**Drought Prediction Across the Continental United States through Statistical and Machine Learning Methods**

Yujie Cai, Michael Fein, Thee Ngamsangrat, Jim Zhang,

Institute of Applied Computer Science, Harvard, Cambridge, United States of America, equal contribution.

May 4, 2022

## **Abstract**

Drought has become one of the leading causes of humanitarian disasters. Drought leads to water shortages, wildfires, loss of crops and livestock, and more. Post-drought aid is time-consuming, and due to the inability to perform long-term drought prediction, irreparable damage has already been done by the time aid has arrived. This paper attempts to extend drought prediction accuracy past currently accepted capabilities using statistical, and machine learning approaches based on several feature variables. Our work also addresses a lack of substantial research into modeling types for drought forecasting. We successfully improved the prediction power of drought forecasting with more advanced model architectures. This work could allow for a more accurate allocation of emergency response resources before the onset of a drought.

## **I. Introduction**

### **1.1 Understanding Drought and Current Tools for Predicting Drought**

Drought is one of the most costly and environmentally damaging climatic events. Better predicting when and where drought will occur is of utmost importance to aiding those affected by drought and understanding how to mitigate oncoming droughts potentially. A prediction period of greater than even a month would allow significant mitigation of drought effects (Ma, 2017).

From a scientific perspective, droughts are difficult to measure with unclear beginnings and ends (Vicente, 2010). Droughts also tend to be slow, developing over weeks and months rather than days or hours, like other natural disasters (Ma, 2017). Defining drought, while not the main topic of this work, is an essential aspect of predicting its onset and effects. It is important to note that drought is not simply a lack of available water since different ecosystem types will naturally experience different precipitation levels; instead, drought is a deviation from the average amount of available water in specific climate regions. We use the U.S. Drought Monitor, which synthesizes expert vetted information, to classify types of drought risk across the United States (USDM, 2022). The USDM uses the Palmer Drought Severity Index, the Standardized Precipitation Index (SPI), and other climatic inputs as well as soil moisture indicators. The figure below shows a recent sample of drought classification across the Continental United States (CONUS).



Due to its relative acceptance as an accurate categorization of the risk of drought severity and its aggregation of different drought indices, we use the USDM as our response variable throughout this work.

NOAA outputs weekly seasonal drought outlooks based on available climate information and historical indications of drought, mainly precipitation and temperature, through its climate prediction center (NOAA, 2022). Based on their published work, their forecasting methods are based on historical data combined with the USDM drought risk index and have increased accuracy steadily over time. Most research thus far has employed empirical methods to attempt forecasting through a set of initial conditions correlated with outcomes (Ma, 2017). As we later discuss in section II, recent work has been done in applying more advanced computational methods to drought prediction, partly because of its complexity as a natural event and partly because of the computational cost of more advanced methods involving large data sets.

In this work, we aim to more accurately predict drought in longer-term periods by applying two main advancements to currently available tools: (1) more significant computational and data complexity through the aggregation of data sources and use of satellite imagery and (2) regionalizing the CONUS into distinct climate regions to understand both how predictability changes across climates and better predict drought on a spatially larger scale. We use a combination of statistical and computational models with multiple X predictors to work towards these goals.

### 1.2. Related Work

Drought remains one of the least understood natural disasters due to its convoluted causal factors and occurrence mechanisms (Sundararajan, 2021). Recent progress has resulted from either a hyper-localization of prediction or a shortened prediction timeline. We review both of these related work fields below and their pros and cons.

Prior work in drought forecasting has focused on specific regions of an area, such as Shaanxi, China (Zhang et al. 2019) or New South Wales, Australia (Dikshit et al. 2020). These papers have demonstrated successful drought prediction accuracies of 90% or greater through machine learning techniques. While successful, however, this work lacks generalizability to more significant regions diverse in their climatic, geographic, and hydrologic features. In this paper, we attempt to predict both within climate regions and across the nation to improve the generalizability of our models.

Other work in drought prediction has accurately predicted drought across more significant regions but over shortened time horizons. For example, in Narapusetty et al. (2020), logistic regression predicts drought onset in CONUS with 1-week time horizons. While successful and ultimately quite similar to our work here, a 1-week look-ahead prediction is not nearly as valuable for drought preparedness as a 2-6 month look-ahead forecast as our work attempts.

Most recent work has introduced computational complexity, rather than statistical complexity, to the problem of drought prediction to improve accuracy. Some more recent examples, such as Poornima and Pushpalatha (2019) and Dikshit et al. (2021), use recurrent neural networks and long short term memory (RNN & LSTM) to study the time series drought indices. This approach has only recently been possible, as Hao et al. (2018) points out, due to expansions in available data. For example, models have improved over time by utilizing factors such as large-scale climate indices, local climate variables, and initial land conditions from advances in sensor technology and computing power.

Thus, our work in the field of drought prediction is novel. We aim to generate a generalized model applicable to a wide area full of diverse climate regions and can forecast further into the future. Our modeling methodology has us test a wide range of statistically and computationally complex models using many observations from a wide range of available data.

### 1.3 Note on Complexity of Drought Forecasting

Current literature suggests that the causes of droughts are rooted deeply in the specific topographical and geographical conditions of an area. Zhang et al. (2020) suggest that the stagnant development of drought forecasting is that the field has not agreed on a unified theoretical framework to explain the causal mechanism of droughts happening across the world or even within the same continent or sub-region.

The factors that shape drought climates are convoluted and involve the interaction of multiple factors (Salehi-Lisar, 2016). Existing work tends to analyze drought in a case-by-case fashion (Marengo, 2021), limiting the scope of work on single droughts in narrow time windows and small geographic regions, typically less than a year.

Our work divides the CONUS into generally agreed-upon climate regions, which might have varying causal mechanisms of drought as suggested in the literature. Through this method, we hope to more accurately capture differences in regional causes of drought and assess the predictability of drought based on our available data across different climate types. Below is a depiction of the climate map of the CONUS provided by NOAA and developed by Karl. and Koss (1984).

