Time Series Analysis of Drought Areas Since 1895

Abstract

Drought is a natural hazard in which an area or region experiences below-normal precipitation during a period. With the increasing occurrence of climate change throughout recent years, drought is expected to become more severe in terms of economic, social, and ecological damage. Therefore, drought has for centuries shaped the societies in the United States and will continue to do so into the future. To understand the patterns of drought in the U.S in the past as well as the future, we propose several time-series models to compare the frequency of severe drought in the U.S before and after 1957. In particular, we consider Holt-Winters, Seasonal Autoregressive Integrated Moving Average (SARIMA), Periodogram, and ARIMA with cosine-sine models as potential models for our drought data. We visualized the models' predictions of the drought pattern before and after 1957, showing substantial differences in the trend and patterns of drought in the United States.

I. Introduction

Drought is a natural phenomenon that occurs when there is a shortage of precipitation throughout a prolonged period. While not typically seen as a natural disaster like hurricanes, snowstorms, and floods, droughts happen more frequently and leave damaging impacts on multiple industries, especially agriculture. Droughts in 2012 cost USD 77.6 billion and the heatwave killed 7,500 lives. This was the second severe drought since the drought in 1988. In September 2012, 65.6% of the US experienced moderate to severe drought. Droughts have direct impacts on agriculture as they could reduce production. Low precipitation leads to a low water supply, thus threatening wildlife and limiting access to water among the public.

Research and studies to monitor and predict droughts met with challenges concerning the definitions of drought and the signals of the beginning and end of a drought period. Droughts can be classified into 5 categories: meteorological (increasingly dry weather patterns), hydrological (evident low water supply), agricultural (crops loss), socioeconomic (supply and demand of commodities), and ecological (natural ecosystem). Whether a lack of precipitation can be seen as a drought event or not depends on the definitions and the geographic areas.

In this study, we investigated historical meteorological drought conditions in the US using the data collected from NOAA's National Centers for Environmental monthly from 1895 to the present. The dataset stores information about the percentage of contiguous US that experienced Abnormally Dry to Severe Drought and Abnormally Wet to Severe Wet. The categorization is used to determine the Standardized Precipitation Index (SPI). Within the scope of this study, we focus on the univariate time-series analysis of the percentage of the US that is abnormally dry. We attempt to try Holt-Winters smoothing models, Seasonal ARIMA and ARIMA with cosine-sine pairs to model this time series. The goals of the study are: (i) to **detect patterns** within drought time series, (ii) to **propose** a model that can capture closely the trend and predictions of the time series, and (iii) to **provide short-term forecasts** of future droughts.

II. Dataset

2.1 Dataset Description

U.S. Gridded Standardized Precipitation Index (SPI) is a collection of Standardized Precipitation Index (SPI), the percentage of areas in the U.S with different levels of drought and wet. Standardized Precipitation Index (SPI) is the number of standard deviations that observed cumulative precipitation deviates from the climatological average to characterize meteorological drought on a range of timescales, ranging from 1 to 72 months. Both dryness and wetness are divided into five categories: Abnormally (Level 0), Moderate (Level 1), Severe (Level 2), Extreme (Level 3), and Exceptional (Level 4). The data is collected monthly by the NOAA's National Centers for Environmental Information from 1985 to 2022. In total, the data contains 1,527 rows with 13 columns.

2.2 Exploratory Data Analysis

Out of the indexes in our data, in this paper, we focus on the percentage of areas in the U.S that are abnormally dry, which is explained by the column "D0" in our dataset mentioned above. From Figure 1, the percentage of abnormal dryness in the U.S appears to be stationary. Our Dicky-Fullers test also results in a p-value of 0.01, which also indicates that our data is stationary.

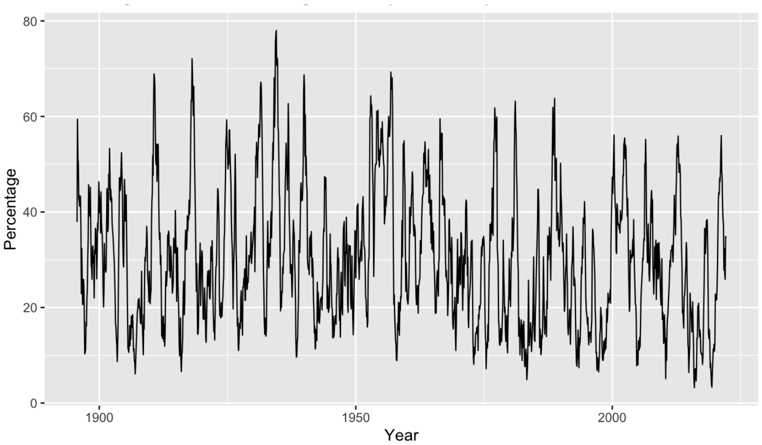


Figure 1: Percentage of abnormally dry areas in the U.S from 09/1895 to 03/2022.

From Figure 2, we can see that the drought data does not have any clear repeating trends. Instead, the trend plot displays an oscillation with random amplitude. From the remainder plot, the remainders, despite displaying alternating signs, have relatively similar amplitude through the years. The values of the percentage of Abnormal Dry in the U.S are concentrated the most in the range from 25 to 40, as displayed in Figure 3.

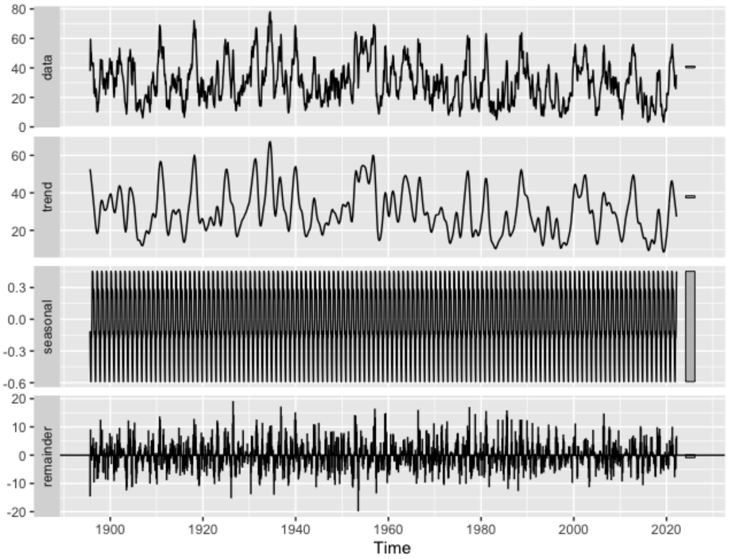


Figure 2: Trend, seasonality, and remainder of abnormally dry time series (09/1895 - 03/2022).

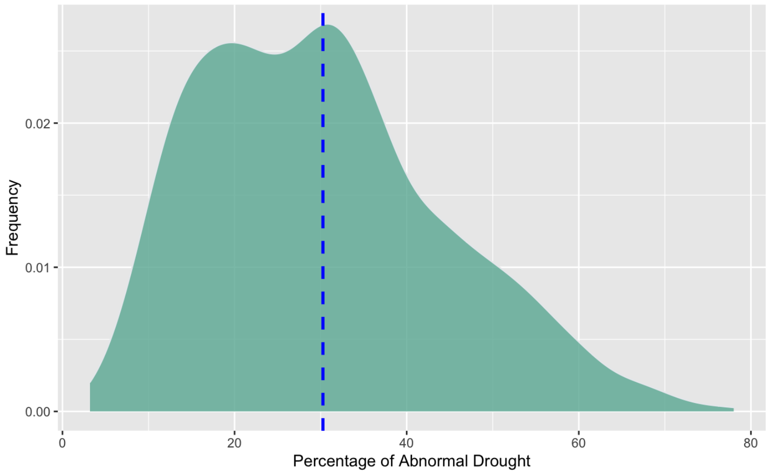


Figure 3: Density plot of the percentage of abnormally dry areas (09/1895 - 03/2022)

2.3 Data Subsetting

We focus on the percentage of areas in the U.S that are abnormally dry, which is explained by the column “D0” in our dataset mentioned above. After considering the time-series plot of drought percentage, we have decided to split the dataset into two separate subsets: One subset from 09/1895 to 12/1956 and the other from 01/1957 to 03/2022. Based on Figure 3, we choose 1957 as the splitting point because, after 12/1956, the percentage of dry areas in the US seems to follow a new cycle after a small period with extremely little drought. Moreover, the percentage of wet areas seems to increase and fluctuate more, causing the percentage of dry and wet areas abnormally close to each other. For both each subset and the original data, we partition the individual datasets last year as validation data for our time-series model.

