California Wildfires Prediction
W207 Final Project

Final Presentation
August 4, 2022

Joe Ritter, Mon Young, Prakash Krishnan

Motivation

predict wildfire size of burned area in California

Persistence

- Climate change is causing wildfires to be longer, frequent and more devastating–a trend likely to continue

Social Utility

- Significant societal and economical impacts

Edification

- Wildfires are a complex phenomenon due to spatial, temporal and non-linear relationships of local meteorology, land-surface characteristics, socio-economic factors and long-term climate patterns
- Predicting wildfires is extremely challenging due to numerous complex relationships

Executive Summary

- Literature review to understand performance of external models and develop domain knowledge

- Leveraged several techniques of ML in the project (Spatial join, PCA, Time series modelling, DNN, Sub-group analysis and Automated Hyper Param Optimization)

- Built baseline shallow models (Linear Regression) to assess baseline metrics
  - Predict burnt area size
  - Predict burnt area class

- Tuned DNN Model, Random Forest Regression, Gradient Boost Regression to improve performance over baseline and external benchmarks

- Developed a “Stretch Model (Prototype)” seeking to predict the rate of change in a burned area

Existing Wildfire ML Models

Item	Paper	Model	Features	Prediction
1.	Data-Driven Wildfire Risk Prediction in Northern California (atmosphere 2021)	Random Forest - 92% Adabost - 91.5% Gradient Boosting Trees - 90.5%	Weather, Terrain, Powerline and Vegetation	Fire / No Fire
2.	Identifying Key Drivers of Wildfires in the contiguous US using Machine Learning and Gaming Theory (Earth's Future, May 2021)	eXtreme Gradient Boosting Model - RMSE 2.04 km squared	Local Meteorology, Large Scale Meteorological Patterns, Land Surface Properties and Socio-Economic	Size of Burnt Area
3.	Wildfire Prediction Through Live Fuel Moisture Content Maps (Civil and Environmental Engineering, Stanford University)	SVM - 65.86% Random Forest - 71.95% CNN - 52.46%	Live Fuel Moisture Content (LFMC) Maps	Fire / No Fire

Our Approach

Literature Review and Understand the Domain

- Literature Review of 5 Research Papers
- **Spoke with lead scientist Dr. Sally Wang (expert in this field)**
- Understand what ML Models have worked

Develop Research Question

- Most models predict fire risk prediction (Binary Classification)
- Predicting wildfire burnt area is hard.
- **We picked predicting wildfire burnt area as the research question for a grid area**

Gather Data and Data Curation

- Spatial join of CA Counties with the USA Wildfire Dataset from the researcher

- **Log Transform of label**
- **Standardize the features**

EDA to understand label and feature distribution

- A variety of EDA techniques employed
- Geospatial Viz
- Geopandas
- Histograms
- Correlation Heatmap - Features and Labels
- Scatterplot - Features and Labels
- Time Series Trends

Build Baseline Weak Models to get a benchmark

- **Linear Regression and a Shallow NN Model**

- **External benchmark from Literature Review**

Identify Ml Models to Tune the Hyper-Params

- Leverage Literature Review and ML Theory to determine ML Models to Tune the Hyper-Params

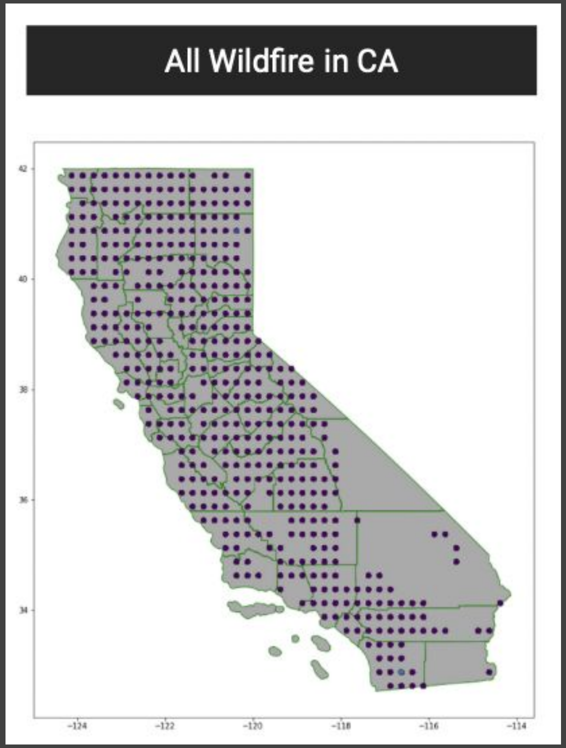
Experiment and Benchmark Compare

- Prepare Evaluation Metrics on Test Data
- **Aggregate and Sub-Groups**
- Determine Model Improvement over Baseline

Conclusion

- Model Takeaways
- Key learnings
- **Future work**

About the Dataset



fires_within_county = gpd.sjoin(geofires, ca, how='inner', op='within')

- Our dataset is based on the paper: *Identifying Key Drivers of Wildfires in the Contiguous US Using Machine Learning and Game Theory Interpretation* by Sally S.-C. Wang.
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8243942/
- The dataset is in the RDS format: downloaded from https://zenodo.org/record/4467161
  - Rows: 1,240,704
  - Cols: 44
- This dataset includes wildfires that happened between 2000-2017 in the United States.
- We use geopandas with the CA_Counties_TIGER2016.shp file and inner join it with our USA wildfires dataset to remove wildfire records outside of California.
  - Rows: 102K
  - Cols: 92

Features and EDA

Land-Surface Properties

Feature Variable	Feature Name	Unit
soilm	Monthly mean surface soil moisture	kg m-2
ET	Monthly mean evapotranspiration	kg m-2
NDVI	Normalized difference vegetation index	unitless
p_1	Water bodies	%
p_2	Grasslands	%
p_3	Shrublands	%
p_4	Broadleaf Croplands	%
p_5	Savannas	%
p_6	Evergreen Broadleaf Forests	%
p_7	Deciduous Broadleaf Forests	%
p_8	Evergreen Needleleaf Forests	%
p_9	Deciduous Needleleaf Forests	%
p.x	Nonvegetated Lands	%
p.y	Urban and Built-up Lands	%
elev	elevation	m
slope	slope	degree

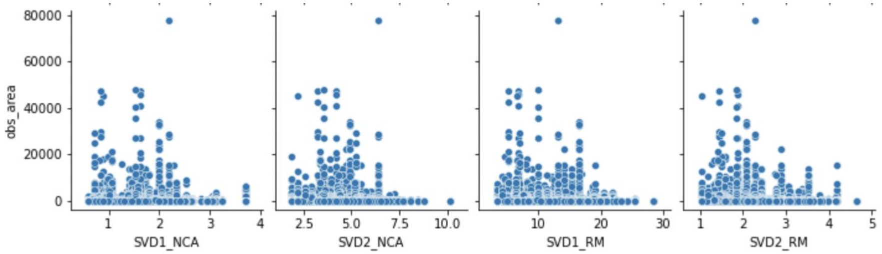
Large-Scale Meteorological Patterns

Feature Variable	Feature Name	Unit
SVD1_NCA	northern California	unitless
SVD2_NCA	northern California	unitless
SVD1_SE	Monthly standard deviation of daily SVD1 for southeastern US (with 2-month lag)	unitless
SVD2_SE	Monthly standard deviation of daily SVD2 for southeastern US (with 2-month lag)	unitless
SVD1_RM	Monthly standard deviation of daily SVD1 for southern Rocky Mountain	unitless
SVD2_RM	Monthly standard deviation of daily SVD2 for southern Rocky Mountain	unitless

incorporated features of local meteorology, land-surface characteristics, and socioeconomic variables to predict wildfire burned area size in California

- P_1 ~ P_7 = land type
- High Monthly Mean Evapotranspiration (ET) and Low Deciduous Broadleaf Forest (P_7) seem to have an effect on wildfires

- Long term patterns in Northern California and Rocky Mountains seem to have an effect of the size of wildfires as evident from the scatter plots



Features and EDA

Local Meteorology

Feature Variable	Feature Name	Unit
apcp	monthly mean of daily precipitation	kg m-2
temp	monthly mean surface temperature	K
rhum	monthly mean relative humidity	%
uwnd	Monthly mean zonal component of wind speed	m/s
vwnd	speed	m/s
ERC	Monthly mean energy release component	
FM1000	Monthly mean 1000-hour dead fuel moisture	%
VPD	Monthly mean vapor pressure deficit	kPa
PDSI	Monthly mean Palmer Drought Severity Index	

- Some features have a high correlation (ex. ERC & FM1000)
- Low Monthly Mean Daily Precipitation, High Monthly Mean Surface Temperature, Low Monthly Mean 1000-Hour Dead Fuel Moisture and Low Monthly Mean Vapor Pressure Deficit have a effect on wildfires

Socio-Economic

Feature Variable	Feature Name	Unit
Lon	Longitude of the grid	degree
Lat	Latitude of the grid	degree
pop2	Population density	population km-2
GDP	GDP per capita	Constance 2011 international US dollar
N_campsite	Number of campsites	

- GDP and Population do not seem to have a clear relationship to burnt area
- One Hot Encoding for Counties
- Large wildfires are restricted to certain grid locations

Baseline Models

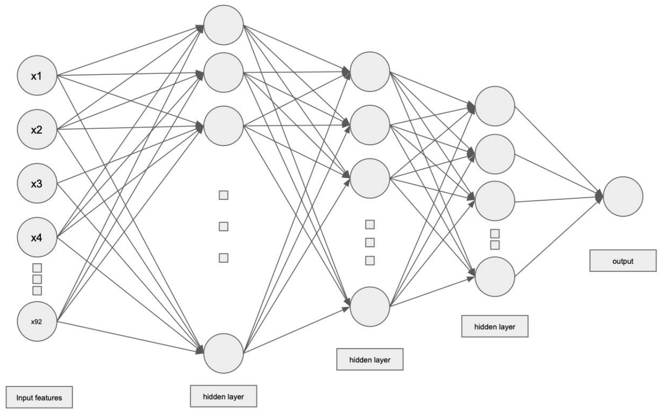
	Model	Features	MSE_Train	MSE_Test	R2_Train	R2_Test
0	Model-1 Baseline Scikit Linear Reg	All Features (Scaled)	7.468	7.479	0.423	0.426
1	Model-2 Baseline Scikit Random Forest Reg	All Features (Scaled)	1.871	4.785	0.855	0.633
2	Model-3 Baseline Keras Linear Reg	All Features (Scaled)	7.646	7.662	0.409	0.412
3	Model-4 Baseline Scikit Decision Tree Linear Reg	All Features (Scaled)	7.507	7.634	0.420	0.415
4	Model-5 Baseline Scikit Gradient Boost Reg	All Features (Scaled)	6.534	6.710	0.495	0.485
5	Model_1 + PCA	8 Principal Comp	8.514	8.591	0.342	0.341
6	Model-2 + PCA	8 Principal Comp	3.477	7.396	0.731	0.433
7	Model-3 + PCA	8 Principal Comp	8.516	8.590	0.342	0.341
8	Model-4 + PCA	8 Principal Comp	8.516	8.590	0.342	0.341

Models for Hyper Parameter Tuning

Model Selection:

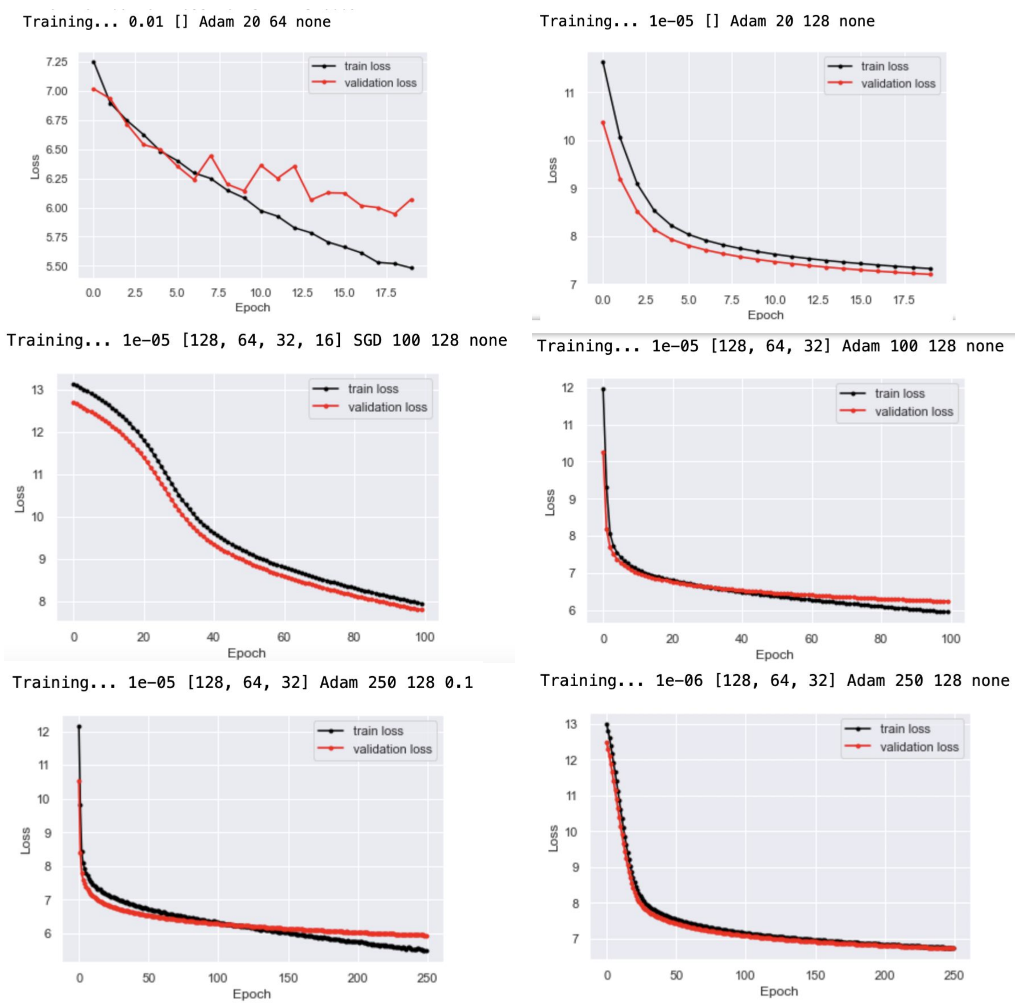
1. **Random Forests: Reduces overfitting, higher accuracy compared to other models, low variance due to multiple decision trees**
2. **Gradient Boosting Regression: Can handle non-linear relationships, multi-collinearity and higher accuracy than other models**
3. **DNN: Can model complex non-linear relationships with right architecture and parameter tuning**

