Received 11 May 2024, accepted 6 June 2024, date of publication 24 June 2024, date of current version 30 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3418139

RESEARCH ARTICLE

Cotton Yield Prediction: A Machine Learning Approach With Field and Synthetic Data

ALAKANANDA MITRA[1,2], (Member, IEEE), SAHILA BEEGUM[1,3], DAVID FLEISHER[1], VANGIMALLA R. REDDY[1], WENGUANG SUN[4], CHITTARANJAN RAY[3], DENNIS TIMLIN[1], AND ARINDAM MALAKAR[3,5]

[1]Adaptive Cropping Systems Laboratory, USDA-ARS, Beltsville, MD 20705, USA
[2]Nebraska Water Center, Institute of Agriculture and Natural Resources, University of Nebraska–Lincoln, Lincoln, NE 68588, USA
[3]Nebraska Water Center, Daugherty Water for Food Global Institute, University of Nebraska–Lincoln, Lincoln, NE 68588, USA
[4]Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA
[5]School of Natural Resources, University of Nebraska–Lincoln, Lincoln, NE 68583, USA

Corresponding author: Alakananda Mitra (Alakananda.Mitra@usda.gov)

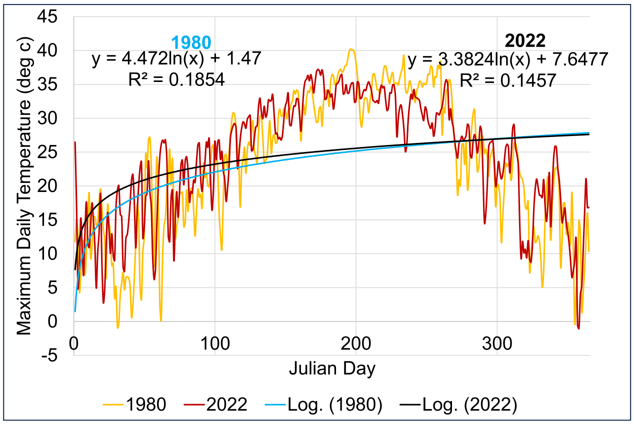
ABSTRACT The United States cotton industry is devoted to sustainable production strategies that reduce water, land, and energy consumption while enhancing soil health and cotton yield. Climate-smart agricultural solutions are being developed to increase yields and reduce operational costs. However, crop yield prediction is challenging because of the complex and nonlinear interactive effects of cultivar, soil type, management, pests and diseases, climate, and weather patterns on crops. To address this challenge, the machine learning (ML) method was used to predict yield, considering climatic change, soil diversity, cultivars, and fertilizer applications. Field data were collected over the southern US cotton belt in the 1980s and the 1990s. A second data source was generated from the process-based cotton model GOSSYM to reflect the most recent effects of climate change over the last six years (2017–2022). We focused on nine locations in three southern states: Texas, Mississippi, and Georgia. The accumulated heat for each set of experimental data was used as an analogue for the time-series weather data to reduce the number of computations. The Random Forest (RF) regressor, Support Vector Regression (SVR), Light Gradient Boosting Machine (LightGBM) regressor, Multiple Linear Regression (MLR), and neural networks were evaluated. Cross-validation was performed to obtain an improved model that did not suffer from overfitting. The RF regressor achieved an accuracy of 97.75%, with an R^2 of roughly 0.98 and a root mean square error of 55.05 kg/ha. The results demonstrate how a simple and robust model can be developed and utilized to help cotton climate-smart efforts.

INDEX TERMS Cotton yield prediction, climate change effect, smart agriculture, machine learning, field data, synthetic data.

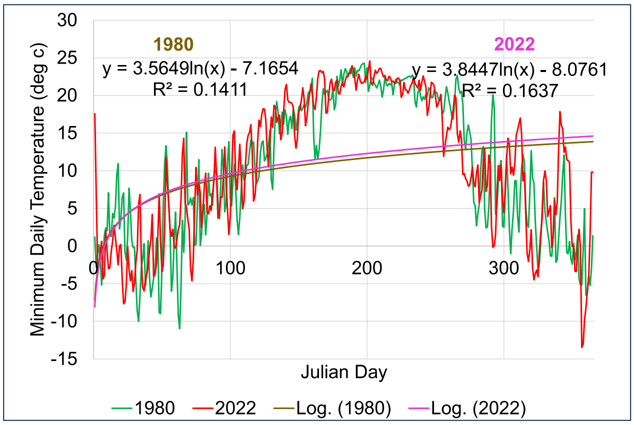
I. INTRODUCTION

Crop yield prediction is crucial for addressing food security challenges amid global climate change. Stakeholders, ranging from farmers to policymakers, emphasize the need for accurate and timely yield prediction [1]. Farmers can make more informed financial decisions and apply appropriate management strategies when accurate yield forecasts are available [2]. However, it can be challenging to precisely predict crop yields because of numerous variables such as crop-specific parameters, management strategies, cultivars, soil types, pests, and diseases. Thus climate change is a major factor. Agriculture is highly dependent on the weather and climate. Changing climate and weather patterns can severely affect crop yields [3], [4] and make them unpredictable. For example, Fig. 1(a) and 1(b) depict how daily maximum and minimum temperatures changed over 42 years, from 1980 to 2022 at a location (33.47° - 88.78°) in Mississippi, one major US cotton production site. Over 42 years, the temperature at this location increased by 0.8° C and the world's temperatures rose by at least 1° C. Increase in greenhouse gas levels and heat accumulation [5] have resulted in higher temperatures than pre-industrial levels. These factors affect crops in complex and nonlinear ways [6], [7], [8]. Hence, building a

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.



(a)



(b)

FIGURE 1. Daily (a) maximum and (b) minimum temperature variations in 1980 and 2022 at R.R. Foil Plant Science Research Center near Starkville, MS (33.47°, −88.78°). The average maximum temperature was 23.5° C and 24.3° C in 1980 and 2022, respectively, whereas the average minimum temperature was 10.4° C and 10.8° C in those years, respectively. The average maximum temperature increased by 0.8° C and the average minimum temperature increased by 0.4° C in the last 42 years.

robust, reliable, and accurate crop yield prediction model is not easy.

Typically, classical process-based crop growth models are based on the agronomic principles of plants and soils, management strategies, crop phenotypes, and weather variables and are used to predict crop yields [3]. It takes considerable effort and time to build such models and requires the availability of substantial data for calibration and validation of the models. Remote sensing-based approaches have also been employed for crop yield predictions. A wide range of devices, such as satellites, drones, LIDAR and RADAR sensors, and Internet-of-Things (IoT) field sensors are used for remote sensing [9], [10], [11]. Images and soil data collected through this approach were used to calculate various parameters (such as evapotranspiration, normalized difference vegetation index (NDVI), soil type, surface temperature (ST), soil moisture (SM), green vegetation index (GVI), enhanced vegetation index (EVI), temperature condition index (TCI), and vegetation condition index (VCI)) [12]. Indices such as NDVI, green normalized difference vegetation index (GNDVI), VCI, and TCI [12], [13], [14] are also used in statistical models [15], machine learning (ML) and DL-based models [16], [17] and hybrid models [18] to predict crop yield.

However, rapid growth in the IoT has changed this scenario. The proliferation of sensors generating huge amounts of data helped to emerge a new technology called "big data." Advancements in information and communication technologies (ICT), the hardware industry, and various computing platforms (cloud, fog, and edge) [19] enabled the data to be processed and computed more efficiently [20]. Advanced analytical tools such as ML technologies, provide a promising avenue to use these data more proficiently. The effectiveness of ML algorithms has already been demonstrated in the fields of healthcare [21], finance [22], multimedia forensics [23], [24], security and surveillance [25], retail [26], manufacturing [27], self-driving cars [28], virtual assistants [29], and plant/crop/fruits disease prediction [30], [31], [32], [33].

ML models are conceptually different from process-based ones. The advantages of applying ML algorithms are as follows:

- They are data-driven and learn patterns and relationships from a large dataset to predict the output. Therefore, tedious calculation or equation validation with data is not required. Once the input is selected, the remainder of the process is automatic.
- ML-based approaches are highly accurate when trained on high-quality datasets. Therefore, precise yield estimation can be achieved using this method.
- ML-based models are much simpler than process-based ones. Therefore, they are faster whereas process-based models take longer to run.
- As process-based models are computationally heavy they cannot be run on portable devices such as mobile phones. However, ML-based models enable mobile applications for crop models.
- ML models can replicate the nonlinear relationships between various inputs and yields accurately [34].
- Additionally, regularization techniques can make the model robust and generalized for noisy data [35].

Nonetheless, the accuracy of a model depends on various data quality aspects [20] such as volume, variety, meaningfulness, correctness, availability, and reliability. Therefore, an ML model can be developed to predict crop yields accurately when trained on a high-quality dataset. Hence, in the last several years, ML and DL-based approaches have been used extensively to predict crop yields. The ML and DL algorithms used for this approach include neural networks [36], random forests [1], support vector machines [37], convolutional neural networks [38], long short-term memory networks [17], autoencoder [39], faster-RCNN [40], etc. The United States (U.S.) cotton industry is devoted to sustainable production strategies that reduce water, land, and energy consumption while enhancing soil health and yield [41], [42]. Climate-smart agriculture solutions are being developed to increase yields and lower operational costs [43], [44], [45].

A. SOLUTION PROPOSED

Our research goal was to reliably predict cotton yield considering the effects of climate change, specifically high temperatures, in the U.S. southern cotton belt. We compared various ML-based and DNN-based methods for cotton yield prediction and used the best-performing method to evaluate yield estimates over multiple locations. Instead of 30 years of historical data, we focused on the last six years of meteorological data to include the climate change effect.

B. SIGNIFICANCE OF THE SOLUTION

- Because temperature is the largest driver among weather variables for plant growth and development, accounting for this relatively rapid climate change is important for incorporating an ML approach to estimate cotton yield. Our study addresses climate change and underscores its significance.
- Our study also showed how synthetic data use could apply state-of-the-art technology such as AI/ML in agriculture. The scarcity of publicly available datasets necessitates this for sustainable agriculture research.
- In this study, we calculated the accumulated heat from temperature variation during the season. This strategy simplifies the entire method and requires less computation. Consequently, it offers a more portable and edge-based solution.

The rest of the paper is organized as follows: Section II describes the methodology and experimental verification. The results are documented, discussed, and compared with existing works in Section III. Finally, Section IV summarizes how the results of this research move the knowledge base regarding ML applications to cotton yield forecasting and the importance of using synthetic data in agricultural research.

