

NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Chapter 11

Coupling AquaCrop and machine learning approaches for cotton yield simulation

Lisa Umutoni and Vidya Samadi

Department of Agricultural Sciences, Clemson University, Clemson, SC, United States

1 Introduction

Accurate seasonal crop yield estimation is one of the most important pieces of information that stakeholders such as farmers, commodity merchants, and government officials can have at their disposal to make strategic choices in their respective positions. However, crop yield prediction is one of the challenging problems in precision agriculture (Bali and Singla, 2022; Basso and Liu, 2019). This problem requires the use of several datasets and analogies since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety (Xu et al., 2019). This indicates that crop yield prediction is not a trivial task; instead, it consists of several intricate steps. To effectively forecast crop production, prior knowledge of the relationship between functional features and interacting variables is needed. To investigate such correlations, large datasets that can be obtained from farm systems, and intelligent algorithms that machine learning (ML) can provide are needed (Murugantham et al., 2022).

ML, which is a branch of artificial intelligence (AI), is a data-driven approach that can provide better yield prediction based on several features. ML can determine patterns and correlations and discover knowledge from datasets. Process-based agricultural models are developed to simulate the crop growth under various environment and management conditions, some of which include irrigation scheduling modules. With the help of calibrated process-based agricultural models from several field treatments, it is possible to generate as many training samples as ML algorithms need.

Intelligence Systems for Earth, Environmental and Planetary Sciences.

<https://doi.org/10.1016/B978-0-443-13293-3.00007-5>

Copyright © 2024 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

ML and simulation crop modeling advancements have offered new approaches to enhance agricultural forecast (Archontoulis et al., 2020; Bogard et al., 2020; Ersoz et al., 2020; Washburn et al., 2020). These technologies have each brought distinct capabilities and considerable improvements in prediction performance; however, they have primarily been evaluated independently, and there may be benefits to combining them to boost more the prediction accuracy (Daw et al., 2022). Crop simulation models estimate yield, flowering time, and water stress employing management, crop cultivar, and environmental inputs, as well as science-based crop physiology, hydrology, and soil C and N cycle equations (Asseng et al., 2014; Basso and Liu, 2019; Shahhosseini et al., 2019). These crop models are pretrained utilizing a varied collection of experimental data from various conditions and are further enhanced (calibrated) in each research for more accurate forecasts (Ahmed et al., 2016; Gaydon et al., 2017). ML, on the other hand, seeks to create predictions by establishing links between input and response variables. ML comprises techniques in which the system learns a transfer function to forecast the intended output based on the available inputs, as opposed to a user providing the transfer function. Furthermore, it is simpler to apply than simulated crop models since calibrating the model does not require expert knowledge or user abilities, has shorter runtimes, and less data storage limits (Shahhosseini et al., 2019).

This study aimed at testing whether ML algorithms including multilayer perceptron (MLP), Random Forest (RF), and gated recurrent unit (GRU) could precisely estimate daily cotton yield during a growing season by learning from data generated by a calibrated process-based agricultural model (AquaCrop) and thereby serve as an appropriate tool for irrigation scheduling to optimize yield and help farmers in planning in advance for equipment, labor, fertilizers, or energy needs to better manage farm operations.

2 Materials and methods

2.1 Data collection and study site

This study was conducted at the Edisto Research and Education Center (EREC) of Clemson University located in South Carolina, USA. The region is categorized as a humid subtropical climatic region with mild winters and hot summers. Climate data including precipitation, relative humidity, maximum and minimum temperature, and solar radiation starting from April 25, 2003 to September 22, 2021 recorded by a National Oceanic and Atmospheric Administration (NOAA) station. The variability of climate data over time is illustrated Fig. 1. The data were then preprocessed to fill any gaps and mismatch using the forward filling method. This method uses the nearest past observed value to fill the gap of the following missing value.

