Predicting Food Shortages in Africa from Satellite Imagery

New Mexico

Supercomputing Challenge

Final Report

April 3, 2018

LAHS56

Los Alamos High School

Lillian Kay Petersen

Mentors: Daniela Moody and Rick Chartrand

Contents

List of Figures

List of Tables

1 Executive Summary

Developing countries often have poor monitoring and reporting of weather and crop health, leading to slow responses to droughts and food shortages. Here, new satellite data analysis tools were created to monitor crop health in Africa. The method was first tested in Illinois where there is reliable, high-resolution crop data. Measures of vegetation health were computed from 120-meter resolution MODIS satellite imagery since 2000. The author wrote 4000 lines of python code to process 12 terabytes of data. Correlations were computed between corn yields and monthly satellite index anomalies for every county and year, and a multivariate regression using every index and month (1600 values) produced a correlation of 0.86 with a p-value <1e-6. The high correlations in Illinois show that this model has good forecasting skill for crop yields. Next, the method was applied to three countries in Africa: Ethiopia, Tunisia, and Morocco for each country's main crop. All three had high correlations with the maximum monthly satellite index during the rainy season, at 0.99, 0.97, and 0.73 respectively. The satellite analysis methods and software tools developed here can be used to predict crop production two to four months before the harvest, and many more months before official crop data is published. Satellite imagery was then processed for every African country, and a publicly viewable interactive map displaying real-time crop predictions was posted online. This method is unique because it can be applied to any location, crop, or climate, making it ideal for African countries with small fields and poor ground observations. The author is actively engaged with several international aid organizations that are interested in using this early warning system of impending food shortages.

2 Introduction

In the United States, there is exceptional monitoring and reporting of weather and crop health, with thousands of weather stations and county-level crop yield data from the USDA that has been recorded since 1910 (Hamer et al., 2017; Menne et al., 2012). With this substantial amount of publicly available data, crop yields may be predicted based on historical records. However, not all parts of the world have open, reliable data (McKinnon, 2016). The availability of weather and crop data depends on the government's ability to collect it, financial resources, and willingness of authorities to share it. Lack of data is an especially important problem in developing countries where crop yields are less stable and droughts can lead to famines, death, government instability, and war. Therefore, there is a major need to monitor crop health in the developing world. Satellites provide coverage over the entire earth and certain bands may be used to assess plant health and drought conditions. This would enable scientists to monitor risks of food shortages and alert governments and international aid organizations in real time.

Crop yields in developing countries do not benefit from the same level of agricultural technology as the US, and therefore have much lower yields. Since 1970, corn yields have doubled in the US from 80 bu/acre to 160 bu/acre due to improvements in agricultural technology such as irrigation, pesticides, herbicides, fertilizers, and plant breeding (Figure 1a). In developing countries, crop yields are both much lower and much more variable than in the US, both geographically and in time (Mann and Warner, 2017b). For example, Ethiopia's corn yield has increased from 15 to 55 bu/acre since 1960 (Figure 1b), which is still one-third the corn yield of the US. Farmers in poor countries lack the financial resources and education to use the advanced technology used by the American and European farm industries. Therefore, crop yields in African countries are much more susceptible to the dangers of heat waves and droughts.

Satellite imagery has been extensively used for crop monitoring for decades. The majority of these studies are in the United States, where there is an immense amount of yield and production data at high resolution. Such data significantly improves agricultural research, but is only affordable by developed countries. The US also has large fields of a small number of individual crops, mainly corn, soybeans, and wheat. This allows research to be specific to individual crops and locations, and uniform crops within each satellite pixel. For example, Johnson (2016) developed algorithms

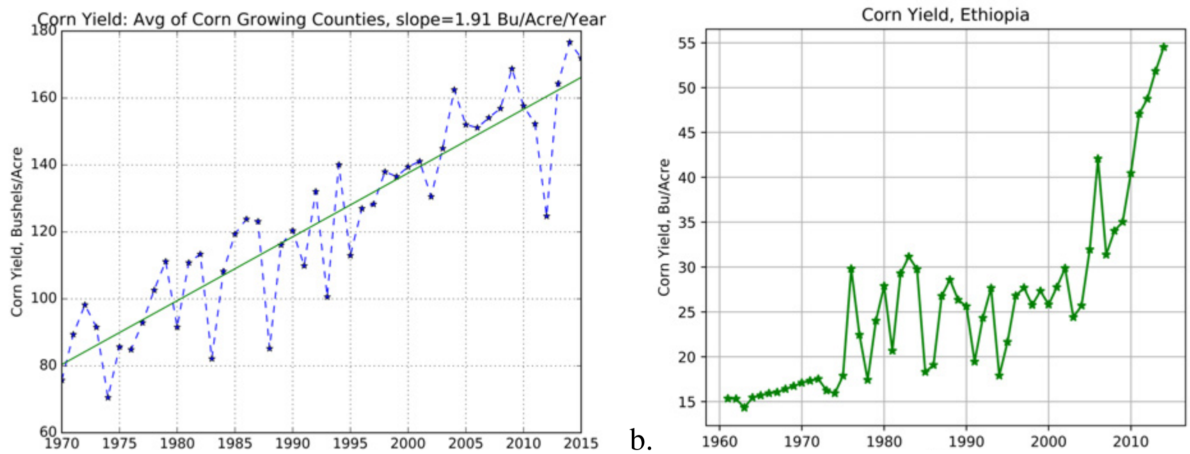
a.  b.

Figure 1. Illinois (a) and Ethiopia (b) corn yield over time. Both have improved significantly, but the yield in Ethiopia is still one-third of the U.S. All plots in this paper were created by the author.

to identify crops in the US from MODIS imagery and analyzed each crop individually. Gao et al. (2017) utilized week-by-week plant growth data in Iowa to design a method to monitor the growth stages of corn and soybeans from satellite imagery.

These types of studies are not possible in Africa because there is minimal reporting of crop health and yields; farms consist of very small plots of varied crops interspersed with buildings; and the continent contains a vast number of different climates, growing seasons, and crops. Despite these difficulties, a few studies have examined the climatology of specific countries or regions in the developing world. Gissila et al. (2004) correlated seasonal rainfall in Ethiopia with sea-surface temperature anomalies in the Indian ocean and the central Pacific. Tadesse et al. (2014) predicted NDVI (Normalized Difference Vegitation Index) 1–3 months in the future from multiple indices (land cover, standardized deviation of NDVI, etc.) as a means of forecasting droughts. Other studies develop models that forecast crop yields. NDVI/yield regressions for cereals at national level have been developed for specific countries in northern Africa (Rembold et al., 2013). Mann and Warner (2017b) use kebele (district) level data collected by the Ethiopian government, including crop damage, elevation, fertilizer use, population density, and road density, to estimate wheat output per hectare. I contacted the Ethiopian Central Statistical Agency, Mann, and Warner in an attempt to obtain this detailed, high-resolution data. Unfortunately, the Ethiopian government refuses to release data, even for agricultural research. Mann and Warner were only able to obtain this data under strict conditions and after years of collaboration (Mann and Warner, 2017a). These factors all contribute to the difficulty of developing predictive tools for crop yields in Africa.

