Predicting droughts using climate data and satellite imagery

Yujie Cai, Michael Fein, Thee Ngamsangrat, Jim Zhang
Institute of Applied Computer Science, Harvard University, Massachusetts, USA

PROBLEM

Drought is one of the leading causes of humanitarian disasters around the globe. Due to the lack of a long-term drought prediction system, in most drought events, by the time aid arrives, irreversible damage has already occurred.

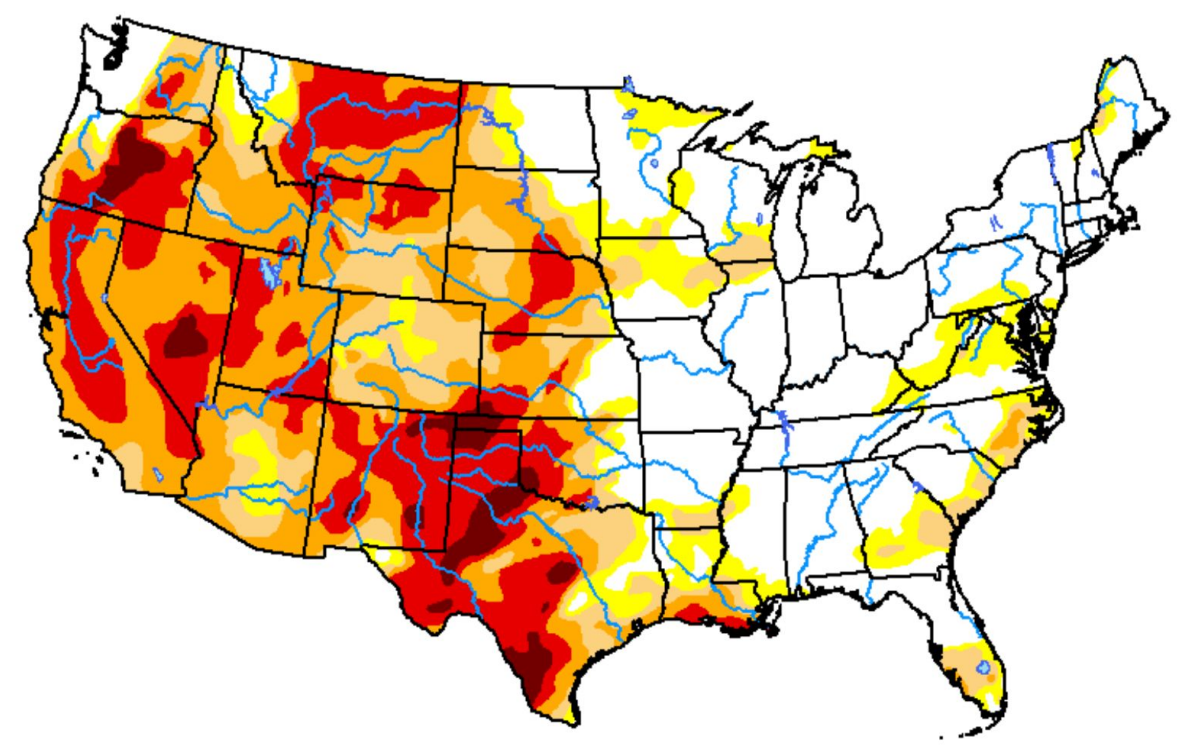
In this project, we design **a long-term drought prediction system** that takes in earth surface climate data and low orbit meteorological satellite image to predict the binary **onset of drought in the continental US** with an **80% accuracy over 6 months.**

DATA

Data sources:

Label (Y):

United States Drought Monitor (USDM) drought index: A categorical index of the severity of drought within a county.



