

# California Wildfires: Predicting Wildfire Spread Using Machine Learning

---

**Gwen Squires**

Dept. of Mathematics | Southeast Missouri State University

*Graduate Mentor: Daniel Breininger | Faculty Advisor: Dr. Nezamoddin N. Kachouie*

Florida Institute of Technology



# Introduction

- **4 million** acres burned, 2020
- **2.6 million** acres burned, 2021
- **\$148.5 billion** in costs, 2018
- **30%** emissions increase, 2020
- **50,000** premature deaths, 2008-18



McKinney Fire in Klamath National Forest, CA. (CNN, 2022)

# Previous Work

Environmental Impact Assessment Review (EIA) 2020 | 18(4)

Estimating the probability of wildfire occurrence in Mediterranean landscapes using Artificial Neural Networks

Marie Elia<sup>1</sup>, Maria D'Este<sup>2</sup>, Davide Acci<sup>3</sup>, Vincenzo Gianico<sup>4</sup>, Giuseppino Spasari<sup>5</sup>, Antonio Ganga<sup>6</sup>, Giuseppe Calogrosso<sup>7</sup>, Raffaele Laferlotta<sup>8</sup>, Giovanni Sansi<sup>9</sup>

**Abstract**

Wildfires are a major disturbance in the Mediterranean Basin and an ecological fire indicator. In this context, it is crucial to understand their dynamics and their effects on the landscape. This study aims to estimate the probability of wildfire occurrence in the Mediterranean area. We used a neural network to estimate the probability of wildfire occurrence in the Mediterranean area. The model was trained on a dataset of 10,000 wildfires and tested on a dataset of 5,000 wildfires. The model achieved a high accuracy of 0.95. The results show that the probability of wildfire occurrence is higher in the Mediterranean area. This study provides a valuable tool for estimating the probability of wildfire occurrence in the Mediterranean area.

**1. Introduction**

Wildfires are a key driver of many natural landscapes and for the delivery of ecosystem services (Carnell et al., 2016; Medhurst et al., 2015). However, wildfires have detrimental effects on natural resources and human life when they occur in urban landscapes (Carnell et al., 2017; Medhurst et al., 2015; van Wieren et al., 2012).

Reports of the European Commission suggest that over the past 30 years Europe has seen an increase of extreme wildfire events generating major socio-economic impacts (Elia et al., 2019; Sansi et al., 2017; Medhurst et al., 2015). In Italy, the majority of the wildfires

occur in similar to that of other Mediterranean (Elia et al., 2019; Sansi et al., 2017). In 2017, wildfires occurred in the peninsula burning an area of 100,000 hectares (Elia et al., 2019). While the mean number of wildfires is 20 per year (Elia et al., 2019). Despite the consistent support of the European Commission and the efforts of national and regional governments to improve fire management policies, these data depict a downward picture. The authors intend to estimate the probability of wildfire occurrence in the Mediterranean area using a neural network. The model was trained on a dataset of 10,000 wildfires and tested on a dataset of 5,000 wildfires. The model achieved a high accuracy of 0.95. The results show that the probability of wildfire occurrence is higher in the Mediterranean area. This study provides a valuable tool for estimating the probability of wildfire occurrence in the Mediterranean area.

Journal of Applied Ecology 2015, 52, 107–115  
© 2015 The Authors. Journal compilation © 2015 British Ecological Society

Climate change and the eco-hydrology of fire: Will area burned increase in a warming western USA?

Donald McKenzie<sup>1,2</sup> and James S. Littell<sup>2</sup>

<sup>1</sup>U.S. Forest Service, Pacific Wildland Fire Science Lab, Seattle, Washington, USA  
<sup>2</sup>USGS, Rocky Mountain Climate Science Center, Anchorage, Alaska, USA

**Abstract**

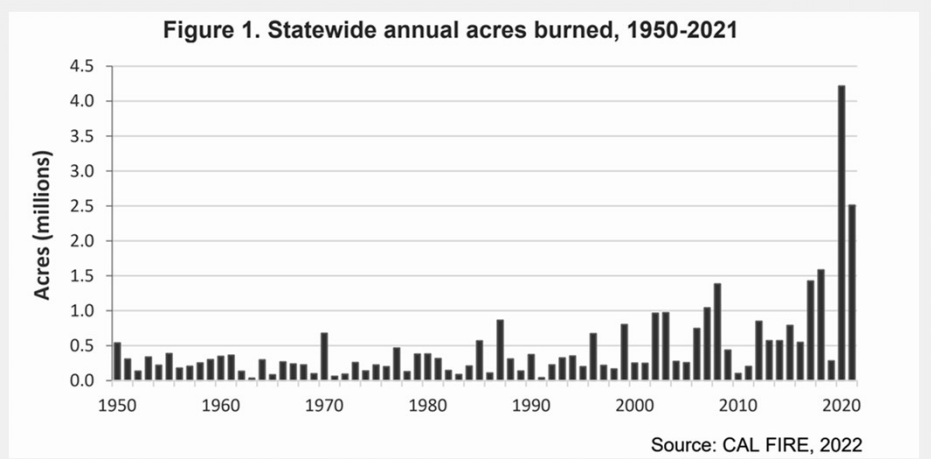
Wildfire area is predicted to increase with global warming. Empirical statistical models and process-based simulations agree globally. The key relationship for this sensitivity, observed at multiple spatial and temporal scales, is between drought and fire. Predictive models often focus on components which this relationship appears to be particularly strong, such as mean and old forest and shrublands with substantial biomass such as chaparral. We examine the drought fire relationship, specifically the correlations between water balance deficit and annual area burned, across the full gradient of deficit in the western USA. From temperate rainforest to desert, in the middle of the gradient, correlations on vegetation (leaf), correlations are strong, but outside this range the equivalence factor and other spatial extent for other biotic drivers or in conjunction with other factors such as previous year climate. This suggests that the regional drought fire dynamics will not be stationary in future climate, nor will other complex contingencies associated with the variation in fire extent. Problems of future wildfire area therefore need to consider not only vegetation changes, as some dynamic vegetation models do, but also potential changes in the drought fire dynamics that will occur in a warming climate.

