# U.S. Drought Prediction Using Meteorological Data

**University id: 2283791**

## Abstract

The prevalence of extreme climate conditions in recent years has been causing disruption to the precipitation cycles and causing shortage of water globally. Drought, a serious threat to mankind, is causing challenges in different areas, from food security and famine to tourism and energy. Accurate forecasting of drought conditions plays an important role in increasing the preparedness for potential crop failure and avioding famine.

## 1 Introduction

Drought can be characterized as a prolonged period of below average precipitation and it can categorized into four main different categories: Meteorological, Agricultural, Hydrological, and Socioeconomic. Each category has impact on different areas of our daily lives. Drought has been linked to drying up rivers in the U.S. [1] to increasing olive prices across Europe [3, 4]. In this paper, we will study the relation between several meteorological variables and drought level and we will assess the suitability of using linear regression to predict drought.

## 2 Datasets

In this work, we will use two datasets:

### 2.1 U.S. Drought Data

This dataset was obtained from Kaggle [5]. The dataset consists of meteorological variables along with drought level at the county level in the Contiguous United States on daily and weekly basis.

Tables 1 and 2 shows sample five rows from the data.

Table 1: First 10 variables for first 5 rows

| fips | date | PRECTOT | PS | QV2M | T2M | T2MDEW | T2MWET | T2M_MAX | T2M_MIN |
|---|---|---|---|---|---|---|---|---|---|
| 1001 | 2000-01-01 | 0.22 | 100.51 | 9.65 | 14.74 | 13.51 | 13.51 | 20.96 | 11.46 |
| 1001 | 2000-01-02 | 0.20 | 100.55 | 10.42 | 16.69 | 14.71 | 14.71 | 22.80 | 12.61 |
| 1001 | 2000-01-03 | 3.65 | 100.15 | 11.76 | 18.49 | 16.52 | 16.52 | 22.73 | 15.32 |
| 1001 | 2000-01-04 | 15.95 | 100.29 | 6.42 | 11.40 | 6.09 | 6.10 | 18.09 | 2.16 |
| 1001 | 2000-01-05 | 0.00 | 101.15 | 2.95 | 3.86 | -3.29 | -3.20 | 10.82 | -2.66 |

Table 2: Last 11 variables for first 5 rows

| T2M_RANGE | TS | WS10M | WS10M_MAX | WS10M_MIN | WS10M_RANGE | WS50M | WS50M_MAX | WS50M_MIN | WS50M_RANGE | drought |
|---|---|---|---|---|---|---|---|---|---|---|
| 9.50 | 14.65 | 2.20 | 2.94 | 1.49 | 1.46 | 4.85 | 6.04 | 3.23 | 2.81 | NaN |
| 10.18 | 16.60 | 2.52 | 3.43 | 1.83 | 1.60 | 5.33 | 6.13 | 3.72 | 2.41 | NaN |
| 7.41 | 18.41 | 4.03 | 5.33 | 2.66 | 2.67 | 7.53 | 9.52 | 5.87 | 3.66 | NaN |
| 15.92 | 11.31 | 3.84 | 5.67 | 2.08 | 3.59 | 6.73 | 9.31 | 3.74 | 5.58 | 1.00 |
| 13.48 | 2.65 | 1.60 | 2.50 | 0.52 | 1.98 | 2.94 | 4.85 | 0.65 | 4.19 | NaN |

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

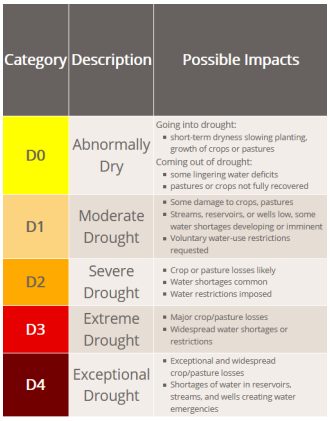


Figure 1: Drought categories and their corresponding impact

As we can see, the data consists of only continuous variables along with one categorical variable, the county fips code. The meteorological variable were gathered from the NASA POWER Project [7] and the drought levels were gathered from U.S. Drought Monitor Project [8]. Both of these datasets were calculated using satellite imagery.

The meteorological variables are available on a daily basis, however, the drought level is available every week (this explains Null values in the drought column). The data spans from *2000-01-01* to *2020-12-31* and contains 23,841,468 records. This large number of records necessitates some pre-processing in order to make the data easier to work with. We call this type of data time series data, as we have observations recorded at equally spaced intervals (daily in our case). We will see later how this type of data differs from traditional tabular data and what challenges it imposes for developing regression model.

The detailed meaning of the meteorological features is specific to climate scientists, however, it's necessary for our analysis to understand at a high-level what do these variables stand for. Table 3 contains description for each different variable:

The drought variable (the variable that we want to predict) takes on values in the range [0-5]. The higher the value, the more severe the drought is. The U.S. drought monitor provides the drought as a categorical variable taking five different values. Figure 1 shows the drought levels and their corresponding impact.

### 2.2 U.S. Geographical Maps

As we have seen, our dataset contains a spatial attribute, county FIPS code. In order to visualizae the data and aggregate it at different levels, we need data about the contiguous U.S. states and their corresponding boundaries. We use the states boundaries data from the U.S. census Bureau [9].