(a visualization of USDM drought index map)

Features (X):

National Oceanic and Atmospheric Administration (NOAA) weather station data: temperature, snowfall, and precipitation
Low orbit satellite data

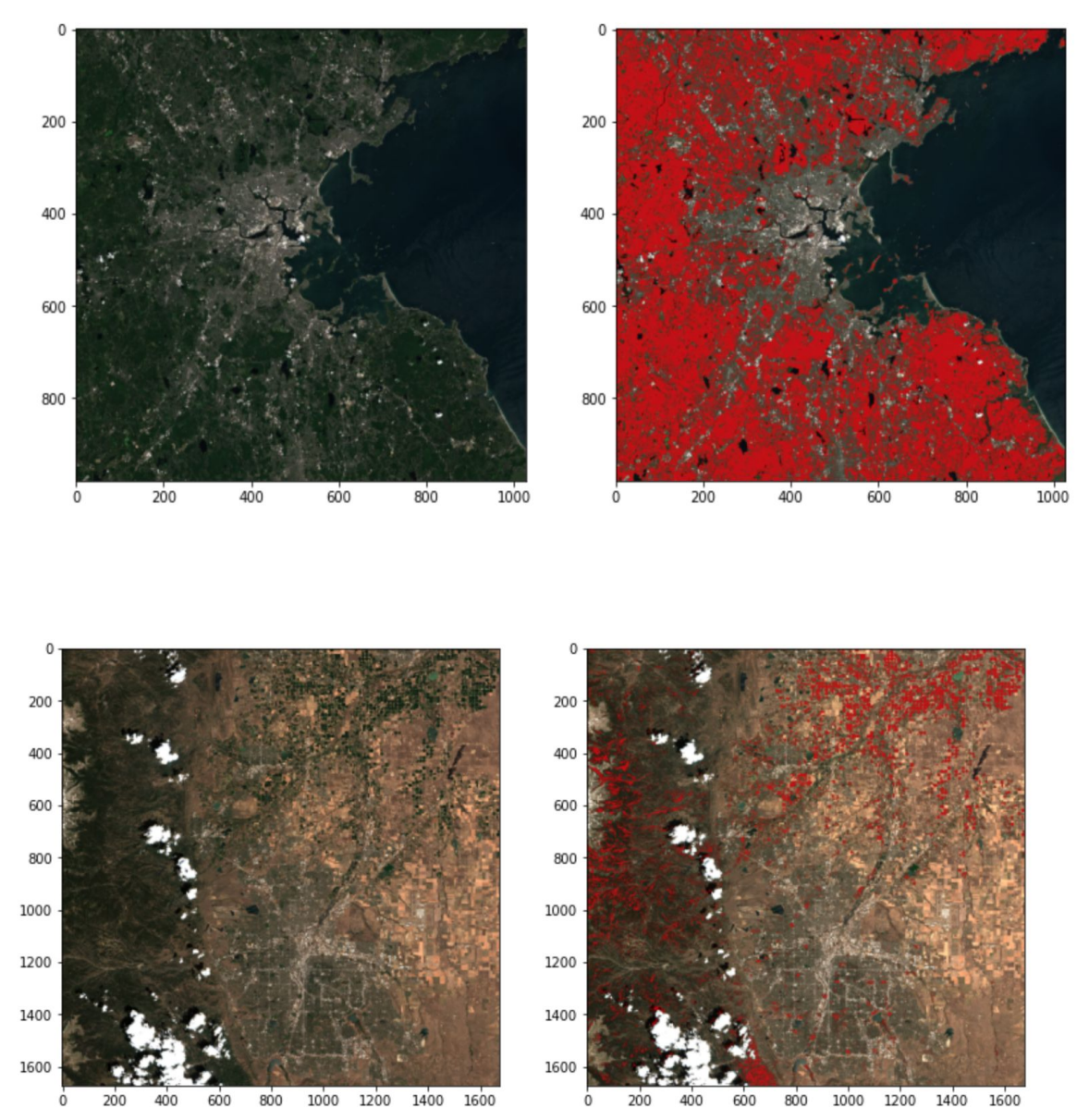
Sentinel L2A Satellite Images – multi-band 60m x 60m images

ECMWF ERA5 Evapotranspiration Data - evaporation and run-off data in 36km x 36km areas

