Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations

DOI: 10.54254/2753-8818/51/2024CH0204

Assessing the impact of natural hazards on U.S. cotton yields: a random forest analysis of drought, wildfires, and hurricanes (2012-2023)

Yasong Qin

College of Agriculture and Natural Resources, University of Maryland, MD, 20740, The United States of America

yasong88@terpmail.umd.edu

Abstract. The cotton industry is essential in the world. This paper investigates the impacts of natural hazards, particularly drought, wildfires, and hurricanes, on cotton yield (LB/Acre) in 14 major cotton-producing states in the United States from 2012 to 2023. The paper uses a random forest model with drought severity and cover index (DSCI), number of wildfire alerts, and number of hurricanes and their interactions as influencing factors on cotton yield. The results shows that wildfires were most likely to affect cotton yield among the three major categories of natural hazards, and hurricanes did not significantly affect cotton yield. Cotton yield decreases significantly in years with severe drought and high wildfire activity. In addition, the study emphasizes the compounding impact of these hazards, suggesting that the interaction between drought and wildfires exacerbates the effects on cotton yields. Overall, cotton growers and local authorities need substantial mitigation and adaptation strategies for wildfires and droughts.

Keywords: Cotton, drought, wildfires, random forest model.

1. Introduction

Ridley stated that the cotton industry in the United States has consistently been one of the largest in the world [1]. Although other formidable cotton exporters have emerged in the 21st century, it remains impractical to swiftly supplant the United States as the foremost global cotton exporter. The U.S. cotton sector generates over $21 billion in annual revenue from its products and services. The United States is a significant provider of raw cotton materials that the rapidly growing global cotton textile industry cannot disregard.

Starbird et al. pointed out that cotton cultivation in the United States first began in the early 17th century in Virginia and South Carolina to plant cotton. Subsequently, with the advancement of cotton processing technology, cotton cultivation gradually expanded to other regions [2]. In 2020, the USDA released data showing that Texas produced 4748,000 balls of cotton, accounting for 31.75% of the U.S. cotton production in the United States first, followed by Georgia and Arkansas, respectively, 14.58% and 8.69%.

Cotton farming in the American continent is predominantly concentrated in the southern region of the continental United States, namely throughout 17 states that serve as the primary cotton-producing

DOI: 10.54254/2753-8818/51/2024CH0204

areas. The cotton-growing region in the continental United States spans from 30 degrees north latitude to 37 degrees north latitude. The renowned Cotton Belt is created at this specific latitude. Jones et al. believe that The Cotton Belt region is characterized by its climatic characteristics, featuring an extended and scorching growth season, along with somewhat arid weather patterns throughout the cotton harvest. The region benefits from favorable meteorological and soil conditions conducive to cotton production, making cotton the primary source of income in the area [3].

Ray et al. found that for Texas, a major cotton-growing region in the U.S., which is chronically susceptible to drought, drought has had a significant impact on the yields and acreage of major crops (cotton, corn, winter wheat, and sorghum) in Texas during the period 2008-2016 [4]. According to the study, drought in Texas had the most significant impact on cotton and corn acreage, which decreased by 21% and 18%, respectively, during 2011-2013 [4]. Drought also creates a flammable environment for the occurrence of hill fires, making the rate and extent of their spread more uncontrollable. Armando estimated that from 2000 to 2011, more than $16 billion was spent on fire suppression alone [5]. Patrick et al., in their study of hill fires in the Midwestern and Eastern United States, found that high levels of smoke from hill fires can negatively impact corn and soybean produce [6].

Klotzbach et al. investigated and found that The East Coast and Gulf Coast of the United States, especially counties near the ocean and inland areas, are vulnerable to hurricanes [7]. In the Florida Cotton Production Guide, it is stated that Hurricanes occurring in August or early September cause the most damage, avoiding having the entire crop in a vulnerable condition (open and defoliated) when a hurricane arrives [8]. Pandya et al. experimentally found that in the case of cotton, heavy rainfall from hurricanes can also cause economic damage to cotton cultivation. Excessive rainfall may lead to soil oversaturation and waterlogging, adversely affecting cotton growth and development [9].

This paper aimed to analyze the relationship between cotton yield and various environmental damage factors through random forest modeling. The main focus is assessing each predictor's importance and model performance. Ultimately, the goal is to comprehensively understand the importance of three natural hazards, drought, wildfire, and hurricanes, and their interactions on cotton yield.

