DEVELOP Technical Report

Texas & Georgia Agriculture

Assessing the Drivers of Cotton Quality and Yield for Improved Crop Forecasting in the Southern United States

Spring 2025 | Maryland – Goddard
April 4th, 2025

Authors: Lorryn Andrade (Analytical Mechanics Associates), Nick Grener (Analytical Mechanics Associates), Kelechi Igwe (Analytical Mechanics Associates), Henry Osei (Analytical Mechanics Associates)

Abstract: Declines in cotton quality and yield across the United States are raising concerns in the cotton industry. NASA DEVELOP partnered with the United States Department of Agriculture (USDA) Agricultural Research Service and USDA Agricultural Marketing Service to identify environmental conditions driving cotton quality and yield in two high-production regions with distinct climate zones: southern Georgia and western Texas. This project used Earth observations from the Harmonized Landsat Sentinel-2 initiative, Global Precipitation Measurement (GPM) Integrated Multi-satellite Retrievals for GPM, and the Soil Moisture Active Passive L-band radiometer to understand how climate and soil parameters correlate with cotton quality. We analyzed the Enhanced Vegetation Index (EVI), precipitation, soil moisture, wind speed, and growing degree days from 2015 to 2024, focusing on the months of May to November. Using multiple regression models, we found that wet autumns were associated with lower brightness in cotton crops across both states, summer EVI readings had moderate-to-strong associations with some quality variables in Georgia, and August soil moisture was an important driver of multiple quality variables in Texas. Models that used Earth observation data generally performed better than models with only data from weather stations, with adjusted R-squared values ranging from 31.6 – 72.5%. A major limitation of the study, however, was the uncertainty about the exact location of the cotton fields that bales were grown in. The results of this feasibility study indicate the promise of incorporating remote sensing observations into predictions for the quality of cotton harvests.

Key Terms: remote sensing, EVI, cotton quality, crop yield, regression model, phenology, HLS, GPM, SMAP

Advisors: Sean McCartney (NASA Goddard Space Flight Center, Science Systems and Applications, Inc.), Thomas Holmes (NASA Goddard Space Flight Center), Manh-Hung Le (NASA Goddard Space Flight Center, Science Applications International Corporation), Yoseline Angel (NASA Goddard Space Flight Center, University of Maryland).

Lead: Spencer Harman (Maryland – Goddard)

1. Introduction

Cotton is the most widely cultivated natural fiber globally, with an estimated annual production of 25 million tons (Khan et al., 2020). The United States plays a key role in the global textile industry as the third-largest producer of cotton fiber (United States Department of Agriculture Economic Research Service, 2022). With global warming expected to adversely affect crop quality (Intergovernmental Panel on Climate Change, 2022), preserving fiber quality is as important as optimizing yields, since low-quality bales can disrupt processing efficiency and lead to economic losses (Beegum et al., 2023). Various abiotic and biotic factors affect cotton yield and quality. Although warm temperatures are vital for cellulose synthesis, which constitutes 85% of cotton fibers, recent studies have shown heat stress as a primary factor impairing reproductive activity and degrading lint quality, including fiber length, micronaire, and strength (Saini et al., 2023; Beegum et al., 2024). Also, extreme weather events such as droughts or floods, and biotic stressors have been shown to have an adverse impact on quality and yield, with production losses estimated at approximately 50 – 60% under such conditions (Noreen et al., 2020).

Within the United States cotton industry, the Agricultural Research Service and the Agricultural Marketing Service (AMS) of the United States Department of Agriculture (USDA) have been concerned about declining cotton quality and yields in recent years. The Agricultural Research Service (ARS) develops new processes, applications, and product-enabling technologies which facilitate the expanded use and enhanced value of US cotton, while the AMS provides weekly cotton quality reports and recommendations to cotton producers on optimizing yield and quality (USDA AMS, n.d.). Considering recent extreme weather events such as Hurricane Harvey, which caused substantial losses to Texas cotton production, these agencies hypothesize climate variables as a principal factor driving cotton production decline. Presently, the ARS conducts physical testing of cotton bales in designated warehouses to assess cotton quality; however, these agencies seek to integrate remote sensing technologies into their quality monitoring practices to better evaluate the factors affecting cotton health.

Research has shown the viability of using vegetation indices derived from the visible to near-infrared regions of the electromagnetic spectrum to track cotton development. Johnson (2016) conducted a national study correlating remote sensing products with county-level yields of ten major crops in the United States and found that most vegetation indices were positively associated with cotton yields throughout the growing season. The Enhanced Vegetation Index (EVI) exhibited the highest correlation (~0.7) of these indices in late July and early August (Johnson, 2016). Ping et al. (2023) applied lasso and random forest regression models, incorporating both historical climate data and vegetation indices (from Terra Moderate Resolution Imaging Spectroradiometer [MODIS] and Sentinel-2 MultiSpectral Instrument [MSI]), to accurately predict cotton yields at the county level in Xinjiang Province, China. Similarly, Le et al. (2024) improved vegetative health prediction by combining early-season Normalized Difference Vegetation Index (NDVI) calculations with NASA's Soil Moisture Active Passive (SMAP) algorithm. This integration enhanced yield forecasts, especially during extreme water events such as floods or droughts.

Research on remote sensing for cotton quality assessment remains limited compared to yield studies, but existing literature shows promise. For example, Xu et al. (2023) identified unique spectral responses for cotton fiber length and strength within the 350 – 920 nm visible light range and the 1400 – 2500 nm shortwave infrared range. In a recent study, Li et al. (2025) demonstrated that nearly 50% of the variation in cotton fiber quality traits, such as strength, length, and micronaire, could be attributed to climate change. They highlighted temperature, precipitation, and radiation as the key climate factors influencing cotton quality, particularly during the later stages of crop growth (Li et al., 2025).

To address the concerns of the ARS and AMS agencies of the USDA, we examined the feasibility of using remote sensing and ground-based data to identify environmental factors influencing cotton quality and yield in western Texas and southern Georgia from 2015 to 2024. We collected and analyzed data on broadband vegetation indices, precipitation, soil moisture, wind speed, and near-surface air temperature from May to

November for each year in the study period. We focused on three main objectives: 1) characterizing the phenological patterns of cotton for each region, 2) identifying the main variables associated with cotton quality, and 3) quantifying the relationship between environmental factors and cotton quality using a regression analysis.

The study areas for this project were western Texas and southern Georgia (Figure 1). Together, these states account for approximately 50% of domestic cotton production in recent years and presented an opportunity to compare trends across two significant but distinct climate regions (USDA Economic Research Service, n.d.). We conducted our analysis at the National Agricultural Statistics Service (NASS) district level, which are groups of neighboring counties delineated by NASS that share similar agricultural practices and climate. The study focused on five NASS districts in western Texas and three NASS districts in southern Georgia (Figure 1).

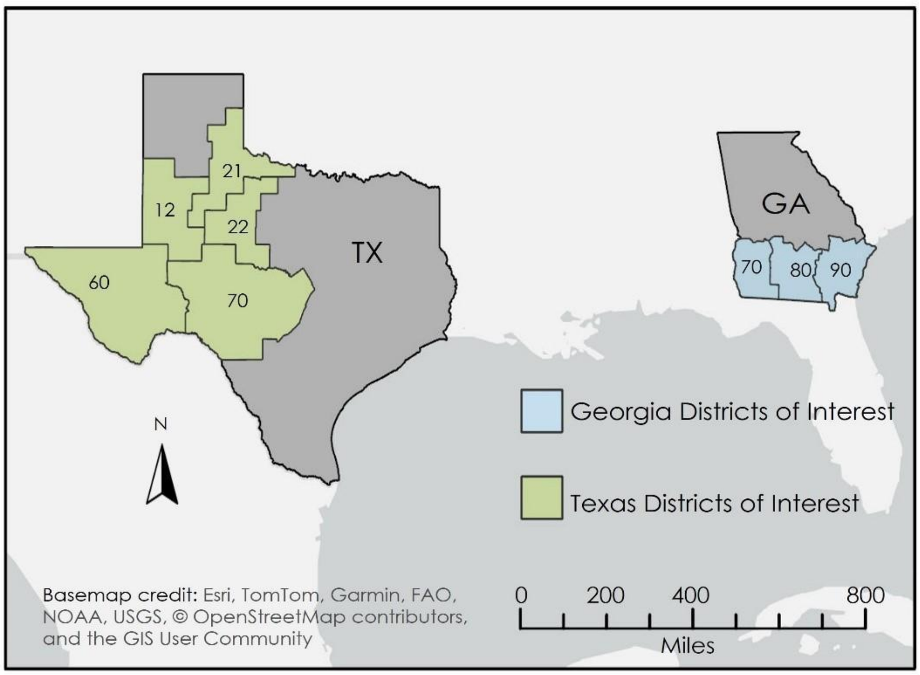

Figure 1. NASS Districts of Interest in Western Texas and Southern Georgia.

