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

Thee Ngamsangrat, Yujie Cai, Jim Zhang, Michael Fein

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

Mar 30, 2022

## **Abstract**

Drought has become one of the leading causes of humanitarian disasters. Drought leads to drinking water shortages, wildfires, loss in 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. In this paper, we attempt to extend drought prediction accuracy past currently accepted capabilities using statistical and machine learning approaches based on a number of feature variables. Our work combines satellite images with climatrophic data at a regional level to gain higher forecast accuracy for larger regions of the Continental United States. Our results show regional and temporal variability in the predictability of drought using different features, and we are able to improve the prediction power with more advanced model architectures. This work will allow for a more accurate allocation of emergency response resources before the onset of a more severe drought. If possible, we also hope to describe certain actions, taken by humans, that may have a correlational relationship with causing drought, such as land use or urban sprawl.

## **I. Introduction**

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

Droughts are one of the most costly and environmentally damaging climatic events. Due to nature's dependency on hydraulic cycles, droughts can lead to the long-term destruction of large-scale ecosystems through positive feedback loops. Understanding and better predicting when and where drought will occur is of utmost importance to aiding those affected by drought and in understanding how to potentially mitigate oncoming droughts. A prediction period of greater than even a month ahead of time would allow a 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 periods of 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 important aspect in predicting its onset as well as its effects. It is important to note that drought is not simply a lack of available water, since different ecosystem types will naturally experience different levels of precipitation; rather, drought is a deviation from what is considered an average amount of available water in specific climate regions. For our purposes, 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 makes use of the Palmer Drought Severity Index, the Standardized Precipitation Index (SPI), and other climatic inputs as well as indicators of soil moisture. 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 body of work.

In terms of what is currently available, NOAA outputs weekly to 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 been increasing in 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, in part because of its complexity as a natural event, and also in part 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) greater 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. Finally, we will conduct a correlational study of human-factors such as urbanization and land use and drought prediction.

### **1.2 Available Data**

For our purposes of drought risk forecasting and prediction, we use data from a number of other sources in an effort to improve 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

Using this data in various modeling capacities we align changes of drought risk (again measured through USDM) over time to climatrophic and satellite data to achieve better accuracy over longer periods of time than currently available. We highlight examples of these data sources below.

The USDM quantifies the drought risk into 5 levels (D0-D5) from the least serious drought to the most severe one. The organization officially suggested a transformation can be done to map the categorical data into the continuous regime. That is, apply a weighted sum across five levels for a location. Below is the figure showing the weighted sum of drought index in five counties, from 2012-1-1 to 2022-2-22.

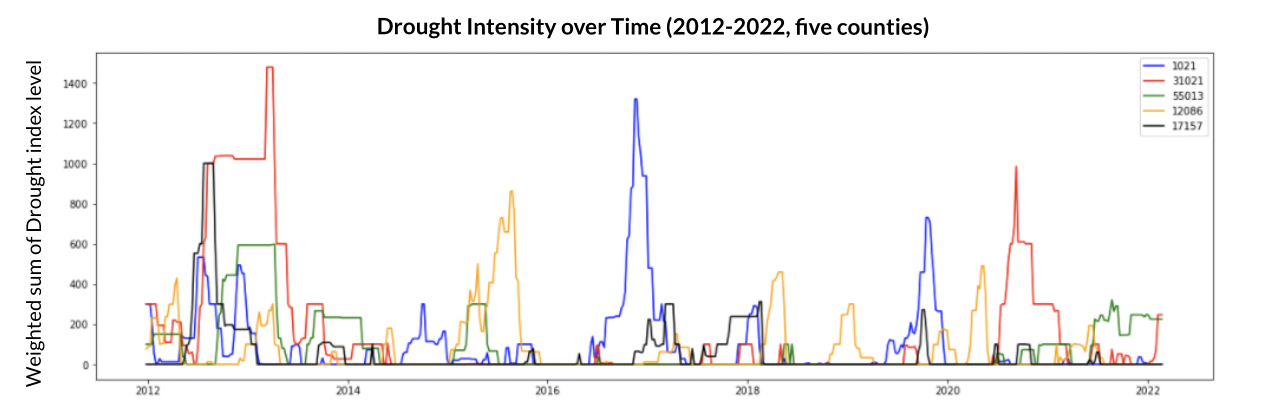


Figure 2. USDM visualization

The figure below demonstrates how SMAP, soil moisture, can change over time:



Figure 3. Difference in SMAP Measurements from 2019 to 2020 across the state of California.

Using Sentinel Satellite images such as those pictured below gives us many options for creating feature sets in our project. For example, using metadata from Sentinel we can measure the amount of vegetation in specific areas as well as the percentage of cloud cover. Through masking for these measurements we can derive quantitative features which can be less computationally costly to use in modeling.