FFNN: Feedforward Neural Network Models



hyperparameters used for tuning:

- learning rate = 0.1, 0.001. 0.0001, 0.00001, 0.000001
- optimization = SGD, Adam
- batch size = 32, 64, 128
- hidden layers = [], [128], [128, 64], [128, 64, 32], [128, 64, 32, 16]
- dropout layers = none, 0.5, 0.1, 0.2, 0.8
- epoch = 10, 100, 250, 500



Findings: In general,

- increase batch_size ⇒ a better loss plot's curve
- Adam optimizer has lower MSE and higher R^2 values than SGD
- smaller learning rate ⇒ a better loss plot's curve, but higher MSE and lower R^2 values
- more hidden layers ⇒ a better loss plot curve, but a higher MSE and lower R^2 values
- adding dropout layers does not help to make our models better

FFNN: Examine Highly Correlated Features

Check When Removing Features with High Collinearity

- The following shows removing FM1000 and rhum features does not make much different in our model

```
ERC      FM1000    -0.969334
FM1000   ERC       -0.969334
ERC      rhum      -0.900941
rhum     ERC       -0.900941
Lon      Lat       -0.795681
Lat      Lon       -0.795681
rhum     VPD       -0.783512
VPD      rhum      -0.783512
FM1000   VPD       -0.777328
VPD      FM1000    -0.777328
dtype: float64
```

removed FM1000, rhun

#PARAMETERS	TRAIN LOSS	VAL LOSS	LOSS DIFF	R2
66561	6.465789	6.536608	-0.070818	0.499345

keep all features

#PARAMETERS	TRAIN LOSS	VAL LOSS	LOSS DIFF	R2	TEST LOSS	TEST R2
67073	6.398295	6.479384	-0.081089	0.5044	6.628367	0.491696

Feedforward Neural Network Model Summary

Model Summary
Train Data: Examples-81,561, Features-92
Test Data: Examples-20,391, Features-92

Hyper Parameter Tuning
learning rate, optimization, batch size, hidden layers, dropout layers and epoch. Manually tried different combinations. Details in JNB.

Best Parameters
learning rate = 0.000001, optimazor = Adam, batch size = 128, hidden layers = [128, 64, 32], dropout layers = none, epoch = 500

Model Evaluation
Continuous Variable Prediction: MSE, R-Square, Residual Plot

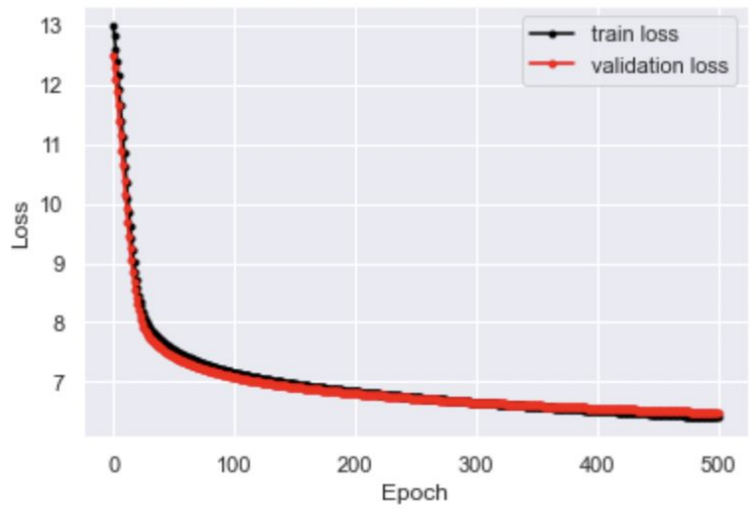
* National Wildfire Coordinating Group (NWCG) size class of fire classifications

https://www.nwcg.gov/term/glossary/size-class-of-fire

Continuous Variable Prediction

TRAIN LOSS	VAL LOSS	LOSS DIFF	R2	TEST LOSS	TEST R2
6.398295	6.479384	-0.081089	0.5044	6.628367	0.491696

```
Training... 1e-06 [128, 64, 32] Adam 500 128 none
```



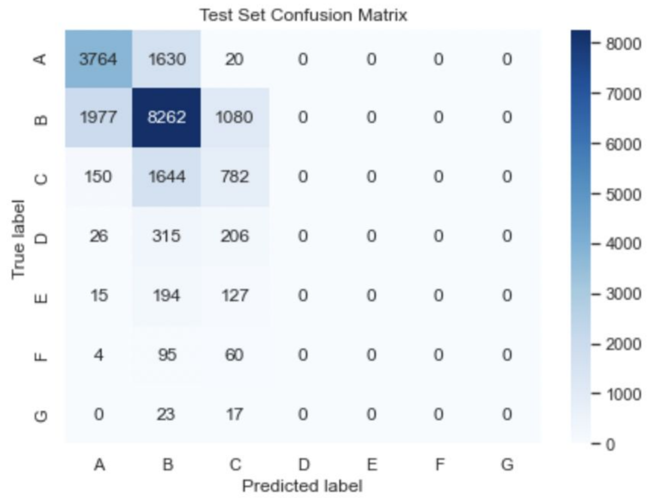

```
Final train loss: 6.3982954025268555
Final validation loss: 6.479383945465088
```

Size of Wildfires Classification

ID	CLASS	TRAIN ACC.	TEST ACC.
	All	0.627518	0.628120
0	A	0.693663	0.695235
1	B	0.729687	0.729923
2	C	0.313394	0.303571
3	D	0.000000	0.000000
4	E	0.000000	0.000000
5	F	0.000000	0.000000
6	G	0.000000	0.000000

in acres

A: < 0.25 acres
B: 0.25~ 10
C: 10 ~100
D: 100 ~ 300
E: 300 ~ 1,000
F: 1,000 ~ 5,000
G: > 5,000 acres



Random Forest Model Summary

Model Summary
Train Data: Examples-81561, Features-92
Test Data: Examples-20391, Features-92

Hyper Parameter Tuning
Random Forest Linear Regression with RandomizedSearchCV for Parameter Tuning
Iterations-40, CV-10

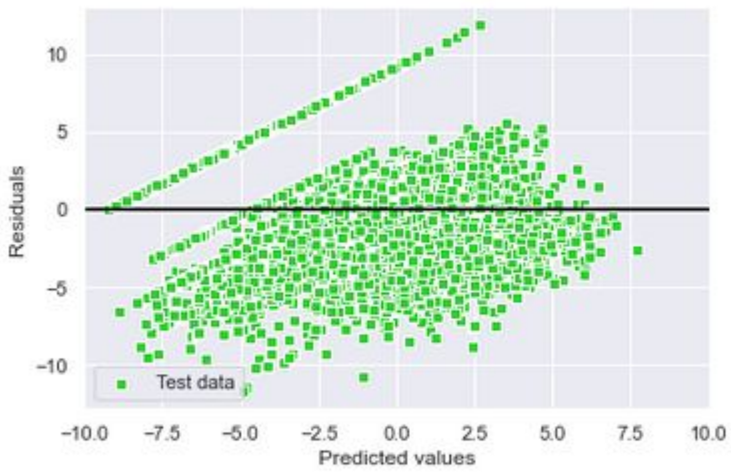
Best Parameters (after running 40 iterations)
{'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 20, 'bootstrap': True}

Model Evaluation
Continuous Variable Prediction: MSE, R-Square, Residual Plot, SHAP Analysis
- Additional Eval on Sub-Groups such as Counties and Regions

Classification Prediction: Accuracy, Confusion Matrix
- Additional Eva Fire Class Prediction

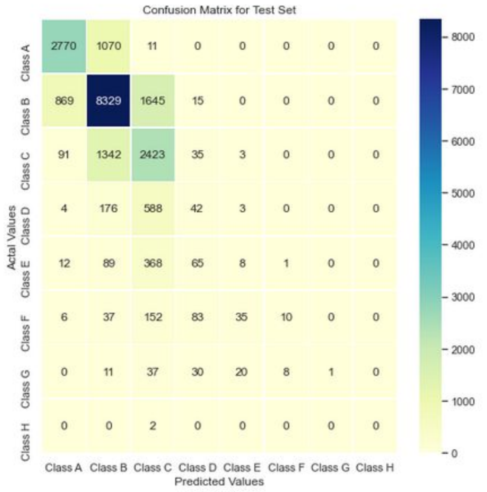
Continuous Variable Prediction

	Fire_Class	Train_MSE	Train_R2	Test_MSE	Test_R2
0	All Fire Class	1.830	0.859	4.849	0.628



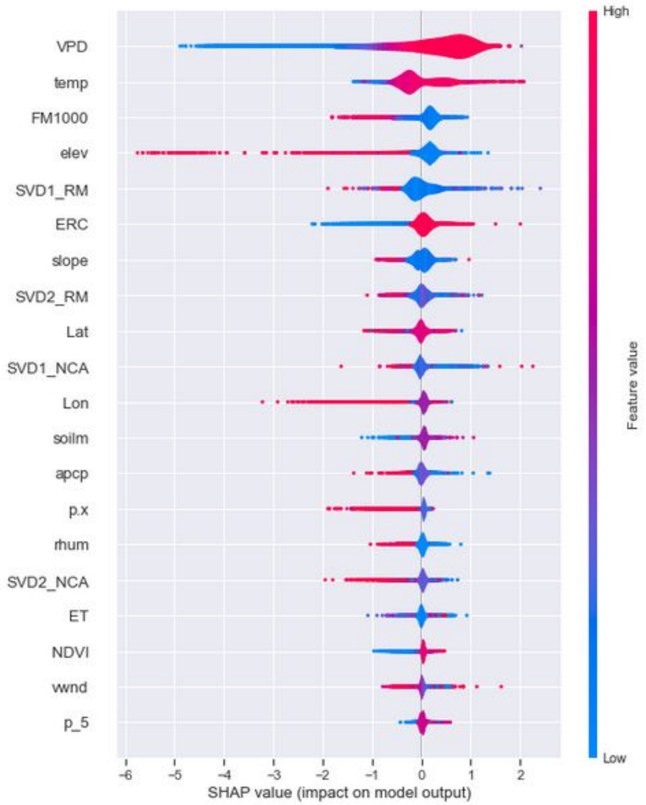
Classification Prediction

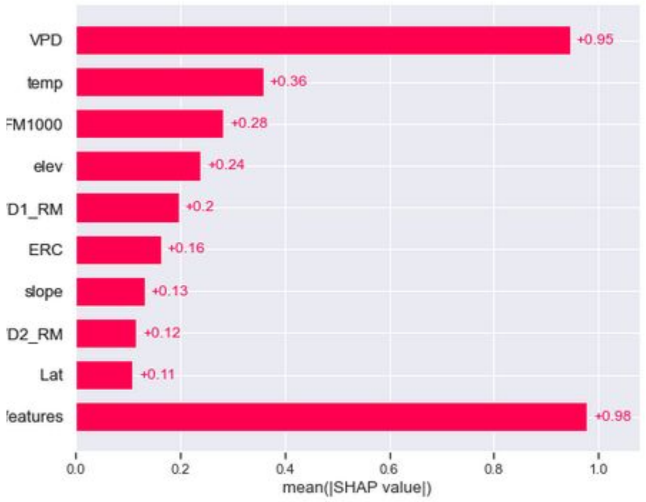
	Fire_Class	Train_Accuracy	Test_Accuracy
0	All Fire Class	0.756	0.666
1	A	0.777	0.777
2	B	0.871	0.871
3	C	0.725	0.725
4	D	0.114	0.114
5	E	0.072	0.072
6	F	0.084	0.084
7	G	0.029	0.029