2. Methodology

2.1 Data Acquisition

2.1.1 Satellite Data

We acquired multispectral imagery from the Harmonized Landsat Sentinel-2 (HLS) initiative, which integrates observations from the Landsat 8 Operational Land Imager (OLI), Landsat 9 OLI-2, and Sentinel-2 MSI (Table 1). We acquired all available HLS data for each district using the bounding boxes of the NASS districts of interest and extracted modeled daily soil moisture estimates from the SMAP Level-4 datasets (Reichle et al., 2022) for the states of Texas and Georgia. SMAP Level-4 includes near-surface and root-zone soil moisture by assimilating satellite-based SMAP soil moisture data into a land surface model, providing uninterrupted soil moisture data every three hours. We obtained precipitation data from the Integrated Multi-Satellite Retrievals (IMERG) product of the Global Precipitation Measurement (GPM) mission (Huffman et al., 2015). We selected IMERG Final Run, Version 07, monthly precipitation data. The specific variable used was the "merged satellite-gauge precipitation estimates", which is the precipitation estimate calibrated with gauge data. The satellite remote sensing data were acquired from May to November of each year of the study period, except for HLS which was acquired from March to November.

Table 1
List of model components, sensors and data products used in this project

Model Component	Platform & Sensor	Variable(s) & Resolution	Source
Enhanced Vegetation Index (EVI)	Landsat 8 OLI Landsat 9 OLI-2 Sentinel-2 MSI	Surface Reflectance Global 30 m v2.0 – HLSS30.002 HLSL30.002	Land Processes Distributed Active Archive Center (LP DAAC)
Root Zone Soil Moisture	SMAP L-band Radiometer	L4 Global 3-hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Analysis Update	NASA National Snow and Ice Data Center Distributed Active Archive Center (NSIDC DAAC)
Precipitation	GPM IMERG	GPM IMERG Final Precipitation (mm) L3 1 month 0.1-degree x 0.1-degree V07	Goddard Earth Sciences Data and Information Services Center (GES DISC)
Standardized Precipitation Index (SPI)	NA	gridMET 30-day SPI, NASS district	gridMET Climate Engine API
Growing Degree Days-60 (DD60)	NA	County Daily Maximum, Minimum, and Mean Temperature	PRISM
Cotton Mask	NA	Crop Sequence Boundary (CSB)	USDA National Agricultural Statistics Service (NASS)
Wind Speed	NA	gridMET Daily Maximum Wind Speed (m/s), NASS district	gridMET Climate Engine API
Cotton Quality	NA	Micronaire, Fiber Length, Fiber Strength, Fiber Uniformity, Reflectance (HVI RD), Yellowness (HVI +b), Surface Area Trash (each bale graded on all variables)	USDA Agricultural Marketing Service (AMS)

Cotton Yield	NA	Yield (lbs./acre)	USDA National Agricultural Statistics Service (NASS) Quick Stats API

2.1.2 Ground-based Weather Data

We retrieved near surface temperature dataset from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) Climate Group website using the data explorer platform. To obtain data for each district of interest, we used the corresponding latitudes and longitudes of each county within the districts to extract the data at the county level. We then filtered the data and selected the daily minimum (T_{min}), mean (T_{mean}), and maximum temperature (T_{max}) in degrees Fahrenheit from March to November for each year of the study period. Additionally, we obtained daily maximum wind speed data and monthly SPI data with a 4 km resolution via Climate Engine for all pixels intersecting the study area (Huntington et al., 2017). SPI data is originally computed based on the previous 30-day rainfall and regularly updated every five days until the end of the month, so we extracted the SPI value of the last day of each month, which accurately represented the drought condition of the given month (Abatzoglou, 2011).

2.1.3 Cotton Data

Our partners at the USDA AMS provided data containing weekly quality metrics on every bale of cotton that was processed and graded for the crop year (which runs from June 1st of the given calendar year into May of the following calendar year) from 2015 to 2024. We retrieved yield data (measured in pounds per acre) for the NASS districts of interest from 2015 to 2019 using the Quick Stats Ad-Hoc Query Tool available on the NASS website. For 2020 to 2024, this data was not available at the NASS district level, so we retrieved yield data for all the counties constituting the districts in our study area and aggregated them at the district level. Additionally, we obtained crop sequence boundary (CSB) data from the NASS website. This land cover dataset provided information on the type of crop cultivated and its corresponding area.

2.2 Data Processing

2.2.1 Cotton Data

To arrive at representative quality metrics for the cotton grown each crop year in each NASS district, we calculated an average of the measurements across all bales for each of the seven quality indicators of interest in each weekly cotton quality data files. We also recorded the number of bales represented in each weekly file to calculate a weighted average, for each quality indicator, of the combination of NASS district and crop year. Since our partners were also interested in comparing the variability of some quality indicators as response variables, we calculated a weekly coefficient of variation – the ratio of the standard deviation to the mean – for the indicators of micronaire, fiber length, and trash surface area within each NASS district. We then averaged these weekly coefficients of variation across the crop year to arrive at a measurement that we called the "Typical Coefficient of Variation" for each of these three indicators, which gave a standardized indication of the amount of variability between bales processed in that NASS district for the given quality indicator.

2.2.2 Satellite Data

To process remote sensing datasets such as HLS and SMAP, we first generated a cotton mask for each region using the CSB dataset. Initially, we identified all agricultural areas documented as cotton cultivation between 2015 and 2023. Subsequently, the final mask delineated cotton-growing areas as fields that had cultivated cotton during at least two of the four most recent growing seasons (2020 – 2023) available in the CSB dataset.

To characterize cotton phenology for each district, we then calculated EVI (Equation 1; Huete et al., 2002) using the near infrared (NIR), red, and blue spectral bands of the HLS dataset. Specifically, our calculations utilized bands 5, 4, and 2 from the HLS Landsat OLI sensor and bands 8A, 4, and 2 from the HLS Sentinel-2 MSI sensor. We mosaicked multiple EVI tiles with the same acquisition date for each NASS district. For each month, we overlaid all EVI raster layers and selected the maximum pixel values to create one EVI raster per month, reducing the impact of cloud cover. Finally, we applied a cotton mask to extract EVI values for cotton-growing areas and averaged the cotton EVI pixels within each district to produce a single monthly EVI value for each district.

$$EVI = 2.5 \times \frac{NIR - Red}{NIR + 6 \times Red - 7.5 \times Blue + 1} \tag{1}$$

Monthly precipitation data from GPM IMERG was originally provided in 10 km grids and at a half-hourly rate. To obtain the mean hourly precipitation rate at the district level, we clipped all the grids that fell within the districts using the bounding box of each district. We then obtained the final precipitation per month for each district by multiplying the hourly precipitation rates by the number of hours for each given month, taking into consideration the varying number of hours per month. To calculate monthly root zone soil moisture, we applied the cotton mask to identify pixels corresponding to cotton-growing areas and averaged them within each district to obtain a single daily value for root zone soil moisture. We then averaged the daily values for each month to derive the monthly root zone soil moisture for each district.

