Comparative Predictive Models of Drought

by Daniel Groneberg, Ben McPeek, and Joey Notaro

Problem Statement

- The California State Water Resources Board is seeking better ways to predict drought throughout the state

- The agency has hired a team of data scientists to investigate better methods to predict i

Problem Statement

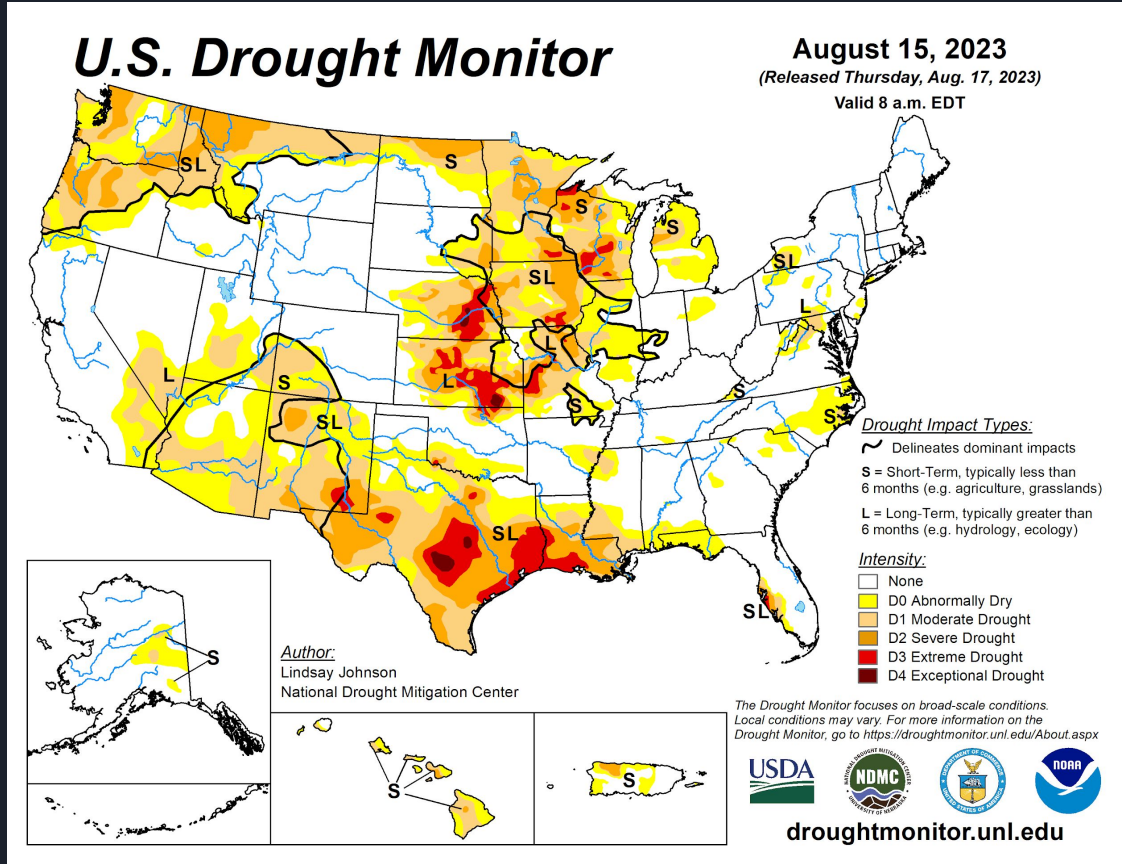
Which predictive model is the best indicator of whether a drought is likely to occur **given historical weather, climate, and satellite image data**?

Does one model perform better than any other in predicting drought or drought-like conditions for the state of California to declare forecasts of droughts more accurately?

Background Information

- Drought prediction has historically relies on precipitation and temperature.
- Climate is inherently variable, these factors maintain notoriously difficult to predict accurately.
- Climate factors are even harder to predict with global climate change.

Modeling Factors - Differential Severity



CA Drought Impacts for Potential Modeling

- Less rainfall, precipitation, and snowpack.
- More frequent and intense wildfires.
- Drier vegetation and soils.

Modeling Methods

- Model 1 - Predict future precipitation based on historical precipitation using SARIMA time series forecasting.
- Model 2 - Predict wildfire areas based satellite images using convolutional neural network.
- Model 3 - Predict soil moisture based on historical soil moisture data using SARIMA time series forecasting.

Model 1 (Time-Series Model)

- **Data:** Daily mean averages of precipitation in Fresno California from 1992-2022

- **Models Used:** RNN, Holt-Winters autoETS, Naive Forecaster, SARIMA, and Ensemble Forecasting

- **Metrics:** RMSE (inches of precipitation)

Model 1 (RNN)

Our target variable's variation can only be explained by 5% of our features of our best performing RNN model (week). This makes using these neural network models not very efficient because our features are not contributing seemingly at all.

