HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 3 (2021)

Predicting Drought Using Bayesian Structural Time Series Model

Connor Lee[1], Sophia Wang[2], and XL Pang[#]

[1] Saratoga High School, Saratoga, CA, USA
[2] Monta Vista High School, Cupertino, CA, USA
[#] Advisor

ABSTRACT

The western U.S. has been experiencing a mega-scale drought since 2000. By killing trees and drying out forests, the drought triggers widespread wildfire activities. In the 2020 California fire season alone, more than 10.3 million acres of land were burned and over 10000 structures were damaged. The estimated cost is over $12 billion. Drought also devastates agriculture and drains the social and emotional well-being of impacted communities. This work aims at predicting the occurrence and severity of drought, and thus helping mitigate drought related adversaries. A machine learning based framework was developed, including time series data collection, model training, forecast and visualization. The data source is from the National Drought Monitor center with FIPS (Federal Information Processing Standards) geographic identification codes. For model training and forecasting, a Bayesian structural time series (BSTS) based statistical model was employed for a time-series forecasting of drought spatially and temporally. In the model, a time-series component captures the general trend and seasonal patterns in the data; a regression component captures the impact of the drought in measurements such as severity of drought, temperature, etc. The statistical measure, Mean Absolute Percentage Error, was used as the model accuracy metric. The last 10 years of drought data up to 2020-09-01 was used for model training and validation. Back-testing was implemented to validate the model . Afterwards, the drought forecast was generated for the upcoming 3 weeks of the United States based on the unit of county level. 2-D heat maps were also integrated for visual reference.

Introduction

Since 2000, the western United States has entered the beginning of a megadrought, the second worst in 1200 years. The U.S. Drought Monitor places 60% of the western states under severe, extreme or exceptional drought conditions [1]. The impacts of drought come in a variety of forms. Lack of water supplies devastates the agricultural industry. Drought creates flammable fuel from dry vegetation, which feeds on rampant wildfires. Prolonged drought also burdens social and emotional wellbeing of the impacted communities, as quoted “if the land is sick, we are sick”.
Drought is by nature complex and stochastic. It is very challenging to determine when a drought will start or end. Drought forecasting can help to establish drought mitigation strategies in advance. However, drought prediction is often marked by uncertainty. Therefore statistical modeling techniques with uncertainty estimation are necessary for a reliable forecast.

Bayesian inference is a well-developed statistical framework that allows practitioners to both intuitively incorporate prior beliefs about certain data into the modeling process, and obtain comprehensive uncertainty estimates about predictions [2]. Bayesian approaches provide such uncertainty quantification by directly producing posterior predictive distributions, rather than the point forecasts generated by traditional Frequentist approaches [2]. Bayesian methods can also be used to generate point forecasts by simply taking the mean of the posterior predictive distribution

output. Because these tasks are naturally incorporated into Bayesian analysis and are thus more intuitive to perform. Accordingly, Bayesian methods are often preferred for use in contexts requiring uncertainty estimation [3-5].

To generate the drought-volume forecasts with uncertainty estimate, we used the Bayesian structural time series (BSTS) model to perform feature selection, time series forecasting, and causal impact inference. Historical climate records such as temperature and precipitation data were applied to the Bayesian statistical model to generate probabilistic drought prediction.

Bayesian structural time series model (BSTS) is a structural or state space model. Such models are defined by two equations. The first, called the observation equation:

Equation 1:

$$Y_t = Z^T{}_t \alpha_t + \epsilon_t$$

ϵ describes the relationship between our observed target, Y_t, and a vector of latent variables α_t, called the *latent state* of the system. The second equation, called the transition equation, models the evolution of this latent state over time:

Equation 2:

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t$$

The vectors Z_t, T_t, R_t are structural parameters that are constructed according to the evolution dynamics of the modeled system. These parameters render structural models extremely flexible. Many classical time-series models such as auto-regressive models can be expressed equivalently as structural models [6].