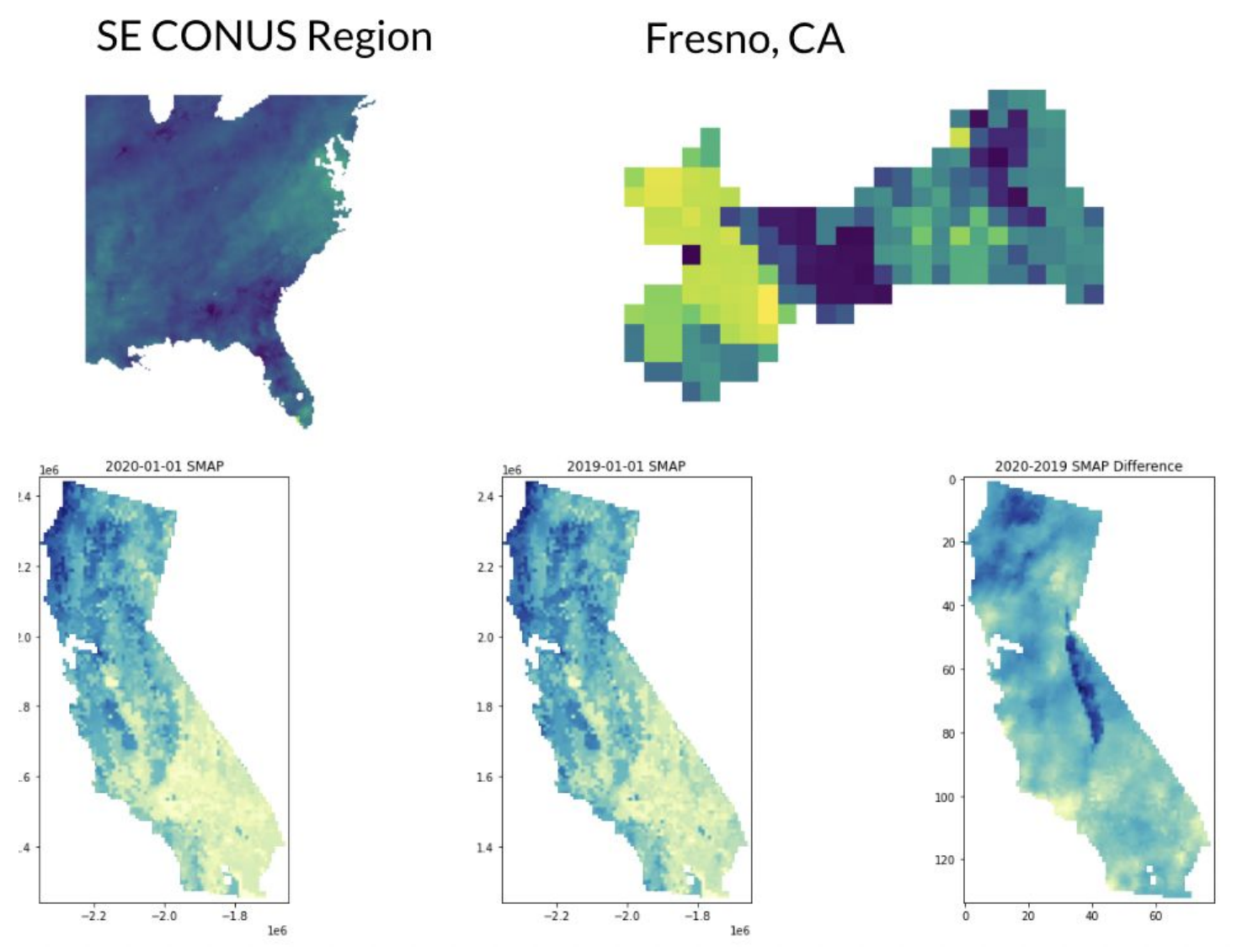
Soil Moisture Active Passive (SMAP) Satellite Measurements - soil moisture measurements in 9km x 9km areas