The method of predicting crop yields developed here differs from previous work in the U.S. and Africa because it is an overall measure of relative vegetation health compared to the mean climate on a per-pixel bases. Unlike previous studies, it may be applied anywhere in the world—it does not depend on special tuning for the particular crop, region, or climate of interest. The method was created for developing countries where detailed monitoring on the ground simply does not exist, but was successfully validated against extensive crop data in Illinois.

3 Methods

The overall goal of this research is to create a predictive measure of crops computed from satellite data. Python code was written by the author in order to obtain satellite images, mask out clouds, calculate vegetation and water indices, compute monthly anomalies since 2000, and correlate the anomalies of the satellite indices with crop yield anomalies for every county in Illinois and then apply the same method to three countries in Africa.

MODIS (Moderate Resolution Imaging Spectroradiometer) imagery was obtained from the Descartes Labs satellite platform at a resolution of 120 meters (Figure 2a, 2b). MODIS, hosted on the satellites Aqua and Terra, has a return time of one day, giving almost continuous imagery of every location on earth since 2000. The instruments capture 36 spectral bands ranging from wavelengths of 0.4 μm to 14.4 μm (Jenner, 2015).

Clouds and snow in images can disrupt data and distort values. In order to account for cloud contamination, clouds were identified based on the values of the bands blue, red, NIR, and SWIR. Pixels with clouds or snow were not included in monthly averages and images with over 80% clouds were thrown out (Figure 2c).

To measure the health of crops throughout the growing season, three indices were computed: NDVI, EVI, and NDWI (Table 1). All three indices range from -1 to 1. Areas containing dense vegetation show high NDVI and EVI values, between 0.4 and 0.8, desert sands will register at about zero, and snow and clouds are negative. NDVI is sensitive to chlorophyll, which absorbs visible light, from 0.4 to 0.7 μm, for use in photosynthesis. In contrast, EVI detects canopy structural variations, including leaf area, canopy type, and canopy architecture (Herring and Weier, 2000).

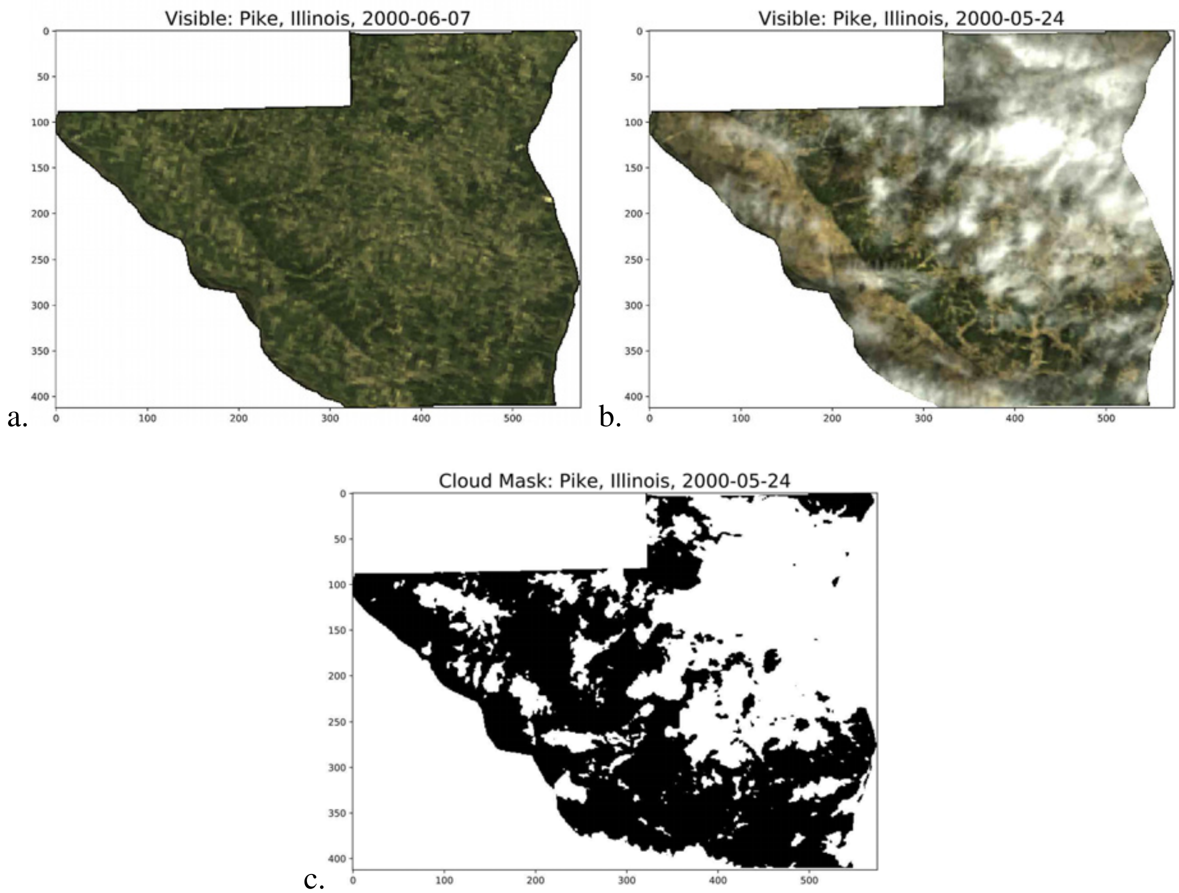


Figure 2. Snapshots of two MODIS satellite passes over Pike county, Illinois (a, b) and the cloud mask for the second image (c).

NDWI detects water content. Combined, all three indices complement each other on the detection of vegetation changes.

For every pixel in Illinois, the NDVI, EVI, and NDWI monthly averages and climatologies were computed. The climatology is defined as the average over years 2000 through 2016 for each month and pixel. Next, the monthly climatology was subtracted from the monthly average for every pixel, resulting in the monthly anomaly. The pixels in each county were then averaged together to find the monthly anomaly for NDVI, EVI, and NDWI.

Annual corn yield data was downloaded for every county in Illinois for years 2000 through 2016 from the USDA (Hamer et al., 2017). Because each county has different growing conditions (soil quality, hills, proximity to large water bodies, etc.), the mean was subtracted out of each county's corn yield to find the yield anomaly. Correlations were found between each county's corn

Index	Description	Formula
NDVI	Normalized Difference Vegetation Index	$NDVI = \frac{NIR-Red}{NIR+Red}$
EVI	Enhanced Vegetation Index	$EVI = G * \frac{NIR-Red}{NIR+C1*Red-C2*Blue+L}$
NDWI	Normalized Difference Water Index	$NDWI = \frac{Green-NIR}{Green+NIR}$

TABLE 1. Definitions of indices to measure crop health. NIR is near infrared, L is the canopy background adjustment that addresses non-linear, differential NIR and red radiant transfer through a canopy, and C1, C2 are the coefficients of the aerosol resistance term, which uses the blue band to correct for aerosol influences in the red band.