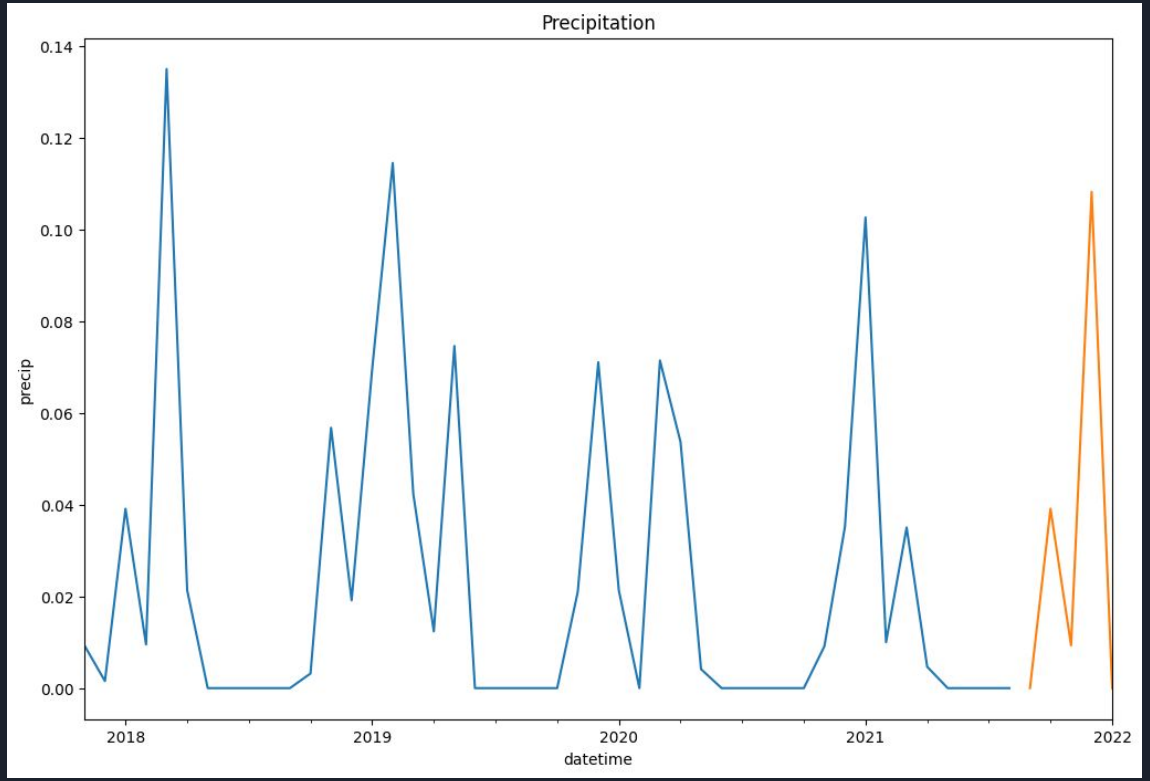
RMSE = 0.037

Model 1 (Splitting Data)

Train: 192 Months
[2005-09-01: 2021-09-01]

Test: 5 Months
[2021-09-02: 2022–02-01]

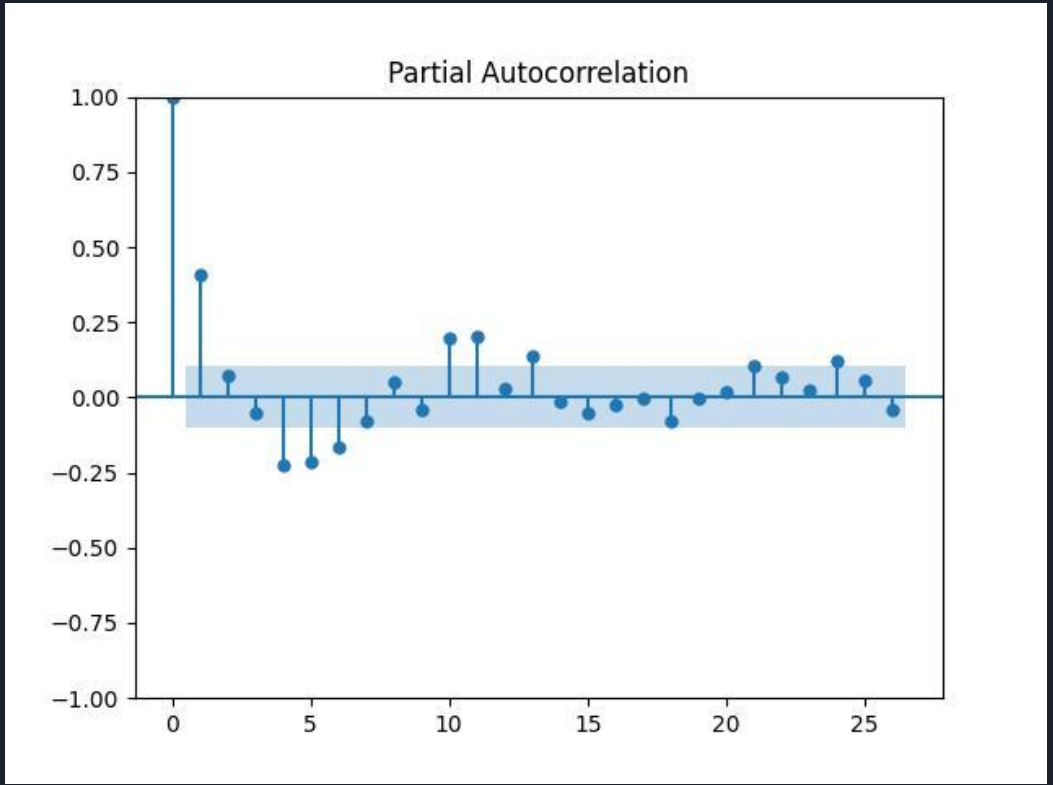
Visually we can see no trend.
Appears to have seasonality at the start of each year (Winter)



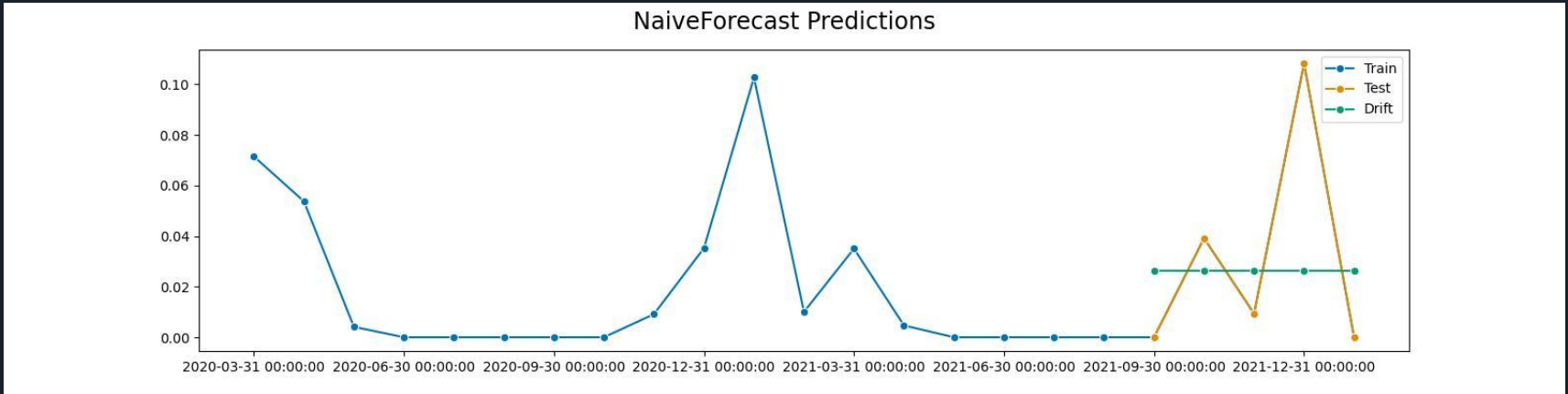
Model 1: Check for Seasonality and Stationarity

We have clear seasonality because there are evident recurring peaks in the lags.