**Key words:** climate change, ecosystem, drought, regime, vegetation, feedback, sustainability, water-balance deficit.

**Introduction**

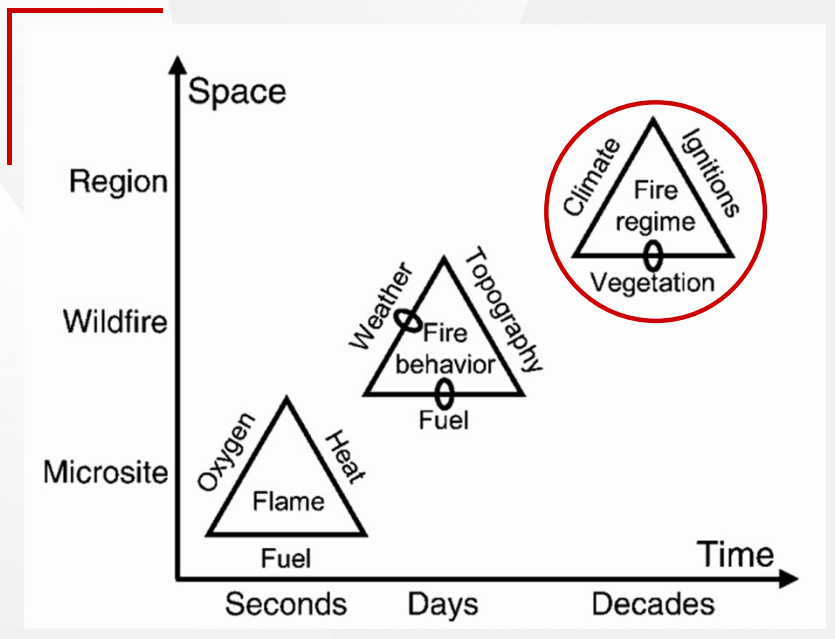
Wildfire area is predicted to increase with global warming. The straightforward nature of warming climate affecting the regimes is compelling and is supported by both empirical evidence and process-based models. For example, Flannigan et al. (2009) reviewed the climate fire literature and found wide agreement on projections of increased area burned in a warmer climate. If statistical models are projected into future climate space that represent even moderate warming scenarios, estimates of future area burned are so large as to imply broad-scale changes to ecosystem composition, structure, and function, with consequences for ecosystem services. For example, the fairly stark statistical models of McKenzie et al. (2004) predicted fire in fire-killed trees in annual area burned in the western USA under a moderate warming scenario. At that scale, and using a simple statistical approach focusing on changes in fire-size distribution, McKenzie et al. (2011) found major implications for the Greater Yellowstone ecoregion: increased area burned shortened fire cycles to the point that smaller forests were expected to change to more to shrublands.

Such projections, even when statistical and robust at process-based algorithms are fully mechanistic, assume stationarity of fire-climate dynamics within the geographic domain for which they are projecting area burned. This is not realistic. The regional drought fire dynamics will not be stationary in future climate, nor will other complex contingencies associated with the variation in fire extent. Problems of future wildfire area therefore need to consider not only vegetation changes, as some dynamic vegetation models do, but also potential changes in the drought fire dynamics that will occur in a warming climate.



Acres burned by year. (OEHPA, 2022)

# Purpose



Spatial and temporal scales. (Parisien & Moritz, 2009)

## Ignition vs. Spread

# Objective

Determine the flammability of California's landscapes by predicting whether fire will spread given an ignition occurs at a specified time and location.



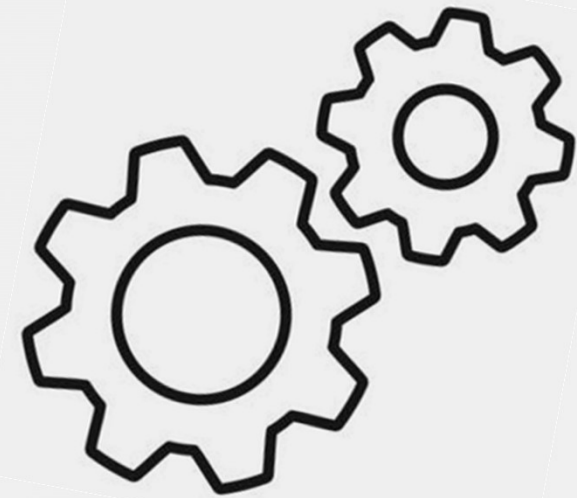
# Procedure

Data Processing

Time Series Analysis

Logistic Regression

Neural Network



# Data Sources

- TerraClimate
- Spatial wildfire occurrence data for the United States
- California Vegetation - WHR13 Types



Region of study, California, USA.

# Data Description

- Actual Evapotranspiration
- Climate Water Deficit
- Potential Evapotranspiration
- Precipitation
- Runoff
- Soil Moisture
- Downward Surface Shortwave Radiation
- Maximum Temperature
- Minimum Temperature
- Vapor Pressure
- Wind Speed
- Vapor Pressure Deficit
- Palmer Drought Severity Index