II. MATERIALS AND METHODS

Temperature is a key factor in cotton development. However, genetic traits can influence crop responses at different developmental stages. Cultivar responses to other environmental and field conditions, such as day length and water or nitrogen stress, vary across time and space, making the scenario dynamic and complex [46]. For example, Fig. 2 shows the variation in the weather elements at a representative site in Hockley, TX, at (33.51°, -102.5°) for the cotton season in 2022.

The primary focus of this study is to build an ML-based, robust cotton model that can precisely predict cotton yield while addressing the dynamic effects of these multiple inputs. Fig. 3 presents an overview of the proposed approach.

A. INPUT SELECTION

Weather [46], cultivar [47], soil type [48], and nitrogen [49] were selected as the input variables. The effects of rainfall, wind speed, and solar radiation were not considered because temperature is the most important weather element to influence cotton growth [44]. However, any variable from a weather dataset is time-series data, which demands more dynamic computation and increases the complexity of any crop model. To reduce complexity, we converted the time series temperature data to a scalar value, the accumulated heat, as described in Section II-D, without sacrificing model accuracy.

B. DATASET DETAILS

Two types of cotton yield data were used: field data and synthetic data. Field data were collected over the southern cotton belt in the 1980s and the early 1990s and archived with the Adaptive Cropping Systems Laboratory, USDA-ARS, Beltsville, Maryland. The dataset included multiple field plots, covering a range of soil types, cultivars, and nitrogen fertilizer concentrations. Plants were irrigated as determined by the farm managers to supplement precipitation to avoid water stress. Table 1 lists the details of the field dataset used in this study [50].

However, more rapid climate change has recently been observed worldwide. For example, the atmospheric CO_2 concentration ($\approx$ 337 ppm) in 1980 surged to 412 ppm in 2019 [51]. Therefore, the most recent effects of climate change over the past several years were not reflected in field data. In addition, more training data are needed to train the ML model without overfitting as ML models predict well when trained on a large diverse dataset [20], [52].

To address this issue, we generated synthetic yield data using a process-based cotton model called GOSSYM [53], [54], [55]. The development and applications of GOSSYM have been extensively documented in multiple scientific articles [54], [56], [57]. It is fundamentally a material balance model that simulates crop growth and development, carbon and nitrogen uptake, movement, and allocation in plants, as well as water and nitrogen in the soil. GOSSYM predicts crop responses to meteorological inputs [58], such as daily total solar radiation, maximum and minimum air temperatures, daily total wind speed, rainfall, fertilizer applications, and irrigation. Over the years, this model has undergone numerous enhancements and adjustments utilizing advanced concepts and insights acquired from experiments conducted in laboratory settings, field-scale scenarios, and controlled environments [56], [57]. The most recent iteration of GOSSYM involved enhancements in the soil, photosynthesis, and transpiration mechanisms [59]. The GOSSYM model incorporates 50 parameters related to weather, management, soil processes, and species- and cultivar-dependent characteristics. These parameters have been extensively discussed and documented in previous studies [56], [60].

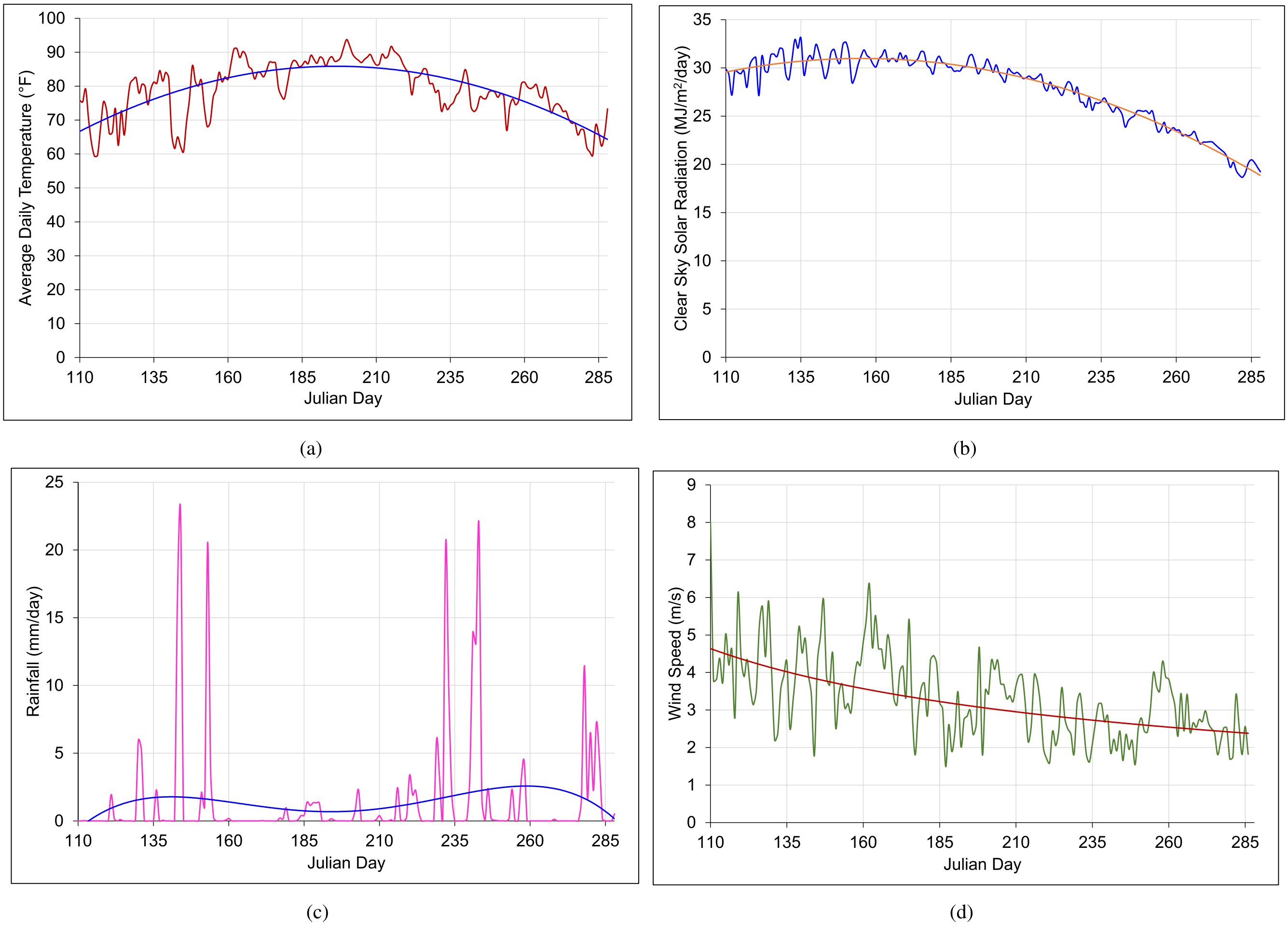


FIGURE 2. Variation in weather variables from Julian day 110 to 288 in 2022 at location Hockley, TX (33.51°, −102.52°). a) Average daily temperature vs Julian day, b) Clear sky solar radiation vs Julian day, c) Rainfall vs Julian day, and d) Wind speed vs Julian day.

GOSSYM was previously calibrated with cultivar parameters for 12 cotton cultivars, including those from the Delta, Acala, Stripper, and PIMA cultivar groups [56], [61]. These parameters were derived from multiple modeling studies of these cultivars. The cultivars used in this study were selected based on identical parameter sets. To include the most recent effects of climate change, the last six years, 2017-2022, were selected as the study period. The POWER Data Access Viewer [62] tool was used to download the daily weather data required to drive the GOSSYM model. We chose the first three leading cotton-producing states: Texas, Mississippi, and Georgia, based on their historical cotton production practices. We also selected three locations within each of the three states, as study areas (Table. 2).

Three different soil types, two cultivars, and four different amounts of applied nitrogen were selected for each location. This range of values was selected based on their presence in field data (Table 1) to ensure compatibility. Table 3 lists the input variations for the generated datasets. Sufficient irrigation was applied in addition to rainfall to avoid water stress. To make the data consistent with that of Table 1 we assumed that the plants were sown on May 1st and harvested on September 30th of each year [63], [64], [65], [66].

TABLE 1. Details of the field data obtained from experimental trials from seven states in the U.S. cotton belt from 1980 through the early 1990s.

Items	Details	Remarks
Study States	California, Texas, Missouri, New Mexico, Mississippi, Tennessee, Alabama	[63]
Number of Locations	48	Field Data
Number of Study Years	7	Field Data
Number of Soil Types	10	Field Data
Number of Cultivars	2	Field Data
Nitrogen Amount	0-300 kg/ha	Field Data

There were 48 instances of field data and 1296 instances of generated data. 1075 data samples (80%) were randomly selected and used for training and validation, and 269 data samples (20%) were used for testing purposes.

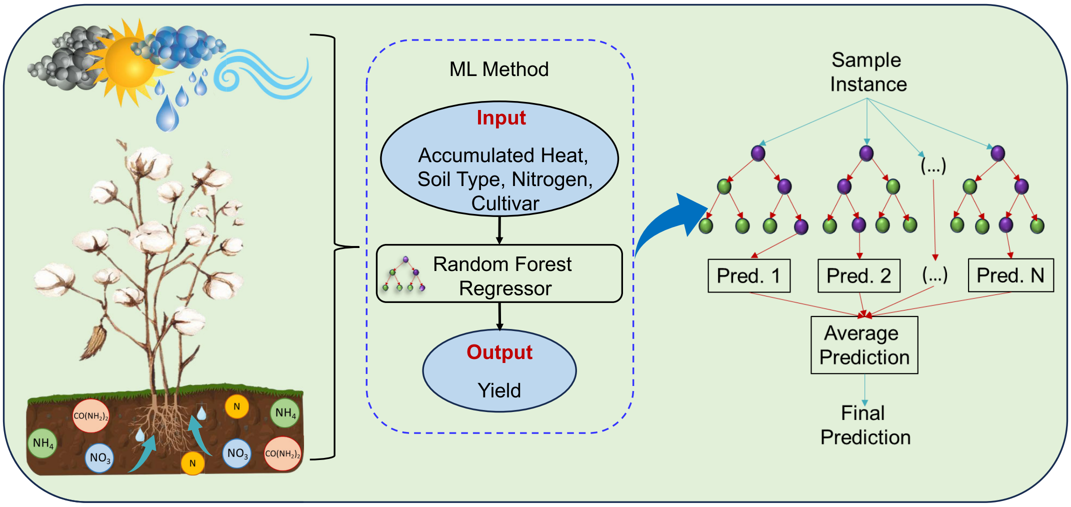


FIGURE 3. Overview of cotton yield prediction using random forest regressor.

TABLE 2. Details of the study area used for generating synthetic yield data using the GOSSYM cotton model.