Because we saw peaks every winter season in the previous slide, we will set our Seasonality to months (sp =12)



Model 1 (Baseline Model)



This Naive Forecast model will take the mean of the last seasonal window as our baseline predictions. All other models, including RNN, will base their success off of having a lower RMSE than our baseline model.

Baseline RMSE Score: 0.0414

Model 1 (Performance)

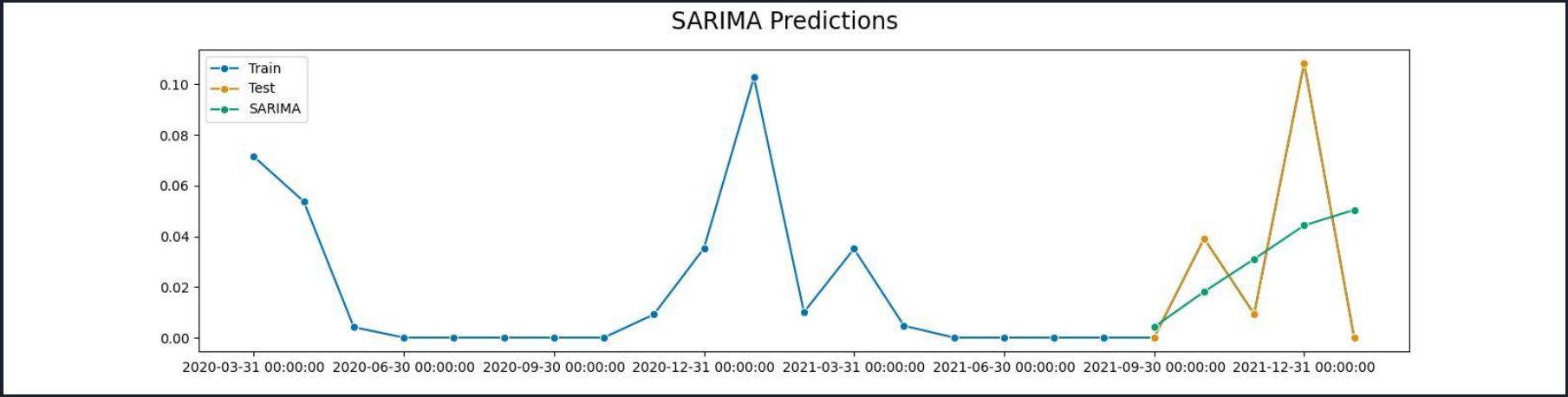
Baseline RMSE Score: 0.0414

HW_EST_Model RMSE Score: 0.0404

SARIMA_Model RMSE Score: 0.0389

Ensemble_Forecast_Model RMSE Score: 0.0395

Model 1 (Best Performance)



SARIMA_Model RMSE Score: 0.0389

Although our improvements in our predictions are minimal, the average measure of precipitation per month is fractions of an inch. This means that any development will also be within fractions of an inch.

Model 2 - Image Classification Convolutional Neural Network (CNN)

- Use labeled satellite images of “no wildfire” and “wildfire” to train a CNN.

- Assumes that landscape conditions for wildfire resemble those for drought (dried vegetation and soils, sparser landscape, etc.)

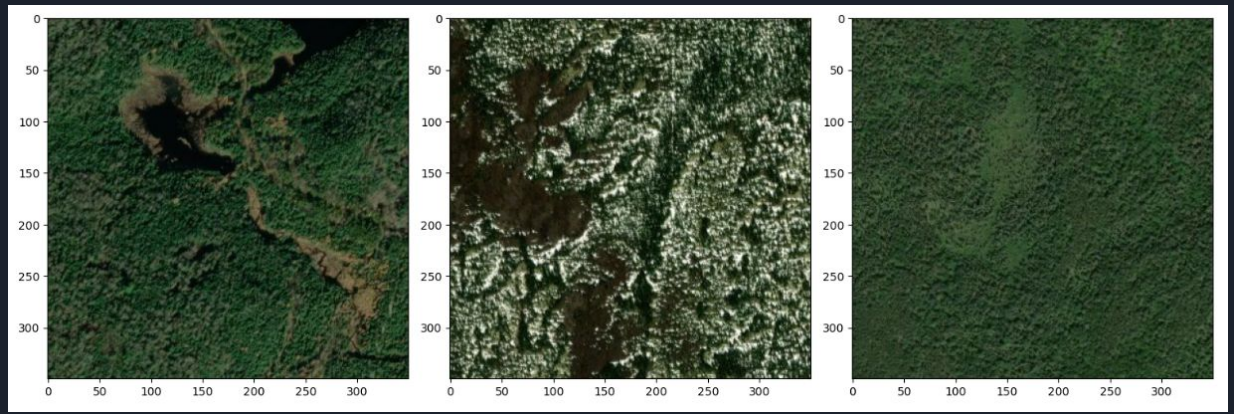
Model 2 - CNN Dataset

- Labeled “no wildfire” and “wildfire” binary classed image dataset procured from Canadian government.
- 14,449 “no wildfire” images (~48%)
- 15,750 “wildfire” images (~52%)

Model 2 - CNN Dataset

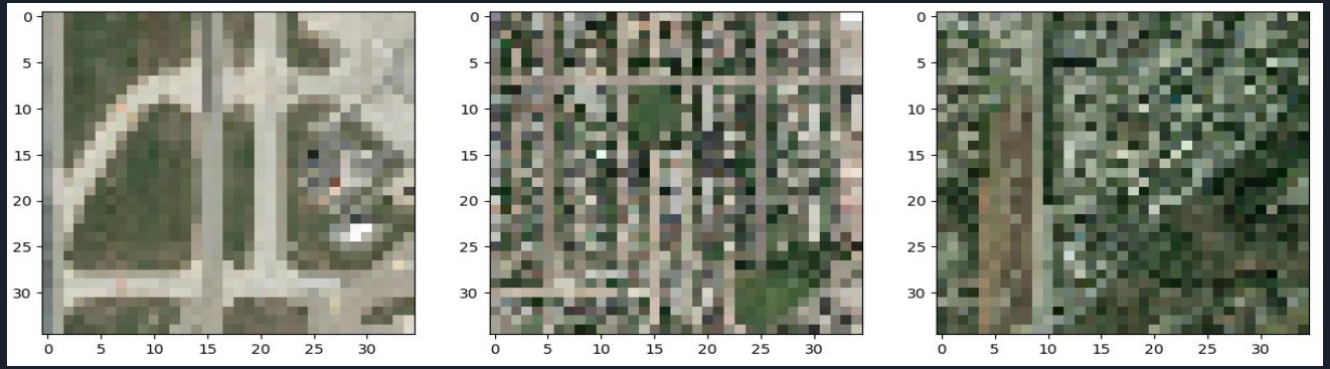
“No Wildfire”