2.2.3 Ground-based Data

We used temperature data from PRISM to calculate growing degree days, a metric that tracks heat accumulation to monitor changes in crop and pest development. The threshold temperature for cotton growth is 60˚F; therefore, growing degree days for cotton are mostly referred to as "DD60's" (Texas A&M AgriLife Extension, n.d.). We calculated DD60's for each day and accumulated DD60's for the counties in each district, starting from May 1st to November 30th for all years of the study period (Equation 2). We started accumulating from May since less than approximately 15% and 10% of cotton is planted before May for Texas and Georgia respectively (USDA NASS, 2025). Lastly, for each month, we averaged the DD60's and cumulative DD60's for all counties in a district to get the monthly DD60's and cumulative DD60's for each district. To derive the monthly maximum wind speed for each district, we selected the highest wind speed recorded in a month among all counties in each district. SPI data was already at the monthly level, so we aggregated the pixels in each district to get the monthly SPI per district.

$$DD60s = \left(\frac{(Tmin + Tmax)}{2}\right) - 60 \tag{2}$$

2.3 Data Analysis

2.3.1 Preliminary Analysis

Monthly EVI values were used to generate time series graphs to determine the start, peak, and end of a growing season for cotton plants in each district. We compared the DD60's cumulative results to the cotton growth stage values (Table 2) to establish an approximate phenology of the cotton grown in each combination of year and NASS district, allowing all subsequent parts of the analysis to be tied to specific stages of the cotton growth cycle. For each of the ten measurements of cotton quality, as well as the cotton yield variable, we generated a visualization of the univariate distribution to determine if any unusual values existed in the dataset for this measurement. These outliers served as the basis for case studies. For each noted outlier, the combination of year and NASS district was recorded and then highlighted in the distribution of each of the predictor variables. We then reviewed each case to see if any extreme values of one or more predictor variables were connected to outliers for an aspect of cotton quality or yield. Also, we ran the full set of 35 predictor variables (the monthly values from May through November for growing degree days, EVI, max wind speed, precipitation, and root zone soil moisture), as well as the full set of average cotton quality and yield metrics through a correlation analysis by state. The idea here was to determine which variables on the respective input and output sides of the models were generally dependent on each other throughout the time frame of the analysis, which would help inform the outputs from the regression models described below.

Table 2
Average number of days and heat units required for various growth stages of cotton in Texas, United States (Texas A&M AgriLife Extension. (n.d.)

Growth Stage Sequence	Days	Heat Units – DD60's
Planting to Emergence	5 – 17	59 – 159
Emergence to First Square	29 – 41	378 – 663
Square to Flower	21 – 27	719 – 1129
Flower to Open Boll	43 – 58	1857 – 2021
Planting to Harvest Ready	79 – 95	2150 – 2420

2.3.2 Regression Analysis
The primary component of the analysis was to determine the main variables associated with cotton quality though regression models; this was completed using RStudio (version 4.3.0). In the absence of a controlled experimental setting, and with so many variables to consider, it was not possible to establish a causal relationship between any of our potential predictors and the various aspects of cotton quality. However, we could at least identify some variables which – at a minimum – were correlated with (perhaps latent) variables that were truly predictive of these outcomes. Due to the large number of predictors (35) relative to observations (the final counts were 40 in Texas, 20 in Georgia), ordinary least squares regression methods were inappropriate, but the alternative of lasso regression models allowed for a theoretically sound way to perform variable selection. We built separate models for each cotton quality indicator, as well as for yield, within each state, creating a total of twenty-two regression models. Through the assignment of a penalty for non-zero coefficients (the optimal value for this penalty parameter being chosen through cross-validation), lasso regression resulted in the elimination from the model of those variables which were only very weakly correlated with the response variable of interest. We ran the lasso regression algorithm one hundred times in each state for each response variable to account for the variability introduced by the choice of penalty parameter. By examining which variables consistently appeared within the resulting regression models within each state, we were able to focus on justifiable correlates of cotton quality in Texas and Georgia, respectively.

Furthermore, since this feasibility study was meant to determine if remote sensing products could be used to inform our partners' decision-making process, we repeated the above lasso regression analysis but with the removal of all predictors that had components with satellite-derived data in them. This left a smaller set of 14 potential predictors, as only the growing degree days and maximum wind speed variables met this definition. For each response variable, we then compared the full lasso regression model to the smaller lasso regression model. Since predictive power can only be increased through the inclusion of more variables in the model, we used adjusted R-squared (which accounts for the loss in parsimony and interpretability as more variables are added to the model) as a quantitative comparison tool to see if the full models including variables based in remote sensing technology were more useful.

3. Results

3.1 Analysis of Results

Two of the original eight NASS districts selected for inclusion were excluded from the analysis. In TX-60, we determined that Pima cotton (a distinct species primarily grown in the Southwestern United States) was the primary type of cotton being grown; this extra-long staple species has known trait differences from the Upland cotton grown in the rest of the study area and would have confounded our analysis. In GA-90, we encountered multiple years of missing data due to the reduction in the number of in-district processing gins to one in 2020. Since the USDA could no longer anonymize the data associated with this one gin, we could not obtain cotton quality measurements for the last half of the study period we were interested in.

As a means of establishing a general phenology for the cotton crops in our study, Figure 2 compares the 2022 EVI time series, as a representative year, to the monthly average and range of the study period for the GA-70 and TX-12 NASS districts. These time series reflect vegetation presence and confirm that in these two districts (which are representative of all districts we examined in their respective states), cotton crops have already emerged and are developing significant green vegetation as early as May, as shown by a sharp rise in EVI. Crops reach peak greenness around August and September before declining in October and November, when bolls are primarily open, and harvest occurs. Comparing peak EVI values reveals that cotton fields in Georgia are greener and healthier than those in Texas. The 2022 EVI time series is notably higher than the average for both districts, aligning with the fact that 2022 recorded the highest yield for GA-70 and the second highest yield for TX-12 within the study time frame. Considering the extreme drought conditions Texas experienced in 2022, our partners hypothesized that this anomaly might be due to irrigation; indeed, observations from EVI images and Google Earth revealed that a sizeable portion of the TX-12 fields consisted of pivot irrigation farms. Whether a field was strictly rain-fed or irrigated was a variable that was not measured in the current study. The monthly average EVI time series of the study period for all the districts is shown in the appendix (Figure C1).

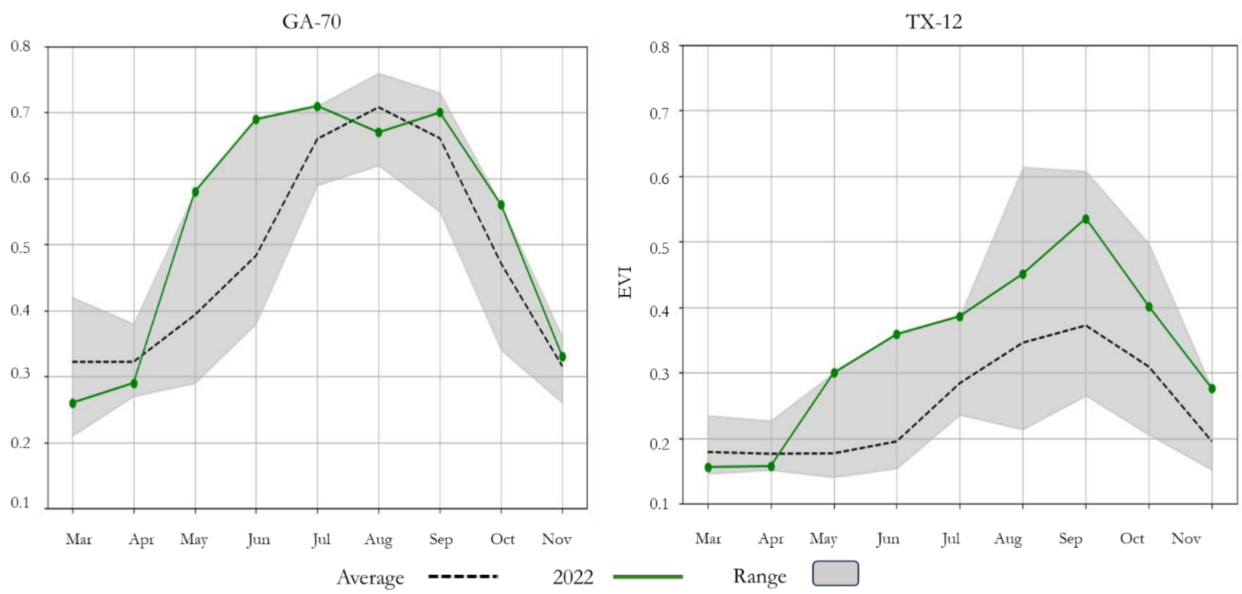

Figure 2. EVI time series for one NASS district in each state.