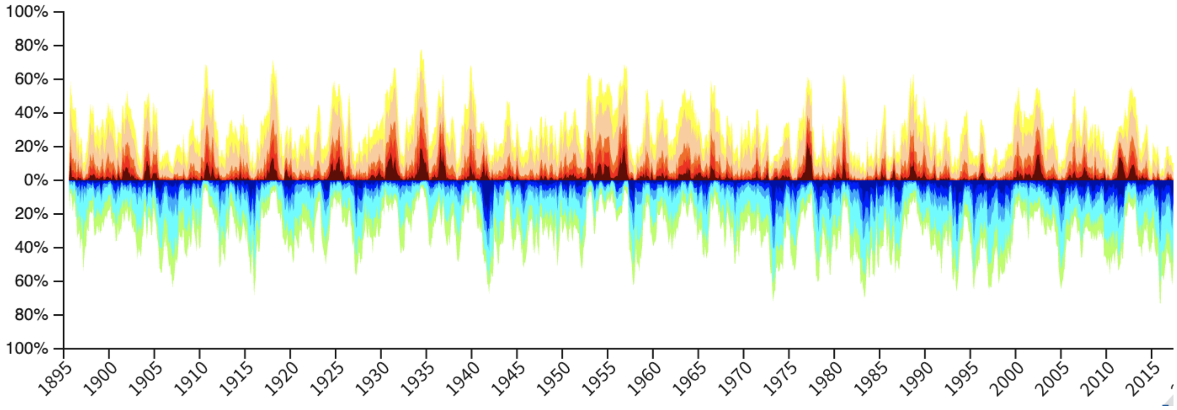


Figure 4: Percentage of dry areas (red, orange) vs percentage of wet areas (blue) in the United States (09/1895 - 03/2022).

(Source: Historical Data and Conditions - National Integrated Drought Information System)

2.4 Hypothesis

Based on the exploratory data analysis, we **hypothesize** that the percentage of US areas that experienced abnormal dry to severe drought conditions becomes more volatile in the later periods. We also **hypothesize** that the percentage of abnormally dry areas could be forecasted using a time series model with seasonal components.

III. Methodology

3.1 Holt-Winters Methodology

Holt-Winters is a modeling method that was created from a generalization of the exponential smoothing method in 1957, being fully codified in 1960. This process will be discussed as described in *The Holt-Winters Forecasting Procedure* (Chatfield). The exponential smoothing method is integral to this process. The exponential smoothing method works by the application of a smoothing component to estimate values. This smoothing term is used as follows:

$$
\begin{gathered}
a = \alpha y_t + (1-\alpha)\alpha_{t-1} \\
a_{t-1} = \alpha y_{t-1} + (1-\alpha)\alpha_{t-2} \\
\dots \\
a_t = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_t \text{ ...} \\
a_t = \sum_{i=0}^{n} \alpha(1-\alpha)^i y_{t-i}
\end{gathered}
$$

As we can see, this method accounts for gradual change in the data but falls flat in some regards. In data that experience structural changes, especially in trend, or fluctuates with seasonality, exponential smoothing will fail. For these reasons, Holt-Winters Modeling was created, allowing the application of exponential smoothing-like techniques on more complicated data.

The method is a combination of three integral equations, an equation for mean, an equation for trend, and an equation for seasonality presented below in that order.

$$
\begin{gathered}
a_t = \alpha(y_t - s_{t-p}) + (1-\alpha)(\alpha_{t-1} + \beta_{t-1}) \\
b_t = \beta(a_t - a_{t-1}) + (1-\beta)(b_{t-1}) \\
s_t = \gamma(y_t - a_t) + (1-\gamma)(s_{t-p})
\end{gathered}
$$

These equations are generated in two methods, additively or multiplicatively. Of these two processes, only the Additive Holt-Winters Model was relevant to this work.

In Additive Holt-Winters, mean, trend, and seasonality components are combined through addition, creating the final Holt-Winters Model as a whole.

$$
y_{t+h} = a_t + h * b_t + s_{t-p+1+(h-1)modp}
$$

Through the additive process, four different models can be generated. The exponential smoothing model, which is without the trend or seasonality components, a model with only trend and mean, a model with seasonality and mean, and a model with all components, which is a Full Holt-Winters model. The ability to decompose trend and seasonality into separate components is one of the key features of the Holt-Winters modeling technique. By attempting to model a time series by its integral components, Holt-Winters is both flexible and accurate.

3.2 Spectral Analysis

Frequency domain analysis, or spectral analysis, concentrates on the frequency properties of the time series. A time series with repeated cycles could be fit with cosine curves. The cosine curve has an equation

$$Rcos(2\pi ft + \phi)$$

The cosine function could be parameterized using trigonometric identity:

$$Rcos(2\pi ft + \phi) + = Acos(2\pi ft) + Bsin(2\pi ft)$$

$$R = \sqrt{A^2 + B^2}, \phi = atan(-B/A)$$

$$A = Rcos(\phi), B = -Rsin(\phi)$$

The final model is a linear combination of m cosine curves:

$$Y_t = A_0 + \sum_{j=1}^{m} [A_j cos(2\pi f_j t) + B_j(2\pi f_j t)]$$

Once identifying the frequencies, OLS regression is used to fit As and Bs. If the sample size is odd, the least square estimates would be:

$$\hat{A}_0 = \bar{Y}$$

$$\hat{A}_j = \frac{2}{n}\sum_{t=1}^{n} Y_t cos(2\pi tj/n) \qquad \hat{B}_j = \frac{2}{n}\sum_{t=1}^{n} Y_t sin(2\pi tj/n)$$

If the sample is even, then estimates are:

$$\hat{A}_k = \frac{1}{n}\sum_{t=1}^{n} (-1)^t Y_t, \hat{B}_k = 0$$

The periodogram, a spectral analysis tool, is used to explore the hidden frequencies in the data. Cran explained in depth this concept in the textbook Time Series Analysis in R: "The periodogram is the sum of squares with two degrees of freedom associated with the coefficient pair (A_j, B_j) at frequency j/n". It is a graph that shows the value of the spectrum against the frequency of the time series. The high spike in the periodogram reflects the significant strength of the cosine-sine pairs at the specified frequency in the overall time series. The interval in the periodogram is between 0 and ½ as the cosine repeats after a frequency f=1/2. For 0 <= f <= ½:

$$I(f) = \frac{n}{2}(\hat{A}_f^2 + \hat{B}_f^2)$$

$$\hat{A}_f = \frac{2}{n}\sum_{t=1}^{n} Y_t cos(2\phi tj/n) \qquad \hat{B}_f = \frac{2}{n}\sum_{t=1}^{n} Y_t sin(2\phi tj/n)$$

Analyzing time-series data using spectral analysis starts with examining the stationarity of the time series and then exploring frequencies using a periodogram. Nonstationary data needs to be detrended to observe the accurate spikes in the periodogram. A large dataset could also cause noisy spikes so a smoothed periodogram may be considered. With the known frequencies, we can fit a regression model for the sine and cosine terms to estimate A's and B's. Similarly to regression analysis, we can drop insignificant terms and check the residuals of the model using the Ljung-Box test and visual plots to see whether the residuals are uncorrelated white noise. If the residuals exhibit some AR or MA patterns, we can consider fitting the ARIMA model in addition to the cosine-sine model. The model selection and validation process are similar to normal ARIMA models. Several models are proposed and compared based on AICc, BIC on the training set, and RMSE on the validation set. The model with the least RMSE and AICc would be chosen for forecasting.