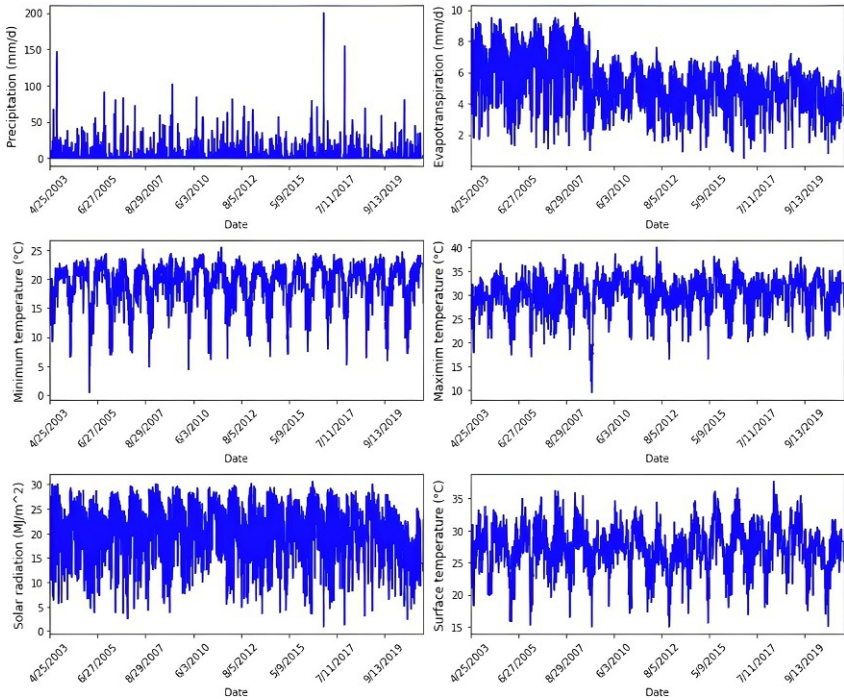


FIG. 1 The variability of climatic data in the study area.

2.2 AquaCrop

2.2.1 Description of AquaCrop

AquaCrop is a process-based model that simulates the attainable yield of crops under different environmental conditions and management practices by employing few conservative crop parameters and a limited number of input variables (Vanuytrecht et al., 2014). AquaCrop can be used to predict crop yield at the farm or regional scale. The model does this through related components of the soil-crop-atmosphere continuum. The soil is subdivided into horizons of variable depths and hydraulic properties such as hydraulic conductivity at saturation (k_{sat}), field capacity (FC), permanent wilting point (PWP), and drainage coefficient (τ). The crop grows by expanding the canopy and accumulating biomass in daily time steps. As it matures, it responds to water changes through four stress coefficients, namely, leaf expansion, canopy senescence, stomata closure, and change in harvest index (Steduto et al., 2009).

AquaCrop uses four steps to simulate crop growth. The first step is the crop development simulation; first-order kinetics equations are employed to describe

this process. The initial canopy cover after emergence (CC_o), the maximum canopy cover reached (CC_x), and the canopy growth coefficient (CGC) parameters are used for simulations of canopy development in nonlimiting conditions (Hsiao et al., 2009). The decrease in canopy cover (CC) is expressed with a canopy decline coefficient (CDC). In case of water stress, the stress coefficient for water expansion ($K_{s_{exp}}$) starts to decrease to reduce the CGC, thus slowing down canopy development. On the other hand, when water depletion in the root zone reaches the maximum level for accelerated canopy senescence, CDC is increased by $K_{s_{sen}}$, causing a rapid decline of the canopy (Raes et al., 2009). The second step involves crop transpiration simulation, which is computed by multiplying ET_o by the crop coefficient ($K_{c_{Tr}}$). Third, the daily above ground biomass (B) production is simulated using the normalized water productivity (WP^*). Last, crop yield is calculated by multiplying the final biomass (B) with the harvest index (HI) (Steduto et al., 2009). For an in-depth description of these steps, refer to Steduto et al. (2009) and Vanuytrecht et al. (2014).

2.2.2 Model development

AquaCrop uses climate, crop, management, and soil data for yield simulation. The climate variables during simulated growth seasons were defined by five daily weather variables. Those are maximum temperature, minimum air temperatures, precipitation, reference evapotranspiration (ET_o), and carbon dioxide concentration (CO_2). ET_o was computed using maximum, minimum and mean temperature, relative humidity, and solar radiation based on the Penman-Monteith method (Allen et al., 1998). The annual CO_2 mean values measured by the Mauna Loa station in Hawaii and provided in AquaCrop were used as CO_2 concentration in the study area. The total length of the cycle was 151 days, starting from 25th of April to the 29th of September each year from 2003 to 2021. Sprinkler irrigation was selected as the irrigation method, and irrigation was set to occur when the allowable depletion reached 80% of the readily available water. Table 1 indicates the input data used in the AquaCrop simulation. The soil texture of the field mainly consists of loamy sand with some sandy clay loam from 20 to 60 cm of depth as indicated in Table 2. Soil water content at saturation, field capacity, and PWP were also defined.

2.3 ML algorithms

Several ML algorithms were tested to evaluate their capability to draw the patterns embedded within the input variables and their relationship with the yield. The tested algorithms include MPL, RF, and GRU. Their performances were compared against each other to identify which model performs best.

2.3.1 Multilayer perceptron

MLP is a class of feedforward artificial neural network consisting of at least three fully connected layers (an input layer, a hidden layer, and an output layer).

TABLE 1 AquaCrop's input data and their sources.