FEATURE ENGINEERING

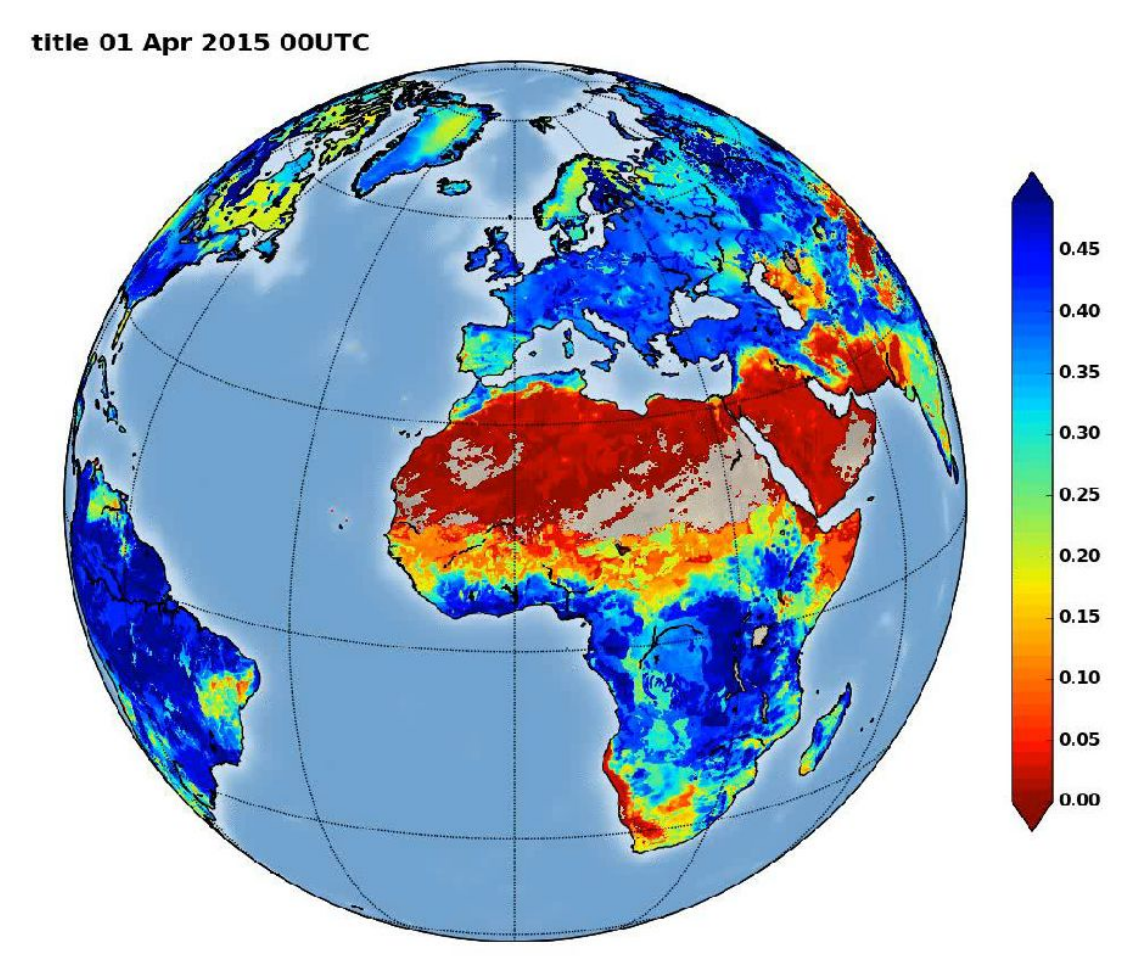
USDM Drought Index provides an artificial score for each county. The score is a weighted sum of scores in each drought class. We perform a binarization on this data: if a county receives 100% in “None drought”, it is labeled as 0, otherwise 1.
Climate data collected by weather stations on earth's surface are used directly
Satellite imagery:
- ECMWF ERA5 Evapotranspiration Data: we obtain the evaporation and runoff measurement near the central location of a county
- Soil Moisture Active Passive (SMAP) Satellite Measurements: we obtain the relative soil moisture level of the whole county and calculate an aggregate score for the county
- Sentinel L2A Satellite Images: we calculate two index to measure the absolute value and change in vegetation, which is usually highly correlated with rainfall potential and land water storage capacity.



