UNIVERSITY OF OXFORD

OKRA

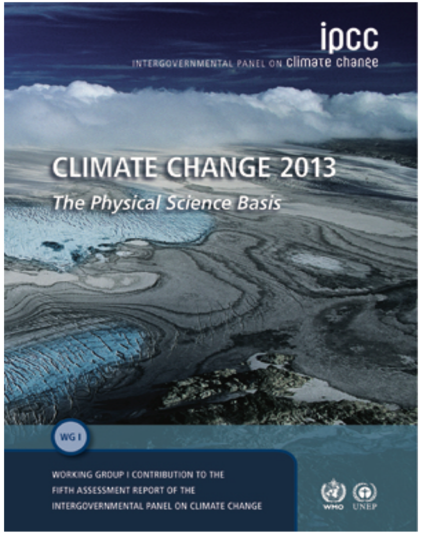
A Machine Learning Pipeline for Drought Prediction

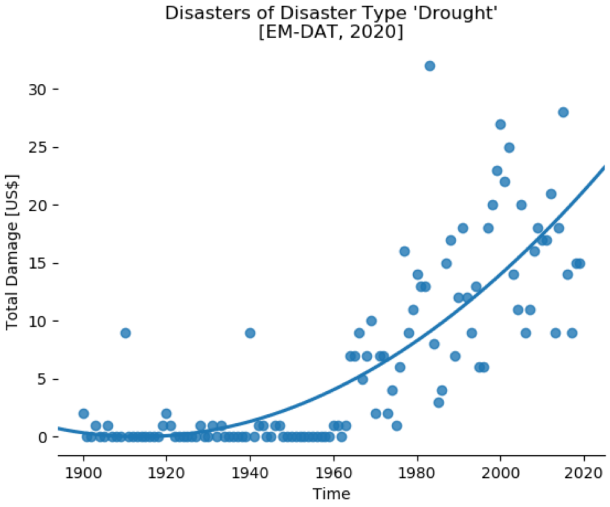
Tommy Lees, **Gabriel Tseng**, Alex Hernandez-Garcia, Clement Atzberger, Simon Dadson, Steven Reece

@tommylees112, @gabrieltseng

tommylees112, gabrieltseng

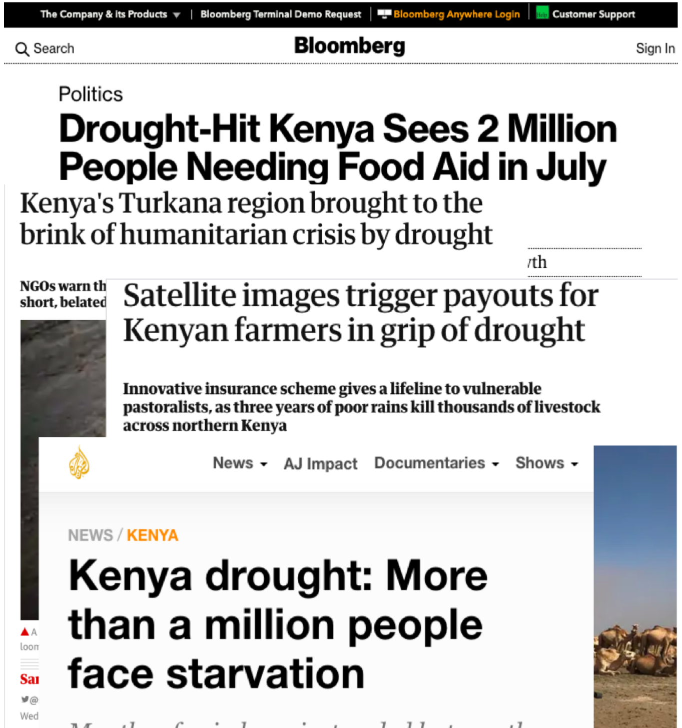
Agricultural drought is a significant global problem, and is getting worse.





$29 billion in losses to developing world agriculture between 2005 and 2015

Kenya distributes emergency funds using a vegetation index, mitigating the impact of droughts.