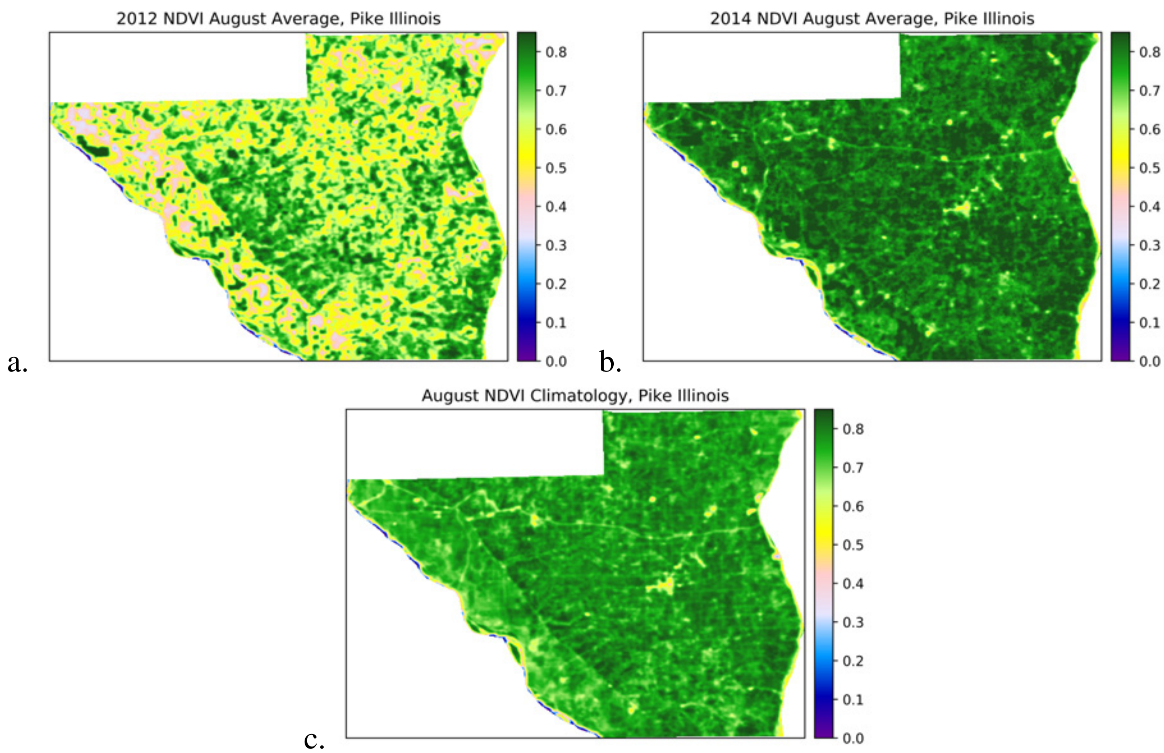


Figure 3. August average NDVI for a drought year (a) and a wet year (b), and the NDVI August climatology (c).

anomaly and the three satellite indices. To find the best prediction measure possible, a multivariate regression was fit to each month and index for a total of 15 variables.

The same method was then applied to three countries in Africa: Ethiopia, Tunisia, and Morocco, and later to every country in Africa. In each country, a box was analyzed where the majority of the crops are grown (Figure 12) and was then correlated to national crop production data from USDA (2018). A total of 4000 lines of code were written to process twelve terabytes of raw data and produce the graphs. A code repository is maintained at the author's GitHub site: `https://github.com/lillianpetersen/CropPredictionFromSatellite2018`.

4 Results

The method was first validated in Illinois and then applied in Africa.

4.1 Illinois

Illinois corn yield is highly correlated with NDVI, EVI, and NDWI. The correlations at the state level are extremely statistically significant at 0.9, 0.85, and -.92 respectively (Figure 4). NDVI and EVI both have a positive relationship to crop yields, and NDWI is inversely related. Strong NDWI in critical growing stages could indicate insufficient evapotranspiriation, resulting in a negative correlation.

In 2012, the central United States was hit by a drought and Illinois had a lower than average corn yield and a negative NDVI anomaly. Yields and NDVI anomalies in 2014 were significantly higher. These two years are used as examples to show corn yield and satellite anomalies at the county level (Figure 5).

Next, the satellite anomalies were plotted against the corn yield anomaly for every county and year, for a total of 1559 points. August has the highest correlations with corn yields at 0.7, 0.71, and -.73 for EVI, NDVI, and NDWI respectively (Figure 6). July has less predictive skill than August, and the other months are almost uncorrelated with yields (Figure 7). All of July and August's correlations have a P value less than 0.000001 (GraphPadSoftware, 2018), meaning there is less than one in a million chance of them occuring through a random process.

Correlations have been computed with three indices (NDVI, EVI, and NDWI) and five months, for a total of fifteen independent variables. In order to create a single predictive measure of corn yields, a multivariate regression was fit to every index and every month using a Python machine learning library. Figure 8 shows an example of the multivariate regression for two of the variables. The multivariate regression improved the individual correlations to 0.86.

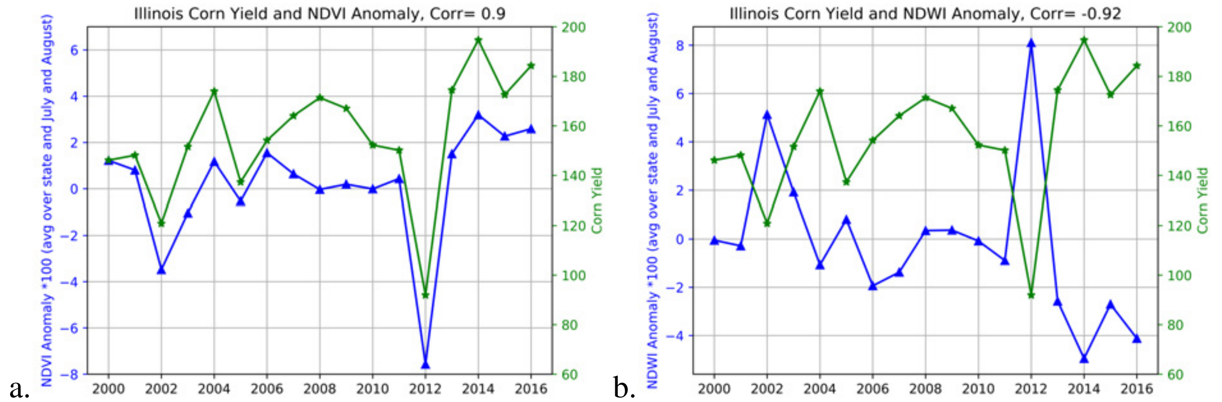


Figure 4. Illinois mean corn yield since 2000 (green) correlated with NDVI (a, blue) and NDWI (b, blue).