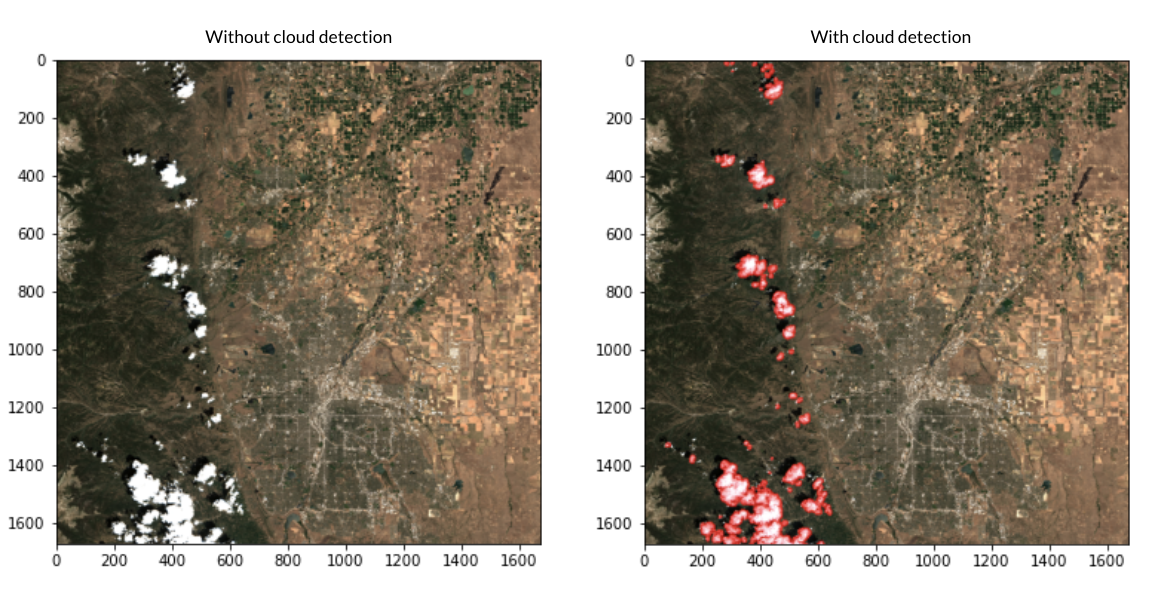


Figure 4.Sample of Satellite images with and without cloud masking

Finally, SET provides us with historical runoff and evaporation data while NOAA provides us with temperature and precipitation data across weather stations. The main difficulty in using these various data sets is understanding how to most effectively merge them. For example, mapping satellite images to SMAP images raises issues along the border of maps and mapping weather stations across the CONUS to counties may not line up precisely.

We aim to combine all of the data sources into one full dataset at weekly time scales and at county level, so we are able to proceed to modeling. However, we acknowledge that NOAA, SMAP, SET, and Sentinel L2 have different availability across location and space. For example, the temperature and precipitation from NOAA is measured by weather stations, and those weather stations are not necessarily aligned with individual county boundaries. SET starts to have full entries since 2015, while the most complete data we have is USDM, which starts from 2012. Having the imbalance panel dataset in mind, when we enter the modeling phase, we want to investigate how the data availability contributes to the model predictability.

## **II. Related Work**

Previous work in the study of drought and arid climate forecasting has formed a major consensus – drought is one of the least understood natural disasters due to its convoluted causal factors and occurrence mechanisms (Sundararajan, 2021). With the flourishing of machine learning and deep neural networks, researchers have made some limited but encouraging progress in forecasting drought and encoding its possible predictors.

### **2.1 Complexity in Droughts’ Causal Mechanisms**

Current literature suggests that causes of droughts are rooted deeply in the specific topographical and geographical conditions of an area. Zhang et al. (2020) suggests that the reason for 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). As such, 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, typically less than a year. Most work also focuses on specific, smaller regions of sub-climates, where hydraulic cycles can be more accurately known, such as river basins, where water, evaporation, and precipitation levels can be closely measured and reported.

This work does not attempt to point to specific causes of drought, which might vary by climate region as suggested in literature, but rather divides the CONUS into generally agreed upon climate regions. Through this method, we can 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).

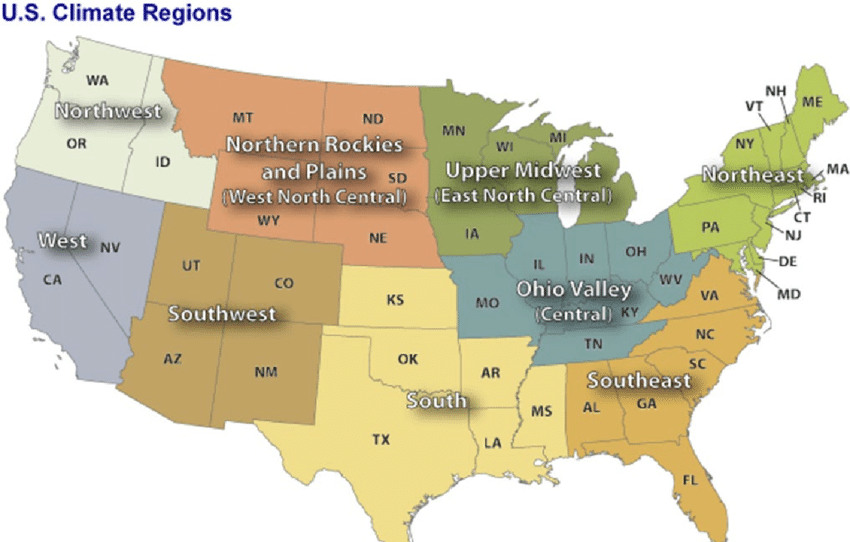


Figure 5. US Climate Regions

### **2.2 Recent Complications in Response to Climate Change**

Previous works culminating in the survey work of Vicente-Serrano et al. (2020) recounted a snowballing problem that makes drought forecasting increasingly complex – droughts react differently to climate change.