The Company & its Products | Bloomberg Terminal Demo Request | Bloomberg Anywhere Login | Customer Support

Search Bloomberg Sign In

Politics

Drought-Hit Kenya Sees 2 Million People Needing Food Aid in July

Kenya's Turkana region brought to the brink of humanitarian crisis by drought

NGOs warn th short, belated

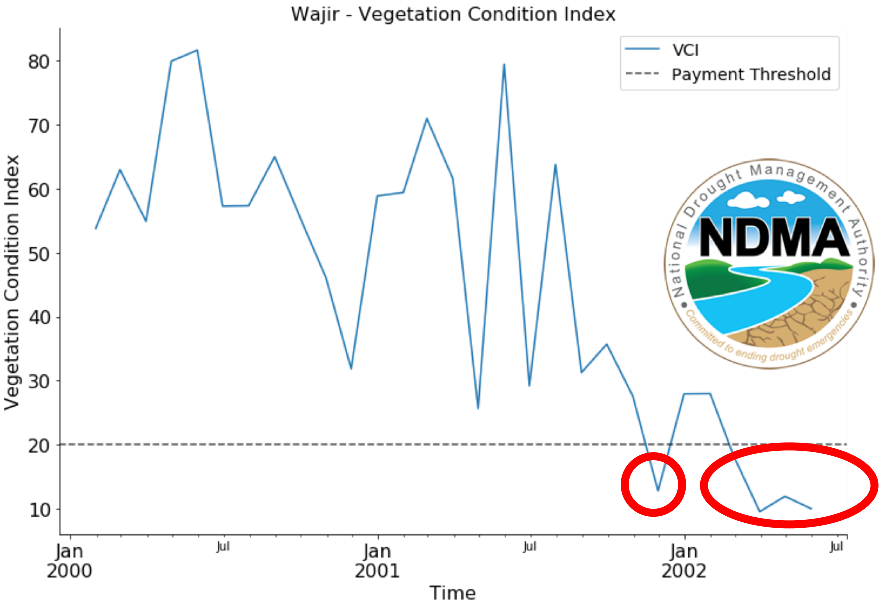
Satellite images trigger payouts for Kenyan farmers in grip of drought

Innovative insurance scheme gives a lifeline to vulnerable pastoralists, as three years of poor rains kill thousands of livestock across northern Kenya

News AJ Impact Documentaries Shows

NEWS / KENYA

Kenya drought: More than a million people face starvation



We sought a machine-learning based approach to forecast vegetation health.

Hydrological & Meteorological Data

$$X_{t-n} = \{x_{static}, x_{dynamic,t-n}\}$$

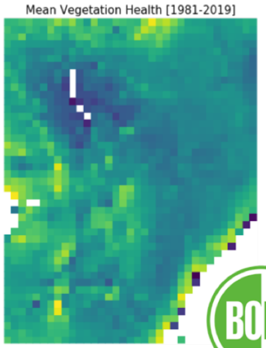
Regression Model

$$y_t = f(X_{t-1,...,t-n})$$

learning a **vegetation health** model

Satellite derived Drought Indicator

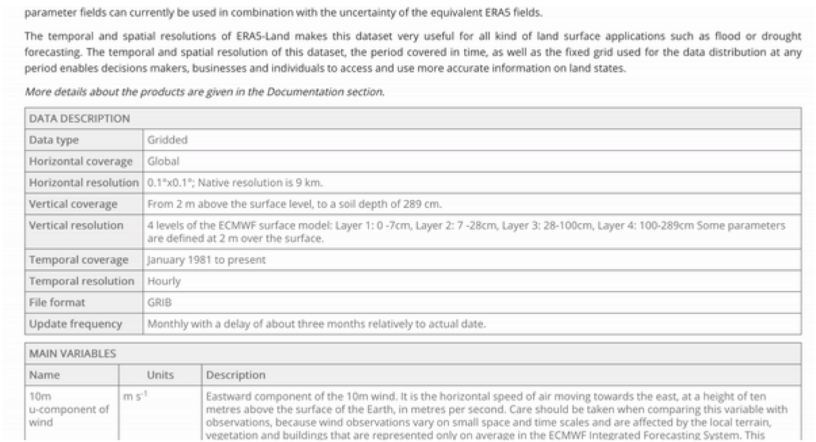
$$y_t \in [0, 100]$$



"Information on certain parameters is not only difficult to access … [we also] lack understanding of the different physical processes. This has led to the widespread use and development of **data-driven models over process-based models**"

Anshuka et. al. 2019

There is friction in applying machine learning to drought forecasting



parameter fields can currently be used in combination with the uncertainty of the equivalent ERA5 fields.