Specific models are constructed with different trend components and seasonality structures. An example of such a model is described below, a local-level trend μ_t, a seasonal pattern τ_t, and external regressors x_t:

Equation 3 (a):

$$Y_t = \mu_t + \tau_t + \beta^T x_t + \epsilon_t, \epsilon_t \sim N(0, \sigma^2 \varepsilon)$$

Equation 3 (b):

$$\mu_{t+1} = \mu_t + \eta_t, \eta_t \sim N(0, \sigma^2{}_\eta)$$

Equation 3 (c):

$$\tau_{j,t+1} = \tau_{j,t} \times \cos(\lambda_j) - \tau^*{}_{j,t} \times \sin(\lambda_j) + \omega_{j,t}$$

Equation 3 (d):

$$\tau^*{}_{j,t+1} = \tau^*{}_{j,t} \times \cos(\lambda_j) - \tau_{j,t} \times \sin(\lambda_j) + \omega^*{}_{j,t}$$

Equation 3 (e):

$$\tau_t = \sum_{j=1}^{k} \quad \tau_{j,t}$$

Equation 3 (f):

$$\lambda_j = 2\pi j/s$$

where $j = 1, .., k$ is the j-th seasonal frequency, s is the length of the longest seasonal cycle in number of time-steps, and Y_t is our observed target. The local-level trend μ_t models the evolution of the latent state of the system, as described in equation (3b), is assumed to evolve following a random walk in levels. This choice reflects a belief in no strong upward or downward trend at a short-term regular level.

In the Bayesian structural time series (BSTS) model, we impose *Spike-and-Slab* prior on the regression coefficients, which enables automatic feature selection *via* parameter shrinkage [7]. For instance, we impose an inclusion probability of 1 on the temperature feature and 0.5 on all others. Additionally we set the elements of the prior mean vector to some value like ± 0.5, with the sign of each element determined by the assumed directionality of the relationship between the corresponding feature and drought volume. Our model also incorporates weekly, monthly, and quarterly seasonality. As are typical with Bayesian statistical models, our forecasting model is fit using a Markov

Chain Monte Carlo (MCMC) method [8].

The statistical measure, Mean Absolute Percentage Error, or *MAPE* (9), is used as the model accuracy metric. It measures the accuracy as a percentage and can be calculated as the average absolute percent error for each time period:

Equation 4:

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T} \quad \left|\frac{ActualDrought - PredictedDrought}{PredictedDrought}\right|$$

where $T = \{t_1, t_2, .., t_n\}$ is the time step t. It is computed across the total time periods of prediction.

Equation 5:

$$\text{Accuracy} = 1 - \text{MAPE}$$

Equation 6:

$$\text{Model Error} = (\text{min\_MAPE}, \text{max\_MAPE})$$

Methods

The system framework consists of the following components, as illustrated in Figure 1.

- Data preparation which includes data loader and data pre-processor. Data loader loads the past data till the latest data into the system
- Data pre-processor performs standard data cleaning procedure
- Data visualization including heat-map visualizes the actual and predicted drought at national and state levels
- Drought forecasting predicts severity of drought at national and state levels for upcoming 3 week periods