Random Forest Model Summary Continued

SHAP Analysis





Top 9 Determinant Features Influencing Prediction:

VPD: Monthly Mean Vapor Pressure Deficit
temp: Monthly mean surface temperature
FM1000: Monthly mean 1000-hour dead fuel moisture
elev: elevation
SVDI_RM: Monthly std deviation of daily SVD1 for Rocky Mountains
ERC: Monthly mean energy release component
slope: slope
SVD2_RM: Monthly std deviation of daily SVD2 for Rocky Mountains
Lat: Latitude

Sub Group Evaluation

	Region	Example_Size	MSE_Test	R2_Test
1	Central Cal	8026	10.977	-0.019
0	Southern Cal	4816	11.483	-0.033
2	Northern Cal	7549	17.479	-0.089

	County	Example_Size	MSE_Test	R2_Test
23	county_Merced	329	4.898	0.017
50	county_Sutter	40	4.975	-0.017
43	county_Santa Cruz	44	5.480	-0.269
29	county_Orange	140	5.627	-0.030
19	county_Madera	472	5.710	0.009
49	county_Stanislaus	246	5.803	0.013
36	county_San Diego	649	5.834	-0.025
57	county_Yuba	38	6.081	0.012
8	county_El Dorado	302	6.423	-0.004
32	county_Riverside	757	6.608	-0.025
15	county_Kings	173	6.813	0.012
3	county_Butte	384	7.109	0.005
42	county_Santa Clara	249	7.113	-0.016
2	county_Amador	128	7.174	-0.009
18	county_Los Angeles	551	7.175	-0.051
21	county_Mariposa	249	7.882	-0.005
34	county_San Benito	180	7.888	-0.204
20	county_Marin	128	7.943	-0.090
30	county_Placer	252	8.017	-0.008

Gradient Boost Model Summary

Model Summary
Train Data: Examples-81561, Features-92
Test Data: Examples-20391, Features-92

Hyper Parameter Tuning
Gradient Boost Regression with RandomizedSearchCV for Parameter Tuning
Iterations-40, CV-5

Best Parameters (after running 40 iterations)
{'n_estimators': 100, 'max_depth': 9, 'learning_rate': 0.1}

Model Evaluation
Continuous Variable Prediction: MSE, R-Square, Residual Plot

Classification Prediction: Accuracy, Confusion Matrix

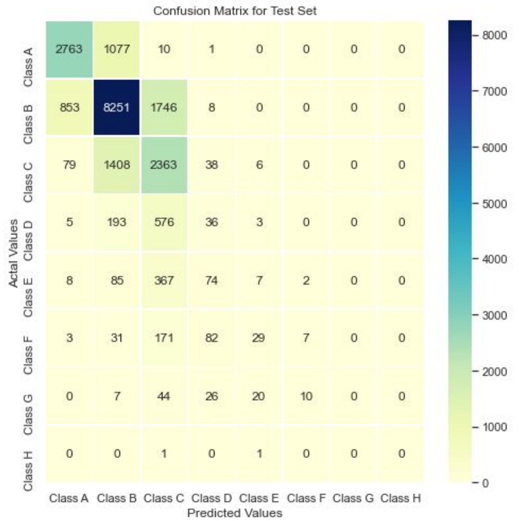
- Additional Eval Fire Class Prediction

Continuous Variable Prediction

	Model	Features	MSE_Train	MSE_Test	R2_Trair	R2_Test
1	Random Grid Search	All Features (Scaled)	2.734	4.717	0.789	0.638

Classification Prediction

	Fire_Class	Train_Accuracy	Test_Accuracy
0	All Fire Class	0.697	0.658
1	A	0.748	0.748
2	B	0.807	0.807
3	C	0.644	0.644
4	D	0.059	0.059
5	E	0.025	0.025
6	F	0.029	0.029
7	G	0.024	0.024

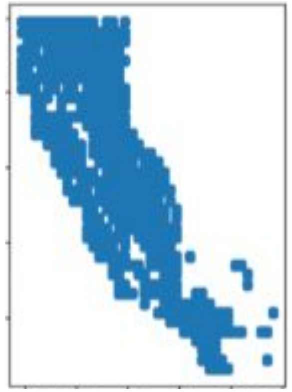


Experiment Summary: Baseline and Advanced

Item	Purpose	ML Model	Test Evaluation Metric	% Improve Over Baseline	Features and Labels	Hyper Parameters
1	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Baseline Scikit Learn Linear Regression	R Squared: 0.426 MSE: 7.479 Overall Accuracy: 0.582	Not Applicable	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	None
2	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Baseline Keras Shallow NN	R Squared: 0.412 MSE: 7.662 Overall Accuracy: 0.577	Not Applicable	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	None
3	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Tuned Scikit Learn Random Forest Regressor using RandomizedSearchCV	R Squared: 0.628 MSE: 4.849 Overall Accuracy: 0.666 Class Prediction Accuracy: (A: 0.777, B: 0.871, C: 0.725, D: 0.114, E: 0.072, F: 0.084, G: 0.029)	+ 47.4% - 35.1% + 14.4%	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	{'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 20, 'bootstrap': True}
4	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Tuned Scikit Learn GradientBoost Regressor using RandomizedSearchCV	R Squared: 0.638 MSE: 4.717 Overall Accuracy: 0.658	+ 49.7% - 40.5% + 13.0%	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	{'n_estimators': 100, 'max_depth': 9, 'learning_rate': 0.1}
5	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Feedforward Neural Network with manual hypermarameter selections	R Squared: 0.4917 MSE: 6.6283 Overall Accuracy: 0.6281	+ 15.4% -11.37% + 7.9%	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	learning rate = 0.000001, optimazor = Adam, batch size = 128, hidden layers = [128, 64, 32], dropout layers = none, epoch = 500

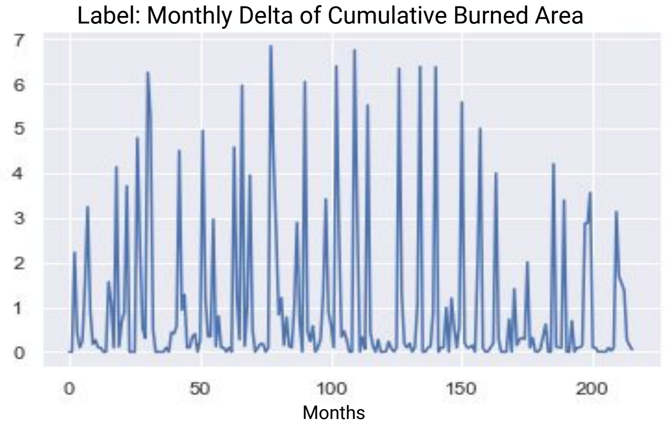
Time Series Models:
Different Question, Same Data, Different Structure

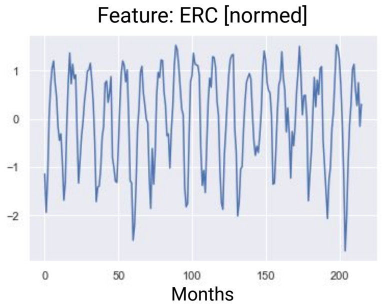
Does the dimension of time provide additional, useful information?
If so, how much and what frequency is most useful?
Requires different research question: pivot from area burned to the rate of change of area burned

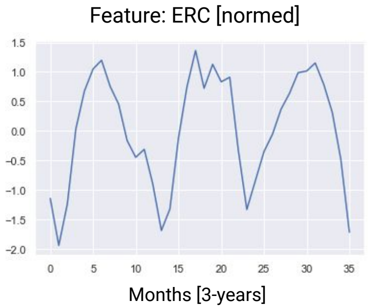


Data: same dataset, different approach

- 472 Lat./Lon. Grids.
- Date range, monthly frequency: 2000-2017 (216 months)
- **Features**:
  - 3 land-surface properties
  - All local and large-scale meteorological patterns
  - One-hot-encodings assigned for each grid
  - Features normalized (z-scored) within each grid (ex-OHE)
- **Label**: *month by month change* of log-transformed *cumulative* burned area, within each grid



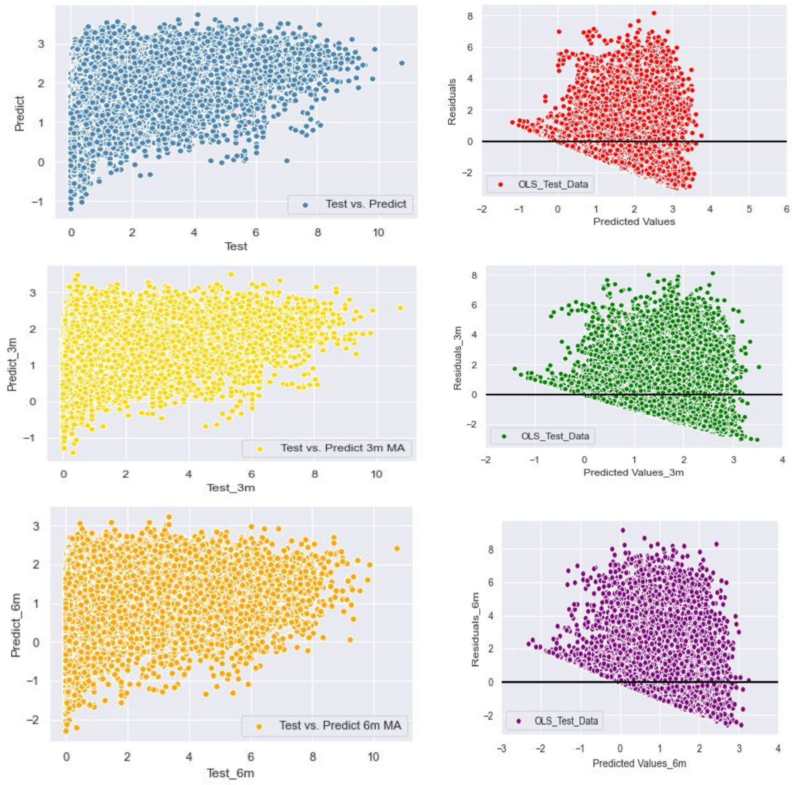




Time Series Models:
Closed Formed [OLS] & TF Baseline [shallow]

OLS

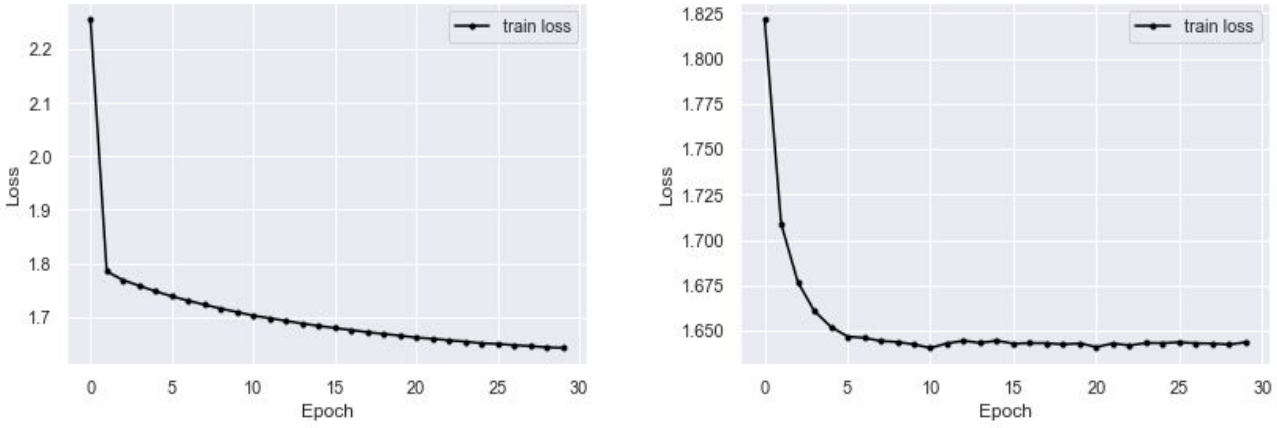
		R^2	MSE	MAE
0	Baseline	0.314	2.04	1.03
1	3-month MA	0.283	2.22	1.02
2	6-month MA	0.25	2.72	1.09