3.3 Seasonal ARIMA Methodology

Multiplicative seasonal ARMA models are the combining idea of seasonal and nonseasonal ARMA that contain autocorrelation for seasonal lags but also low lags of neighboring series value (Cryer & Chan, 2011). Like ARMA, multiplicative seasonal ARMA requires the time series to be stationary so the first difference and a seasonal difference might be applied if the time series is not stationary.

Multiplicative seasonal ARMA(p,q)×(P, Q)s model with seasonal periods as a model with AR characteristic polynomial φ(x)Φ(x) and MA characteristic polynomial θ(x)Θ(x) are defined as:

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p$$
$$\Phi(x) = 1 - \Phi_1 x^s - \Phi_2 x^{2s} - \cdots - \Phi_p x^{ps}$$

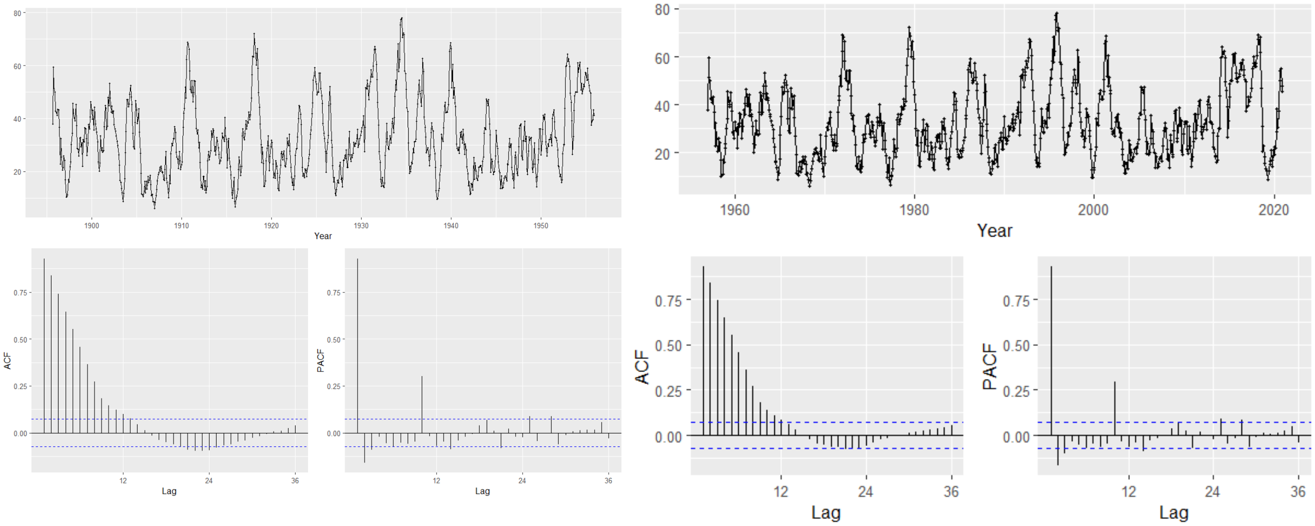
and

$$\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^q$$
$$\Theta(x) = 1 - \Theta_1 x^s - \Theta_2 x^{2s} - \cdots - \Theta_q x^{qs}$$

The first step in modeling using SARIMA is a model specification that requires a careful inspection of the time series plot. The plot of the time series, ACF, PACF, and Augmented Dickey-Fuller test would be used to determine the stationarity of the data and the need for first or seasonal difference. After that, we would perform model fitting and determine the optimal model based on AIC, AICc, and BIC. Then, we perform model diagnostics by looking at the time series, ACF, and PACF plot of the residuals as well as perform Ljung-Box to officially test the autocorrelation among the residuals. We would prefer a high p-value from the Ljung-Box test so that we would fail to reject the null hypothesis that the residuals are independently distributed. Finally, the model would be used to make forecasts and cross-validate with the testing set. The best model would be used to make forecasts for the future.

IV. Application

4.1 Stationarity



Figures 5 and 6: Time series, ACF, PACF plots of percentage drought in 1895-1956 (left) and 1957 - 2022 (right)

The dataset is split into 2 subsets to allow for a better fit. The time series of the drought percentage in the 1895-1965 period is not stationary since the variance appears more frequently and increases as time passes. Similarly, the drought percentage in 1957-2022 has the same patterns but with more volatilities.

4.2 Holt-Winters

Holt-Winters modeling, with its ability to separate trend and seasonality, was a necessary modeling type to consider with our data. To generate separate models for both the years 1895-1956 and 1956-2022, the full modeling approach was performed twice. Once for the oldest set of data, and once for the more recent past.

a. Period 1895-1956

In the time series from 1895-to 1956, the first step was to create all forms of Holt-Winters models and then eliminate them comparatively. Additive Holt-Winters was determined to be the best approach to seasonality in this data. Since all Holt-Winters models must contain a mean component, the models will be differentiated by their inclusion of seasonality and trend terms.

The first model considered was a model with no seasonality or trend, an exponential smoothing model. This model will serve as a baseline to compare against and is not expected to perform the best. If this is the case, then Holt-Winters modeling would be useless in this dataset. The next model considered was a model with only the addition of a trend component. Third, a model with only the addition of seasonality. These models serve to balance each other, perhaps the data has discernible trends but no seasonality, or perhaps there is no trend, but a presence of seasonality. The flexibility of Holt-Winters allows us to simply account for both. Finally, there is a full Holt Winter model, with both the additionality of seasonality and a trend component.

These models were created on the training set for the 1895-1956 data, and were then used to predict against the validation set using the forecast() function. To determine which model was to be selected, both a visual and mathematical inspection was required. A graph of these models compared against both each other and the validation data is included below.

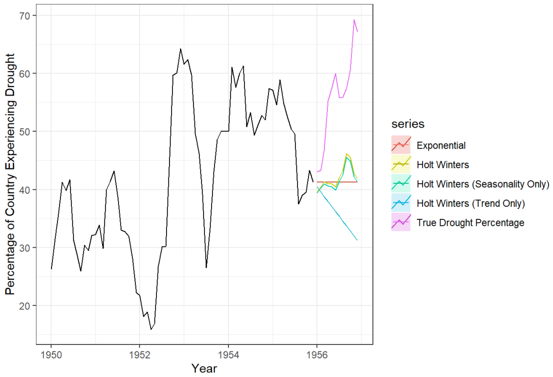
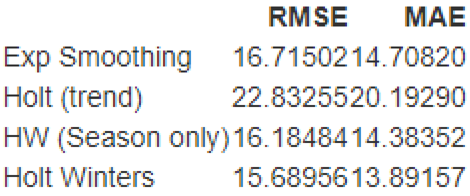


Figure 7: Comparison between Holt-Winters models and 1895 -1956 data

From a brief visual inspection, it appears that while none of the models is a perfect fit for the data, two of the models appear to capture the spirit of the data. To come to a more sound conclusion, mathematical results are required.