Input variable	Source/value
Daily maximum and minimum temperature and rainfall	NOAA station
Daily ET_o	Computed by ET_o calculator
Mean annual CO_2 concentrations	Historical data from Mauna Loa Observatory (Hawaii)
Planting date	April 25
Harvest date	September 22
Plant density	13.6 plants/m ²
Normalized water productivity	14.5 g/m ²
HI	0.27
Length of crop cycle	151 days

TABLE 2 Soil texture and properties of the study area.

Depth (cm)	Soil type	Water content at saturation (%)	Field capacity (%)	Permanent wilting point (%)	k_{sat} (mm/day)
0–20	Loamy Sand	38	16	8	800
20–40	Sandy Clay loam	47	32	20	125
40–60	Sandy Clay loam	50	39	27	75

Each node uses a nonlinear activation function except the input layer. The input layer receives the data to train and test the model; the output layer gives its final prediction using classification or regression functions. The hidden layers transfer the input data to the output layer and are the computational brain of the MLP. The links between adjacent layers connecting the neurons are known as weights; these are updated during the learning phase to minimize the prediction error. For training all neurons, MLP uses the backpropagation algorithm, a chain rule-based supervised learning approach. MLP differs from linear perceptron in that it has multiple layers and uses nonlinear activation functions unlike

linear perceptron. MLP can, thus, solve problems that are nonlinear. The computation done by each neuron in the hidden and output layers is formulated in Eqs. (1) and (2).

$$h(r) = \Phi(r) = \rho(w_1g(r) + b_1) \quad (1)$$

$$g(r) = \rho(w_2h(r) + b_2) \quad (2)$$

where the parameters to be learned are the bias vectors (b_1 and b_2) and the weight matrices (w_1 and w_2); Φ and ρ represent the activation functions.

The data were split into three sets: training, validation, and test sets. The training dataset is used to train the model to learn the underlying pattern in the data. The validation dataset is used to evaluate different model architecture and to find the best set of parameters without overfitting at the learning stage. The test set is employed to assess the performance of the model on unseen data and is disregarded during model training and parameter tuning. The dataset was split into 50%, 20%, and 30% of the train, validation, and test sets, respectively. Before being fitted to the model, data were scaled between 0 and 1 using the min-max scaler for all the data points to be in the same range. The input features were precipitation, minimum and maximum temperature, solar radiation, surface temperature, ET_o , cotton growth stage, irrigation amount, and total water content, whereas cotton yield was the predictand. Fig. 2 depicts the structure of MLP. The developed model consists of an input layer, two hidden layers of 128 nodes each, and an output layer. The number of layers and the number of nodes in each layer are model hyperparameters that need to be tuned. In this paper, the hyperparameters were selected manually. We used tanh as the activation function, the learning rate was set to $10e^{-3}$, and Adam optimization method was used as the optimizer.

2.3.2 Random Forest

RF, which is a machine learning algorithm introduced by Breiman (2001), relies on the concept of model aggregation to solve classification problems by

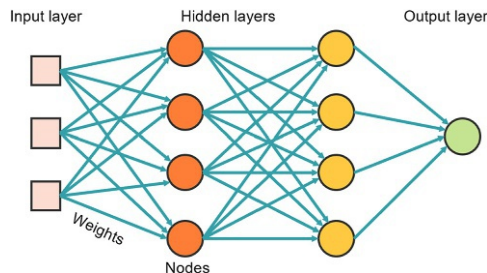


FIG. 2 The MLP architecture. The data flow in the structure from the input layer to hidden layers to the output layer. We used an MLP structure with one input layer, two hidden layers of 128 nodes each, and an output layer.

predicting a categorical response variable or to provide a continuous response variable as a solution for regression problems. RF solves these two types of problems by combining binary decision trees built from the training subset of a dataset. To increase the predictability of the response variable, the data for each tree are iteratively separated into more homogeneous units known as nodes. Split points are determined by the values of predictor variables. As a result, the factors utilized to separate the data are regarded as major explanatory variables. A categorical response's predicted value is the mode of the classes from all the individual fitted decision trees, and a continuous response's predicted value is the mean fitted response from all the individual trees that emerged from each bootstrapped sample. The main hyperparameter to be tuned is the number of decision trees which was set to be 200 after a series of trial and error to identify the optimum value. Fig. 3 demonstrates the workflow of RF.