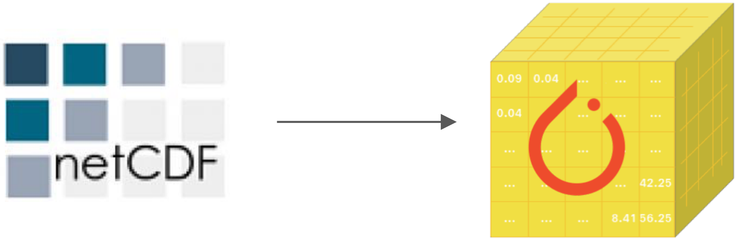
The temporal and spatial resolutions of ERA5-Land makes this dataset very useful for all kind of land surface applications such as flood or drought forecasting. The temporal and spatial resolution of this dataset, the period covered in time, as well as the fixed grid used for the data distribution at any period enables decisions makers, businesses and individuals to access and use more accurate information on land states.

More details about the products are given in the Documentation section.

DATA DESCRIPTION	
Data type	Gridded
Horizontal coverage	Global
Horizontal resolution	0.1°x0.1°; Native resolution is 9 km.
Vertical coverage	From 2 m above the surface level, to a soil depth of 289 cm.
Vertical resolution	4 levels of the ECMWF surface model: Layer 1: 0 -7cm, Layer 2: 7 -28cm, Layer 3: 28-100cm, Layer 4: 100-289cm Some parameters are defined at 2 m over the surface.
Temporal coverage	January 1981 to present
Temporal resolution	Hourly
File format	GRIB
Update frequency	Monthly with a delay of about three months relatively to actual date.

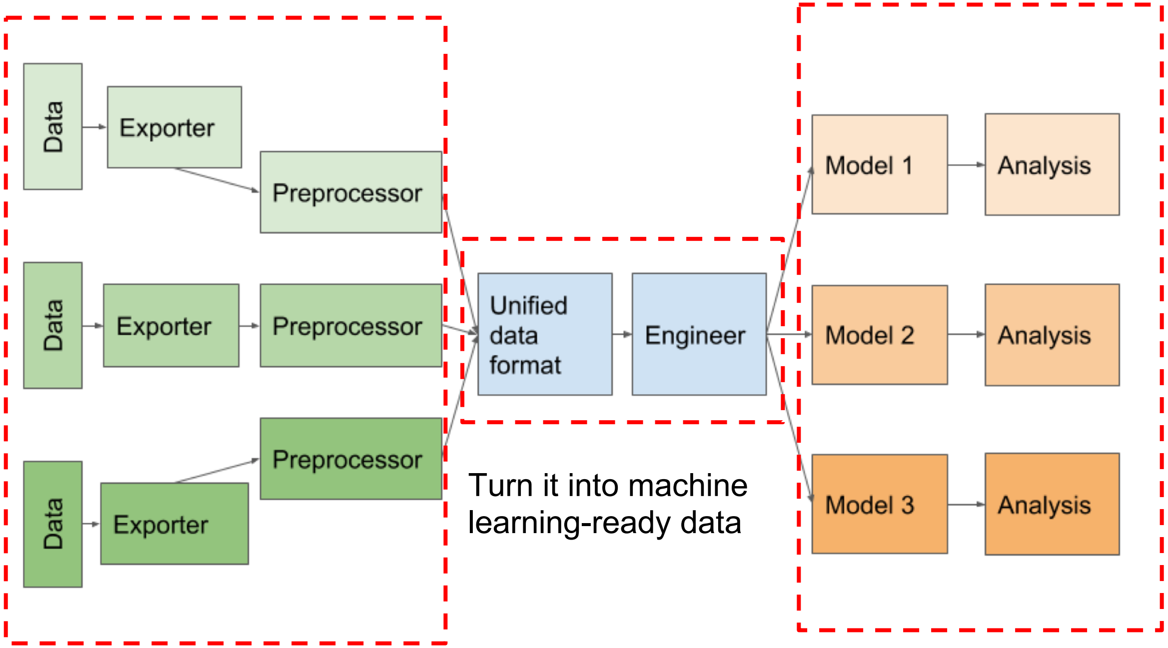
MAIN VARIABLES		
Name	Units	Description
10m u-component of wind	$m\ s^{-1}$	Eastward component of the 10m wind. It is the horizontal speed of air moving towards the east, at a height of ten metres above the surface of the Earth, in metres per second. Care should be taken when comparing this variable with observations, because wind observations vary on small space and time scales and are affected by the local terrain, vegetation and buildings that are represented only on average in the ECMWF Integrated Forecasting System. This