	RMSE	MAE
Exp Smoothing	16.71502	14.70820
Holt (trend)	22.83255	20.19290
HW (Season only)	16.18484	14.38352
Holt Winters	15.68956	13.89157

Table 1: RMSE, MAE of smoothing methods on 1895 -1956 time series

From these results, we can see that the Full Holt-Winters model has the joint lowest RMSE and MAE. This model contains both trend and seasonality components, and since this model is the most accurate both seem to affect its ability to model the overall data. The Full Holt-Winters Model is the model selected for the oldest time frame (1895-1956).

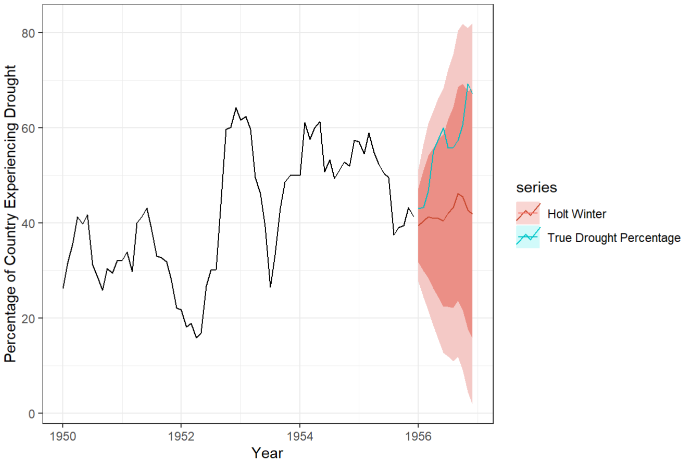


Figure 8: Comparison between 1895-1956 validation set and full holt winters model

By comparing the Full Holt-Winters model against the validation set with prediction intervals, we can see the data is contained within the prediction bands. Sadly, the prediction bands are quite wide, but this should still give us an idea of the mean predicted values, which are important to discern if the change has occurred between the distant and recent past.

b. Period 1957-2022

The second set to be modeled with Holt-Winters is the recent past, the period of droughts from 1956-to 2022. The methodology behind the generation of a final model in this time frame is nearly identical to that used for the prior interval, that is the generation of four separate kinds of Holt-Winters models.

The first model to be generated is the Exponential Smoothing model, which serves as a baseline. This is followed by the Holt-Winters with seasonality and the Holt-Winters with the trend. This diametrically opposed set allows us to discern the absence of a significant key component of prediction. Finally, the Full Holt-Winters model. All these are once again generated against a training set from the majority of the 1956-2022 interval and are compared against a small validation set.

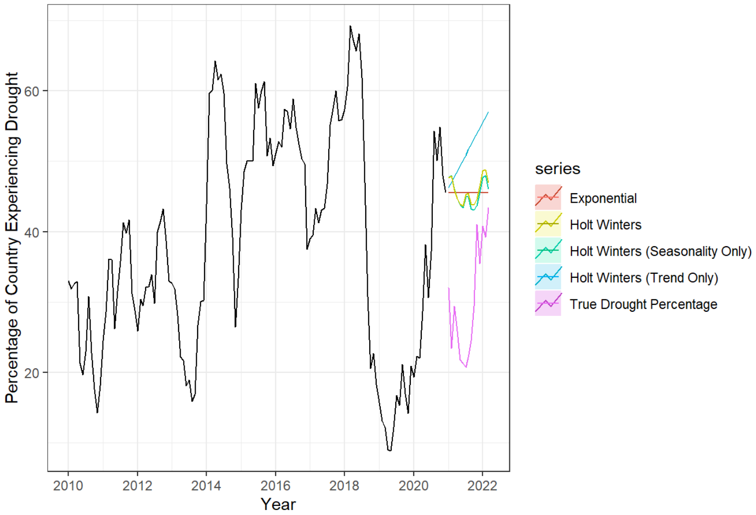
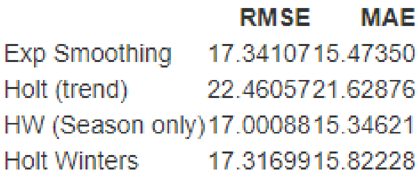


Figure 9: Comparison between Holt-Winters models and 1956-2022 data

When all models are presented on the same graph, we can see that while no model is a great representation of reality, the Holt-Winters and Holt-Winters with only seasonality appear to follow the overall gist of the data. That being said, a mathematical comparison will allow us to see information that may not be present in a visual inspection.



	RMSE	MAE
Exp Smoothing	17.34107	15.47350
Holt (trend)	22.46057	21.62876
HW (Season only)	17.00088	15.34621
Holt Winters	17.31699	15.82228

Table 2: RMSE, MAE of smoothing methods on 1957 - 2022 time series

From the chart, we can see that Holt-Winters with seasonality only has the joint lowest RMSE and MAE values. This will be the model selected.

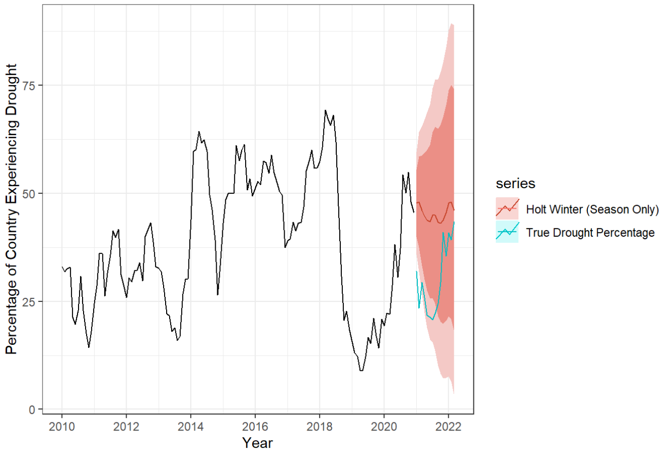
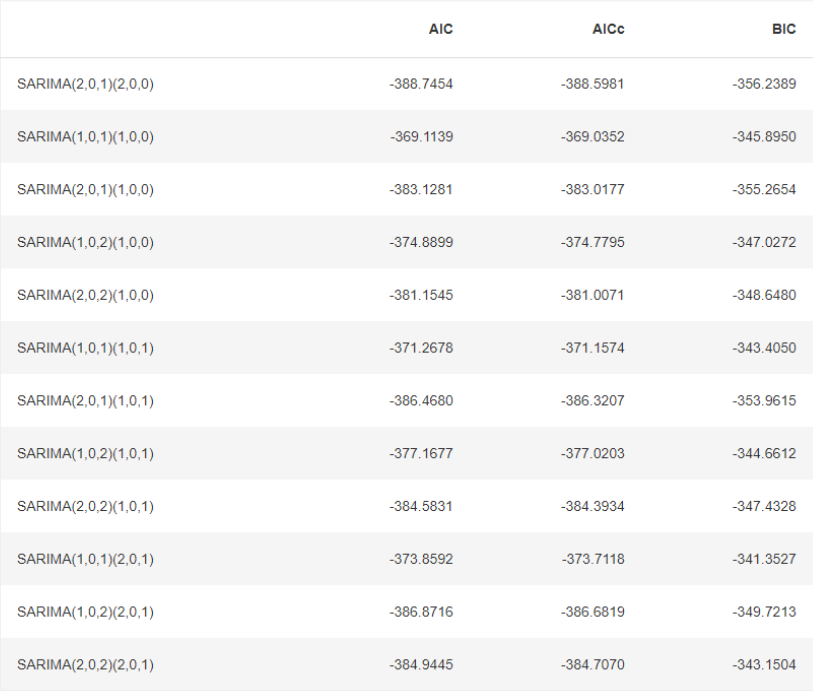


Figure 10: Comparison between 1957-2022 validation set and full Holt-Winters model

Once again, our results are characterized by wide prediction bands. This is indicative of a lack of confidence in overall model activity. While this may be the case, our model still allows us to understand the average value predicted, which can be useful in the future to past comparisons. Overall, our Holt-Winters modeling has yielded promising models that, although their predictions may not be fully accurate, contain with their confidence bounds the data, which means that they can serve as a prediction in the range of true values.