One collection of extremes appeared in our outlier analysis; in the year 2017, three NASS districts in Texas reported unusually low average micronaire values (3.50 in TX-12, 3.76 in TX-21, 3.81 in TX-22). Because our partners at the USDA had mentioned this as a concerning situation, we conducted a case study analysis that is summarized in the figures located in Appendix A. For each of the 35 predictor variables, we graphed the distribution across all years and districts in Texas, highlighting the values of the three districts in 2017 with the troubling micronaire values. Mostly, these three values did not appear as extreme outliers in any of the predictor variable distributions; however, it was noted that the three points corresponded to some of the lowest August growing degree days values (Figure A2) and highest August precipitation values (Figure A3) over the study years in Texas.

The correlation matrices of the response variables in each state (Appendix B) produced two interesting patterns. In Texas, average fiber length, average fiber strength, and average uniformity formed a cluster of positively associated measurements (Figure B1). Both states had a robust negative linear relationship between reflectance and surface area trash. In Georgia, the correlation coefficient was -0.69 (Figure B2), while in Texas, the correlation coefficient was -0.83 (Figure B1).

The primary piece of our analysis to find drivers of cotton quality was a multiple regression analysis. The criteria for inclusion as an "important correlate" of cotton quality in Texas was twofold: first, the predictor had to appear with a non-zero coefficient in all 100 iterations that we conducted of the lasso regression algorithm (each of which had a unique tuning parameter based on the cross-validation randomization used), and second, the predictor had to have a correlation coefficient magnitude of at least 0.50 with the quality metric in question (Table 3). During the crop's early stages through June, no variables appeared as meaningful predictors in our regression models. In July, growing degree days showed a strong positive correlation with eventual average fiber length. In August, which roughly corresponds to the boll development stage for much of the crop in western Texas, precipitation was strongly associated with crops that were lower in yellowness, while root zone soil moisture was negatively associated with fiber strength and uniformity. EVI readings in August showed a positive association with the brightness of the crop at the time of processing. In the back half of the growing season, heat in September (as measured by growing degree days) generally led to thicker fibers. At the end of the season, high root zone soil moisture (0 – 100 cm vertical average) in a district was associated with crops with more surface area trash, lower fiber strength, and darker color. When precipitation was high at the end of the growing season, there was a similar correlation with crops that were high in trash matter and darker color.

Table 3

Important predictors found to be consistently associated with specific cotton quality measurements at various times of the growing season in Texas

Month	**Predictor**	**Response**	**Correlation Coefficient**
July	Growing Degree Days	Fiber Length	0.58
August	Enhanced Vegetation Index	Reflectance	0.55
August	Precipitation	Yellowness	-0.68
August	Root Zone Soil Moisture	Fiber Strength	-0.56
August	Root Zone Soil Moisture	Fiber Length Uniformity	-0.54
September	Growing Degree Days	Micronaire	0.54
October	Precipitation	Surface Area Trash	0.60
October	Precipitation	Reflectance	-0.54
October	Root Zone Soil Moisture	Fiber Strength	-0.51
November	Root Zone Soil Moisture	Surface Area Trash	0.60
November	Root Zone Soil Moisture	Reflectance	-0.57

There were three negative results worth noting here as well. The partners were interested in a potential association between high wind events and poor cotton quality in Texas, particularly unwelcome levels of surface area trash. Our regression models did not identify a relationship between these variables. The correlation coefficient between the maximum wind speed in each month and the average surface area trash of the crop in a district was typically near zero, with only the maximum wind speed in November ($r = 0.37$) having a hint of association with surface area trash. Also, the regression models that attempted to predict the typical coefficient of variation for micronaire, fiber length, and trash surface area had relatively low adjusted R-squared values (in the 15% to 30% range) and did not identify important environmental correlates as predictors, so we could not establish any associations involving the bale-to-bale variability of these quality measurements. Finally, the regression models built to predict yield had low explanatory power (adjusted R-squared = 0.12) and did not return any significant predictors in Texas.

In Georgia, we had difficulty establishing strong associations between our predictors and the cotton variables due to low statistical power in our study. However, we were still able to perform correlation analyses and noted that EVI demonstrated a moderate correlation with three important cotton measurements throughout the summer months. As early as July, EVI showed a positive correlation with both fiber strength (r = 0.51) and fiber length (r = 0.52), and this persisted throughout August (r = 0.18 for strength, r = 0.61 for length) and September (r = 0.48 for strength, r = 0.62 for length). Over the summer months in Georgia, EVI demonstrated a persistent negative correlation with the average micronaire of the eventual crop (r = -0.58 for June, -0.54 for July, -0.50 for August, and -0.64 for September). One further interesting finding in Georgia was that, like Texas, there was a strong negative correlation between the late-season precipitation and root zone soil moisture readings and the reflectance of the crop.

As a final component of our analysis, we compared the predictive power of our regression models that incorporated remote sensing observations to those models that used only modeled data from weather stations. The smaller models used only growing degree days and maximum wind speeds. In comparison, the larger models used these variables and the EVI, root zone soil moisture estimates, and IMERG precipitation estimates. Since adding more variables to a linear regression model can only increase the amount of variation in the response variable that the model is able to explain, we used adjusted R-squared as the metric to compare the two different types of models. Adjusted R-squared penalizes models that have irrelevant or highly collinear variables as predictors and is generally regarded as a reliable tool for evaluating the performance of models with differing numbers of predictors. Table 4 shows that for each of the primary quality metrics we examined (except for yellowness), the model that used predictors based on satellite data outperformed the simpler models, and by a large margin in some cases.

Table 4

Comparison of performance of lasso regression models in Texas

Quality Metric	**Adjusted R^2 without RS Observations**	**Adjusted R^2 with RS Observations**	**Increase in Percent of Variation Explained**
Fiber Strength	36.9	54.4	+17.5
Fiber Length	15.5	51.2	+35.7
Micronaire	57.7	72.5	+14.8
Length Uniformity	0	41.4	+41.4
Trash Surface Area	26.6	63.5	+36.9
Color Reflectance	18.6	60.2	+41.6
Color Yellowness	38.1	31.6	-6.5

3.2 Errors & Uncertainties

As is typical in a spatial analysis, the modifiable areal unit problem potentially biased our calculations. The point-based readings of cotton quality – each bale represented output from a particular farm – were aggregated to the spatial level of NASS district, a relatively coarse scale at which to analyze agricultural data. This choice of areal unit was driven by the fact that the AMS weekly files of cotton quality data identified the NASS district where each cotton bale was ginned but gave no finer information in terms of spatial identification. To achieve consistency throughout our analysis, we aggregated all environmental data to this same spatial scale as well, which may have introduced statistical bias in estimating our regression model coefficients.

Similarly, there was evidence of missing and misleading cotton quality data. Due to privacy concerns, the AMS was unable to release data regarding cotton quality in situations where that data could be tied to a specific private business. As such, when a single cotton gin was in operation within a particular area for a year, the quality data was withheld to maintain the anonymity of private business operations. We were also made aware of, but unable to pinpoint, the precise situations in which bales processed in a particular NASS district were grown on land contained within a neighboring NASS district. Cotton farmers have their choice of which gin to process their crop at, and that choice does not necessarily obey NASS district boundaries.

This created a mismatch between the environmental data that governed the cotton's development and the resulting quality data for an unknown number of bales in our study.

While planting dates can vary at the field level and by year due to local climate conditions and logistical factors, we assumed a uniform planting date of May 1st across all districts and seasons in the study period, based on historical records. This simplification may have introduced bias into the use of DD60s for estimating cotton growth stages by district and season, and potentially affected its relationship with other variables—particularly EVI, which was also used to model cotton phenology.

