Drought Prediction in the Western U.S. Using a Long Short-Term Memory (LSTM) Model

Bennett Bolen
Department of Mechanical Engineering
bbolen@stanford.edu

Mo Sodwatana
Department of Energy Resources Engineering
jarupas@stanford.edu

Hannah Hampson
Department of Civil and Environmental Engineering
hhampson@stanford.edu

Abstract

Droughts are a high-priority climate threat that can result in vast economic damage while threatening food and water systems. With the drought vulnerable hydrologic patterns of the Western United States, it is especially critical to understand their behavior in this region. Droughts are multifaceted and difficult to predict, in part due to a large number of contributing factors [1]. While prior work using machine learning models to predict droughts has been done in other countries, we propose using a novel many-to-one Long Short-Term Memory (LSTM) model approach across the Western U.S. Our model predicts the severity of droughts given sequential measurements of precipitation, temperature, and soil moisture, outputting the drought index forecast for the following season. The linear model gave a test set error of 0.137 while the LSTM model gave a test set error of 0.0699. Further tuning of hyperparameters and an increase in training set size could help to reduce this error.

1 Introduction

Droughts are amongst one of the most complex geological hazards, given the many intricate contributing factors, such as precipitation and soil moisture, and the various characteristics of droughts, from intensity to spatial extent. Drought forecasting with long lead time is critical for detecting early warning systems and risk management strategies, especially in the Western U.S. where it is extremely susceptible to droughts. There are several types of models used in forecasting, most notable being data-driven, physical and hybrid. There are many advantages and disadvantages to each approach, but the popularity of artificial neural networks in the past decade has given rise to effective data-driven models [1]. In particular, LSTM models, with its ability to retain information for longer periods, can be considered very effective in drought prediction.

The purpose of this work is to develop and validate the utility of a many-to-one LSTM architecture for seasonal drought forecasting in the Western U.S. While droughts are dependent on numerous factors, our goal is to explore the possibility of drought forecasting with few variable inputs. Our approach is to utilize the sequential measurements of precipitation, temperature and soil moisture during the rainy season to predict the drought index in the following dry season.

CS230: Deep Learning, Winter 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related Work

Prior work using machine learning models to predict droughts has been done with models such as decision trees, random forest, conventional artificial neural networks, coupled-wavelet artificial neural networks, support vector regression, among others [2, 3, 4, 5]. In one study, a stacked LSTM model was used, coupled with lagged climate variables, to perform long lead time drought forecasting [6]. These studies have all taken place outside of the U.S. Much of the existing literature focuses on comparing the effectiveness of different models and the accuracy of different lead-up times [7, 8], while our focus is on using inputs from the most precipitation heavy half of the year to make a forecast in the drought-prone half of the year, taking into account the key hydrologic seasonality at play in this region of the United States.

3 Dataset and Features

Our dataset includes the drought index, precipitation, soil moisture, and minimum and maximum temperature.

Drought Index: The drought index dataset [9] provides global, gridded Standardized Precipitation-Evapotranspiration Index (SPEI) values. This dataset gives 6-month temporally averaged values from years 1900-2018 over a 0.5° resolution. The value represents the degree of drought that an area is experiencing, ranging from -2.33 to +2.33, where a negative value denotes drier than average conditions, and a positive value indicates wetter than average conditions.

Precipitation: The precipitation dataset is generated by Climate Prediction Center (CPC) Unified Gauge-Based Analysis of Daily Precipitation over CONUS provided by the National Oceanic and Atmospheric Administration (NOAA) [10]. This dataset is an interpolation of precipitation gauges over the continental US, providing daily precipitation values at 0.25° resolution from 1948 to present.

Soil Moisture: The soil moisture dataset is generated by NASA's Soil Moisture Active Passive (SMAP) satellite [11]. The SMAP observatory provides a measurement of the moisture in the top 5 cm of soil everywhere in the world every 2-3 hours at a 9km resolution. The data is downloadable as a HDF5 file and includes the latitude and longitude of each measurement. SMAP has been in operation since January 2015.

Temperature: The maximum and minimum temperature dataset is extracted from CPC Global Daily Surface Air Temperature provided by NOAA [12]. The data is a global GTS data and is gridded using Shepard Algorithm. The temporal coverage is daily from 1979 to present and the spatial coverage at 0.5° resolution.