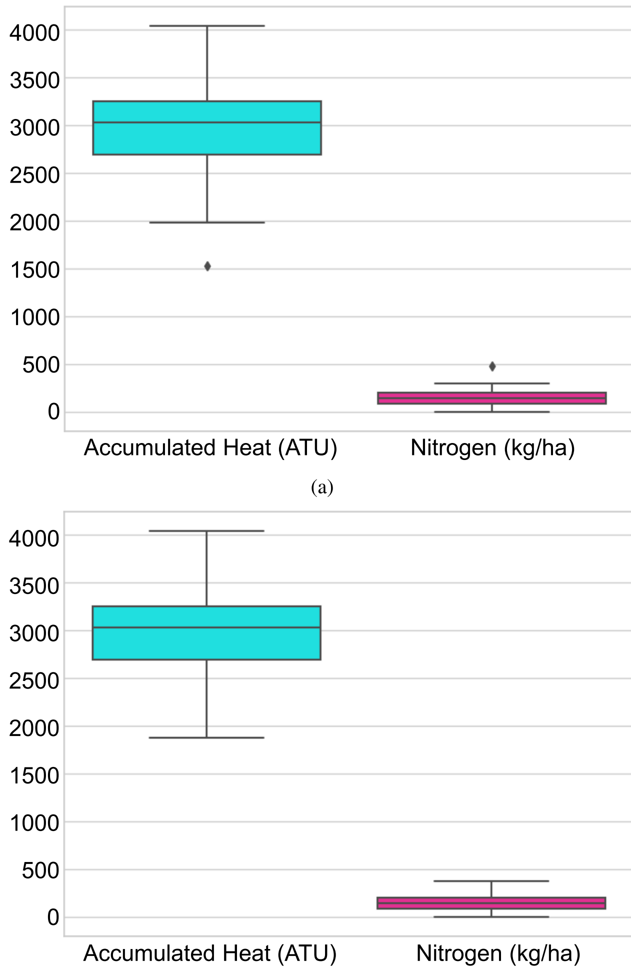
State	County	Lat.	Long.	Alt. (m)
Texas	Hockley	33.51	-102.52	1095.1
	Cameron	25.88	-97.40	5.48
	Calhoun	28.48	-96.62	4.27
Georgia	Bulloch	32.25	-81.74	56
	Mitchell	32.22	-84.46	44
	Dooly	32.14	-83.72	118
Mississippi	Starkville	33.28	-88.46	50
	Coahoma	34.26	-90.55	53
	Monroe	33.77	-88.67	83

TABLE 3. Details of the synthetic dataset generated through GOSSYM.

Study Location	Year	Cultivar	Soil	Nitrogen
Locations (From Table.2)	2017 - 2022	2 Upland Varieties (DPL90, NuCot33)	Clay, Sandy Loam, Sandy Clay Loam	4 Different values (0, 100, 200, 300 kg/ha) match the field data.

C. DATA FEATURE ENGINEERING

Data feature engineering and input data feature selection are crucial for the performance of an ML model. We applied two feature engineering techniques: transforming categorical data into numeric forms and removing outliers. Fig. 4(a) shows the data distribution for the accumulated heat and nitrogen inputs. The diamond shape (⧫) represents the outliers for each variable, which can have detrimental effects on the accuracy of ML methods. Hence, we removed all outliers from the inputs. Fig. 4(b) shows the data distribution of the inputs after removing outliers. There were two categorical inputs in the dataset: soil type and cultivar. These were changed to the representative numerical values.



(b)

FIGURE 4. Data distribution for the inputs: Accumulated Heat (ATU) and Nitrogen (kg/ha). (a) shows the outliers (shown in black diamond shape ⧫) present in the dataset for the inputs. (b) shows the data distribution after removing the outliers.

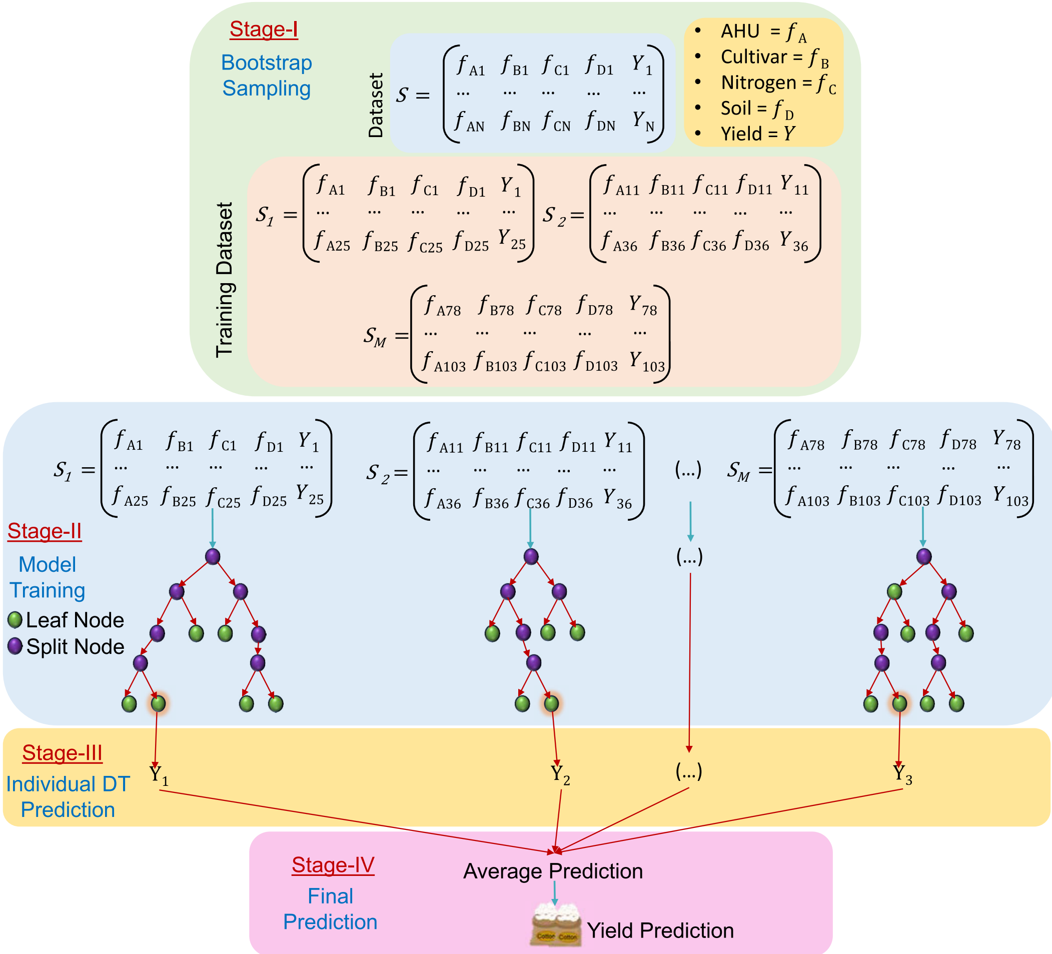


FIGURE 5. The process of how random forest regressor predicts the yield.

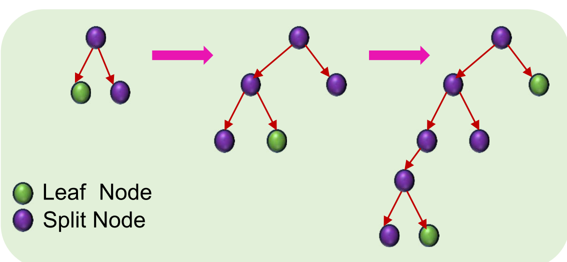


FIGURE 6. Leaf-wise expansion of decision tree for LightGBM.

D. ACCUMULATED HEAT CALCULATION

Cotton growth and the rate of development are primarily temperature-driven [46], [67]. Because cotton crops develop more slowly on days with cool temperatures than on days with warm temperatures, temperature measurements during the cropping season are frequently recorded on-farm or at nearby weather stations to help estimate when a crop reaches a particular developmental stage (Main). The heat unit for cotton, DD_{60}, indicates that the accumulated temperature effect occurs over the course of a day. DD_{60} was calculated by taking the daily average of the highest and lowest temperatures recorded in Fahrenheit ($^{\circ}F_{max}$ and $^{\circ}F_{min}$ respectively) as in Eq. 1.

$$DD_{60} = \frac{(^{\circ}F_{max} + ^{\circ}F_{min})}{2} - 60 \tag{1}$$

The total accumulated heat (AH) was calculated using Eq. 2 summing over the season.

$$AH = \sum_{n=1}^{N} DD_{60_n} \tag{2}$$

where $N = (Earliest\ Day\ of\ Harvest - Day\ of\ Sowing)$.

To calculate the accumulated heat for a past year, that weather data of a specific year is downloaded from POWER [62]; however, we can use 10 years of historical data for a future prediction.

E. MODELS

This study evaluated five different ML algorithms to determine the best-performing algorithm for cotton yield

```
Workflow
1. Load the dataset.
2. Remove the outliers from the
   input and output variables.
3. Transform the categorical values
   to numerical values.
4. Randomly divide the dataset into
   Training and Test parts.
5. Split the Training dataset into
   Train and Valid parts.
6. Select the ML model.
7. Select the loss function and
   optimizer.
8. Train the model.
9. Cross-validate the model.
10.Test the model with unseen data.
```

(a)

```
Algorithm-1
1. Download weather data.
2. For the past year, download the
   specific year's data.
3. For a future year, take the last
   10 year's average data.
4. Get the daily maximum and minimum
   temperatures.
5. Calculate Accumulated Heat.
6. Provide the soil type, cultivar,
   and amount of nitrogen to the
   trained model.
7. The model will predict the cotton
   yield.
```

(b)

FIGURE 7. (a) Workflow of cotton yield prediction model development. (b) Algorithm for cotton yield prediction.

prediction. As cotton yield is a nonlinear function of weather parameters, cultivar, soil, and fertilizers added, we selected several ML algorithms that work well with nonlinear functions.

Statisticians employ polynomial regression to model the nonlinearity between y and x using n^{th}-degree polynomial of x. The nonlinear connection between x and the conditional mean of y, $E(y|x)$, is fitted using a *polynomial regression*. *Polynomial regression* fits a nonlinear model well; however it is a linear statistical estimation problem because the unknown parameters are linear. When there is more than one independent variable, the problem becomes a linear statistical estimation problem that involves multiple variables. Therefore, *linear regression* is a subset of *polynomial regression*, which is regarded as a specific instance of *multiple linear regression* [68], [69], [70]. Additionally, in this case, inputs were not highly correlated. Hence, we used *multiple linear regression* (*MLR*) as the base method.