2. Methodology

2.1. Data source

The dataset utilized in this paper was compiled from 169 records spanning the years 2012 to 2023 across 14 cotton-growing states in the United States. The data was sourced from publicly accessible databases provided by the United States Department of Agriculture (USDA), the National Oceanic and Atmospheric Administration (NOAA), the U.S. Drought Monitor (USDM), and Global Forest Watch (GFW).

2.2. Variable selection

The dataset contains 9 explanatory variables: DSCI, FireAlertTimes, Hurricane, DSCI_ FireAlertTimes, DSCI_Hurricane, FireAlertTimes_Hurricane, DSCI_squared, FireAlertTimes_ squared, Hurricane_squared and one dependent variable (Yield). The specific description of this dataset is shown in Table 1:

Table 1. Variable Name and Meaning

Variable Name	Logogram	Meaning
DSCI	x_1	Drought Severity and Coverage Index
FiresAlertTimes	x_2	Number of Wildfires Alert Times in a year
Hurricane	x_3	Number of Hurricanes in a year
DSCI_FireAlertTimes	x_4	Interaction between DSCI and Wild Fires Alert Times
DSCI_hurricane	x_5	Interaction between DSCI and Hurricanes

DOI: 10.54254/2753-8818/51/2024CH0204

Table 1. (continued).

FireAlertTimes_hurricane	x_6	Interaction between Wild Fires Alert Times and hurricanes
DSCI_squared	x_7	Squared Drought Severity Index
FireAlertTimes_squared	x_8	Squared Number of Wildfires Alert Times
hurricane_squared	x_9	Squared Number of Hurricanes
Yield	Y	Cotton Yield (LB/Acre)

2.3. Method introduction

This paper employs a random forest model to evaluate the importance of each predictor and assess the accuracy of the model's predictions for cotton yield. Rigatti et al. stated that the random forest technique is a regression tree technique that uses bootstrap aggregation and randomization of predictors to achieve a high degree of predictive accuracy [10]. The key features of Random Forest include:

Bootstrap Aggregation: Each tree is trained on a random subset of data, reducing variance and mitigating overfitting.

Random Feature Selection: At each split, a random subset of features is considered, which enhances model diversity.

Performance Metrics: Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) are used to assess the model's performance.

The random forest model used in this paper can be mathematically represented as:

$$\hat{y} = \frac{1}{\text{ntree}} \sum_{t=1}^{T} h_t(\mathbf{x}; \text{mtry}) \quad (1)$$

Where $\hat{y}$ represents the predicted value the model generates, corresponding to the estimated cotton yield or other target variables derived from the input feature vector x. The parameter ntree specifies the number of decision trees utilized in the random forest. T indicates the total number of decision trees contributing to the final prediction. The function $h_t(\mathbf{x}; \text{mtry})$denotes the prediction made by the t-th decision tree based on the input vector x, with mtry defining the number of features randomly selected at each split. Vector **x** encompasses the input features, such as DSCI, wildfire alert times, and hurricane frequency, which are relevant to predicting cotton yield. mtry governs the randomness in feature selection during the tree construction process.

3. Results and discussion

3.1. Descriptive analysis

Table 2 presents the descriptive statistics for the study variables. The Drought Severity and Coverage Index (DSCI) and the number of wildfire alert times in a year (FireAlertTimes) exhibit substantial variability, with means of 99.588 and 1044.167, respectively, and standard deviations of 96.499 and 2099.923. These statistics highlight significant fluctuations in drought severity and wildfire alerts across the study period, with DSCI ranging from 1.654 to 405.558 and FireAlertTimes from 28 to 20232.

Hurricane occurrences were relatively rare, reflected by a low mean of 0.095 and a standard deviation of 0.294, indicating variability but infrequent hurricane events during the study period.

Cotton yield (Yield) values varied significantly throughout the study period, with a mean of 1026.280, a standard deviation of 301.510, and a range from 531 to 2025 LB/Acre. These data illustrate the differences in cotton yields that may result from differences in physical geography and climate, as well as differences in the degree of exposure to natural hazards, in different states spread across the southern United States, and provide an important basis for subsequent analyses (table 2).