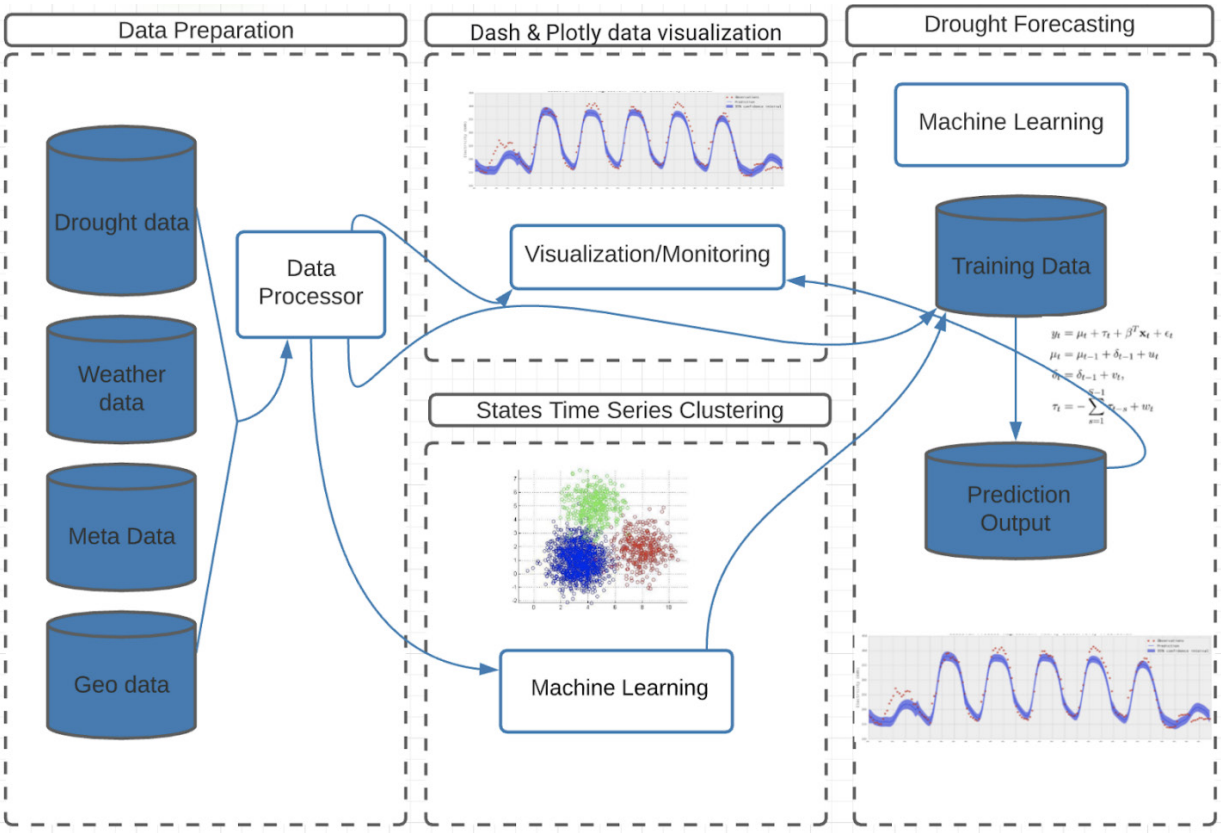


Figure 1: System Framework

Data source

The historical climate data source is the United States Drought Time-Series Data (USDM). It describes the extent and severity of the drought at national and state levels. The data is updated weekly. The severity level of the drought is defined from D0 to D4, with D0 representing the percentage of the county that is abnormally dry, D1 moderate drought, D2 severe drought, D3 extreme drought and D4 exceptional drought. The FIPS (Federal Information Processing Standard) code is used to uniquely identify each county. The data source is obtained from https://droughtmonitor.unl.edu/ (2015-2020).

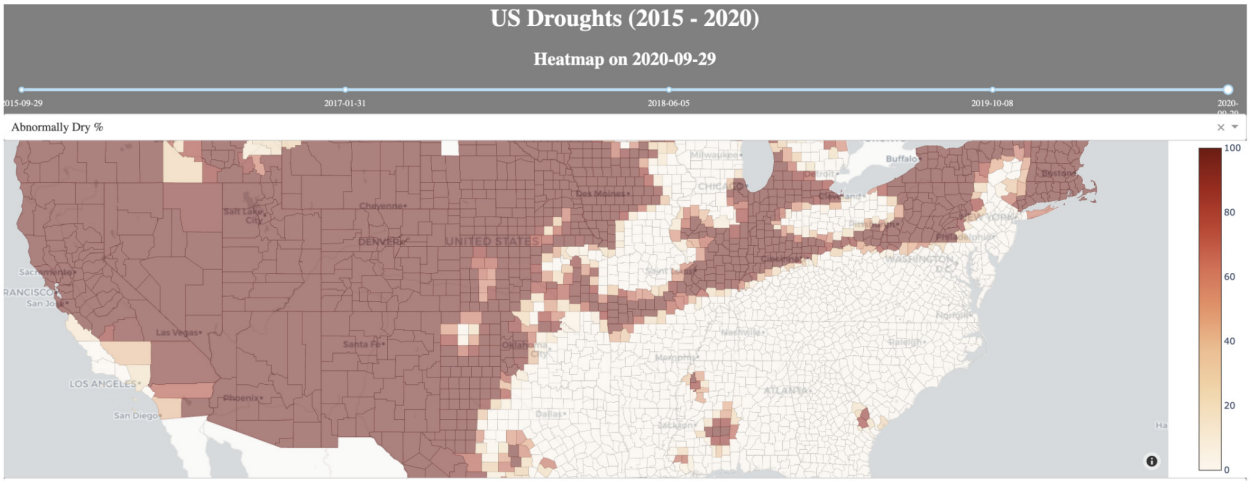
User Interface components

The open source Dash framework (https://plotly.com/dash/open-source/) is used for data analysis. The underlying code for this analysis is written in Python. A Plotly Python graphing library is used to visualize the data and generate 2D drought heat-maps. The application is deployed on the Heroku Cloud Application Platform.