A very large input space (above: ERA5 land variables from the Copernicus Climate Data Store)



Going from climate data formats (e.g. NetCDF) and storage conventions to something which a machine learning model can ingest

Our pipeline* aimed to reduce this friction



Dataset selection and integration

Turn it into machine learning-ready data

Plug and play machine learning models

```
from pathlib import Path
from src.models import LinearRegression

model = LinearRegression(Path("path_to_data"))
model.train()
```

*https://github.com/ml-clim/drought-prediction

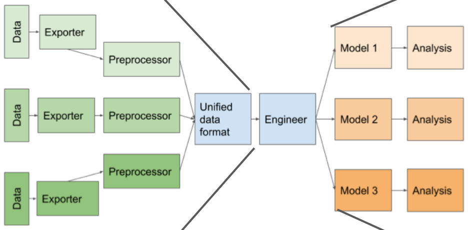
We used it with the following datasets and models:

Copernicus Climate Data Store
ERA5 Climate Reanalysis data

Climate Hazards Group
InfraRed Precipitation data (CHIRPS)

Global Land Evaporation Amsterdam Model
Evapotranspiration and Soil Moisture

CGIAR-CSI
Shuttle Radar Topography Mission Data



Persistence Baseline

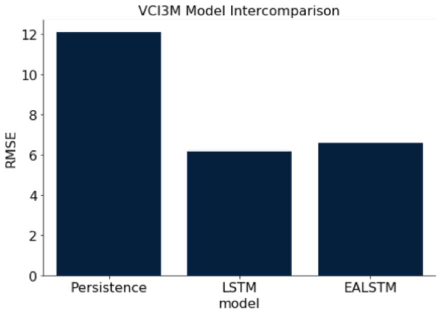
Linear Regression

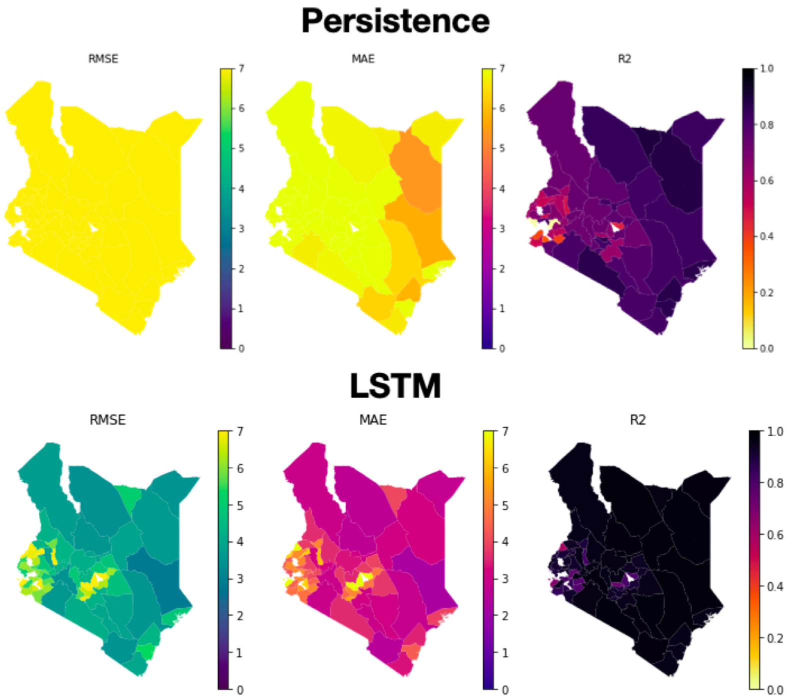
Linear Neural Network

LSTM

Entity-Aware LSTM

Using this pipeline, we were able to achieve results competitive with SOTA to predict vegetation health in Kenya.