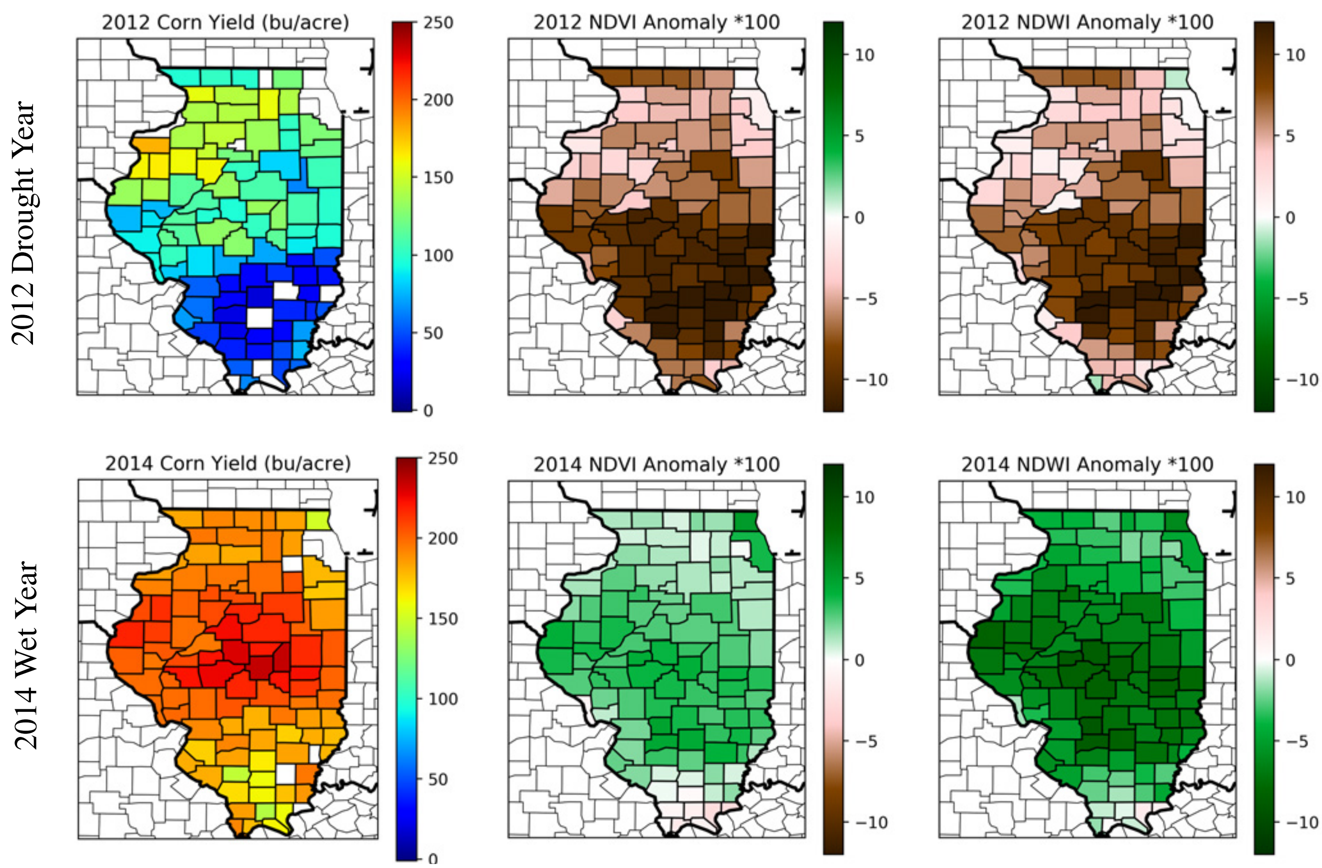


Figure 5. Corn yield (left), NDVI anomaly (center), and NDWI anomaly (right) by county in Illinois for the drought year 2012 (top) and for the wet year 2014 (bottom). During the drought year, there are low yields, low NDVI anomalies, and high NDWI anomalies, while the drought year is opposite.

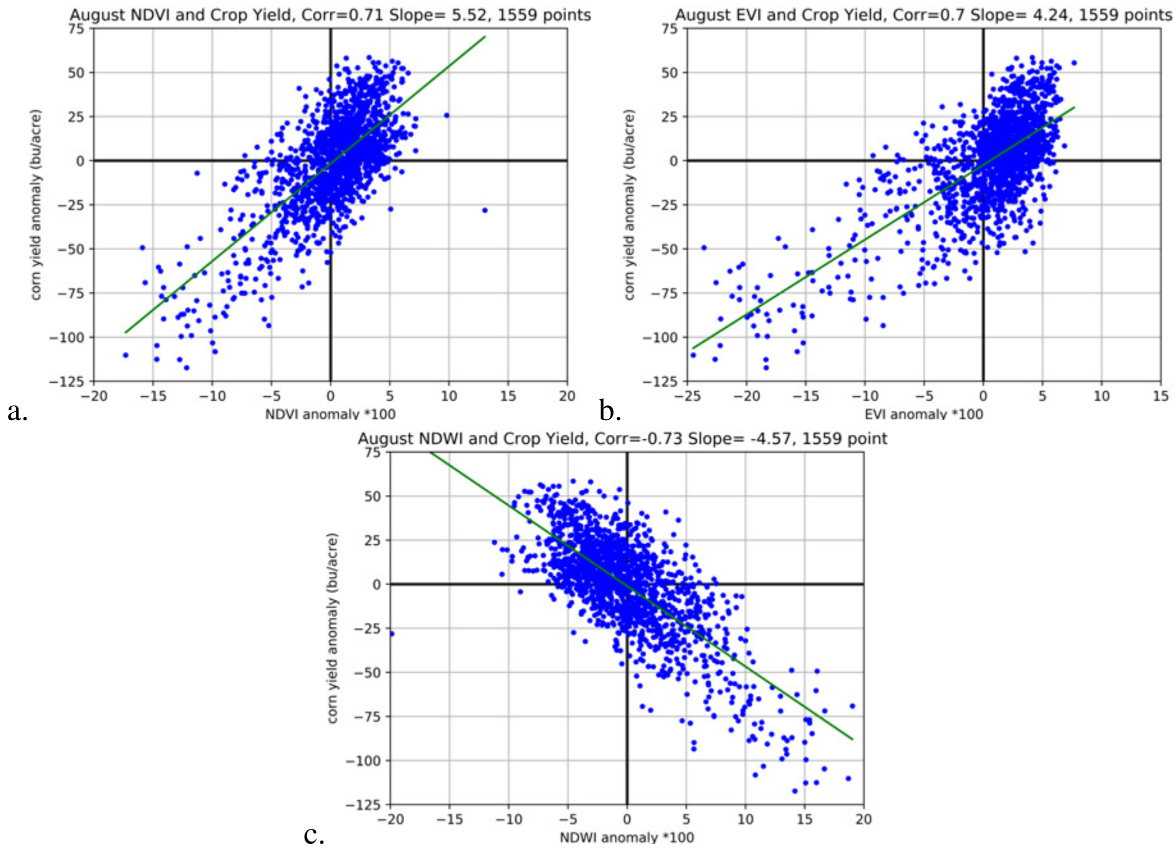


Figure 6. Correlations between Illinois corn yield and August average NDVI (a), EVI (b), and NDWI (c). All correlations are extremely significant with P values of <0.000001. August had the highest correlations to yields out of all the months.