Plotly Graphing Components

Three main graphing libraries are used for analyzing the drought data collected through the US Drought Monitor (USDM, https://droughtmonitor.unl.edu/). A basic table and trend-line is used to display core statistics and project future trends generated by the predictive algorithms. A choropleth map is selected to represent the spatial variation of

drought severity using colored polygons. This map is configured with standard Federal Information Processing Series (FIPS) codes for geometric information and the USDM drought data identified by FIPS code as the input data. Different input options such as drop down menu and slider bar are provided as controls to promote an ease of data selection. One example of a generated 2D heatmap and corresponding table are shown in Figure 2.



Release Date	County	State	Drought %	Abnormally Dry %	Moderate Drought %	Severe Drought %	Extreme Drought %	Exceptional Drought %	Valid Start	Valid End
20191112	Abbeville County	SC	0	100	100	4.61	0	0	2019-11-12	2019-11-18
20191105	Abbeville County	SC	0	100	100	33.46	0	0	2019-11-05	2019-11-11
20191029	Abbeville County	SC	0	100	100	95.96	0	0	2019-10-29	2019-11-04
20191022	Abbeville County	SC	0	100	100	95.96	1.15	0	2019-10-22	2019-10-28
20191015	Abbeville County	SC	0	100	100	95.96	1.15	0	2019-10-15	2019-10-21
20191008	Abbeville County	SC	0	100	100	63.44	0	0	2019-10-08	2019-10-14
20191001	Abbeville County	SC	0	100	65.57	0	0	0	2019-10-01	2019-10-07
20190924	Abbeville County	SC	0	100	65.57	0	0	0	2019-09-24	2019-09-30
20190917	Abbeville County	SC	31.33	68.67	27.8	0	0	0	2019-09-17	2019-09-23
20190910	Abbeville County	SC	32.22	67.78	0	0	0	0	2019-09-10	2019-09-16

Figure 2: Example of 2D heatmap to visualize drought distribution

Results

The model was trained using weekly historical time-series data collected from the last 10 years' available data up to 2020-09-01. To evaluate whether and to what degree the model gives an accurate projection of drought volume against actual volume, back tests were conducted using available data from 2020-06-01 to 2020-09-01 to answer these questions: Is the forecast different from the actual performance data? If so, what is the error percentage? As shown in Table 1 is the accuracy of our suite of models applied to national and different states.

Table 1: Model Evaluation using national and three states as examples

States	Accuracy	MAPE	Model Error
National	93.2%	6.8%	4-7%
California	91.9%	8.1%	5-9%
Colorado	90.3%	9.7%	3-8%
Oklahoma	89.2%	10.8%	3-7%

Representative graphs are generated to describe the backtesting of our model suite applied to the national and states. Figure 3 (a) shows one example at national level, where the drought index is plotted from 2016 to 2020 time period, where the data of 2016 up to September 2020 is from historical climate data (United States Drought Time-Series Data, USDM), and the data of October 2020 is generated from this statistical model. Backtesting method was applied to verify the trained model by comparing actual data with predicted data from June to Sept of 2020. The labels with *original*, *pre_lowerbound*, *pre_upperbound* and *pre* are the ground truth value, lower bound, upper bound and predicted values respectively. Figure 3(b) are magnified graphs depicting the comparison among actual data (i.e. original) and predicted data (i.e., pre_, pre_upperbound, and pre_lowerbound) from June, July, August and September 2020 respectively. The calculated Mean Absolute Percentage Error (MAPE) at national level is 4-7%. Figure 3(c) is the projected drought of upcoming three weeks at national level using the validated model including first, second and third weeks of October 2020.