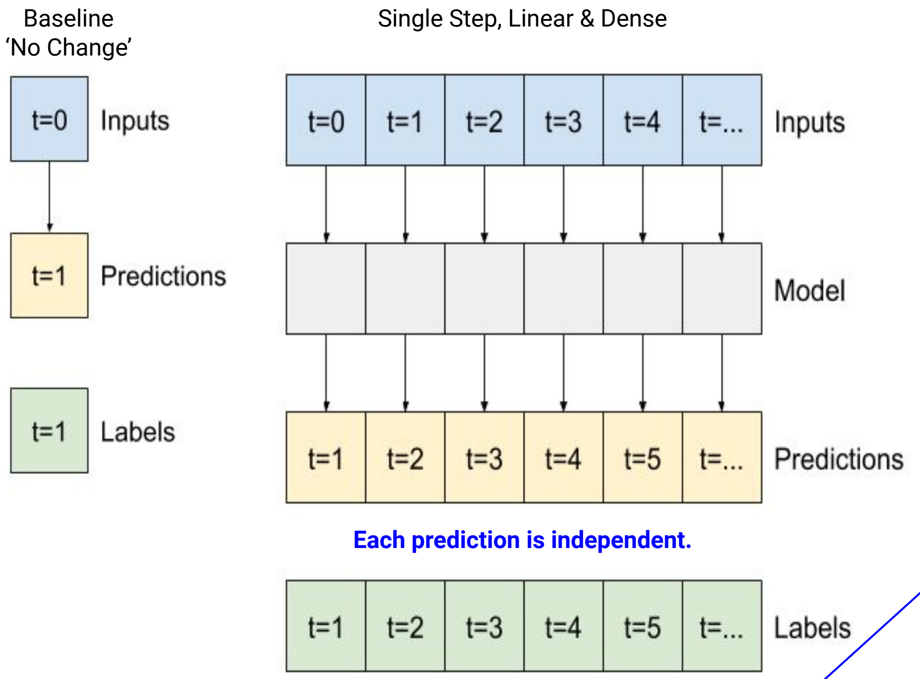
Shallow TF Model

Learning Rates	MSE	MAE
0.0001	2.01	0.98
0.001	2.05	1.03
0.01	2.18	1.03

SGD Optimizer
76,464 Parameters
30 Epochs

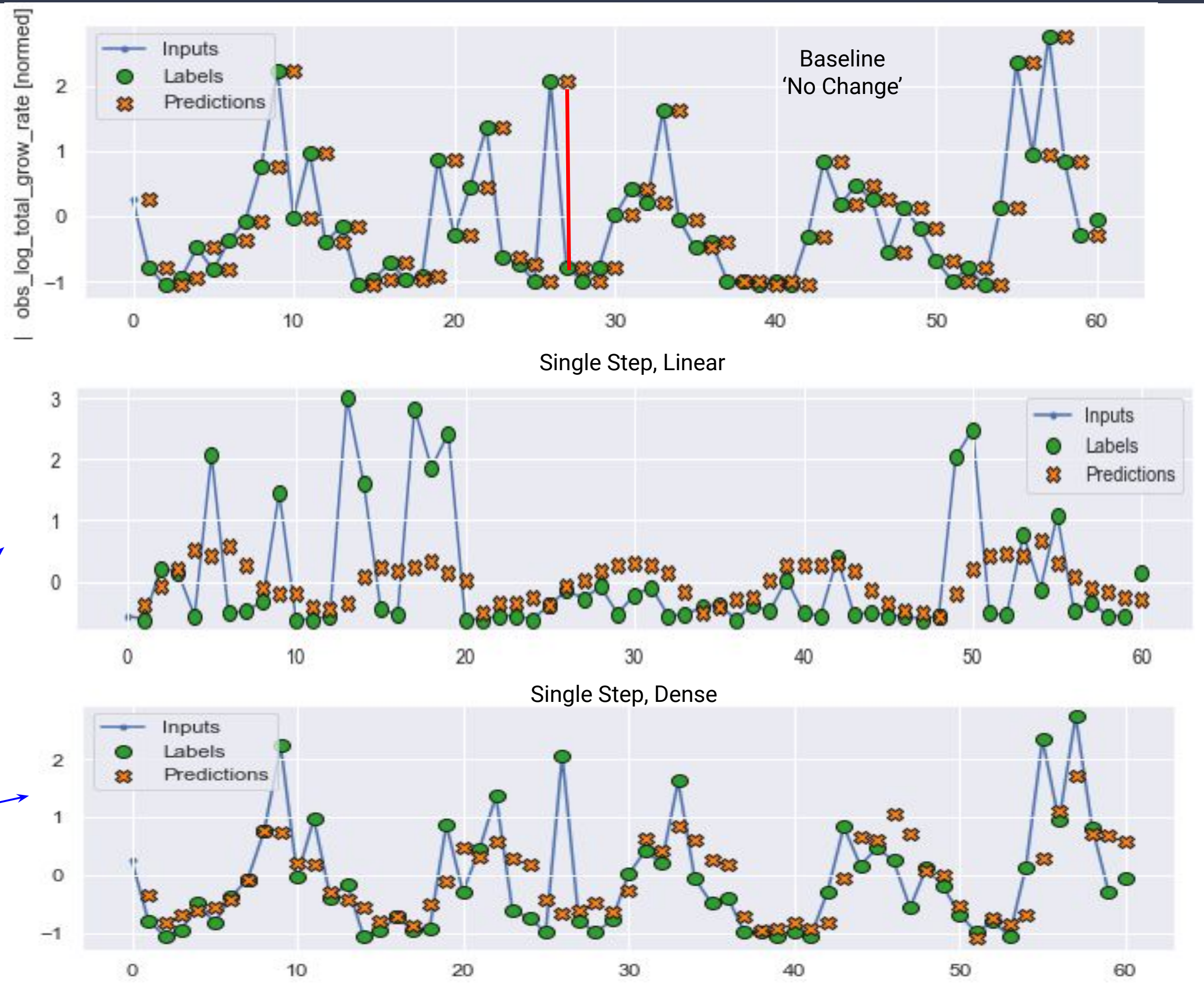


Time Series Models:
Various TF Single & Multi-Step Models




```
linear = tf.keras.Sequential([
    tf.keras.layers.Dense(units=1)
```

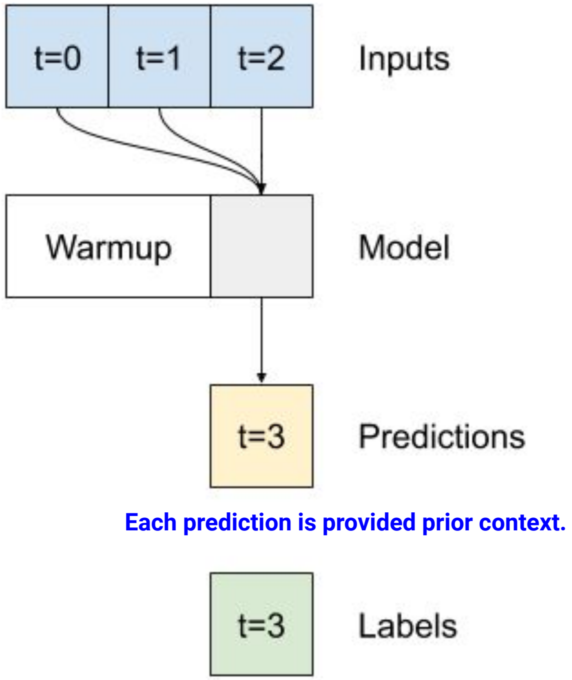
```
dense = tf.keras.Sequential([
    tf.keras.layers.Dense(units=64, activation='relu'),
    tf.keras.layers.Dense(units=64, activation='relu'),
    tf.keras.layers.Dense(units=1)
```



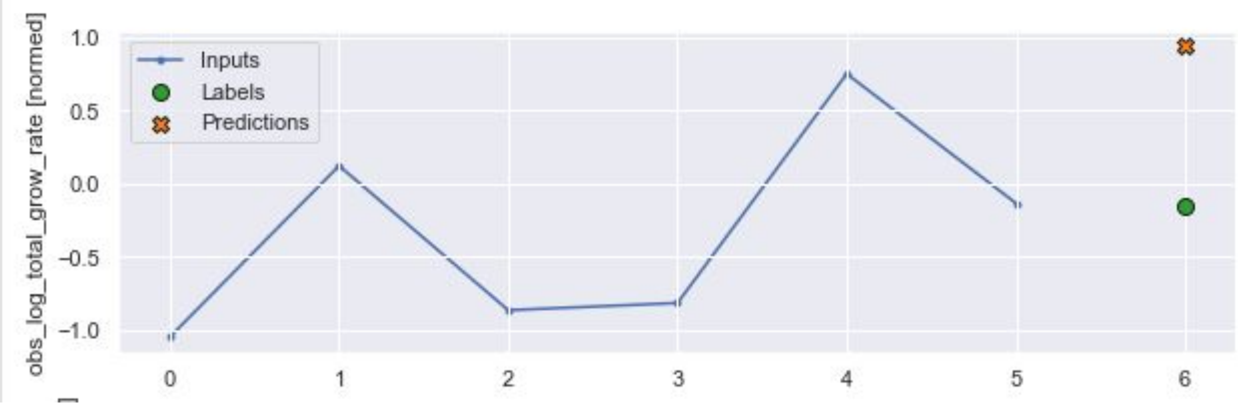
Model diagram from TensorFlow

Time Series Models:
Various TF Single & Multi-Step Models

Multi-Step, Dense



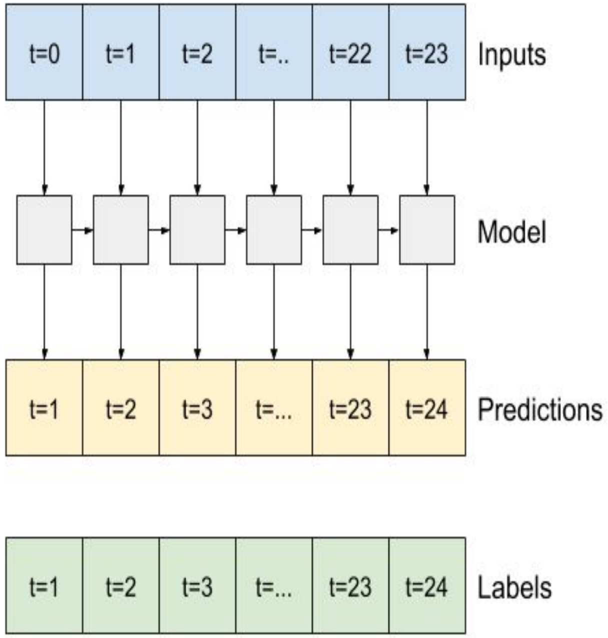
Multi-Step, Dense
t=6 prediction, requires inputs from t=0 … t=5

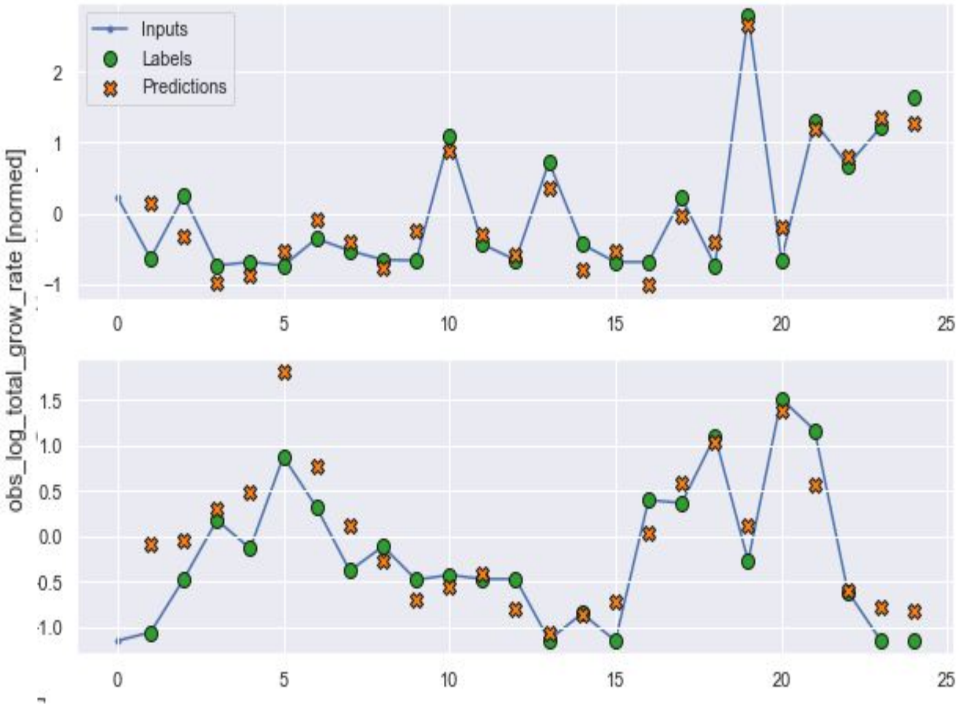


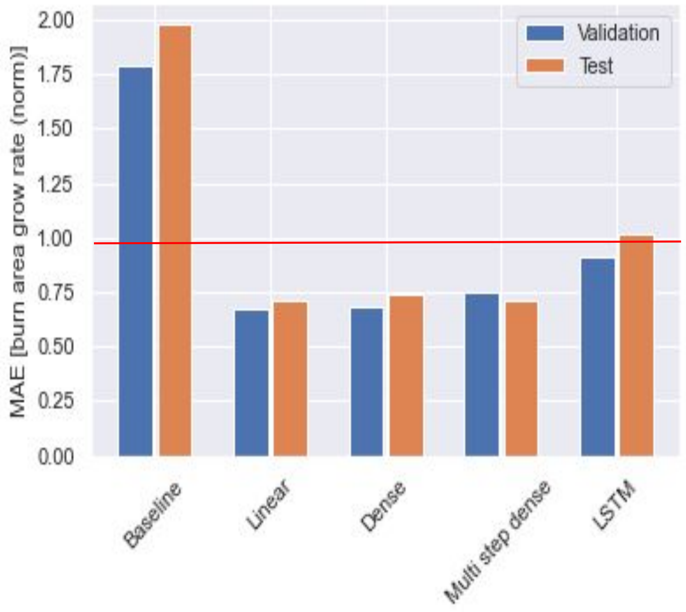
Model diagram from TensorFlow

Time Series Models:
Various TF Single & Multi-Step Models

LSTM







Model diagram from TensorFlow

Conclusions

- ❖ Random Forest ML Model outperforms the baseline linear regression by +47% improvement on R-Squared, -35% reduction in MSE and +14% improvement in accuracy predicting a fire
- ❖ Random Forest accuracy prediction of 66% compares with 72% as reported in the Wildfire Prediction Through Live Fuel Moisture Content Maps (Civil and Environmental Engineering, Stanford University)
- ❖ DNN model with proper architecture and parameter tuning can potentially outperform Random Forest Model
- ❖ Temporal effect of burnt area growth on future fires requires more research and higher frequency of data
- ❖ Integrate satellite images as a feature (Capstone, anyone?)