2.3.3 Gated recurrent unit

GRU is a newest type of recurrent neural network (RNN) developed by Cho et al. (2014) to solve the vanishing gradient problem of RNN (Bengio et al., 1993; Hochreiter, 1998). In GRU, the information flow is regulated by two gates: the update gate that determines the amount of information that will be transmitted from the previous timestep to the current timestep and reset gate that decides what to forget from the past information. Two fully connected layers with sigmoid activation functions provide the gates' outputs. Fig. 4

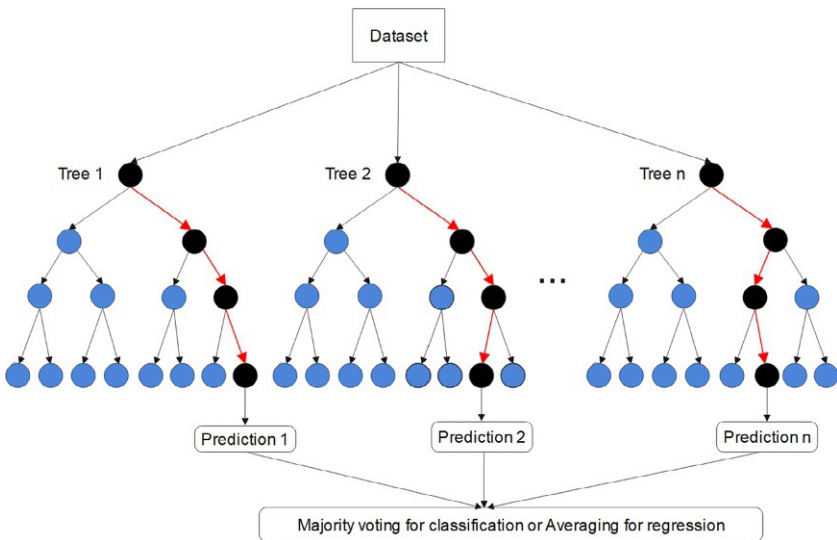


FIG. 3 RF architecture. (Adapted from Umutoni, L., Samadi, V., 2024. Application of machine learning approaches in supporting irrigation decision making: a review. *Agric. Water Manag.* 294.)

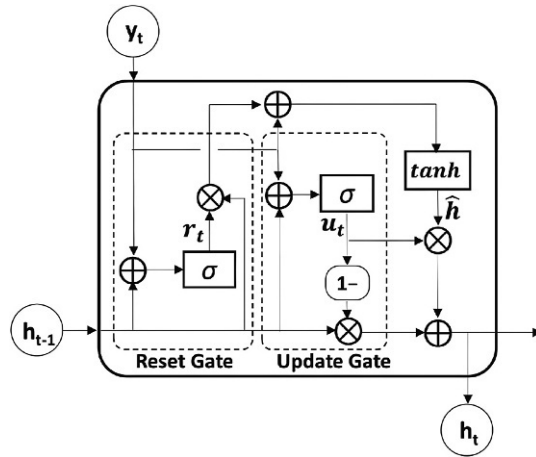


FIG. 4 The GRU structure. \otimes symbolizes element-wise multiplication, \oplus is element-wise concatenation, σ is the sigmoid function, and \tanh is the hyperbolic tangent function.

illustrates the structure of a GRU network, and Eqs. (3) and (4) describe the algorithmic workflow process:

$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r) \quad (3)$$

$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \quad (4)$$

where X_t is the input vector, H_{t-1} is the hidden state and contains the information of the previous time step, R_t and Z_t are the reset gate and update gate, respectively, σ is the activation function, W_{xr} , W_{hr} , W_{xz} , W_{hz} are weights, and b_r and b_z are biases.

The process of training a neural network model starts with initializing the values of some specific variables known as hyperparameters. Some of these hyperparameters such as the activation function and number of hidden layers define the structure of a model while parameters such as the learning rate, decay time, etc. define how it is trained (Patterson and Gibson, 2017). Multiple runs of the same algorithm with various hyperparameters will provide different results. The number of layers, number of hidden units, the learning rate, and the dropout size are the hyperparameters adjusted in this study. The learning rate is a hyperparameter that controls the rate at which the model's weights are updated with respect to the gradient of the loss function. It has a substantial influence on the deep learning model training process. A very low learning rate slows network learning, whereas a very high learning rate produces variations in training and hinders learning process convergence (Patterson and Gibson, 2017). The developed GRU model consists of one input layer, three hidden layers, and an output layer. Each of the four layers has 64 hidden units to avoid overfitting, the dropout parameter was set to 0.2 and the batch size to 64, and the learning rate was $10e^{-3}$. The input data are the same as the ones used for the MLP and RF

models; similarly, the dataset was split into 50% training set, 20% validation set, and 30% testing set.

2.4 Performance evaluation

To assess how well the models predicted the yield, we used several performance measures, namely, mean absolute error (MAE), root mean squared error (MSE), and R-squared (R^2). MAE measures the discrepancies between the observed and simulated values of a variable and returns the mean of their absolute values; RMSE, on the other hand, gives the square root of the mean of the squared errors. R^2 evaluates the predictive performance of a model by measuring the average squared difference between the observed and predicted values (see Eqs. 5–7).

