



Predicting high-magnitude, low-frequency crop losses using machine learning: an application to cereal crops in Ethiopia

Michael L. Mann¹  · James M. Warner² · Arun S. Malik³

Received: 15 October 2018 / Accepted: 26 March 2019 / Published online: 23 April 2019
© The Author(s) 2019

Abstract

Timely and accurate agricultural impact assessments for droughts are critical for designing appropriate interventions and policy. These assessments are often ad hoc, late, or spatially imprecise, with reporting at the zonal or regional level. This is problematic as we find substantial variability in losses at the village-level, which is missing when reporting at the zonal level. In this paper, we propose a new data fusion method—combining remotely sensed data with agricultural survey data—that might address these limitations. We apply the method to Ethiopia, which is regularly hit by droughts and is a substantial recipient of ad hoc imported food aid. We then utilize remotely sensed data obtained near mid-season to predict substantial crop losses of greater than or equal to 25% due to drought at the village level for five primary cereal crops. We train machine learning models to predict the likelihood of losses and explore the most influential variables. On independent samples, the models identify substantial drought loss cases with up to 81% accuracy by mid- to late-September. We believe the proposed models could be used to help monitor and predict yields for disaster response teams and policy makers, particularly with further development of the models and integration of soon-to-be available high-resolution, remotely sensed data such as the Harmonized Landsat Sentinel (HLS) data set.

1 Introduction

Projections of climate change point to increasing severity and duration of extreme weather events, including drought, in large portions of the developing world (Parry et al. 2007). These

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10584-019-02432-7>) contains supplementary material, which is available to authorized users.

✉ Michael L. Mann
mmann1123@gwu.edu

¹ Department of Geography, The George Washington University, Washington, DC, USA

² International Food Policy Research Institute, Addis Ababa, Ethiopia

³ Department of Economics, The George Washington University, Washington, DC, USA

changes are anticipated to have substantial, adverse effects on agriculture (Rosenzweig et al. 2014; Costinot et al. 2016), with regions in which rainfed, smallholder agriculture is predominant being among the most vulnerable (Evangelista et al. 2013; Lindoso et al. 2014; Ramirez-Villegas et al. 2012). Given the anticipated adverse effects of climate change, the ability to monitor and forecast agricultural outcomes in developing countries is clearly important. Forecasts of agricultural losses from drought can help communities, government agencies, and NGOs put in place measures to mitigate the effects of reduced agricultural output before food insecurity or famine strike. They can also help crop insurers anticipate the magnitudes of payouts. These benefits are especially pronounced if the forecasts can be made at a fine spatial scale and with short lead times. At a more fundamental level, monitoring of losses from droughts can be beneficial in identifying appropriate measures for adapting to changing weather patterns, in the form of changing crop varieties or management techniques, such as planting dates, among others.

Remote sensing-based efforts to characterize the extent and productivity of croplands have a long history. Substantial progress has been made in mapping cropland extent, crop types, irrigation status, cropping intensity, and productivity from remotely sensed imagery, particularly for developed countries. Initial efforts, e.g., LACIE and AgRISTARS, primarily utilized remotely sensed imagery to characterize the spatial extent and growth stage of crops, relying on models driven largely by meteorological information to predict crop yields (Idso et al. 1977; Doraiswamy et al. 2003). Subsequent efforts also exploited the well-established biophysical link between canopy spectral reflectance and net primary production (Tucker and Sellers 1986), demonstrating that satellite measurements can play a role in measuring crop yield directly. A large number of studies have documented highly explanatory empirical relationships between satellite measures of plant phenology such as Normalized Difference Vegetation Index (NDVI) and yields for a variety of crops, particularly at regional scales (Rasmussen 1992; Mann and Warner 2017; Benedetti and Rossini 1993; Funk and Budde 2009; Becker-Reshef et al. 2010a, b; Mkhabela et al. 2011). Because certain crop growth stages are critical for final yield (Butler and Huybers 2015), improved results are often seen when remotely sensed data are used to characterize crop phenology (Bolton and Friedl 2013).

The ability to monitor agricultural outcomes is as yet limited in the more complex environments common to many developing countries. As an example, Ethiopia's agriculture is characterized by smallholder farmers dependent on rain-fed crops with only one to 2% of farmers having access to irrigation (Mann and Warner 2015). Improving our ability to monitor and forecast changes in agricultural outcomes in such environments is especially important given that projections of climate change point to increasing severity and duration of droughts in many parts of the developing world (Parry et al. 2007).

The main objective of this paper is to develop scalable, machine-learning models that forecast substantial crop losses from drought in Ethiopia at the village level, relying solely on remotely sensed data from early in the growing season. Such forecasts allow potential shortfalls in agricultural output to be anticipated before the end of the growing season. The models are applied to predict substantial losses from drought for each of the five principal cereal crops in Ethiopia at the level of a sub-kebele. A sub-kebele corresponds to a village of approximately 200 households, covering an average area of 24 km². We define substantial crop losses to be losses that equal or exceed a sub-kebele average of 25%, as reported by farmers. A high-damage categorical variable, or trigger, is needed to indicate where significant damage is likely to be occurring and where further on-the-ground assessment and intervention is likely needed. We discuss the choice of a 25% loss cutoff in Sect. 3.2 below.

Our models allow losses to be forecast by the date of peak greenness in any sub-kebele. The dates vary by location, therefore influencing the timing of when predictions can be made. Looking at the data from 2010 to 2015, 67% of sub-kebele losses for the main meher growing season can be estimated by September 6th and 93% by September 30th.

The data employed to generate predictions of crop loss come from three sources: (1) precipitation data are from the Climate Hazards Group Infrared Precipitation Station (CHIRPS) database, (2) data on hydrological availability of water and available energy for plant growth are from FEWS NET (2017), and (3) NDVI data are derived from two 16-day MODIS products (MOD13Q1, MYD13Q1) from the Aqua and Terra satellites (Didan and Huete 2006). The models we develop provide a framework for utilizing even higher resolution vegetative index imagery, such as the forthcoming Harmonized Landsat Sentinel data, at a national scale. We also augment a suite of algorithms used to extract, summarize, and organize remotely sensed data and prepare it for spatiotemporal analysis (Mann et al. 2018).