## 3 Data Preprocessing

### 3.1 Software

The drought data is available as three CSV files and the U.S. map is available as shapefile. In this section, we list the software tools that will be used to analyze this data and to build and evaluate regression models.

### 3.2 Jupyter Notebook

Jupyter notebook provide an interactive-based environment for analyzing and visualizing data. We use the Python language given the large number of packages for data analysis in this language.

Table 3: Meteorological variables description

| variable_name | short_description | long_description |
|---|---|---|
| PRECTOT | Precipitation (mm day-1) | The bias corrected average of total precipitation at the surface of the earth in water mass. |
| PS | Surface Pressure (kPa) | The average of surface pressure at the surface of the earth. |
| QV2M | Specific Humidity at 2 Meters (g/kg) | The ratio of the mass of water vapor to the total mass of air at 2 meters (kg water/kg total air). |
| T2M | Temperature at 2 Meters (C) | The average air (dry bulb) temperature at 2 meters above the surface of the earth. |
| T2M_MAX | Maximum Temperature at 2 Meters (C) | The maximum hourly air (dry bulb) temperature at 2 meters above the surface of the earth in the period of interest. |
| T2M_MIN | Minimum Temperature at 2 Meters (C) | The minimum hourly air (dry bulb) temperature at 2 meters above the surface of the earth in the period of interest. |
| T2M_RANGE | Temperature Range at 2 Meters (C) | The minimum and maximum hourly air (dry bulb) temperature range at 2 meters above the surface of the earth in the period of interest. (Max - Min) |
| T2MDEW | Dew/Frost Point at 2 Meters (C) | The dew/frost point temperature at 2 meters above the surface of the earth. |
| T2MWET | Wet Bulb Temperature at 2 Meters (C) | The adiabatic saturation temperature which can be measured by a thermometer covered in a water-soaked cloth over which air is passed at 2 meters above the surface of the earth. |
| TS | Earth Skin Temperature (C) | The average temperature at the earth's surface. |
| WS10M | Wind Speed at 10 Meters (m/s) | The average of wind speed at 10 meters above the surface of the earth. |
| WS10M_MAX | Maximum Wind Speed at 10 Meters (m/s) | The maximum hourly wind speed at 10 meters above the surface of the earth. |
| WS10M_MIN | Minimum Wind Speed at 10 Meters (m/s) | The minimum hourly wind speed at 10 meters above the surface of the earth. |
| WS10M_RANGE | Wind Speed Range at 10 Meters (m/s) | The minimum and maximum hourly wind speed range at 10 meters above the surface of the earth. |
| WS50M | Wind Speed at 50 Meters (m/s) | The average of wind speed at 50 meters above the surface of the earth. |
| WS50M_MAX | Maximum Wind Speed at 50 Meters (m/s) | The maximum hourly wind speed at 50 meters above the surface of the earth. |
| WS50M_MIN | Minimum Wind Speed at 50 Meters (m/s) | The minimum hourly wind speed at 50 meters above the surface of the earth. |
| WS50M_RANGE | Wind Speed Range at 50 Meters (m/s) | The minimum and maximum hourly wind speed range at 50 meters above the surface of the earth. |

### 3.3 Pandas

Pandas is a tool that allows us to perform different wrangling steps, specifically working with DataFrames. This includes cleaning, transformation, joining, and so on. We use it to load and clean the drought data from CSV files.

### 3.4 GeoPandas

GeoPandas is an extension to the pandas package. It provides support for working with geospatial data. Here, we use it to read shapefile of the U.S. map. It also includes features for visualizing maps as we will see later.

### 3.5 matplotlib and seaborn

These are data visualization packages. seaborn is specifically tailored to statistical charts, which are essential to our exploratory data analysis/

### 3.6 scikit-learn

Scikit-learn contains a wide range of machine learning algorithms along with evaluation metrics. We use it to build and evaluate regression models.

### 3.7 sktime

Sktime contains routines specific for working with time series data. From splitting data to extracting suitable features from time series data.

### 3.8 Data Cleaning

### 3.9 Replace FIPS with corresponding state

Our data is clean and doesn't require extensive cleaning steps. We only have missing values in the **drought** column, but we will fix this in the next section.

The FIPS column is stored as integer while it should be a string of five digits. We add leading zeros to the column and convert it to string.

### 3.10 Replace FIPS with U.S. state

The FIPS column contains 3108 distinct values. In order to make the data easier to visualize and to analyze, we replace the FIPS code with its corresponding state. This way, we decrease the granularity of our data, going from data at the county level to data at the state level.

We do this replacement using the U.S. map we downloaded earlier.

After this replacement, we end up with 49 different values for the state column. This is much easier because it will allow us to, for example, study how the drought varies among different states and visualize this variation.

### 3.11 Spatio-temporal aggregation

Our data is still relatively large and hard to work with. We propose to aggregate the data both spatially and temporally. For each different state in the 49 available states, we aggregate all the meteorological features and the drought value and take the average.

Figure 2 illustrates the aggregation for a single state and for a single month.