**Difference**

**Original  
observation**

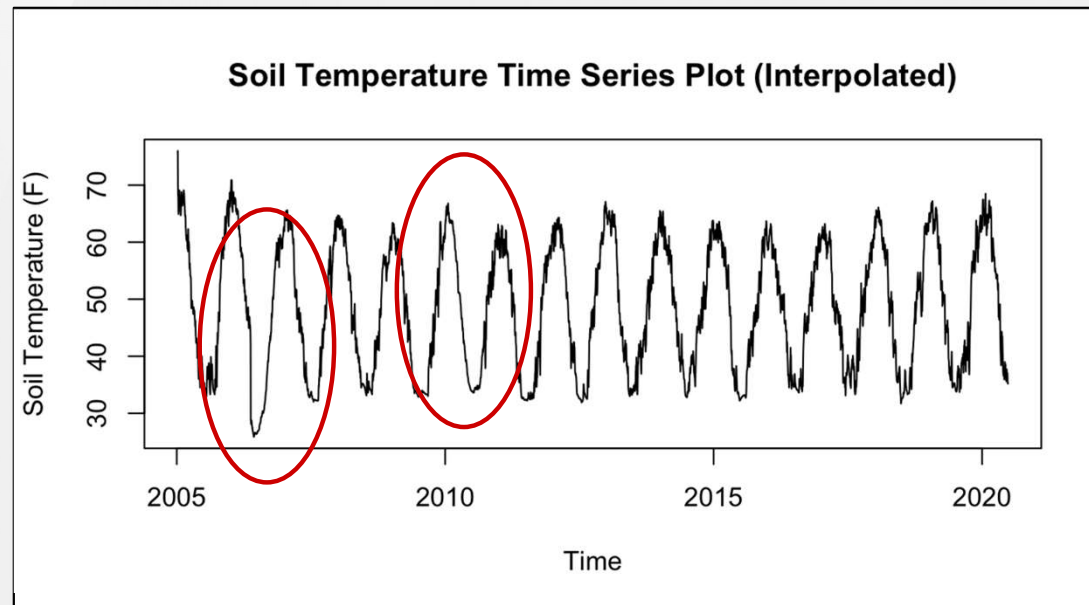
**Anomaly**

**Anomaly  
Lags**



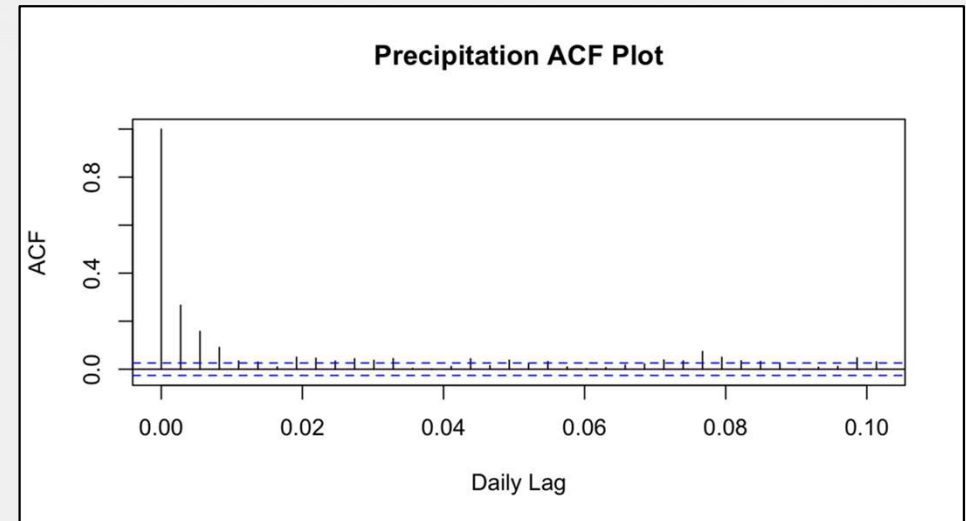
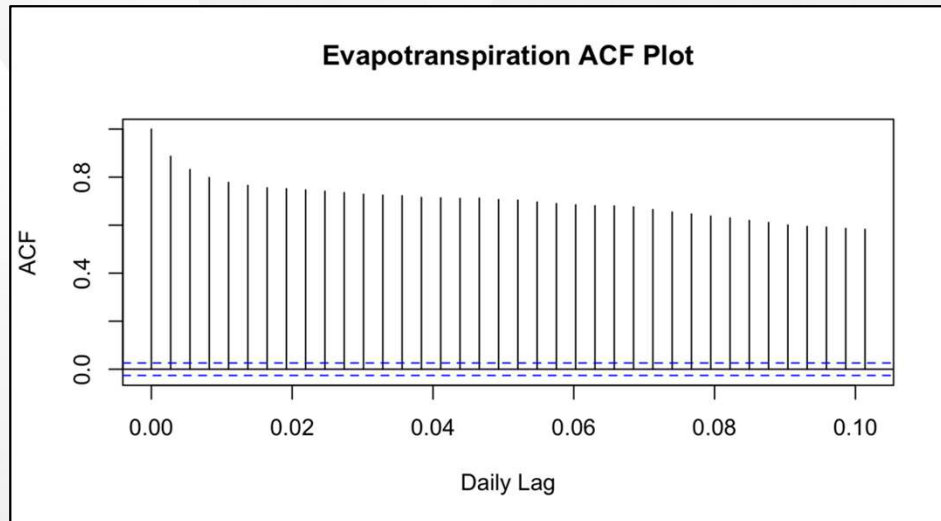
# Time Series Analysis

- California Irrigation Management Information System (CIMIS) data

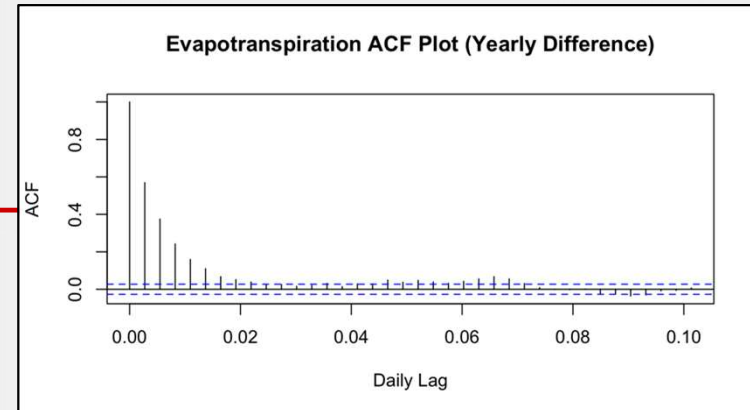
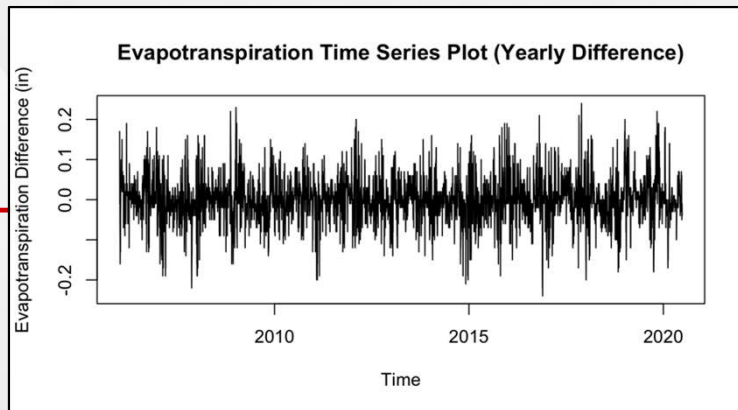
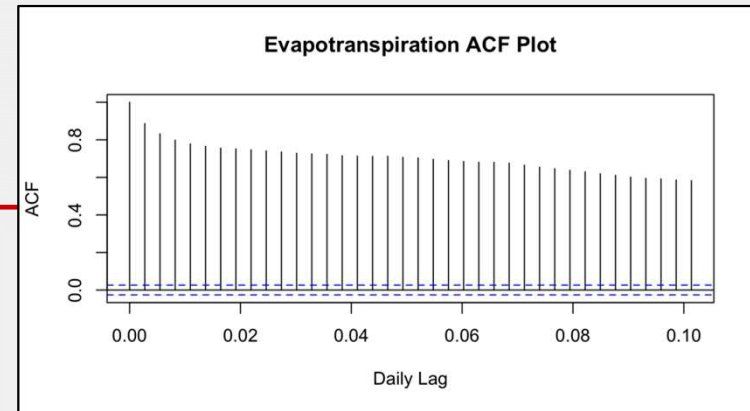
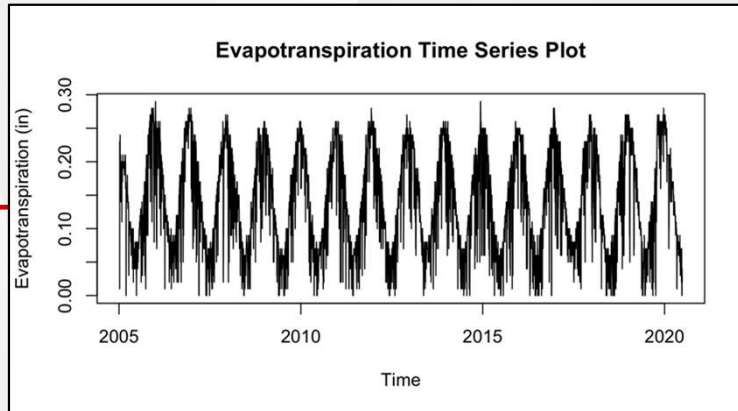