Contributions / Primary Areas of Focus

	Prakash Krishnan	Joe Ritter	Mon Young
Theoretical Research	✅	✅	✅
Data Cleaning	✅	✅	✅
Exploratory Data Analysis	✅	✅	✅
Data Splitting	✅	✅	✅
Hyper Parameter Tuning	✅	✅	✅
Augmentations	✅	✅	✅
Presentation Slides	✅	✅	✅

github: https://github.com/mon203/w207-final-project-sum2022

Appendix

Appendix

github: https://github.com/mon203/w207-final-project-sum2022

Our final report
https://github.com/mon203/w207-final-project-sum2022/blob/main/w207_Final_Project_Report.ipynb

Our Team

Joe Ritter

Prakash Krishnan

Mon Young

Machine Learning Techniques Leveraged

1. GeoPandas Visualization
2. GeoPandas Spatial Join for Feature Data Set
3. EDA - Scatter Plot, Heatmap, Correlation Plot, Histogram

1. SciKit Learn Linear Regression
2. SciKit Learn Random Forest Regression
3. Scikit Learn Gradient Boost Regression
4. Scikit Learn Decision Tree Regression
5. Scikit Learn Principal Component Analysis
6. RandomizedSearchCV for Parameter Tuning
7. Test Set Stratification by Sub Groups

1. FF DNN with hidden layers
2. FF DNN Parameter Tuning
3. FF DNN Regression and Logistic Regression
4. Time Series Modelling of Temporal Effect of Burnt Area

Features and Labels

Local Meteorology

monthly mean of daily precipitation
monthly mean surface temperature
monthly mean relative humidity
Monthly mean zonal component of wind speed
Monthly mean meridional component of wind speed
Monthly mean energy release component
Monthly mean 1000-hour dead fuel moisture
Monthly mean vapor pressure deficit
Monthly mean Palmer Drought Severity Index

Observed burned area
Observed normalized burned area
Predicted normalized burned area

Label

Label options:

- Burnt Area Size
- Fire Class Based on Burnt Area Size

Large Scale Meteorological Patterns

Monthly standard deviation of daily SVD1 for northern California
Monthly standard deviation of daily SVD2 for northern California
Monthly standard deviation of daily SVD1 for southeastern US (with 2-month lag)
Monthly standard deviation of daily SVD2 for southeastern US (with 2-month lag)
Monthly standard deviation of daily SVD1 for southern Rocky Mountain
Monthly standard deviation of daily SVD2 for southern Rocky Mountain

Land Surface Properties

Monthly mean surface soil moisture
Monthly mean evapotranspiration
Normalized difference vegetation index
Water bodies
Grasslands
Shrublands
Broadleaf Croplands
Savannas
Evergreen Broadleaf Forests
Deciduous Broadleaf Forests
Evergreen Needleleaf Forests
Deciduous Needleleaf Forests
Nonvegetated Lands
Urban and Built-up Lands
elevation
slope

Socio-Economic Properties

Longitude of the grid
Latitude of the grid
Population density
GDP per capita
Number of campsites

* Each example row represent one grid (0.25 degree by 0.25 degree centroid) for each month and year

Features

Land-Surface Properties

Feature Variable	Feature Name	Unit
soilm	Monthly mean surface soil moisture	kg m-2
ET	Monthly mean evapotranspiration	kg m-2
NDVI	Normalized difference vegetation index	unitless
p_1	Water bodies	%
p_2	Grasslands	%
p_3	Shrublands	%
p_4	Broadleaf Croplands	%
p_5	Savannas	%
p_6	Evergreen Broadleaf Forests	%
p_7	Deciduous Broadleaf Forests	%
p_8	Evergreen Needleleaf Forests	%
p_9	Deciduous Needleleaf Forests	%
p.x	Nonvegetated Lands	%
p.y	Urban and Built-up Lands	%
elev	elevation	m
slope	slope	degree

Local Meteorology

Feature Variable	Feature Name	Unit
apcp	monthly mean of daily precipitation	kg m-2
temp	monthly mean surface temperature	K
rhum	monthly mean relative humidity	%
uwnd	Monthly mean zonal component of wind speed	m/s
vwnd	speed	m/s
ERC	Monthly mean energy release component	
FM1000	Monthly mean 1000-hour dead fuel moisture	%
VPD	Monthly mean vapor pressure deficit	kPa
PDSI	Monthly mean Palmer Drought Severity Index	

Large-Scale Meteorological Patterns

Feature Variable	Feature Name	Unit
SVD1_NCA	northern California	unitless
SVD2_NCA	northern California	unitless
SVD1_SE	Monthly standard deviation of daily SVD1 for southeastern US (with 2-month lag)	unitless
SVD2_SE	Monthly standard deviation of daily SVD2 for southeastern US (with 2-month lag)	unitless
SVD1_RM	Monthly standard deviation of daily SVD1 for southern Rocky Mountain	unitless
SVD2_RM	Monthly standard deviation of daily SVD2 for southern Rocky Mountain	unitless

Socio-Economic

Feature Variable	Feature Name	Unit
Lon	Longitude of the grid	degree
Lat	Latitude of the grid	degree
pop2	Population density	population km-2
GDP	GDP per capita	Constance 2011 international US dollar
N_campsite	Number of campsites	

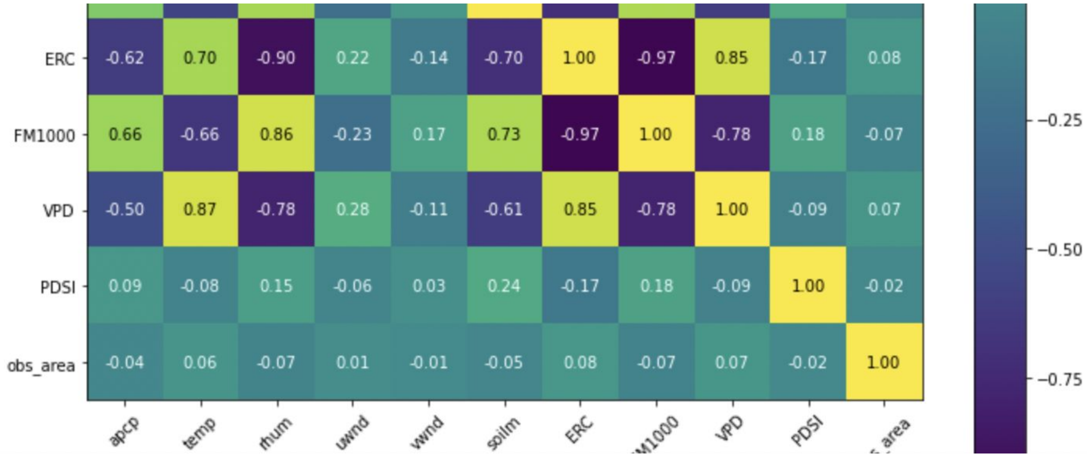
Key Takeaways from Feature Distributions

Item	Observation	Conclusion
Local Meteorology Variables	• Scatter plots demonstrate a highly non-linear relationship between features and obs_area • Some features have a high correlation (ex. ERC & FM1000). • Low Monthly Mean Daily Precipitation, High Monthly Mean Surface Temperature, Low Monthly Mean 1000-Hour Dead Fuel Moisture and Low Monthly Mean Vapor Pressure Deficit have a effect on wildfires	• The target label (obs_area) is highly skewed -> Log transformation. • Can be determinant features for the ML model. Validate via SHAP Analysis on Final Model • Linear Regression -> poor results • Need a ML model such as Neural Network, Random Forest Regression or Gradient Boost Regression
Land Surface Property Variables	• High Monthly Mean Evapotranspiration and Low Deciduous Broadleaf Forest have an effect on wildfires	• Can be determinant features for the ML model. Validate via SHAP Analysis on Final Model
Socio-Economic and Location Variables	• GDP and Population do not seem to have a clear relationship to burnt area	• Left in the final model due to findings from Literature Review
Large Scale Meteorological Patterns	• Long term patterns in Northern California and Rocky Mountains seem to have an effect of the size of wildfires as evident from the scatter plots	• Included in the Final Model
Location Variable (Lat/Lon)	• Wildfires occur all over CA • Large wildfires are restricted to certain grid locations	• Will be a key feature
Time Series Trends	• No appreciable long term trend observed • Seasonal patterns exist as expected	• Can potentially use a random shuffle for a train/test split. Researcher recommended this. • Also included 10 Fold CV