The aggregation reduces the number of records from 23,841,468 to 12348. While there is definitely loss of important information, but the assumption is that counties within the same state will have identical or very similar values.

## 4 Exploratory Analysis and Correlation study

In this section, we will dig deeper into the data aiming at understanding the main characteristics underlying its generating process.

We can think of each column in our dataset as a continuous random variable. Then, we can measure its central tendency and other summary statistics, plot its Probability Density Function (PDF) and Cumulative Distribution Function (CDF), and calculate its quantiles.

### 4.1 Univariate Analysis

We start with quick summary table of all the variables. Table 4 shows the summary statistics for all variables.

| date | PRECTOT | QV2M | drought |
|---|---|---|---|
| 2000-01-01 | 0.22 | 9.65 | NaN |
| 2000-01-02 | 0.20 | 10.42 | NaN |
| 2000-01-03 | 3.65 | 11.76 | NaN |
| 2000-01-04 | 15.95 | 6.42 | 1.0 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 2000-01-29 | 8.39 | 4.56 | NaN |
| 2000-01-30 | 0.87 | 3.88 | NaN |
| 2000-01-31 | 0.00 | 2.81 | NaN |

| date | PRECTOT | QV2M | drought |
|---|---|---|---|
| 2000-01-31 | 3.756452 | 5.65129 | 1.75 |

Figure 2: Example of applying mean aggregation over one month

Table 4: Summary statistics

| | mean | std | min | max | variance | skewnewss | kurtosis |
|---|---|---|---|---|---|---|---|
| PRECTOT | 2.6 | 1.7 | 0.0 | 12.5 | 2.8 | 0.9 | 1.4 |
| PS | 95.9 | 6.0 | 77.9 | 102.1 | 36.2 | -1.6 | 1.3 |
| QV2M | 7.3 | 4.0 | 0.9 | 19.2 | 16.0 | 0.6 | -0.6 |
| T2M | 11.7 | 10.0 | -17.5 | 34.0 | 99.8 | -0.2 | -0.9 |
| T2MDEW | 5.8 | 9.0 | -19.8 | 24.6 | 80.1 | -0.0 | -1.0 |
| T2MWET | 5.8 | 8.9 | -18.3 | 24.6 | 79.0 | -0.0 | -1.0 |
| T2M_MAX | 17.3 | 10.5 | -12.7 | 41.0 | 110.1 | -0.3 | -0.9 |
| T2M_MIN | 6.4 | 9.5 | -22.4 | 26.8 | 90.6 | -0.2 | -0.8 |
| T2M_RANGE | 10.9 | 2.4 | 4.3 | 18.5 | 5.8 | 0.3 | -0.1 |
| TS | 11.7 | 10.2 | -17.6 | 35.8 | 104.6 | -0.2 | -0.9 |
| WS10M | 3.4 | 1.1 | 1.2 | 6.8 | 1.1 | 0.5 | -0.6 |
| WS10M_MAX | 5.0 | 1.6 | 1.8 | 10.2 | 2.5 | 0.4 | -0.6 |
| WS10M_MIN | 1.8 | 0.6 | 0.7 | 4.2 | 0.3 | 0.5 | -0.3 |
| WS10M_RANGE | 3.2 | 1.1 | 1.1 | 7.0 | 1.2 | 0.4 | -0.6 |
| WS50M | 5.4 | 1.2 | 2.7 | 9.3 | 1.6 | 0.3 | -0.7 |
| WS50M_MAX | 7.6 | 1.7 | 4.1 | 13.4 | 2.8 | 0.4 | -0.6 |
| WS50M_MIN | 3.0 | 0.8 | 1.1 | 5.9 | 0.7 | 0.3 | -0.6 |
| WS50M_RANGE | 4.6 | 1.0 | 2.6 | 8.7 | 1.1 | 0.5 | -0.3 |
| drought | 0.7 | 1.0 | 0.0 | 4.8 | 1.0 | 1.6 | 1.8 |

We can gather the following insights from the table:

1. Variables **PRECTOT**, **QV2M**, and **drought** are strongly positively skewed, meaning most of their values are concentrated to the left (small values) with few higher values in the right. Other variables are also positively skewed, but with less intensity. On the other hand, the PS variable is strongly negatively skewed, meaning most of its values are large with few small values in the left.
2. Variables **PRECTOT**, **PS**, and **drought** have high positive kurtosis values which entail that their distribution is heavy-tailed, while variables **T2M**, **T2M_MAX**, and **T2M_MIN**, **TS** have low kurtosis values which entail that their distribution is light-tailed
3. Only temperature-related variables have negative values.
4. Wind-related variables at different spatial resolutions (10 meters and 50 meters) have similar skewness and kurtosis values. This suggests that we can use one group of the variables since they have similar distribution.

Figure 3 shows the distributions of 7 variables.

### 4.2 Bivariate Analysis

In the previous section, we looked at each variable independently. In this section, we will study the relation between two random variables.

Most important of all is correlation. Correlation coefficient quantifies the relation between random variables.

Figure 4 shows the Pearson correlation coefficient between the variables.