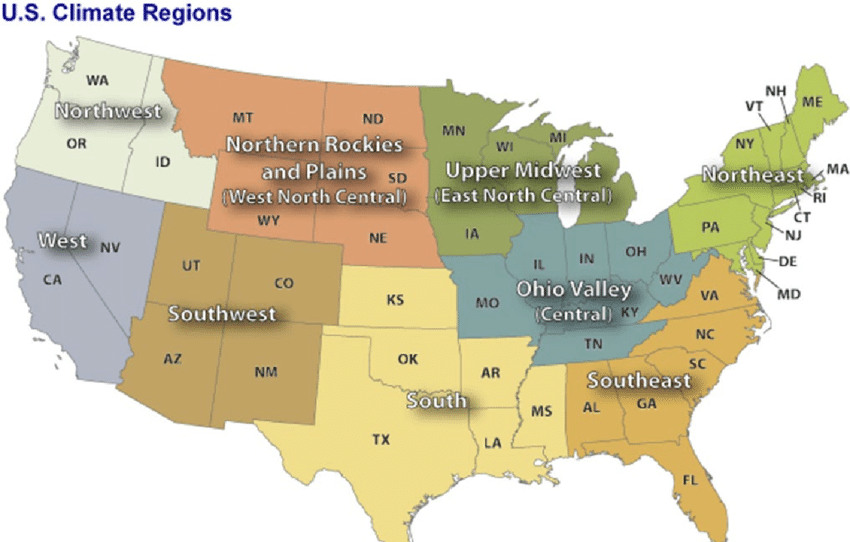


Figure 2. US Climate Regions

## **II. Problem Formulation**

In this section, we formulate the problem of drought prediction and explain why we build our model with continental US data and define this problem as a binary classification prediction problem.

Our motivation is to design an accurate, long period, and generalizable model which will not be limited to a particular climate region for drought relief organizations and government agencies. To develop such a drought prediction model, we need to study an area on the earth's surface that is sufficiently large and developed enough to have the ability to monitor drought at the continental level. Therefore, we choose the United States to study the drought prediction problem. First, the continental US is large enough to encompass nine climate regions. Suppose accurate drought predictions can be made on the whole continental US. In that case, this will show that our model may have strong generalization capabilities and can be applied to other countries and regions on earth. Second, among all drought data we found globally, the Drought Severity and Coverage Index (DSCI) provided by the US Drought Monitor project (USDM) – an index that measures the area of land within each county that is indifferent classes of drought – has the highest coverage and the neatest organization. We also surveyed other regions and countries with large land areas and major drought problems, such as Sub-Saharan Africa, Australia, the western part of China, etc. None of them has a national or continental level drought monitor system as good as USDM.

However, comprehensive data coverage does not guarantee high quality at all data points for machine learning. Drought research has been facing a severe challenge - drought lacks a well-defined numerical boundary that is sharp. In contrast to rainfall, temperature, and surface evaporation rates, substantial quantities that can be measured precisely by instruments, drought is dynamic and complex. What is traditionally referred to as drought means the inability to obtain "sufficient" water. "Sufficient water" is often related to the population, farmland, and other objects with water needs. In short, the definition of drought is highly dependent on the subjective perception of human beings. Unfortunately, DSCI, as a drought measurement, also contains a considerable amount of subjective perception.

Also, our extensive EDA on the DSCI dataset, including linear regression, PCA, and TSNE clustering, found that the raw numerical DSCI index is very noisy. The way this data is aggregated results in a poor signal-to-noise ratio for our learning purposes. For example, when 50% of a county's land area is in "Abnormally Dry" (which is not even considered an onset of drought), its DSCI may be equivalent to another county with 10% of its land area in Exceptional Drought (highest class for drought). Only the latter is the kind of drought we want to predict. So, building a model on such data is not well aligned with our motivation, "drought disaster relief." Luckily, USDM also published break-downed drought data, based on which we will define our learning task.

Based on the "NONE" data provided by USDM for each county (i.e., the area of land in each county that is not in any degree of drought, thus NONE), we defined the label (Y). When NONE=100, i.e., when 100% of the land in the county has strictly no drought of any degree, the label is marked as 0. Otherwise, any degree of drought is treated as a drought occurrence and is labeled as 1. Based on this formulation, we build binary classification models with various visual and numerical features to achieve an accuracy as high as possible.

## **III. Models and Methodology**

### 3.1 Data Preprocessing

#### 3.1.1 Available Data

For our drought risk forecasting and prediction purposes, we use data from several other sources to improve the longer-term predictability of drought onset. These data sources include:

* United States Drought Monitor (USDM) drought risk index
* National Oceanic and Atmospheric Administration (NOAA) Weather Station temperature, snowfall, and precipitation
* Sentinel L2A Satellite Images - 60m x 60m images
* SET Evapotranspiration Data - evaporation and run-off data over time
* Soil Moisture Active Passive (SMAP) Satellite Measurements - soil moisture measurements in 9km x 9km areas

We combined all of these data sources on a weekly time scale at the county level. We acknowledge that NOAA, SMAP, SET, and Sentinel L2 vary across locations and spaces. For example, the temperature and precipitation from NOAA are measured by weather stations, and those weather stations are not necessarily aligned with individual county boundaries. Datasets do not all begin at the same point in time. SET data, for example, began in 2015, while USDM data began in 2012. As we review in our methods section, we dealt with variations in data availability through imputation and experimenting with maximizing predictability.

#### 3.1.2 Sampling and Train-Test Split

As mentioned above, we have county-level data for 3,143 counties across the whole CONUS. Suppose we pull climate and satellite data from each county for ten years (531 weeks), which requires much work. Therefore, we apply stratified sampling according to the climate region division. We sample 50 counties from each region out of 450 counties. We argue that our sample is representative of the CONUS, and we test that counties in each region can be viewed as independent entries (no information leakage).

We build our machine learning model using a train-test split, as demonstrated in the figure below, to test our model's robustness across time and space. As per the general process for dealing with cross-sectional data, we first take 80% of the counties to form our training set and stratify them by climate region. Then we can subdivide training and testing periods by specifying the start date of training and the start date of prediction (testing).