4.3 Seasonal ARIMA

When looking at the time series plot of drought, we observed a high degree of variability, thus we apply a logarithmic transformation to the time series to stabilize the variance. Since the data is stationary, the SARIMA model we start with is SARIMA(p,0,q)(P,0, Q). We increase the AR and MA terms of both the seasonal and nonseasonal subsequently and use AIC, AICc, and BIC to search for the best model. Based on the three criteria, SARIMA(2,0,1)(2,0,0) is the best model for both the new and old-time series of drought percentages (Table 3 and Table 4).



	AIC	AICc	BIC
SARIMA(2,0,1)(2,0,0)	-388.7454	-388.5981	-356.2389
SARIMA(1,0,1)(1,0,0)	-369.1139	-369.0352	-345.8950
SARIMA(2,0,1)(1,0,0)	-383.1281	-383.0177	-355.2654
SARIMA(1,0,2)(1,0,0)	-374.8899	-374.7795	-347.0272
SARIMA(2,0,2)(1,0,0)	-381.1545	-381.0071	-348.6480
SARIMA(1,0,1)(1,0,1)	-371.2678	-371.1574	-343.4050
SARIMA(2,0,1)(1,0,1)	-386.4680	-386.3207	-353.9615
SARIMA(1,0,2)(1,0,1)	-377.1677	-377.0203	-344.6612
SARIMA(2,0,2)(1,0,1)	-384.5831	-384.3934	-347.4328
SARIMA(1,0,1)(2,0,1)	-373.8592	-373.7118	-341.3527
SARIMA(1,0,2)(2,0,1)	-386.8716	-386.6819	-349.7213
SARIMA(2,0,2)(2,0,1)	-384.9445	-384.7070	-343.1504

Table 3: SARIMA model AIC, AICc, and BIC comparison for 1957-2022

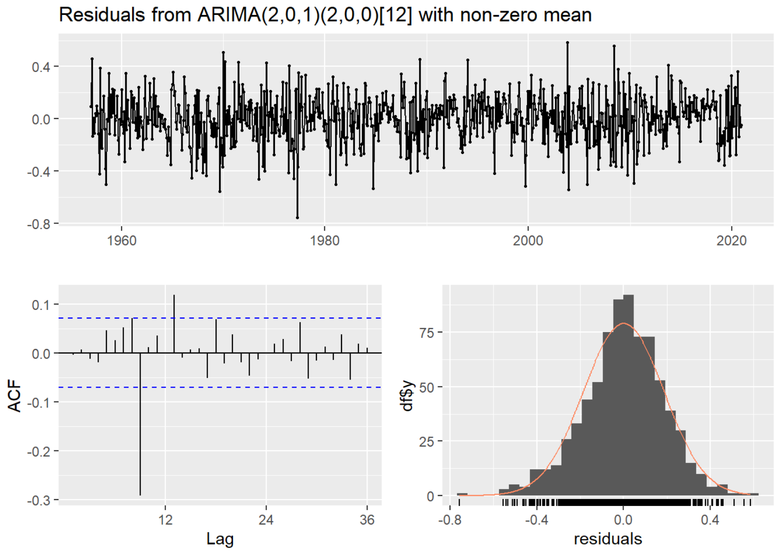


Figure 11: SARIMA(2,0,1)(2,0,0) residuals diagnostics for 1957-2022

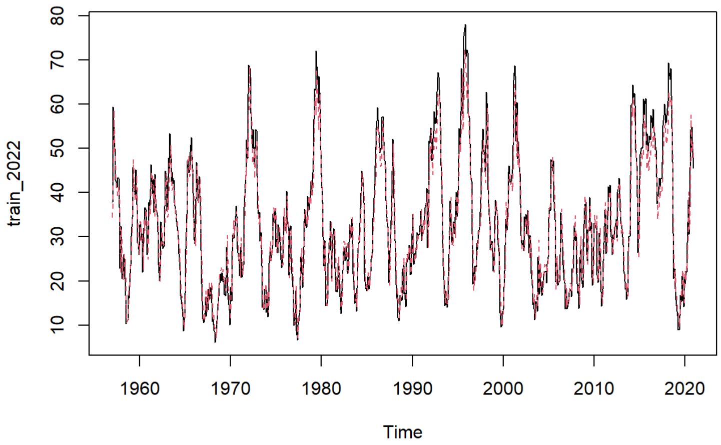


Figure 12: SARIMA(2,0,1)(2,0,0) fitted value against actual drought percentage for 1957-2022

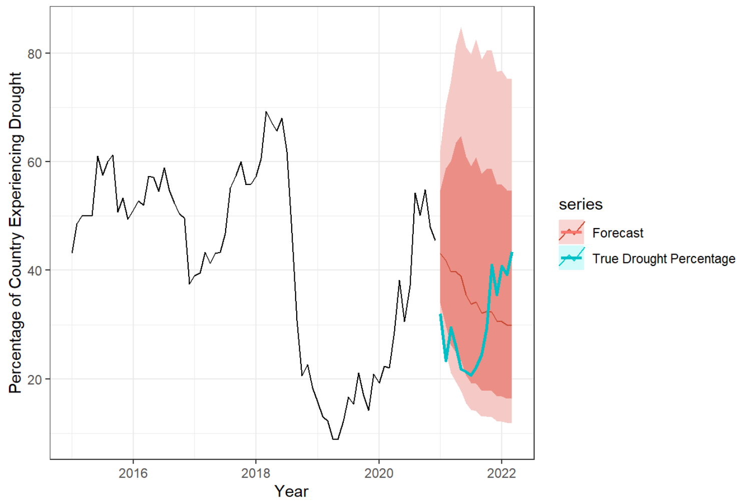
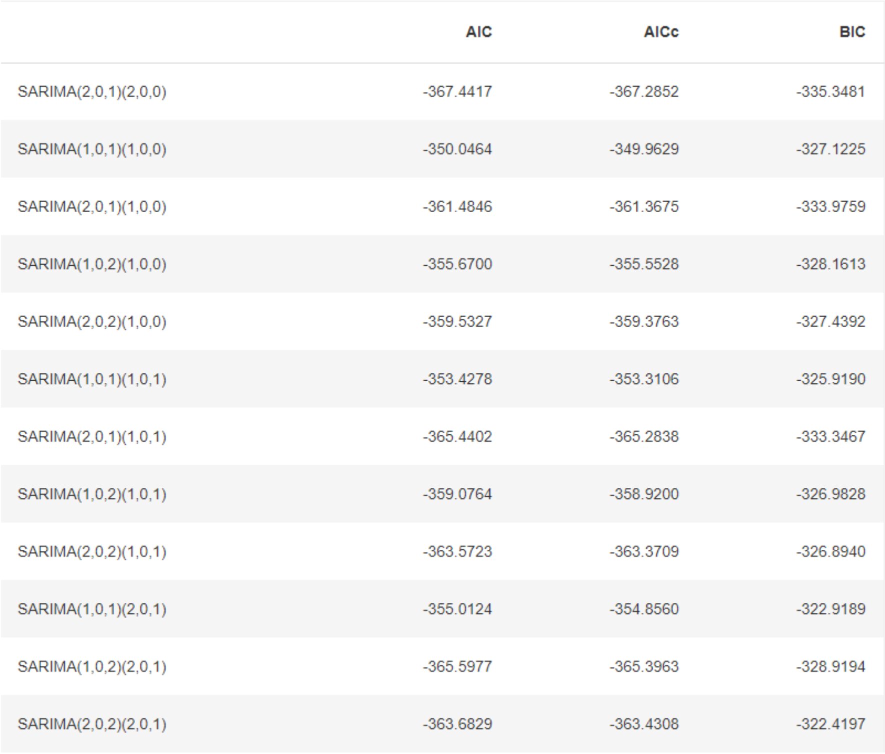