DOI: 10.54254/2753-8818/51/2024CH0204

Table 2. Descriptive Statistics

Variable Name	Average	Std. Dev.	Minimum	Maximum
DSCI	99.588	96.499	1.654	405.558
FireAlertTimes	1044.167	2099.923	28.000	20232.000
Hurricane	0.095	0.294	0.000	1.000
Yield	1026.280	301.510	531.000	2025.000

Figure 1 illustrates the trends in cotton yield per acre from 2012 to 2023 across six major cotton-producing states. California consistently exhibits the highest yields, suggesting more favorable growing conditions or advanced agricultural practices than the other states. New Mexico and Oklahoma generally show lower yields, which may indicate less optimal conditions or challenges in cotton production. The observed fluctuations in yield across the states reflect the varying impacts of climatic and environmental factors over the years, with some states experiencing more stability while others show significant variability. These trends underscore the importance of region-specific strategies in managing cotton production effectively.

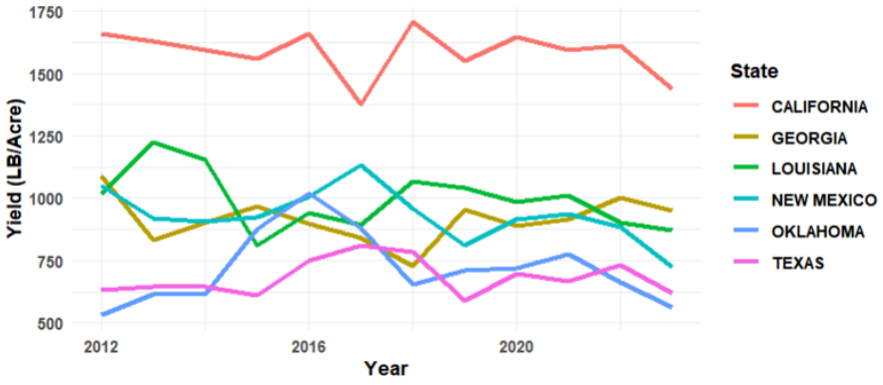

Figure 1. Cotton Yield Per Acre (2012-2023)

3.2. Feature importance analysis

Figure 2 shows that FireAlertTimes_squared, DSCI_FireAlertTimes, and FireAlertTimes are the most influential features in the random forest model, indicating that FireAlertTimes and DSCI indicators are critical for predicting cotton yields. In contrast, Hurricane and its related terms have much lower importance, suggesting they have a minimal impact on yield in comparison.

DOI: 10.54254/2753-8818/51/2024CH0204

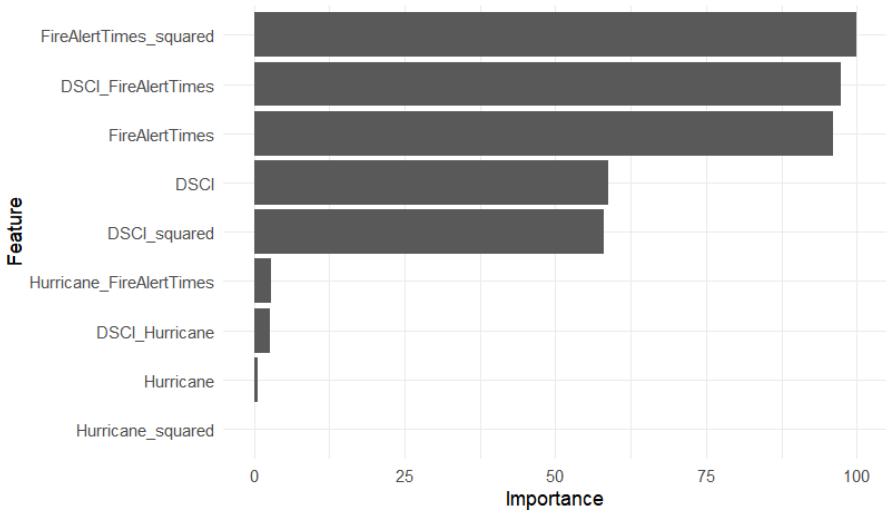

Figure 2. Feature Importance

3.3. Variable interaction strength analysis

Figure 3 illustrates the overall interaction strength between the key features affecting cotton yield in the random forest model. The interaction between DSCI and FireAlertTimes exhibits the highest interaction strength, indicating that the combination of drought severity and wildfire alerts plays a crucial role in influencing cotton yield. Other significant features include FireAlertTimes, FireAlertTimes_squared, and DSCI, highlighting the importance of both the individual and combined effects of drought and wildfire. In contrast, features related to hurricanes show much lower interaction strength, suggesting that hurricanes have a comparatively minor impact on cotton yield.