TABLE 4. Details of the loss functions of the models.

Model / Loss Function	Loss Function Expression
Random Forest / MSE [76]	$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$
LightGBM / Huber [77]	$L_\delta(a) = \begin{cases} \frac{1}{2}(\hat{y}_i - y_i)^2 & for\|\hat{y}_i - y_i\| \leq \delta, \\ \delta.(\|\hat{y}_i - y_i\| - \frac{1}{2}\delta), & otherwise. \end{cases}$
Support Vector Regressor / Epsilon Insensitive [78]	$L_\epsilon = \begin{cases} 0 & for\|\hat{y}_i - y_i\| \leq \epsilon, \\ \|\hat{y}_i - y_i\| - \epsilon, & otherwise. \end{cases}$
Multiple Linear Regression / MAE [79]	$MAE = \frac{1}{N}\sum_{i=1}^{N}\|\hat{y}_i - y_i\|$
Neural Network / MAE [79]	$MAE = \frac{1}{N}\sum_{i=1}^{N}\|\hat{y}_i - y_i\|$

$\hat{y}_i \rightarrow$ predicted value; $y_i \rightarrow$ true value
$N \rightarrow$ number of samples; $\delta \rightarrow$ threshold parameter
MAE $\rightarrow$ Mean Absolute Error; MSE $\rightarrow$ Mean Squared Error.

1) MULTIPLE LINEAR REGRESSION

As mentioned earlier *multiple linear regression* (*MLR*) is the same as a basic linear regression algorithm with multiple inputs. MLR fits the best-fit line in the data distribution. The MLR algorithm satisfies the following linear equation:

$$y = \sum_{n=1}^{N} \alpha_i x + \epsilon \tag{3}$$

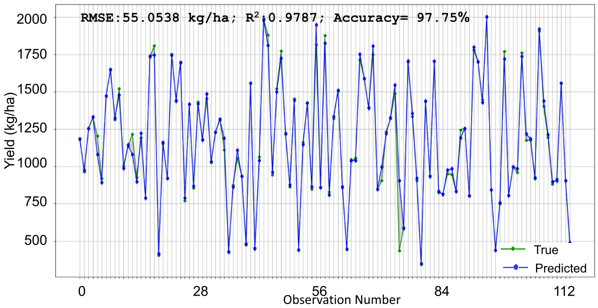
where α_i is the regression coefficient of the i-th independent variable, ϵ the model error, x is the input variable, N is the number of input variables, and y is the output variable.

2) SUPPORT VECTOR MACHINE

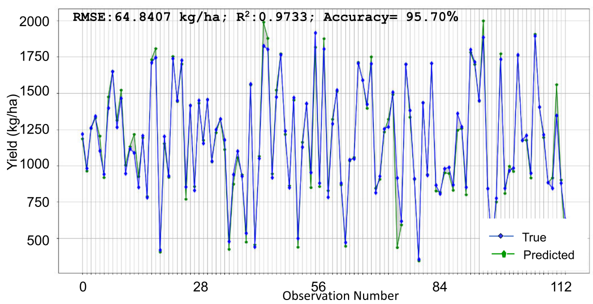
Support Vector Machine [71] is another supervised learning method that predicts the outcome by finding the correct hyperplane in the M-dimensional features space of N samples, maximizing the margin, and minimizing the prediction error. A *kernel* in an SVM maps the data into a higher-dimensional space, making it suitable for use in cases where there are nonlinear relationships between inputs and outputs. We used a support vector regression (SVR) algorithm to predict cotton yields.

TABLE 5. Performance metrics of four ML-type models for the cotton yield estimated from test dataset formed with field and simulated data.

Model	*RMSE* (kg/ha)	R^2	*NSE*	Accuracy (%)
Random Forest	**55.05**	**0.98**	**0.98**	**97.75**
LightGBM	64.84	0.97	0.97	95.70
SVR	341.15	0.18	NA	71.89
MLR	347.59	0.17	NA	69.13



(a)



(b)

FIGURE 8. Yield Prediction for test cases by (a) Random forest regressor (b) LightGBM regressor. Results show true and predicted yields for part of the test dataset. Most of the predicted yields comply with the true yields for both cases.

3) RANDOM FOREST

Random Forest [72] is a robust ML algorithm consisting of an ensemble of decision trees. It learns through supervised learning methods and uses a bootstrap aggregating (bagging) technique to determine the result. For N training samples, a Random Forest (RF) is built with N decision trees. Each unpruned decision tree uses slightly varied training samples, resulting in slightly different but overfitted performance. This makes the trees distinct, thereby reducing the forecast error or variance. Given the N samples and M features, the training samples of size n are repeatedly subsampled, where any sample may be present more than once and $n \leq N$. Additionally, a small subset m of features is selected from M features $m \leq M$. Consequently, using the bootstrap technique, each tree obtains n samples with m features. Finally, an average of each decision tree's results is used to generate the final prediction for the regression problem using aggregation. The branching techniques of RF regressors enable their use in nonlinear input-output relationships. Fig. 5 shows how the cotton yield was predicted using the RF algorithm.

4) LIGHTGBM

Light Gradient Boosting Machine (LightGBM) is a gradient boosting decision tree-based ensemble algorithm [73]. It expands leaf-wise (Fig. 6) instead of level-wise compared to other Decision Trees. It uses histogram-based and cost-effective algorithms because their time complexity is related to the number of bins but not to the data volume once histograms are created. Various boosting types can be used with LightGBM, for example, GBDT, DART, GOSS, etc.

5) NEURAL NETWORK

Our last ML model is an artificial neural network, specifically a multilayer perceptron (MLP). This feed-forward network was constructed with four input nodes and one output node to match the number of input features and outputs. The numbers of hidden layers and nodes in these layers varied. The ReLU activation function was used for the hidden layers, to include the nonlinearity of the inputs in the model. The mean absolute error loss function and Adam optimizer [74] were used.

F. YIELD PREDICTION

Fig. 7(a) depicts the workflow for model development and Fig. 7(b) shows the pseudocode for cotton yield prediction. The algorithms were trained using supervised learning. The best-performing algorithm was cross-validated 10-fold. We used a random grid search method to identify the most appropriate RF parameters from the estimator values: max depth, min samples split, min samples leaf, and bootstrap set to True and False. The best RF was obtained with 4000 estimators with bootstrapping True, and the maximum depth of the tree was 31. For the LightGBM regressor, the number of leaves was 100 with a maximum tree depth of 10, the boosting type was GBDT, and the metric was *Huber*.

For the linear regressor, a two-step Keras [75] sequential model with a normalizer as the first layer, and a linear layer with one output as the second layer was used with multiple inputs. A radial basis function kernel was used for the SVR algorithm. Various DNNs with varying numbers of layers were evaluated. Finally, the trained model was used for the yield prediction. Table 4 presents the various loss functions used in this study.

We implemented this work using different packages and frameworks such as scikit-learn, Numpy, Pandas, and Keras [75]. The models were trained and tested on an Intel Xeon server with 16 cores CPU, 64 GB RAM, and an NVIDIA RTX A4000 GPU.

G. PERFORMANCE METRICS

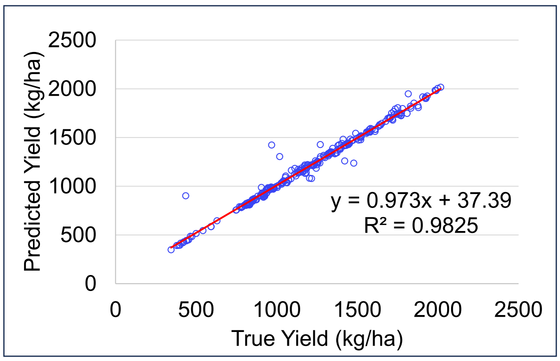
Several performance metrics have been calculated to evaluate the performance of the proposed method. The calculated metrics are *root mean square error* (*RMSE*), R^2, *Nash-Sutcliffe efficiency* (*NSE*), and *accuracy*. *RMSE* quantifies how well the regression line fits a data distribution. This is the average difference between the model predictions

TABLE 6. Performance metrics of the two Neural Networks for the cotton yield estimated from the test dataset formed with field and simulated data.

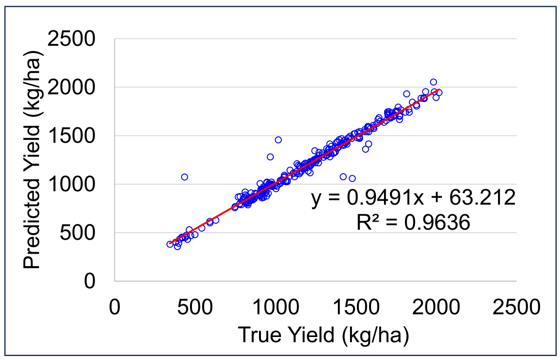
Network	Network Architecture	Number of Trainable Parameters	Accuracy(%)	RMSE (kg/ha)	R^2	Data Provided	Minimum Data Needed
NN1	FC (16, ReLU) FC (16, ReLU) FC (8, Sig) FC (4, Sig) FC (2, Sig) FC (1)	537	63.42	406.34	-0.0007	1075	5370
NN2	FC (8, ReLU) FC (4, ReLU) FC (2, ReLU) FC (1)	89	71.35	340.18	0.3000	1075	890

TABLE 7. Comparative analysis between the current results with the RF regressor and other related research associated with estimating cotton yields.

Crop	Method	Study Area	Study Period	*RMSE* (kg/ha)	R^2	Remarks
Cotton	MODIS + RF	5 locations in Maharashtra, India	2001-2017	62.77	0.69	[82]
Crop	Rainfall, pH, Temperature, Season, Crop, Nitrogen, and Electrical Conductivity + RF	537 locations in India	N.A.	-	0.882	[83]
Rice	Satellite Data+ DT + RF	1 location in India	1981-2030	281	0.67	[84]
Cotton	Spatial-Temporal M.T.L.	A 48-ha location in west TX, USA	2001-2003	83.7	-	[85]
Cotton	Spatial Temporal + RF + GBM	2 locations in New South Wales, Australia	2014, 2016, 2017	170 (RF), 190 (GBM)	0.44 0.39	[86]
Cotton	Accumulated Heat + RF	9 locations in USA	2017-2022	55.05	0.98	Current Work



(a)



(b)

FIGURE 9. Yield Prediction by a) Random Forest Regressor and b) LightGBM Regressor. Results show true versus predicted yields for the test dataset. Most of the predicted yields comply with the true yields.