Figure 3 (a)

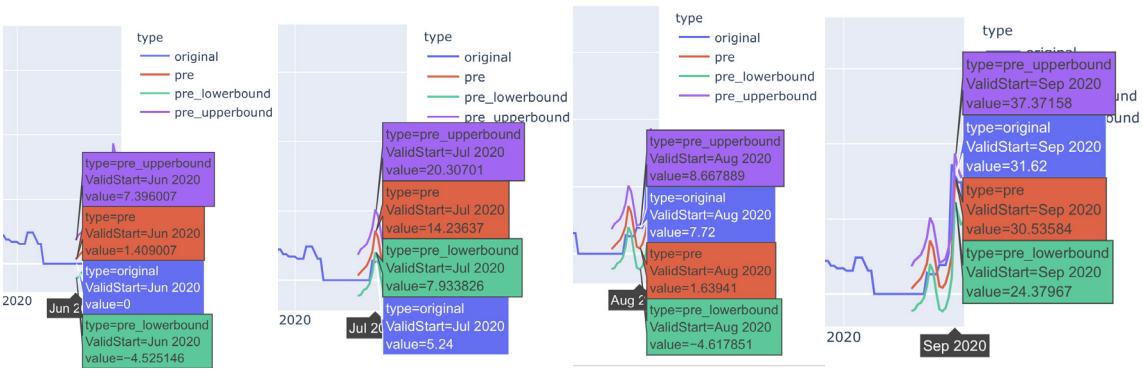


Figure 3 (b)

Figure 3 (c)

Figure 3. national level drought data training and backtesting (a), high magnification graphs describing training and backtesting data of June, July, August and September, 2020 (b) and next three weeks' forecast from first, second and third weeks of October 2020 (c)

Figure 4. drought data training and backtesting (a), and next three weeks' forecast (b) of Colorado State, the forecast data ranges from first, second to third weeks of October 2020.

Figure 4. describes one example of the data training and backtesting (a) and forecast (b) for Colorado state respectively. It is worth noting that the upcoming three weeks trends toward more dryness in the category of severe dry and maybe worth considering drought alertness and preventative actions for the state of Colorado.

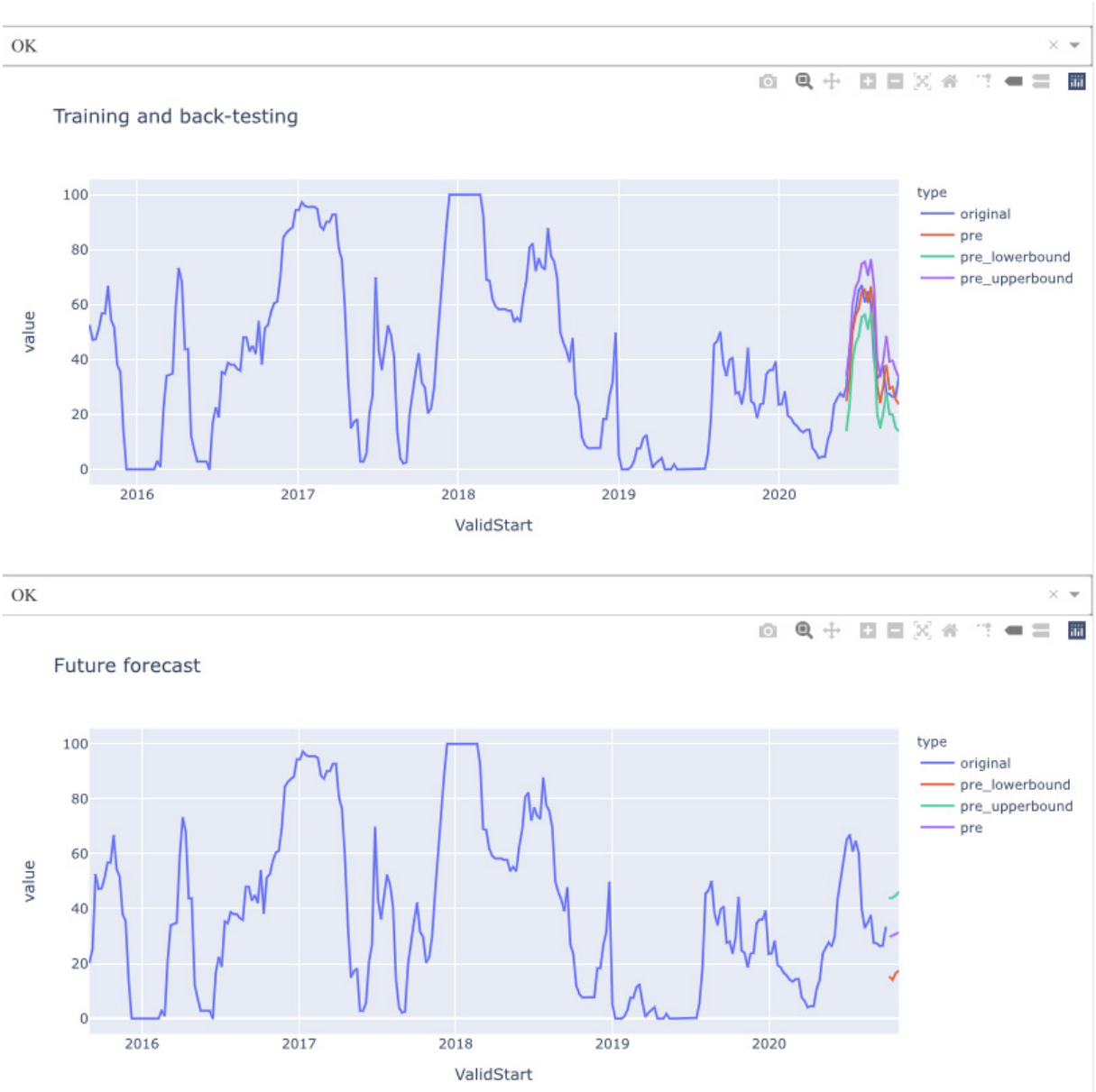
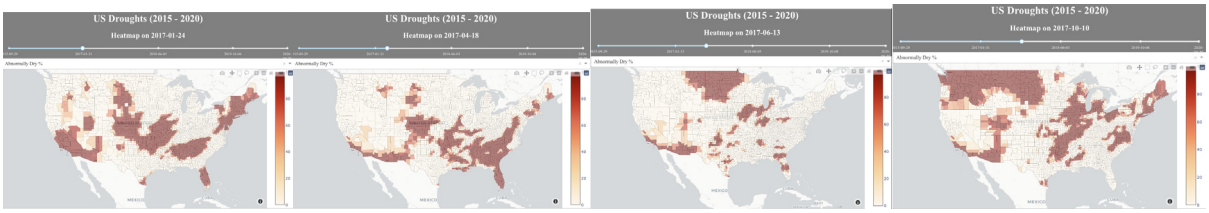