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

where y_i is the actual cotton yield, \hat{y}_i is the simulated yield, and \bar{y} is the actual yield mean value.

Following these proposed algorithms and data, we developed data-driven models to simulate yield at an irrigation field in the EREC. The development of an ML-based crop yield estimation model was done in two steps as illustrated by Fig. 5. First, AquaCrop was used to simulate the yield based on climate, crop, soil, and management practices. A study conducted by Qiao (2012) from 2009 to 2011 at EREC established the optimum parameters for cotton simulations in AquaCrop. From 2009 to 2010, irrigation experiments were conducted under field environments, while in 2011, a rainout shelter was used to control the environment. The study's goal was to parameterize and validate AquaCrop for cotton growth simulations in humid climate of the southeast United States. Adjusted parameters were CGC, CDC, water depletion thresholds (p factors), water productivity (WP), and reference harvest index (HI_o). Building on this research, cotton growth seasons from 2003 to 2021 were simulated using the AquaCrop model with the aim of generating sufficient data to train an ML model to replicate AquaCrop in simulating cotton yield. Weather data were obtained from a NOAA weather station located in Blackville, SC. Prior to being used in the AquaCrop, the collected weather data were preprocessed to correct

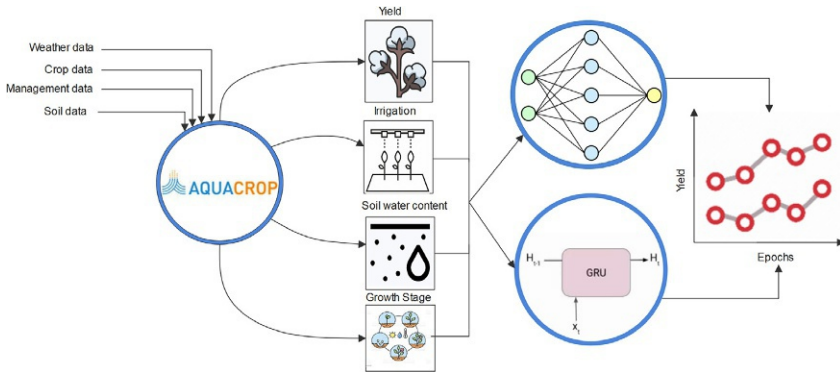


FIG. 5 Workflow followed to simulate yield with crop and data-driven models.

gaps and identify missing values and outliers. Missing data values were filled using the forward filling method of *numpy*. The method uses a previously recorded value to fill the following missing data points. This technique was chosen because the dataset did not have a significant number of consecutively missing values.

Second, the outputs of AquaCrop were used as input data to ML models. After setting up and calibrating the AquaCrop model, simulation results were used to train MLP and GRU models to predict cotton yield. In addition to weather data (rainfall, minimum and maximum temperature, and evapotranspiration, solar radiation, and surface temperature), crop growth stage, total water content, and applied irrigation data obtained from AquaCrop were used to train these ML models. To investigate the most influencing variables on yield, a correlation analysis was performed by using the Pearson correlation coefficient (r) value as an indicator of the strength of correlation between each feature and the yield. Values of r presented in Table 3 show that growth stage, solar radiation, evapotranspiration, and minimum temperature are the most important features for yield prediction with r ranging from 0.59 to 0.18, while precipitation, temperature, water content, surface temperature, and irrigation are least correlated to yield ($0.00 < r < 0.07$). The highest correlation of yield to temperature, solar radiation, and evapotranspiration compared to other climatic variables can be explained by the fact that for cotton developmental events occur much more rapidly as temperature increases (Reddy et al., 1996). This implies that the ML models for yield simulation are likely to extract the most meaningful patterns from these features.

3 Results and discussion

3.1 AquaCrop results

After setting up and parameterizing the AquaCrop model, each season was run to simulate the yield given weather, crop, and management data. One of the

TABLE 3 Input features and their respective correlation coefficient.

Predictor variable	Pearson correlation coefficient (<i>r</i>)
Growth stage	0.59
Solar radiation	-0.29
Evapotranspiration	-0.24
Minimum temperature	0.18
Maximum temperature	0.05
Irrigation	-0.07
Surface temperature	-0.06
Soil water content	-0.06
Precipitation	0.00

results AquaCrop produces is the total water content which is the soil moisture in the defined soil profile (0.0–0.6 m). The soil moisture during the simulation period ranges from 135.2 to 269.8 mm as illustrated by Fig. 6. The average soil moisture is 167.7 mm, and the higher soil moisture levels were mostly observed on days with rainfall or irrigation.