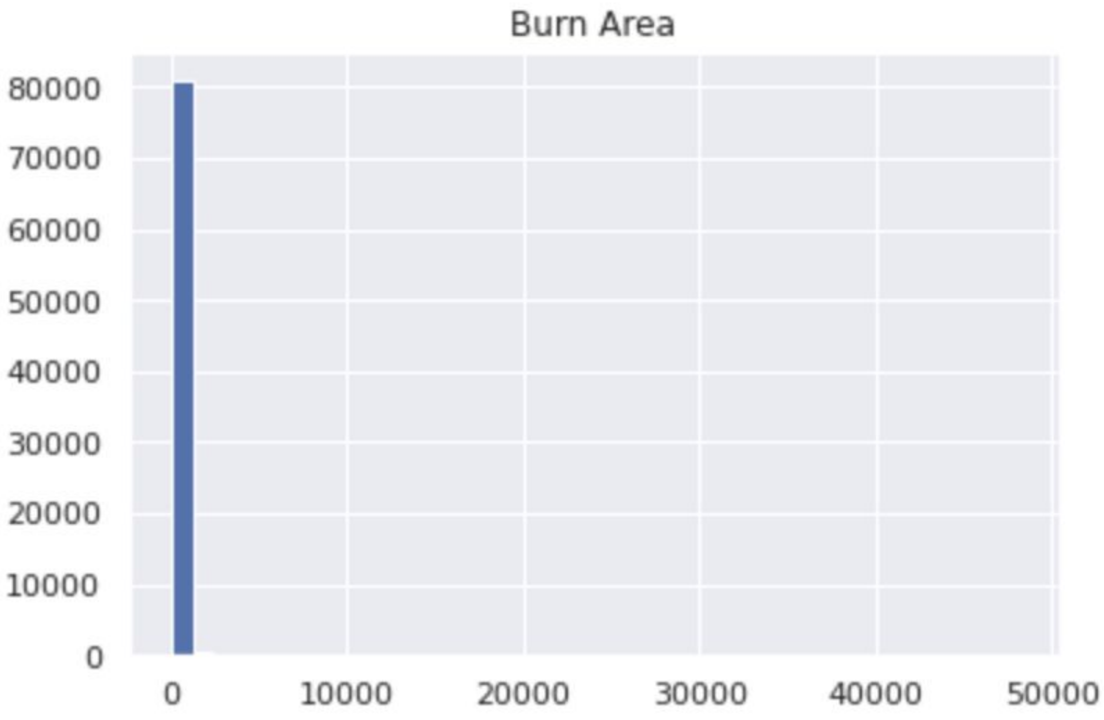
Our dataset

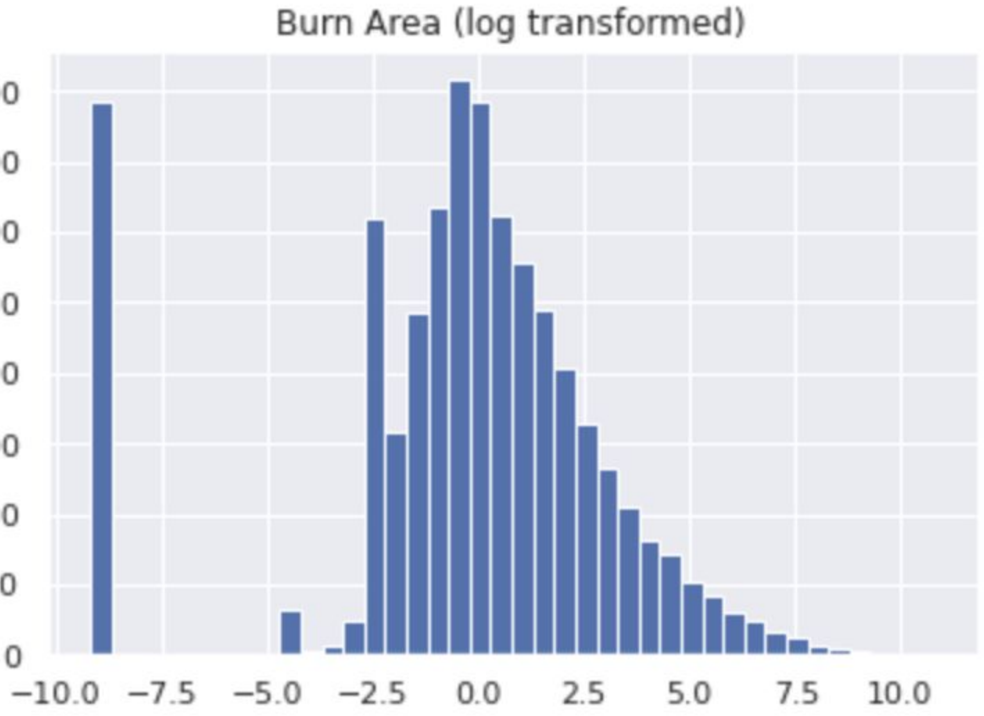
- Our dataset is a a structured dataset. We examine histograms, scatter plots, correlations and heatmaps.
- Colinearality
  - ERC & FM1000 = -0.97, ERC & rhum = -0.90
  - We have fewer than 100 features, having them in the machine learning model should not impact our result.
  - We will note these highly correlated features and examine them further in our model to verify our assumption.

```
ERC     FM1000   -0.969334
FM1000  ERC      -0.969334
ERC     rhum     -0.900941
rhum    ERC      -0.900941
Lon     Lat      -0.795681
Lat     Lon      -0.795681
rhum    VPD      -0.783512
VPD     rhum     -0.783512
FM1000  VPD      -0.777328
VPD     FM1000   -0.777328
dtype: float64
```



Outcome Labels with Log Transformed





Conclusion: Key Results

Item	Purpose	ML Model	Test Evaluation Metric	% Improve Over Baseline	Features and Labels	Hyper Parameters
1	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Baseline Scikit Learn Linear Regression	R Squared: 0.426 MSE: 7.479 Overall Accuracy: 0.582	Not Applicable	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	None
Advance Models						
2	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Tuned Scikit Learn Random Forest Regressor using RandomizedSearchCV	R Squared: 0.628 MSE: 4.849 Overall Accuracy: 0.666 Class Prediction Accuracy: (A: 0.777, B: 0.871, C: 0.725, D: 0.114, E: 0.072, F: 0.084, G: 0.029)	+ 47.4% - 35.1% + 14.4%	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	{'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 20, 'bootstrap': True}
3	Log(Burnt Area) Prediction in a Grid Fire Class Prediction in a Grid	Feedforward Neural Network with manual hypermarameter selections	R Squared: 0.4917 MSE: 6.6283 Overall Accuracy: 0.6281 Class Prediction Accuracy: (A: 0.695, B: 0.730, C: 0.304, D: 0.0, E: 0.0, F: 0.0, G: 0.0)	+ 15.4% -11.37% + 7.9%	Label: Log(Burnt Area) Features: Local Meteorology, Land Surface Properties, Large Scale Meteorological Patterns, Socio-Economic	{learning rate = 0.000001, optimazor = Adam, batch size = 128, hidden layers = [128, 64, 32], dropout layers = none, epoch = 500}

Executive Summary

- Extensive Literature Review to understand performance of external models and develop domain knowledge
- Leveraged several advanced techniques of ML in the project (Spatial Join, PCA, Time Series Modelling, DNN, Sub-Group Analysis and Hyper Param Optimization)
- Built baseline shallow models (Linear Regression) to assess baseline metrics
- Tuned DNN Model and Random Forest Regression to improve performance over baseline:
  - Random Forest ML Model outperforms the baseline linear regression by +47.4% improvement on R-Squared, -35.1% reduction in MSE and +14.4% improvement in accuracy predicting a fire
  - Random Forest accuracy prediction of 66.6% compares with 71.95% as reported in the Wildfire Prediction Through Live Fuel Moisture Content Maps (Civil and Environmental Engineering, Stanford University)
  - DNN model with proper architecture and parameter tuning can potentially outperform Random Forest Model
- Developed a “Stretch Model” to integrate Temporal effect of burnt area. Good intro to a capstone

Machine Learning Models

Item	Algorithm	Baseline	Advanced	Rational	Evaluation
1.	Linear Regression Predicting a Continuous Variable (“Observed Burnt Area”)	Local Meteorology and Location Features	Add Socio-Economic and Large Scale Patterns	Provides a baseline prediction of burnt area	RMSE
2.	Logistic Regression Predicting a Binary Classification (Fire or Not)	Local Meteorology and Location Features	Add Socio-Economic and Large Scale Patterns	Provides a baseline prediction of fire or not	Accuracy, Precision, Recall
3.	Decision Tree	Local Meteorology and Location Features	Add Socio-Economic and Large Scale Patterns	Provides a baseline understanding of feature importance	Information Gain
4.	Deep Neural Network	All features considered		Expect better performance	RMSE Accuracy, Precision, Recall
5.	Gradient Boosting Regression to predict a Continuous Variable (“Observed Burnt Area) or a Binary Classification (Fire or Not)	All features considered		Better accuracy than linear and logistic regression Can handle non-linear relationship and multi-collinearity	RMSE

Research Question

Given a set of conditions is it possible to determine the:

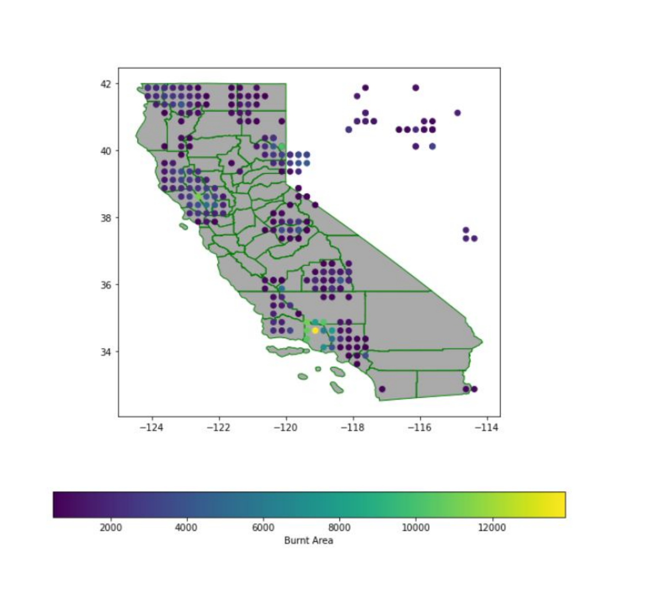
- probability of a wildfire
  (classification)
- size of burnt area
  (continuous variable)

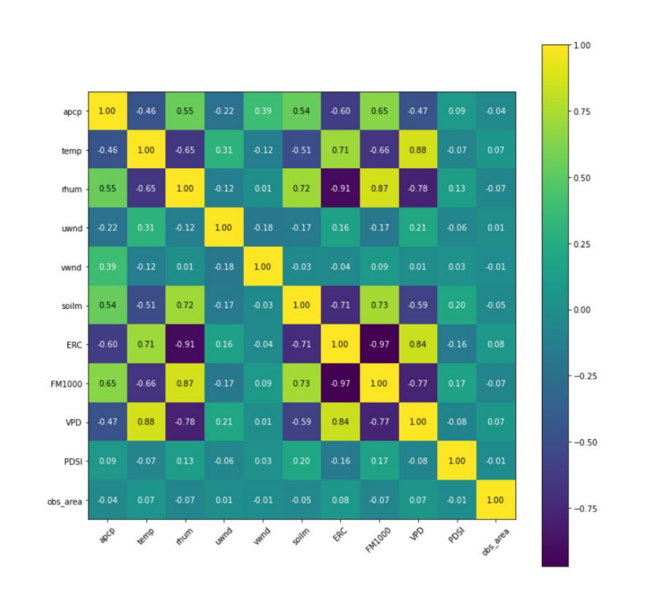
Project Schedule

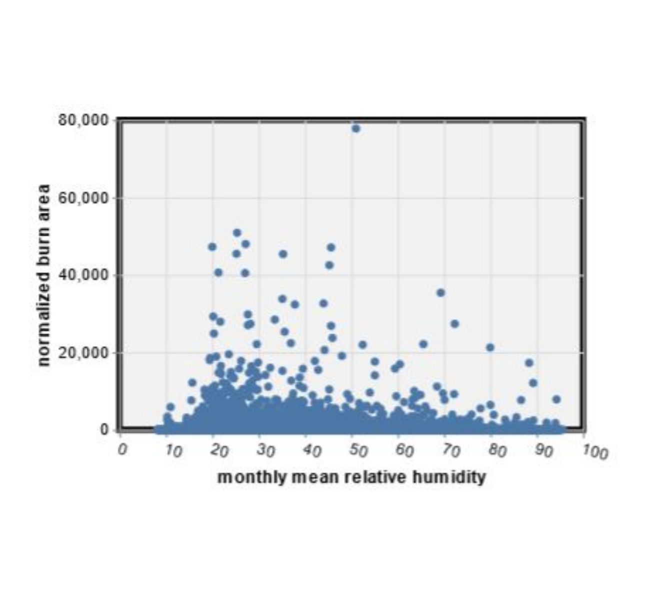
	June-13	June-20	June-27	July- 4	July-11	July-18	July-25	Aug- 1
Data preprocessing								
Read papers and talk to researcher								
Data Visualization								
Build baseline model								
Additional model								
Prepare summary and conclusions								
Prepare presentation								

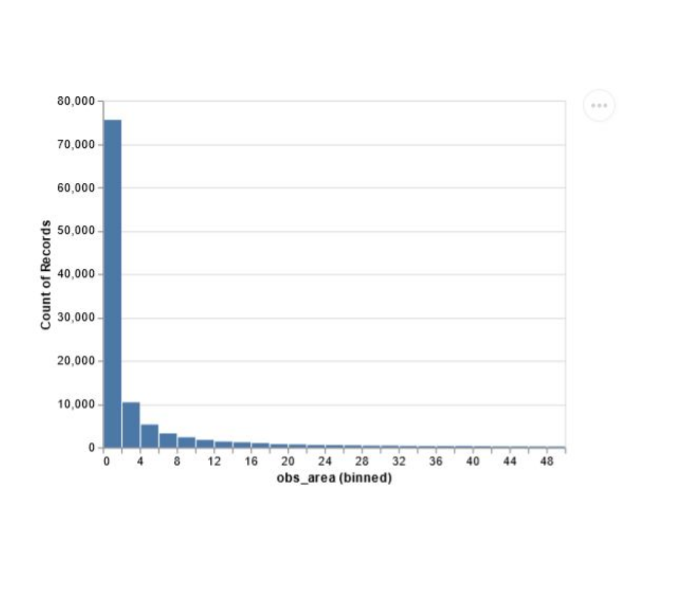
Exploratory Data Analyses

- Geospatial Viz - Geopandas
- Histograms
- Correlation Heatmap - Features and Labels
- Scatterplot - Features and Labels
- Time Series Trends