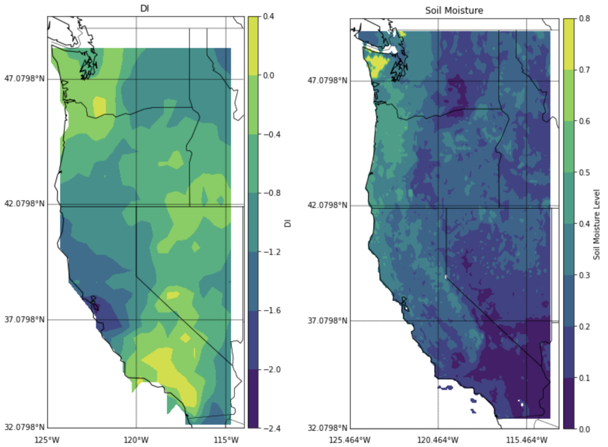

Figure 1: Contour map of the gridded drought index and soil moisture over our region on a single day.

3.1 Data Preprocessing

In order to match the spatial resolution of our dataset, we downsampled all finer resolution datasets to 0.5° resolution. Figure 1 illustrates an example of this difference in spatial resolution between our drought index and soil moisture data. This downsampling was done by taking an average over all inputs except for precipitation, where instead a sum was taken to reflect cumulative precipitation. Our data inputs were downsampled temporally as well to output monthly values. Similar to the spatial downsampling, over all inputs except for precipitation a temporal average was taken, with the precipitation matrix again taking a cumulative value instead.

Across our five datasets all but the precipitation data was normally distributed, as pictured in Figure 2. To account for this, we used a lognormal transformation on the precipitation values, outputting a normal distribution to match those of the other datasets (Figure 3).

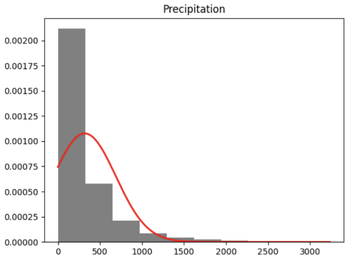

Figure 2: Precipitation distribution pre-lognormal transformation. The normally fitted probability density function (red line) poorly represents the true distribution captured by the histogram.

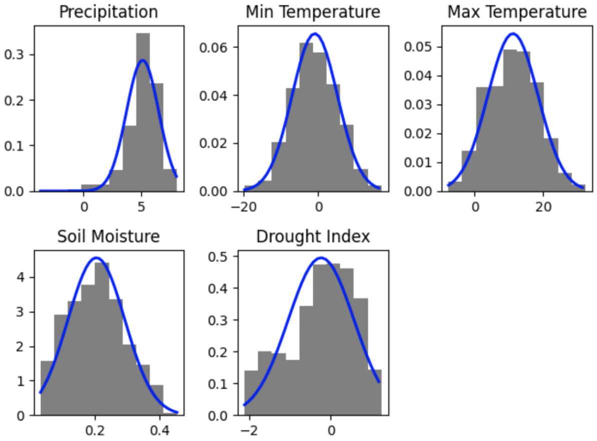

Figure 3: Histograms and fitted normal probability density functions of the data following a lognormal transformation on precipitation.

Prior to outputting our X and Y matrices, any training example with an invalid input or prediction value (marked as NaN) was dropped from the sample. Following this step, our X and Y matrices were output for use in the models.

4 Methods

The preprocessed dataset is then split into an input array and an output array. The input array contains four features: precipitation, soil moisture, and minimum and maximum temperatures. These features are sequential, monthly averages of the "rainy season" from the beginning of November to the end of April. In the case of precipitation, the feature represents monthly summations. The output array is the drought index of the following dry season at each corresponding latitude and longitude.

The total number of samples available after data processing is 888. The samples are shuffled and split into 75% training set and 25% testing set.

4.1 Linear Regression

To provide a linear baseline upon which the LSTM could improve, we developed an ordinary least square linear regression model in scikit-learn. For the linear regression model, the input dimension is (training set, sequence length * input size = 24). As previously mentioned, the training set has 666 samples and the test set has 222.