“Wildfire”

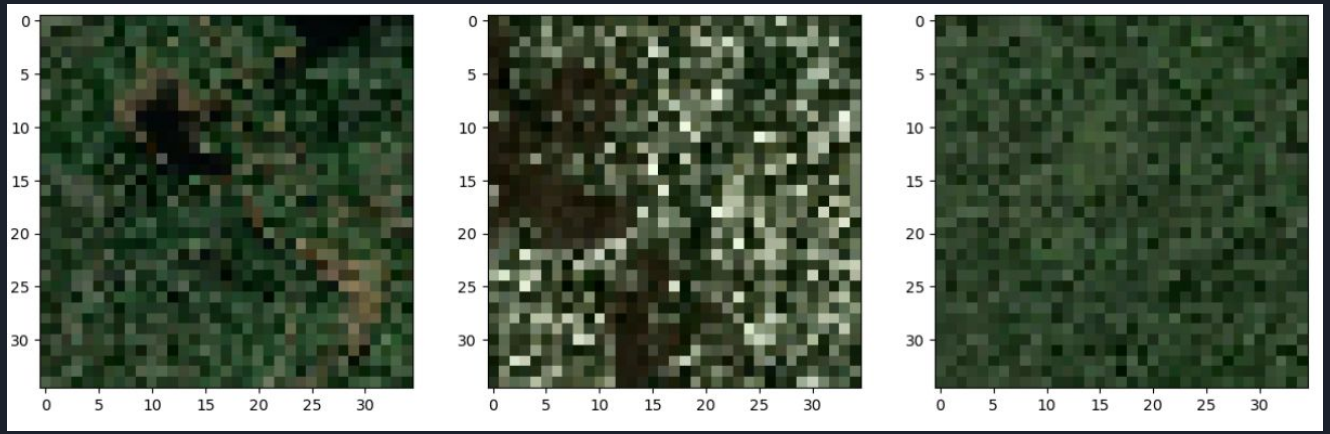


Model 2 - CNN Exploratory Data Analysis

“No Wildfire”

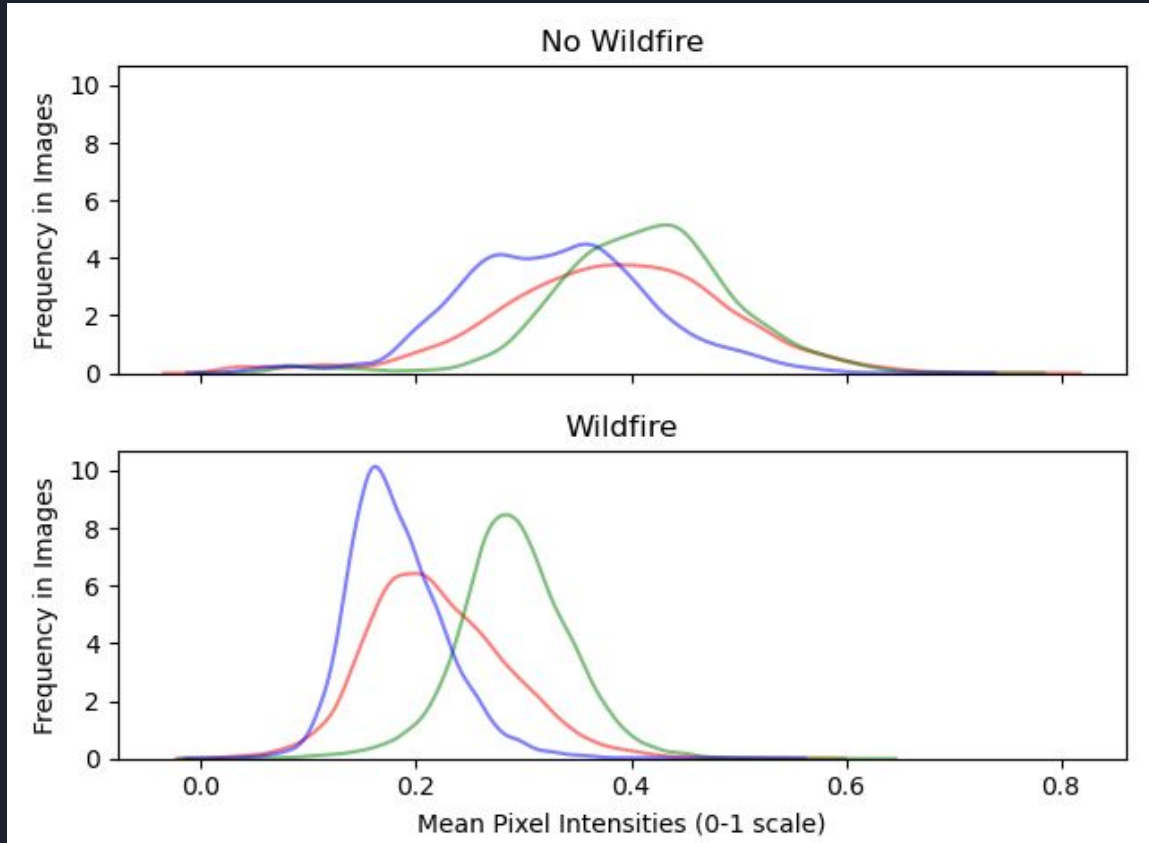


“Wildfire”