This heatmap unlocks very importants insights:

1. **T2MDEW** and **T2MWET** are perfectly positively correlated. Further, we can see that the temperature-related variables form a cluster, where the rows from **T2M** to **T2M_MIN** mostly colored in red, indicating higher correlation.
2. Similarly, the **WS10M** and **WS50M** have correlation value of 0.97. And the wind-related features also form some kind of cluster with the group of red squares next to each other in the bottom right representing all different wind-related features.
3. The **drought** variable is *negatively* correlated with **PRECTOT** and **PS**, and *positively* correlated with **T2M_MAX**, **T2M_RANGE**, and **WS10M_RANGE**. In other words, higher precipitation and air pressure is correlated with (or result in) lower drought levels, and higher air temperature and speed correlate with (or result in) higher drought levels.

#### 4.2.1 Hypothesis Testing: correlation significance

Our analysis led us to the question: Is the correlation between the **drought** variable and **PRECTOT**, **PS**, **T2M_MAX**, **T2M_RANGE**, **WS10M_RANGE** statistically significant (significantly different from zero)?

Let's define the null hypothesis as the true population correlation is zero, and the alternative hypothesis as the true population correlation is not equal to zero. We set significance level to 0.05. Table 5 shows that the correlation is actually significant and we reject the null hypothesis.

Table 5: Correlation significance testing

| | Correlation | P-value | Null Hypothesis |
|---|---|---|---|
| PRECTOT | -0.3 | 0.0 | Reject |
| PS | -0.3 | 0.0 | Reject |
| T2M_MAX | 0.1 | 0.0 | Reject |
| T2M_RANGE | 0.4 | 0.0 | Reject |
| WS10M_RANGE | 0.2 | 0.0 | Reject |

### 4.3 Maps

Instead of looking at tables summarizing the drought level across the different states, we can plot them all in a single map, making it easier to understand if there are any patterns.

Figure 5 shows the U.S. choropleth map encoding the average drought per state for the year 2020.

## 5 Regression Analysis

After cleaning the data and making important observations about the independent variables and their relation with the dependent variable, we know focus on regression modelling.

Our data consists of 49 states, each with monthly meteorological and drought data. To design a regression model, we have two choices:

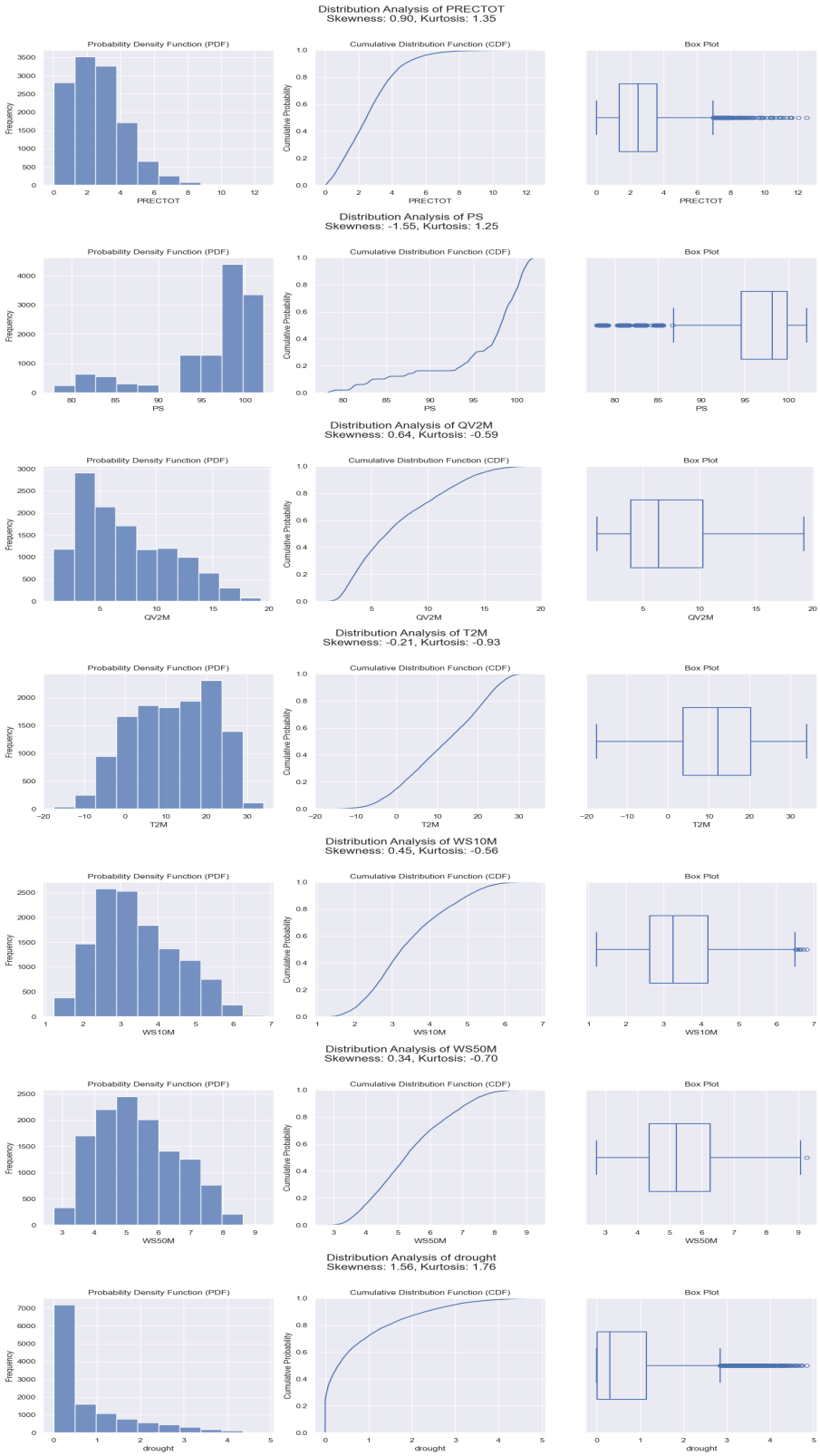


Figure 3: PDF - CDF - Box Plot

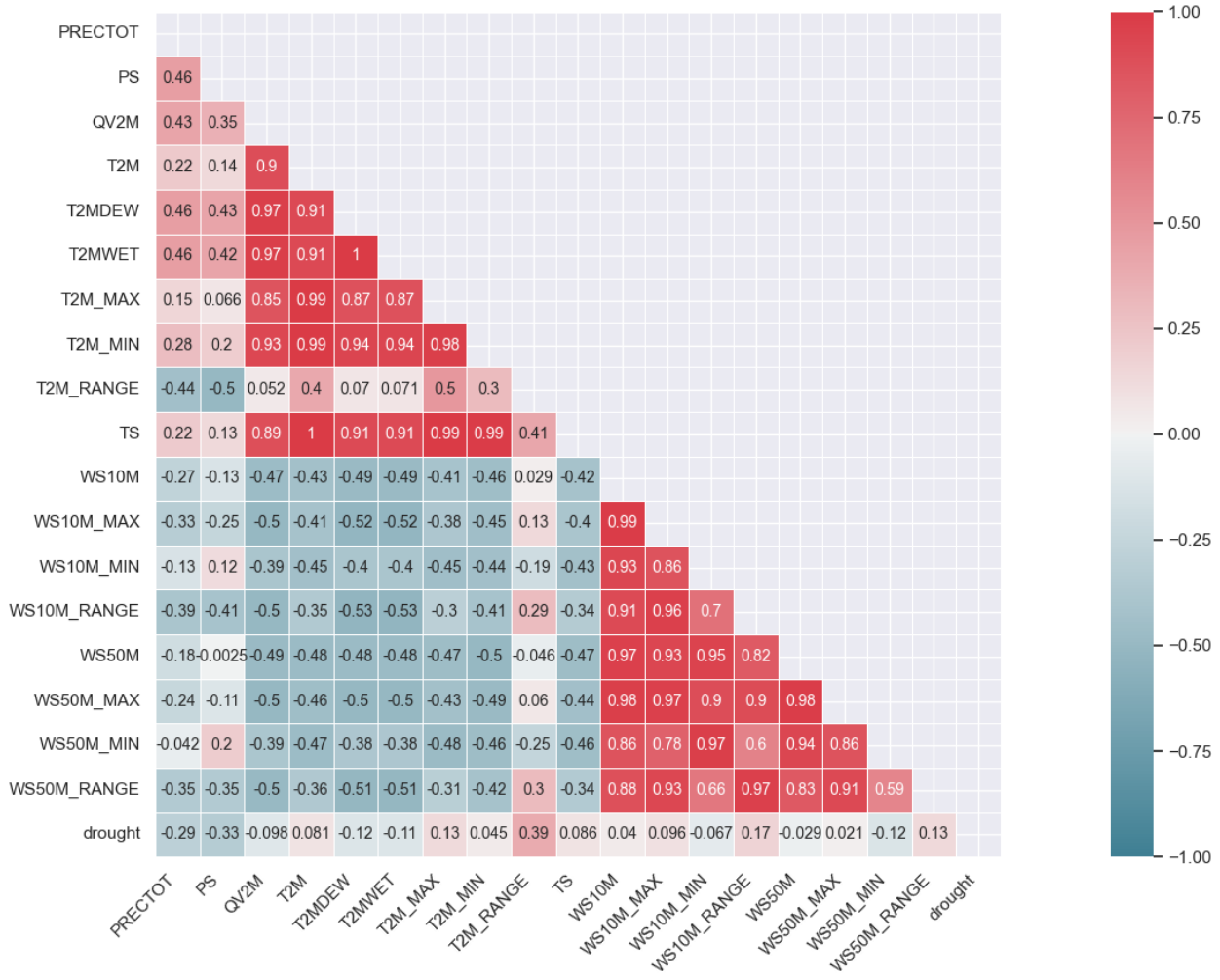


Figure 4: Diagonal correlation heatmap. Red indicates positive correlation and blue indicates negative correlation

1. Aggregate all states data and treat the problem as predicting drought for the entire country. This way we have only *one* dependent variable we want to predict, the average drought over the whole country. However, this has the potential problem of treating different states with different climate conditions as the same. For example, as we have seen from the previous map, drought tend to be higher in the south region than in the north.