The NASS districts chosen for inclusion in the study were initially selected for an "apples-to-apples" comparison of primarily rain-fed operations in two distinct climatic zones in the south. However, this assumption was called into question as the study progressed and, we suspect that irrigation was a confounding variable in the analysis. This made it particularly difficult to interpret findings related to precipitation or root zone soil moisture. If low precipitation was associated with a quality measurement, we did not know if the crop was responding to the low precipitation levels, or perhaps to high levels of moisture provided by compensatory irrigation.

Finally, the lasso regression models we built all assume a roughly linear response in the outcome variable, as well as the independence and normality of the residuals from the model. While these assumptions were checked and verified for some of the regression models we built (through residual plots and Q-Q plots), we were unable to do this for all models we constructed.

4. Conclusions

4.1 Interpretation of Results

While we were unable to build reliable regression models in Georgia due to insufficient data, the main take-away from the analysis is that indices derived from satellite observations taken in mid-summer can provide some indication of the average quality readings of the crops that will be harvested in autumn. The moderate positive correlation between EVI readings in July, August, and September on the one hand, and cotton fiber length on the other, makes sense given what we know about cotton development. Cotton fiber length is completely determined during the first three weeks after the flower has been fertilized and the closed boll is developing, which roughly corresponds to July and August in western Texas (depending on the exact planting date). When we pick up on higher-than-normal EVI readings from remotely sensed cotton fields during the flowering and boll development stages – indicating vigorous crops – we can anticipate that the fiber lengths at harvest time will generally be higher-than-normal as well.

The biggest takeaway from this feasibility study is the promise of incorporating remote sensing observations into predictions for the quality of cotton harvests generally. Even at the coarse spatial scale of multi-county agricultural districts and the coarse temporal scale of crop season, we were able to show that regression models that incorporated remote sensing observations explained more of the variability in average cotton quality measurements than models that used only weather data (Table 4). Our analysis of the NASS districts in Texas identified some variables that, over the last decade, have been associated with various quality measurements in the cotton crops. The earliest seasonal relationship we identified was between growing degree days in July and the eventual average length of the cotton fibers, which is plausible given the discussion of the Georgia results. The other major finding was the cluster of moisture-related variables associated with quality measurements (Table 3). We caution against reading too much into these associations for two main reasons: the first being the lack of information surrounding the level of irrigation used in the Texas NASS districts in question and the second reason applies more generally to the full set of results in this study, which is that all the analyses done here could only establish associations, not causal relationships, between variables.

4.2 Feasibility & Partner Implementation

The relationships uncovered in this study between environmental measurements and average values of cotton quality metrics are helpful, as the USDA can use information about (for instance) growing degree days recorded in July in Texas to have a general sense of the average fiber length of the crop that will arrive at the end of the season. Even more so, this feasibility study serves as an encouraging indication of more that could be discovered with access to the appropriate data. A more detailed study undertaken at the county or (ideally) field level has the potential to pin down key variable-time combinations that impact cotton quality. EVI shows promise as a remote sensing tool for making predictions of cotton quality, thanks in large part to the fine temporal (every 2–3 days) and spatial (30 m) resolution of the estimates that can be obtained based on data from the HLS dataset. Root zone soil moisture can be considered as an alternative to precipitation measurements when predicting cotton quality, as it indirectly measures rainfall and irrigation, as well as the persistence of this moisture for crops to access. The frequency with which root zone soil moisture showed up as a correlate of quality metrics in our Texas regression analysis is a suggestion of its utility moving forward. Care must be taken with the differing levels of temporal and spatial resolution available for all the variables of interest; generally, the finer resolution that an analysis can be conducted at, the more reliable the results will be, as the modifiable areal unit problem can be avoided at finer scales. Future research should also pay close attention to the matter of irrigation and either control for this variable or ensure that it is accurately measured for incorporation in any analysis (perhaps by using root zone soil moisture instead). Overall, we have established that regression models that incorporate remote sensing products are useful for identifying correlates of – and making in-season predictions about – cotton quality.

5. Acknowledgements

Project Partners:

- Dr. Michael Cosh – Research Hydrologist, Hydrology and Remote Sensing Laboratory, USDA ARS
- Chris Delhom – Research Leader, Sustainable Water Management Research Unit, USDA ARS
- Barbara Meredith – Director, Market News Division, USDA ARS
- Vikki Martin – Vice President, Fiber Competition, Cotton Incorporated
- Ed Barnes – Senior Director, Agricultural & Environmental Research, Cotton Incorporated

Science Advisors:

- Sean McCartney – Goddard Space Flight Center, Science Systems and Applications Inc.
- Thomas Holmes – Goddard Space Flight Center
- Manh-Hung Le – Goddard Space Flight Center, Science Applications International Corporation
- Yoseline Angel Lopez – Goddard Space Flight Center, University of Maryland

Special Thanks:

- Spencer Harman – NASA DEVELOP Center Lead, Maryland – Goddard
- Marisa Smedsrud– NASA DEVELOP Geoinformatics Fellow, Maryland – Goddard
- Dr. Kent Ross– NASA DEVELOP Program Manager, NASA Langley Research Center

This material contains modified Copernicus Sentinel data (2015 – 2024), processed by ESA.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration.

This material is based upon work supported by NASA through contract 80LARC23FA024.

6. Glossary

Crop Sequence Boundary (CSB) – Geospatial algorithm-based field polygons

Earth observations – Satellites and sensors that collect information about the Earth's physical, chemical, and biological systems over space and time
Enhanced Vegetation Index (EVI) – A metric used to quantify vegetation greenness and density, offering improved sensitivity in regions with high biomass by minimizing atmospheric and canopy background influences
Global Precipitation Measurement (GPM) – An international satellite mission between NASA and JAXA designed to unify and advance precipitation measurements from research and operational microwave sensors for delivering next-generation global precipitation data products
Growing Degree Days – Heat units that measure the amount of useful heat energy a plant accumulates each day, month, and season, the plant must accumulate a specified level of heat units to reach each development stage and to achieve complete physiological maturity
Harmonized Landsat Sentinel-2 (HLS) – A NASA initiative designed to generate consistent surface reflectance dataset by harmonizing observations from Landsat 8 Operational Land Imager, Landsat 9 Operational Land Imager-2, and Sentinel-2 MultiSpectral Instrument
Integrated Multi-satellitE Retrievals for GPM (IMERG) – Product that combines information from the GPM satellite constellation to estimate precipitation over the majority of the Earth's surface
MODIS – Moderate Resolution Imaging Spectroradiometer
NASS – National Agricultural Statistics Service
NASS Districts – NASS uses Agricultural Statistics Districts to group neighboring counties for statistical analysis and reporting
Soil Moisture Active Passive (SMAP) – Earth satellite mission designed to measure and map Earth's soil moisture and freeze/thaw state to better understand terrestrial water, carbon and energy cycles
Standardized Precipitation Index (SPI) – Index used for estimating wet or dry conditions based on precipitation variable

7. References

Abatzoglou, J. T. (2011). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology, 33*(1), 121 – 131. https://doi.org/10.1002/joc.3413

Beegum, S., Reddy, K. R., Ambinakudige, S., & Reddy, V. (2024). Planting for perfection: How to maximize cotton fiber quality with the right planting dates in the face of climate change. *Field Crops Research, 315*, 109483. https://doi.org/10.1016/j.fcr.2024.109483

Beegum, S., Reddy, V., & Reddy, K. R. (2023). Development of a cotton fiber quality simulation module and its incorporation into cotton crop growth and development model: GOSSYM. *Computers and Electronics in Agriculture, 212,* 108080. https://doi.org/10.1016/j.compag.2023.108080

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment, 83*(1-2), 195 – 213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2023). *GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree* (Version V07) [Data set]. Sciences Data and Information Services Center. Accessed February 19, 2025, from https://doi.org/10.5067/GPM/IMERG/3B-MONTH/07

Huntington, J., Hegewisch, K., Daudert, B., Morton, C., Abatzoglou, J., McEvoy, D., & Erickson, T. (2017). Climate Engine: Cloud Computing of Climate and Remote Sensing Data for Advanced Natural Resource Monitoring and Process Understanding. *Bulletin of the American Meteorological Society*, *98*(11), 2397 – 2410. https://doi.org/10.1175/BAMS-D-15-00324.1