The machine learning models we employ are of the random forest variety. They are trained using data on farmer-reported average crop losses at the sub-kebele level for the years 2010 to 2015 collected by Ethiopia's Central Statistical Agency (CSA). The Agricultural Sample Survey (AgSS) captures data from over 45,000 households, making it one of the largest annual household agricultural surveys in the world. We focus on the five principal cereal crops grown in Ethiopia: maize, wheat, sorghum, barley, and teff.

The AgSS data together with the remotely sensed data from the weather stations and satellites mentioned above provide rich information, permitting the coupling of observations of agricultural outcomes with dense observations of variables such as plant greenness and the timing of rains. This data fusion allows us to create models that, once trained, can predict crop losses using only remotely sensed observations from satellites and weather stations.

2 Background

2.1 The 2015 drought and other localized losses

Motivated by a long history of monsoon failures across the Sahel, forecasters and aid teams anxiously watched eastern Africa at the onset of a significant El Niño event in late 2014. Below-average rainfalls between 2015 and 2017, particularly across the low-lying pastoral communities of southeastern Ethiopia, triggered an international effort to provide food aid to roughly 8.5 million people in the summer of 2017 (ReliefWeb 2017; UNOCHA 2017). The understated response from the Ethiopian government had led to accusations, as early as 2016, that the government was downplaying the severity of the drought (Schemm 2016; Schemm 2017). However, there are many indications that the drought was less severe than expected and that it primarily affected areas of lesser significance for national grain production.

Looking at the deviations of rainfall for 2015 compared to longer term patterns (Fig. A1, Appendix), we can see that lower than expected rains began in June, peaked in July, and subsided by September. The largest shortfalls are centered on the northern and eastern areas of the country with much of it in lower-lying pastoral areas near Mekele and near Dire Dawa or in high elevation upland areas near Debre Birhan. The timing of these shortfalls is critical, because planting for the major meher growing season typically starts in May and continues into late June, depending on crop type. Setbacks early in the season could undermine the germination of seed and thus potentially lead to failed crops. Early crop failure is often

followed by farmers attempting to plant drought-resistant or shorter-duration crops that reach peak growth later in the season.

Despite the reality of a severe and crippling drought in food-insecure, low-lying areas, there was scant evidence of food shortages nationally. A recent study, for example, observed no significant change in grain prices in the country as a whole (Bachewe et al. 2016). The lack of a price response is consistent with food insecure areas being hit hard, with the most productive highland areas being largely unaffected or experiencing only minor losses. An early rapid assessment performed by Tufts University researchers indicated that losses of production would exceed 18% nationally (Tufts 2016). However, this was contradicted by the arguably more accurate measure of a 2% loss, estimated from AgSS crop cuts, released later in 2016 (AgSS 2016).

Nevertheless, there was a significant up-tick in farmer-reported drought losses in 2015. Looking at the top panel of Fig. 1, the average reported crop loss due to drought increased fivefold in Tigray and fourfold in Amhara, primarily in their northern and eastern areas (as seen in Fig. 2 below).

The diversity and composition of crop types are shown in the graphs of percentage of area planted by crop in the bottom panel of Fig. 1. Interestingly, likely in response to the drought, Tigray had a sudden up-tick in 2015 in the planting of sorghum, a naturally drought-resistant plant.

Looking at the spatial patterns of losses in Tigray compared to the national average in Fig. 2, we can see two general patterns: a clear clustering of losses on the eastern border of the province and scattered, but substantial, losses over the eastern zone. The latter pattern of losses, displaying substantial intra-zonal variability, presents a considerable challenge for current early-warning systems, which aggregate above-average and below-average losses at the zonal level. Looking at the zonal means, which are reported in black in Fig. 2, we can see that Western Tigray, the western-most zone, has pockets of well-above-average losses, but these are averaged out by a larger number of areas with below-average losses.

