected because of their large influence on fires in the western US and because they are assumed to be uncorrelated, i.e. I did not choose precipitation in combination with VPD because they vary too closely and will generate autocorrelation. Vapor-pressure deficit (VPD) was an especially important variable, as this variable has been repeatedly shown to be a dominant correlate with forest-fire area burned in the western US and regions within, and acts as a proxy for aridity because it can represent the physical process of vegetation drying out to become fuel for fire (Abatzoglou & Williams, 2016; Williams et al., 2019). Wind speed is important because extreme winds can potentially lead to rapid growth of burned area within seconds to days. Maximum temperature is important to the drying of fuels similar to VPD, but also to potentially encouraging fuel growth.

Table 1: Variables for model inputs and outputs. The data preparation column describes the additional steps needed after all data were preprocessed into 12-km resolution gridded products.

Driver input variables	Unit	Data preparation in code
Past area burned	km^2	None, except to sum per month.
VPDz	hPa	Separate all data into EPA Level II ecoregions. For each ecoregion, calculate (current-month VPD) – (1984-2019 average VPD).
Average wind speed	m/s	None
Maximum temperature anomaly (Tmaxz)	°C	Separate all data into EPA Level II ecoregions. For each ecoregion, calculate (current-month VPD) – (1984-2019 average VPD).
Months since fire last burned (return months)	Months	Using the burned area dataset, create a new variable that counts up from months since the fire last burned, and resets the counter when the fire burned in current-month.
Variable to predict	**Unit**	**Preprocessing**
Probability an area burned	Probability, 0 (unburned) or 1 (burned)	Using the burned area dataset, set all gridcells that burned in a given month to 1, and all gridcells with 0 burned area to 0.

VPD is the difference between saturation vapor pressure (e_s) and actual vapor pressure (e_a). Monthly mean e_a was calculated from monthly mean dew point from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) Climate Group (Daly et al., 2004). Monthly mean e_s was calculated from mean maximum daily temperature (T_{max}) and mean minimum daily temperature (T_{min}), obtained from the National Oceanic and Atmospheric Administration (NOAA) Climgrid dataset (Vose et al., 2014). Wind speed data at 10-m were acquired from the ERA-5 reanalysis product (Hersbach, et al., 2020).

Data preprocessing and assumptions

Additional steps were taken to prepare the data before it can be added to the statistical model. Given that the western US is comprised of varying and sometimes opposite climates at different times of the year, I decided to transform VPD and maximum temperature (Tmax) into

the VPDz and Tmaxz variables. This was done by separating the variables into North America level II ecological regions ("ecoregions") (Commission for Environmental Cooperation, 1997). Within each region, the 1984-2019 average for VPD or Tmax was subtracted from the current-month's observed value. Therefore, the model's input is a deviation from the average for VPD and Tmax. A new variable was also created, the "return months". This variable counts the number of months since a fire last returned in each gridcell, and will reset the counter to zero when a fire burns in the current month, counting up from 1984 to 2019. One assumption is that the counter all starts at zero in 1984.

When the model underwent preprocessing in the next setup step, the data was converted from a NetCDF-like file format in the *xarray* data library to a *numpy* array. Furthermore, in order to run the data through a machine learning model, I converted all NaNs in the original dataset to 0. These NaNs originally defined the boundaries of the western US and so the NaNs are located over small parts of the ocean, Canada, Mexico, and the midwestern states. By transforming these into zeros, the model may possibly pick up that there is no fire in these locations and also no wind/no change in VPD/no change in temperature/fire currently burned in that month.