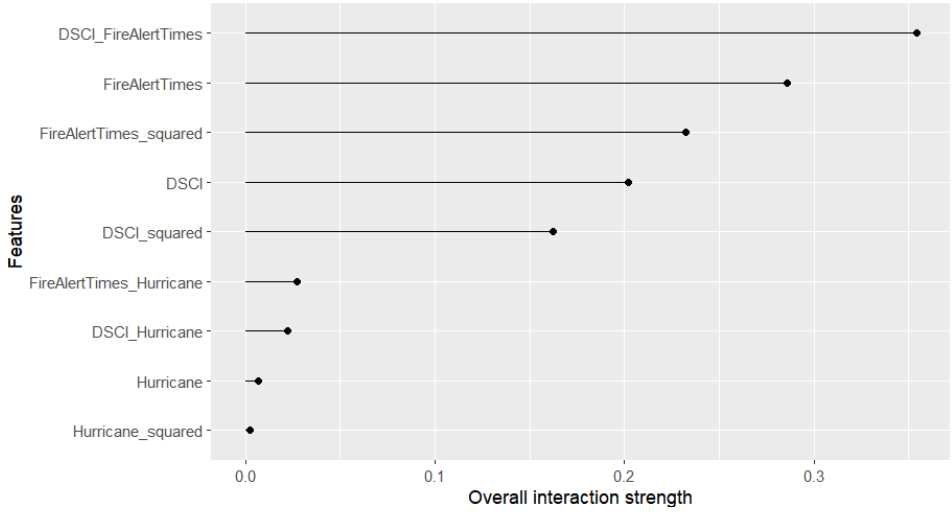

Figure 3. Variable interaction strength

3.4. Performance metrics

The performance metrics in Table 3 indicate that the random forest model demonstrates a strong predictive ability for cotton yields, with an R^2 of 0.6575 indicating that 65.75% of the variance in yields is explained by the model. The RMSE of 175.9177 indicated a reasonable mean error in prediction, with some variance that could not be accounted for by these three types of natural hazards and the interrelationships. This is in line with the objective reality that cotton yield is also influenced by other environmental stressors. Overall, the model performed well in determining the importance of drought, wildfires, and hurricanes on cotton yields.

DOI: 10.54254/2753-8818/51/2024CH0204

Table 3. Performance Metrics

	VALUE
MSE	30947.0260
RMSE	175.9177
RAE	134.4533
R^2	0.6575

3.5. Residuals analysis

Figure 4 shows a plot of residuals versus fitted values, revealing that although the Random Forest model was generally effective in predicting cotton yields, significant bias occurred at higher fitted values. The higher the fitted values, the more dispersed the residuals are, indicating the model's limitations in capturing extreme yield variations. This observation suggests that while the model remains robust, its accuracy in predicting extreme yields may be higher if more local production factors and comprehensive data are incorporated.

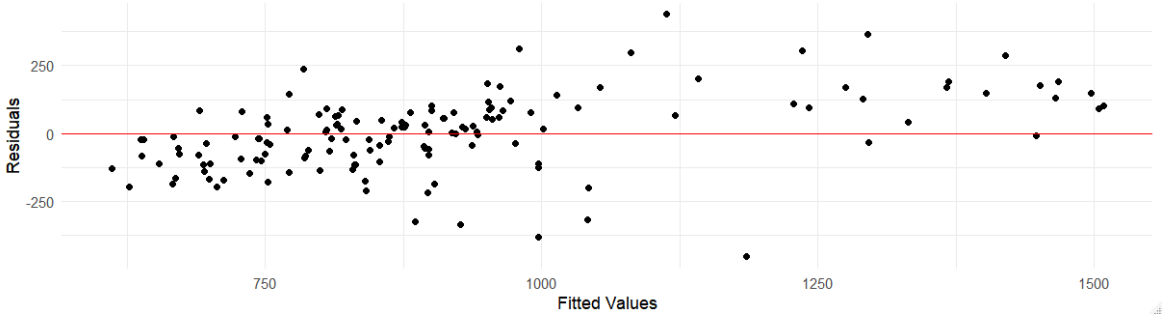

Figure 4. Residuals VS Fitted Values

Figure 5 provides the Q-Q plot of residuals. The Q-Q plot shows that the residuals of the random forest model are generally normally distributed, indicating that the model captures the overall data structure well. Deviations from extreme values indicate that the model has difficulty in predicting extreme values or outliers, which suggests that there are other factors affecting cotton yields in this case and that improvements need to be made to deal with these situations by appropriately incorporating local specifics.