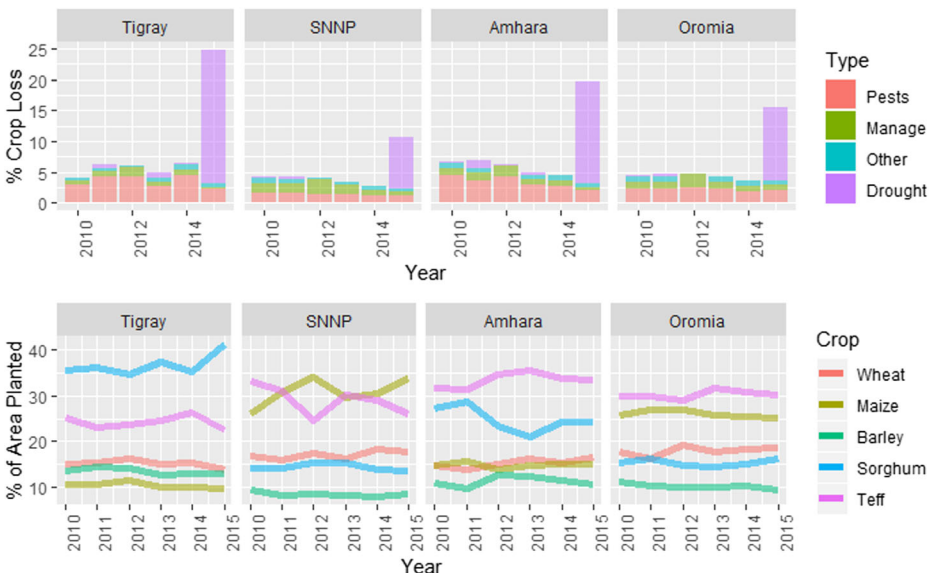


Fig. 1 Percent planted area and losses by crop and region

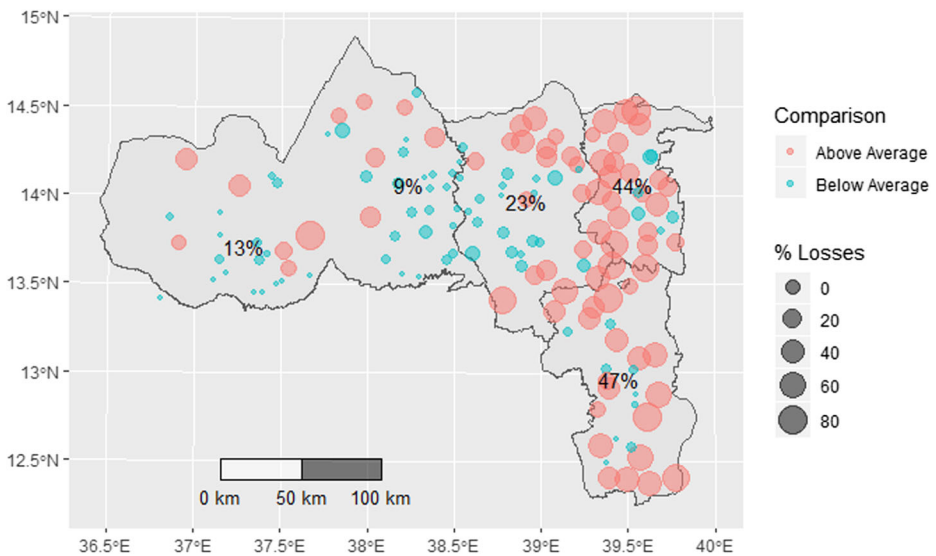


Fig. 2 Mean percentage losses in 2015 reported for wheat, maize, and teff in Tigray region relative to the national average

2.2 Motivation

This paper attempts to develop a replicable methodology for identifying multiple types of crop losses by mid-season at a high level of spatial resolution. This acuity will provide better insights into the effects of droughts at multiple scales: local, zonal, and regional. We believe these tools can be used to better design and target local, national, and international responses and can help guide spatial allocation of resources for more in-depth studies such as rapid assessments.

3 Methods

3.1 Data sources

3.1.1 Survey data—agricultural sample survey data (2010–2015)

Data on crop losses for the years 2010 to 2015 was obtained from Ethiopia’s Central Statistical Agency’s Agricultural Sample Survey and was chosen for its annual collection, spatial coverage, and unique sampling structure.

For this study, we construct a sub-kebele-level panel data set for the meher growing season for five crops for the years 2010 to 2015.¹ We were able to construct the panel data set because approximately 75% of all the sub-kebeles were sampled by CSA in each of the six meher crop

¹ The two main crop seasons in Ethiopia are the meher and belg seasons. The meher season is the main season and produces 90–95% of the nation’s total cereals output (AgSS 2016).

seasons.² This effectively comprises a base of five relatively favorable crop seasons, from 2010 to 2014, followed by a drought year, 2015.

For each household surveyed and for each crop, CSA collects farmer reports on the percentage of total crop production lost due to a range of causes, including drought, management error, pests, and weather (e.g., winds or flooding). We focus on crop losses due to drought because such losses can be most readily assessed from satellite imagery by characterizing crop phenology.

We study the five principal Ethiopian field crops (maize, wheat, sorghum, barley, teff), because of their importance in value terms and their geographically widespread adoption. Accordingly, our estimates of crop losses are for these five crops alone and not for total agricultural output. We have no reason to believe that our estimates for these five crops are not generally representative of all crop production in terms of the effects of droughts, but a more detailed analysis would be needed to substantiate this presumption.

For each crop, data on (drought-related) crop losses was aggregated to the sub-kebele level by averaging losses across the 20 households in a sub-kebele. This yielded approximately 1780 observations at the sub-kebele level for each year and a total of 10,680 observations for all 6 years. The number of observations varies by crop since not all sub-kebeles grow all five crops. Data for each crop year was checked for consistency and cleaned. In the course of data cleaning, we dropped any sub-kebeles for which crop loss data was not available for at least 4 years.

3.1.2 Remotely sensed data—greenness, precipitation, and evaporation

Normalized Difference Vegetation Index NDVI is sensitive to the amount of chlorophyll in any given pixel and is commonly used to estimate plant productivity and health in agricultural applications. Given the relatively small scale of agriculture in Ethiopia and the lack of high frequency observations from other satellites (such as HLS), we derive the NDVI using the 250 m vegetation products from the MODIS satellites, specifically the two 16-day MODIS products (MOD13Q1, MYD13Q1) from the Aqua and Terra satellites (Didan and Huete 2006). We find that the combination of these two products provides a stable and informative time series. Critical metrics like the timing of planting dates and peak greenness are calculated as follows. The “planting date” is estimated as the date of the global minimum of NDVI values during the planting season, which is estimated as ranging from March to May. Ties in minima were broken by storing the later of the two dates. The date of peak greenness was calculated as the maximum value of the meher growing season (March to December), where ties were broken by storing the earlier date.

Water availability variables Data on precipitation (PPT) was obtained from the Climate Hazards Group Infrared Precipitation Station (CHIRPS) database. Data is collected as precipitation by dekad (Funk et al. 2015). There are three dekads in a month, the first two being 10 days long and the third being the remaining days in the month. Previous research has shown hydrological availability of water and available energy for plant growth to be important factors in crop growth. Accordingly, we include monthly estimates of potential evaporation (PET) and actual evapotranspiration anomaly (ETA).

² For each sub-kebele, CSA samples households with replacement. Therefore, the households surveyed in a given sub-kebele can vary from 1 year to the next. Thus, our data set does not constitute a panel at the household level, though it does at the sub-kebele level.

Actual evapotranspiration (AET) is the sum of transpired water, water moved through plants, plus evaporation from soils and water surfaces. Effectively, AET is the total volume of water, supplied by rainfall, which moves through a system. The amount of AET is driven by two forces: (1) the supply of water through rainfall and (2) the demand for water, driven primarily by energy from the sun, which allows for evaporation and movement of water through plants but is also affected by wind, vapor pressure, and soil properties. AET is correlated with vascular plant productivity, biomass accumulation, and regrowth (Major 1967; Mann et al. 2016). High values of AET tend to be found in highly productive ecosystems. We utilize actual evapotranspiration anomaly (ETA), which compares current AET values to the historical norm. As such, AET can be used as a proxy for drought, as low current values of AET would correspond to reductions in water availability.

We also include estimates of PET, which is a measure of the ability of the atmosphere to remove water through the processes of evaporation and transpiration, assuming unlimited water availability. PET is largely a measure of energy in the form of heat from the sun, as well as the effects of soils, e.g., sandy vs loamy, and wind. High PET values can be thought of largely as a proxy for high temperatures and soil types that allow for rapid evaporation and transpiration.