Machine Learning Model Setup

In this study, I created a ConvLSTM framework using an example designed for frame-by-frame video prediction on the Moving MNIST dataset in Keras in TensorFlow (Joshi, 2021). The Long-Short Term Memory (LSTM) part of the algorithm identifies patterns in a temporal sequence of data, created first by Hochreiter & Schmidhuber (1997). The LSTM unit is then run through a convolutional neural network (Shi et al., 2015). The benefit of ConvLSTM is that the algorithm will first capture temporal correlations in the data inputs, and then can learn the spatial information by stacking the LSTM cells in a convolution operation (Shi et al., 2015; Panda,

2021). The difference between the Moving MNIST dataset example and my example is that instead of inputting only the previous timesteps of the image as 1 channel, I am instead inputting 5 channels for the fire drivers (Table 1).

The dataset underwent preprocessing to ultimately create a single dataset that has a 5-dimensional shape of (300, 432, 10, 10, 5). There are 300 groups of 10x10-gridcell spatial datasets, 432 timesteps for all months in the time period, and 5 channels representing each of the variables. Each channel was first refined to 150x200 spatial gridcells, representing X and Y directions in space in the entire western US. The larger grid was grouped into 10x10 gridcells (with each gridcell of 12 km resolution), resulting in 300 groups.

For training and testing, the dataset was split in consecutive order, allowing the earlier characterization of western US fire activity to predict the future fire regime. The first 232 months were used as training data, and the last 200 months were used as validation data. The variable to predict for both training and testing is a binary dataset of 0s and 1s, where 0 means the fire did no burn in a gridcell in the current month, and a 1 signifies the fire burned.

The simple schematic for putting the data through the ConvLSTM model is illustrated in Figure 2. The model is run through 3 ConvLSTM2D layers with batch normalization, and then a single ConvLSTM3D layer. The details are expressed in Table 2. I added regularization forms of early stopping with a patience of 10, and reduced the learning rate when a metric has stopped improving, in order to avoid overfitting. The model was trained on 20 epochs with a batch size of 5.

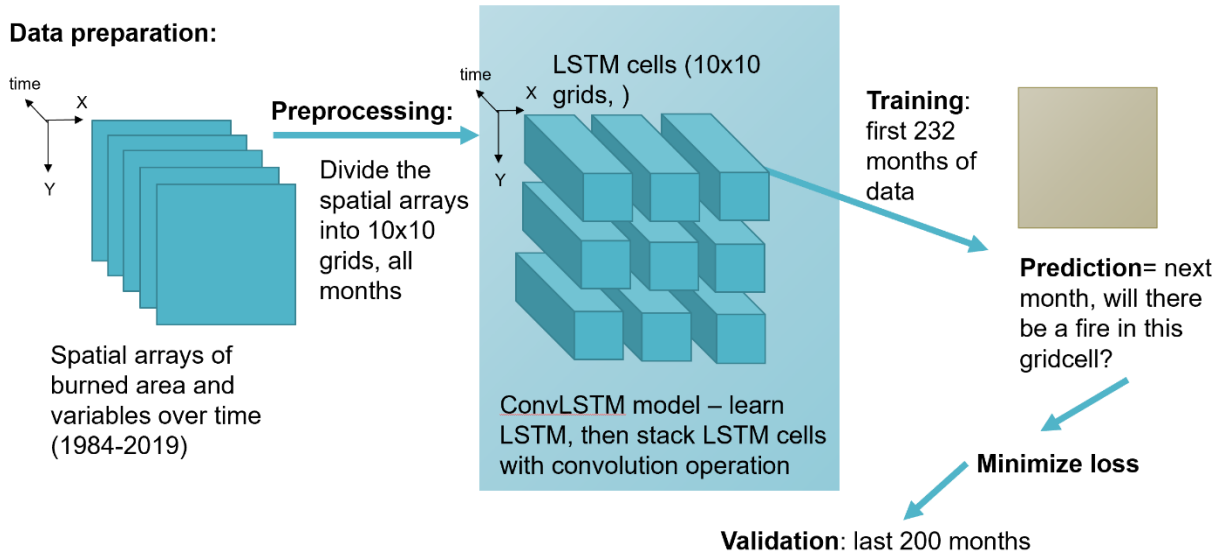


Figure 2: Schematic of the ConvLSTM setup. The five variables used in the project are first prepared, then all data is preprocessed into groups of 10x10 grids (with each grid 12km by 12km in resolution). The groups are fed through the ConvLSTM model, which is trained on the first 232 months (first ~19 years) of data in order to output the predicted value, the probability an area burned. The model is validated on the last 200 months of data.