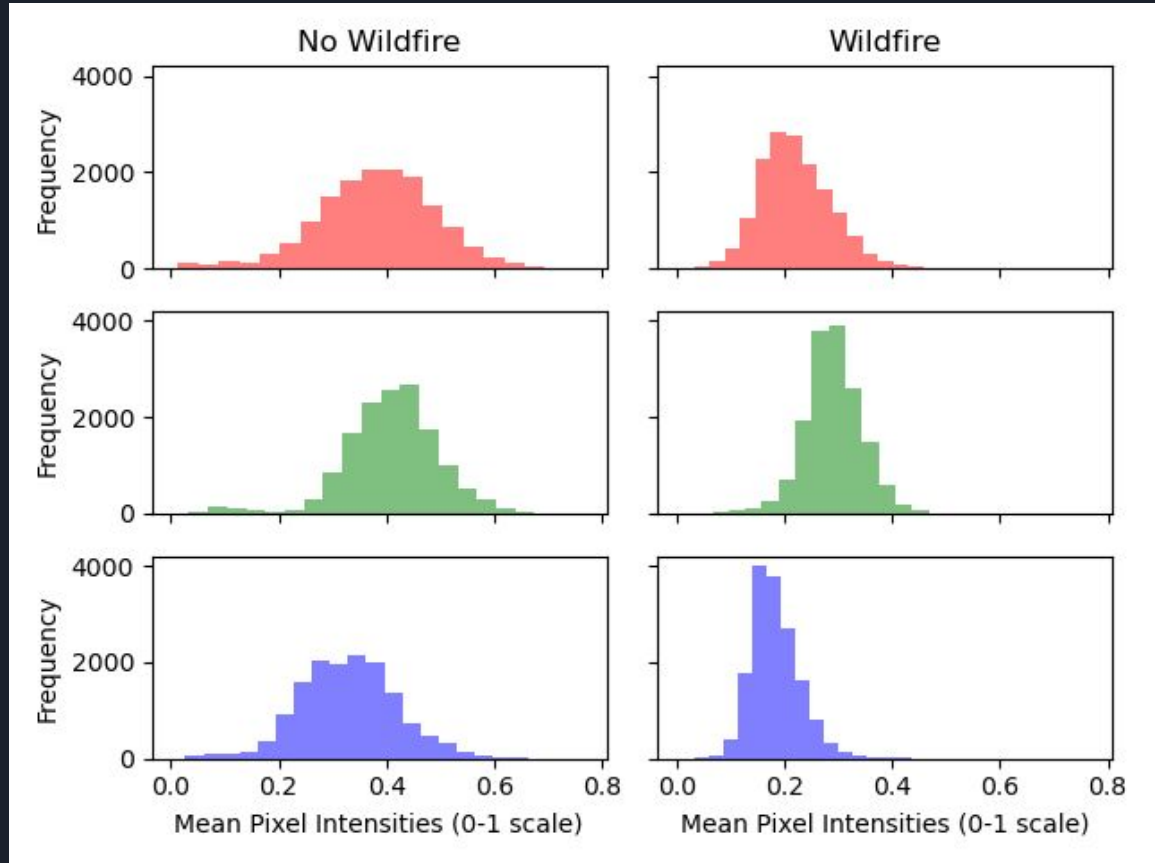
Model 2 - Image Pixel Intensities

- Wider distribution of red, green, and blue pixel intensities for “no wildfire” images.
- Higher pixel intensity on average for “no wildfire” images.



Model 2 - Image Pixel Intensities

- Wildfire images has lower average pixel intensity for all three color channels.
- Wildfire images on average are darker.



Model 2 - CNN Baseline Model

- Null Baseline Accuracy is 0.52

- Baseline CNN Model:
  - One Input Layer with 16 Nodes
  - One Output Layer
  - Trains on Ten Epochs

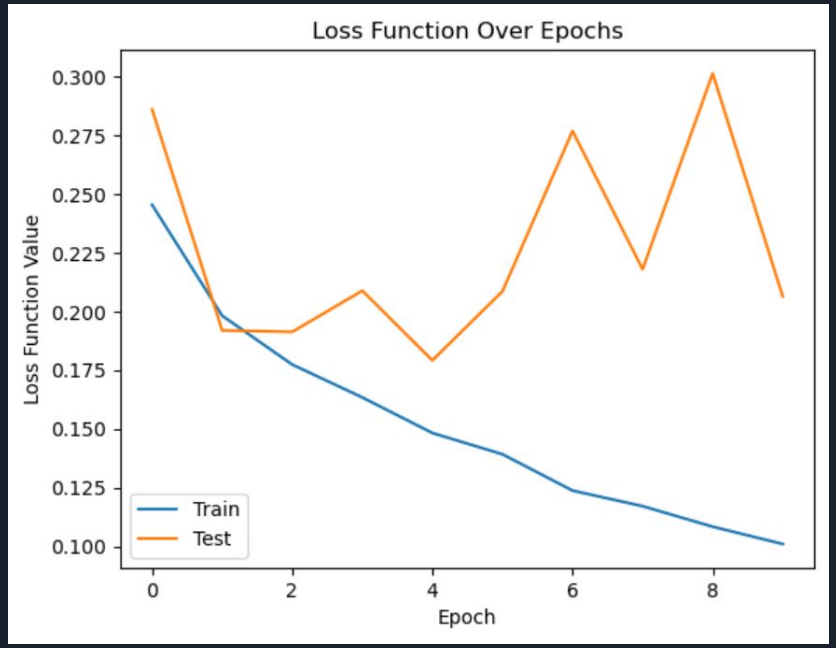
Model 2 - CNN Baseline Model

- Null Baseline Accuracy is 0.52

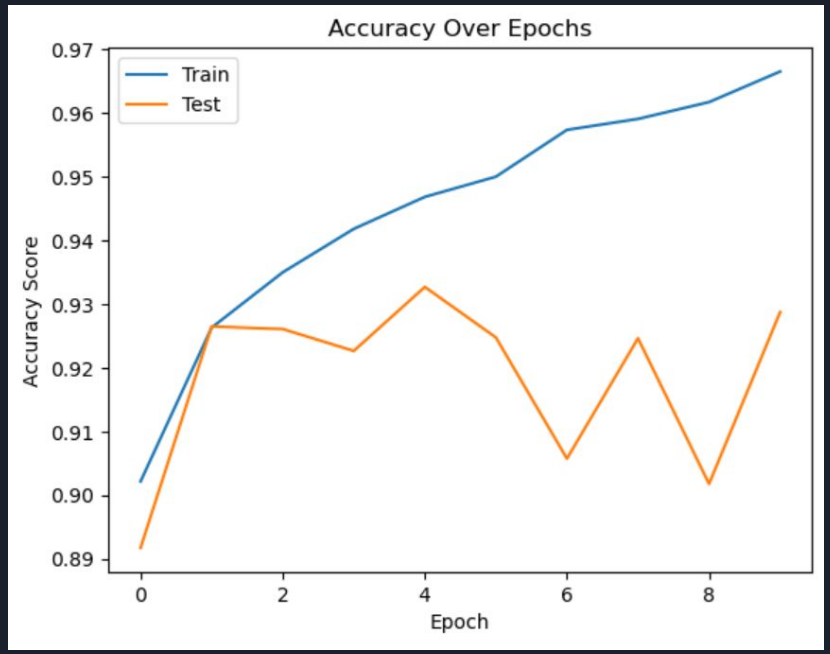
- Baseline CNN Model:
  - One Input Layer with 16 Nodes
  - One Output Layer
  - Trains on Ten Epochs