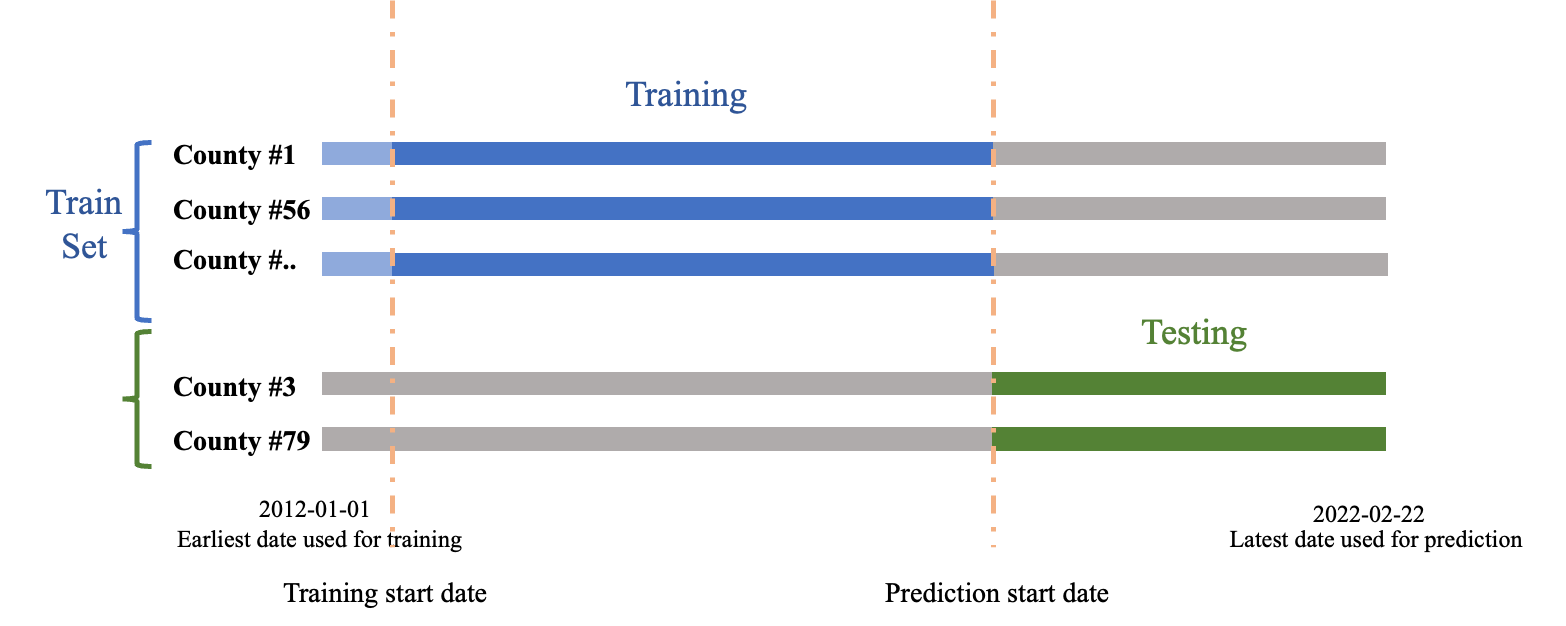


Figure X. Train-test split

#### 3.1.3 Drought classification (YC)

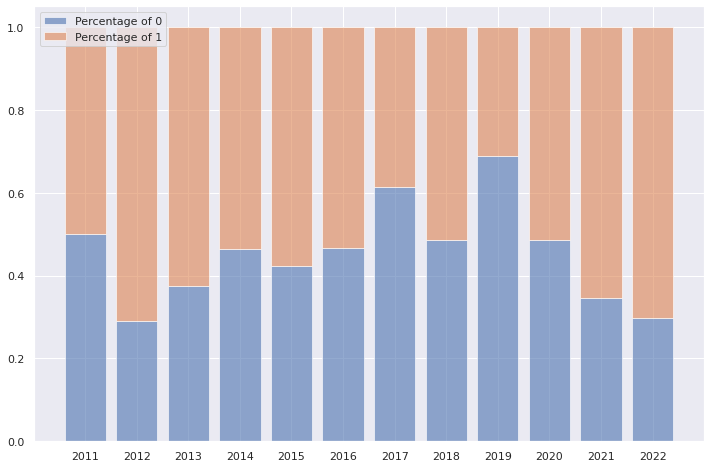
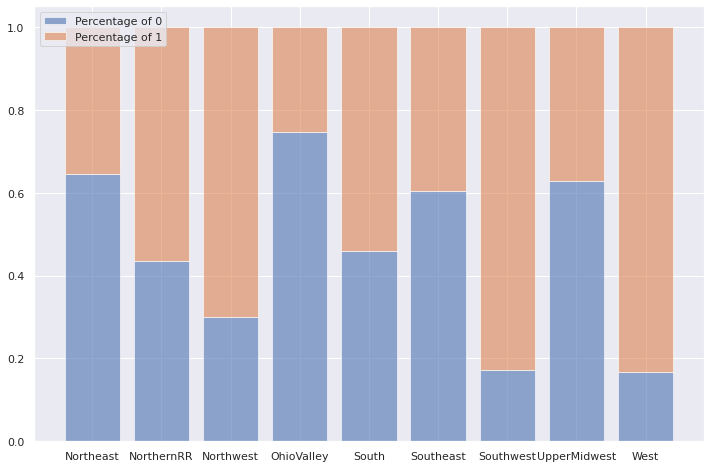
As mentioned in section II, we use "NONE" data provided by USDM for each county to define our categorical target variable "has drought (1)" and "no drought (0)". The USDM dataset also includes multinomial classification, i.e., levels D0, D1, D2, D3, and D4, to indicate the different severity of drought. However, our primary task for this project is to identify whether there will be drought or not in a locale. Before diving into building prediction models, we first examine the temporal and spatial balance of the target variables. 

Figure X: target variable distribution across year (left) and region (right)

In the left plot of figure X, we do not see a muscular class imbalance across years. It suggests that we do not have to worry too much about the time-fixed effect. We see a robust regional imbalance on the right, especially in the Southwest and West region; less than 20% of the targets are "no drought." The imbalance is expected, in any case. When we fit regional models later, we shall be aware of the lack of data in these regions.

#### 3.1.4 Obtaining features from SMAP and SET

Both SMAP and SET are more accessible features to obtain. To extract soil moisture measurement from the SMAP dataset, we used an API provided by George Mason University, which returns the soil moisture map within each county. Each pixel of returned images stands for the moisture level at 9 x 9 km resolution. We then aggregated pixel values in each image and calculated the moisture index.

ECMWF, the publisher of the SET dataset, also provides a python package called Metview as an official tool to load, extract, and roughly analyze the data. We obtain evaporation and runoff data at 36 x 36 km resolution at the center of each county.

#### 3.1.5 Vegetation feature extraction (Thee, please edit that)

Sentinel L2A contains Bottom of the Atmosphere (BOA) global satellite images starting in January 2017. The satellite revisits earth every five days with an image resolution of 10-60 meters, dependent on one of thirteen bands. Those bands that we identify as relevant to our prediction tasks include coded color (RGB), contain a red vegetation edge, NIR, and cloud probability and mask bands. We combine these bands and compute indices, including the Normalized Vegetation Index (NDVI) and Normalized Difference Red Edge index (NDRE), using the formula below. These indices are primarily constructed to detect the rapid change in vegetation reflectance in the electromagnetic spectrum's near-infrared range. We use them to accurately identify vegetation areas in images, as shown in Figure 1. This method allows us to measure the percentage of vegetation in a given image. We then calculate the change in the size of vegetation area over time and incorporate this feature as a predictor to predict the drought risk.