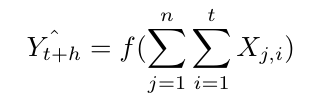
Some persistent and cyclical droughts have been paradoxically mitigated by the melting of the polar ice caps, leading to a number of sporadic misallocations of drought alleviation supplies (Scheff, 2018). Meanwhile, Bhaga et al. (2020) warned that global climate change has worsened the already severe droughts in sub-saharan Africa. Even if a stable mechanism of drought occurrence existed, global climate change would likely disrupt its processes , making drought prediction more difficult.

In this work, we attempt to avoid potential confounding from prior data by focusing our forecasting efforts on more recent data, beginning in 2012, rather than incorporating some available long-term historical data, such as precipitation and temperature. Since climate change has been occurring at an increasing rate, even faster than originally expected, as a result of positive feedback cycles in its root causes (Tollefson, 2022), focusing on recent changes in climate and forecasting off of more recent measurements from the last decade will hopefully reduce the potential overfitting of longer-term cyclical patterns allowing for more accurate predictions in the present and into the near-term future.

### **2.3 Overview of Drought Forecasting Methods**

Despite the central challenge of difficult to understand causal mechanisms and the disruption of mechanisms by global climate change, scholars have made some recent promising progress in attempts to predict drought. Most existing work in drought forecasting can be categorized into two types: reduced-form approach and structural approach.

The reduced-form approach adapts and applies the State-of-the-Art (SOTA) machine learning models to drought forecasting, maximizing accuracy in a machine learning optimization setting without being overly concerned with the theoretical interpretability of the drought prediction models. In other words, it creates black-box models to attempt to embed hidden patterns within available data to enable more accurate prediction. Usually, this approach trains such a model:



where are features and observable history and it seeks to learn a projection and predict steps into the future. Reduced-form studies do not necessarily understand how the features of interact in a causal way; they simply tackle this as an optimization task that aims to minimize the selected loss function. During training, these works usually use a mean squared error (MSE) loss for continuous drought responses or a categorical cross entropy (CCE) loss for index-level drought responses.

Some more recent examples of the reduced-form approach, such as Poornima and Pushpalatha (2019), Dikshit et al. (2021) use recurrent neural networks and long short term memory (RNN & LSTM) to study the time series drought indices. Dikshit et al. (2022) , one of the most recent scientometric analyses in the field of drought prediction, confirms that reduced-form machine learning is achieving steadily increasing accuracy in drought prediction.

The structural approach is its conceptual counterpart. The structural approach starts from the theoretical model and strives for the correctness and rationality of the theories and causal mechanisms at each step; machine learning methods or probabilistic modeling are auxiliary tools that help achieving better numerical performance.

For example, a structural model would start by specifying the curve between temperature and moisture, and then, how moisture factors into precipitation, and ultimately, drought. Researchers build these model structures based on domain knowledge and use statistical and computational methods to optimize the prediction performance.

Structural modeling, thus far, has made relatively slow progress. As summarized above, the causes of drought are complex and the conclusions of theoretical models from different geographical areas lack interchangeability and generalizability. Hao et al. (2018) point out that most of the breakthroughs in structural modeling have occurred through expansions in available data. For example, by utilizing factors such as large-scale climate indices, local climate variables, and land initial conditions from advances in sensor technology and computing power, models have improved over time.

Ideologically, there is no advantage or disadvantage to these two types of prediction methods. Reduced-form models face lighter constraints and can usually produce more accurate predictions thus far while structural modeling yields higher interpretability in the possible causes of drought. In this work, we create both reduced-form models, through black-box methods, and structural models, through statistical methods.

## **III. Methodology**

### **3.1 Preprocessing of Satellite Images**

Sentinel L2A contains Bottom of the Atmosphere (BOA) satellite images globally starting 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 which we identify as relevant to our prediction tasks include those which are 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 reflectance of vegetation in the near infrared range of the electromagnetic spectrum. We use them to identify vegetation areas in images accurately as shown in Figure 1. This method allows us to measure the percentage 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 nature of cloud coverage between different images. This problem is best illustrated in Figure 2, where the two images are taken at the same 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 directly measure the percentage change of vegetation area between the two images. We tackle this problem by using the cloud probability band to detect pixels that are likely to be covered by clouds in either image and ignore those pixels. We then compare the percentages of vegetation areas in the remaining pixels in the two images. Using this method, we are able to generate a weekly percentage change of vegetation area in each county over the past 5 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.