2. Build a separate regression model for each state. While this solution seem more computationally expensive, but it's more appealing because it treats each state as an independent variable with its distributions. This solution also has the problem of not taking into account the interaction in the drought variable across different states. For example, if a states is affected by severe drought caused by natural disaster or man-made, then it's plausible to assume that nearby states will also exhibit similar conditions. We will discuss in the future work part how this problem is typically addressed.

In our work, we will follow the second approach and build 8 different regression models for eight different states that have varying levels of drought. Table 6 shows the states that we will build regression models for.

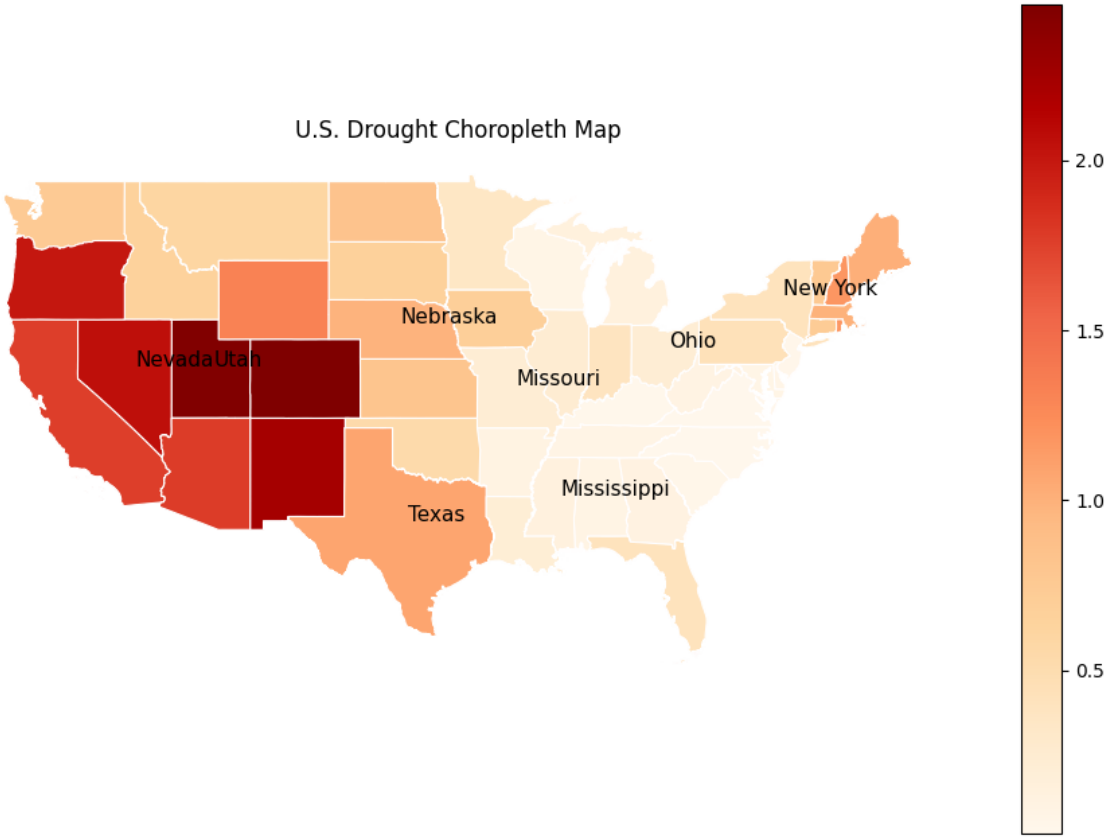


Figure 5: Average Drought by state for year 2020

Table 6: States drought level

| state | drought_level |
|---|---|
| Nevada | very high |
| Utah | very high |
| Nebraska | high |
| Texas | high |
| Mississippi | medium |
| Missouri | medium |
| New York | low |
| Ohio | low |

### 5.1 Data Leakage in Time Series

Unlike tabular data, in time series data we can't split the data randomly into train and test. The main idea for using test data is to measure the model ability to generalize on unseen data points. In the context of time series data, we have to make sure that the test data represents a time frame that wasn't part of the training data.

We utilize the **temporal_train_test_split** function from sktime package to split our data *temporally*. The training data consists of 8624 records (70%) and the test data consists of 3724 records (30%). The train and test data have the following ranges:

1. Range of dates in training data: **2000-01-31** to **2014-08-31**
2. Range of dates in testing data: **2014-09-30** to **2020-12-31**

This way of splitting the data ensures that there will be no leakage between train and test data.

### 5.2 Feature Engineering

Let's first formally define the problem: given the meteorological data for a state $X_{state}$ and its corresponding drought level $y_{state}$, we are interested in learning the function:

$$y_{state} = f_{state}(X_{state})$$

$X_{state}$ is the independent variable (also called exogenous variable) and $y_{state}$ is the dependent variable (also called endogenous)

Both $X_{state}$ and $y_{state}$ consists of timely recorded data points.

We can't use this data directly as input for linear regression model because it need tabular data. To address this, we use the Window Summarizer from sktime. This function generates *features* from the time series data.