However, quantifying the change in vegetation area over time is non-trivial due to the inconsistent cloud coverage between different images. This problem is best illustrated in Figure 2, where the two images are taken at the exact location but one year apart in time. The amount and spatial location of cloud coverage in the two images are very different, making it difficult to measure the percentage change of vegetation area between them. We tackle this problem by using the cloud probability band to detect pixels likely to be covered by clouds in either image and then ignore those pixels. We then compare the percentage of vegetation areas in the remaining pixels in the two images. Using this method, we can generate a weekly percentage change of vegetation area in each county over the past five years. We will incorporate this feature to improve our predictive model. We also plan to use a pre-trained model (Cordeiro et al., 2020) to detect the water area in the images and extract the changes in water percentage over time.

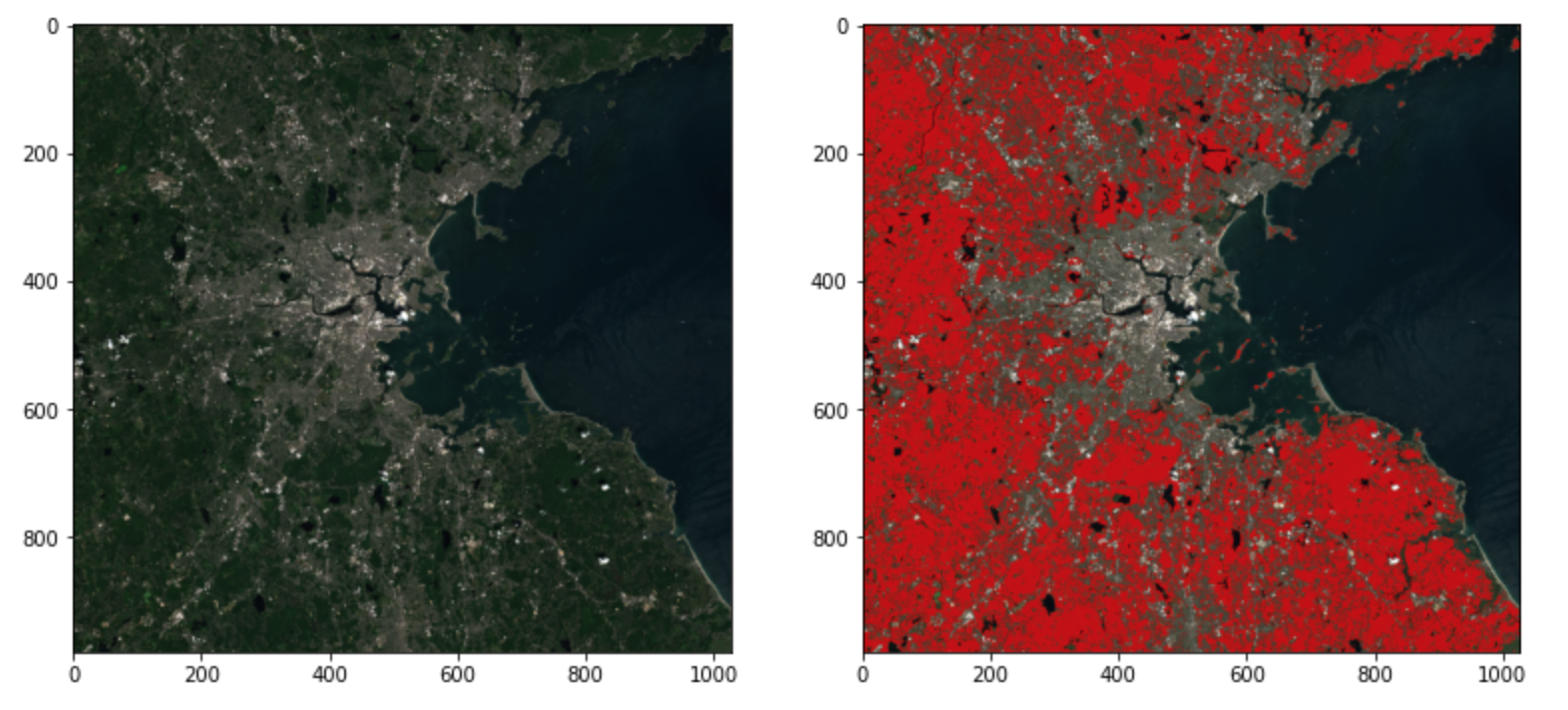


Figure X: Suffolk, MA 2020-07-15, Highlighted by Red when NDVI > 0.8 and NDRE > 0.65