In addition, the amount of water applied to plants through irrigation is shown by Fig. 7. Fig. 8 illustrates the yearly irrigation amount against precipitation. Like the two figures depict it, the irrigation amount was higher for periods of low rainfall such as the years 2006, 2007, and 2008 which had a total rainfall of 466.4, 375.9, and 506.6 mm during the simulation period consequently requiring more water to grow plants.

Precipitation amount over time determines the soil water dynamics. In the EREC, precipitation varies significantly during the growing season facilitating a period of utilization of stored water during the early growing season or irrigation if the region experienced a prolonged deficit. Fig. 8 illustrates the fluctuations in soil water content, irrigation, and rainfall. The comparison of soil water content, irrigation, and rainfall provides an example of how soil moisture dynamics and precipitation timing govern water utilization and movement in soil. The timing of precipitation is equally as important as its magnitude when considering soil water dynamics. As evidence, it appears that higher irrigation amount was applied by the model for years with low rainfall to overcome the water deficit.

3.2 MLP results

Using weather data, namely, precipitation, evapotranspiration, maximum and minimum temperature, and AquaCrop's results, including irrigation, water

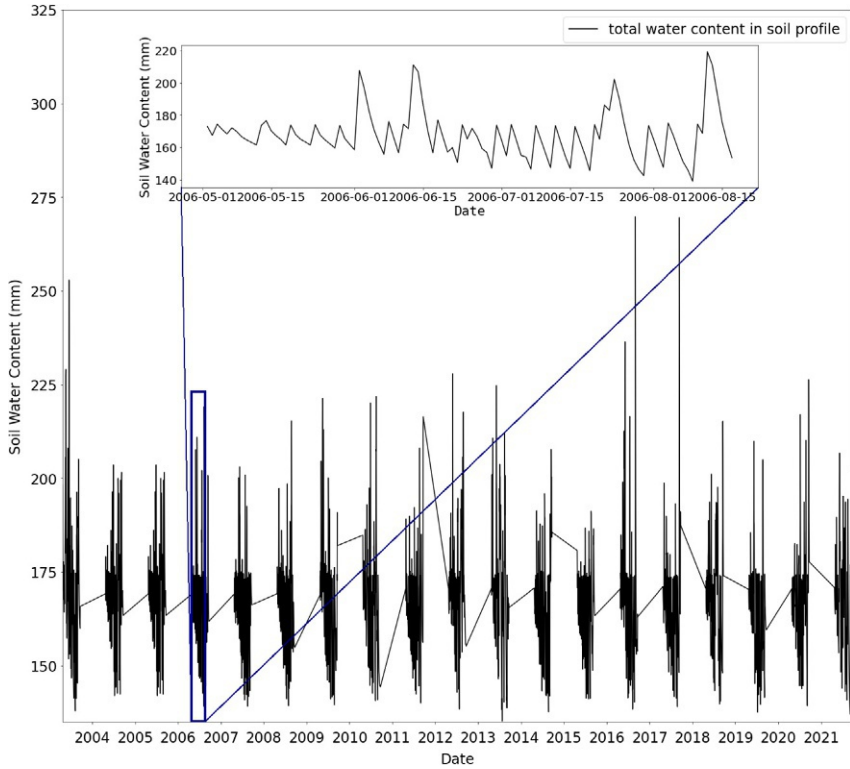


FIG. 6 Total soil water content.

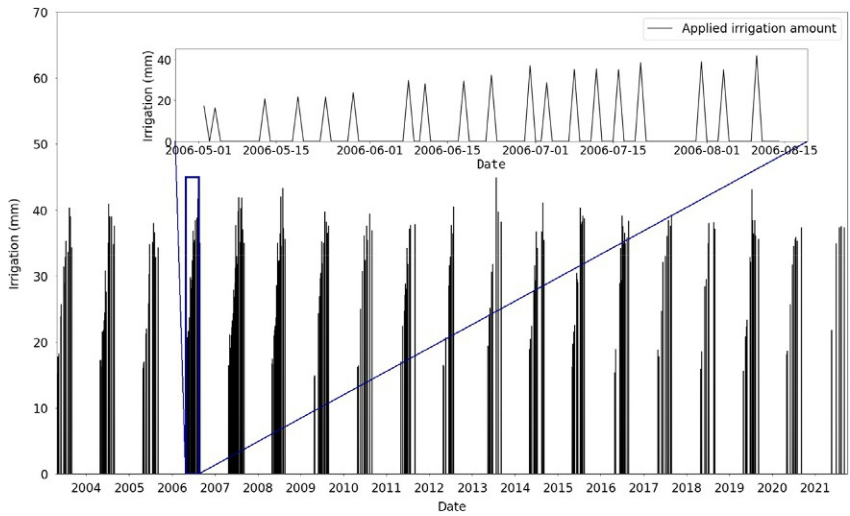


FIG. 7 Applied irrigation amount during the simulated period.