coverage (red) for Suffolk, MA and Broomfield, CO

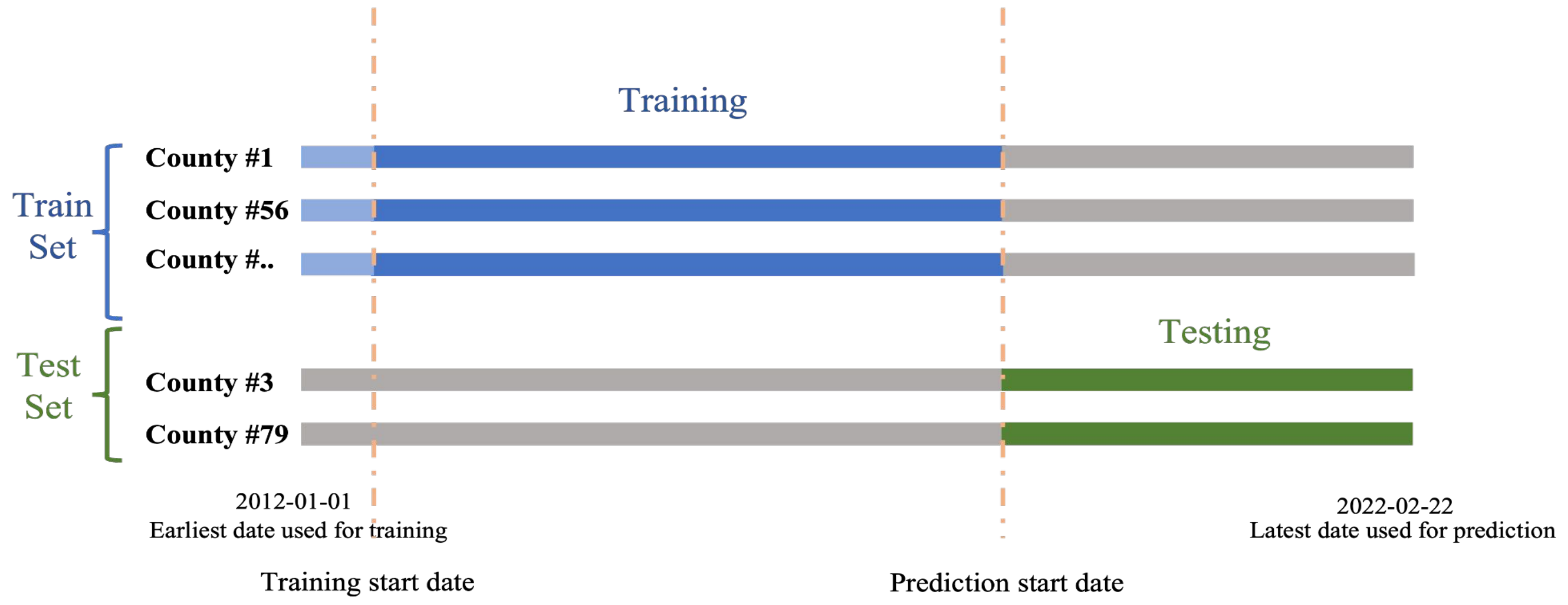


Visualization of SMAP data for California and Fresno county



Visualization of SET data for the globe
Copyright Joaquin Muñoz Sabater

MODELS



- Sample 50 counties from each climate region from a total of 3,143 continental US counties
- Train-Test split step 1: 80-20 spatial split
- Train-Test split step 2: 80-20 temporal split
- Double split to ensure no information leakage

Model	Specification
Logistic Regression	Default, no regularization
Decision Tree	Max_depth = 10
Random Forest	Max_depth = 10
XGBoost	Default
CNN on top of Multitask MLP	2 CNN layers with 64 and 32 filters
LSTM	16 layers*

Logistic Regression(LR) Model: used as a baseline due to its simplicity and disinclination to overfit.

Tree-based Model: Decision tree (DT) offers great interpretability; Random forest (RF) combines multiple models with low correlation to increase generalizability.

Long Short Term Memory (LSTM): effective for capturing the time-varying state of weather-related trends. LSTM also offers a control over the flow of inputs and alleviates exploding and vanishing gradients.

CNN-Multitask Networks: used to capture the spatial relationship between features across times.

RESULTS & DISCUSSION

Model Comparison

Model	Accuracy	AUC	Feature used
LR	0.68	0.646	roff, evap, smap
DT	0.679	0.632	smap, prcp, snwd, tmax, tmin
RF	0.676	0.642	smap, evap, roff, prcp, snwd, tmax, tmin
XGBoost	0.706	0.694	smap, evap, roff, prcp, snwd, tmax, tmin
CNN	0.815	0.91	roff, evap, lagged Y
LSTM	0.756	0.809	All features

