A Machine Learning Pipeline to Predict Vegetation Health

Thomas Lees*
University of Oxford
thomas.lees@chch.ox.ac.uk

Gabriel Tseng*
Okra Solar
gabriel@okrasolar.com

Alex Hernandez-Garcia
University of Osnabrück
ahernandez@uos.de

Clement Atzberger
University of Natural Resources
and Life Sciences, Vienna
clement.atzberger@boku.ac.at

Simon Dadson
Steven Reece
University of Oxford
simon.dadson@ouce.ox.ac.uk
reece@robots.ox.ac.uk

Abstract

Agricultural droughts can exacerbate poverty and lead to famine. Timely distribution of disaster relief funds is essential to help minimise the impact of drought. Indices of vegetation health are indicative of higher risk of agricultural drought, but their prediction remains challenging, particularly in Africa. Here, we present an open-source machine learning pipeline for climate-related data. Specifically, we train and analyse a recurrent model to predict pixel-wise vegetation health in Kenya.

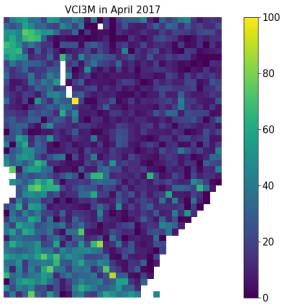
1 Introduction

Drought is estimated to be one of the world's most costly hazards, accounting for 22% of damage from natural disasters (Wilhite et al., 2007). Since 1980, East Africa has experienced a number of severe droughts. There is evidence to suggest that droughts are becoming longer and more frequent in the region due to climate change (Nicholson, 2017).

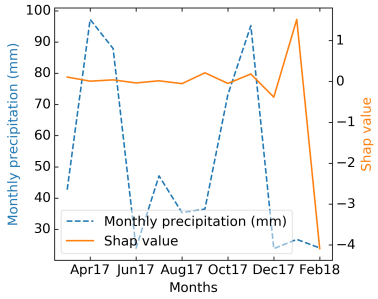
The timely prediction of drought can help reduce the risk of a hazard turning into a social or environmental disaster, by improving the response times of NGOs and governments (Hillier and Dempsey, 2012). In Kenya, the National Drought Management Authority (NDMA) has distributed emergency funds through the Drought Contingency Fund since 2014 (Klisch and Atzberger, 2016). Currently, the funds are distributed depending on near-real time indices drawn from satellite imagery. Despite the importance of predicting drought occurrence and impact, it remains challenging, particularly in East Africa, due to the complex interactions of large scale atmospheric circulation with local orography (Gebremeskel Haile et al., 2019). In this paper we demonstrate the use of a machine learning pipeline for improving the prediction of the vegetation health index used by the NDMA.

In particular, we train a regular Long Short Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) and an *Entity-Aware* LSTM (Kratzert et al., 2019) to predict seasonal drought conditions in Kenya, measured by the Vegetation Condition Index averaged over 3 months (VCI3M) (Kogan, 1997). The VCI3M is a satellite-derived measurement of anomalous vegetation health. Our models outperform a persistence baseline for predictions one month ahead. In addition, they are competitive with a model implemented by the NDMA (Adede et al., 2019) while producing much more spatially granular predictions. In this paper, we focus on the same four arid districts as Adede et al. (2019). As an additional contribution, we present an open-source pipeline for training machine learning algorithms with multiple sources of climate-related data.

*Equal Contributions



(a) Observed 3 Month Vegetation Condition Index (VCI3M) in April 2017, showing low vegetation health across most of the region



(b) The effect (Shap value) of spatially aggregated precipitation on the model predictions for a prediction made for March 2018.

Figure 1: (a) An example target variable for April 2017. Input data from the previous three months, January, February and March 2017, are used as input to the model to make a prediction as close as possible to the observed VCI3M in April 2017 displayed above. (b) Shap values assign higher absolute values to inputs which had a greater impact on the model prediction. For this model, trained on 12 months of input data, data from more than 3 months away than the prediction date had very little effect on the model's prediction.

2 Data and Methods

2.1 A pipeline for climate science and machine learning

To easily combine data from multiple sources, such as gridded climate data and satellite imagery, we developed an open source pipeline[1] for training and evaluating machine learning algorithms with these geospatial datasets. Although this pipeline was developed for drought forecasting, many aspects of the pipeline are applicable to other hydrological and climatological problems, since we can easily change the predictors and target variables.

The aspects that are generally applicable include:

- downloading and working with geospatial data in raster format.
- the ability to combine time-varying (dynamic) and non-time varying (static) data.
- the importance of spatial and temporal aggregations to the models.

The pipeline was written to be flexible and extensible. We use multiple data sources and demonstrate the potential of our method to aid disaster management and relief efforts.

2.2 Data

We leveraged the pipeline to combine a combination of publicly available datasets to train the models. For this research we focused on VCI3M in Kenya, using data within latitudes (-5.202, 6.002) and longitudes (33.501, 42.283) (see an example in Figure 1a). We used data from 2001 to 2015 as training data, and evaluated the trained models on data from 2016 to 2018. We are using the same VCI data developed by Klisch and Atzberger (2016) which is in operational use by the NDMA in Kenya.