3.1.3 Variable definitions

Tables 1 and 2 provide a key to the variables obtained and derived from the AgSS, GIS, and remotely sensed data.

3.2 Estimating crop losses

Data Because our unit of analysis is the sub-kebele, remotely sensed data has to be spatially aggregated to match the sub-kebeles in the country. We calculate the mean of the raw remotely sensed data values for all available agricultural pixels at the sub-kebele-level and further extract statistics of interest from the resulting 8-day or dekadal time series.

Table 1 Ethiopian Agricultural Sample Survey variable names and descriptions

Name	Description
Prefixes	
MAIZE_	Variables based on maize plots
WHEAT_	Variables based on wheat plots
BARLEY_	Variables based on barley plots
SORGHUM_	Variables based on sorghum plots
TEFF_	Variables based on teff plots
Suffixes	
_diff	First difference or change in value from one period to the next ($X_t - X_{t-1}$)
_diff_md	Difference from the median of all observations
AgSS variables	
DAMAGE_DROUGHT_P	Percentage of crop production reported lost due to drought
Other geographic variables	
elevation	Elevation (m)
z_code	Zonal codes

Table 2 Remotely sensed variable names and descriptions

Name	Description
Prefixes	
G_	Calculated from meher growing season values
ETA_	Actual evapotranspiration anomaly
PET_	Potential evapotranspiration
PPT_	Precipitation
Suffixes	
_diff	First difference or change in value from one period to the next ($X_t - X_{t-1}$)
_diff_md	Difference from the median value for all observations
NDVI basic summary statistics	
G_max	Meher maximum value of NDVI
G_Qnt	Meher 90th percentile value of NDVI
G_plant_date	Estimated date of meher planting based on NDVI phenology
G_mx_dates	Estimated date of maximum NDVI during meher season
Integrated summary statistics	
G (PET,ETA,PPT)_AUC_leading	Area under the curve of NDVI (or PET, ETA, PPT) for the ascending part of the NDVI curve during the meher season
Comparison to norms	
G (PET,ETA,PPT)_diff_md	Difference between the median meher NDVI (or ETA, PET, PPT) and observed value
G_mx_dates_diff_md	Difference between median date of maximum NDVI and observation date

Summarizing temporal data One of the primary challenges in utilizing high-frequency time series data, such as 8-day NDVI, to estimate low-frequency events (seasonal crop losses from drought) is reconciling this temporal mismatch when formalizing the relationship between sources of data. Low-frequency properties must be extracted from the high-frequency time series. In our setting, the low-frequency properties extracted must be relevant to identifying and characterizing the aspects of plant phenology that affect crop yields until the midpoint of the growing season as determined by the time of peak NDVI (greenness). We start by capturing the relevant phenological features of each crop through 41 metrics summarizing 8-day NDVI data as well as 10-day PPT, PET, and AET. These measures span the portion of the meher growing season from the estimated planting date to the harvest date. The mean planting date is estimated at the sub-kebele level so as to capture any differences in timing due to elevation, crop choice, or other management considerations. It is worth emphasizing that given our interest in forecasting losses early in the growing season, we limit the variables in our models to those that rely only on data up to the midpoint of the growing season.

We utilized phenological timing observed from NDVI to summarize PPT, PET, and AET for the relevant portion of the growing season, based on estimated planting date and the date of the maximum observed NDVI value. For instance, the variable that includes “PPT_leading” in its name is calculated using the PPT time series only for the period between the estimated planting date and the date of the maximum observed NDVI value.

We estimate three classes of statistics: summary statistics, e.g., mean, maximum, variance; integrated summary statistics, e.g., area under the curve for the leading half of the growing season; and deviation-from-the-norm statistics, e.g., deviations of a given statistic from its 90th

historical percentile. Deviations from the norm provide a comparison between expected values and the current year and, thus, should provide insight into low-frequency events like drought.

For the dependent variable, we choose drought losses of 25% or more as a cutoff value given the decreasing number of observations at higher loss levels. Table A1 (online Appendix) shows that moving the cutoff above 25% significantly reduces the number of observations available for training for wheat. Additionally, research from Ethiopia indicates that for smallholder farmers, a 40% crop loss initiated a downward spiral of food insecurity and distress sales that took years to recover from (Cater et al. 2007). Combining this information with our actual observed reported crop losses that clustered between 20 and 33%, we considered 25% to be a realistic value for potentially serious impact on farmer welfare.

3.3 Predicting crop losses from drought

As noted, our objective is to determine if crop losses due to drought at the sub-kebele level can be predicted at mid-season using remotely sensed data alone. To do this we make use of machine-learning methods in the form of random forests. We define our dependent variable to be a binary variable that reflects whether or not losses due to drought for a given crop were 25% or higher. Given the binary nature of our dependent variable, our prediction problem is one of classification: Can we predict whether crop losses due to drought will be 25% or higher for a given sub-kebele? Our model takes the following form:

$$Y_{ci} = f \left(\begin{array}{l} \alpha, \alpha_diff, \alpha_diff_md, \alpha_auc_leading, \rho_auc_leading, \\ \sigma_dates, \sigma_dates_diff_md, elevation, zone_code \end{array} \right)$$

where

$$Y_{ci} = \begin{cases} 1, & \text{if drought loss} \geq 0.25 \\ 0, & \text{otherwise} \end{cases}$$

Y_{ci} is a binary variable that captures whether losses for crop c due to drought in sub-kebele i are greater than or equal to 25%. The model is estimated using a non-linear function of α which includes two measures of NDVI, G_mx and G_qnt ; ρ includes the sum of G , PET , ETA , and PPT up until peak greenness; and σ includes the dates³ of initial greenup (planting date) and of peak NDVI and their difference from the median date; also included are elevation and zonal codes.

Please refer to the [Appendix](#) for a full description of random forests, training and cross-validation, and methods for handling spatial autocorrelation.

4 Results and discussion

4.1 Predicting drought losses

In this section, we present the out-of-sample prediction results from our machine learning models. For each crop, the models predict whether a sub-kebele experienced crop losses of

³ Datetime variables are measured in days since some arbitrary origin date, in this case, days since January 1, 1960.

25% or more due to drought in a given year, with the predictions based solely on remotely sensed vegetation and weather data obtained by mid-season.