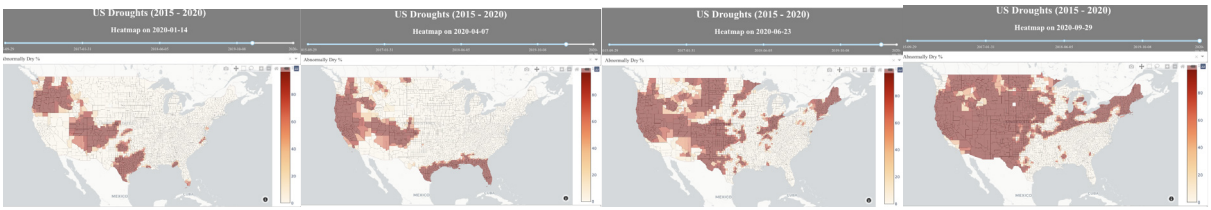
Figure 5. drought data training and backtesting (a), and next three weeks' forecast (b) of Oklahoma State, the forecast data ranges from first, second to third weeks of October 2020.

Figure 5. shows another example of training and forecasting data for Oklahoma state. For both these two examples, model training and forecasting data is for the time period of first, second and third week of October 2020, and the rest of the data is from historical data base. The drought index shown in y axis indicates a moderate drought condition and is considered as low risk for the state of Oklahoma for the first, second and third weeks of October 2020.

2-D heat maps were generated to visualize the drought contour in the United States at different time periods using a Dash framework. Units are by county area. Figure 6 shows examples of snapshots of drought heat-maps at the national levels in 2017 and 2020 respectively, where the model generated data refers to October 2020, and the rest of the data is from a historical database (United States Drought Time-Series Data, USDM).



Drought heatmap, Jan, April, June and October, 2017



Drought heatmap, Jan, April, June and October, 2020

Figure 6. 2D heatmap to visualize drought at 2017 (a) and 2020 (b), where the heatmap of October 2020 is from model prediction whereas other plots from historical data.

Though a few examples are demonstrated in this main context, this work was carried out for the entire United State with the modeling resolution at county level. The complete dataset including model training and forecasting was deployed at https://cal-droughts.herokuapp.com/.