Model 2 - CNN Baseline Model

Loss Function ~ 0.21



Accuracy ~ 0.93



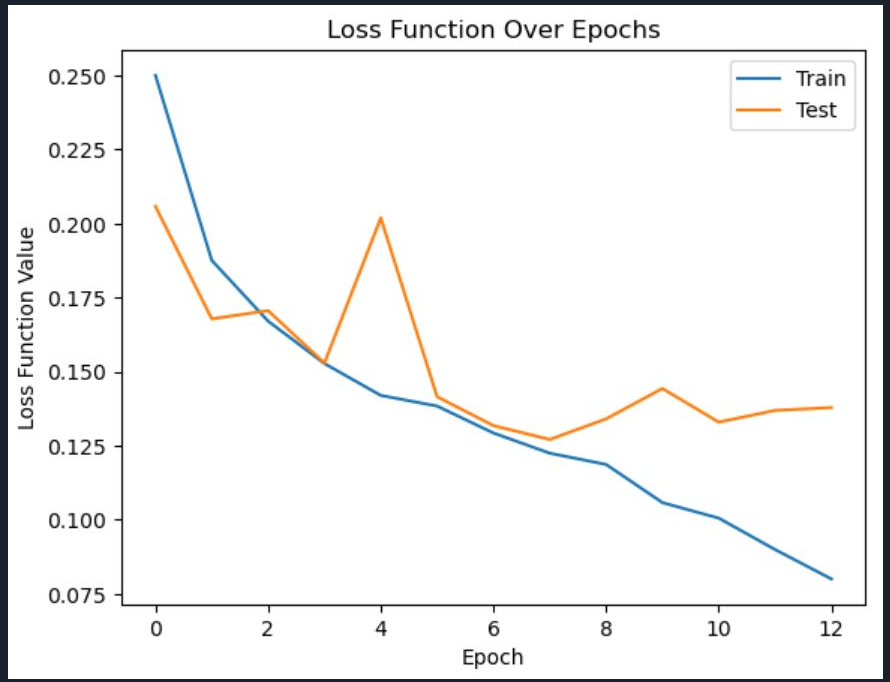
Model 2 - CNN Model Tuning

- Iterated over five different models

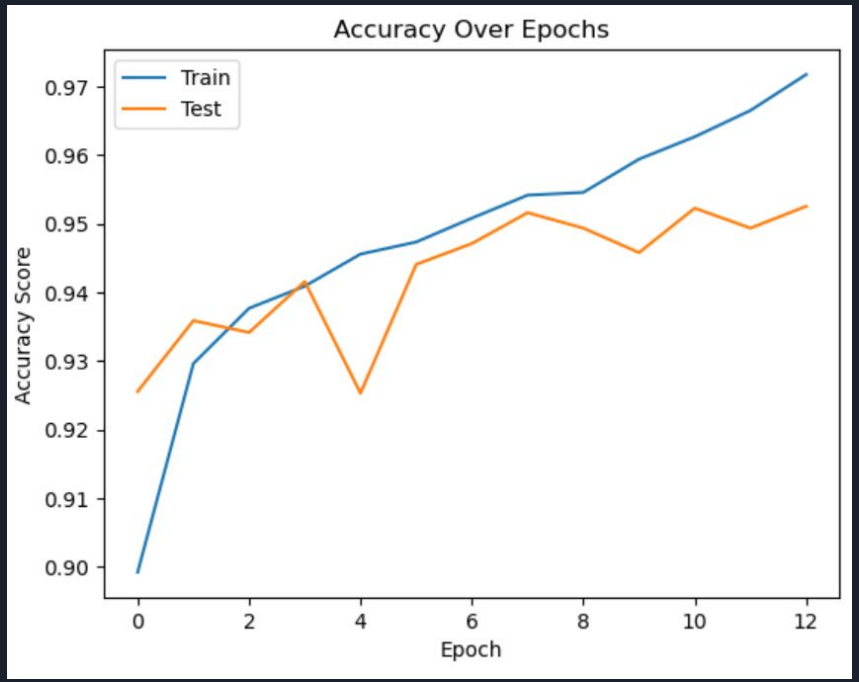
- Tune and Regularize with:
  - Hidden Layers
  - Early Stopping
  - Ridge Regularization
  - Dropout Layers

Model 2 - CNN Production Model

Loss Function ~ 0.13



Accuracy ~ 0.95



Model 2 - CNN Production Model

- Predicts “no wildfire” and “wildfire” images with 95% accuracy

- Production Model includes:
  - Input Layer of 16 Nodes
  - Two Hidden Layers of 64 and 128 Nodes
  - Early Stopping with a Patience of 5 Epochs
  - No Dropout Layers of Regularizations

Model 2 - Findings & Recommendations

- Finding: Image Classification CNN model predicts wildfire images (>0.01 acres) with 95% accuracy

- Recommendation: The agency should invest long-term in procuring labeled satellite images of drought landscapes, including of different severity to train a multi-class CNN drought prediction model.