All datasets were spatially regridded to match the ERA-5 spatial grid (Hersbach and Dee, 2016), which has a spatial grid of 0.3 degrees (roughly 30 km), and normalised by subtracting the mean and dividing by the standard deviation. The models received as input the previous 3 months of data to the month being predicted, predicting VCI3M one month into the future.

The data was divided into (temporally) *static data*, and *dynamic data*. This allows us to transfer knowledge between locations with similar characteristics (Kratzert et al., 2019).

[1] `https://github.com/ml-clim/drought-prediction`

Table 1: RMSE results of the baseline persistence model, a standard LSTM and the EALSTM model. We present both the pixel-wise RMSE, and the RMSE of the models when their predictions are averaged across a district. The best results are highlighted.

	Pixel wise			Spatially Aggregated		
District	**Persistence**	**LSTM**	**EALSTM**	**Persistence**	**LSTM**	**EALSTM**
Mandera	8.68	**4.96**	5.48	8.51	**4.13**	4.69
Marsabit	8.65	**4.69**	5.03	8.90	3.25	**3.25**
Turkana	10.22	**4.70**	4.90	9.78	**2.32**	2.92
Wajir	7.77	**4.47**	4.70	7.82	**3.77**	3.92

Dynamic Data: The following raw variables were used as dynamic data: (i) VCI and VCI3M (Klisch and Atzberger, 2016), (ii) evapotranspiration and soil moisture (Martens et al., 2017), (iii) precipitation (Funk et al., 2015), (iv) potential evaporation and temperature (Hersbach and Dee, 2016). In addition to being spatially regridded, datasets were also temporally resampled to monthly time steps, using a monthly mean.

In addition to the raw dynamic data, we passed to the models: (v) spatial means of the dynamic data, collapsing spatial variability and producing one value for each time step, constant across all pixels.

Static Data: The following raw variables were used as static data: (i) topography (Jarvis et al., 2008), and (ii) soil type (Hersbach and Dee, 2016).

In addition to the raw static data, we passed to the model: (iii) a one hot encoding of the month being predicted, (iv) the latitude and longitude of the pixel being predicted, and (v) spatio-temporal means of the dynamic data (collapsing space and time to a single value).

2.3 Models

As predictive models, we trained an LSTM and an *Entity-Aware* LSTM. EALSTMs were first used for rainfall-runoff modelling by Kratzert et al. (2019), and their use in this other hydrological application motivated us to apply them to drought prediction. The models were trained to predict pixel-wise VCI3M for four arid and semi-arid districts in Kenya. As a baseline we employ a persistence model, which predicts VCI3M to be VCI3M one month ago. Persistence is a common baseline in meteorological and hydrological forecasting (Wilks, 2011). To train the models, we used the smooth L1 loss function, also known as the Huber loss function, with $\delta = 1$. The smooth L1 loss is less sensitive to outliers than the mean squared error loss (Girshick, 2015).

EALSTM: Receives both a dynamic and static input. The dynamic data, $X^{dynamic}$, is fed into the network sequentially. The static data, X^{static}, is the same for each time step but unique for each 'entity' (here each entity is a pixel) (Kratzert et al., 2019). We expect the model to learn how different pixel VCI3M values respond to dynamic data (meteorological forcing) differently conditional on the static data (such as topography).

LSTM: Receives the static data appended to every time step of the dynamic data.

3 Results and Analysis

3.1 Model Performance

We measured the performance of the models by calculating the root mean square error (RMSE) of the model predictions. VCI (and therefore VCI3M) is on a scale of 0-100, where 0 is the least healthy vegetation observed, and 100 is the most healthy vegetation observed. The results of the experiment are presented in Table 1. The EALSTM and LSTM both outperform the persistence baseline, in both the pixel-wise and spatially aggregated case.

We experimented with feeding longer time-series to the model, but using model interpretability techniques (specifically, Shap values (Lundberg and Lee, 2017) calculated using DeepLIFT

Table 2: R^2 values reported by Adede et al. (2019), a persistence baseline, and our LSTM and EALSTM models. The best results are highlighted.

District	Adede et al. (2019)	Persistence	LSTM	EALSTM
Mandera	**0.94**	0.66	0.88	**0.94**
Marsabit	**0.94**	0.74	0.93	0.93
Turkana	0.91	0.74	**0.98**	0.95
Wajir	**0.96**	0.72	0.84	0.92

(Shrikumar et al., 2017)), we determined that the model was not using information from more than 3 months before the prediction date (Figure 1b). Reducing the length of the input time-series significantly reduced training time without penalising performance.

3.1.1 Comparison with Adede et al. (2019) vegetation health model

In addition to comparing against a persistence baseline, we compared our models to the Adede et al. (2019) model, developed by researchers at the NDMA and considered state-of-the-art. Using indices derived from temperature, vegetation health, evapotranspiration, potential evapotranspiration and precipitation Adede et al. (2019) train an ensemble of 111 linear neural networks and support vector regressions.