# Time Series Analysis Cont.

- Checking for stationarity

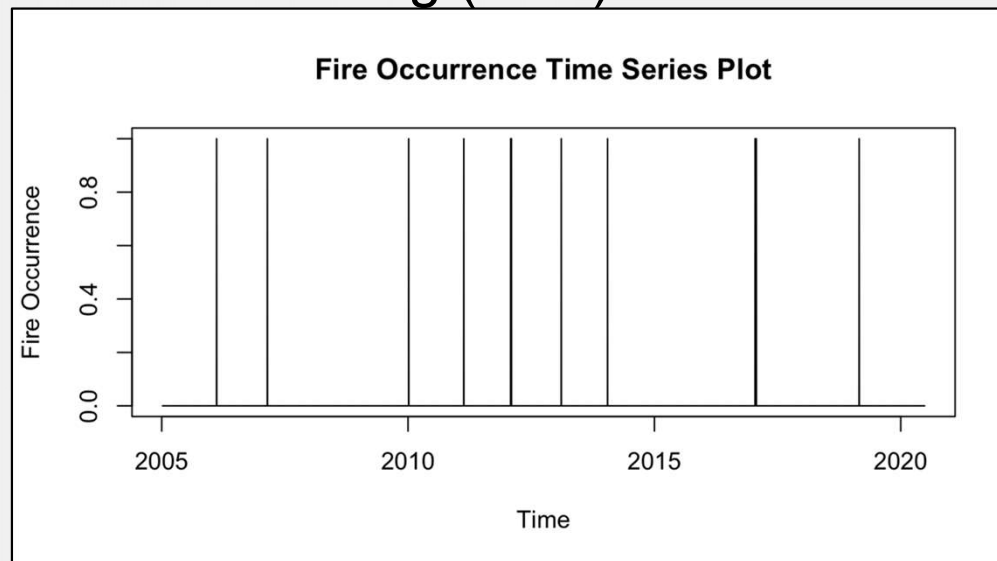


# Time Series Analysis Cont.

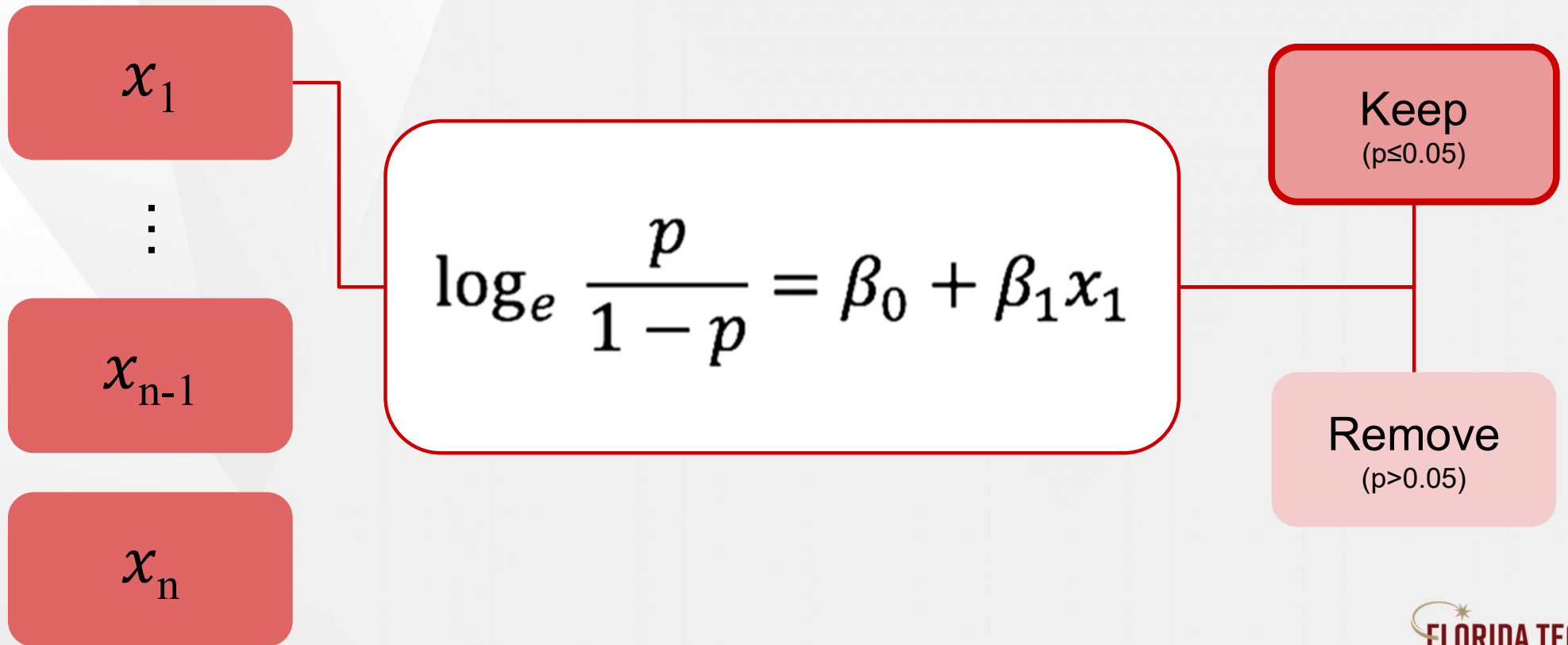


# Time Series Analysis Cont.

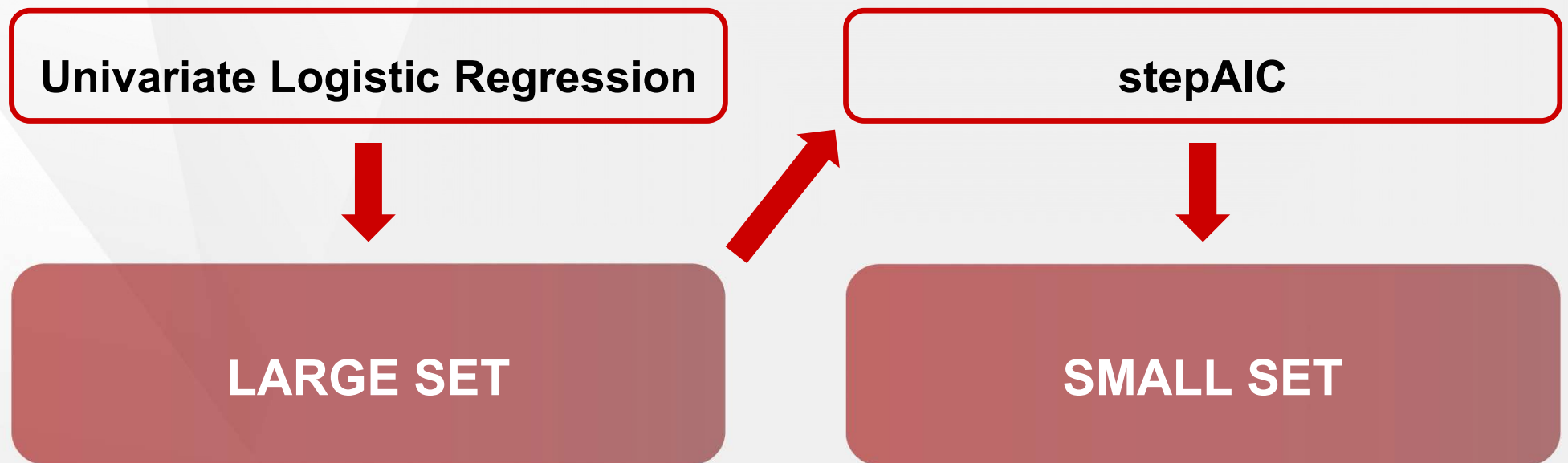
- Autoregressive (AR) model
- Autoregressive Distributed Lag (ADL) model