Figure 13: SARIMA(2,0,1)(2,0,0) forecasts against true drought percentage for 1957-2022



	AIC	AICc	BIC
SARIMA(2,0,1)(2,0,0)	-367.4417	-367.2852	-335.3481
SARIMA(1,0,1)(1,0,0)	-350.0464	-349.9629	-327.1225
SARIMA(2,0,1)(1,0,0)	-361.4846	-361.3675	-333.9759
SARIMA(1,0,2)(1,0,0)	-355.6700	-355.5528	-328.1613
SARIMA(2,0,2)(1,0,0)	-359.5327	-359.3763	-327.4392
SARIMA(1,0,1)(1,0,1)	-353.4278	-353.3106	-325.9190
SARIMA(2,0,1)(1,0,1)	-365.4402	-365.2838	-333.3467
SARIMA(1,0,2)(1,0,1)	-359.0764	-358.9200	-326.9828
SARIMA(2,0,2)(1,0,1)	-363.5723	-363.3709	-326.8940
SARIMA(1,0,1)(2,0,1)	-355.0124	-354.8560	-322.9189
SARIMA(1,0,2)(2,0,1)	-365.5977	-365.3963	-328.9194
SARIMA(2,0,2)(2,0,1)	-363.6829	-363.4308	-322.4197

Table 4: SARIMA model AIC, AICc, and BIC comparison for 1895-1956

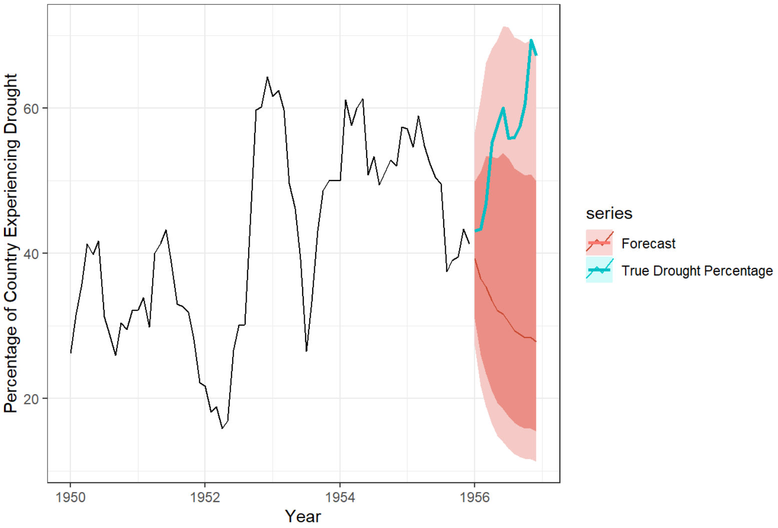


Figure 14: SARIMA(2,0,1)(2,0,0) forecasts against true drought percentage for 1895-1956

Full model equation for SARIMA(2,0,1)(2,0,0) for 1895-1956:

$$Y_t = 0.96Y_{t-1} - 1.273Y_{t-2} - 0.009Y_{t-12} - 0.009Y_{t-13} + 0.012Y_{t-14} - 0.01Y_{t-24} - 0.01Y_{t-25} + 0.013Y_{t-26} + \varepsilon_t - 0.7\varepsilon_{t-1}$$

Full model equation for SARIMA(2,0,1)(2,0,0) for 1957-2022:

$$Y_t = -0.08Y_{t-1} - 0.717Y_{t-2} - 0.02Y_{t-12} - 0.019Y_{t-13} + 0.026Y_{t-14} - 0.008Y_{t-24} - 0.008Y_{t-25} + 0.011Y_{t-26} + \varepsilon_t - 0.717\varepsilon_{t-1}$$

From Figure 12, we can see that the SARIMA(2,0,1)(2,0,0) model fits the data very well. However, when using the model to forecast, the predicted values are far off from the actual values. This might be due to the significant lag at 9 remaining in the residuals diagnostics plot (Figure 11) that we are unable to capture using our model.

The 95% confidence interval can capture the actual value for the new data but for the old data the drought percentage is far outside the 95% confidence interval, but still being captured within the 80% confidence interval. Indeed, we observe a high variation in our forecast that makes our confidence interval too large as well as reduces the effectiveness and usefulness of the model in forecasting future drought percentages in the US.

4.4 Spectral Analysis - ARIMA with Cosine/Sine regressors

To fit the ARIMA model with cosine-sine pairs, we work with single differenced data. The differenced data in both periods appear more stationary at the first-order difference. With both the 1895-1956 train data and 1957-2022 train data, there are some significant lags at k =1 and 2 and an unusual spike at around lag 9 in the ACF/PACF plots. Since there are lots of data points, we use a smooth periodogram with a span of 9 to only capture the most significant frequencies.

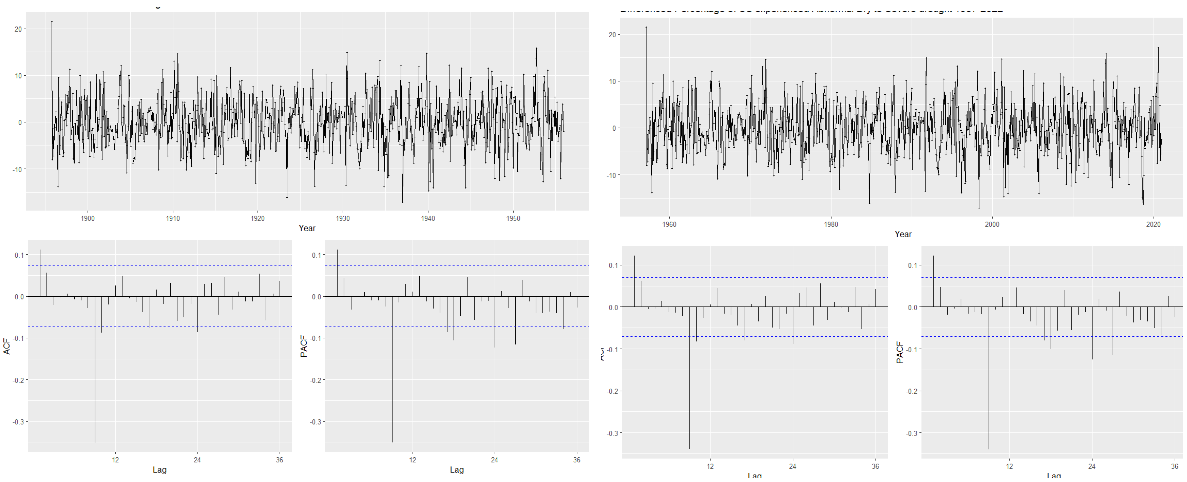


Figure 15: Time series plot, ACF, PACF plots of 1st-difference data 1895-1956 (left) and 1957-2022(right)

a. Period 1895 - 1956

The smooth periodogram shows 4 significant spikes at 0.05, 0.155, 0.275, 0.39. We fit a regression model on the training data with 4 cosine-sine pairs.