From its name, this function takes a specified window range (one week, one month, etc) and a window or aggregation function (lag, mean, max, etc) and apply the function for each window in the data. Thus, *summarizing* the temporal data and converting it to tabular data ready to be used by regression models.

In our case, we select from the meteorological data only the columns that had high correlation with the target column: PRECTOT, PS, T2M_MAX, T2M_RANGE, WS10M_RANGE. At each time point t, we generate lag features from these column along with the drought column for previous month, two, three, and four months. Table 7 shows the generated features and their description. At a particular month, each of these features are calculated by looking back one, two, three, and four months back. We apply this transformation on the training and testing datasets separately.

Table 7: Time series features

| column | description |
|---|---|
| PRECTOT_lag_1 | PRECTOT value for one month before |
| PRECTOT_lag_2 | PRECTOT value for two months before |
| PRECTOT_lag_3 | PRECTOT value for three months before |
| PRECTOT_lag_4 | PRECTOT value for four months before |
| PS_lag_1 | PS value for one month before |
| PS_lag_2 | PS value for two months before |
| PS_lag_3 | PS value for three months before |
| PS_lag_4 | PS value for four months before |
| T2M_MAX_lag_1 | T2M_MAX value for one month before |
| T2M_MAX_lag_2 | T2M_MAX value for two months before |
| T2M_MAX_lag_3 | T2M_MAX value for three months before |
| T2M_MAX_lag_4 | T2M_MAX value for four months before |
| T2M_RANGE_lag_1 | T2M_RANGE value for one month before |
| T2M_RANGE_lag_2 | T2M_RANGE value for two months before |
| T2M_RANGE_lag_3 | T2M_RANGE value for three months before |
| T2M_RANGE_lag_4 | T2M_RANGE value for four months before |
| WS10M_RANGE_lag_1 | WS10M_RANGE value for one month before |
| WS10M_RANGE_lag_2 | WS10M_RANGE value for two months before |
| WS10M_RANGE_lag_3 | WS10M_RANGE value for three months before |
| WS10M_RANGE_lag_4 | WS10M_RANGE value for four months before |
| drought_lag_1 | drought value for one month before |
| drought_lag_2 | drought value for two months before |
| drought_lag_3 | drought value for three months before |
| drought_lag_4 | drought value for four months before |

After generating these window features for each different state, we build a simple linear regression model for each state and evaluate against the test data.

## 6 Performance evaluation

Now we have 8 different regression models for each state. We evaluate each model using several evaluation metrics. Table 8 shows the different metrics for each model. We can see that most models have good $R^2$ score (close 1) except for Ohio.

Figure 6 shows the training, test, and prediction for the 8 different states.

Table 8: Evaluation metrics

| State | Explained Variance | Mean Absolute Error | Mean Squared Error | R$\hat{2}$ Score | Residual Sum of Errors | Root Mean Squared Error |
|---|---|---|---|---|---|---|
| Nevada | 0.9 | 0.3 | 0.2 | 0.9 | 16.0 | 0.5 |
| Utah | 0.9 | 0.2 | 0.1 | 0.9 | 4.7 | 0.2 |
| Nebraska | 0.6 | 0.2 | 0.1 | 0.6 | 8.5 | 0.3 |
| Texas | 0.8 | 0.3 | 0.1 | 0.8 | 8.0 | 0.3 |
| Mississippi | 0.5 | 0.3 | 0.2 | 0.5 | 14.2 | 0.4 |
| Missouri | 0.6 | 0.2 | 0.1 | 0.6 | 9.4 | 0.4 |
| New York | 0.8 | 0.2 | 0.1 | 0.8 | 5.3 | 0.3 |
| Ohio | 0.1 | 0.2 | 0.1 | 0.1 | 4.4 | 0.2 |

## 7 Conclusion and Future work

In this work, we showed how we can use simple linear regression model with time series feature engineering to predict drought at the state level. The drawbacks of our approach is that we couldn't handle the time-series data as is and we couldn't capture the inter-relation of the drought variable between states.

Additionally, in our work, we predicted the drought values for 6 years interval. In reality, we need to have real-time predictions to provide us with up-to-date monitoring on severe conditions.

In time series literature, this type of data is called Panel Data, and it has its corresponding regression algorithms, including PanelOLS[10]

In [6] and [2] the authors utilized novel LSTM Deep Learning models to solve the prediction task. They used similar meteorological data in addition to oceanic and vegetation indices, and land cover data to improve the performance. The first paper addresses the spatial-challenges by predicting drought at the county-level without the need to perform aggregation to the state-level.

## References


[1] These NASA images show the staggering impact of drought — weforum.org. https://www.weforum.org/agenda/2022/08/nasa-lake-mead-water-drought/. [Accessed 11-01-2024].

[2] Akanksha Ahuja and Xin Rong Chua. Forecasting global drought severity and duration using deep learning. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022.

[3] https://www.facebook.com/bbcnews. Olive oil price skyrockets as Spanish drought bites — bbc.co.uk. https://www.bbc.co.uk/news/world-europe-67565503. [Accessed 11-01-2024].