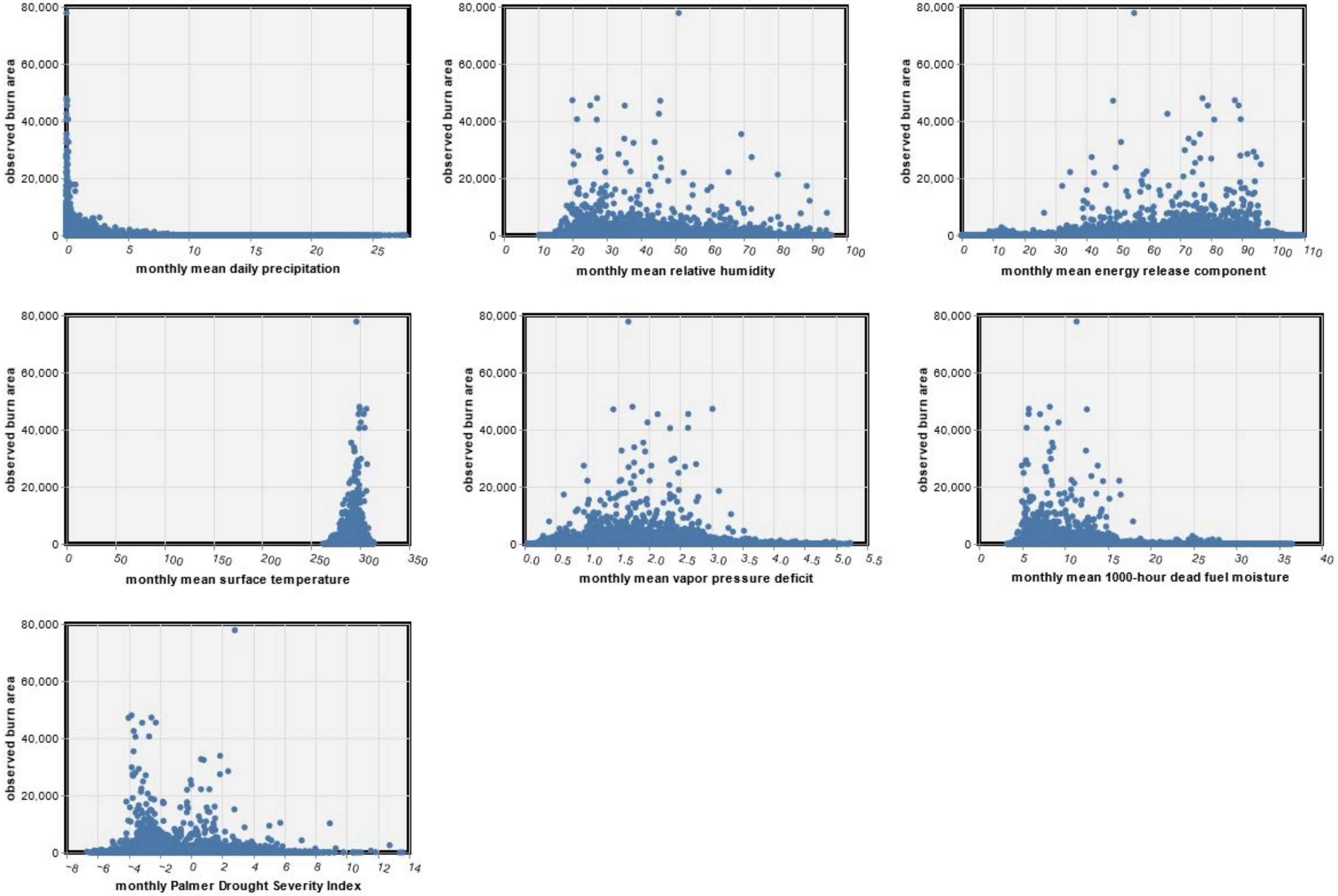


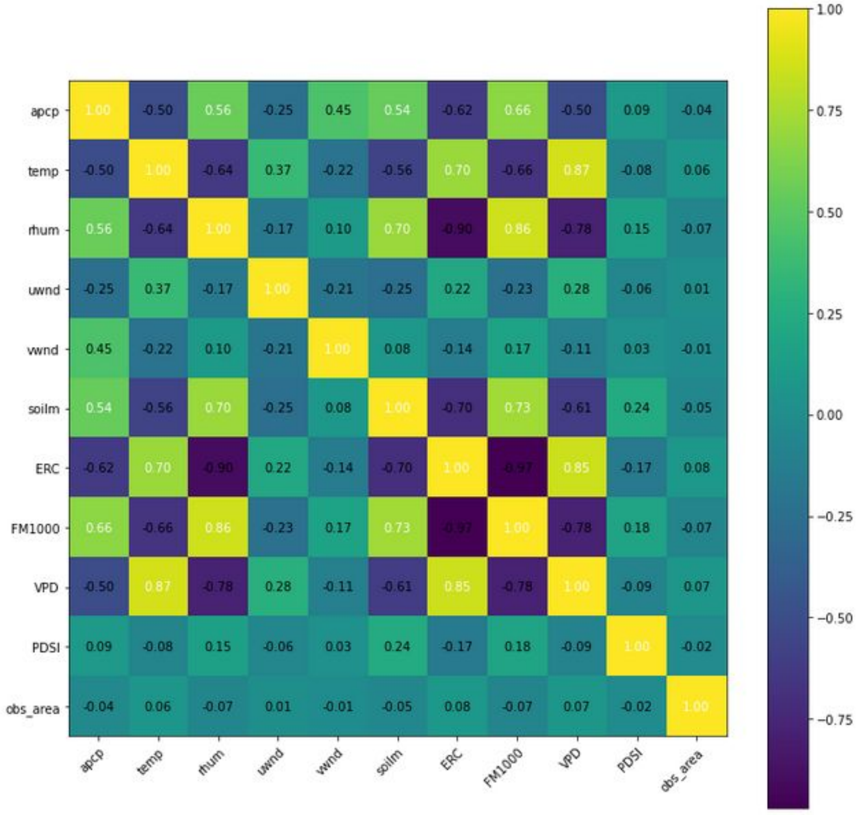




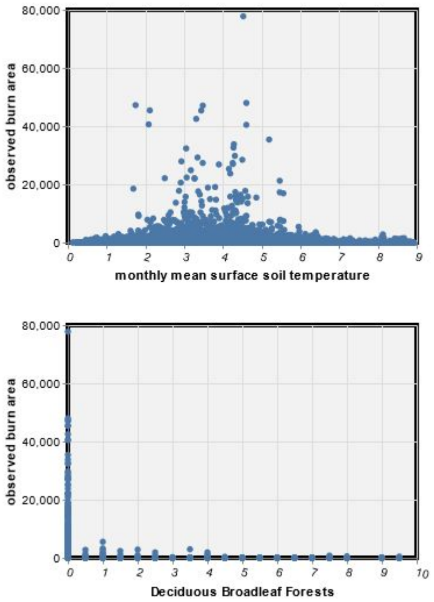


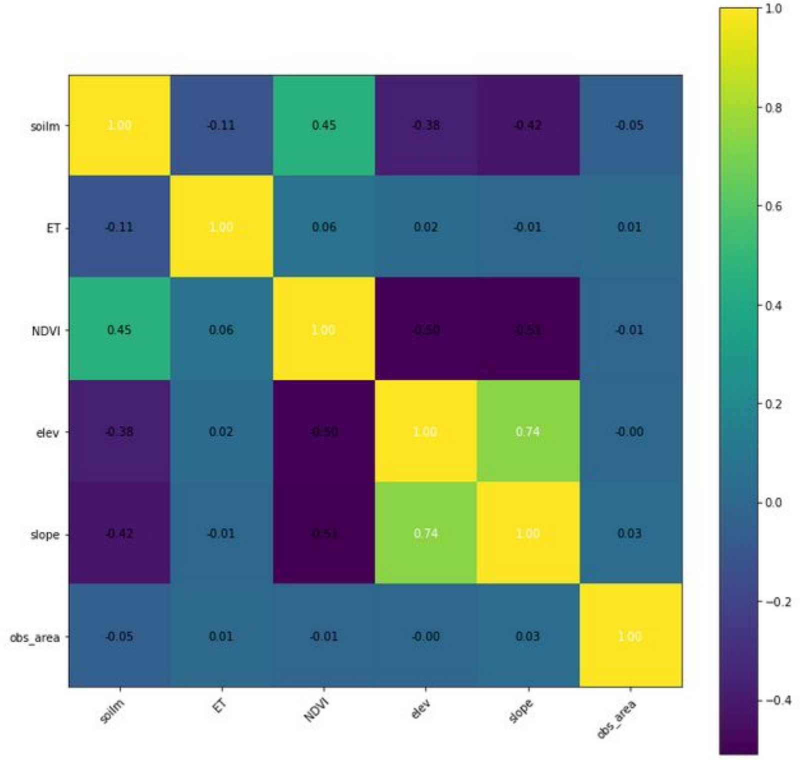
EDA: Local Meteorology Features



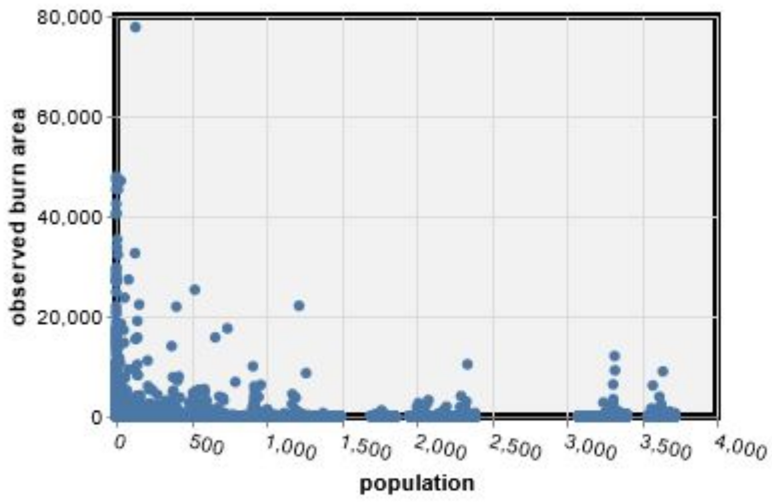


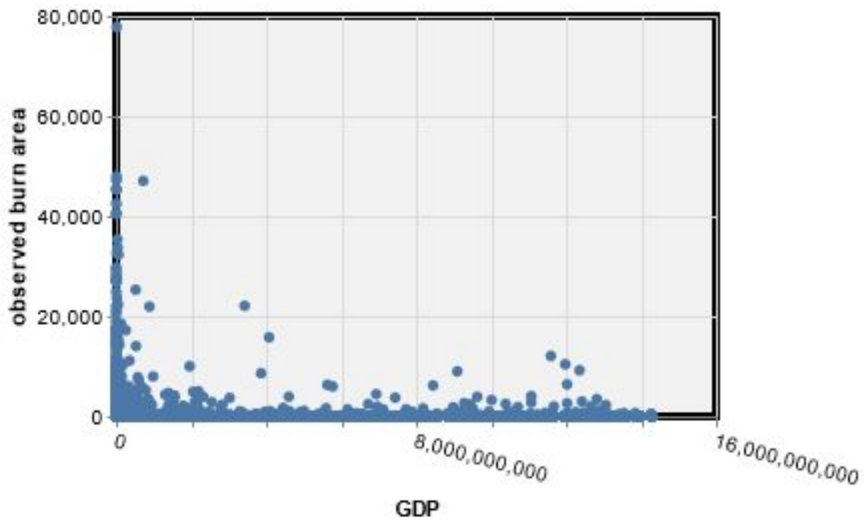
EDA: Land Surface Property Features

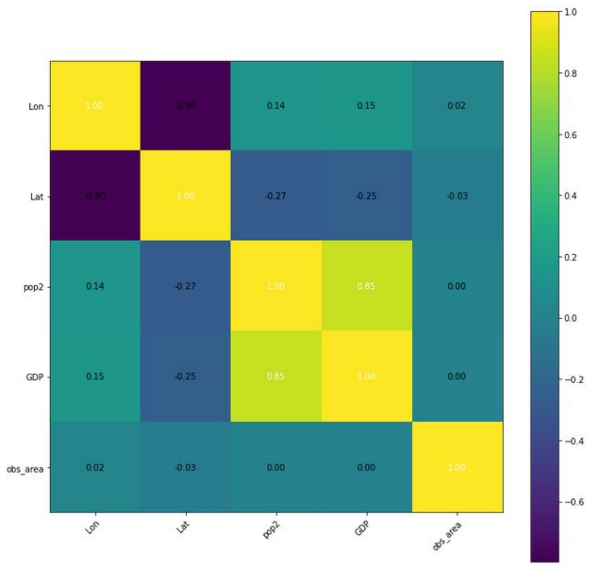




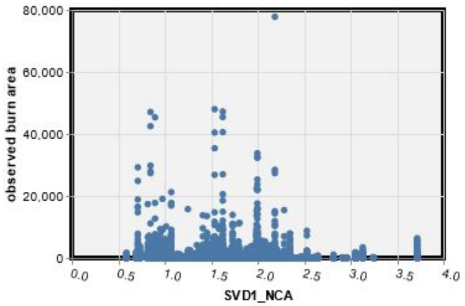
EDA: Socio Economic Features

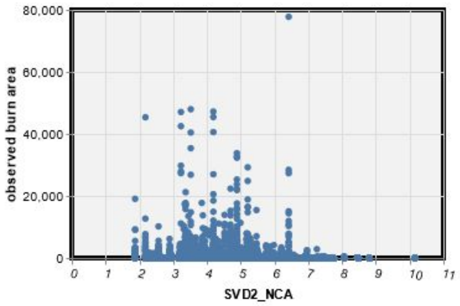