For the sake of brevity, from hereon we will refer to crop losses of 25% or more due to drought as “substantial losses.” We convey the accuracy of our predictions by presenting confusion matrices as well as the values of the three performance metrics described in the online [Appendix](#) (Sect. 3.3) for out-of-sample predictions.

4.1.1 Maize losses

The confusion matrix and the values of the three performance metrics for our out-of-sample predictions of maize losses are presented in Table 3.

Encouragingly, we see a high level of accuracy for our out-of-sample predictions, with an overall accuracy of 96%. We see good identification of reported substantial losses with a recall rate of 0.81, indicating that 81% of all substantial loss cases were correctly predicted or 57 out of 70 such cases. Thirty-six cases were falsely predicted as substantial loss cases. Bear in mind that recall is more important than overall accuracy. Although we were able to obtain a 96% accuracy rate, this overstates model performance because we could simply classify all observations as “healthy” and obtain a 95+ percent accuracy rate.

Clearly, further work could help better identify low-probability, substantial loss events. For maize, our model would have predicted substantially more loss events than actually occurred. Any early warning system, including this one, would likely have to be linked to substantive on-the-ground verification services. Warnings could be provided to local administrative services such as agricultural extension agents and humanitarian organizations that could help verify and address any losses. However, the importance of this research should not be understated. Early identification of localized crop losses could be very useful for policy action. Moreover, adoption of the more accurate 30 m HLS satellite data could markedly improve results using our methodology.

We can also examine the relative importance of each variable to the model, as well as its estimated non-linear relationship with the dependent variable through partial dependence plots. We first examine the relative importance of each variable in the random forest through the role it played in Gini coefficient loss.

From Fig. A2, Appendix, we can see that the most important variables are primarily date variables. A large weight is placed on date of maximum greenness (G_{mx}), as measured by NDVI, the date of initial greenup ($plant_dates$), and their differences from their median values ($G_{mx_dates_diff_md}$ and $plant_dates_diff_md$). The importance of these variables can be explained as follows. If the July rains fail, farmers are forced to adjust their planting dates, which in turn delays when maximum greenness is reached.

Table 3 Out-of-sample confusion matrix and performance metrics for reported losses $\geq 5\%$ for maize

		Reference			
		False	True		
Predicted	False	1182	13		
	True	36	57		
	Value	Accuracy	Kappa	Recall	
		0.96	0.68	0.81	

True = crop losses $\geq 25\%$ for maize

Other variables of interest include the sum of precipitation from the planting date up until the date of maximal greenness (PPT_G_AUC_leading). Precipitation during this early period in the growing season has clear implications for successful germination. Moreover, this variable likely provides a good indication, after controlling for the date of initial greenup, of whether or not the planting was successful. G_Qnt_diff, the first-difference in 90th percentile greenness, and the difference between G_Qnt and its median value are also of importance. Note that “_diff” indicates a first-difference or the difference between the current year’s value of a variable and the previous year’s value. Drops in near maximal greenness likely provide some evidence that plant health suffered throughout the growing season. Each one of these variables will have a specific non-linear response, with curves being estimated by actual plant and weather patterns observed remotely. Moreover, the emergent properties of some farmer management practices, such as delayed planting, can also be examined.

We now examine partial dependence plots⁴ to help understand the response of reported maize losses to variables of importance. The first panel of Fig. 3 shows the plot for G_mx_dates_diff_md, which is the difference between the date of the current year’s maximum NDVI value and the median value of the date. As mentioned above, this variable would strongly track the date of successful planting and germination, as well as plant vigor following successful germination. Note that higher values on the vertical axis indicate a greater likelihood of substantial crop losses.

Figure 3 represents the change in the percent likelihood of a substantial loss on the *y*-axis, where more negative values indicate a lower probability of loss. The *x*-axis represents the full range of variable observations, where the observation is indicated by the plot title in gray. The number of observations at each segment of partial dependence plot is indicated by line width and hue, where darker and thicker portions of the curve should be given greater weight in interpretation.

The likelihood of reporting substantial wheat losses decreases slowly as G_mx_dates_diff_md approaches zero and increases dramatically after the value of 1. Looking at the number of observations, indicated by a darker color and thicker line, we can surmise that the erratic step-like function on the left is due to the NDVI observations being collected every 8 days. Generally, sub-kebeles with earlier than normal maximum greenness dates have mixed success, likely depending on geography and weather, with some being successful and others not. According to regional experts, having been given advanced warning of potential failed rains, the Ethiopian government encouraged farmers to plant earlier than usual. From the curve above, we see indications of a mixed outcome, with those planting the earliest seeing a relatively high likelihood of loss and those planting a few weeks early seeing improved but mixed outcomes. The lowest likelihood of losses was reached after the median plant date. This indicates that many areas of the country experienced normal rains and successfully kept to normal or slightly later-than-average planting dates. On the other hand, those sub-kebeles with much later than expected maximum greenness dates experienced the greatest losses. As discussed earlier, late maximum dates point to areas with one or more failed plantings. In these areas, farmers would have likely tried to replant later in the season.

In the second panel, PPT_G_AUC_leading tracks the effects of total precipitation from initial greenup until the maximal NDVI value. Note that this variable is automatically centered around its mean. We can see from the figure that the likelihood of crop losses declines dramatically up until its centered mean value, 0. Values to the left of 0 clearly correspond to

⁴ These plots are graphical visualizations of the marginal effect of a given variable (or multiple variables) on an outcome. Higher values on the *y*-axis indicate a higher likelihood of loss.