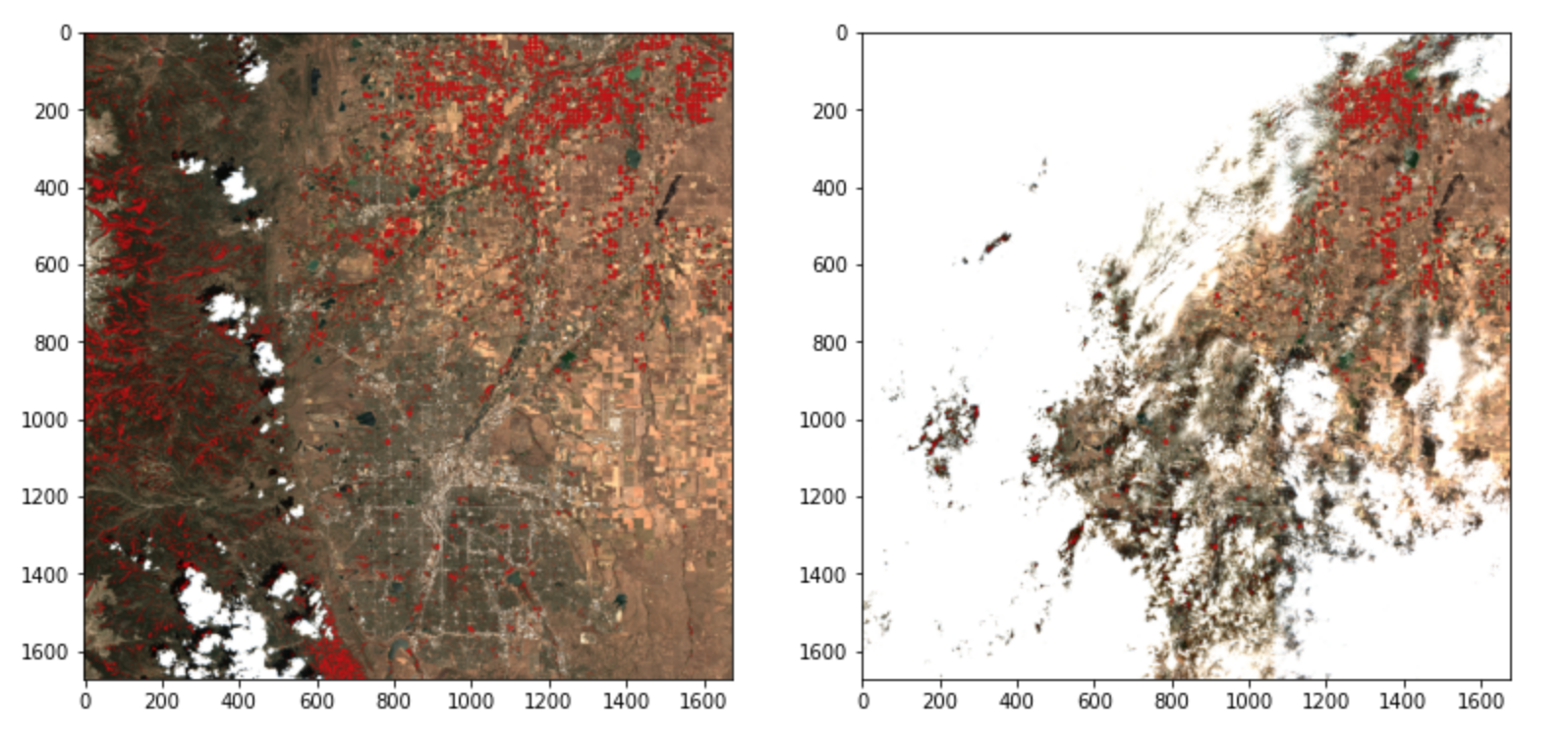


Figure X: Broomfield, CO in July 2020 and in July 2021

### 3.2 Modeling

| **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 + Dense |
| LSTM | 128 uni-directional LSTM + 2 layers of CNN + Dense |

Table X. Model Summary

*Logistic Regression (LR).* We are predicting categorical variables in the drought prediction context. After preprocessing the data, we assign 0 to be counties with no drought in a given week and 1 to be counties with any drought in a given week. Multinomial classification is omitted. We first implemented a logistic regression model because it is easy to implement and interpret.

*Tree-based model. Decision Tree (DT) and Random Forest (RF)*. A natural progression for making categorical predictions is to implement tree-based models. DT offers great interpretability; RF combines multiple models with low correlation to increase generalizability. In the project, we implement both decision trees and random forests. To improve the performance (Mehr et al., 2020), we use XGBoost with default parameters.

*Long Short Term Memory (LSTM)*. LSTM is a promising candidate model for this prediction task because it can capture the time-varying state of weather-related factors and trends. Compared to a plain Recurrent Neural Network (RNN), LSTM offers more control over the flow and mixing of inputs. It also alleviates the exploding and vanishing gradient problems when fitting the model.

*Convolutional time series multitask model.* We develop a convolutional neural network to find the mapping between previous climates and the onset of drought in the future. We use 52 weeks of climate history to predict the onset of drought in the future 26 weeks. Since we use 78 weeks of 3 features, each data point in the feature space can represent an image of size 26x12. Furthermore, for such training data, it was natural to think of applying a CNN model to it. We built our DroughtNet CNN with 356,298 trainable parameters. We lay two 2D convolutional layers on a deep and densely connected net.

### 3.3 Metrics

We choose MSE as our evaluation metric for time series modeling due to its simplicity and ease of comparison with the rest of the literature. We also consider R-squared but discard it due to its susceptibility to influence by the number of observations and predictors. We mainly choose accuracy, AUC, precision, and recall graph for linear prediction. Precision is the fraction of relevant instances among the retrieved instances (TP / TP + FP), while recall is the fraction of relevant instances that were retrieved (TP / TP + FN).

## **IV. Result and Discussion**

### 4**.1 Results and comparisons**

| **Model** | **Best Accuracy** | **Associated AUC** |
| --- | --- | --- |
| Logistic Regression (LR) | 0.680 | 0.646 |
| Single Decision Tree (DT) | 0.679 | 0.632 |
| Plain Random Forest (RF) | 0.676 | 0.642 |
| XGBoost | 0.706 | 0.694 |
| CNN | 0.815 | 0.910 |
| LSTM | 0.801 | 0.875 |

Table X. Prediction results

We implemented two categories of machine learning models to model the real-world drought problem. Since we predict a categorical target, logistic regression and tree-based model serve as reasonable methods to start the experiment. Putting other experiment parameters, such as the prediction horizons and the features we used aside for later discussion, the most accurate results from these models are around 0.6-0.7, with the boosting random forest (XGBoost) being the best one getting 0.706. The associated AUC ranges from 0.63 to 0.69. The boosting model produces the best result. The boosting random forest model could predict drought and no drought with 70% accuracy and can distinguish between two classes (true positive and true negative) with 69.4% chances. The real improvement is achieved when we switch to deep learning methods. When implementing an LSTM network, the accuracy is boosted to 0.801, and the AUC increases to 0.879. Using a CNN on the top of a multitasking MLP, the highest accuracy we get is 0.9, with an excellent AUC of 0.815.

Results show the usefulness of introducing advanced learning models to drought prediction problems. The logistic regression and the tree models are expected to be stuck in the mediocre accuracy range. We conjecture that the dataset is too big and noisy, so the simplest models can only tell the big picture but not enough to explain detailed variations. In other words, a very slight variation across counties or time could affect the metrics. However, with deep neural networks, though the interpretability is low, we achieve better results in all metrics. We shall ask ourselves whether we prefer the prediction accuracy/precision or interpretability for the drought prediction case.

### 4**.2 Prediction Sensitivity in Time Length**

The temporal hyperparameters are critical for us to experiment and understand mid-to-long-term drought prediction. For each type of model, we need to know how far into the future we can make accurate predictions based on limited information and how far backward we need to go to make accurate predictions for settled prediction horizons. In addition, we want to know the time horizon of our training period to ensure higher performance.