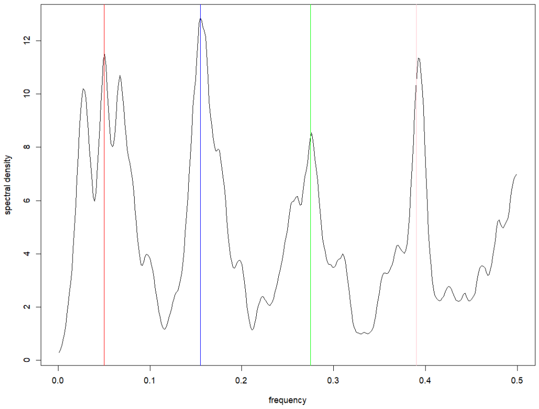
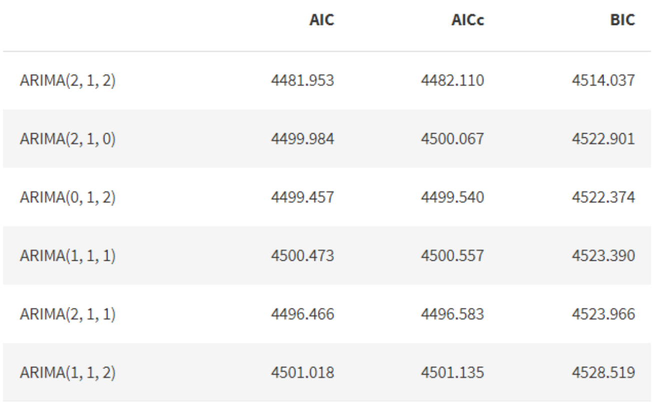


Figure 16: Smooth periodogram (Span 9)

In this first model, only sine components at 0.05 and 0.39 have a p-value under 0.05. We examine the residuals of the new regression model with only sine(0.05). PACF/ACF plots of the residuals of the sin(0.05) model still show a high spike at lag 1 and lag 9 and the residuals failed the Ljung-Box test. Thus, we add ARIMA components to fit this data better. Based on the aforementioned analysis of order in Seasonal ARIMA sections, we can test a variety of ARIMA with AR and MA orders between 0 and 2. We fit 6 ARIMA models with sine(0.05) regression components to the training data and compare their fit using AIC, AICc, and BIC in the below table:



	AIC	AICc	BIC
ARIMA(2, 1, 2)	4481.953	4482.110	4514.037
ARIMA(2, 1, 0)	4499.984	4500.067	4522.901
ARIMA(0, 1, 2)	4499.457	4499.540	4522.374
ARIMA(1, 1, 1)	4500.473	4500.557	4523.390
ARIMA(2, 1, 1)	4496.466	4496.583	4523.966
ARIMA(1, 1, 2)	4501.018	4501.135	4528.519

Table 5: AIC, AICc, BIC of ARIMA models with sine component

ARIMA(2, 1, 2) achieves the best fit to training data, followed by ARIMA(2,1,1). We proceed with these two models and the sine pair to fit testing data of the last 2 years of the period 1955-1956. ARIMA(2,1,2) with sin(0.05) has RMSE of 16.42314 while ARIMA(2,1,1) with sine has RMSE of 16. 76. Thus, ARIMA(2, 1, 2) with sine(0.05) is the most suitable model for the training set. However, the forecasted points from ARIMA(2, 1, 2) with sine are much lower than the true value and can not capture the high volatilities.

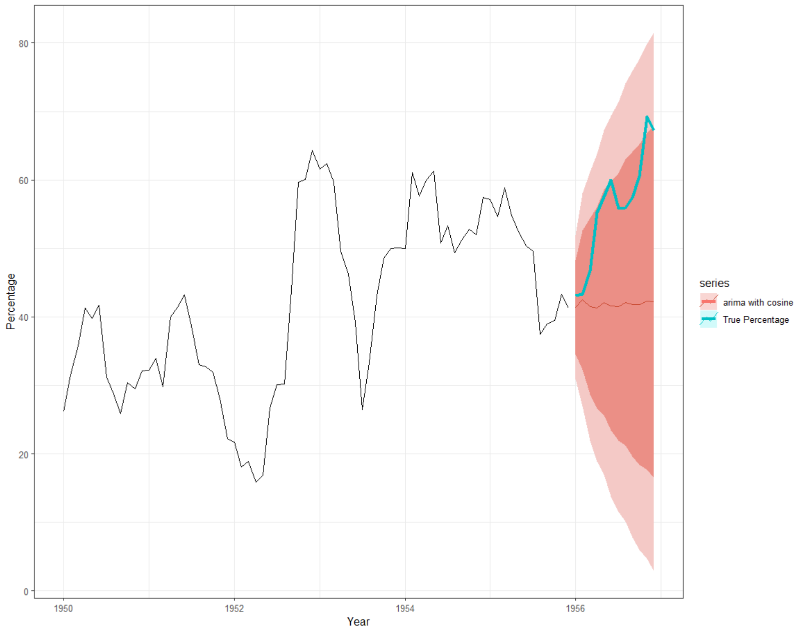
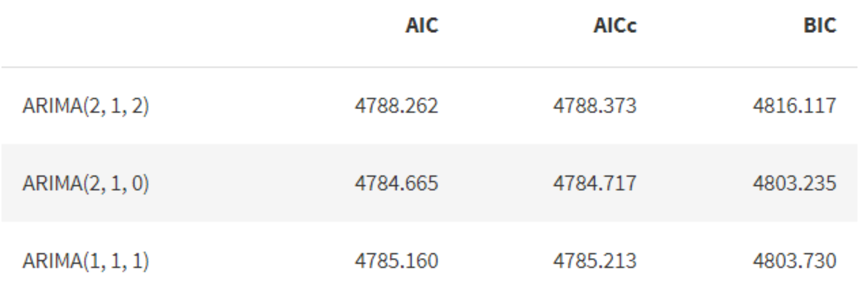


Figure 17: Drought percentage in 1895-1957 overlaid with prediction intervals of 1955-1956

Even though the actual values are still within the 95% prediction intervals, the range is too wide to produce a significant forecast.

b. Period 1957 - 2022

The smooth periodogram of the current period has very similar patterns to the previous periods. Thus we fit the same cosine-sine regression model to the training dataset. We only find sine at 0.05 significant. This means that we expect to see a repeated cycle after every 20 months. Following the same model fitting procedure as the training data 1895-1956, We fit 3 ARIMA models with sine(0.05) regression components to the training data and compare their fit using AIC, AICc, and BIC in the below table:



	AIC	AICc	BIC
ARIMA(2, 1, 2)	4788.262	4788.373	4816.117
ARIMA(2, 1, 0)	4784.665	4784.717	4803.235
ARIMA(1, 1, 1)	4785.160	4785.213	4803.730

Table 6: AIC, AICc, BIC of ARIMA models with sine component

ARIMA(2, 1, 0) achieves the best fit to training data, followed by ARIMA(1,1,1). We proceed with these two models and the sine pair to fit testing data of the last 2 years of the period 2020-2022. ARIMA(2, 1, 0) with sin(0.05) has RMSE of 16.593924 while ARIMA(1,1,1) with sine has RMSE of 16.718574 . Thus, ARIMA(2, 1, 0) with sine(0.05) is the most suitable model for the training set. However, the forecasted points from ARIMA(2, 1, 0) are not sensitive to the high volatility of drought percentage.