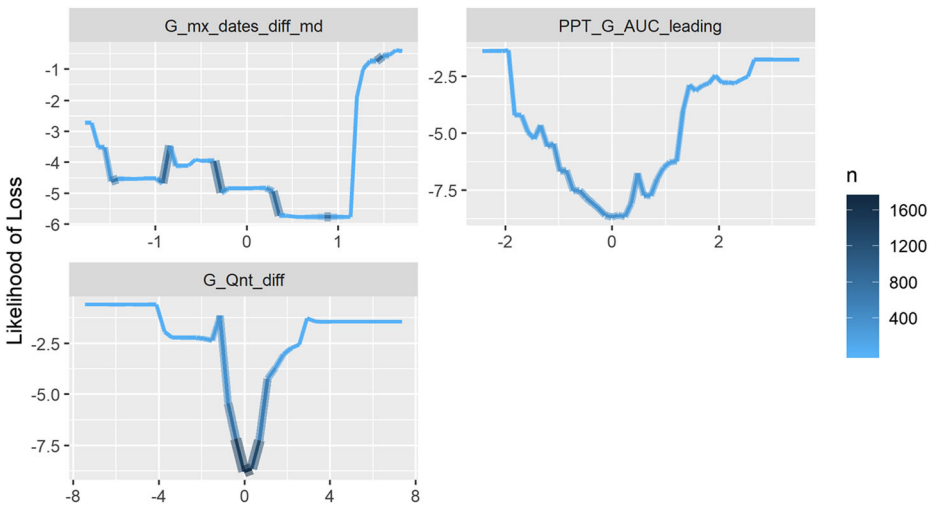


Fig. 3 Partial dependence plot for influence of $G_mx_dates_diff_md$, $PPT_G_AUC_leading$, and G_Qnt_diff on maize loss prediction

below average or drought conditions. Additionally, years with above average levels of precipitation also correspond to increased losses due to excess precipitation or damaging storms, although this is likely the result of only having a few observations. The fact that above average precipitation would be considered “drought losses” likely relates to the fact that farmer response to questionnaires is imperfect. Farmers might well conflate losses related to drought with losses related to weather or rain more generally.

The last panel shows how the first difference in G_Qnt (G_Qnt_diff), the 90th percentile value of NDVI, influences substantial losses of maize crops. The highest likelihood of losses is observed with large declines in maximal values of NDVI relative to the previous year. This change likely reflects declines in plant health, due to water stress or other climatic stressors such as high heat. Positive changes in maximal values might reflect poor management, e.g., failure to control weeds, narrow row spacing due to broadcast seeding, or excess rain.

4.1.2 Wheat losses

We now turn from maize to wheat. Table 4 presents the confusion matrix for our out-of-sample predictions of wheat losses, along with our three performance metrics.

Overall, we find a high level of accuracy for our out-of-sample predictions, with overall accuracy of 97%. Compared to maize, we obtain somewhat weaker prediction of substantial wheat losses, with a recall rate of 0.65 instead of 0.81, i.e., 65% of all substantial wheat loss cases, or 35 out of 54 such cases, were correctly predicted. Eight cases were falsely predicted as substantial crop losses.

Figure A3, Appendix, shows the 10 most important variables in our random forests, as measured by the role each variable played in Gini coefficient loss. The set of variables is similar to that for maize. However, the absolute and relative importance of the variables differs. This suggests that our models are able to utilize the same broad set of phenological and weather variables to capture differences between crop types and the factors that influence

Table 4 Out-of-sample confusion matrix and performance metrics for reported losses $\geq 25\%$ for wheat

		Reference			
		False	True		
Predicted	False	832	19	Accuracy 0.97	Kappa 0.71
	True	8	35		
Value					

True = crop losses \geq or equal to 25% for wheat

losses for each crop. The variables *G_Qnt_diff* and *G_mx_dates* are once again important, suggesting a robustness in their predictive ability

We can see that *G_Qnt* has the highest relative importance. *G_Qnt* is the 90th percentile of NDVI values in the meher season. High NDVI values reflect a measure of plant health, with higher *G_Qnt* values corresponding to better plant health (such as larger leaf area and lower water stress) at some point in the early growing season. We can also see that many of the most important variables are variables summarizing NDVI, with the exception of the dates of these phenological indicators. Again, the model has identified and emphasized variables that help distinguish losses related to wheat plantings.

Interestingly, comparing Figs. A2 and A3 Appendix, we can see that planting and maximal NDVI dates are of less importance for predicting substantial wheat losses than maize losses. This might relate to the fact that many of the most productive wheat-growing areas of Ethiopia were largely unaffected by the failure of the July rains during the 2015 drought. Instead, these areas were more affected by declines in rainfall in August. As a result, many wheat growing sub-kebeles did not need to change their planting dates. Instead, plant health was more affected by water stress after a successful germination.

We now turn to the partial dependence plots for the prediction of substantial losses in wheat and start by examining, in Fig. 4, the plot for *G_Qnt*, the 90th percentile value of NDVI.

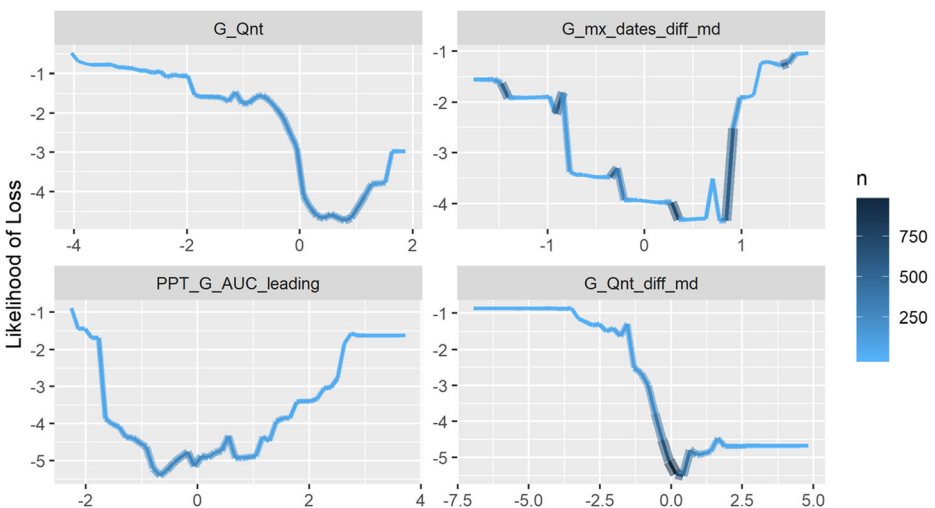


Fig. 4 Partial dependence plot for influence of *G_Qnt*, *G_mx_dates_diff_md*, and *PPT_G_AUC_leading* on wheat loss prediction

In the first panel of Fig. 4, the likelihood of substantial maize losses decreases rapidly as maximal NDVI values (G_Qnt) approach the mean, with a minimum around 0.5. The likelihood of losses increases again at the upper end of the range of values for G_Qnt . These losses at the upper end might reflect poor management, such as failure to control weeds or narrow row spacing. The consistency across the two different crop models (maize and wheat) is an indication that these models are picking up meaningful nonlinear relationships between plant phenology and the likelihood of loss.