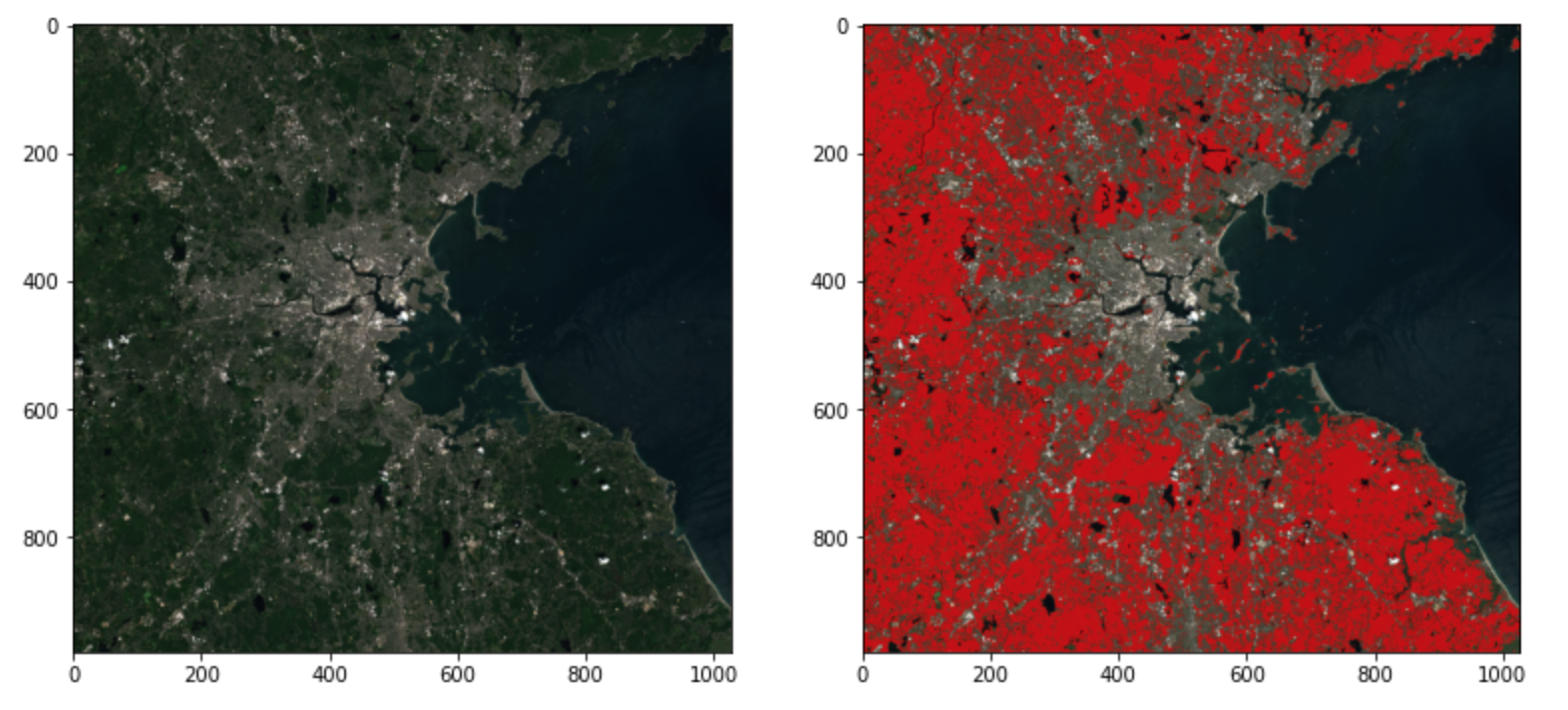


Figure 6: Suffolk, MA 2020-07-15, Highlighted by Red when NDVI > 0.8 and NDRE > 0.65