Intergovernmental Panel on Climate Change. (2022). *Climate change and land: IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. Cambridge University Press. https://doi.org/10.1017/9781009157988

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. https://doi.org/10.1016/j.jag.2016.05.010

Khan, M. A., Wahid, A., Ahmad, M., Tahir, M. T., Ahmed, M., Ahmad, S., & Hasanuzzaman, M. (2020). World Cotton Production and Consumption: An Overview. In S. Ahmad & M. Hasanuzzaman (Eds.), *Cotton Production and Uses* (pp. 1–7). Springer. https://doi.org/10.1007/978-981-15-1472-2_1

Le, M-H., Bolten, J. D., Whitney, K. M., Johnson, D. M., Mueller, R., & Mladenova, I. E. (2024). On the Use of SMAP Soil Moisture for Forecasting NDVI Over CONUS Cropland Regions. *Geophysical Research Letters, 51*(20). https://doi.org/10.1029/2024GL111187

Li, X., Zhang, Z., Pan, Z., Sun, G., Li, P., Chen, J., Wang, L., Wang, K., Li, A., Li, J., Zhang, Y., Zhai, M., Zhao, W., Wang, J., & Wang, Z. (2025). Demonstrating almost half of cotton fiber quality variation is attributed to climate change using a hybrid machine learning-enabled approach. *European Journal of Agronomy, 162*, 127426. https://doi.org/10.1016/j.eja.2024.127426

Masek, J., Ju, J., Roger, J., Skakun, S., Vermote, E., Claverie, M., Dungan, J., Yin, Z., Freitag, B., & Justice, C. (2021). *HLS Operational Land Imager Surface Reflectance and TOA Brightness Daily Global 30 m* (Version v2.0). [Dataset]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/HLS/HLSS30.002

Noreen, S., Ahmad, S., Fatima, Z., Zakir, I., Iqbal, P., Nahar, K., & Hasanuzzaman, M. (2020). Abiotic Stresses Mediated Changes in Morphophysiology of Cotton Plant. In S. Ahmad & M. Hasanuzzaman (Eds.), *Cotton Production and Uses* (pp. 377–393). Springer. https://doi.org/10.1007/978-981-15-1472-2_18

Ping, L., Lifu, Z., Huang, C., Chen, J., Kang, X., Zhang, Z., & Tong, Q. (2023). Integrating environmental and satellite data to estimate county-level cotton yield in Xinjiang Province. *Frontiers in Plant Science, 13*. https://doi.org/10.3389/fpls.2022.1048479

PRISM Climate Group, Oregon State University. (2025) *PRISM Monthly Average Temperature, Minimum Temperature and Maximum Temperature* (AN91d). [Dataset]. Accessed February 21, 2025, from https://prism.oregonstate.edu/explorer/

Reichle, R., De Lannoy, G., Koster, R. D., Crow, W. T., Kimball, J. S., Liu, Q., & Bechtold, M. (2022). *SMAP L4 Global 3-hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Analysis Update* (Version 7) [Dataset]. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/LWJ6TF5SZRG3

Saini, D. K., Impa, S. M., McCallister, D., Patil, G. B., Abidi, N., Ritchie, G., Jaconis, S. Y., & Jagadish, K. S. (2023). High day and night temperatures impact on cotton yield and quality—current status and future research direction. *Journal of Cotton Research, 6*(1), 16. https://doi.org/10.1186/s42397-023-00154-x

Texas A&M AgriLife Extension. (n.d.). *Cotton Insect Management Guide- Development and Growth Monitoring of the Cotton Plant.* https://cottonbugs.tamu.edu/development-and-growth-monitoring-of-the-cotton-plant/

United States Department of Agriculture Economic Research Service. (2022). *Cotton and Wool - cotton sector at a glance.* https://www.ers.usda.gov/topics/crops/cotton-and-wool/cotton-sector-at-a-glance/

United States Department of Agriculture Agricultural Marketing Service. (n.d.). *Cotton.* https://www.ams.usda.gov/grades-standards/cotton

United States Department of Agriculture Agricultural Marketing Service. (2025). Weekly Cotton Quality Data Files. Accessed March 3, 2025 from https://apps.ams.usda.gov/Cotton/Weekly_Cotton_Quality_Data_files_by_NASS_Ag_Districts/

United States Department of Agriculture National Agricultural Statistics Service. (n.d.) *QuickStats Ad-hoc Query Tool.* [Dataset]. US Department of Agriculture. Accessed March 6, 2025, from https://quickstats.nass.usda.gov/

United States Department of Agriculture National Agricultural Statistics Service. (2025). *Crop Sequence Boundaries* (Version v2024) [Dataset]. https://www.nass.usda.gov/Research_and_Science/Crop-Sequence-Boundaries/index.php?form=MG0AV3

Xu, W., Yang, W., Chen, P., Zhan, Y., Zhang, L., & Lan, Y. (2023). Cotton Fiber Quality Estimation Based on Machine Learning Using Time Series UAV Remote Sensing Data. *Remote Sensing, 15*(3), 3. https://doi.org/10.3390/rs15030586

8. Appendices

Appendix A: *Outlier Analysis: Low Micronaire in Texas, 2017*

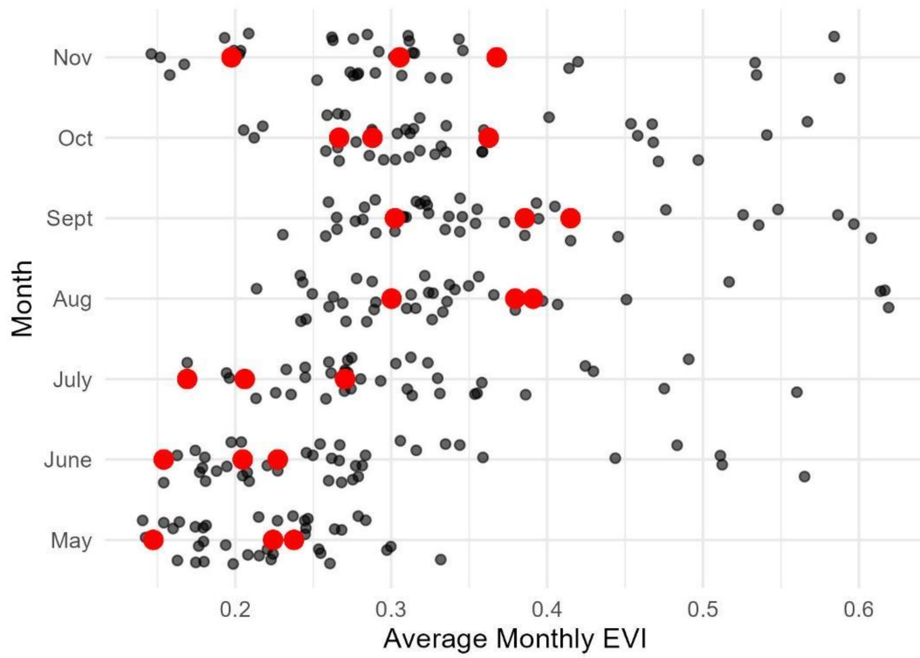

Figure A1. Monthly EVI in Texas.
The 3 districts with extremely low average micronaire in 2017 are highlighted in red.