To better compare the models, we retrained our models to receive the same raw inputs as the Adede et al. (2019) model, using raw precipitation, temperature, evapotranspiration, potential evaporation, VCI and VCI3M as input variables (keeping the aggregations, the one hot encoding of the month being predicted and the latitude and longitude being predicted).

Our results are presented in Table 2. Because Adede et al. (2019) only report test results from 2016-2017, we only used those years to calculate the VCI3M R^2 score. In addition, the Adede et al. (2019) model predicts a single VCI3M value per district. To compare their model with ours, we took a district-wide average of our pixel-wise predictions. Our models are competitive with Adede et al. (2019)'s ensemble of 111 models, and produce much more spatially granular predictions. This is particularly encouraging because during the 2016-2017 period being compared, there were droughts in Wajir (NDMA, 2017) and in Turkana and Marsabit (Uhe et al., 2017). These conditions therefore represent the conditions in which we most want the models to perform well.

We ultimately hope our models can supplement the NDMA's disaster relief efforts.

4 Conclusion and Future Work

We have introduced an open-source pipeline for training and evaluating machine learning models with climate and hydrological data. We demonstrate the use of our pipeline by predicting pixel-wise vegetation health in Kenya, producing results that outperform a baseline persistence model and are competitive with the current state of the art model developed by Kenya's National Drought Management Authority (Adede et al., 2019).

We anticipate three future research avenues:

1. Explore the difference in information content between the drought indices used as input to the Adede et al. (2019) model and the raw meteorological and hydrological variables used as input for our models. Is information lost when using current indices?
2. Utilise the tools from interpretable machine learning to better understand the relationship between climate factors and vegetation health. We have integrated the ability to calculate importance scores of input features into our pipeline, using DeepLIFT. We intend to experiment with this further.
3. Incorporate forecasted weather data into our predictions as predictor variables, combining the predictions from physics-based models with machine learning techniques.

References

C. Adede, R. Oboko, P. W. Wagacha, and C. Atzberger. Model ensembles of artificial neural networks and support vector regression for improved accuracy in the prediction of vegetation conditions and droughts in four northern kenya counties. *International Journal of Geo-Information*, 8(12), 2019. doi: 10.3390/ijgi8120562. URL https://www.mdpi.com/2220-9964/8/12/562.

C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2, 2015.

G. Gebremeskel Haile, Q. Tang, S. Sun, Z. Huang, X. Zhang, and X. Liu. Droughts in East Africa: Causes, impacts and resilience. *Earth-Science Reviews*, 193:146–161, jun 2019. ISSN 0012-8252. doi: 10.1016/J.EARSCIREV.2019.04.015. URL https://www.sciencedirect.com/science/article/pii/S0012825218303519.

R. Girshick. Fast R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

H. Hersbach and D. Dee. Era5 reanalysis is in production. *ECMWF newsletter*, 147(7):5–6, 2016.

D. Hillier and B. Dempsey. A dangerous delay: The cost of late response to early warnings in the 2011 drought in the horn of africa. *Oxfam Policy and Practice: Agriculture, Food and Land*, 12 (1):1–34, 2012.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

A. Jarvis, H. Reuter, A. Nelson, and E. Guevara. Hole-filled seamless srtm data v4. *International Centre for Tropical Agriculture (CIAT)*, 2008.

A. Klisch and C. Atzberger. Operational drought monitoring in Kenya using MODIS NDVI time series. *Remote Sensing*, 8(4), 2016. ISSN 20724292. doi: 10.3390/rs8040267.

F. N. Kogan. Global drought watch from space. *Bulletin of the American Meteorological Society*, 78 (4):621–636, 1997. doi: 10.1175/1520-0477(1997)078<0621:GDWFS>2.0.CO;2.

F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. Benchmarking a catchment-aware long short-term memorynetwork (lstm) for large-scale hydrological modeling. *submitted to Hydrol. Earth Syst. Sci. Discussions*, 2019.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

B. Martens, D. Miralles, H. Lievens, R. van der Schalie, R. de Jeu, D. Fernández-Prieto, H. Beck, W. Dorigo, and N. Verhoest. Gleam v3: satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, 10:1903–1925, 2017.

NDMA. Wajir county: Drought early warning bulletin for august 2017. Technical report, Government of Kenya, 2017. URL https://reliefweb.int/report/kenya/wajir-county-drought-early-warning-bulletin-august-2017.

S. E. Nicholson. Climate and climatic variability of rainfall over eastern Africa. *Reviews of Geophysics*, 55(3):590–635, 2017. ISSN 19449208. doi: 10.1002/2016RG000544.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. 2017.

P. Uhe, S. Philip, S. Kew, K. Shah, J. Kimutai, F. Otto, G. J. V. Oldenborgh, R. Singh, J. Arrighi, and H. Cullen. The drought in kenya, 2016-2017. Technical report, Climate and Development Knowledge Network, 2017. URL https://reliefweb.int/report/kenya/drought-kenya-2016-2017.

D. Wilhite, M. Svoboda, and M. Hayes. Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water Resources Management*, 21(5):763–774, 2007.

D. S. Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.