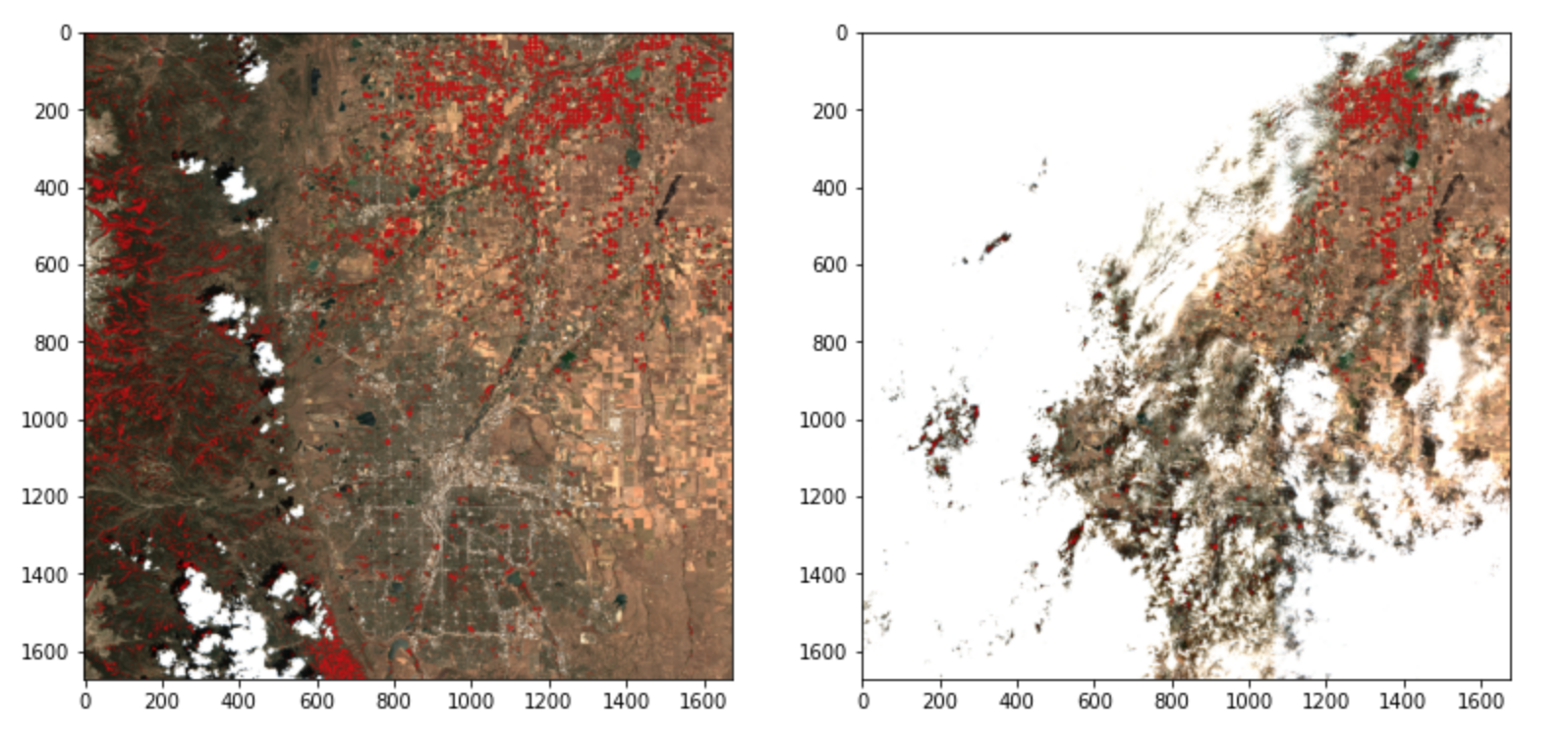


Figure 7: Broomfield, CO in July 2020 and in July 2021

### **3.2 Sampling and Train-Test Split**

As mentioned above, we have county level data for 3,143 counties across the whole CONUS. If we pull climate and satellite data from each county for 10 years (531 weeks), which requires too much work to be feasible. Therefore, we apply stratified sampling according to the climate region division. We sample 50 counties from each region out of a total of 450 counties. We argue our sample is representative of the CONUS and we test that the 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, in order to test our model’s robustness across time and space. As per the general process for dealing with cross-sectional data, we randomly assign a portion of our data to be the training data and the other one to be the testing data, ensuring that the two groups are independent. The panel data is handled differently using our two time and space dimensions. We first take 80% of the counties to form our training set and stratify them by climate region. Within the training set, we use several two-year windows to do training and validation, and make predictions starting from 2018-03-20. Next, we test on our control data.

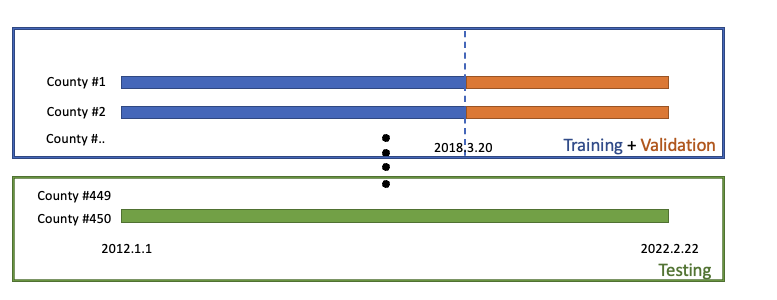


Figure 8. Train/test split

### **3.3 Modeling**

We divided our modeling approaches into two main categories: time series based and linear model based. Based on our review of previous work, both types of models have their advantages and disadvantages. Time series based models have the advantage of detecting longer-term trends through seasonality cyclicality while having access to more limited data. Linear models can incorporate a greater diversity of data but lack the ability to pick-up longer term cyclicality.