[4] https://www.theguardian.com/profile/helenasmith https://www.theguardian.com/profile/sarahbutler, https://www.theguardian.com/profile/samjones. Europe's olive oil supply running out after drought – and the odd hailstorm — theguardian.com. https://www.theguardian.com/world/2023/sep/28/europes-local-olive-oil-supply-runs-almost-dry-after-summer-of-extreme-weather. [Accessed 11-01-2024].

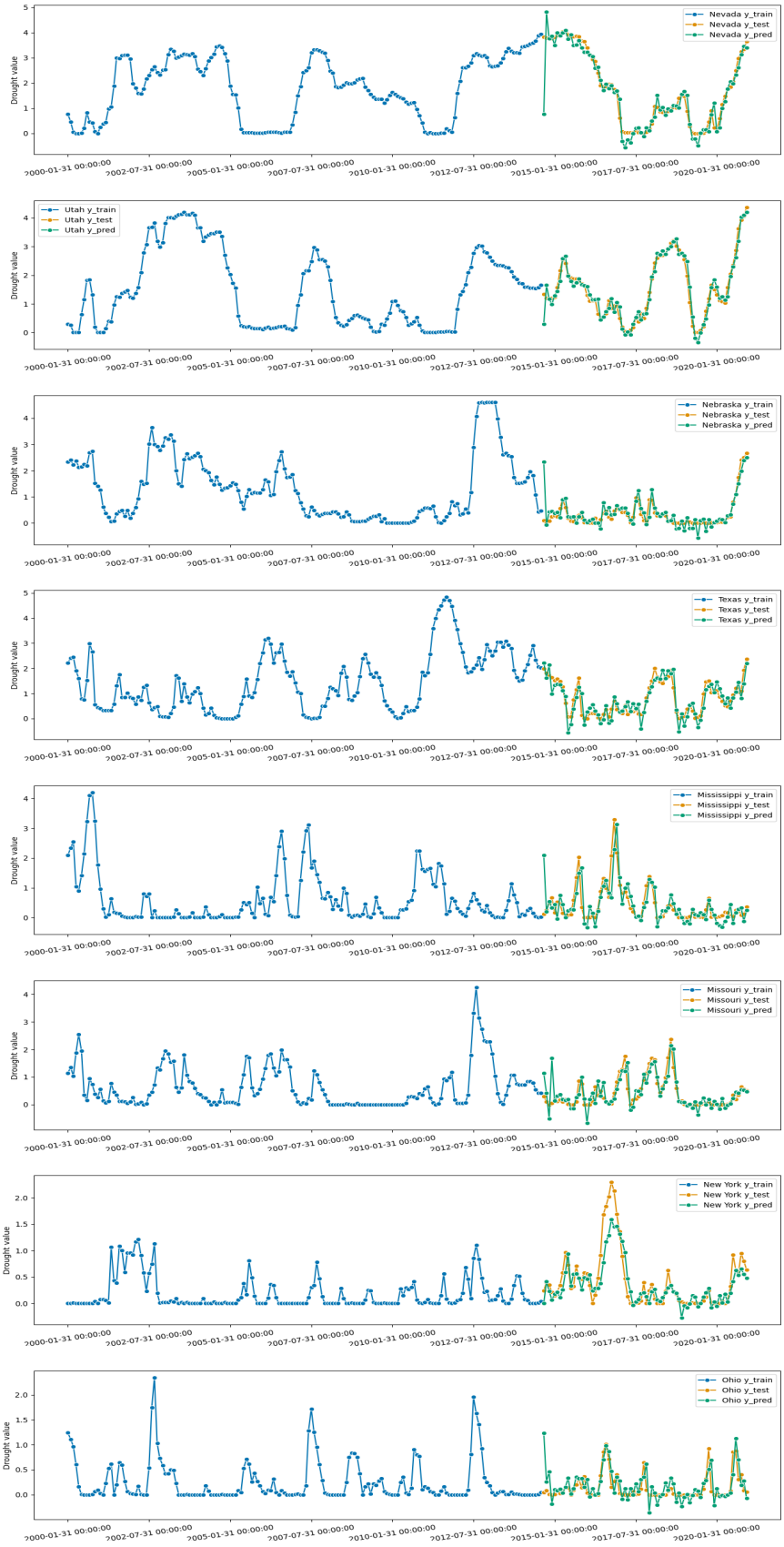


Figure 6: Time series plot for each state

[5] Christoph Minixhofer. Predict Droughts using Weather & Soil Data — kaggle.com. https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data/data, 2021. [Accessed 10-01-2024].

[6] Christoph D Minixhofer, Mark Swan, Calum McMeekin, and Pavlos Andreadis. Droughted: A dataset and methodology for drought forecasting spanning multiple climate zones. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

[7] National Aeronautics NASA Langley Research Center Atmospheric Science Data Center GIS, Atmospheric Sciences Data Center and 2018 Space Administration, Langley Research Center. Nasa/power agroclimatology data. 2018.

[8] University of Nebraska-Lincoln National Drought Mitigation Center. U.s. drought monitor. 2014.

[9] United States Census Bureau. U.s. census bureau.

[10] Jeffrey M Wooldridge. Econometric analysis of cross section and panel data. *MIT press Cambridge*, 2010.