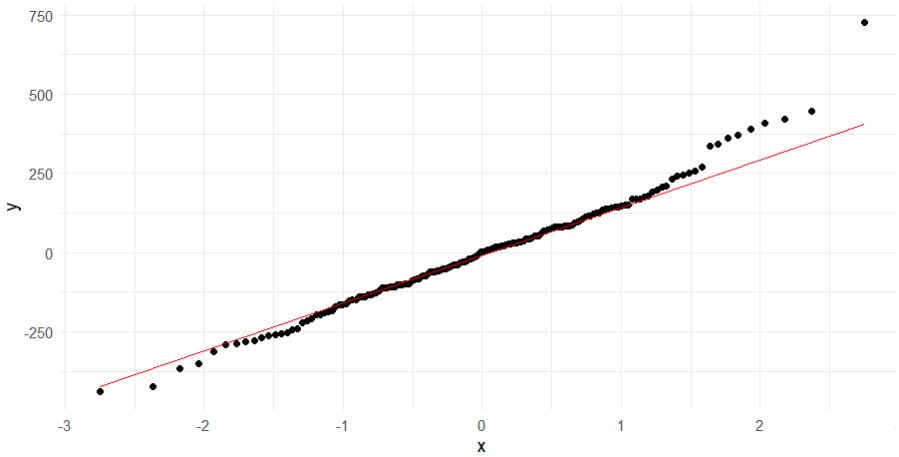

Figure 5. Normal Q-Q plot of residuals

DOI: 10.54254/2753-8818/51/2024CH0204

4. Conclusion

This paper delved into the extent to which drought, wildfires, and hurricanes affect cotton yield per acre in 14 major cotton-producing states across the United States from 2012 to 2023. To figure out more, this research considered the interaction effect of the three natural disasters and added interaction terms to the equation.

The low RMSE and high R^2 values of the random forest model demonstrate its robustness and excellent predictive accuracy, highlighting its ability to capture the intricate interactions and nonlinear correlations among the variables. Residual analysis further validated the reliability of the model. No significant pattern or bias was found in the fitted values for cotton acres ranging from 700 to 1000.

The analysis revealed that wildfires are the most critical factors affecting cotton yields, followed by the interaction between drought severity and wildfires. Importantly, hurricanes were found to have a negligible impact on cotton yields, providing a sense of relief from the potential damage caused by this natural disaster.

To mitigate the substantial impact of drought and wildfires on cotton production, local and state authorities must adopt and implement more effective management strategies, including advanced water management systems, modern irrigation technologies, and preventive measures such as firebreaks and controlled burns. Additionally, it is crucial to develop tailored insurance programs that address the specific risks posed by drought and wildfires, ensuring that farmers receive adequate financial protection and support in the event of crop damage. Emphasizing wildfire protection education, particularly in high-risk areas, is essential to reduce the likelihood of human-caused fires. Furthermore, ongoing investment in research and technology is necessary to improve prediction models and advance adaptive agricultural practices, such as developing drought-resistant cotton strains. States should also collaborate to share best practices and resources, potentially creating a joint fund for hazard response or implementing a coordinated water resource management strategy across state lines.

References

[1] Ridley W and Devadoss S 2023 Competition and trade policy in the World Cotton Market: Implications for us cotton exports. American Journal of Agricultural Economics, 105(5), 1365-1387.

[2] Starbird I R 1987 The U.S. cotton industry. U.S. Dept. of Agriculture, Economic Research Service.

[3] Jones L A and Durand D 1954 The cotton belt. In Mortgage lending experience in agriculture. Princeton University Press/NBER, 90-110.

[4] Ray R L, Fares A and Risch E 2018 Effects of drought on crop production and cropping areas in Texas. Agricultural & Environmental Letters, 3(1).

[5] González-Cabán A 2013 17 The Economic Dimension of Wildland Fires. Vegetation Fires and Global Change, 229.

[6] Behrer A P and Wang S 2022 Current benefits of wildfire smoke for yields in the US Midwest may dissipate by 2050. Policy Research Working Papers, 5-6.

[7] Klotzbach P J, et al. 2018 Continental U.S. hurricane landfall frequency and associated damage: Observations and future risks. Bulletin of the American Meteorological Society, 99(7), 1359-1376.

[8] Collins G, Edmisten K and Wright D 2021 Cotton producers guide. U.S. Department of Agriculture Forest Service, Southern Research Station.

[9] Pandya P A, et al. 2020 Effect of rainfall on productivity of Cotton. Emergent Life Sciences Research, 6(1), 56-63.

[10] Rigatti S J 2017 Random Forest. Journal of Insurance Medicine, 47(1), 31-39.