#### **3.3.1 Time Series Forecasting**

##### 3.3.1.1 Baseline: Autoregressive Models

Treating the drought prediction as a time series problem, we first try the simple, autoregressive model using only the USDM drought index data. As we can see in section 1.2, the USDM data is continuous. To account for possible seasonality and non-stationarity, we adapt Meta’s Prophet open source package (Letham and Taylor, 2017) to construct our autoregressive model. We aim to stretch the prediction horizon from one to eight week. The package also allows us to tune two hyper-parameters that work as L1 and L2 regularizations, adding penalties to abrupt change and seasonality in the series. We train the model using 2012-1-1 to 2018-03-20 (2190 days) data and start making predictions and cross-validation afterwards. The model is finally evaluated on the test set.

We first apply the autoregressive model on the country-level. We average the drought index for all 450 counties in order to generate a single number for each week. Acknowledging the fact that we might be overlooking the difference between climate regions and data imbalance issues, we then add more granularity into the model construction. We also model each climate region individually and compare the models’ performances.

##### 3.3.1.2 Long-Short-Term Memory

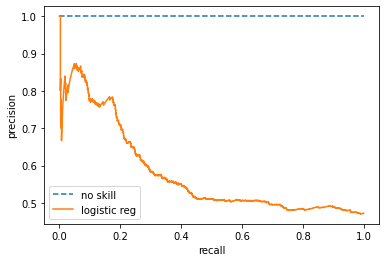
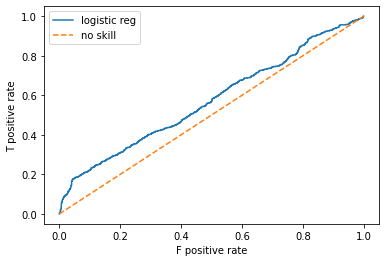
##### **3.3.1.3 Recursive Neural Network**

#### 

#### **3.3.2 Linear Prediction**

##### 3.3.2.1 Baseline: Logistic Regression

We first implemented a logistic regression model due to its ease of implementation and interpretation. We experimented with various predictors and training hyperparameters to find a baseline accuracy score for long-range predictions. After this fine tuning process, the best logistic model achieved an accuracy of ~67% with an AUC and Precision Recall graph shown below:



Logistic regression worked best with a limited set of predictors: runoff, evaporation, precipitation, snow, temperature max and temperature min.

##### 3.3.2.2 Tree Based

##### 3.3.2.3 Deep Learning

We further hope to improve prediction accuracy. After studying what sort of feature combination works better for certain regions, we are interested in seeing how model complexity or model class can affect the predictability. Specifically, **we select XXX features on XXX regions** and apply deep learning methods (LSTM, RNN).

## **IV. Result and Discussion**

### **4.1 Metrics**

We choose MSE as our evaluation metric 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 ease of influence by the number of observations and predictors.

**4.2 Regional Variability in Predictability (FUTURE also)**

As 3.3.1 suggested, we have tried ensemble and regional models. We find that the baseline model is catching the trend in the ensemble methods: there is an increase in national drought risk following a decrease in drought risk till 2017; drought happens more in dry seasons (autumn and winter) and less in humid seasons (spring and summer). However, when comparing the performance across time horizons, in general, we observe high levels of MSE (in the order of 10^4) in near (1 week) and far future (5-8 weeks). We think the high MSE might be due to too much averaging that we have done, so the model is not well-generalizable to the test set. We then move onward and experiment with regional models.