# Feature Reduction



# Feature Reduction Cont.



# Feature Reduction Cont.

**LARGE SET**



**Subset**



**Train-Test Split**

**SMALL SET**



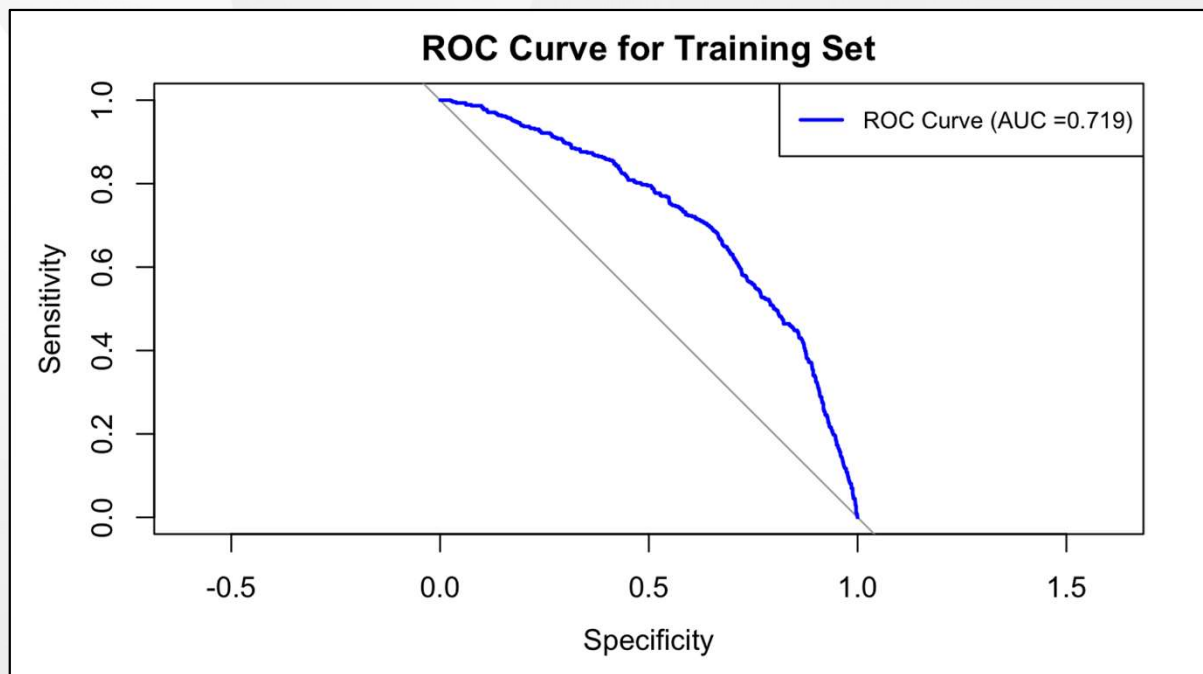
**Subset**



**Train-Test Split**

# Logistic Regression

## Large Variable Set



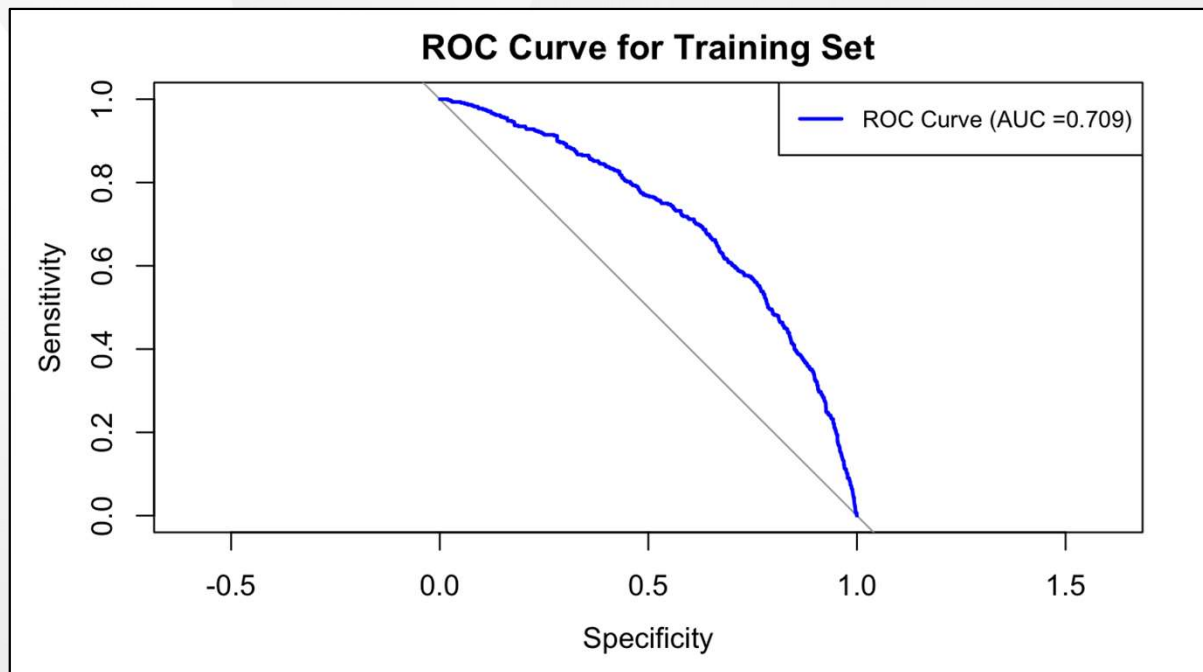
Accuracy	<b>0.633</b>
Sensitivity	<b>0.612</b>
Specificity	<b>0.653</b>
AUC	<b>0.719</b>