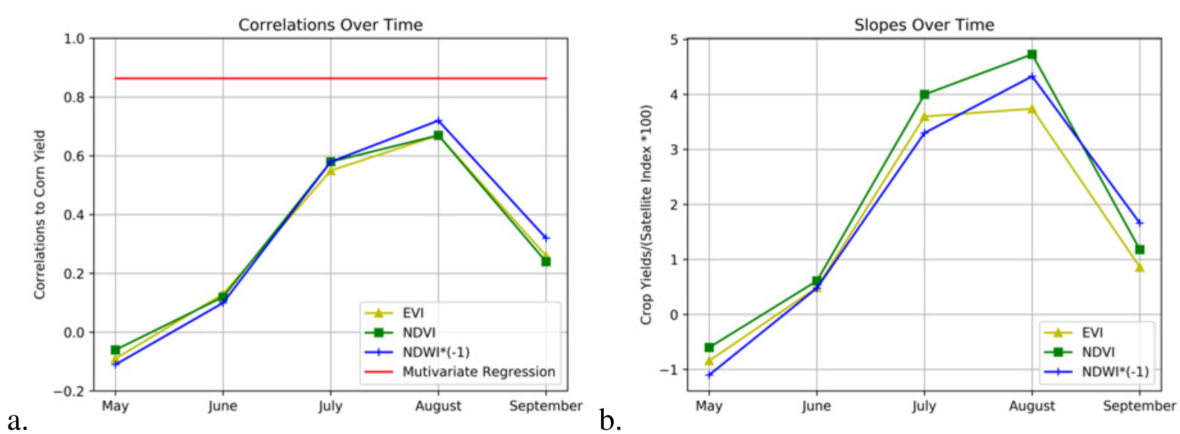


Figure 7. The absolute value of the correlations (a) and slopes of the linear regressions (b) for each month between Illinois corn yield and NDVI (green), EVI (yellow), and NDWI (blue). July and August have the highest predictive skill for crop yields which are harvested in October, meaning there is a two to three month lead time on yield estimates. The red line shows the correlation of the multivariate regression, which is higher than any individual month.

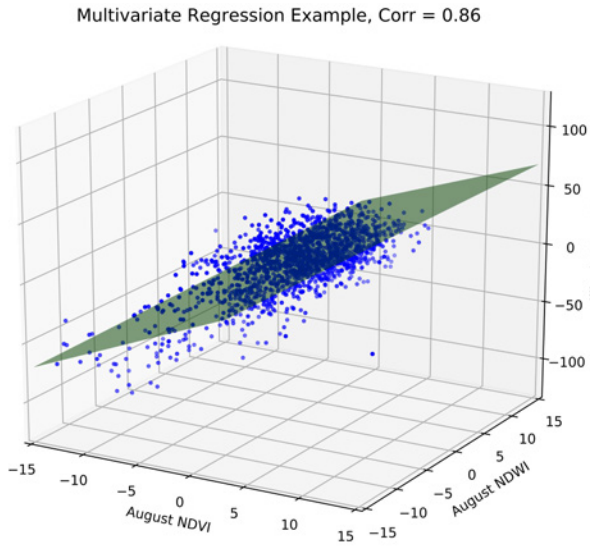


Figure 8. An example of the multivariate regression comprised of all three satellite indices and months. The multivariate regression improved the individual correlations to 0.86.

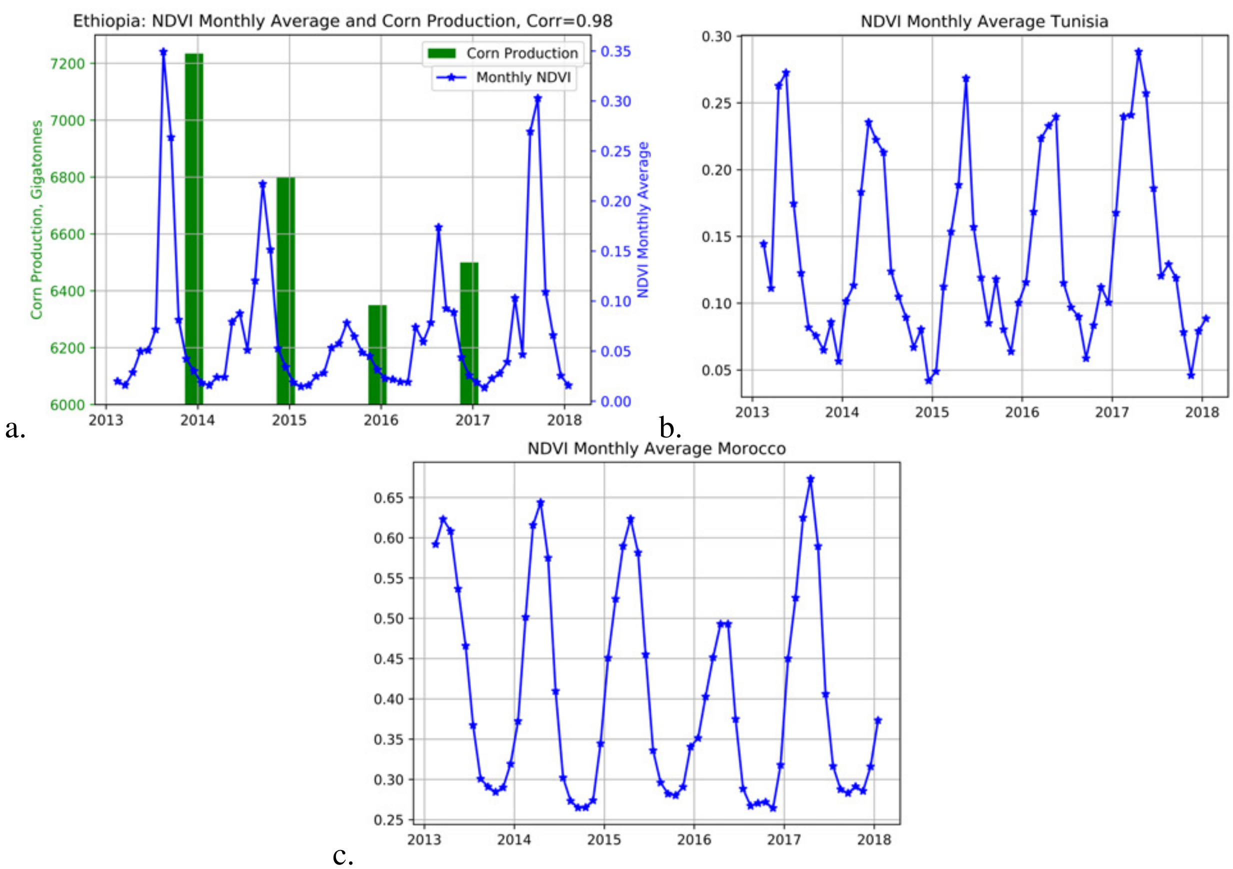


Figure 9. NDVI monthly average for Ethiopia (a), Tunisia (b), and Morocco (c). The annual rainy season produces high NDVI values and corresponds to the crop-growing months. Ethiopia also has the corn production overlayed, which has an almost perfect correlation to maximum NDVI at 0.98.