4.2 LSTM Model

To take advantage of the temporally sequential nature of the weather and drought forecasting, we use Long Short Term Memory (LSTM). The training framework is developed in PyTorch. The shape of the input to feed into the model is (batch size, sequence length = 6, input size = 4). The model consists of a many-to-one LSTM architecture with input size of 4 and hidden layer of 32. The next layer is a linear layer with 32 input features and 1 output feature with bias. The formulas of LSTM for each layer and each element in the input sequence are shown in (Eqn. 1) [13]:

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
\tag{1}
$$

Here h_t is the hidden state, c_t is the cell state and x_t is the input at time *t*. h_{t-1} is the hidden state at time *t-1* or initial state at time 0. i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively. Figure 4 illustrates the structure of an LSTM unit and the many-to-one architecture.

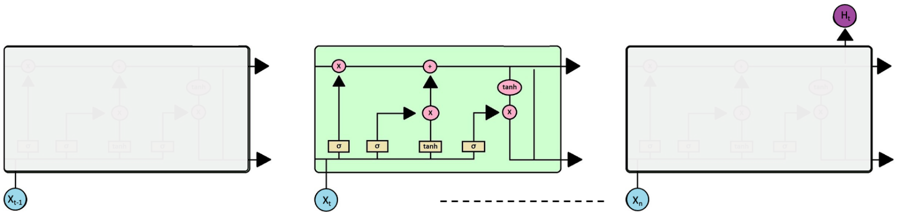

Figure 4: Structure of a many-to-one LSTM unit.

The optimizer function used is the Adam optimizer. The number of hidden layers, learning rate, batch size and epochs are tuned hyperparameters and will be discussed in the next section. In both the linear regression and the LSTM model, error was evaluated using the Mean Square Error (MSE) (Eqn. 2) [14]:

$$
\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2}
$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

5 Results

5.1 Linear Regression

After fitting the linear regression model to the 666 example training set, predictions were generated for the 222 example test set. The MSE calculation yielded an error of 0.137, which is a large amount of error for this application. It is therefore likely the linear regression model is unable to accurately model all of the features of the drought prediction.

5.2 LSTM Model

The purpose of the LSTM model was to achieve higher drought performance than with the linear regression model. Since the size of the dataset was relatively small, a batch size of 1 was used in the LSTM model. This resulted in slower training, but higher performance on the training set. The network used 32 hidden layers and a learning rate of 0.001. Training with these hyperparameters for 150 epochs yielded an MSE of 0.0149 on the training set and 0.0699 on the test set. The training error rate is shown in Figure 5.

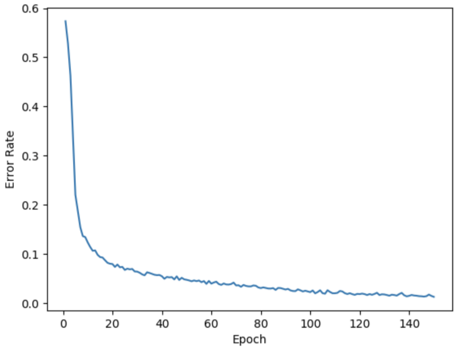

Figure 5: Error rate during training for the LSTM model.

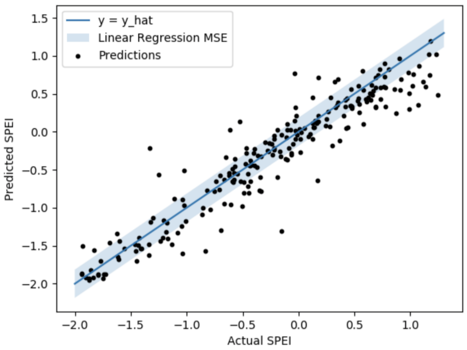

Figure 6: Predicted vs actual drought index for the LSTM Model.

In the LSTM model, we achieve relatively good performance on our training set and worse performance on our test set. This high variance indicates that we may have overfit our model. One solution that could improve test set performance is using a larger dataset for training. While we have used all available years of data for California, more data could be added by expanding our study to a larger geographical area.

The LSTM model had a higher accuracy than the linear regression model. Further tuning of hyperparameters would likely yield a further accuracy improvement. Figure 6 shows a comparison between the predicted and actual drought index for the LSTM model, overlaid on the mean error for the linear regression model.