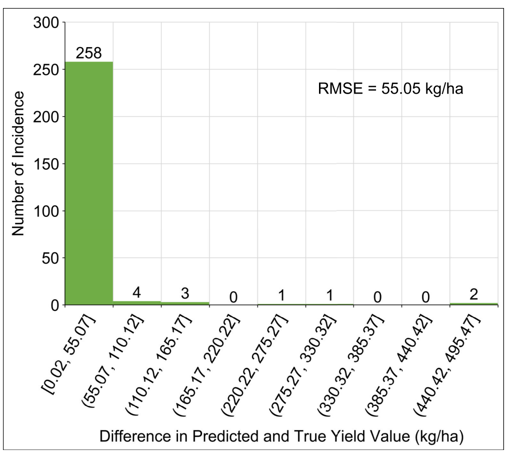


FIGURE 10. Performance histogram for RF regressor.

and the dataset values. A lower value of *RMSE* corresponds to a better-performing model. Eq. 4 is the measure of *RMSE*.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}} \tag{4}$$

where $\hat{y}_i$ is the predicted value for the i^{th} sample and the actual value is y_i. N is the total number of samples.

R^2 or R squared is coefficient of determination. This indicates the percentage of the dependent variable's variability accounted for by the regression of the independent variables. This indicates how well the model explains the variability in the observed data.

Another performance metric is the *Nash-Sutcliffe* efficiency (*NSE*). This normalized metric quantifies how well a model predicts based on the observed data. R^2 is expressed by Eq. 5 and *NSE* is expressed as in Eq. 6.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{5}$$

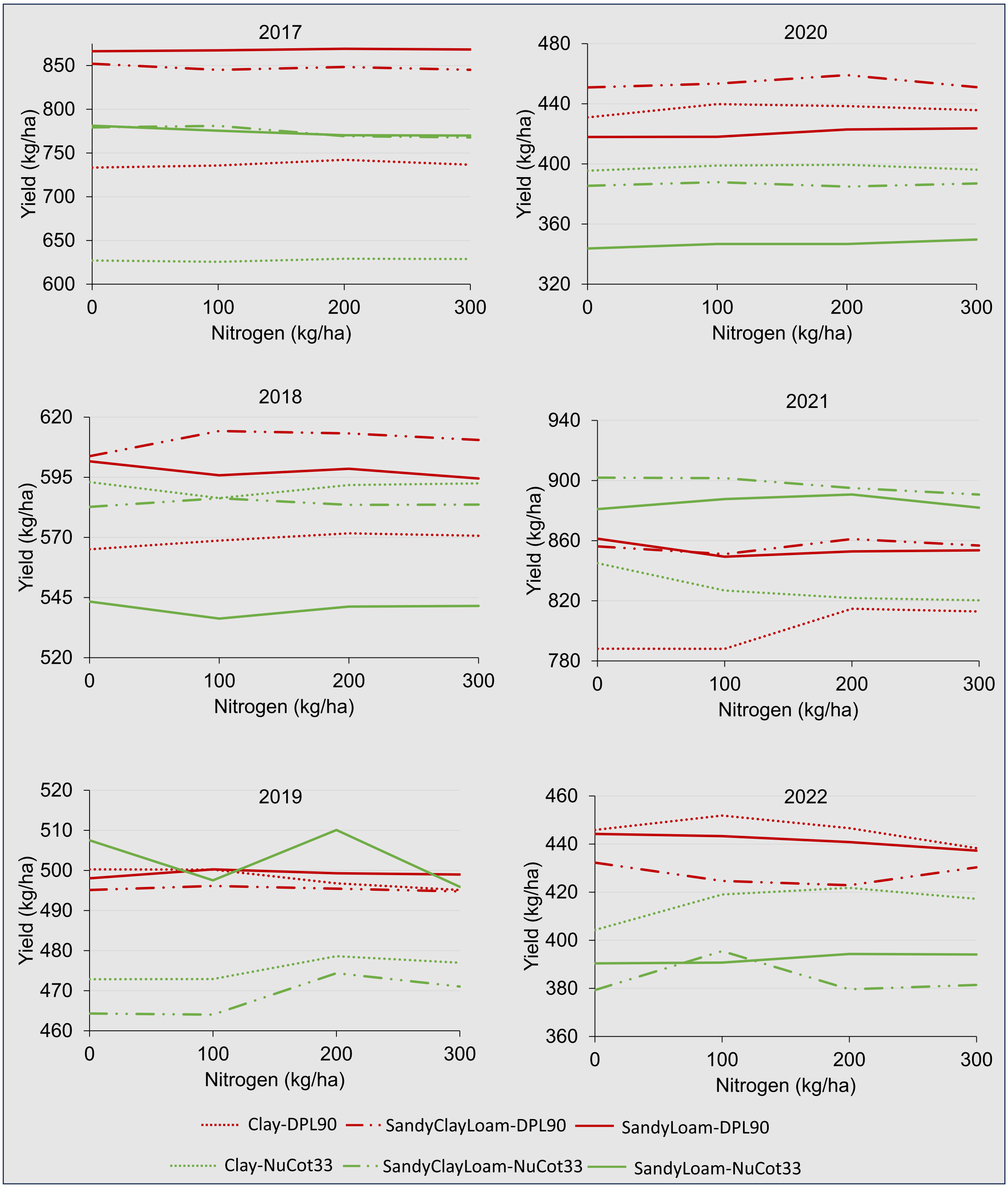


FIGURE 11. Yield vs Nitrogen plots for the years 2017-2022, and all three soil types and two cultivars form the Synthetic dataset for the location Hockley, TX.

$$NSE = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (6)$$

where $\hat{y}_i$ is the predicted value for the i^{th} sample, y_i is the actual value, $\bar{y}$ is the average y value, and N is the total number of samples. Although both metrics are expressed with similar formulas, they are used from different perspectives. NSE focuses on accuracy whereas R^2 indicates the fitness of the model.

III. RESULTS AND DISCUSSIONS

Initially, we experimented with five models and selected the best-performing model using existing data. We evaluated the models using the test dataset, which was kept aside before training. Table 5 lists the performance metrics derived from all the algorithms. Random Forest performed best among all models with 97.75% accuracy and a high R^2 of 0.98. The next-best-performing algorithm was LightGBM. It also had a high R^2 value of 0.97 and a high accuracy of 95.70%. Both algorithms achieved high NSE of 0.98 and 0.97 respectively. This proves that these two models are a good fit for the observed data. A perfect fit between the model and the observed data is represented by $NSE = 1$. When the observed mean is a better predictor than the model, then $\infty < NSE < 0$ is true, and a value of $NSE = 0$ means that the model predictions are equally accurate as the mean of the observed data [80], [81]. However, SVR and MLR did not perform well. An accuracy of only ≈70% was obtained with a very low R^2, and $RMSE$ values were nearly 5× compared with those of RF and LightGBM. The NSE was not calculated for these two algorithms.

The metrics for the neural network are shown in Table 6 which shows the two tested scenarios. Here, NN2 performed better than NN1 which performed randomly (a negative R^2 proves that) with the existing dataset. To train a neural network with considerable accuracy, additional data are required. the amount of data required depends on several factors: the complexity of the problem, the number of trainable parameters of the network, the number of input features, etc. [20]. The number of trainable parameters for NN2 was 89; therefore as a rule of thumb, a minimum of (89×10)=890 training samples were needed while assuming that the problem was not complex. However, for NN1, the minimum number of training data samples required was 5370, which was much higher than the provided training data samples of 1075.

We chose the two best-performing models, RF and LightGBM regressors, from the five models listed in Tables 5 and 6. The true and predicted yield plots for the RF and LightGBM regressors, respectively are shown in Fig. 8(a) and 8(b). The RF regressor predicted 95.91% of the test samples well below the $RMSE$ value of 55.05 kg/ha. The LightGBM predicted 88.10% of the test samples well below the $RMSE$ value of 64.84 kg/ha.

The predicted yield versus true yield is plotted in Fig. 9 where Fig. 9(a) depicts the yield plot for the RF regressor and Fig. 9(b) shows that of the LightGBM regressor. Although both the RF and LightGBM regressors performed well, the RF regressor performed marginally better than the LightGBM regressor. The deviation from the fitting line was less for the RF regressor than for the LightGBM regressor. Appendix describes the yield predicted by the RF model and the corresponding actual yield for the test dataset. Fig. 10 shows that among the 269 test samples, the prediction error of 258 samples was $\leq RMSE$ in the case of RF regressor.

Fig. 11 shows the yield vs nitrogen plots for 2017-2022, all three soil types, and two cultivars from the synthetic dataset for Hockley, TX. In the last six years, the maximum yield has been achieved when the nitrogen amount was 200 kg/ha. Table 7 compares our study with other literature sources that use the RF regressor as a predictive tool for simulating crop yields. These studies used a variety of data types as inputs, including satellite images, spatiotemporal data, and numerical data. RF performs well with noisy data, supporting its use as a popular ML algorithm for crop yield prediction.

In our study, the RF regressor was the best-performing ML algorithm for cotton yield prediction. However, another ML algorithm, LightGBM, also performed very well. This ensures the robustness of the method when using the accumulated heat to predict cotton yield and suggests an ML ensemble approach. Comparing the results in Table 7, our study was spatially and temporally diverse. More recently weather data have been used to address recent climate change effects. A very small part of the weather data was from earlier years to avoid bias in weather data. Hence, the proposed method was considerably more robust.

IV. CONCLUSION

In this study, we predicted cotton yields in U.S. locations with a high accuracy using simple ML approaches with accumulated heat information and an RF regressor. This study stands out by reducing the computational effort by converting time-series weather data to a scalar value, AH, without affecting the accuracy of the model. We introduced another alternative ML algorithm, the LightGBM regressor, which performs competitively with an RF regressor.

Machine and deep learning techniques perform well when trained on large datasets. However, it is not always possible to access large publicly available datasets, which hinders the application of ML/DL-based approaches. Synthetic data use in training AI/ML models is a common method in the case of data scarcity; however, it is not widely used in agriculture. Our study used generated and field data to build and test an AI-based cotton model. It demonstrates the potential use of synthetic data in training ML/DL models in agriculture.

This method is suitable for practical applications because of its simplicity compared to process-level model approaches. This method is also much simpler to use than process-based models, with less computational overhead in terms of eventually porting to mobile applications which can benefit farmers to use our application. However, our model has some limitations too:

- The method is likely limited to interpolating yields within the same range of the training data space as in any other supervised learning method. Retraining the model using data from new locations can address this limitation.
- It is built at the local level. More regional locations need to be included to obtain a regional-level model.
- However, we must add more locations and significantly more diverse soil, cultivar, and nitrogen variations to the training data to create a more global model.
- The goal of this study was to validate the proposed method. Because our field data were from the 1990s, we used the same cultivars to generate recent synthetic

TABLE 8. Test results for Random Forest Regressor.

