An Ensemble Precipitation Prediction Model for Sub-Saharan Africa using a Dendrochronological Reconstruction and Satellite Data

Leif Speer
Freshmen
Terre Haute, Indiana

Sampling in Africa

Study site marked and labeled with a blue pin

I was able to travel to Zambia under a grant to help collect dendrochronological samples, in an effort to gather historical climate data for the region, as well as to help expand the dendrochronological research going on in the sub-Saharan region. Our research group traveled across much of Zambia and collected chronologies at three different sites. In this study I will only be looking at a *Julbernardia paniculata* chronology that was sampled in Zambia's Ngoma Forest in Kafue National Park.

Chronology processing

Most methods used in this project are standard dendrochronological methods. For each chronology in this study, we collected 2-3 cores from each tree, from roughly 20 trees, using 5 millimeter increment bores. We then mounted these cores, and sanded them down using progressively finer sandpaper. With this done, we scanned each core with a resolution of 2,400 DPI. I then cross dated each chronology in the CooRecorder software package and then compiled all cores into a final chronology in the CDendro software package.

An example of the digital scans of the cores

Reconstruction creation

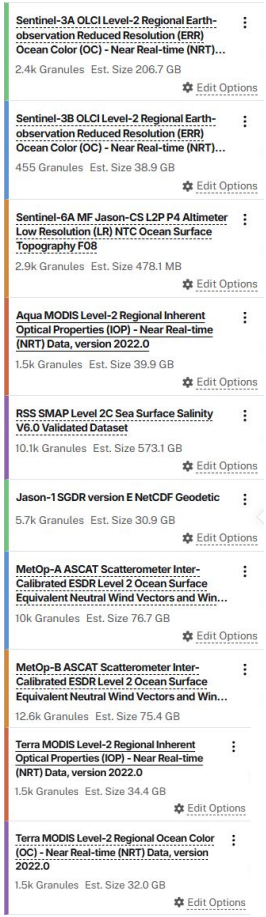
	A	B	C	D	E	F	G	H	I	J	K	L
1	Researche	Species	ID	Start_year	End_year	Series_int	Average_n	# of Cores	Mean Len	EPS_Cutof	Series_int	Keep
2	Slotta	BAO	BWA001	1960	2010	0.341	0.555	16	40.6	N/A		0
3	Trouet	OBA	NAM001	1812	1998	0.045	0.546	15	82.7		N/A	0
4	Trouet	OPA	NAM002	1876	2000	0.086	0.461	9	87.2		N/A	0
5	Stahle	BAO	ZIMB001	1846	1994	0.733	0.619	22	66.5	1930		1
6	Mushove	MZO	ZIMB002	1796	1997	0.691	0.547	31	100.1	1796		1
7	Stahle	SIK	ZIMB003	1870	1996	0.725	0.586	21	94.4	1870		1
8	Ngoma	SES	ZMB001	1958	2013	0.449	0.527	8	43.4	N/A		0
9	Ngoma	KAB	ZMB002	1970	2013	0.603	0.329	11	38.6	N/A		0
10	Ngoma	NAM	ZMB003	1958	2013	0.604	0.484	10	37.5	N/A		0
11	Trouet	NDL	ZMB004	1893	2000	0.023	0.366	6	99		N/A	0
12	Trouet	SLW	ZMB005	1854	2000	-0.048	0.392	9	106.2		N/A	0
13	Trouet	LIV	ZMB006	1917	2002	0.108	0.462	11	73.9		N/A	0
14	Trouet	CH	ZMB007	1896	2000	0.029	0.387	8	92.5		N/A	0
15	Trouet	MBW	ZMB008	1947	2002	0.11	0.52	14	131.8		N/A	0
16	Trouet	MNG	ZMB009	1940	2002	0.09	0.362	18	42.3		N/A	0
17	Trouet	MNG	ZMB010	1953	2000	0.192	0.349	16	36		N/A	0
18	Trouet	MN	ZMB011	1968	2000	0.274	0.421	10	25.2		N/A	0
19	Trouet	PK	ZMB012	1856	2002	0.099	0.487	29	93.1		N/A	0
20	Trouet	NDEA	ZMB013	1903	2000	0.057	0.305	5	78		N/A	0
21	Speer	JUPA	ZMB014m	1885	2024	0.336	0.478	39	99.2	1885		1
22	Speer	AEL	ZMB015	1955	2022	0.457	0.465	20	38.7	1955		1
23	Malubeni	BRLO	ZMB016	1869	2023	0.338	0.574	13	84.9	N/A		0
24	Maxwell	ZFBB	ZMB017	1920	2021	0.395	0.483	24	74.1	1920		1

All of the chronologies that I tested for the reconstruction. A 1 under keep means I used it, a 0 means that I did not.

To create my precipitation reconstruction, I downloaded 19 chronologies from the region from the International Tree Ring Databank (ITRDB), and ran said chronologies, as well as my own, through the COFECHA software package developed by the University of Arizona Laboratory of Tree-ring Research, to validate the quality of the crossdating using the series intercorrelation statistic. In addition to these chronologies, I also used chronologies from Dr. Stockton Maxwell and Collins Malubeni that are not publicly available. Of these chronologies, my chronology, Dr. Maxwell's chronology, and three chronologies from the ITRDB met my standards to be used in the reconstruction. I then ran the chronologies through the ARSTAN program developed by Dr. Ed Cook to remove the age related growth trend by applying an age-dependent spline with a 50-year kernel to the ring width index, and determine the EPS date when the signal dominates the noise. I combined the files with a nested principal component analysis, and finally ran it through an R script by Dr. Shantos Shaw to transform ring width to precipitation.