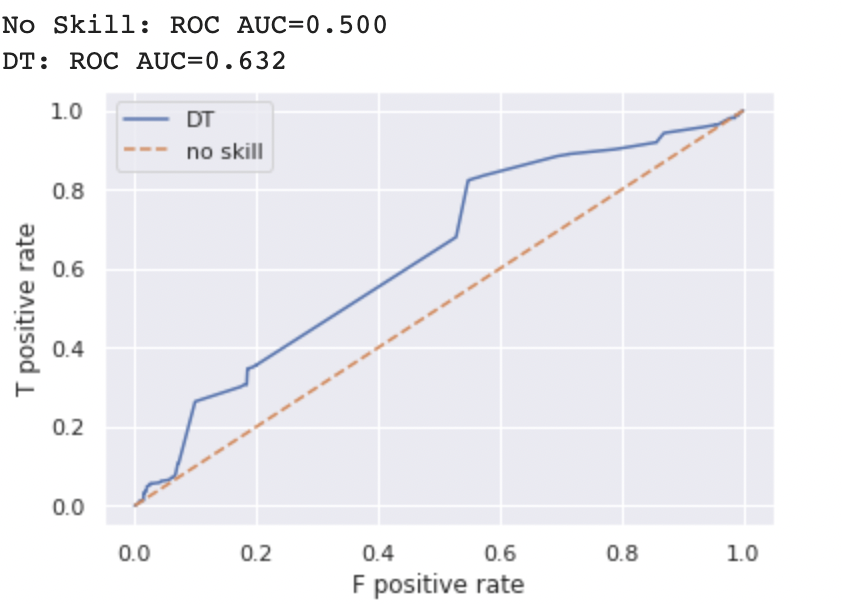
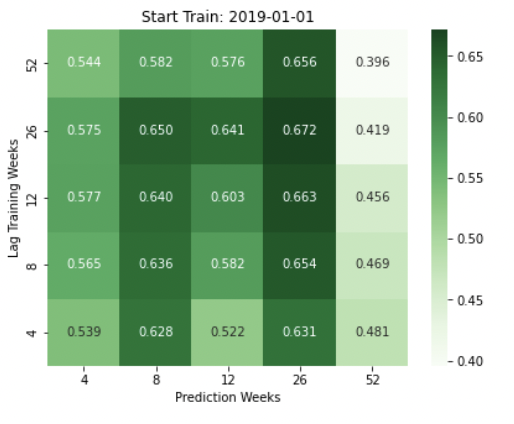


Figure X. left: Prediction and lagging horizon heatmap - DT example, start training from 2019-01-01; right: AUC result for DT example of best accuracy

In the experiments, we focus on varying three temporal hyperparameters. The prediction week defines the Length of time we are predicting the future. The lag training week corresponds to how far backward we take for predicting the first week in the forward prediction. Also, we vary the time of the data inclusion, that is, the start train variable. The hyperparameter tuned for the best results (accuracy, Table X) is shown below (Table X).

| **Model** | **Prediction Week (Best)** | **Lag (Best)** | **Start date** |
| --- | --- | --- | --- |
| LR | 26 | 8 | 2017-01-01 |
| DT | 26 | 26 | 2015-01-01 |
| RF | 26 | 26 | 2015-01-01 |
| XGBoost | 4 | 12 | 2015-01-01 |
| CNN | 26 | 52 | 2015-01-01 |
| LSTM | 26 | 52 | 2012-01-01 |

Table X. Prediction and Lagging Horizon

Five out of six models suggest that the best accuracy happens when we predict 26 weeks. XGBoost can predict 4-week the best. For lagging, most models need mid-term to long-term lagging to get the best predictions. Four out of six models include the data for training from 2015-01-01. In addition, LR starts the latest, and LSTM can include all available data for the prediction task.

So, how far should we go into the future? In general, as we go further into the future, we decrease the predictability across all types of models, which is manifested through our example in Figure X in the prediction accuracy. For the same lag, we observed that the longer the prediction weeks horizon in logistic and tree-based models tends to increase prediction accuracy. The observation matches Nourani and Molajou’s (2017) results that the decision tree model gets better results in seasonal (3-month) and half-year (6-month) predictions, which corresponds to our 12-week-ahead, roughly 26 week-ahead predictions. This probably suggests that the training data from a set period contains much information. The past is highly correlated, so mid-to-long-term predictions (12 to 26) weeks yield better results. Nevertheless, notice that by applying boosting on random forests using the XGBoost algorithm, we can improve the performance of short-term prediction to 4-week ahead. Nevertheless, the mid-term prediction now does not outperform the short-run ones.

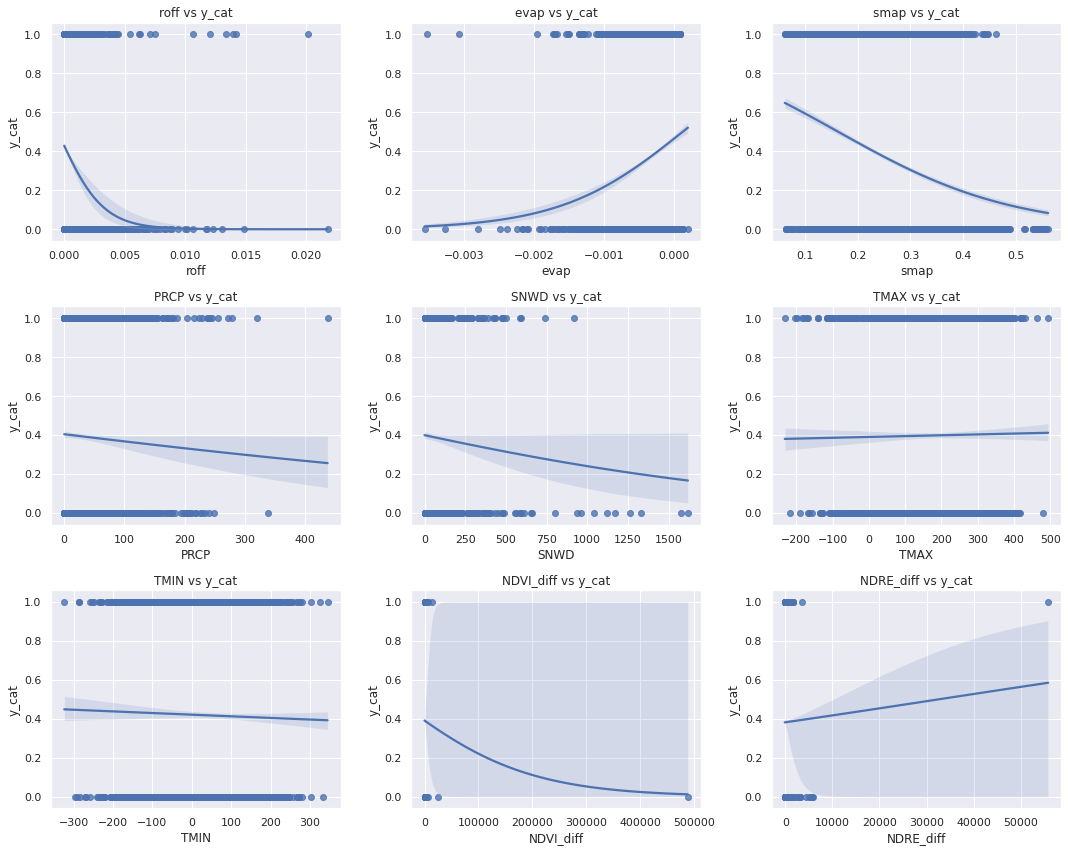
However, the pattern is not so apparent in the deep learning case. For LSTM models, when looking at the heatmap, we see that for the same lag, the prediction accuracy drops when we are predicting more weeks. There is no significant increase in the mid-term range. Partially, it might be that the predictors (features) differ across deep learning models and other models.