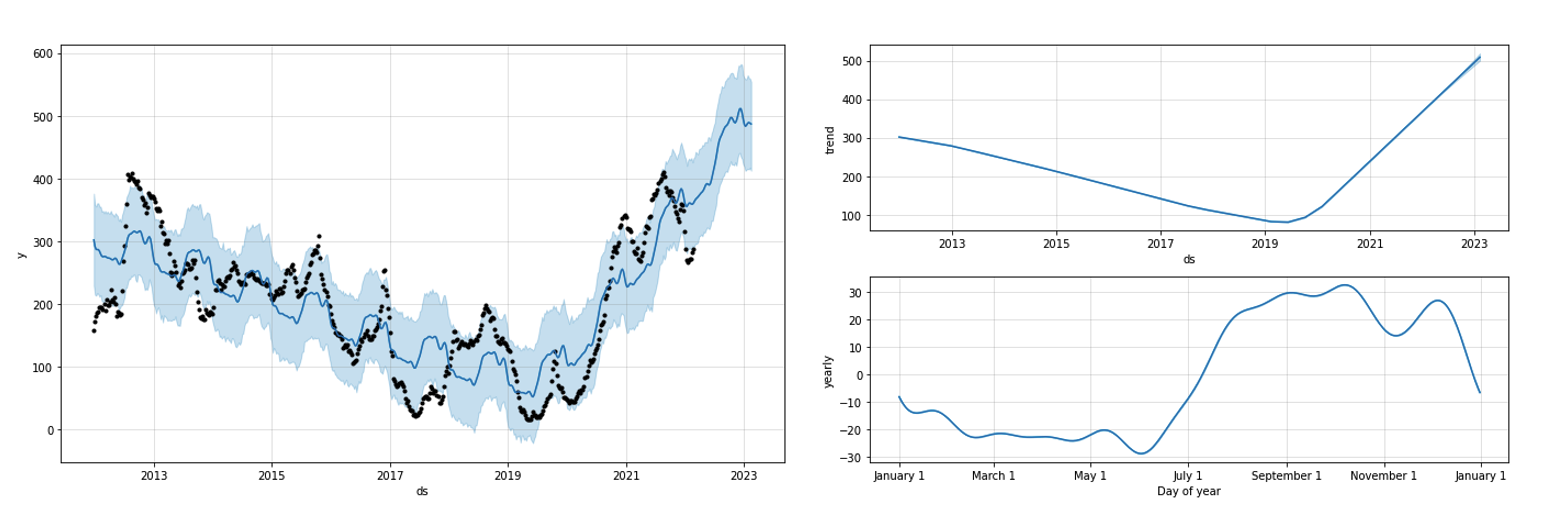


Figure 9. Meta (Facebook)’s Prophet model

The following table outputs a cross-regional comparison in MSE for mid-term predictions. We have omitted the near future ( 1 week) and far future (5-8 weeks) predictions as we are still getting higher order MSE.

| Region | 2 weeks | 3 weeks | 4 weeks |
| --- | --- | --- | --- |
| **Ohio Valley** | 543.28 | 2.67 | 31.71 |
| West | 3002.43 | 534.54 | 2596.50 |
| **Northwest** | 450.73 | 177.55 | 49.76 |
| **Northeast** | 115.04 | 459.88 | 2.68 |
| Northern R & R | 7743.07 | 1069.49 | 1154.16 |
| South | 3869.22 | 1126.15 | 11606.85 |
| **Southeast** | 14.55 | 1612.00 | 327.98 |
| **Upper Midwest** | 9357.98 | 8.194 | 79.37 |
| Southwest | 21820.43 | 2677.14 | 10085.16 |

Table1. Prediction MSEs across climate regions

The results from regional, autoregressive models indicate that we do have variability across climate regions and across time horizons. With regards to time, we find that some regions are more predictive of the nearer future while others are better for the more distant future. The mechanism might again be due to the averaging effect as we discussed in the previous paragraph.

Spatially, Ohio Valley (a), Northwest (b), Northeast, Southeast, and Upper Midwest regions are associated with greater predictive power. Southwest (c) and South (d) are relatively poor in terms of MSE. For our better models (a and b), the original data is more stable – either with regards to constant variance (a) or moderate variance (b). Both of our poor models (c and d) see considerable variance over time.