Copyright © 2024, Elsevier. All rights reserved.

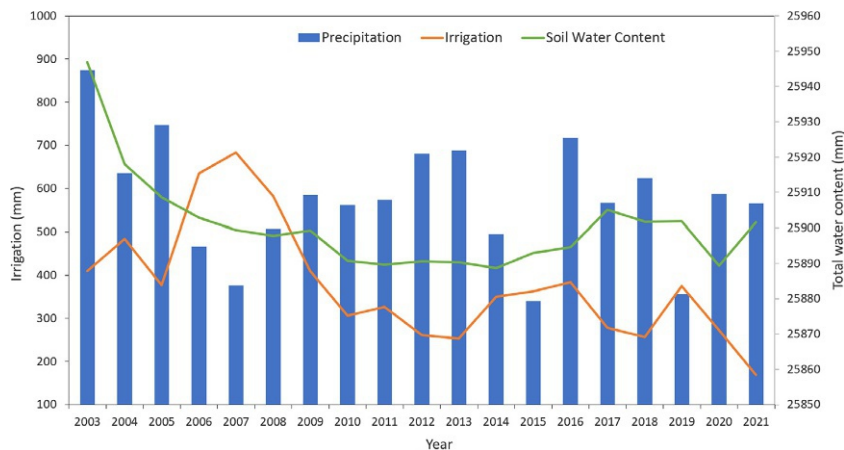


FIG. 8 Seasonal precipitation, irrigation, and soil water content.

content, and growth stage, MLP was trained up to 257 epochs with a different combination of input variables to evaluate the effect of input features on the model's ability to predict yield. First, all the input variables were used, and the obtained results showed that MLP accurately simulated the yield with an MAE of 0.23, RMSE of 0.34, and R^2 of 0.88. The model was able to emulate AquaCrop in simulating cotton's yield by simulating the yield values quite close to the actual yield (from AquaCrop). The observed yield was, however, slightly higher than the predicted yield throughout the simulation period (see Fig. 9), and it appears that the model performed slightly better at predicting peak yield values than lower values.

Furthermore, input variables were changed to identify the best set of inputs, i.e., input variables that give the minimum error metric. The results of the correlation analysis showed that growth stage has the highest correlation compared to other inputs; thus, the model was run without this variable to assess its accuracy when the growth stage data are unavailable and the impact this would have on the simulation. Omission of growth stage resulted in a reduced performance as RMSE increased from 0.34 to 0.9 and R^2 decreased from 0.88 to 0.2 while the MAE became 0.7. Based on the graphical performance of this model (see Fig. 10), it is evident that it performs poorly and does not capture the trends in the output yield.

Second, solar radiation and evapotranspiration were omitted one at the time to assess the impact on the simulation as they are the second and the third significant variables after growth stage. Without solar radiation as an input, the model still had an acceptable performance as the MAE was 0.24, RMSE was 0.33, and R^2 was 0.88. The performance when evapotranspiration was excluded was an MAE of 0.25, a RMSE of 0.38, and R^2 was 0.85. Additionally, the model was run with the variables that have a high r ($r > 0.15$); the error metrics are

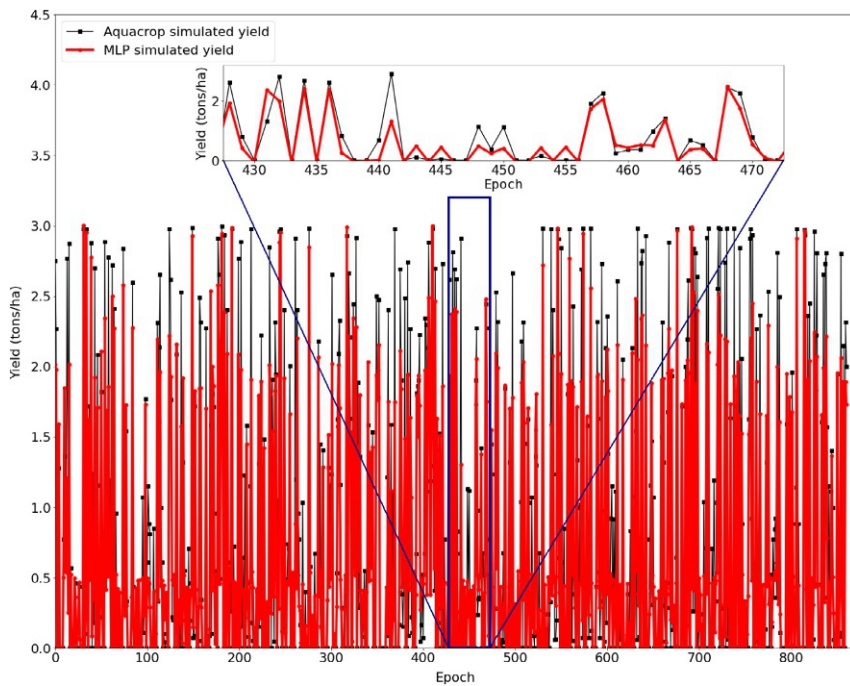


FIG. 9 Simulated yield using all variables.