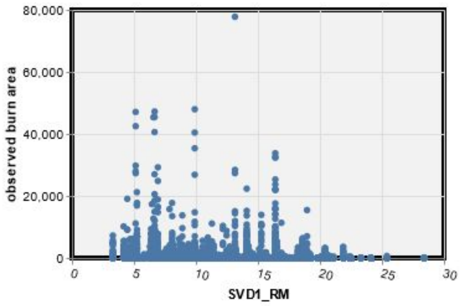


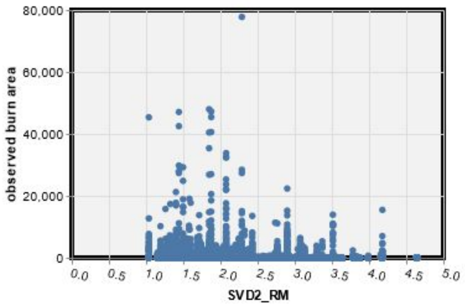


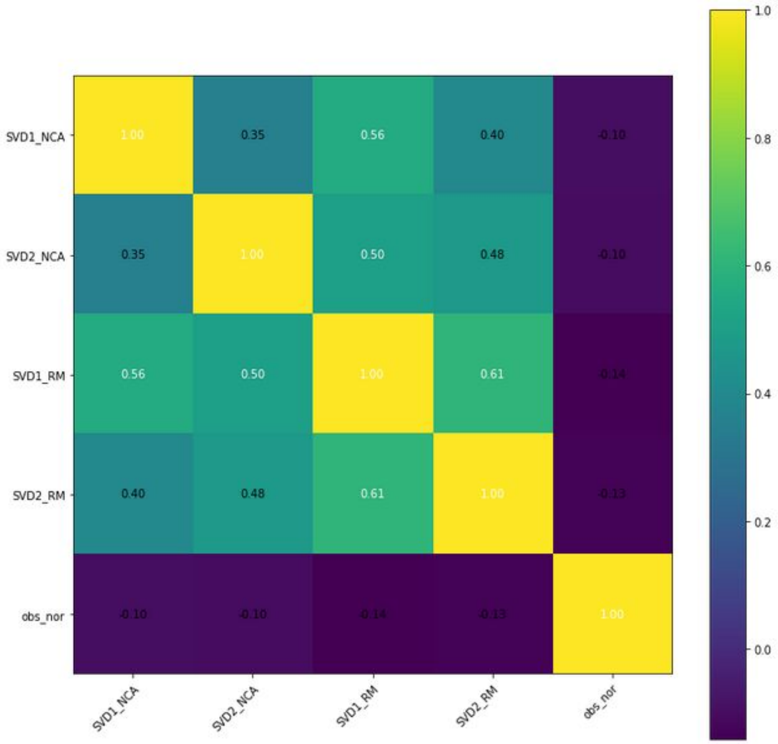
EDA: Large Scale Meteorological Patterns



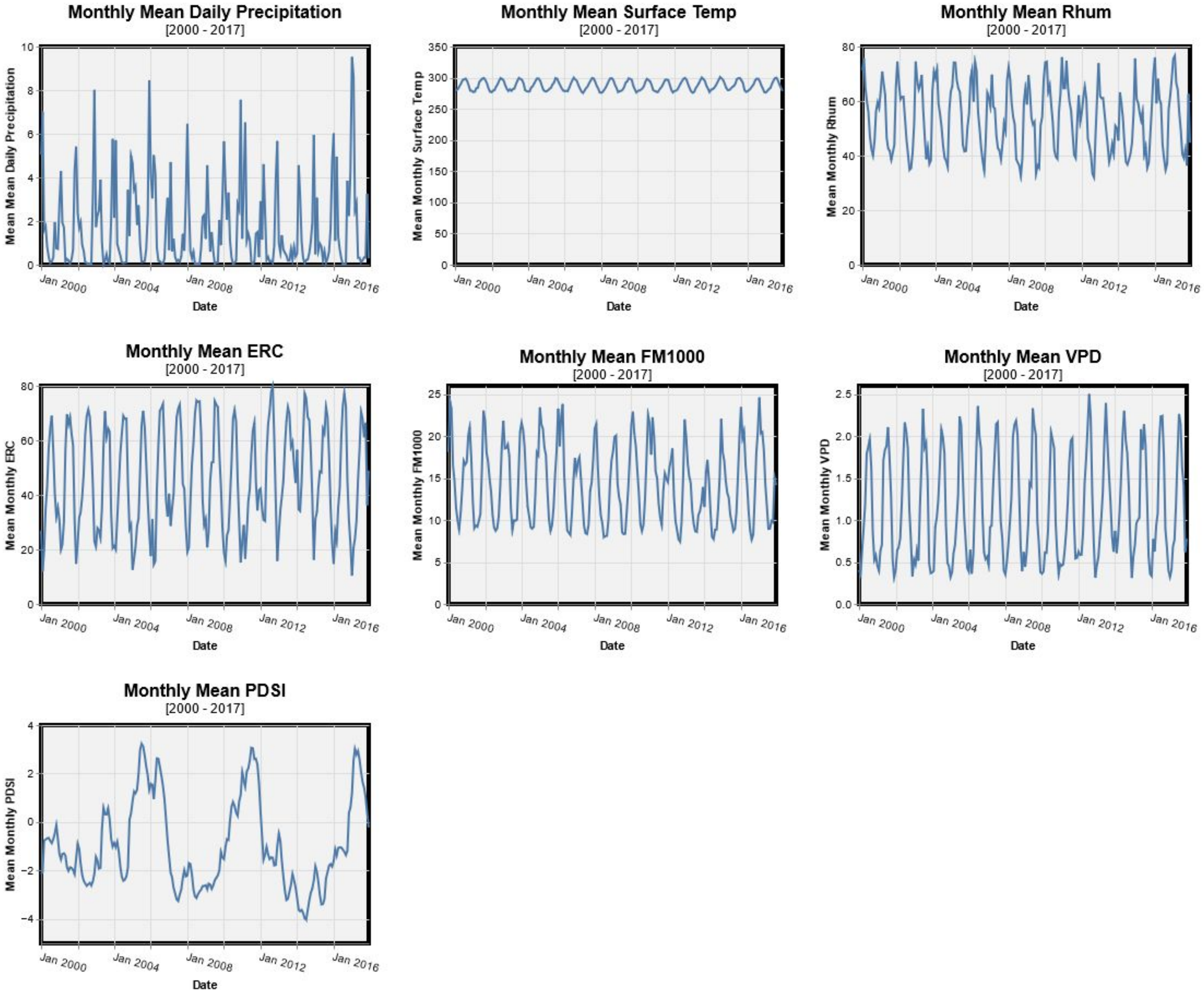






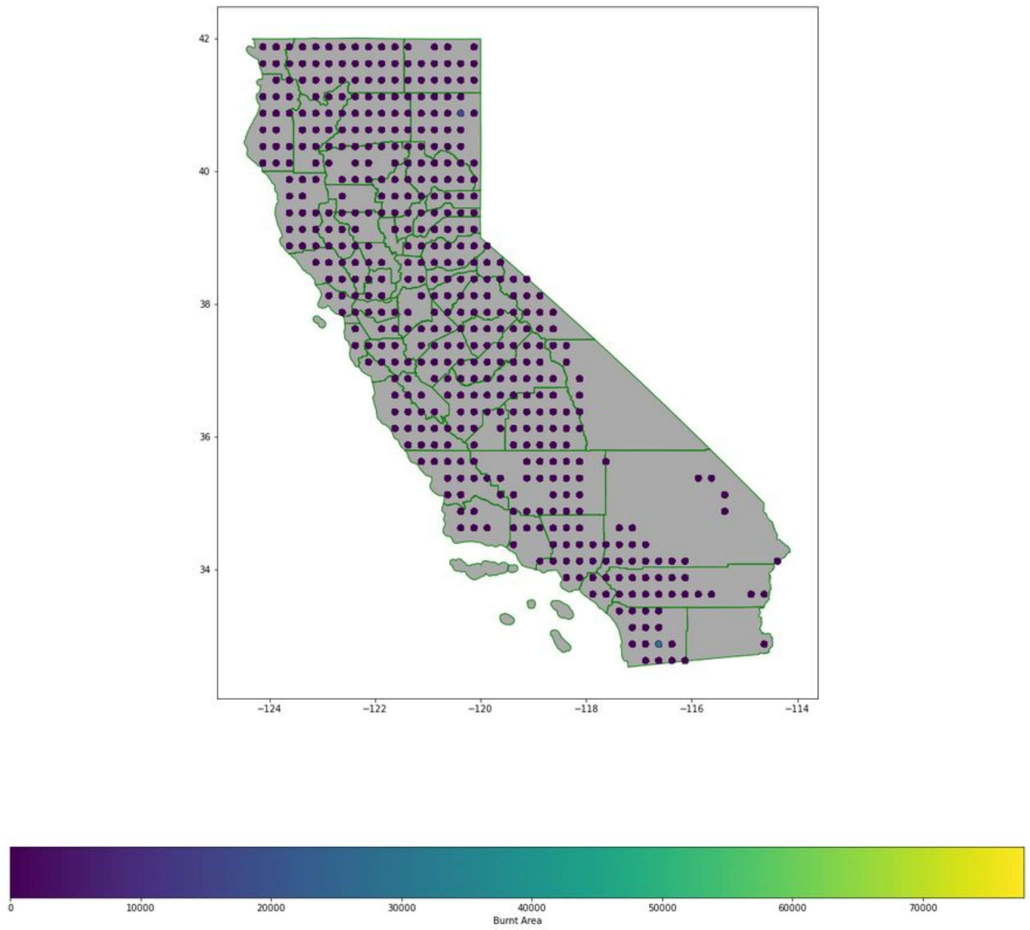


Time Series Trends

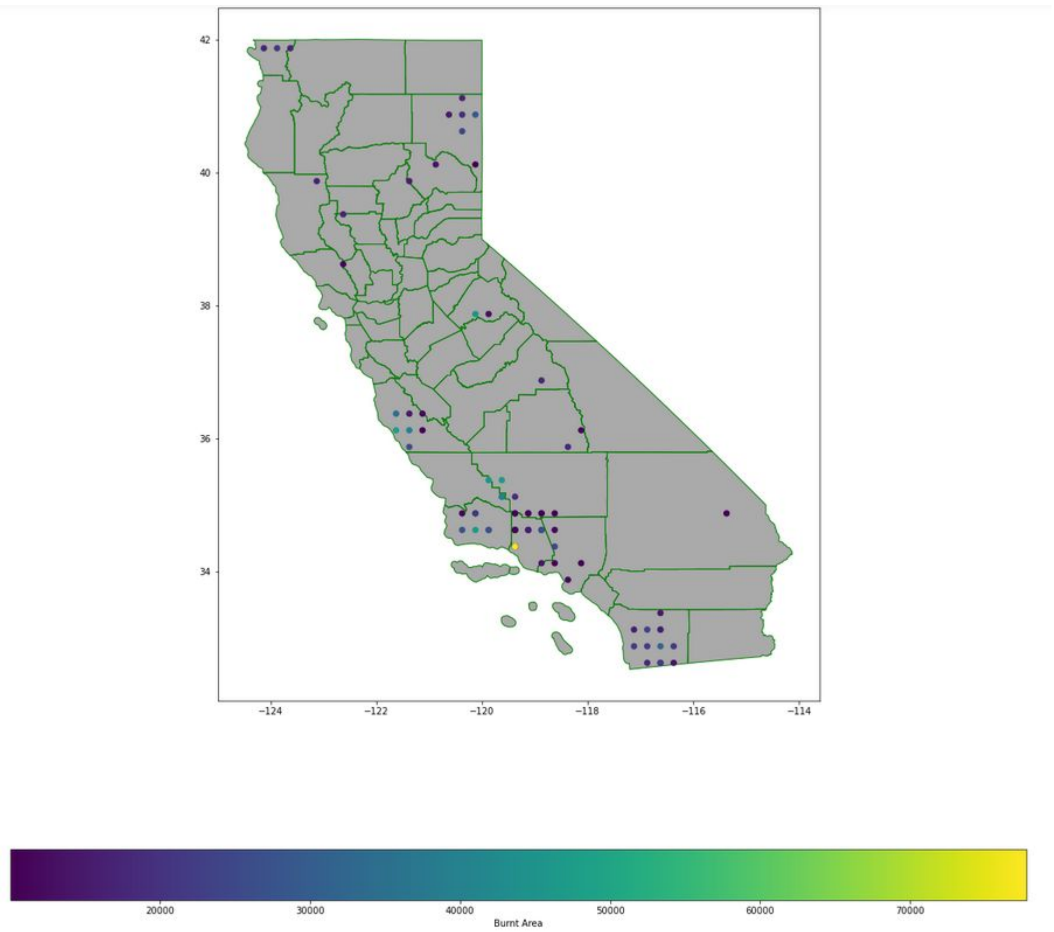


Wildfire locations over last 10 years

All Wildfire in CA



Large Wildfires in CA (>10k hectares)



Certain Counties Experience High Fire Danger