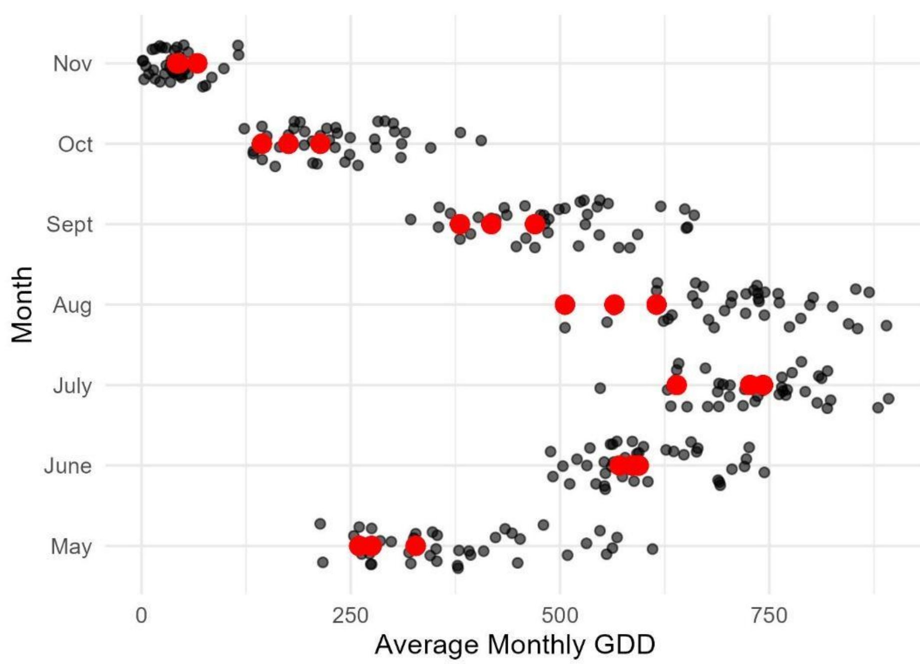

Figure A2. Monthly growing degree days in Texas.
The 3 districts with extremely low average micronaire in 2017 are highlighted in red.

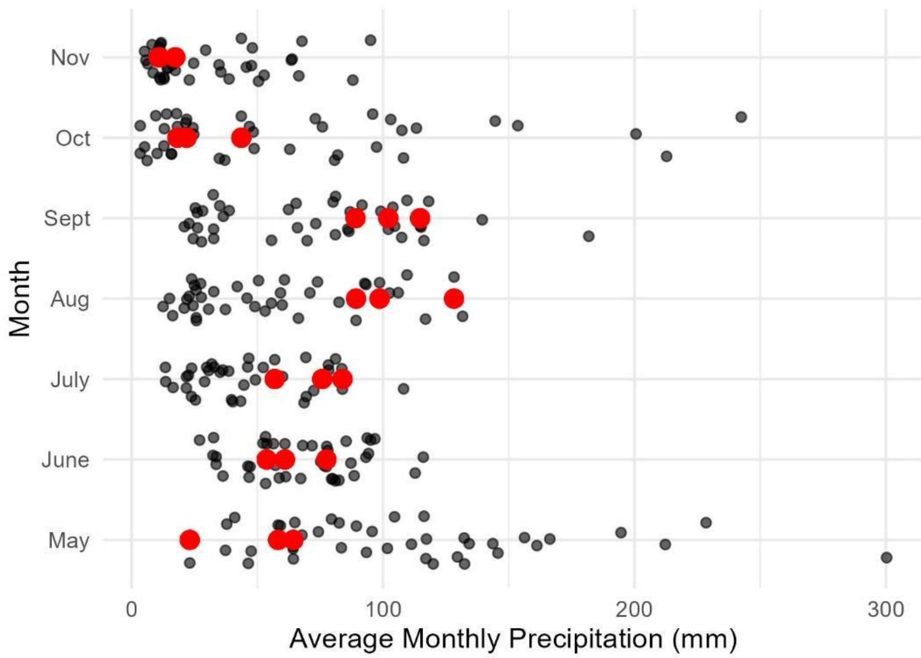

Figure A3. Monthly precipitation in Texas.
The 3 districts with extremely low average micronaire in 2017 are highlighted in red.

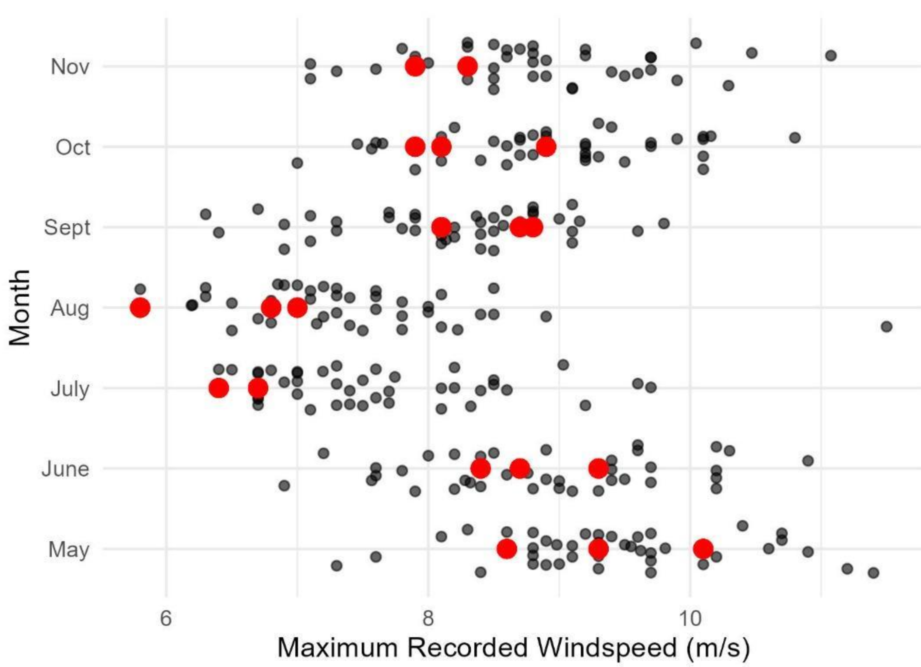

Figure A4. Monthly maximum windspeed in Texas.
The 3 districts with extremely low average micronaire in 2017 are highlighted in red.

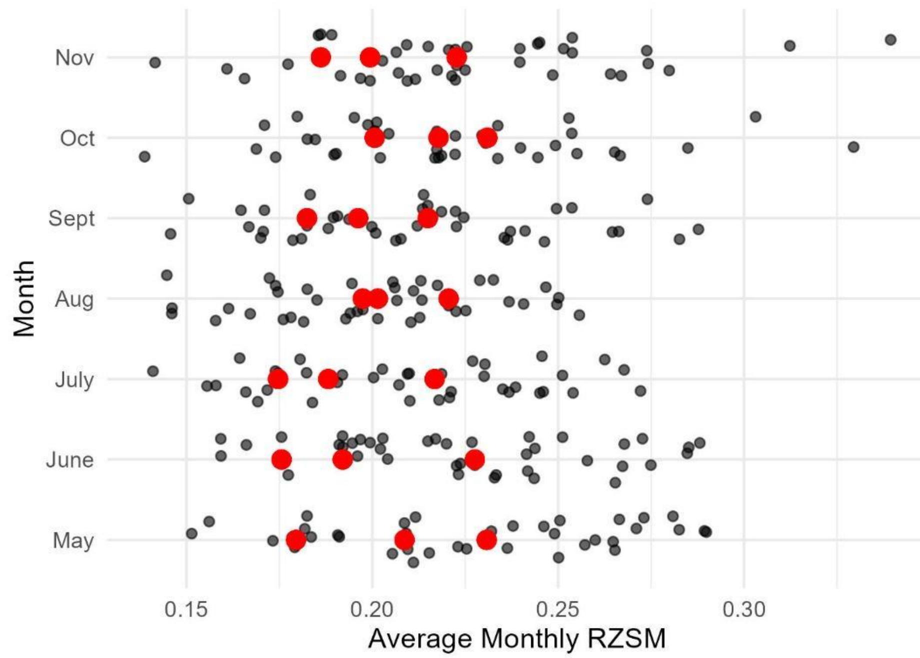

Figure A5. Monthly root zone soil moisture in Texas.
The 3 districts with extremely low average micronaire in 2017 are highlighted in red.