Table 2: ConvLSTM model setup, consisting of three ConvLSTM2D layers and a single ConvLSTM3D layer.

Layer	Hyperparameters
(1) ConvLSTM2D	Filters=8 Kernel size=(5,5) Padding= “same” Return_sequences=true Activation= “sigmoid”
(2) ConvLSTM2D	Filters=8 Kernel size=(3,3) Padding= “same” Return_sequences=true Activation= “sigmoid”
(3) ConvLSTM2D	Filters=8 Kernel size=(1,1) Padding= “same” Return_sequences=true Activation= “sigmoid”
(4) ConvLSTM3D	Filters=1 Kernel size=(3,3,3) Padding= “same” Activation= “sigmoid”

Results

The model was fairly efficient, and ran in about 30 minutes. The model successfully reduced the loss as it progressed in epochs (Figure 3). The validation loss was less than the model loss, showing that the model's regularization successfully avoided overfitting.

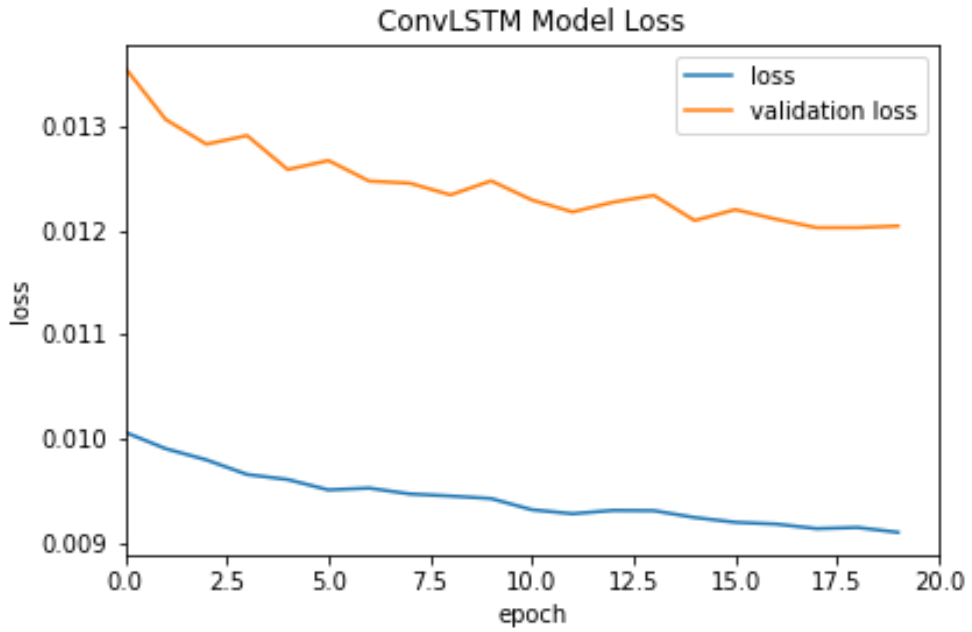
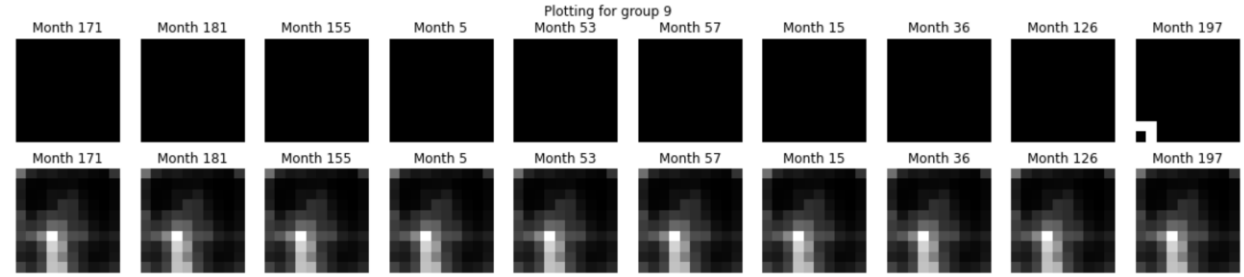