Satellite Data

The datasets that I downloaded, representing 9 satellites and 10 datasets (2 from the Terra satellite)

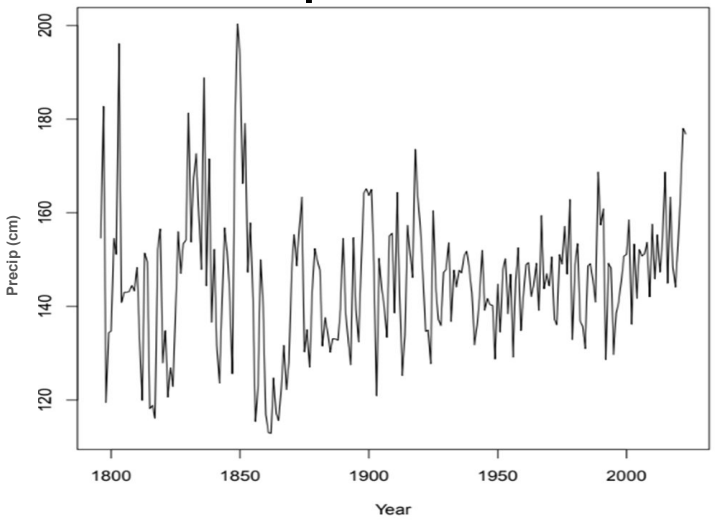


To help create the prediction model, I decided to incorporate satellite data, as satellites can provide modern, high resolution data, while my reconstruction has only an annual resolution. To do this, I manually downloaded data from NASA's EOSDIS Earthdata website. This was a rather difficult task, as many of the datasets were hundreds of terabytes in size. I had the option of using NASA's Amazon Web Service, but decided against that, because the storage cost was still to great. Instead I downloaded the data onto an external hard drive. To select the datasets that I would test, I researched satellites that studied hydrological systems on the planet, as I figured this would be the most applicable for predicting precipitation. I came up with a list of over 20 satellites. When I went to download data, this list was sized down to 9 satellites. These satellites were selected for having smaller datasets that I could feasibly do analysis on, as well as for having data stored in NetCDF format. After I downloaded these datasets, together over a terabyte of data, I further narrowed these down to just 4 satellites for having a time dimension with which to concatenate. Finally I resampled at a monthly resolution, as this was the resolution I was predicting at, and incorporated lags in the datasets based on correlations with what little instrumental precipitation data from the region there was. With this, I was ready to do analysis.

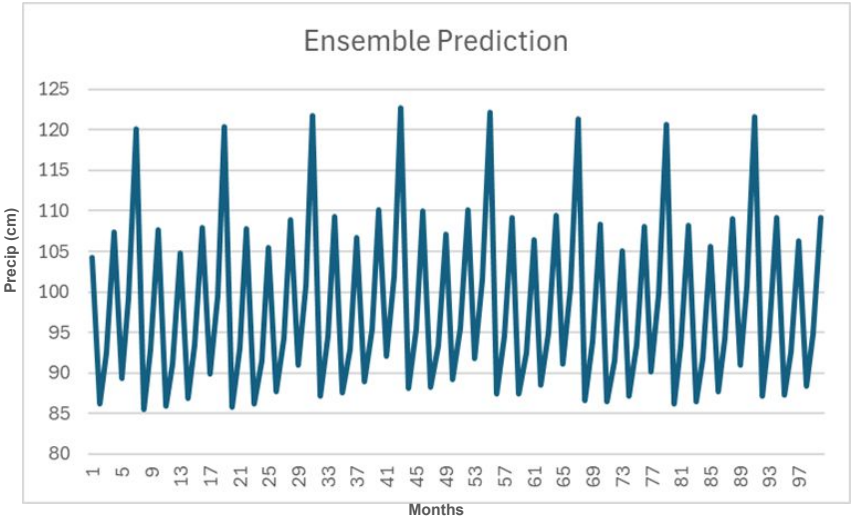
Creating the prediction model

In this study, I exclusively used Gradient Boosting models, sometimes called Decision Tree models. I chose to do this largely because of computation limitation, as Neural Network or Transformer models are much more computationally costly that Gradient Boosting models are. The three models that I used were eXtreme Gradient Boosting model (XGBoost), Light Gradient Boosting model (LightGBM), and Categorical Boosting model (CatBoost). These models were selected for their high performance metrics in review papers. I ran each model for each satellite, conducted hyperparameter optimization in each run using a Grid Search method, evaluated each model with the Mean Squared Error (MSE) statistic, and stacked the outputs using a standard weighted stacking method.

Notable products



This is my reconstruction going back 200 years. Variability increases before 1870, though this is likely due to low sample depth at that age. The reconstruction also, somewhat surprisingly to me, shows an increasing trend in the modern era. My reconstruction was statistically significant ($p < 0.05$).



This is my stacked prediction 100 months into the future. As you can see, seasonality is represented highly in the prediction, while low frequency variability was not so much. In general, this is what we would expect to see, although I was surprised by the extent that seasonality dominated the projection. This result represents three models trained on both satellite data and my reconstruction.