The second panel shows the partial dependence plot for $G_mx_dates_diff_md$, the difference between the date of maximum NDVI in the current year and the median date. Comparing this panel with the corresponding panel in Fig. 3 reveals similarities and differences between the likelihood of losses for wheat and maize. We see that the likelihood of losses at very negative levels of $G_mx_dates_diff_md$ is higher for wheat (note the differences in the y-axis intercept values) and declines more rapidly as the change in maximal greenness date increases. However, for both crops, the curve reaches a minimum above 0. Losses increase again with much later than average planting and therefore later maximal NDVI values.

The third panel of Fig. 4 has the partial dependence plot for $PPT_G_AUC_leading$. This variable captures total rainfall until maximum greenness. As for maize (Fig. 3), we see a U-shaped response, with the highest drought losses observed with below average (centered around 0) precipitation. Higher than average precipitation is also linked to reported drought losses. Again, this may be a reflection of confusion in survey responses, or it might point to the fact that the timing of rains, rather than the quantity of rainfall, is the key driver of substantial losses. The importance of this timing is reflected across models by the $plant_date$ and G_mx_date variables.

Although the variables we have focused on thus far have a common U-shaped partial dependence plot, there are other possible shapes. For instance, the difference between the current 90th percentile value and the median 90th percentile of previous years, $G_Qnt_diff_md$ in the fourth panel, shows a strong downward trend. The likelihood of significant loss is extremely high with 90th percentile greenness values significantly lower than normal. The likelihood of loss drops rapidly as 90th percentile greenness approaches 0 and tapers up slightly thereafter.

4.1.3 Teff, barley, and sorghum losses

The results of the model for the last three major crops are more mixed (Table 5). While the models accurately identify 75% substantial losses for sorghum, this decreases to 58% and 57% for barley and teff respectively. Given the importance of teff in Ethiopian diet, this is disappointing but not surprising. This model seems to perform best for crops planted earlier in the season (February for maize and sorghum) with a mean recall of 84.25%, compared to those

Table 5 Out-of-sample performance metrics for reported losses $\geq 25\%$ for sorghum, barley, and teff

	Accuracy	Kappa	Recall
Sorghum	0.97	0.81	0.75
Barley	0.94	0.62	0.58
Teff	0.95	0.6	0.57

Please refer to the [Appendix](#) available online for more details on these three crops

planted late in the season (May for barley, teff, and wheat) at 60%. This implies that for barley, teff, and wheat, some salient temporal features were missing in our analysis. Extensive efforts are underway by the lead author to improve time series feature extraction for remotely sensed data.⁵

5 Conclusions

Given the anticipated increase in the frequency and severity of droughts in many parts of the developing world, there is obvious benefit to having methods for forecasting agricultural losses due to droughts in a timely manner and at a fine spatial scale. In this study, we examine the feasibility of using remotely sensed data and machine learning models to predict agricultural losses due to drought in Ethiopia. Specifically, we predict, by the midpoint of the growing season, which sub-kebeles will suffer substantial crop losses at the time of harvest. We develop a custom set of algorithms to summarize changes in plant phenology, as measured by NDVI, precipitation, and potential and actual evapotranspiration up until the date of maximal greenness, which corresponds approximately to the middle of the growing season. The algorithms to generate these statistics have all been made freely available through Github.⁶

Considering the difficulty of the task and the nature of the available data, our results are very encouraging. At the sub-kebele level, the out-of-sample recall rate, or the ability to correctly predict substantial crop losses out-of-sample, ranged from 57 to 81%, with a median value of 65%. The overall accuracy of the predictions ranged from 94 to 97%.

Our models allow crop losses to be predicted sooner than can be done with existing models and at a higher level of spatial resolution. Specifically, we can predict losses by crop in any sub-kebele by the date of peak greenness. The date varies by location, therefore influencing precisely when predictions can be made. Looking at data between 2010 and 2015, we can predict 67% of sub-kebele crop losses by September 6th and 93% by September 30th.

We also provide evidence that our models pick up meaningful nonlinear relationships that respond to both crop loss and differences in crop types. With more development, models like these could be used to provide accurate early-season assessments of agricultural losses in many settings. Moreover, these predictions made in the first half of the growing season provide a critical opportunity for identifying the need for, and subsequent planning for, early national and local interventions. We believe efforts such as these can augment and enhance current early warning systems and help guide field-based rapid assessments.

Though the predictive accuracy of our models is fairly high, we believe that the predictions of these models could be significantly improved with the use of 30 m Harmonized Landsat Sentinel (HLS) data. This data was planned to be released in late 2018, but its timing now appears unclear.⁷ Once released, HLS data will have 5-day coverage. This likely would provide a high enough frequency to avoid clouds and provide accurate estimates of NDVI curves at a much higher resolution than the 250 m cells used in this research. The methods developed in this study have been specifically designed to transfer to the new HLS data stream, allowing for likely much-improved model accuracy with minimal further development time.

⁵ <https://github.com/mmann1123/ts-raster> “ts-raster” is a Python package under development by the authors for analyzing time series characteristics from raster data. It allows feature extraction, dimension reduction, and applications of machine learning techniques for geospatial data.

⁶ https://github.com/mmann1123/IFPRI_Ethiopia_Drought_2016_public/

⁷ HLS data is currently available for North America, large portions of Europe, South Africa, Tanzania, part of South East Asia, and Eastern Australia.

Acknowledgments We are grateful to the Ethiopian Central Statistical Agency for use of the Agricultural Sample Survey data.