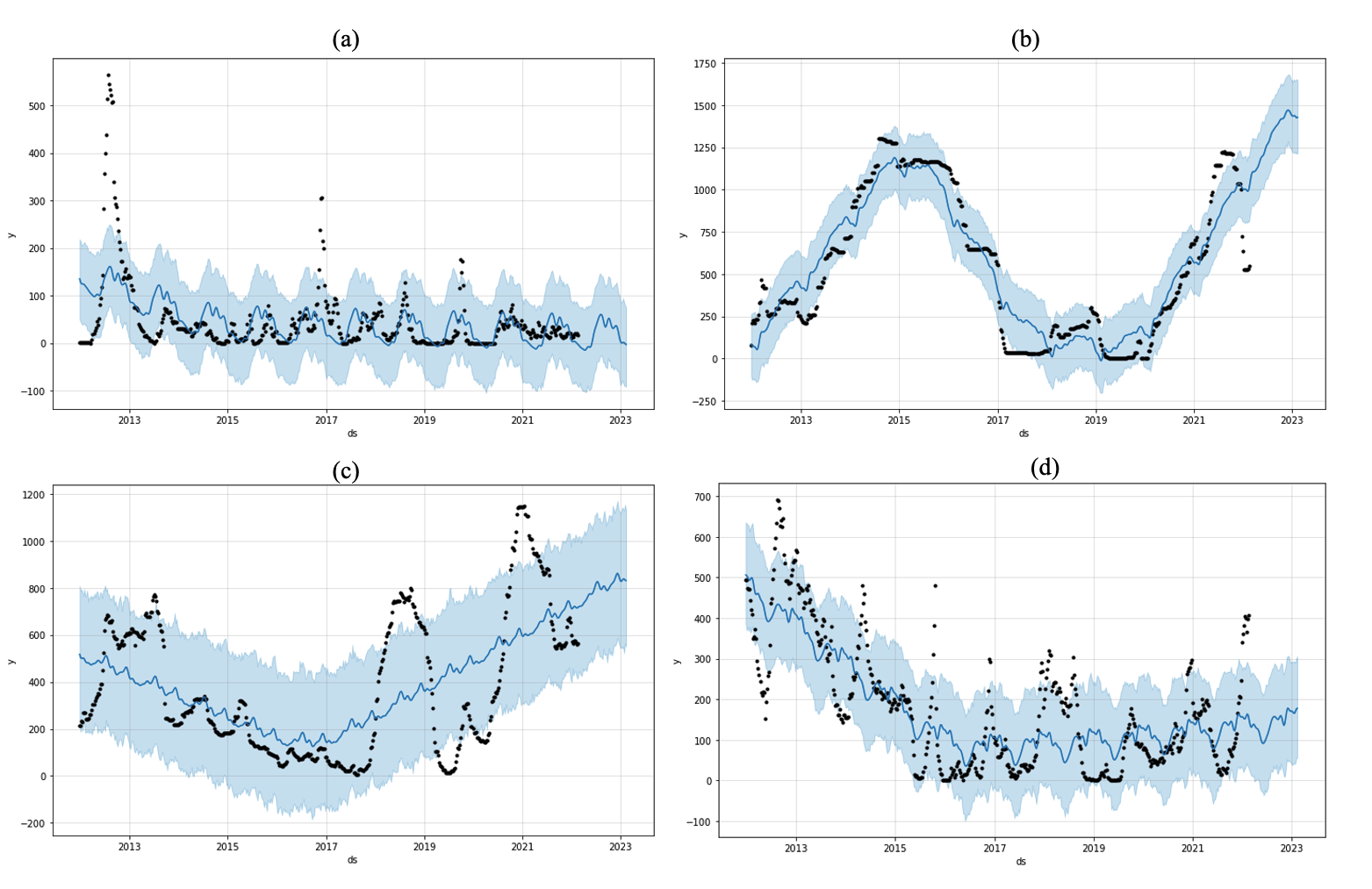


Figure 10.

The USDM dataset is balanced across all time and space, removing any concerns regarding data availability and the predictive power. However, this may prove to be a concern as we move across OLS models, requiring different explanations.

### **4.3 The Shape of Target Variable in Predictability (FUTURE)**

### **4.4 \*Limitation (FUTURE)**

## **V. Work Cited**

Bhaga, Dube, T., & Shoko, C. (2021). Satellite monitoring of surface water variability in the drought prone Western Cape, South Africa. *Physics and Chemistry of the Earth. Parts A/B/C*, *124*, 102914. <https://doi.org/10.1016/j.pce.2020.102914>

Cordeiro, M.C., Martinez, J.M. and Peña-Luque, S., 2021. Automatic water detection from multidimensional hierarchical clustering for Sentinel-2 images and a comparison with Level 2A processors. *Remote Sensing of Environment*, *253*, p.112209.

<https://doi.org/10.1016/j.rse.2020.112209>

Dikshit, & Pradhan, B. (2021). Interpretable and explainable AI (XAI) model for spatial drought prediction. *The Science of the Total Environment*, *801*, 149797–149797. <https://doi.org/10.1016/j.scitotenv.2021.149797>

Dikshit, & Pradhan, B. (2021). Explainable AI in drought forecasting. *Machine Learning with Applications*, *6*, 100192. https://doi.org/10.1016/j.mlwa.2021.100192

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

Jimenez, Marengo, J. A., Alves, L. M., Sulca, J. C., Takahashi, K., Ferrett, S., & Collins, M. (2021). The role of ENSO flavours and TNA on recent droughts over Amazon forests and the Northeast Brazil region. *International Journal of Climatology*, *41*(7), 3761–3780. <https://doi.org/10.1002/joc.6453>

National Drought Mitigation Center. USDM Dataset. (2022). Retrieved form <https://droughtmonitor.unl.edu/Data.aspx>

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>

Salehi-Lisar, & Bakhshayeshan-Agdam, H. (2016). Drought Stress in Plants: Causes, Consequences, and Tolerance. In *Drought Stress Tolerance in Plants, Vol 1* (pp. 1–16). Springer International Publishing. <https://doi.org/10.1007/978-3-319-28899-4_1>

Scheff. (2018). Drought Indices, Drought Impacts, CO2, and Warming: a Historical and Geologic Perspective. *Current Climate Change Reports*, *4*(2), 202–209. <https://doi.org/10.1007/s40641-018-0094-1>

Sharafati, Nabaei, S., & Shahid, S. (2020). Spatial assessment of meteorological drought features over different climate regions in Iran. *International Journal of Climatology*, *40*(3), 1864–1884. <https://doi.org/10.1002/joc.6307>

Sundararajan, Garg, L., Srinivasan, K., Bashir, A. K., Kaliappan, J., Ganapathy, G. P., Selvaraj, S. K., & Meena. (2021). A Contemporary Review on Drought Modeling Using Machine Learning Approaches. *Computer Modeling in Engineering & Sciences*, *128*(2), 447–487. <https://doi.org/10.32604/cmes.2021.015528>

Vincente-Serrano, Begueria, S., & Lopez-Moreno, J. I. (2010). A Multiscalar Drought Index Sensitive to Global Warming. *Journal of Climate*, *23*(7), 1696–1718. <https://doi.org/10.1175/2009JCLI2909.1>

Zhang, Qu, Y., Zhang, X., Wu, X., Zhou, X., Ren, B., Zeng, J., & Wang, Q. (2021). Spatiotemporal variability of annual drought severity, duration, and frequency from 1901 to 2020. *Climate Research*. <https://doi.org/10.3354/cr01680>