6 Conclusion

The results of our model prove the promising nature of LSTM models in hydrologic forecasting. Our model provides a starting point for future work in drought forecasting for this region and applied elsewhere. Promising next steps include the tuning of hyperparameters, increasing training examples (through expanding the model over a greater area, or using higher resolution data), increasing the sequence length (such as inputting 12 months of data versus 6), and attempting to forecast droughts further into the future. Error analysis could also point to ways of improving our model, such as through investigating whether performance is greater achieved for some geographical regions over others.

7 Contributions

Hannah worked on downloading and processing drought and precipitation data, and later preprocessing steps with other data inputs to create X and Y matrices.

Mo gathered and processed the maximum and minimum temperature data, fine tuned the linear regression model, and worked on building the LSTM training model.

Bennett gathered and processed the soil moisture data, created the initial linear regression model, and developed the test set validation portion of the LSTM model.

Our Github repository can be found here:
https://github.com/hhampson/cs230 _where _da _droughts _at/tree/main

References

[1] Hao, Z., Singh, V. P., and Xia, Y. (2018). Seasonal drought prediction: Advances, challenges, and future prospects. *Reviews of Geophysics*, 56, 108– 141. https://doi.org/10.1002/2016RG000549

[2] Rhee, J. and Im, J. (2017). Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. *Agricultural and Forest Meteorology*. Volumes 237–238, Pages 105-122, ISSN 0168-1923, https://doi.org/10.1016/j.agrformet.2017.02.011.

[3] Dayal, K., Deo, R., and Apan, A. (2017). Drought Modelling Based on Artificial Intelligence and Neural Network Algorithms: A Case Study in Queensland, Australia. *Climate Change Adaptation In Pacific Countries*, 177-198. doi: 10.1007/978-3-319-50094-2_ 11.

[4] Deo, R., and Şahin, M. (2015). Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. *Atmospheric Research*, 153, 512-525. doi: 10.1016/j.atmosres.2014.10.016.

[5] Adamowski, J. and Belayneh, A. (2016). New Approaches in Drought Forecasting using Artificial Intelligence Methods. *Encyclopedia of Natural Hazards.*

[6] Dikshit, A., Pradhan, B. and Alamri, A. (2007). Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model. *Science of the Total Environment.* https://doi.org/10.1016/j.scitotenv.2020.142638.

[7] Belayneh, A. and Adamowski, J. (2013). Drought forecasting using new machine learning methods. *Journal of Water and Land Development.* DOI:10.2478/jwld-2013-0001.

[8] Jalalkamali, A., Moradi, M. and Moradi, N. (2015) Application of several artificial intelligence models and ARIMAX model for forecasting drought using Standardized Precipitation Index. *Int. J. Environ. Sci. Technol.* DOI 10.1007/s13762-014-0717-6.

[9] Vicente-Serrano, S., Beguería, S., Moreno, J., Angulo-Martinez, M. and Kenawy, A. (2010). A New Global 0.5° Gridded Dataset (1901–2006) of a Multiscalar Drought Index: Comparison with Current Drought Index Datasets Based on the Palmer Drought Severity Index. *Journal of Hydrometeorology.* 11. 1033-1043. 10.1175/2010JHM1224.1. Retrieved from: https://spei.csic.es/database.html.

[10] CPC US Unified Precipitation data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA. Retrieved from: https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html.

[11] Reichle, R., G. De Lannoy, R. D. Koster, W. T. Crow, J. S. Kimball, and Q. Liu. 2020. SMAP L4 Global 3-hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Analysis Update, Version 5. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: https://doi.org/10.5067/0D8JT6S27BS9. Accessed: 02-20-2021.

[12] CPC Global Daily Air Surface Temperature data provided by NOAA/OAR/ESRL PSL, Boulder, Colorado, USA. Retrieved from: https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html

[13] PyTorch Documentation. LSTM. Retrieved March 17, 2021 from
https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html.

[14] Probability Course. Mean Square Error. Retrieved March 17, 2021 from
https://www.probabilitycourse.com/chapter9/9_ 1_ 5_ mean_ squared_ error_ MSE.php.