Then, how far should we go back? Fixing the prediction week and vertically looking at the heatmaps, we find that we got better predictions using more negligible lags across different models, which means nearer past. Also, we observe that the majority of the tree-based models have better performance when lagging the feature to 12 to 26 weeks before. CNN can take the longest lagging, and LR and LSTM only take eight lags.

Finally, how long should we train on the data? We experimented with four dates for our training start point (YYYY-MM-DD): 2012-01-01 (the beginning of all data entries), 2015-01-01 (the beginning of runoff and evaporation features), 2017-01-01 (the beginning of satellite features, i.e., vegetation band), and 2019-01-01 (the more recent pick). Table X shows that different models have different preferences on the start date. As we mentioned before, LSTM is insensitive to data inclusion, and LR needs the shortest but most features. Comparing different heatmaps of the same model, we observe higher accuracy and AUC results if we move more forward than 2012, with most models preferring 2015. We achieve almost the same level of accuracy when iterating over 2017 and 2019. That means we will have more predictability when we have more features included. However, what we should include as predictors is worth discussing in section 4.3.

### 4**.3 Feature Importance**

We have contributed to this project to concatenate useful data sources and convert them into predictive features in drought risk. At the end, our feature lists contain temperature measurements (temperature max {TMAX}, temperature min{TMIN}, snow {SNWD}, soil moisture {smap}, evaporation {evap}, runoff {roff}, and vegetation band {NDVI\_1, NDVI\_2, NDVI\_diff, NDRE\_1, NDRE\_2, NDRE\_diff }. In time-series models that take lagged labels as a feature, we also included target variables {0,1} as our predictors. The lists remain long, and we are afraid that too many features will increase the noise and offset others' predictability. Therefore, we must investigate the importance of each.

****Figure X . Separation of features, the year 2017

We first examine the separation of features based on our binary target variables (0: no drought; 1: drought). We get similar results for every year in the dataset, and for the sake of simplicity, we output the year 2017 in the presentation here. As shown, runoff, evaporation, soil moisture, precipitation, and snow show some level of separation, indicating that they are more useful in the prediction task. Meanwhile, two temperature measures show no clear trend in the 0 and 1 classes, and the vegetation band measures have obvious outliers that lead us to extreme caution when using them.

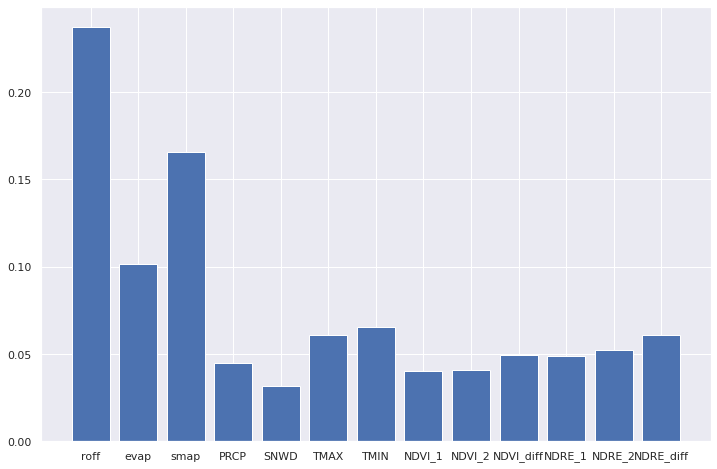


Figure X . Feature Importance for the decision tree model (max\_depth = 10)

We output the feature importance plot in one of the decision tree models with all features included. The top five crucial features are runoff, soil moisture, evaporation, temperature max, and min, almost in consensus with the previous exploration. However, by looking at the numerical value of importance, the top two features have the importance of just over 15%, while the lowest few features have less than 5%. It is not surprising that the model does not give us better results when we use all the features. The noise created by some features might be too large. We shall consider using regularization such as LASSO if we want to include all features.

| **Model** | **Feature Included** |
| --- | --- |
| LR | Evaporation, runoff, soil moisture |
| DT | Soil moisture, precipitation, snow, temperature |
| RF | Evaporation, runoff, soil moisture, precipitation, snow, temperature |
| XGBoost | Evaporation, runoff, soil moisture, precipitation, snow, temperature |
| CNN | Evaporation, runoff, drought index (autoregressive self) |
| LSTM | Evaporation, runoff, drought index (autoregressive self) |

Table X. Features included to produce best result

Table X summarizes the most valuable features in each type of model. For logistic regression, the most valuable features (that give the best accuracy and corresponding AUC) are runoff, evaporation, precipitation, snow, temperature max, and temperature min. For tree models, features are soil moisture, precipitation, snow, temperature max, and temperature min. We choose to use runoff, evaporation, and the target variable for deep learning neural networks. The result matches the separation plot of figure X that evaporation, runoff from the SET dataset, and soil moisture from SMAP could be the most helpful indicator of drought, which makes sense since it is related to the water level in the soil.

In previous section 4.3, "feature importance," we classify soil moistures, evaporation, runoff, and vegetation band features as satellite imagery. Despite the vegetation band extracted directly from the raw satellite images turning out to be weak, SMAP and SET features indicate that satellite mapping shows strong predicting power. NOAA's climate data, including temperature measurements, precipitation, and snow level, are relevant to the prediction task.