Discussion

A Bayesian structural time series (BSTS) based statistical model was used to analyze historical drought data and then predict upcoming drought situations. Mean Absolute Percentage Error (e.g. MAPE), was chosen to measure the modeling prediction accuracy. As shown in Table 1, 6-10% MAPE is calculated from back-testing and validation with model error rate of 3-9%.

The prediction of the upcoming three weeks' drought was plotted for national average as well as individual states using the calibrated BSTS model. Figure 3-5 shows a few examples of drought at national and state levels such as Colorado and Oklahoma. Similar graphs can be plotted for the rest of states as well. Take one example of drought prediction data from Colorado state, represented by Figure 4(b) we can expect that the drought trends toward more severe as quantified by y axis drought index with a range indicating upper and lower bounds of this prediction.

2-D heat maps (Figure 6) clearly shows that the degree of drought in each season grows progressively severer from 2017 to 2020. The extent of drought area and severity of drought concentrates on the western United States and expands further into the midwest in the calendar year of 2020, indicating the overall trend of more alarming drought conditions. The forecasting graph (Figure 6b, October) shows more drought across the United States map, with higher concentration in the Western State, by comparing actual historical data in September 2020. The data visualization through Heroku platform enables users to access the current and future drought situation of their specific interest, and thus support local community based drought mitigation strategies.

Currently, this BSTS model is developed for short-term forecasting of three weeks. In the future, longer term forecasting is possible if we develop the framework to incorporate the effect of seasonality, trend and more regressors from various time series data.

In the future, this research framework will be further developed to predict wildfire occurrence which is a more complex environmental issue. The interaction of heat, drought, and probability of fire ignition are the major factors for the model input, followed by the model calibration and execution. Once proved effective, this research could be connected with a US satellite backed database to provide an on-time alerting system for drought and fire risk in the long term.

Conclusion

Drought devastates the western United States socially and economically with immeasurable long term effects. This work focuses on predicting when and where drought will happen. The potentially impacted communities can use this data to proactively mitigate drought induced disasters. A Bayesian structural time series model (BSTS) was employed to evaluate the severity of drought at national and state levels. The historical data was obtained from the National Drought Monitor center (NDMC). First, a "Training and back-testing" method was applied to validate the model using available data. The model accuracy was estimated to be 91-97% based on the training data. Then, an upcoming 3 weeks of drought forecast data was generated using the validated model at state and national levels. 2-D heat maps were also plotted out to visualize the severity, distribution and evolution of droughts.

Acknowledgements

The authors would like to thank Dr. XL Pang, for guiding us through the procedure of literature research, brainstorming and paper writing.

References

1. U. S. Drought Monitor, West, National Drought Mitigation Center, University of Nebraska-Lincoln, www.droughtmonitor.unl.edu/CurrentMap/StateDroughtMonitor.aspx?West
2. Seeing Theory, Chapter 4, Frequentist Inference, Chapter 5, Bayesian Inference, Brown University, www.Seeing-theory.brown.edu
3. A. K. Mishra and V. R. Desai, DroDrought-forecasting-using-stochastic-models.pdf, Stoch Environ Res Risk Assess, 2005, 19: 326-339
4. JiYae Shin, Muham mad Ajmal, and Jiyoung Yoo, A Bayesian Network-Based Probabilistic Framework for Drought Forecasting and Outlook and Tae-Woong Kim, Advances in Meteorology, 2016
5. Steven L. Scott and Hal Varian, Predicting the Present with Bayesian Structural Time Series, International Journal of Mathematical Modelling and Numerical Optimisation, Vol. 5 , 4-23, 2014.
6. Rob J Hyndman and George Athanasopoulos, Forecasting: Principles and Practice, Chapter 8, ARIMA models.
7. By Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy and Steven L. Scott, Inferring Causal Impact Using Bayesian Structural Time-series Models, Annals of Applied Statistics, Vol 9, No 1, 247-274, 2015.
8. Stephen Brooks, Markov chain Monte Carlo method and its application, Journal of the Royal Statistical Society: Series D (The Statistician), Vol 47, No 1, 69-100, 2002.
9. Swamidass P.M. (eds), MAPE (mean absolute percentage error), 462, Encyclopedia of Production and Manufacturing Management. Springer, Boston, MA.