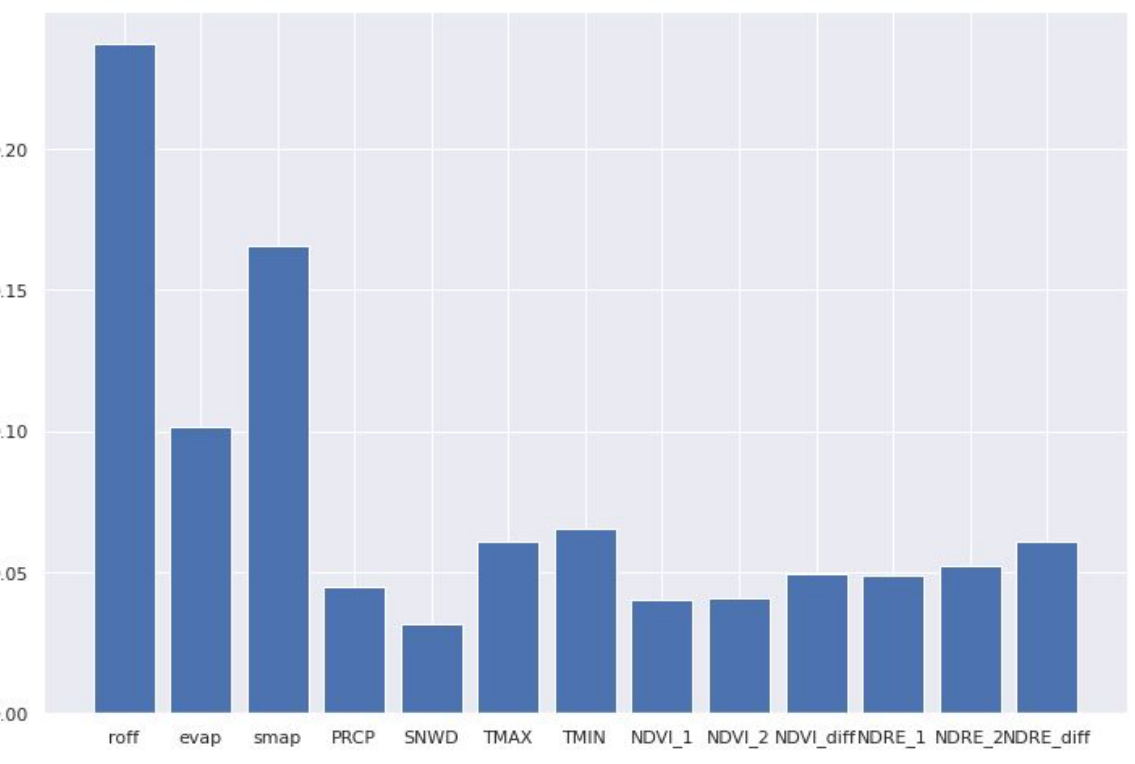
The better-performing CNN and LSTM models achieved 75 ~ 80% prediction accuracy on a six-month prediction horizon. In LR and tree-based models, longer prediction horizon or training lag tends to increase the accuracy and AUC.

Regional variability

Model	Prediction Week	History length	Start date	Accuracy	AUC
Northeast	8	26	2015-01-01	53.8	61.3
UpperMidwest	8	26	2017-01-01	79.6	51.3
OhioValley	8	4	2019-01-01	45	48
Southeast	52	4	2015-01-01	47.5	52.1
NorthernRP	12	8	2015-01-01	87.5	71.8
South	12	4	2017-01-01	72.5	63.6
Southwest	-	-	-	-	-
Northwest	4	26	2019-01-01	50.4	61.7
West	-	-	-	-	-

Due to data imbalance, regional variability in optimal prediction week, training lag, start date exists.

Feature Importance



Feature importance plot from decision tree model shows that **evaporation**, **runoff**, **soil moisture** are the three most predictive features.

CONCLUSION

Our project suggests that long-term droughts can be meaningfully predicted and provides drought relief agencies with a longer prediction horizon.

For future steps, we hope to predict the severity of droughts in a more granular way.

REFERENCE

Hao, Singh, V. P., & Xia, Y. (2018). Seasonal Drought Prediction: Advances, Challenges, and Future Prospects. Reviews of Geophysics (1985), 56(1), 108–141. https://doi.org/10.1002/2016RG000549

Poornima, & Pushpalatha, M. (2019). Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units. Atmosphere, 10(11), 668. https://doi.org/10.3390/atmos10110668

QUANTUMBLACK
A MCKINSEY COMPANY

Harvard John A. Paulson School of Engineering and Applied Sciences