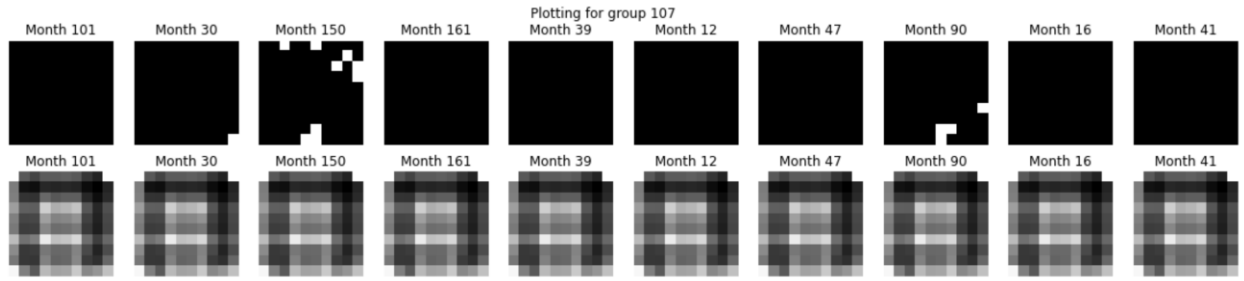
Figure 3: Model training loss (in blue) and validation loss (in orange).

Next, I analyzed the model results visually by checking through several random groups of 10x10 gridcells, which are recorded in Figure 4. These gridcell groups are small samples of the entire western US region at 12km resolution per gridcell, and the group number is arbitrary except to express the location of the group in the map of the western US. In each month frame along the bottom row, I expect to see predictions of fire occurrence, with a value closer to 0 indicating that a fire will not occur (and shaded black in Figure 4), and a value closer to 1 indicating a fire will occur in that month (and shaded white in Figure 4). The most significant feature is that the predictions for fire occurrence vary between the groups; however, the fire predictions do not vary month-to-month. Even in months in the original data where fire burned, like Group 107-Month 150, the prediction does not change.

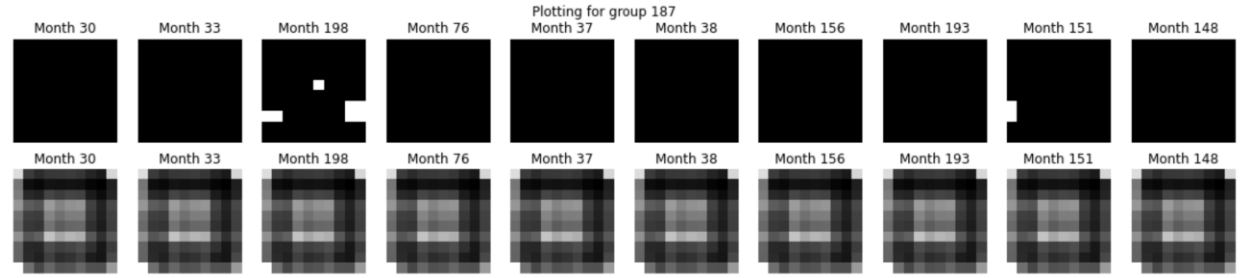
Group 9:



Group 107:



Group 187:



Group 200:

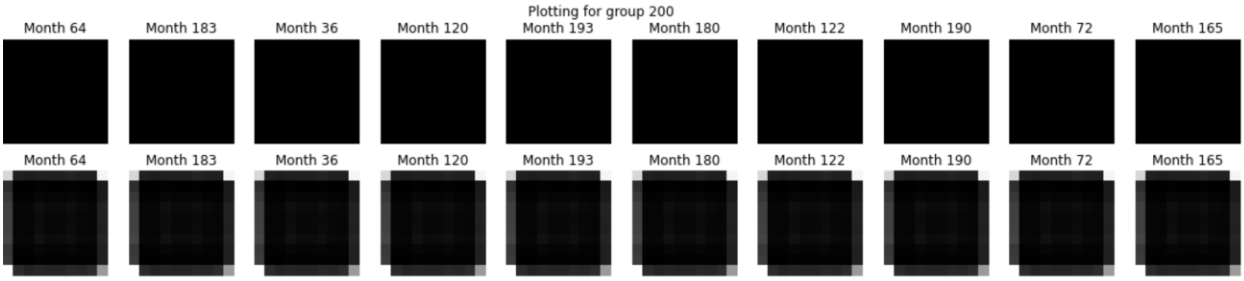


Figure 4: Visual validation of model, through random sampling of 10x10 groups. Both original data (top rows) and predicted data (bottom rows) are plotted for each group. In each gridcell, a black gridcell color equals 0 which indicates no fire. A white gridcell color equals 1 which indicates fire burned.

Discussion and Future Work

The results indicate that the model currently does not work as intended, but there are many promising paths to improvement. Based on the prediction differences between groups in Figure 4, the model is using the inputs of forest-fire drivers to identify spatial patterns. However, it is not good at predicting actual occurrence of fire and differentiating from non-fires. I hypothesize the ratio of fire to non-fire gridcells is too large, and therefore future efforts should reduce the number of groups containing zero fires. Listed, are also several steps along the process where assumptions should be tested:

Preparing data:

- Train the model again without using anomaly values for VPD and Tmax; instead, run these with their observed values.
- Include topographic data (e.g. elevation) and human activity data (e.g. distance to wildland-urban interface) to account for these large drivers of fire.
- Add filter layers to differentiate NaN gridcells from true 0s in the data (which would show “0” for NaN, and “1” for data).
- Consider whether including the number of months since the fire returned adds value as a driver of future fire occurrence. Determine if there’s a variable that accurately characterizes how fire modifies its future fire activity.
- Consider whether including past burned area helps determine future fire occurrence.

Preprocessing the data:

- Change the size of the grouped gridcells, from 10x10 gridcells to a larger number (e.g. 20x20 gridcells) and see if including a larger spatial area adds value to training the model.

- Change the amount of time that is included in the LSTM part of the model, if possible.

Model training:

- In this study, I could not use the "ReLu" activation function for the ConvLSTM2D layers and had to use "sigmoid". This could affect the model results because it affects the hidden layers.
- Remove the current regularization to see what happens if the model is overfit to the data.

ConvLSTM is not commonly used for predicting fire susceptibility. This research has the potential to contribute value to the community by applying this well-known computer vision technique to understanding the spatiotemporal patterns of fire. At its current stage, other machine learning methods are more successful at predicting fires than ConvLSTM, but there are many paths to improving this current model. I plan to continue to explore these paths in the near future in my Ph.D. research.

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

Abatzoglou, J. T., Battisti, D. S., Williams, A. P., Hansen, W. D., Harvey, B. J., & Kolden, C. A. (2021). Projected increases in western US forest fire despite growing fuel constraints. *Communications Earth & Environment*, *2*(1), 1-8. https://doi.org/10.1038/s43247-021-00299-0

Aspegren, E. (2021, July 13). Over 50 large fires are burning in 10 states across the western US. Here's how they stack in history. *USA Today.* Retrieved December 23, 2021. https://www.usatoday.com/story/news/weather/2021/07/12/western-united-states-wildfires-heat-waves-temperatures/7945217002/.