True Yield	Predicted Yield	True Yield	Predicted Yield	True Yield	Predicted Yield	True Yield	Predicted Yield
1185.93	1181.58	472.91	484.56	915.83	926.62	1156.36	1190.09
963.14	974.9	1567.92	1556.14	1905.69	1917.96	1175.12	1164.82
1255.19	1254.62	1557.47	1557.67	1403.22	1440.58	1386.65	1425.24
1329.12	1331.1	453.4	450.33	1193.76	1212.09	1256.29	1242.28
1203.04	1082.9	1063.8	1038.23	881.95	898.2	1610.06	1563.15
918.68	892.39	1988.13	1982.04	915.72	905.22	1982.43	1986.05
1473.64	1473.49	1878.2	1808.9	1559.83	1555.16	1726	1767.85
1651.32	1646.84	942.7	961.04	901.57	904.92	1008.94	1008.27
1313.37	1323.74	1520.82	1499.71	464.32	486.91	1222.72	1221.83
1521.21	1478.53	1771.48	1730.03	1701.03	1701.87	1347.5	1389.24
1003.33	988.58	1215.8	1220.44	917.08	887.28	1851.15	1854.1
1131.49	1146.37	861.3	873.79	1922.35	1902.41	1171.65	1205.34
1215.62	1080.37	1451.46	1440.53	955.25	963.49	1579.99	1566.05
925.23	897.9	438.3	443.91	1332.77	1348.76	1216.46	1218.74
1191.4	1222.39	1160.81	1146.13	898.62	903.35	1018.08	1304.94
787.68	790.61	1426.58	1424.32	1743.06	1738.88	1716.03	1689.66
1731.67	1739.08	848.35	864.24	806.97	811.96	1557.47	1557.67
1808.18	1748.16	1815.6	1949.57	1248.79	1251.46	453.4	450.33
404.26	416.86	856.44	859.31	947.06	978.1	1063.8	1038.23
1150.86	1161.57	1875.97	1827.61	1423.87	1422.16	1988.13	1982.04
919.03	920.33	828.45	809.59	379.62	392.92	1878.2	1808.9
1752.48	1743.61	1320.72	1335.92	979.84	983.33	942.7	961.04
1442.5	1439.87	1512.02	1506.8	1290.84	1316.9	1134.21	1174.55
1700.05	1696.04	868.06	860.47	1643.23	1627.82	1270.94	1304.36
768.91	789.65	445.87	446.87	923.61	912.8	1581.7	1594.08
1417.19	1417.67	1034.36	1041.94	1556.77	1543.77	1090.39	1163.98
856.18	869.19	1056.85	1039.36	1697.14	1701.33	813.6	811.3
1430.43	1418.97	1712.62	1751.57	1933.92	1927.47	1219.93	1219.95
1174.7	1180.95	1590.54	1585.7	923.82	903.48	819.22	815.72
1456.16	1483.61	1396.02	1392.55	1719.4	1678.61	975.67	984.75
1026.84	1031.98	1750.13	1806.73	1611.6	1582.48	2017.4	2016.61
1228.43	1227.32	845.12	850.95	1248.65	1221.61	830.64	833.36
1320.58	1313.57	904.96	988.07	869.07	863.81	1338.08	1347.1
1112	1184.03	1232.88	1218.23	775.49	785.31	1110.78	1124.93
422.88	429.9	1320.89	1327.86	926.39	939.53	1175.96	1133.25
873.78	862.1	1489.07	1543.67	1572.88	1558.16	1118.82	1107.79
1054.98	1104.92	434.896	902.52	856.89	869.65	1269.8	1429.12
936.11	932.36	592.46	585.33	1228.79	1225.48	628.84	645.67
343.77	349.45	1700.26	1707.93	1421	1259.78	850.03	845.91
1434.59	1438.35	1335.1	1355.33	1150.82	1166.54	746.44	756.26
938.2	934.13	1579.64	1556.77	780.72	789.62	926.39	932.7
1706.05	1703.06	974.14	982.05	926.39	932.7	972.86	979.93
824.4	833.9	1832.28	1802	1072.93	1067.45	1047.96	1053.26
816.97	812.77	932.49	961.69	1522.74	1513.22	823.28	810.53
950.05	975.74	1053.46	1090.42	968.3	1424.01	1148.19	1146.25
945.34	985.78	1375.44	1370.71	1360.56	1334.21	926.39	932.7
830.41	832.6	779.3	780.92	909.54	932.07	1522.6	1517.96
1244.32	1184.69	978.1	970.43	1267.58	1281.14	1133.01	1135.96
1256.12	1251.66	390.71	393.23	1411.82	1482.69	1373.52	1343.89
799.61	806.21	1184.01	1188.08	1523.5	1511.06	1437.12	1438.99
1780.74	1797.38	1734.81	1792.64	1158.49	1149.9	984.68	982.02
1698.34	1703.34	810.92	824.52	904.19	904.53	1437.12	1438.99
1444.76	1427.8	1142.27	1166.58	1063.54	1038.19	1676.62	1667.82
1999.26	2004.67	1318.84	1335.55	398.91	395.49	1210.65	1214.8
842.33	843.69	1291.45	1304.47	1921.98	1914.22	1365.19	1340.31
438.48	439.79	1430.89	1424.28	1344.65	1351.06	1215.26	1220.24
749.55	757.4	1387.29	1384.67	591.77	585.36	1073.23	1063.71
1770.98	1724.02	944.77	963.58	1508.65	1475.16	1266.95	1238.62
808.93	804.06	1414.62	1413.84	1041.2	1056.06	1464.55	1466.89
995.98	997.53	1029.47	1021.32	1688.54	1696.68	1224.39	1210.66
959.91	985.55	936.61	929.8	1608.84	1585.06	591.77	585.36
1760.46	1734.66	1299.02	1294.53	1033.38	1036.84	1508.65	1475.16
1174.06	1218.15	1239.84	1261.96	1380.11	1417.36	1041.2	1056.06
1176.32	1188.7	1829.82	1821.97	1574.56	1582.97	1688.54	1696.68

data to maintain the parity. However, we need to retrain the model with the current data for application in today's field.

As part of future research, we will apply this method to new locations in the southern cotton belt in the U.S. and scale it up to a regional and national level. We will also work on extending the methods to other crops in the U.S. to establish the validity of the method.

We believe that the high accuracy of this study will encourage further research in agriculture to use synthetic data to develop AI-based crop models and reduce the gap between advanced technologies and the agricultural industry.

APPENDIX

See Table 8.

ACKNOWLEDGMENT

The authors would like to thank the Adaptive Cropping Systems Laboratory (USDA-ARS) for providing access to the field dataset. This initial work was submitted as an abstract at the American Geophysical Union Fall Meeting, in 2023.

REFERENCES

[1] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy, and S.-H. Kim, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, Jun. 2016, Art. no. e0156571.

[2] J. Ansarifar, L. Wang, and S. V. Archontoulis, "An interaction regression model for crop yield prediction," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Sep. 2021.

[3] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting," *Agricult. Syst.*, vol. 187, Feb. 2021, Art. no. 103016.

[4] R. A. Fischer, "Definitions and determination of crop yield, yield gaps, and of rates of change," *Field Crops Res.*, vol. 182, pp. 9–18, Oct. 2015.

[5] United Nations. (2022). *Confirmed as One of Warmest Years on Record: WMO*. Accessed: 2023. [Online]. Available: https://news.un.org/en/story/2023/01/1132387

[6] D. B. Lobell, M. Bänziger, C. Magorokosho, and B. Vivek, "Nonlinear heat effects on African maize as evidenced by historical yield trials," *Nature Climate Change*, vol. 1, no. 1, pp. 42–45, Apr. 2011.

[7] M. C. Broberg, P. Högy, Z. Feng, and H. Pleijel, "Effects of elevated CO_2 on wheat yield: Non-linear response and relation to site productivity," *Agronomy*, vol. 9, no. 5, p. 243, May 2019.

[8] K. Lamsal, G. N. Paudyal, and M. Saeed, "Model for assessing impact of salinity on soil water availability and crop yield," *Agricult. Water Manage.*, vol. 41, no. 1, pp. 57–70, Jun. 1999.

[9] A. M. Ali, M. Abouelghar, A. A. Belal, N. Saleh, M. Yones, A. I. Selim, M. E. S. Amin, A. Elwesemy, D. E. Kucher, S. Maginan, and I. Savin, "Crop yield prediction using multi sensors remote sensing (review article)," *Egyptian J. Remote Sens. Space Sci.*, vol. 25, no. 3, pp. 711–716, Dec. 2022.

[10] A. Hamza, M. A. Khan, S. U. Rehman, H. M. Albarakati, R. Alroobaea, A. M. Baqasah, M. Alhaisoni, and A. Masood, "An integrated parallel inner deep learning models information fusion with Bayesian optimization for land scene classification in satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9888–9903, 2023.

[11] A. Hamza, M. A. Khan, S. U. Rehman, M. Al-Khalidi, A. I. Alzahrani, N. Alalwan, and A. Masood, "A novel bottleneck residual and self-attention fusion-assisted architecture for land use recognition in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2995–3009, 2024.

[12] A. K. Prasad, L. Chai, R. P. Singh, and M. Kafatos, "Crop yield estimation model for Iowa using remote sensing and surface parameters," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 8, no. 1, pp. 26–33, Jan. 2006.

[13] Z. Ji, Y. Pan, X. Zhu, J. Wang, and Q. Li, "Prediction of crop yield using phenological information extracted from remote sensing vegetation index," *Sensors*, vol. 21, no. 4, p. 1406, Feb. 2021.

[14] A. Kayad, M. Sozzi, S. Gatto, F. Marinello, and F. Pirotti, "Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques," *Remote Sens.*, vol. 11, no. 23, p. 2873, Dec. 2019.

[15] E. Panek and D. Gozdowski, "Analysis of relationship between cereal yield and NDVI for selected regions of central Europe based on MODIS satellite data," *Remote Sens. Appl., Soc. Environ.*, vol. 17, Jan. 2020, Art. no. 100286.

[16] M. Galphade, N. More, A. Wagh, and V. Nikam, *Crop Yield Prediction Using Weather Data and NDVI Time Series Data*, vol. 106. Cham, Switzerland: Springer, 2022, pp. 261–271.

[17] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," *Sensors*, vol. 19, no. 20, p. 4363, Oct. 2019.

[18] P. Feng, B. Wang, D. L. Liu, C. Waters, D. Xiao, L. Shi, and Q. Yu, "Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique," *Agricult. Forest Meteorol.*, vols. 285–286, May 2020, Art. no. 107922.