Optimal Cutoff: 0.214



# Logistic Regression Cont.

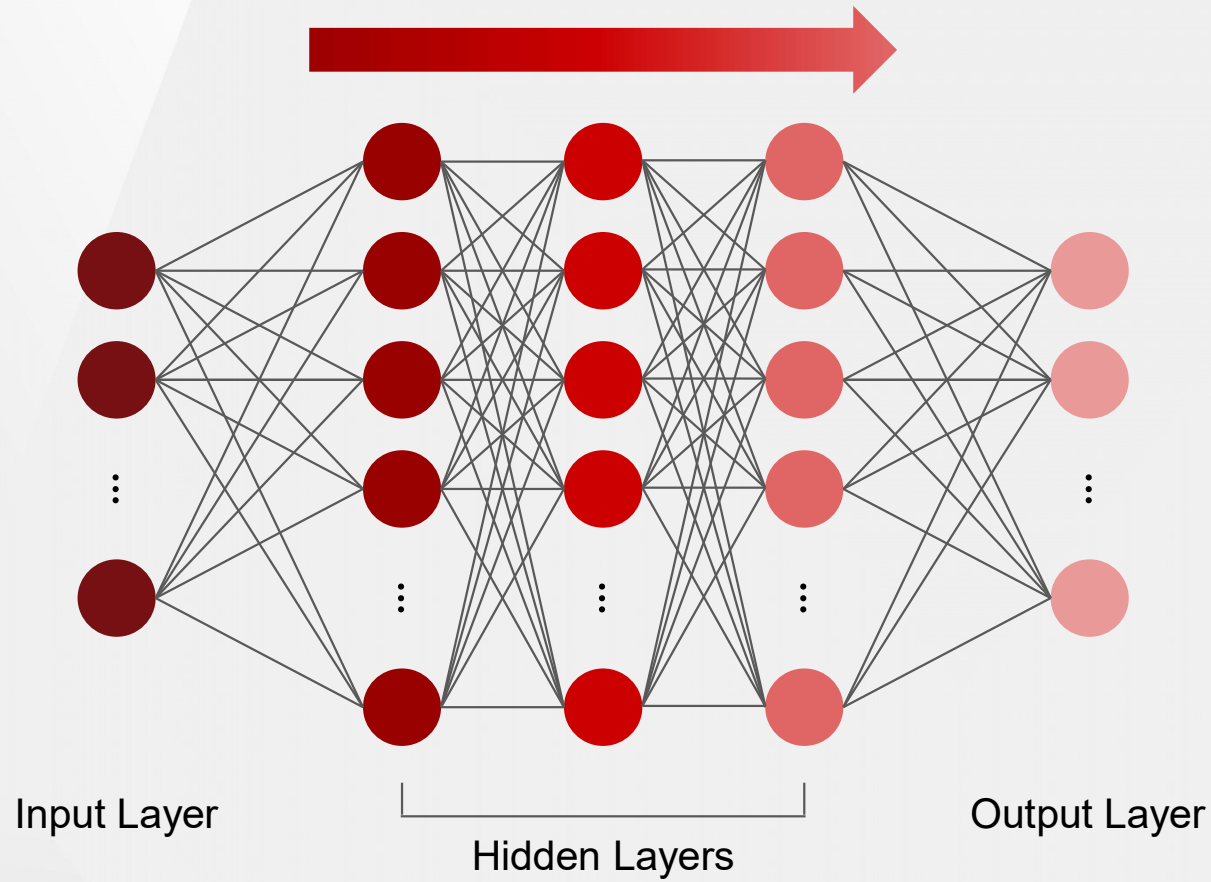
## Small Variable Set



Accuracy	<b>0.643</b>
Sensitivity	<b>0.674</b>
Specificity	<b>0.612</b>
AUC	<b>0.709</b>

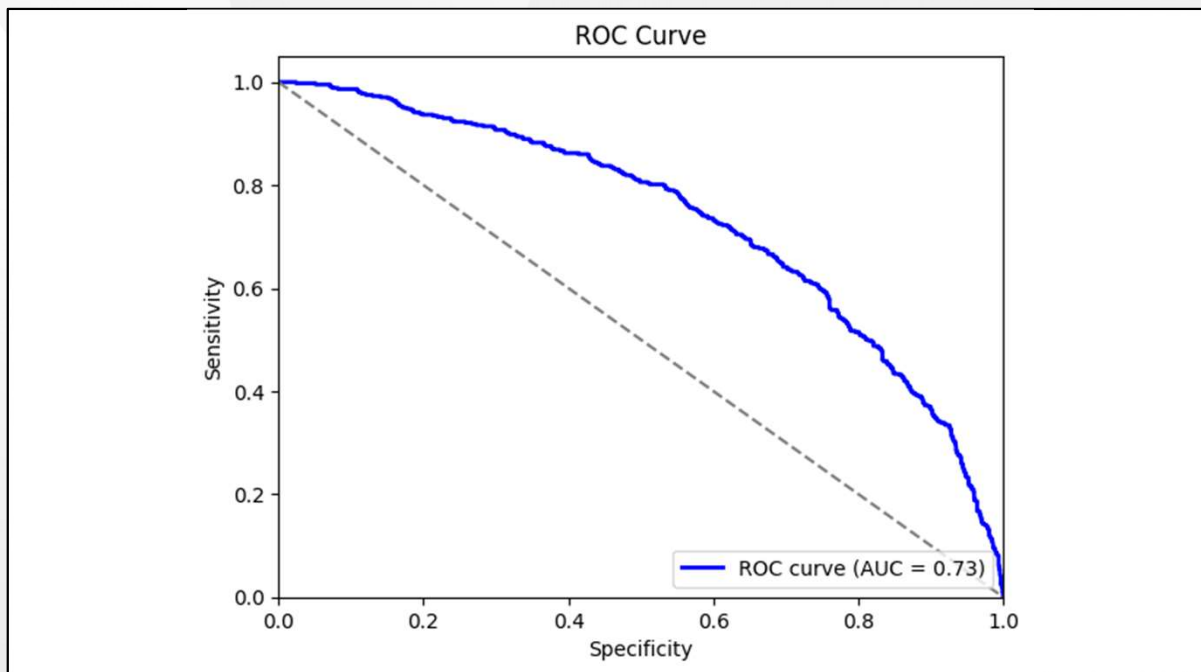
Optimal Cutoff: 0.203

# Feed-Forward Neural Network



# Feed-Forward Neural Network Cont.

## Large Variable Set

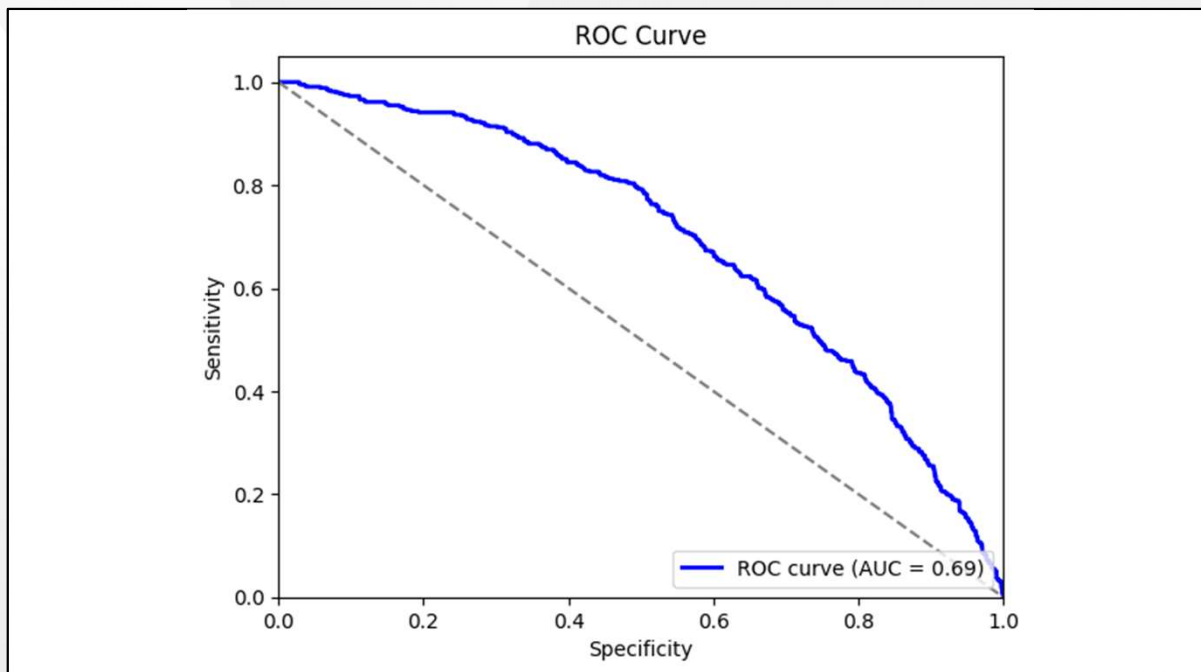


Accuracy	<b>0.740</b>
Sensitivity	<b>0.640</b>
Specificity	<b>0.840</b>
AUC	<b>0.733</b>

Optimal Cutoff: 0.398

# Feed-Forward Neural Network Cont.

## Small Variable Set



Accuracy	<b>0.680</b>
Sensitivity	<b>0.800</b>
Specificity	<b>0.560</b>
AUC	<b>0.691</b>