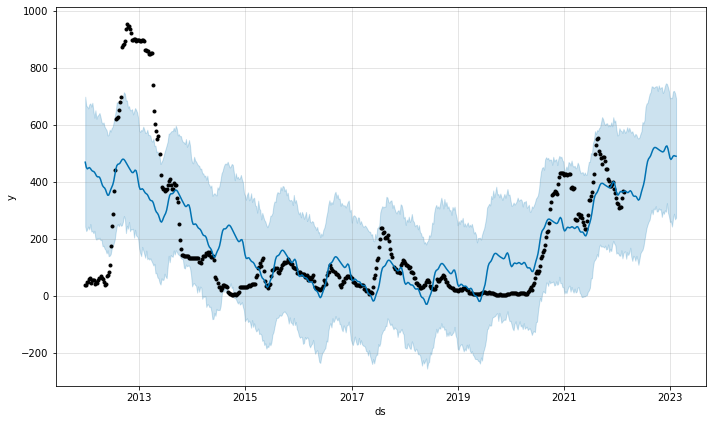
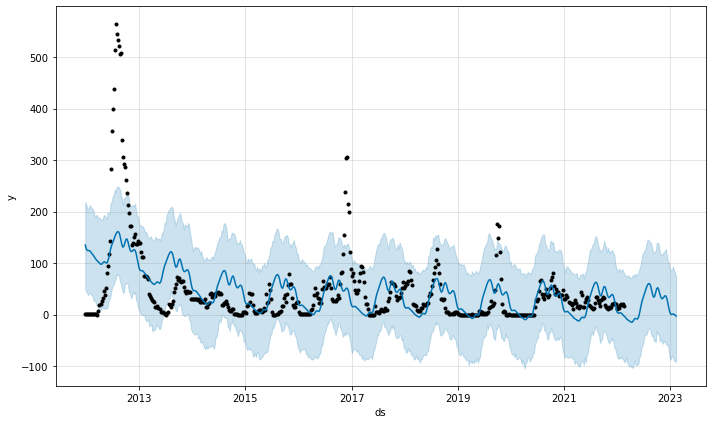
### 4**.4 Regional Variability**

As Nourani and Molaijou (2017) and Li et al. (2019) have mentioned, regional variability in categorical drought risk prediction is evident across different models. Here we used the decision tree model to experiment. We partition the CONUS into nine climate regions: northeast, upper midwest, Ohio valley, southeast, northern Rockies and plains, south, southwest, northwest, and west (Drought.Gov, cite). In each region, we fit a DT and record the prediction week, history length (training lag), data start date, accuracy, and AUC. Features used are soil moisture, precipitation, snow, and temperature measures.

| **Region** | **Prediction Week** | **Lag** | **Start date** | **Accuracy** | **AUC** |
| --- | --- | --- | --- | --- | --- |
| Northeast | 8 | 26 | 2015-01-01 | 0.538 | 0.613 |
| Upper Midwest | 8 | 26 | 2017-01-01 | 0.796 | 0.513 |
| Ohio Valley | 8 | 4 | 2019-01-01 | 0.450 | 0.480 |
| Southeast | 52 | 4 | 2015-01-01 | 0.475 | 0.521 |
| Northern R & P | 12 | 8 | 2015-01-01 | 0.875 | 0.718 |
| South | 12 | 4 | 2017-01-01 | 0.725 | 0.636 |
| Southwest | - | - | - | - | - |
| Northwest | 4 | 26 | 2019-01-01 | 0.504 | 0.617 |
| West | - | - | - | - | - |

Table X. Regional Decision Tree Model (Max\_depth = 10)

In each region, we have data for 50 counties in 10 years. After the train-test split, ten counties remain in the testing set. In the southwest and west regions, all randomly selected data are in the same category (class 1 or 0 only), so the model is trivial and thus has been left blank. We see variabilities across prediction horizons, training lags, and start dates for all other regions. The best accuracy happens in the northern Rockies-sand plains regional model (87.5%) and with over 70% AUC, followed by the upper midwest (79.6%) and south regions (72.5%). The worst performance is observed in the Ohio Valley region, with less than 50% accuracy.

Figure X . left: USDM class weighted sum in Northern R & P; right: USDM weighted sum in Ohio Valley region

We see different time-series patterns from the plots above, the weighted sum of USDM drought index from 6 original classes (No drought, D0, D1, D2, D3, D4). There is an upward trend in northern R & P starting from 2019. In the Ohio Valley, the series oscillates as white noise most of the time. Although the weighted sum plots do not indicate the prediction accuracy on binary drought class as we used in every modeling mentioned earlier, at least they show data's intrinsical difference from region to region, making models from the same modeling class diverge in prediction result.

## **V. Conclusion and Discussion**

### 5.1 Conclusion

This project has collected and concatenated meteorological data and satellite images within CONUS to extract information and predict binary drought risk at the weekly county level. We discovered that machine learning and statistical models could make meaningful predictions by experimenting with different models, including baseline logistic regression, decision trees, random forests, long short-term memory, and convolution neural networks. We are particularly interested in the temporal sensitivity of our models' predictions. We noticed that in logistic and tree-based models, longer prediction weeks horizon tends to increase the prediction accuracy. For neural networks, more extended prediction is less accurate, so the AUC measures the probability of the model's ability to distinguish between classes. The manipulation of training lags also suggests that we achieve a better result in mid-term prediction, which we think results from noisy datasets or strong correlation for closer lags. Also, we found that the start date for data inclusion matters that we tend to have better performance when we get to 2015 when water evaporation and runoff (from soil surface) data kicks in. We also test out feature importance in decision tree models. The higher importance of SMAP and SET features indicate that satellite mapping is proper. Lastly, there exist regional differences in CONUS. Due to data imbalance, when fitting models for each of the CONUS climate regions, we observed variability in optimal prediction week, training lag, and the start date due to data imbalance. This suggests that we shall formulate future prediction tasks with more granularity.

### 5**.**2 **Limitation**

*Data.* This research aims to build meaningful prediction models on free and open-source datasets. The accessible nature of the datasets determines that they may contain a specific rate of fallacies. We do not have the resources to ensure the input of our model is completely accurate despite the enormous amount of time and effort we spent on collecting, cross-checking, and cleaning up data. One major issue in the dataset is the imbalance across time and location. Different sets of features kick in at different times. USDM is complete from 2012 to 2022, and it serves as a benchmark for aligning and merging datasets. SMAP started in 2012, but most entries before 2015 are empty. SET started in 2015, and satellite features started in 2017. Besides, NOAA does not have a precise measurement for every county, simply because there is no weather station in each, and so one county has to draw data from its closest neighbor. For each feature, we impute missing entries using the mean value of that county; if the whole county is missing the data, we impute it with the global average.

*Model choices.* We seek to survey and experiment with a broad spectrum of modeling choices, from basic logistic regression to more advanced LSTM and deep Convolutional Neural networks on Multitasking dense classification (CNM). However, due to time constraints, we can only explore a limited number of model architectures. It is noteworthy that other functional model choices exist in statistical time-series studies. E.g., Vector Autoregressive models (VAR) and state-space models (Hidden Markov Chain) could be good modeling choices. Nevertheless, we view this project as machine learning work, leaving room for our fellow statisticians' future studies.

## VI. Works Cited

**TBA**