[19] A. Mitra, S. P. Mohanty, and E. Kougianos, "ToEFL: A novel approach for training on edge in smart agriculture," in *Proc. Great Lakes Symp. VLSI 2024*, 2024, pp. 657–662, doi: 10.1145/3649476.3660381.

[20] A. Mitra, "Machine learning methods for data quality aspects in edge computing platforms," Ph.D. dissertation, Dept. Comput. Sci. Eng., Univ. North Texas, Denton, TX, USA, 2022.

[21] M. Nasr, Md. M. Islam, S. Shehata, F. Karray, and Y. Quintana, "Smart healthcare in the age of AI: Recent advances, challenges, and future prospects," *IEEE Access*, vol. 9, pp. 145248–145270, 2021.

[22] N. Nazareth and Y. V. Ramana Reddy, "Financial applications of machine learning: A literature review," *Expert Syst. Appl.*, vol. 219, Jun. 2023, Art. no. 119640.

[23] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A machine learning based approach for deepfake detection in social media through key video frame extraction," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–18, Apr. 2021, doi: 10.1007/s42979-021-00495-x.

[24] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "EasyDeep: An IoT friendly robust detection method for GAN generated deepfake images in social media," in *Internet of Things. Technology and Applications*. Cham, Switzerland: Springer, 2022, pp. 217–236.

[25] S. Feldstein, "Evaluating Europe's push to enact AI regulations: How will this influence global norms?" *Democratization*, vol. 17, pp. 1–18, Apr. 2023.

[26] A. Guha, D. Grewal, P. K. Kopalle, M. Haenlein, M. J. Schneider, H. Jung, R. Moustafa, D. R. Hegde, and G. Hawkins, "How artificial intelligence will affect the future of retailing," *J. Retailing*, vol. 97, no. 1, pp. 28–41, Mar. 2021.

[27] J. F. Arinez, Q. Chang, R. X. Gao, C. Xu, and J. Zhang, "Artificial intelligence in advanced manufacturing: Current status and future outlook," *J. Manuf. Sci. Eng.*, vol. 142, no. 11, Nov. 2020, Art. no. 110804.

[28] I. Yaqoob, L. U. Khan, S. M. A. Kazmi, M. Imran, N. Guizani, and C. S. Hong, "Autonomous driving cars in smart cities: Recent advances, requirements, and challenges," *IEEE Netw.*, vol. 34, no. 1, pp. 174–181, Jan. 2020.

[29] A. M. Garcıa-Serrano, P. Martınez, and J. Z. Hernández, "Using AI techniques to support advanced interaction capabilities in a virtual assistant for e-commerce," *Expert Syst. Appl.*, vol. 26, no. 3, pp. 413–426, Apr. 2004.

[30] U. Zahra, M. A. Khan, M. Alhaisoni, A. Alasiry, M. Marzougui, and A. Masood, "An integrated framework of two-stream deep learning models optimal information fusion for fruits disease recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3038–3052, 2024.

[31] A. Mitra, S. P. Mohanty, and E. Kougianos, "AGROdet: A novel framework for plant disease detection and leaf damage estimation," in *Internet of Things. IoT Through a Multi-disciplinary Perspective*. Cham, Switzerland: Springer, 2022, pp. 3–22.

[32] A. Mitra, S. P. Mohanty, and E. Kougianos, "AGRodet 2.0: An automated real-time approach for multiclass plant disease detection," *Social Netw. Comput. Sci.*, vol. 4, no. 5, p. 657, Aug. 2023, doi: 10.1007/s42979-023-02076-6.

[33] C. Dockendorf, A. Mitra, S. P. Mohanty, and E. Kougianos, "Lite-Agro: Exploring light-duty computing platforms for IoAT-edge AI in plant disease identification," in *Proc. IFIP Int. Internet Things Conf.*, 2023, pp. 371–380.

[34] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Comput. Electron. Agricult.*, vol. 151, pp. 61–69, Aug. 2018.

[35] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Cham, Switzerland: Springer, 2013.

[36] Y. Dahikar, D. Chikmurge, and S. Kharat, "Sketch captioning using LSTM and BiLSTM," in *Proc. Int. Conf. Netw., Multimedia Inf. Technol. (NMITCON)*, Bengaluru, India, 2023, pp. 1–9.

[37] N. Gandhi, L. Armstrong, O. Petkar, and A. Tripathy, *Rice Crop Yield Prediction in India Using Support Vector Machines*. Piscataway, NJ, USA: IEEE Press, 2016.

[38] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," *Frontiers Plant Sci.*, vol. 10, p. 1750, Jan. 2020.

[39] J.-W. Ma, C.-H. Nguyen, K. Lee, and J. Heo, "Regional-scale rice-yield estimation using stacked auto-encoder with climatic and MODIS data: A case study of South Korea," *Int. J. Remote Sens.*, vol. 40, no. 1, pp. 51–71, Jan. 2019.

[40] S. K. Behera, A. K. Rath, and P. K. Sethy, "Fruits yield estimation using faster R-CNN with MIoU," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 19043–19056, May 2021.

[41] (2023). *U.S. Cotton Has Always Been an Innovator in Sustainability*. Accessed: Oct. 24, 2023. [Online]. Available: https://www.cottonusa.org/sustainability

[42] (2022). *New National Report Demonstrates U.S. Cotton's Environmental Progress, Opportunities*. Accessed: Oct. 24, 2023. [Online]. Available: https://cottonleads.org/news/new-national-i-report-demonstrates-u-s-cottons-environmental-progress-opportunities/

[43] *USDA to Invest $1 Billion in Climate Smart Commodities, Expanding Markets, Strengthening Rural America*, USDA, 2022. [Online]. Available: https://www.usda.gov/media/press-releases/2022/02/07/usda-invest-1-billion-climate-smart-commodities-expanding-markets

[44] (2022). *NCC Welcomes USDA's Climate-Smart Pilot Projects*. Accessed: Oct. 24, 2023. [Online]. Available: https://www.cotton.org/news/releases/2022/climgr.cfm

[45] A. Mitra, S. P. Mohanty, and E. Kougianos, "Smart agriculture–demystified," in *Internet of Things. Advances in Information and Communication Technology*. Cham, Switzerland: Springer, 2024, pp. 405–411.

[46] K. R. Reddy, H. F. Hodges, and V. R. Reddy, "Temperature effects on cotton fruit retention," *Agronomy J.*, vol. 84, no. 1, pp. 26–30, Jan. 1992.

[47] R. R. Bridge and W. R. Meredith, "Comparative performance of obsolete and current cotton Cultivars[1]," *Crop Sci.*, vol. 23, no. 5, pp. 949–952, Sep. 1983.

[48] S. H. Moore, "Optimum soil-applied nitrogen levels for cotton on a high pH alluvial soil," *J. Plant Nutrition*, vol. 21, no. 6, pp. 1139–1144, Jun. 1998.

[49] M. Saleem, M. Bilal, M. Awais, M. Shahid, and S, "Effect of nitrogen on seed cotton yield and fiber qualities of cotton (*Gossypium hirsutum* L.) cultivars," *J. Anim. Plant Sci.*, vol. 20, no. 1, pp. 23–27, 2010.

[50] B. E. Howard, "Nonlinear system simulation," *Simulation*, vol. 7, no. 4, pp. 205–211, Oct. 1966.

[51] P. Riley, "Timeline of climate change," Britannica, 2022. [Online]. Availaible: https://www.britannica.com/story/timeline-of-climate-change

[52] R. Magar and A. Barati Farimani, "Learning from mistakes: Sampling strategies to efficiently train machine learning models for material property prediction," *Comput. Mater. Sci.*, vol. 224, May 2023, Art. no. 112167.

[53] D. N. Baker, J. D. Hesketh, and W. G. Duncan, "Simulation of growth and yield in cotton: I. gross photosynthesis, respiration, and Growth[1]," *Crop Sci.*, vol. 12, no. 4, pp. 431–435, Jul. 1972.

[54] D. Baker, J. Lambert, and J. Mckinion, "Gossym: A simulator of cotton crop growth and yield," South Carolina Agricult. Exp. Station, USA, Tech. Bull., 1983.

[55] F. Whisler, B. Acock, D. Baker, R. Fye, H. Hodges, J. Lambert, H. Lemmon, J. Mckinion, and V. Reddy, *Crop Simulation Models in Agronomic Systems*, vol. 40. Amsterdam, The Netherlands: Elsevier, 1986, pp. 141–208.

[56] M. Boone, D. Porter, and J. Mckinion, "Calibration of gossym: Theory and practice," *Comput. Electron. Agricult.*, vol. 9, no. 3, pp. 193–203, 1993.

[57] V. Reddy, K. Reddy, K. Sailaja, A. Richardson, V. Kakani, and D. Zhao, "Cotton modeling: Advances and gaps in our ability to assess climate change, crop management, economic and environmental policy decisions," in *Proc. World Cotton Research Conf.*, 2003, pp. 9–13.

[58] V. R. Reddy and D. N. Baker, "Application of GOSSYM to analysis of the effects of weather on cotton yields," *Agricult. Syst.*, vol. 32, no. 1, pp. 83–95, Jan. 1990.

[59] S. Beegum, D. Timlin, K. R. Reddy, V. Reddy, W. Sun, Z. Wang, D. Fleisher, and C. Ray, "Improving the cotton simulation model, GOSSYM, for soil, photosynthesis, and transpiration processes," *Sci. Rep.*, vol. 13, no. 1, p. 7314, May 2023.

[60] K. Reddy, H. Hodges, and J. Mckinion, "Crop modeling and applications: A cotton example," *Adv. Agronomy*, vol. 59, pp. 226–290, Oct. 1997.

[61] V. R. Reddy and D. N. Baker, "Estimation of parameters for the cotton simulation model GOSSYM: Cultivar differences," *Agricult. Syst.*, vol. 26, no. 2, pp. 111–122, Jan. 1988.

[62] NASA. (2023). *Power Data Acces Viewer*. [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/

[63] S. Mauget, M. Ulloa, and J. Dever, "Planting date effects on cotton lint yield and fiber quality in the U.S. southern high plains," *Agriculture*, vol. 9, no. 4, p. 82, Apr. 2019.

[64] W. T. Pettigrew, "Improved yield potential with an early planting cotton production system," *Agronomy J.*, vol. 94, no. 5, p. 997, 2002.

[65] T. P. Cooke and S. Cooke, "Behavior modification: Answers to some ethical issues," *Psychol. Schools*, vol. 11, no. 1, pp. 5–10, Jan. 1974.