Optimal Cutoff: 0.335

# Results

	Large Logistic	Small Logistic	Large NN	Small NN
Accuracy	0.633	0.643	<b>0.740</b>	0.680
Sensitivity	0.612	0.674	0.640	<b>0.800</b>
Specificity	0.653	0.612	<b>0.840</b>	0.560
AUC	0.719	0.709	<b>0.733</b>	0.691

# Discussion



Strong winds. (Delbert, 2022)



Former wetland near Tulelake, California. (NPR, 2022)



# Conclusion

- *Summary*
  - Neural networks outperformed logistic regression models
  - Subset approach disregards time series element
- *Future work*
  - Other neural networks
  - New variable combinations and selection methods
  - Need for complementary ignition model



Fighting a wildfire. (WHO, 2024)

**Thank you!**

**Questions?**



# References

- Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015 [Dataset]. Climatology Lab. <https://www.climatologylab.org/terraclimate.html>
- Esri. (2024). California Multi-Source Vegetation Layer [Data set]. ArcGIS. Retrieved May 20, 2024, from <https://www.arcgis.com/home/item.html?id=b7ec5d68d8114b1fb2bfbf4665989eb3>
- Halofsky, J. E., Peterson, D. L., & Harvey, B. J. (2020). Changing wildfire, changing forests: The effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA. *Fire Ecology*, 16(4). <https://doi.org/10.1186/s42408-019-0062-8>
- Office of Environmental Health Hazard Assessment [OEHHA]. (2022). Wildfires. In *Indicators of Climate Change in California*. California Environmental Protection Agency. <https://oehha.ca.gov/media/downloads/climate-change/document/04wildfires.pdf>
- 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>
- Pham, K., Ward, D., Rubio, S., Shin, D., Zlotikman, L., Ramirez, S., ... & Jiang, X. (2022, December). California wildfire prediction using machine learning. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 525-530). IEEE.
- Satir, O., Berberoglu, S., & Donmez, C. (2016). Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomatics, Natural Hazards and Risk*, 7(5), 1645-1658. <https://doi.org/10.1080/19475705.2015.1084541>
- Short, K. C. (2022). Spatial wildfire occurrence data for the United States, 1992-2020 (FPA\_FOD\_20221014) (6th ed.). Fort Collins, CO: Forest Service Research Data Archive. <https://doi.org/10.2737/RDS-2013-0009.6>

:)





### Data

- 12 months for 29 years; 348 months; Jan 1992-Dec 2020
- 217,500 total rows of observations (348 x 625)
- 25x25 region = 625 cells (16 km<sup>2</sup>)
  - region with habitat variability; has enough fires to be relevant to study
  - near LA and important parks



aet: (Actual Evapotranspiration, monthly total), units = mm  
 def: (Climate Water Deficit, monthly total), units = mm  
 pet: (Potential evapotranspiration, monthly total), units = mm  
 ppt: (Precipitation, monthly total), units = mm  
 q: (Runoff, monthly total), units = mm  
 soil  
 soil: (Soil Moisture, total column - at end of month), units = mm  
 srad: (Downward surface shortwave radiation), units = W/m<sup>2</sup>  
 tmax: (Max Temperature, average for month), units = C  
 tmin: (Min Temperature, average for month), units = C  
 vap: (Vapor pressure, average for month), units = kPa  
 ws: (Wind speed, average for month), units = m/s  
 vpd: (Vapor Pressure Deficit, average for month), units = kpa  
 PDSI: (Palmer Drought Severity Index, at end of month), units = unitless neg=dry  
 fires: total number of fires, count  
 fire\_total: sum of fire area in given month and given cell  
 habitat: specific habitat classification (most common specific habitat in that cell)  
 habitat\_g: general habitat classification (most common general habitat in that cell)  
 cell: concatenated lon0 and lat0 with comma separator  
 fire\_events: 1 when at least one natural fire; 0 when no natural fire  
 fire\_spread: 1 when fire area >0.1 acres; 0 when smaller or none

water leaving soil  
 pet minus aet  
 water could transpire

not absorbed by

water in soil  
 sunlight

humidity

dryness



# Logistic including lag anomalies

-90-10 train-test split  
-all variables; reduce based on p-val >0.05  
-remove:

```
#"q"  
#"diff_aet"  
#"diff_ppt"  
#"diff_q"  
#"diff_soil"  
#"diff_PDSI"  
#"anom_q"  
#"anom_srad"  
#"lag_anom_srad"
```

-reduce with stepAIC to:

#Step: AIC=2028.84

```
aet + def + ppt + srad + vap + ws + vpd +  
PDSI + diff_pet + diff_tmin + diff_vap +  
diff_vpd + anom_aet + anom_def + anom_pet  
+ anom_ppt + anom_vap + anom_ws +  
anom_vpd + anom_PDSI + lag_anom_def +  
lag_anom_pet + lag_anom_vap
```

-optcutoff on train data;

confusion/accuracy/pred for test data

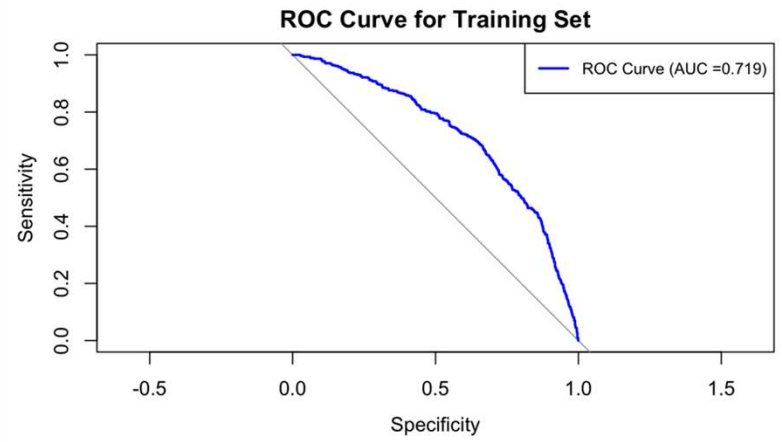
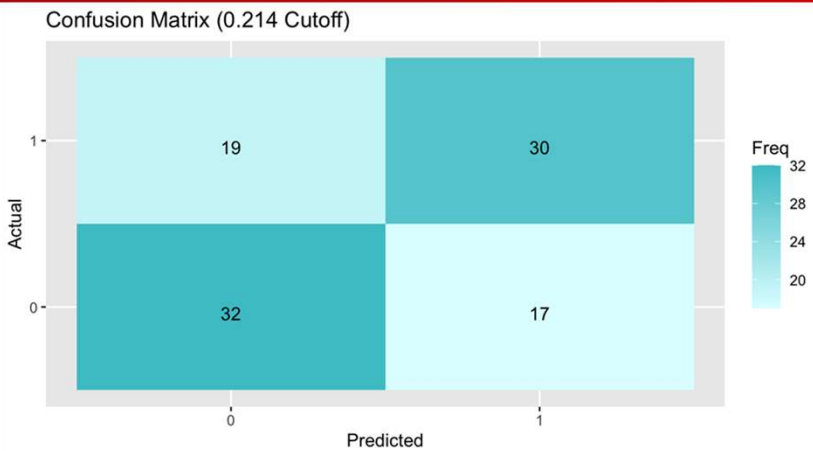
-STEP model has higher accuracy, sensitivity  
-FULL model has higher specificity, AUC