Appendix B: *Correlation Analysis*

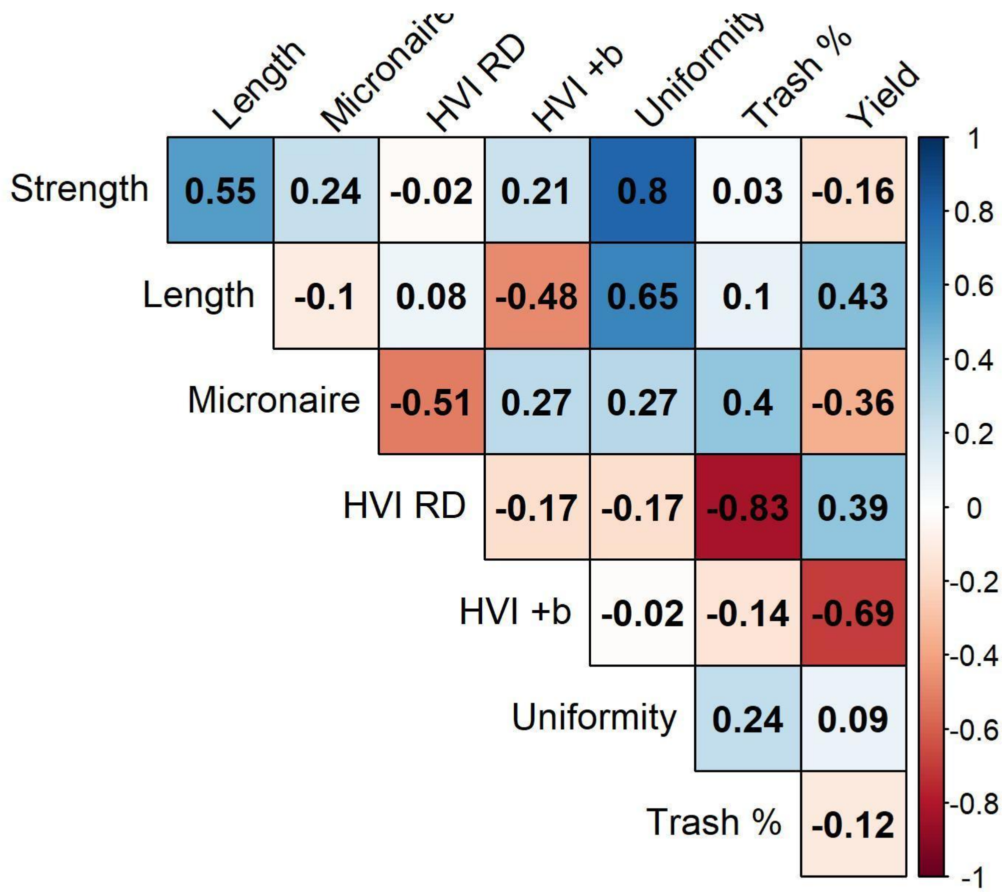

Figure B1. Correlation amongst average values of cotton metrics in Texas NASS districts of interest, 2015 – 2024

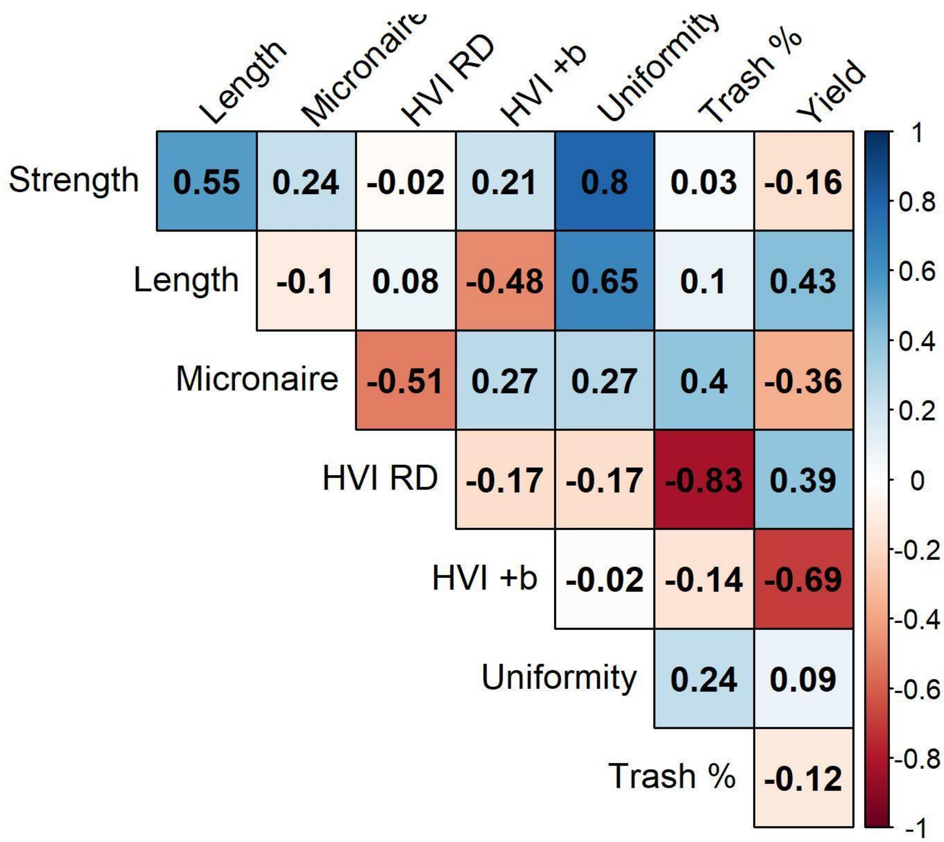

Figure B2. Correlation amongst average values of cotton metrics in Georgia NASS districts of interest, 2015 – 2024

Appendix C: *Cotton Phenology*

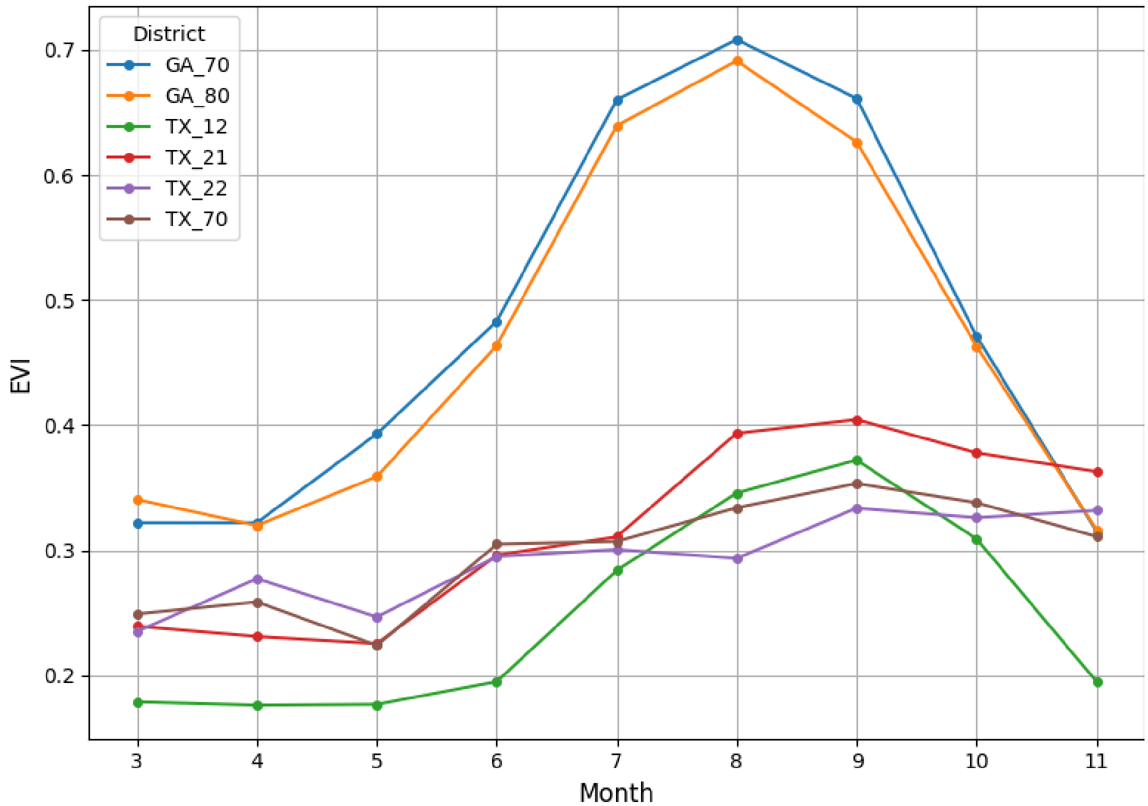

Figure C1. Monthly average EVI for the study period (2015 to 2024) for the NASS districts of interest