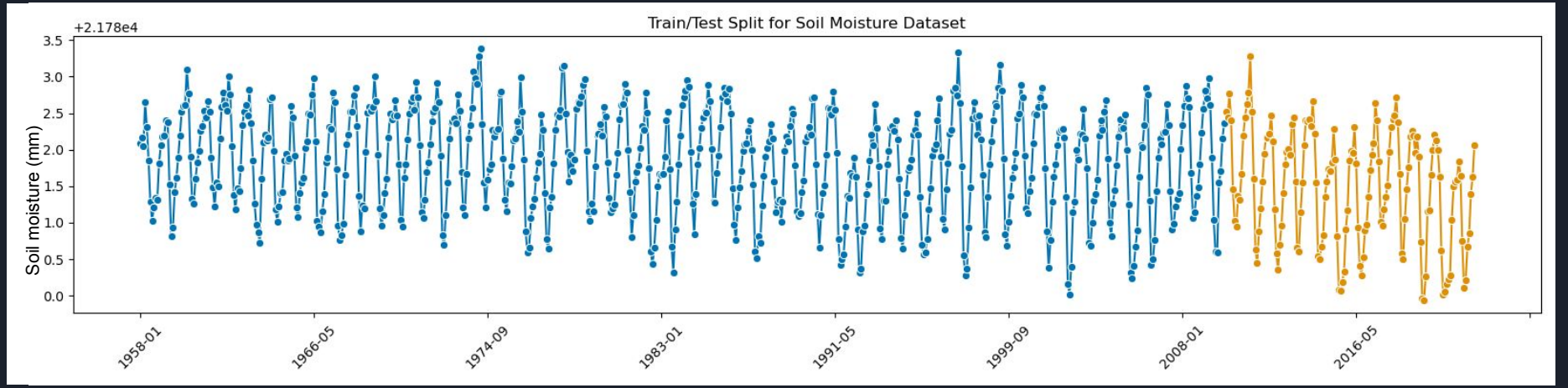
Model 3 (Time-Series Model)