[66] K. Reddy, P. Doma, L. Mearns, M. Boone, H. Hodges, A. Richardson, and V. Kakani, "Simulating the impacts of climate change on cotton production in the Mississippi delta," *Climate Res.*, vol. 22, pp. 271–281, 2002.

[67] K. R. Reddy, D. Brand, C. Wijewardana, and W. Gao, "Temperature effects on cotton seedling emergence, growth, and development," *Agronomy J.*, vol. 109, no. 4, pp. 1379–1387, Jul. 2017.

[68] J. Fox, "Linear models, problems," *Encyclopedia social Meas.*, vol. 2, pp. 515–522, Jul. 2005.

[69] D. Banks and S. Fienberg, "Statistics, multivariate," in *Encyclopedia of Physical Science and Technology*, Dec. 2003, pp. 851–889.

[70] B. Hidalgo and M. Goodman, "Multivariate or multivariable regression?" *Amer. J. Public Health*, vol. 103, no. 1, pp. 39–40, Jan. 2013.

[71] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[72] L. Breiman, M. Last, and J. Rice, *Random Forests: Finding Quasars*, vol. 45. Cham, Switzerland: Springer, 2001, pp. 243–254.

[73] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[74] Y. J. Kingma, "Design method for digital dead beat controllers," *Math. Comput. Simul.*, vol. 10, no. 2, pp. 76–79, Apr. 1968.

[75] F. Chollet et al., "Keras," 2015. [Online]. Available: https://keras.i

[76] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction To Statistical Learning: With Applications in Python*. Cham, Switzerland: Springer, 2023.

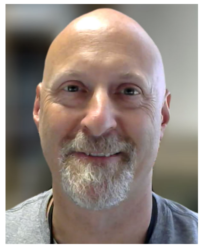
[77] P. J. Huber, "Robust estimation of a location parameter," in *Series in Statistics*. Cham, Switzerland: Springer, 1992, pp. 492–518.

[78] (2024). *Understanding Support Vector Machine Regression*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

[79] (2024). *Metrics and Scoring: Quantifying the Quality of Predictions*. Accessed: Feb. 12, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html
[80] R. E. Criss and W. E. Winston, "Do Nash values have value? Discussion and alternate proposals," *Hydrological Processes*, vol. 22, no. 14, pp. 2723–2725, Jul. 2008.
[81] D. N. Moriasi, M. W. Gitau, N. Pai, and P. Daggupati, "Hydrologic and water quality models: Performance measures and evaluation criteria," *Transactions ASABE*, vol. 58, no. 6, pp. 1763–1785, 2015.
[82] N. R. Prasad, N. R. Patel, and A. Danodia, "Crop yield prediction in cotton for regional level using random forest approach," *Spatial Inf. Res.*, vol. 29, no. 2, pp. 195–206, Apr. 2021.
[83] N. Prasath, J. Sreemathy, N. Krishnaraj, and P. Vigneshwaran, "Analysis of crop yield prediction using random forest regression model," in *Information Systems for Intelligent Systems. Smart Innovation, Systems, and Technologies*, vol. 324. Springer, 2023.
[84] R. Bhatnagar and G. Gohain, "Crop yield estimation using decision trees and random forest," in *Machine Learning and Data Mining in Aerospace Technology. Studies in Computational Intelligence*, vol. 836. Springer, 2020.
[85] L. H. Nguyen, J. Zhu, Z. Lin, H. Du, Z. Yang, W. Guo, and F. Jin, "Spatial-temporal multi-task learning for within-field cotton yield prediction," in *Advances in Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer, 2019, pp. 343–354.
[86] S. Leo, M. De Antoni Migliorati, and P. R. Grace, "Predicting within-field cotton yields using publicly available datasets and machine learning," *Agronomy J.*, vol. 113, no. 2, pp. 1150–1163, Mar. 2021.

ALAKANANDA MITRA (Member, IEEE) received the B.Sc. degree (Hons.) in physics from the Presidency College, University of Calcutta, in 2001, the B.Tech. and M.Tech. degrees in radiophysics and electronics from the Institute of Radio Physics and Electronics, University of Calcutta, in 2004 and 2006, respectively, and the Ph.D. degree in computer science and engineering from the University of North Texas, Denton, TX, USA, in 2022. She is currently a Research Assistant Professor with Nebraska Water Center, Institute of Agriculture and Natural Resources, University of Nebraska–Lincoln, Lincoln, NE, USA. Since March 2023, she has been a Visiting Computer Scientist with the USDA-ARS Adaptive Cropping Systems Laboratory, Beltsville Agricultural Research Center, Beltsville, MD, USA. She is working on AI-based crop models, tinyML devices for plant disease detection, and the application of federated learning in smart agriculture. She is also working on a project to develop crop and soil simulation models, graphical user interfaces, databases, and other suitable agro-climatology modeling tools. Her research interests include artificial intelligence, machine learning, deep learning, computer vision, and edge AI in smart agriculture and multimedia forensics. She received numerous academic awards, honors, and travel grants throughout her academic career. During the Ph.D. research, she received an Outstanding Early-Stage Doctoral Student Award. She has also received several best paper awards. She has one U.S. patent (pending) and one U.S. provisional patent.

SAHILA BEEGUM received the bachelor's degree in civil engineering and the master's and Ph.D. degrees in hydraulics and water resources engineering from Indian Institute of Technology Madras, India. She is currently a Research Assistant Professor with Nebraska Water Center, University of Nebraska–Lincoln. In collaboration with the Adaptive Cropping Systems Laboratory, USDA-ARS, Beltsville, MD, USA, she contributes to several projects focused on soil-plant-atmospheric continuum research and the development of process-based models. She is dedicated to developing and refining the mathematical representation of interactions between soil, crops, and the atmosphere. Her expertise spans soil, groundwater, and mechanistic-process-based crop modeling.

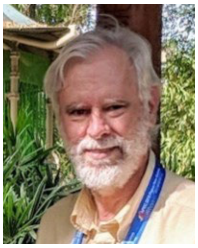


DAVID FLEISHER received the interdisciplinary Ph.D. degree in plant biology and mechanical engineering from Rutgers University, in 2001. He is currently a Lead Scientist and a Research Agricultural Engineer with the Adaptive Cropping Systems Laboratory, USDA-Agricultural Research Service, Beltsville, MD, USA. His research consists of over 25 years of controlled environment and field experiments associated with climate stress on crops including corn, cotton, potato, rice, soybean, strawberry, and wheat. These experimental studies connect with the development of novel conceptual and mathematical models to explain how factors influencing plant production at one level of hierarchy scale up to the entire plant and larger spatial scales. He also leads a team of scientists in the development, testing, and improvement of process-level crop models which he integrated with geospatial methods to evaluate local and regional climate impacts and adaptation options. He is a fellow of the American Society of Agronomy, in 2019. He is internationally recognized for his expertise in agricultural systems modeling, quantification of abiotic stress on crop growth and development, and regional food security studies.

VANGIMALLA R. REDDY is currently the Research Leader and a Supervisory Plant Physiologist with the Adaptive Cropping Systems Laboratory (ACSL), USDA-ARS, Beltsville, MD, USA. Over the years, he served in various professional and administrative positions, as the Acting Associate Director for NEA, BARC, and ANRI, and as a Beltsville Area Representative on the RL Advisory Council. He has authored over 180 publications in peer-reviewed journals, several book chapters, and books. His research interests include crop responses to climate change, especially processes like photosynthesis, respiration, transpiration, carbon and nitrogen metabolism, and growth analysis of cotton, soybean, and other crops. He is a fellow of the American Society of Agronomy and the Crop Science Society of America, and a member of several editorial boards of the *International Scientific Journal*. He recently served as a member of the Scientific Advisory Board (SAB), Organization for Economic Co-operation and Development (OECD), Paris, France.

WENGUANG SUN received the M.S. degree in environment science from the University of the Chinese Academy of Sciences, China, and the Ph.D. degree in agronomy from Louisiana State University, Baton Rouge, LA, USA. He is currently a Research Scientist with Colorado State University. Previously, he was a Postdoctoral Fellow with the Adaptive Cropping Systems Laboratory, USDA-ARS. His research interests include crop modeling and climate change.

CHITTARANJAN RAY is currently a Professor of civil and environmental engineering with the University of Nebraska–Lincoln, Lincoln, NE, USA, and the Director of Nebraska Water Center, University of Nebraska. He is also the Director of the university's Environmental Center and the Chief Environmental Engineer with the Applied Research Laboratory, a U.S. Navy-sponsored facility at the University of Hawai'i at Mānoa. He previously served as the Interim Director of the Water Resources Research Center, University of Hawai'i at Mānoa. He has held positions in industry and Illinois State Water Survey. He has extensive experience in many facets of managing both water quantity and water quality issues, particularly in the areas of chemical and pathogen impacts on groundwater quality; flow and transport processes in the vadose zone, technologies for low-cost water supply, and the agriculture-water/energy nexus.



DENNIS TIMLIN received the B.A. degree in biology from The State University of New York at Buffalo (SUNY-Buffalo), Buffalo, NY, USA, in 1974, and the M.S. and Ph.D. degrees in soil physics from Cornell University, in 1987. He is currently a Research Soil Scientist with the Adaptive Cropping Systems Laboratory, Beltsville Agricultural Research Center, USDA-ARS, Beltsville, MD, USA. Since 1991, he has been at Beltsville. His program is directed toward quantifying the effects of environmental variables on crop growth and soil processes. His interests are in how plants and their environment (soil, temperature, and carbon dioxide) interact, how to quantify that interaction in simulation models, and developing computer-based tools for scientists and growers. His experimental research utilizes sunlit growth chambers to study carbon assimilation, growth, and development of plants. Some of the models that he has worked on include 2DSOIL and GLYCIM and two new models for maize (MAIZSIM) and potato (SPUDSIM). His current research interests include the development of plant and soil simulation models for use in climate change and agricultural management. He is a fellow of the American Society of Agronomy and the Soil Science Society of America.

ARINDAM MALAKAR received the Ph.D. degree, in April 2017. He is currently a Faculty Member with the University of Nebraska–Lincoln, where he leads the Critical Zone Research Group. His academic home is with the School of Natural Resources, University of Nebraska–Lincoln. He works with the local stakeholders and leads research efforts to understand the impact of land surface processes influenced by agricultural activities on the state groundwater quality and validate and verify next-generation cropping system models. He was a recipient of the prestigious Water Advanced Research and Innovation (WARI) Internship Program, supported by the Department of Science and Technology, Government of India, the University of Nebraska–Lincoln, the Daugherty Water for Food Institute (DWFI), and the Indo-U.S. Science and Technology Forum (IUSSTF). He has also received the Young Scientist Award from the Materials Research Society of India for his excellent research in improving water quality.