4.2 Africa

The high correlations in Illinois show that this model has good forecasting skill for crop yields. Next, this method was applied to three countries in Africa: Ethiopia, Morocco, and Tunisia. For each country, a box was analyzed where the majority of the crops are grown (Figure 11a).

In most places in Africa, there is a wet and a dry season. For example, the wet season in Ethiopia spans from June to September, and crops are harvested in December. This is known as the Meher growing season. Ethiopia's core agriculture and food economy is comprised of five major cereals: corn, teff, wheat, sorghum, and barley. These cereals accounted for about three-quarters of total area cultivated and 29 percent of the agricultural GDP in 2005/06 (Taffesse, 2012).

The wet and dry seasons are evident in the monthly NDVI values for all three countries (Figure 9). During the wet season, the crops green and the NDVI values spike. During the harvest, the values drop. The crop examined in each country was chosen based on the production quantity. Corn and sorghum were evaluated in Ethiopia, and wheat was examined in Tunisia and Morocco. It was found that Ethiopia and Morocco have the best correlation to the maximum NDVI value of the growing season, while Tunisia has the highest correlations to NDWI.

There was a major drought in Ethiopia in 2015, and 2013 was a very wet year by comparison. These vegetation differences can also be seen on the pixel level (Figure 11). The anomalies are especially evident in the Rift Valley where most of the crops are grown.

Ethiopia's maximum NDVI values, which usually occur in August, are extremely well correlated with grain production, at 0.98 and 0.99 for corn and sorghum respectively (Figures 9a, 10a). That is an almost perfect correlation between the crop production harvested in December and satellite imagery four months earlier. Tunisia has a correlation of 0.97 and Morocco has a correlation of 0.73 for wheat (Figure 10b, 10c), showing high predictive skill of satellite indices in all three countries.

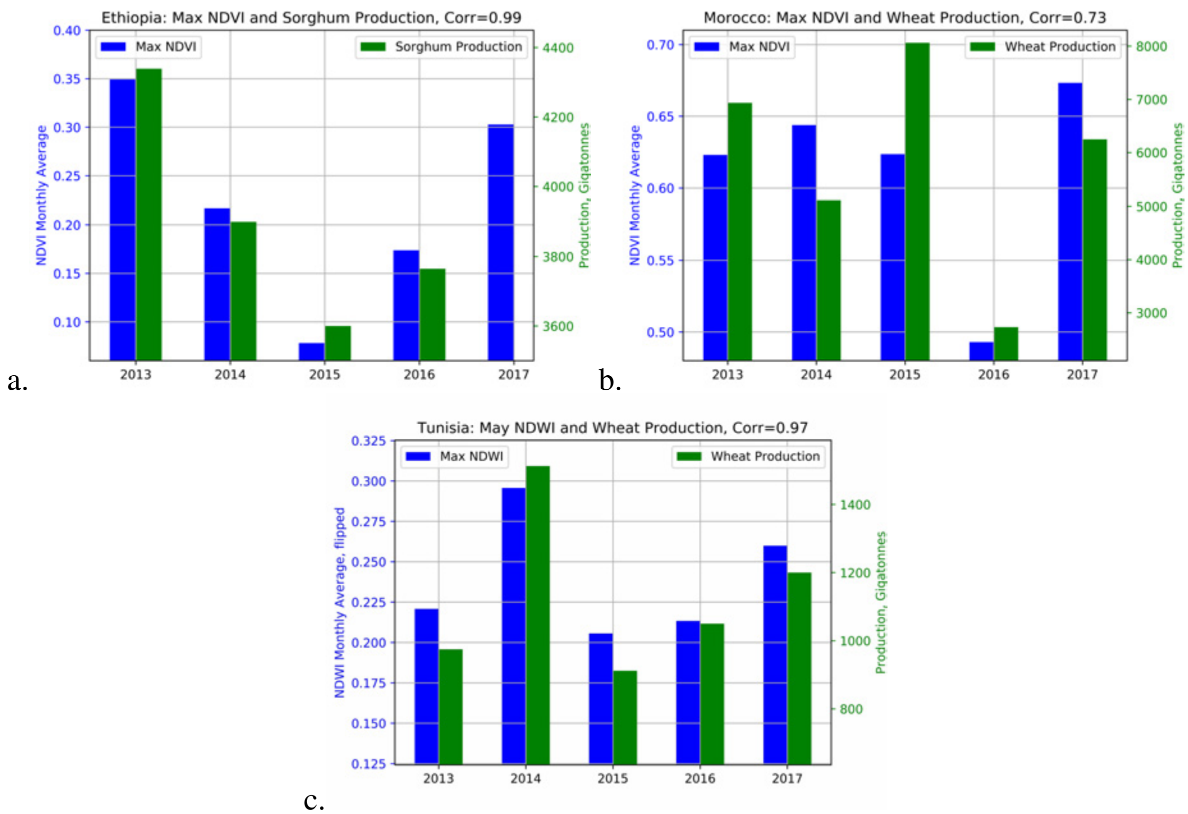


Figure 10. Maximum NDVI value of the growing season (green) with crop production (blue). All countries have significant correlations ranging from 0.99 to 0.73. Ethiopian producition data for 2017/2018 has not been published because crops are harvested from November to February.

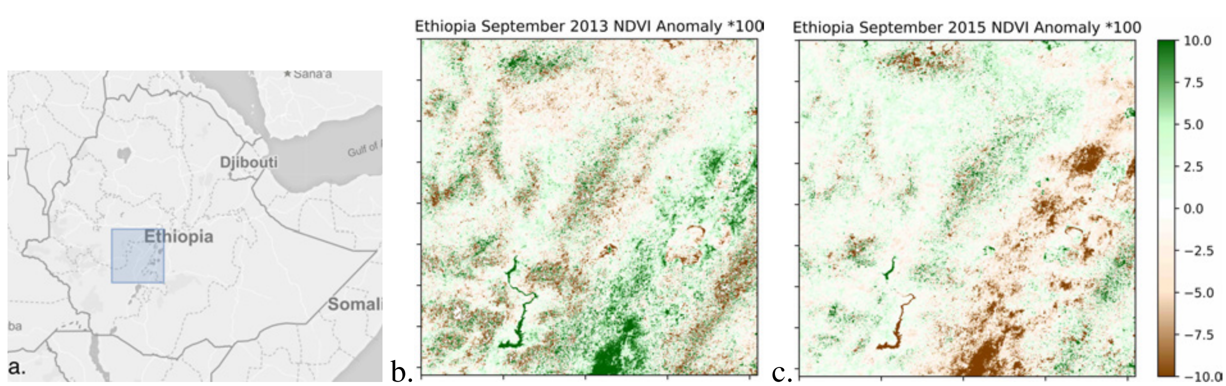


Figure 11. The box examined in Ethiopia (a) during a wet year (b) and a dry year (c). The NDVI anomalies are especially high in the rift valley, where farming is the most dense.