Funding This research was conducted under the Ethiopia Strategy Support Program, which is managed by the International Food Policy Research Institute and is financially supported by the United States Agency for International Development, the Department for International Development of the Government of the UK, and the European Union.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- AgSS (2016) Agricultural sample survey 2016/2017 (2009 E.C.) report on area and production of major crops. Feed the Future. THE FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA CENTRAL STATISTICAL AGENCY
- Bachewe, Fantu, Feiruz Yimer, and Bart Minten. 2016. Agricultural prices during drought in Ethiopia: an updated assessment using national producer data (January 2014 to June 2016).
- Becker-Reshef I, Justice C, Sullivan M, Vermote E, Tucker C, Anyamba A, Small J et al (2010a) Monitoring global croplands with coarse resolution earth observations: the Global Agriculture Monitoring (glam) project. *Remote Sens* 2(6). Molecular Diversity Preservation International:1589–1609
- Becker-Reshef, Inbal, Eric Vermote, Mark Lindeman, and Christopher Justice. 2010b. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using Modis data. *Remote Sens Environ* 114 (6). Elsevier: 1312–1323
- Benedetti, Roberto, and Paolo Rossini. 1993. On the use of Ndvi profiles as a tool for agricultural statistics: the case study of wheat yield estimate and forecast in Emilia Romagna. *Remote Sens Environ* 45 (3). Elsevier: 311–326
- Bolton, Douglas K, and Mark A Friedl. 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric For Meteorol* 173. Elsevier: 74–84
- Butler EE, Huybers P (2015) Variations in the sensitivity of us maize yield to extreme temperatures by region and growth phase. *Environ Res Lett* 10(3). IOP Publishing):034009
- Costinot, Arnaud, Dave Donaldson, and Cory Smith. 2016. Evolving comparative advantage and the impact of climate change in agricultural markets: evidence from 1.7 million fields around the world. *J Polit Econ* 124 (1). University of Chicago Press Chicago: 205–248
- Didan K, Huete A (2006) MODIS vegetation index product series collection 5 change summary. TBRS Lab, the University of Arizona
- Doraiswamy PC, Moulin S, Cook PW, Stern A (2003) Crop yield assessment from remote sensing. *Photogramm Eng Remote Sens* 69(6). American Society for Photogrammetry; Remote Sensing):665–674
- Evangelista, Paul, Nicholas Young, and Jonathan Burnett. 2013. How will climate change spatially affect agriculture production in Ethiopia? Case studies of important cereal crops. *Clim Chang* 119 (3–4). Springer: 855–873
- (2017) Data Portals. FEWS NET. <https://earlywarning.usgs.gov/fews/downloads/index.php>
- Funk C, Budde ME (2009) Phenologically-tuned Modis Ndvi-based production anomaly estimates for Zimbabwe. *Remote Sens Environ* 113(1). Elsevier):115–125
- Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, Husak G et al (2015) The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci Data* 2. Nature Publishing Group:150066
- Idso SB, Jackson RD, Reginato RJ (1977) Remote-sensing of crop yields. *Science* 196(4285). American Association for the Advancement of Science):19–25
- Lindoso DP, Rocha JD, Debortoli N, Parente II, Eiró F, Bursztyrn M, Rodrigues-Filho S (2014) Integrated assessment of smallholder farming's vulnerability to drought in the Brazilian semi-arid: a case study in Ceará. *Clim Chang* 127(1). Springer):93–105
- Major JACK (1967) Potential evapotranspiration and plant distribution in western states with emphasis on California. *Am Assoc Adv Sci Publ* 86
- Mann MLL, Warner J (2015) Ethiopian wheat yield and yield gap estimation: a small area integrated data approach. International Food Policy Research Institute, Addis Ababa

- Mann ML, Batllori E, Moritz MA, Waller EK, Berck P, Flint AL, Flint LE, Dolfi E (2016) Incorporating anthropogenic influences into fire probability models: effects of human activity and climate change on fire activity in California. *PLoS One* 11(4). Public Library of Science):e0153589
- Mann, ML, B Kramer, F Ceballos, M Robles, J Gray, and E Melaas. 2018. Modeling wheat yields in India: describing & exploiting spatiotemporal variability with panel regression. Under Submission
- Mkhabela MS, Bullock P, Raj S, Wang S, Yang Y (2011) Crop yield forecasting on the Canadian prairies using Modis Ndvi data. *Agric For Meteorol* 151(3). Elsevier):385–393
- Parry M, Canziani O, Palutikof J, van der Linden PJ, Hanson CE et al (2007) *Climate change 2007: impacts, adaptation and vulnerability*. Vol. 4. Cambridge University Press, Cambridge
- Ramirez-Villegas J, Salazar M, Jarvis A, Carlos E, Navarro-Racines (2012) A way forward on adaptation to climate change in Colombian agriculture: perspectives towards 2050. *Clim Chang* 115(3–4). Springer):611–628
- Rasmussen, Michael S. 1992. Assessment of millet yields and production in Northern Burkina Faso using integrated Ndvi from the Avhrr. *Int J Remote Sens* 13 (18). Taylor & Francis: 3431–3442
- ReliefWeb (2017) Ethiopia: drought—2015–2017. ReliefWeb. <https://reliefweb.int/disaster/dr-2015-000109-eth>. Accessed 10 June 2018
- Rosenzweig C, Elliott J, Deryng D, Ruane AC, Müller C, Arneth A, Boote KJ et al (2014) Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proc Natl Acad Sci* 111(9). National Acad Sciences):3268–3273
- Schemm P (2016) Ethiopia is facing a devastating drought, and food aid may soon run out. The Washington Post. WP Company. http://www.washingtonpost.com/sf/world/2016/02/22/history-repeats-itself-in-ethiopia/?utm_term=.2b7bb38f3a3e. Accessed 21 July 2018
- Schemm P (2017) Ethiopia is facing a killer drought. But it's going almost unnoticed. The Washington Post. WP Company https://www.washingtonpost.com/news/worldviews/wp/2017/05/01/ethiopia-is-facing-a-killer-drought-but-its-going-almost-unnoticed/?utm_term=.ad44d46b0631. Accessed 21 July 2018
- Tucker CJ, Sellers PJ (1986) Satellite remote sensing of primary production. *Int J Remote Sens* 7(11). Taylor & Francis):1395–1416
- Tufts (2016) FOOD security in Ethiopia in 2016: analysing crop production and market function after the main Meher agricultural season. Feed the future. USAID
- UNOCHA (2017) Mid-year review, Ethiopia humanitarian requirements document, July 2017. ReliefWeb. <https://reliefweb.int/report/ethiopia/mid-year-review-ethiopia-humanitarian-requirements-document-july-2017>. Accessed 10 June 2018

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.