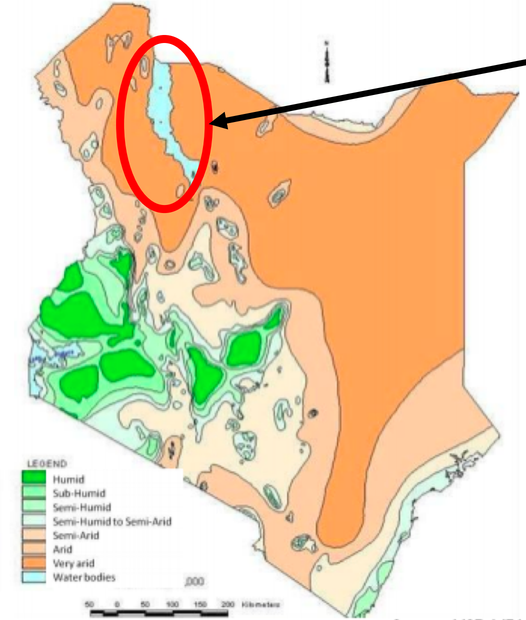


District	Adede et al. (2019)	Persistence	LSTM	EALSTM *
Mandera	**0.94**	0.66	0.88	**0.94**
Marsabit	**0.94**	0.74	0.93	0.93
Turkana	0.91	0.74	**0.98**	0.95
Wajir	**0.96**	0.72	0.84	0.92

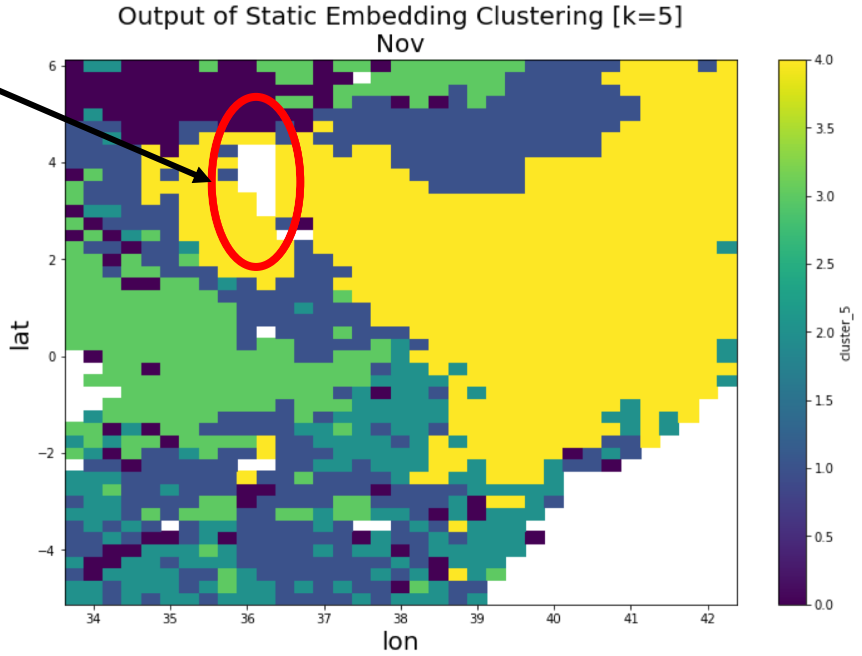
* Our predictions are much more spatially granular (pixel wise vs. district wide) than the current SOTA. In order to make models comparable we downscale our predictions to district-level and compare results at this scale. Here we show a table of results for four arid counties in the North of Kenya.

We have started using trained models to investigate the relationships between vegetation health and weather.

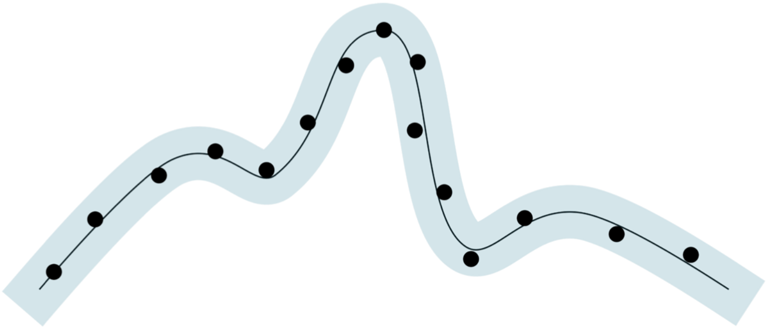
Lake Turkana



Source: USDA/FAS

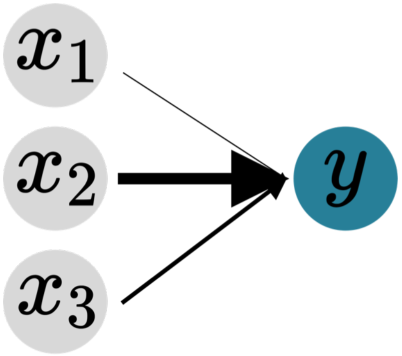


This is how our models need to improve to become operationally useful.



Uncertainty Quantification

Predict the Extremes



Validate Response to Forcings

https://github.com/ml-clim/drought-prediction