4.3 Africa: Prediction of Future Crop Production

Satellite imagery was processed for every African country. First, a box in an agricultural region was selected in every one of the 46 countries in Africa and a total of 12 terabytes of daily satellite imagery was processed according to the method above. Correlations and linear regressions were computed in every country and every crop. Difficulties in finding accurate correlations include:

- False reporting of production in some countries, due to lack of resources, poor oversight, or corruption (e.g. DR Congo, Eritrea, Libya)
- Multiple growing seasons in central countries (Rwanda, Somalia)
- Growing seasons across the December - January year boundary (Tanzania, Botswana)
- Clouds every day for months at a time in central African countries (Gabon, Cameroon)
- Time delays and misclassification of harvests in October–December, where production is incorrectly reported in the following calendar year (Nigeria, Sudan)

In every African country, correlations were computed between six indices (NDVI, EVI, NDWI, averages and anomalies) and for every crop. The highest correlation in each country was examined. Despite the above difficulties, two thirds of the correlations are considered to be statistically significant ($r>=0.75$ for five years, Figure 13)

Satellite imagery was then processed up to the current date for countries that are in growing season. Real-time predictions were computed for each of these countries and their heighest correlating crop from the linear regressions (Figure 14). Next, an interactive map of the predicted production for harvests in the next few months was created and is now publicly viewable through https://lillianpetersen.github.io/africa_satellite. This map can give governments and aid organizations advance notice to see which countries are at the highest risk of a food shortage in order to better prepare supplies, transportation, and manpower for a rapid response.

The author is currently engaged with several international aid organizations who are interested in this product, including the International Food Policy Research Institute (IFPRI), the US Dept. of Agriculture (USDA), and the Global Agricultural Monitoring (GEOGLAM) group. The author has been invited to give hour-long talks to these institutions in Washington, DC.

Figure 12. A box was chosen in the densest agricultural region for each country in Africa.

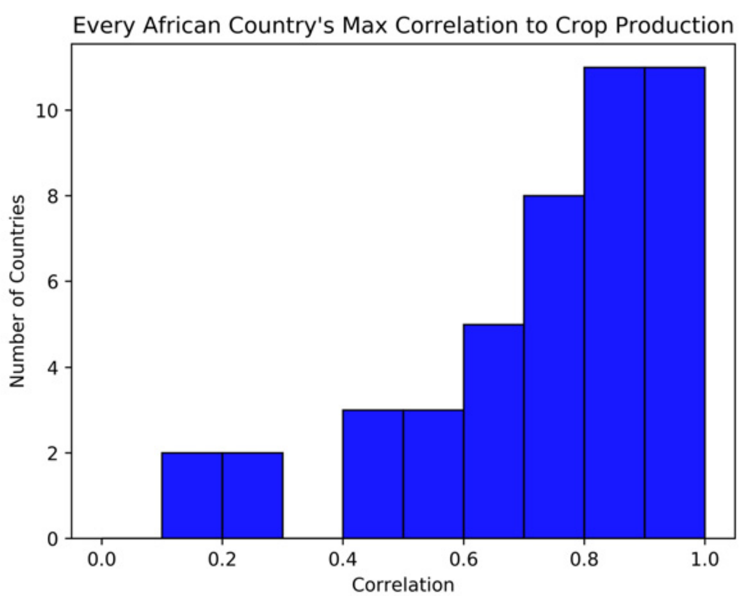


Figure 13. The highest correlations in every country in Africa. Two thirds are considered to be statistically significant.

Production Predictions:
March 2018

For countries with height of growing season in October - February

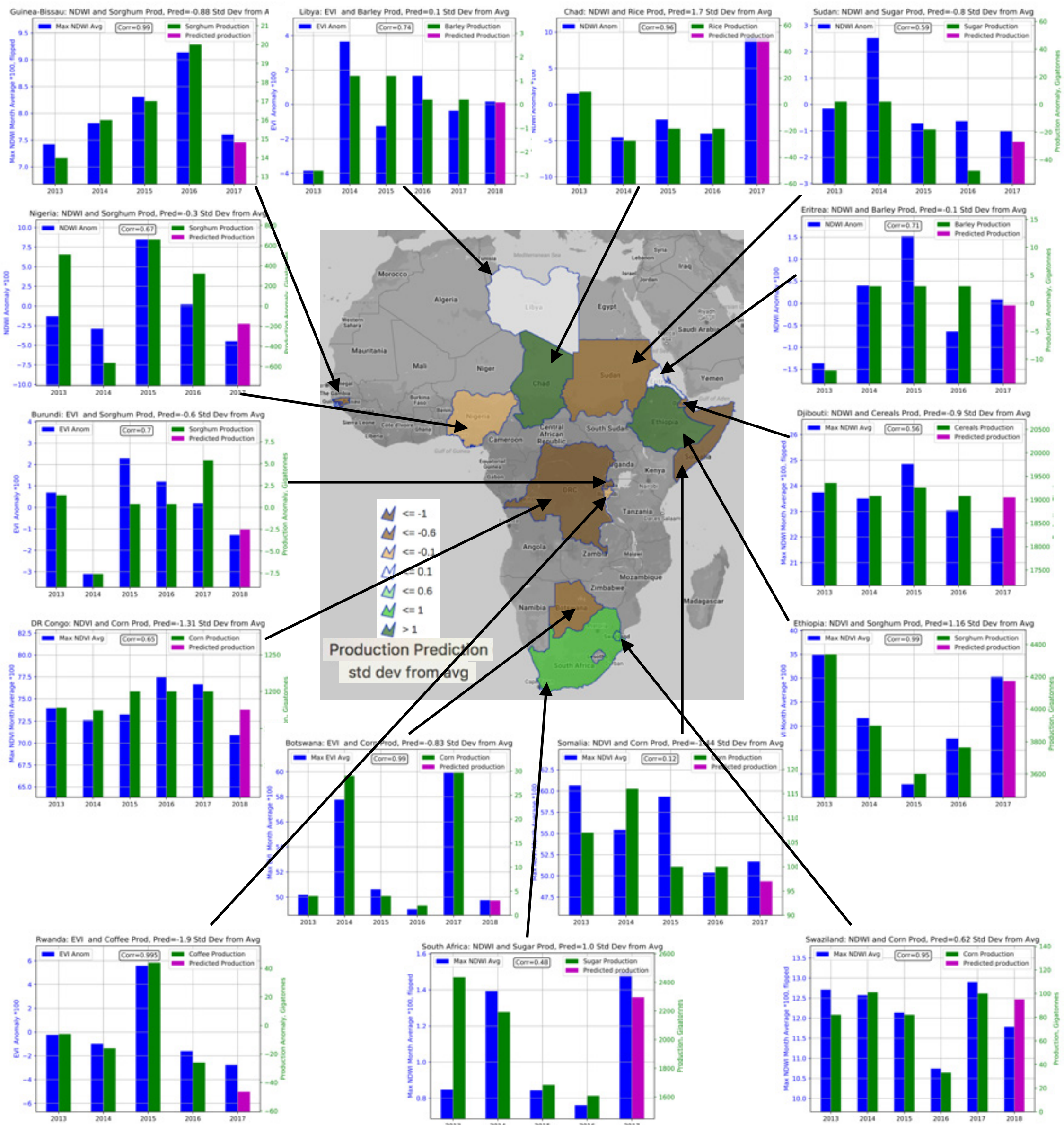


Figure 14. The map displaying the predicted production for every country currently in season in standard deviations from the average (middle). Surrounding the map are plots showing each country's highest correlation (crop and satellite index, green and blue) and predicted production (pink).