- **Data:** Monthly means of soil moisture data from 1958-2023

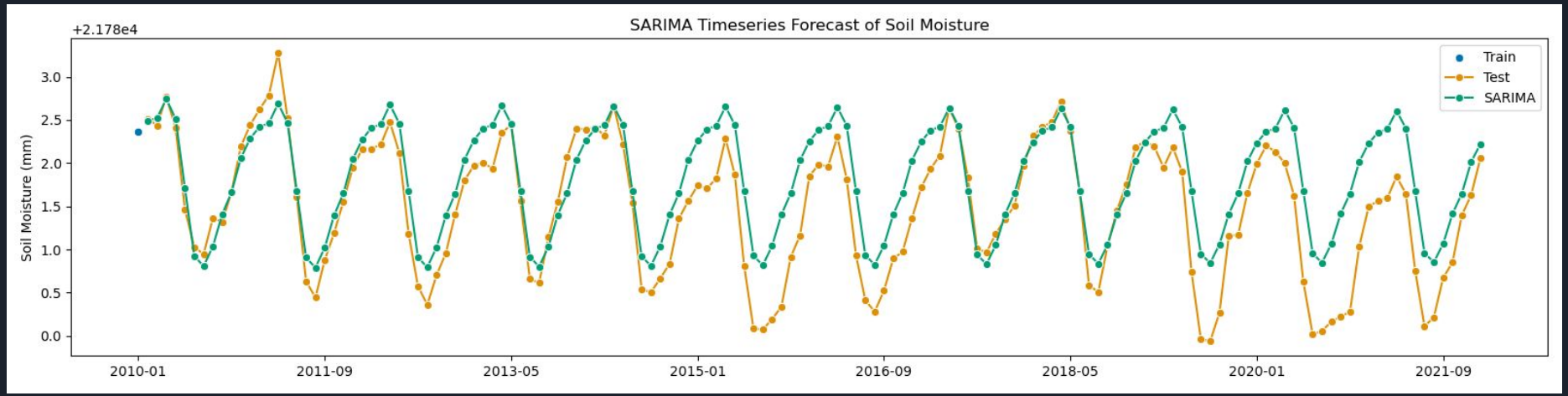
- **Models Used:** SARIMA, Holt-Winters

- **Metrics:** RMSE (mm of soil moisture in column), R2

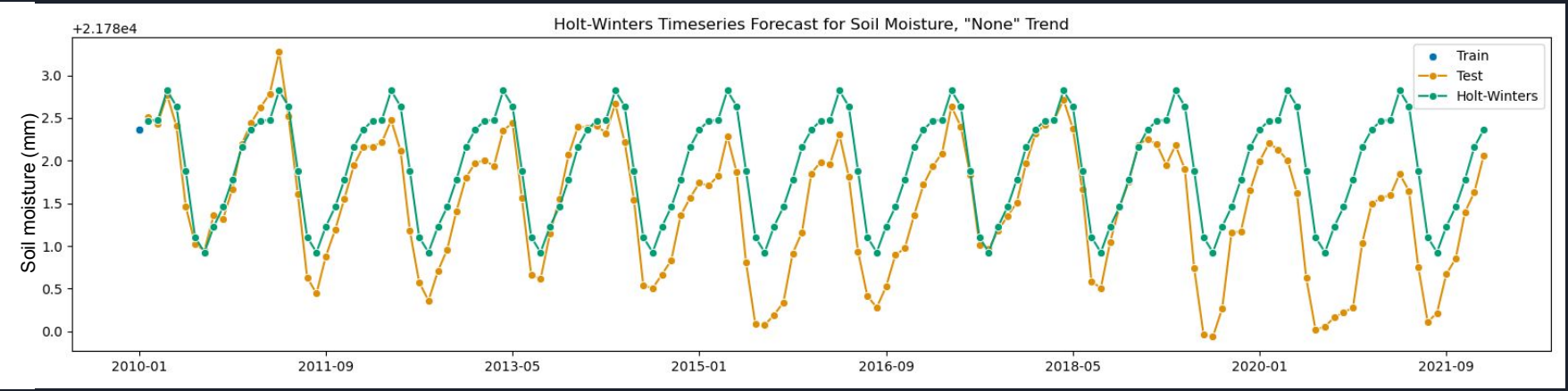
Model 3 Train/Test Split



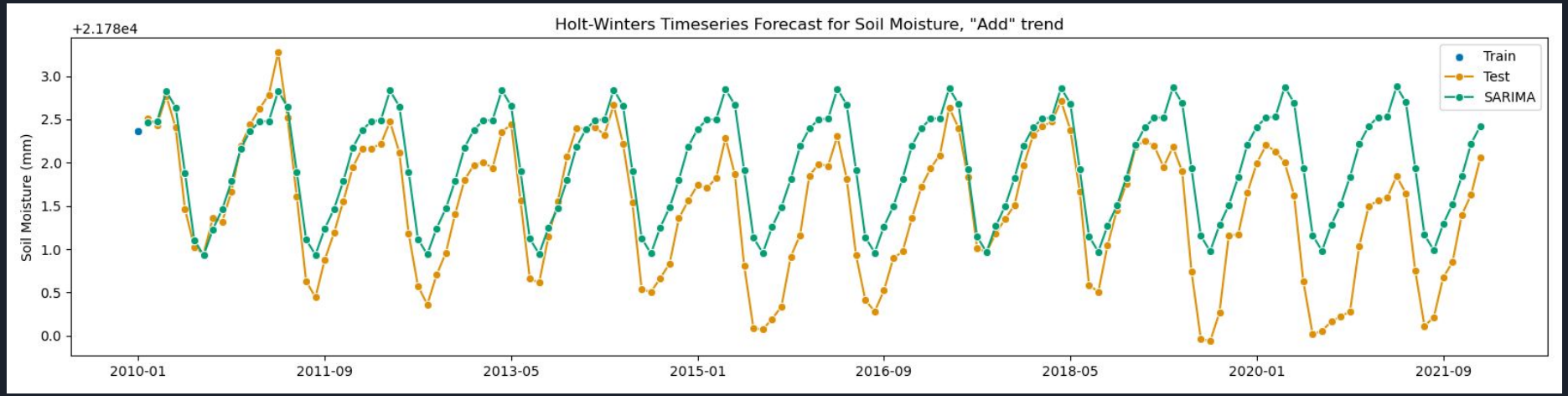
First model: SARIMA



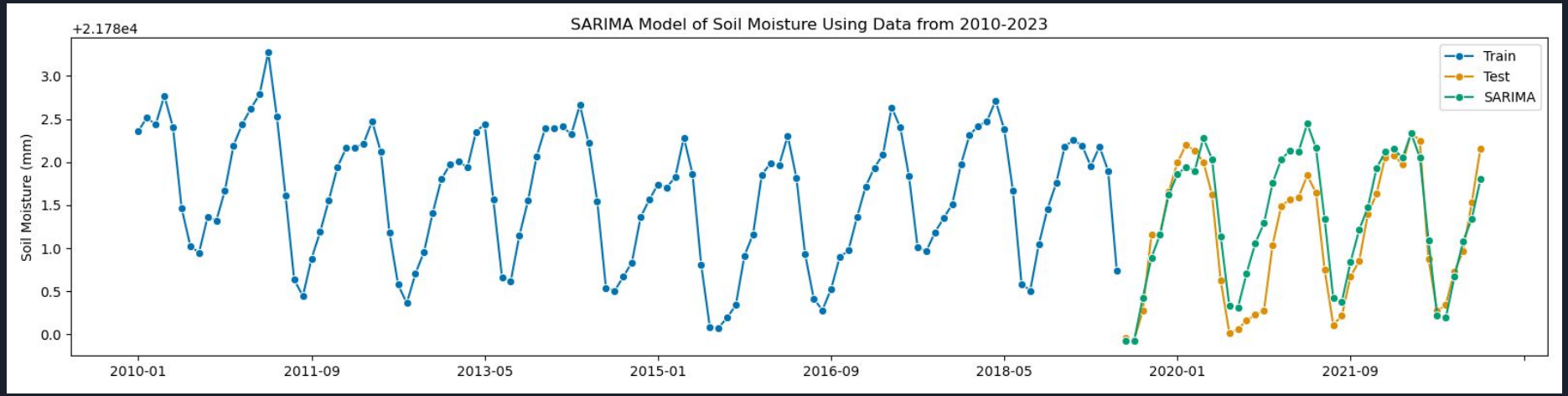
Second model: Holt-Winters



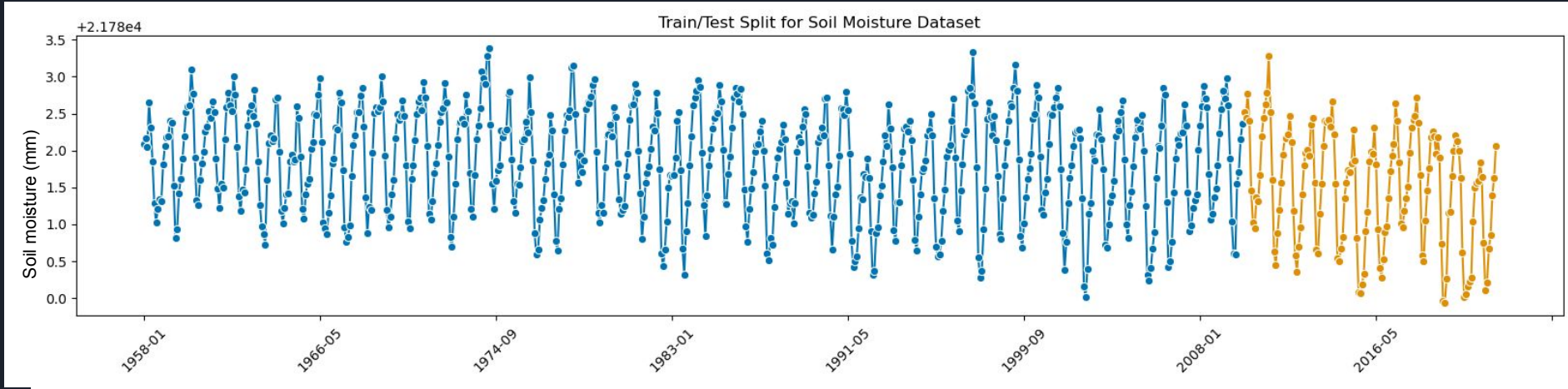
Third model: Holt-Winters, trend = ‘add’



Fourth model: SARIMA 2010-2023



Soil Model Conclusions



Which Model is the Best Predictor?

- Image Classification CNN has extremely high accuracy at 95%. Shows promise for long-term modeling with enough drought image data.

- For climate and environmental features, SARIMA time series models showed the lowest RMSE of ~0.04 inches for precipitation and ~0.38 mm for soil moisture.

Conclusions & Future Directions

- Long-Term Project ($$$): Invest in labeled satellite image data of droughts in different severity and train a multi-class CNN.

- Short-Term Project ($): Invest in more recent, smaller climate and environmental datasets (i.e. precipitation, soil moisture) to train SARIMA forecasting models.