# Logistic including lag anomalies (LARGE)

-90-10 train-test split  
-all variables; reduce based on p-val >0.05  
-remove:

- #"q"
- #"diff\_aet"
- #"diff\_ppt"
- #"diff\_q"
- #"diff\_soil"
- #"diff\_PDSI"
- #"anom\_q"
- #"anom\_srad"
- #"lag\_anom\_srad"

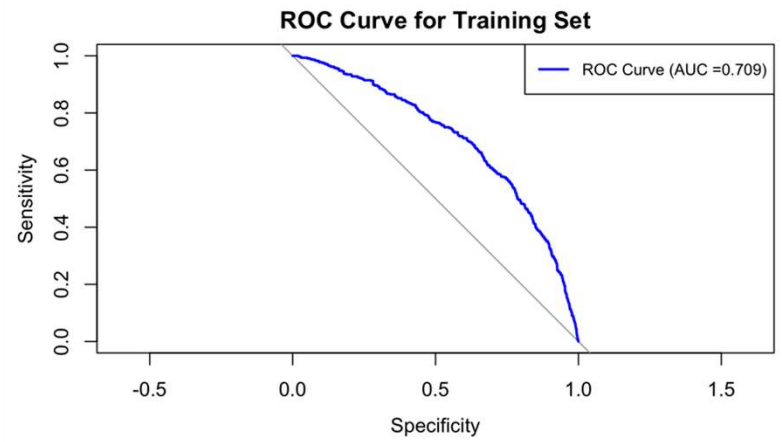
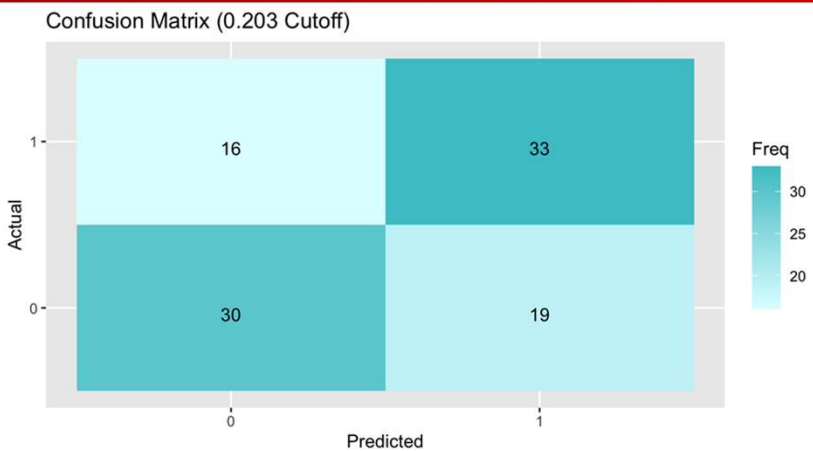
-full model results:  
-using optcutoff (YJS): 0.2136121  
#Accuracy : **0.6327**  
#Sensitivity : 0.6122  
#Specificity : 0.6531  
#AUC : 0.719



# Logistic including lag anomalies (SMALL)

-90-10 train-test split  
-all variables; reduce based on p-val >0.05  
-remove:  
  #"q"  
  #"diff\_aet"  
  #"diff\_ppt"  
  #"diff\_q"  
  #"diff\_soil"  
  #"diff\_PDSI"  
  #"anom\_q"  
  #"anom\_srad"  
  #"lag\_anom\_srad"  
-reduce with stepAIC to:  
#Step: AIC=2028.84  
aet + def + ppt + srad + vap + ws + vpd +  
PDSI + diff\_pet + diff\_tmin + diff\_vap +  
diff\_vpd + anom\_aet + anom\_def + anom\_pet  
+ anom\_ppt + anom\_vap + anom\_ws +  
anom\_vpd + anom\_PDSI + lag\_anom\_def +  
lag\_anom\_pet + lag\_anom\_vap  
-optcutoff on train data;  
confusion/accuracy/pred for test data

-step model results:  
-using optcutoff (YJS): 0.2031919  
#Accuracy : 0.6429  
#Sensitivity : 0.6735  
#Specificity : 0.6122  
#AUC : 0.709



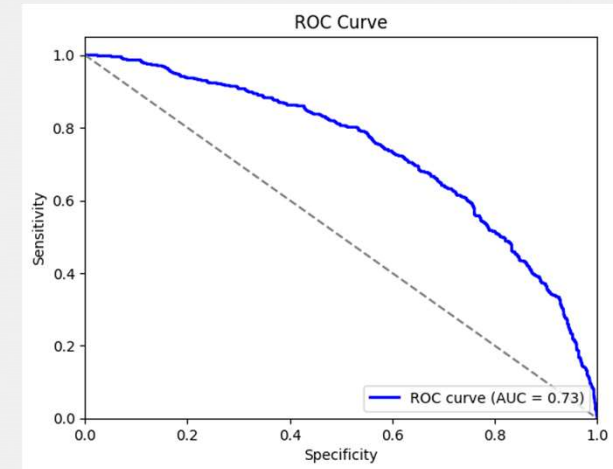
# NN including lag anomalies (LARGE)

-90-10 train-test split  
-all variables; reduce based on p-val >0.05  
-remove:

```
#"q"  
#"diff_aet"  
#"diff_ppt"  
#"diff_q"  
#"diff_soil"  
#"diff_PDSI"  
#"anom_q"  
#"anom_srad"  
#"lag_anom_srad"
```

-activation: leaky\_relu  
-5 layers: 80, 60, 40, 20, 1  
-binary focal cross-entropy (alpha=0.1)  
-optimizer: adam  
-normalized  
-30 epochs, batch size 10

-full model results:  
-using optcutoff (YJS): 0.39797372  
#Accuracy : 0.740  
#Sensitivity : 0.640  
#Specificity : 0.840  
#AUC : 0.733





# NN including lag anomalies (SMALL)

-90-10 train-test split  
-all variables; reduce based on p-val >0.05  
-remove:  
  #"q"  
  #"diff\_aet"  
  #"diff\_ppt"  
  #"diff\_q"  
  #"diff\_soil"  
  #"diff\_PDSI"  
  #"anom\_q"  
  #"anom\_srad"  
  #"lag\_anom\_srad"  
-ALSO reduce with stepAIC to:  
#Step: AIC=2028.84  
aet + def + ppt + srad + vap + ws + vpd +  
PDSI + diff\_pet + diff\_tmin + diff\_vap +  
diff\_vpd + anom\_aet + anom\_def + anom\_pet  
+ anom\_ppt + anom\_vap + anom\_ws +  
anom\_vpd + anom\_PDSI + lag\_anom\_def +  
lag\_anom\_pet + lag\_anom\_vap  
-optcutoff on train data;  
confusion/accuracy/pred for test data

-full model results:  
-using optcutoff (YJS): 0.33526808  
#Accuracy : 0.680  
#Sensitivity : 0.800  
#Specificity : 0.560  
#AUC : 0.691  
  
-activation: leaky\_relu  
-5 layers: 80, 60, 40, 20, 1  
-binary focal cross-entropy (alpha=0.1)  
-optimizer: adam  
-normalized  
-30 epochs, batch size 10