Figure 15. Farm fields by satellite in Ethiopia and Illinois at the same resolution. The small farm fields (smaller than a MODIS pixel) and poor ground truth data increase the difficulty of analyzing and predicting crop yields in Africa.

5 Conclusions

In this research, a method was developed to use three measures of crop health computed from daily MODIS satellite imagery as a predictive tool for crop yields 2–4 months before the harvest. The model was first validated in Illinois, where there is high-resolution yield data, by computing the linear fit between harvest yields in October (USDA, 2010) and the satellite indices in July and August. That is a three month prediction window, which could give farmers and insurance companies valuable information on the market months in advance. When a multivariate regression was fit to all months of the growing season and all three indices, the correlation peaked at 0.86 for 1600 data points. Next, the method was applied to three countries in Africa (Ethiopia, Tunisia, Morocco), all with different climates and crops. High correlations between maximum satellite indices and crop production were calculated in all three countries, with Ethiopia the highest at 0.99 to sorghum. After this success, satellite imagery was analyzed in every African country, and two thirds of the correlations proved to be statistically significant. Real-time crop predictions are now computed for every African country and are displayed on an interactive online map at https://lillianpetersen.github.io/africa_satellite.

Satellite imagery has been used to monitor and predict crop yields since the mid-1990s. However, most of these studies are completed in developed countries (e.g. US and Europe) because of large amounts of ground truth data and large crop fields. Therefore, the method can be tuned to specific crops and growing seasons. In Africa, it is almost impossible to tune the method because of numerous crop types, climates, and growing seasons, as well as small farms and little to no crop yield data (Figure 15). In the literature, there is no general measure of crop prediction that can be applied to any crop, location, or climate. The method developed in this research is unique because of its versatility, and has been shown to accuratly predict crop yields across an entire continent. It can be applied anywhere because it computes an overall measure of relative vegetation health compared to the mean climate on a per-pixel bases.

In Ethiopia in 2015 and 2016, there was a major drought and food shortage, and eight million people were at risk of starvation. However, the Ethiopian government did not have sufficient monitoring and reporting of drought and crop conditions during the growing season, "leading to a crucial delay in the international response." (Laing, 2016). The satellite analysis tools developed for this project can observe drought conditions as they develop and predict crop failures up to four months before the harvest and many more months before the Ethiopian government publishes the crop production data. This could give aid organizations advance notice to organize an early response to famine. Luckily, 2017 has had much higher NDVI values, indicating healthy crop conditions, and hopefully an end to the current crises.

6 Personal Statement

The most significant result of this project was creating a predictive system of crop yields for every country in Africa two to four months before the harvest. I have been told by researchers in this area that my method is unique because it may be applied for any location, crop, or climate. After seeing the predictive results of my model, several international aid organizations invited me to visit their offices. For example, I will be giving an institution-wide seminar to the International Food Policy Research Institute on May 2, which will be advertised to the larger Washington DC metro area. I hope that they will be able to use my model in order to better predict future famines and save lives through a faster response.

7 Acknowledgments

I would like to acknowledge to assistance of: Daniela Moody and Rick Chartrand, my mentors, who advised me on satellite data retrieval and data analysis methods; Mark Petersen, who helped me choose my research topic; and Phillip Wolfram for recommendations on my plots and presentation.

REFERENCES

Gao, F., Anderson, M. C., Zhang, X., Yang, Z., Alfieri, J. G., Kustas, W. P., Mueller, R., Johnson, D. M., and Prueger, J. H. (2017). “Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery.” *Remote Sensing of Environment*, 188, 9–25.

Gissila, T., Black, E., Grimes, D. I. F., and Slingo, J. M. (2004). “Seasonal forecasting of the Ethiopian summer rains.” *International Journal of Climatology*, 24(11), 1345–1358.

GraphPadSoftware (2018). “P value calculator, <https://www.graphpad.com/quickcalcs/pvalue1.cfm>.

Hamer, H., Picanso, R., Prusacki, J. J., Rater, B., Johnson, J., Barnes, K., Parsons, J., and Young, D. L. (2017). “USDA/NASS QuickStats US crop data, <https://quickstats.nass.usda.gov>.

Herring, D. and Weier, J. (2000). “Measuring Vegetation (NDVI & EVI) : Feature Articles.” *NASA*, <https://earthobservatory.nasa.gov/Features/MeasuringVegetation/> (August).

Jenner, L. (2015). “Terra Spacecraft.” *NASA* (March).

Johnson, D. M. (2016). “A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products.” *International Journal of Applied Earth Observation and Geoinformation*, 52, 65–81.

Laing, A. (2016). “Ethiopia struggles with worst drought for 50 years leaving 18 million people in need of aid.” *The Telegraph*.

Mann, M. L. and Warner, J. M. (2017a). “AgSS Ethiopia wheat yield data.” *personal communication* (December).

Mann, M. L. and Warner, J. M. (2017b). "Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach." *Field Crops Research*, 201, 60–74.

McKinnon, K. (2016). "GHCN-D: Global Historical climatology Network daily temperatures | NCAR - Climate Data Guide." *Climate Data Guide*, <https://climatedataguide.ucar.edu/climate-data/ghcn-d-global-historical-climatology-network-daily-temperatures>.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). "An Overview of the Global Historical Climatology Network-Daily Database." *Journal of Atmospheric and Oceanic Technology*, 29(7), 897–910.

Rembold, F., Atzberger, C., Savin, I., and Rojas, O. (2013). "Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection." *Remote Sensing*, 5(4), 1704–1733.

Tadesse, T., Demisse, G. B., Zaitchik, B., and Dinku, T. (2014). "Satellite-based hybrid drought monitoring tool for prediction of vegetation condition in Eastern Africa: A case study for Ethiopia." *Water Resources Research*, 50(3), 2176–2190.

Taffesse, A. S. (2012). "Crop production in Ethiopia." *Food and Agriculture in Ethiopia Progress and Policy Challenges*, University of Pennsylvania Press, 53–83.

USDA (2010). "Field Crops Usual Planting and Harvesting Dates." *Report No. 628*, <http://usda.mannlib.cornell.edu/usda/current/planting/planting-10-29-2010.pdf> (October).

USDA (2018). "Agricultural Production Statistics by Country, <https://www.indexmundi.com/agriculture/>.