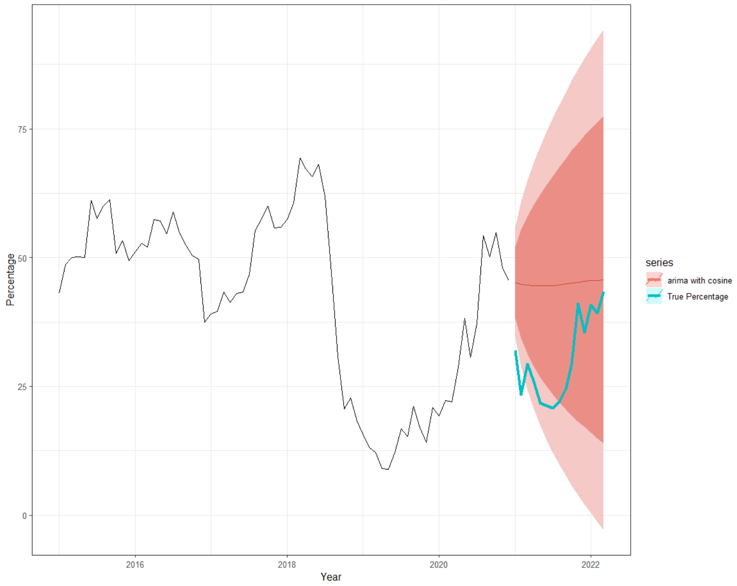


Figure 18: Drought percentage in 1957-2022 overlaid with prediction intervals of 2020-2022

ARIMA models with sine/cosine pairs are not able to provide robust predictions despite having a good fit for the training data.

V. Model Discussion

With all modeling done, what remains is a final comparison. The best of the three integral modeling techniques (Holt-Winters, Cosine-Sine, SARIMA), will be compared against each other to select the best model to represent each time interval. A final model for the distant and recent past drought activity.

The first model we will finalize is the model of drought percentages for the distant past, the period from 1895 to 1956. As done in prior modeling techniques, both a visual and mathematical inspection will yield a final conclusion.

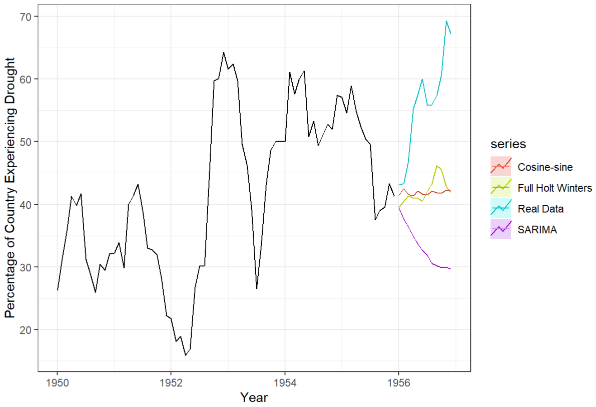
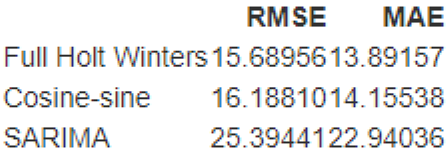


Figure 18: Comparison between final modeling techniques for 1895-1956

Visually, it appears that while no model is a perfect representation of the true data, the Cosine-sine model and the Full Holt-Winters model appear to perform similarly.



	RMSE	MAE
Full Holt Winters	15.68956	13.89157
Cosine-sine	16.18810	14.15538
SARIMA	25.39441	22.94036

Table 7: RMSE, MAE of models for 1895-1956 data

After a mathematical inspection, the Full Holt-Winters model possesses the joint lowest RMSE and MAE values, indicating that this is the best fit for our data. With that in mind, the Holt-Winters model will serve as our mathematical representation of the distant past, with it serving as the baseline to compare the recent past's drought activity.

The equation for the full Holt-Winters model is

$$y_{t+h} = (40.201 + 0.051h)s_{t-p+1+(h-1) \bmod p}$$

Seasonality Coefficient	Value
s1	-0.738
s2	0.183
s3	0.955
s4	0.581
s5	0.513
s6	-0.009
s7	1.520

s8	2.625
s9	5.517
s10	4.802
s11	1.997
s12	1.099

Table 8: Full Holt-Winters' Seasonality coefficients

Next, a model will be selected for the period of drought activity from 1956 to 2022. The arduous process of modeling has led us to a choice between 3 final models, the Cosine-sine model, the SARIMA(2,0,1)(2,0,0) model, and the Holt-Winters with only Seasonality.

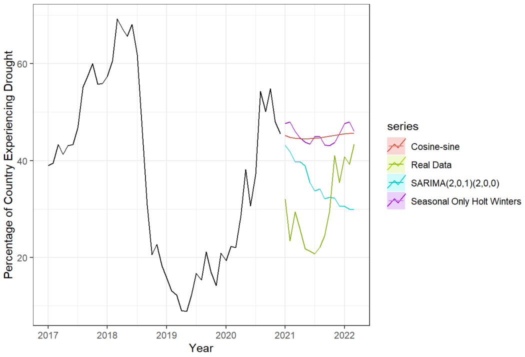


Figure 19: Comparison between final modeling techniques for 1956-2022

A visual inspection of this choice would appear to indicate that while no one model is a perfect fit, the SARIMA model appears to be the closest to reality. A mathematical inspection may confirm this suspicion.

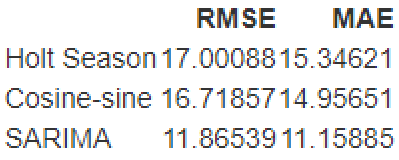

```
                 RMSE      MAE
Holt Season 17.0008815.34621
Cosine-sine 16.7185714.95651
SARIMA      11.8653911.15885
```

Table 9: RMSE, MAE of models for 1957-2022 data

Comparing RMSE and MAE, the decision is clear, the SARIMA(2,0,1)(2,0,0) is the best representation by far of the data. It will serve as the model for our recent past and will be compared against the baseline, the distant past, to see whether there appear to be significant differences between the two models. These differences will indicate whether or not change has occurred in drought conditions between these two periods. Final Model equation of SARIMA (2, 0, 1) (2, 0, 0)

$$Y_t = -0.08Y_{t-1} - 0.717Y_{t-2} - 0.02Y_{t-12} - 0.019Y_{t-13} + 0.026Y_{t-14} - 0.008Y_{t-24} - 0.008Y_{t-25} + 0.011Y_{t-26} + \varepsilon_t - 0.717\varepsilon_{t-1}$$

VI. Conclusion

Various methods of predicting future patterns & occurrences of abnormal drought are presented in this study. Using Holt-Winters, SARIMA, and ARIMA with cosine-sine patterns, we have explored the predictability of drought in different periods. From the information from the real data, we can conclude that the predictability of drought for recent periods is much more difficult than that in the past, as displayed by the oscillating patterns from 2021 to 2022. These characteristics made fitting and choosing the right model from 1957 to 2022 more difficult. In particular, for the model from 1895 to 1956, the full Holt-Winters model (the best model for the period) is still able to capture the increasing trend of drought. In contrast, for the model from 1957 to 2022, the SARIMA(2,0,1)(2,0,0) model, similar to other models, can predict the trend for a small section of the period. However, the model is still incapable of predicting the sudden upward trend in abnormal trends, similar to other models. This unpredictability in the data or abnormal drought indicates that natural hazard is getting harder to predict, which manifests in the various methods we have experimented with. The inconsistency in natural hazards in recent times can be attributed to various reasons, notably climate change. The modeling technique presented in this study will further facilitate the research in the areas of natural hazard predictions, specifically in abnormal drought frequency. However, further studies are required to better understand and capture the inconsistency in the data through improving our models or exploring more options for time-series modeling.

References

Chatfield, C. (1978). The Holt-Winters Forecasting Procedure. Journal of the Royal Statistical Society. Series C (Applied Statistics), 27(3), 264–279. https://doi.org/10.2307/2347162

Cryer, J. D., & Chan, K.-sik. (2011). Time series analysis with applications in R. Springer.

Historical Data and Conditions - National Integrated Drought Information System. https://www.drought.gov/historical-information?dataset=1&selectedDateUSDM=20110301&selectedDateSpi=20070601&selectedDatePaleo=2017