Commission for Environmental Cooperation (1997). Ecological Regions of North America: Toward a Common Perspective. *Commission for Environmental Cooperation*. Accessed on December 31, 2021. https://gaftp.epa.gov/EPADataCommons/ORD/Ecoregions/cec_na/CEC_NAeco.pdf.

Daly, C., Gibson, W., Doggett, M., Smith, J., & Taylor, G. (2004). Up-to-date monthly climate maps for the coterminous United States. *14th AMS Conference on Applied Climatology*.

Eidenshink, J., Schwind, B., Brewer, K., Zhu, Z., Quayle, B., & Howard, S. (2007). A Project for Monitoring Trends in Burn Severity. *Ecology*, *3*(1), 3–21.

Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

Higuera, P. E., Abatzoglou, J. T., Littell, J. S., & Morgan, P. (2015). The changing strength and nature of fire-climate relationships in the northern Rocky Mountains, USA, 1902-2008. *PloS one*, *10*(6), e0127563. https://doi.org/10.1371/journal.pone.0127563.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Joshi, A. (2021). Next-Frame Video Prediction with Convolutional LSTMs. *Keras Code Examples: Computer Vision.* Retrieved December 20, 2021. https://keras.io/examples/vision/conv_lstm/.

Juang, C.S., A. P. Williams, J. T. Abatzoglou, J. K. Balch, M. D. Hurteau, & M. A. Moritz (2021). Causes of Exponentially Increasing Burned Area with Rising Aridity in the Western United States. *Geophysical Research Letters.* In review.

Littell, J. S., McKenzie, D., Wan, H. Y., & Cushman, S. A. (2018). Climate change and future wildfire in the western United States: an ecological approach to nonstationarity. *Earth's Future*, *6*(8), 1097-1111. https://doi.org/10.1029/2018EF000878

Panda, R. (2021, June 14). Video Frame Prediction using ConvLSTM Network in PyTorch. *Medium*. Retrieved December 20, 2021. https://sladewinter.medium.com/video-frame-prediction-using-convlstm-network-in-pytorch-b5210a6ce582.

Parisien, M. A., & Moritz, M. A. (2009). Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecological Monographs*, *79*(1), 127-154. https://doi.org/10.1890/07-1289.1.

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802-810). https://arxiv.org/abs/1506.04214.

Tehrany, M. S., Jones, S., Shabani, F., Martínez-Álvarez, F., & Bui, D. T. (2019). A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial

data. *Theoretical and Applied Climatology*, *137*(1), 637-653. https://doi.org/10.1007/s00704-018-2628-9.

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, M. J., Williams, C. N., Fenimore, C. Jr., Gleason, K., & Arndt, D. (2014). Improved historical temperature and precipitation time series for U.S. climate divisions. *Journal of Applied Meteorology and Climatology*, **53**(5), 1232–1251. https://doi.org/10.1175/JAMC-D-13-0248.1

Westerling, A. L., Turner, M. G., Smithwick, E. A., Romme, W. H., & Ryan, M. G. (2011). Continued warming could transform Greater Yellowstone fire regimes by mid-21st century. *Proceedings of the National Academy of Sciences*, *108*(32), 13165-13170. https://doi.org/10.1073/pnas.1110199108.

Williams, A. P., & Abatzoglou, J. T. (2016). Recent advances and remaining uncertainties in resolving past and future climate effects on global fire activity. *Current Climate Change Reports*, *2*(1), 1-14. https://doi.org/10.1007/s40641-016-0031-0

Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman-Morales, J., Bishop, D. A., Balch, J. K., & Lettenmaier, D. P. (2019). Observed impacts of anthropogenic climate change on wildfire in California. *Earth's Future*, *7*, 1–19. https://doi.org/10.1029/2019EF001210.