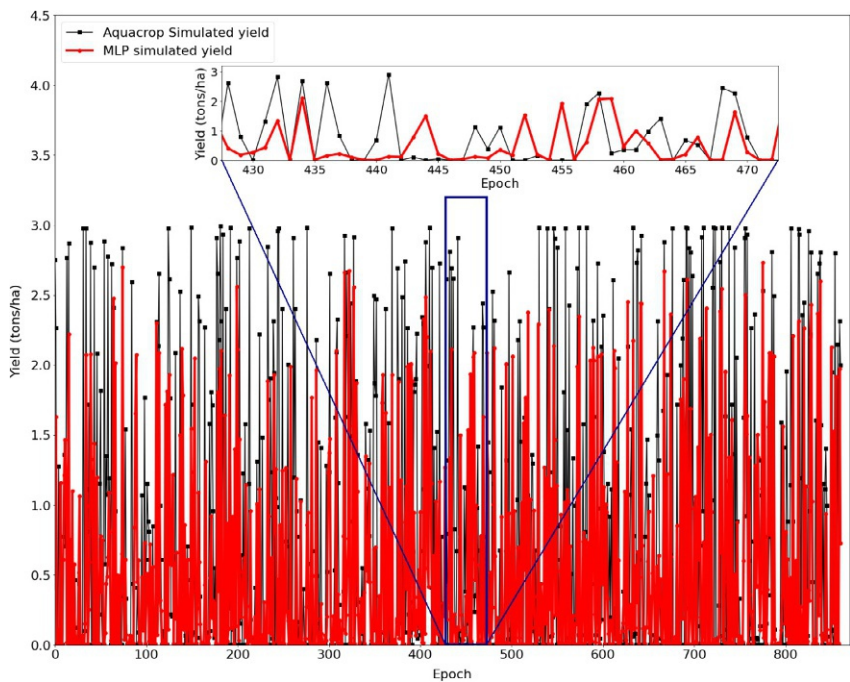


FIG. 10 Simulated yield without the growth stage.

TABLE 4 MLP performance metrics.

Predictor	Performance metrics		
	MAE	RMSE	R^2
Precipitation, maximum and minimum temperature, evapotranspiration, solar radiation, surface temperature, water content, irrigation, and growth stage	0.23	0.34	0.88
Precipitation, maximum and minimum temperature, evapotranspiration, solar radiation, surface temperature, water content, and irrigation	0.7	0.9	0.2
Precipitation, maximum and minimum temperature, evapotranspiration, surface temperature, water content, irrigation, and growth stage	0.24	0.33	0.88
Precipitation, maximum and minimum temperature, solar radiation, surface temperature, water content, irrigation, and growth stage	0.25	0.38	0.85
Growth stage, solar radiation, evapotranspiration, and minimum temperature	0.22	0.32	0.89

Note: The boldface numbers indicate the best performing model.

shown in Table 4 and indicate that for the study area, the most important variables for yield estimation are growth stage, solar radiation, evapotranspiration, and minimum temperature.

3.3 GRU results

GRU, a more advanced ML algorithm, was also tested, and its results were compared to the MLP model. Likewise, GRU performance was evaluated with respect to various input variables. When all variables were included, the performance scores were 0.22, 0.34, and 0.88 for MAE, RMSE, and R^2 , respectively. The model was able to learn the pattern in the data and simulate the results well (see Fig. 11), although low yield values were underestimated like for MLP. Running the model without solar radiation data was excluded; the performance was 0.22, 0.35, and 0.87 for MAE, RMSE, and R^2 , respectively. Thus, no considerable difference was observed between the two GRU models. On the other hand, excluding the growth stage from the input variables resulted in an MAE of 0.64, a RMSE of 1, and a R^2 of 0. This shows that the model's performance in predicting daily crop yield is highly dependent on data of the growth stage of a crop as depicted by Fig. 12. Simulating yield with variables highly correlated to the yield and omitting the remainder of the training dataset did not hinder the model performance as 0.21, 0.33, and 0.89 were obtained for MAE, RMSE, R^2 , and respectively. Table 5 presents the model performance when different input data combinations were used.

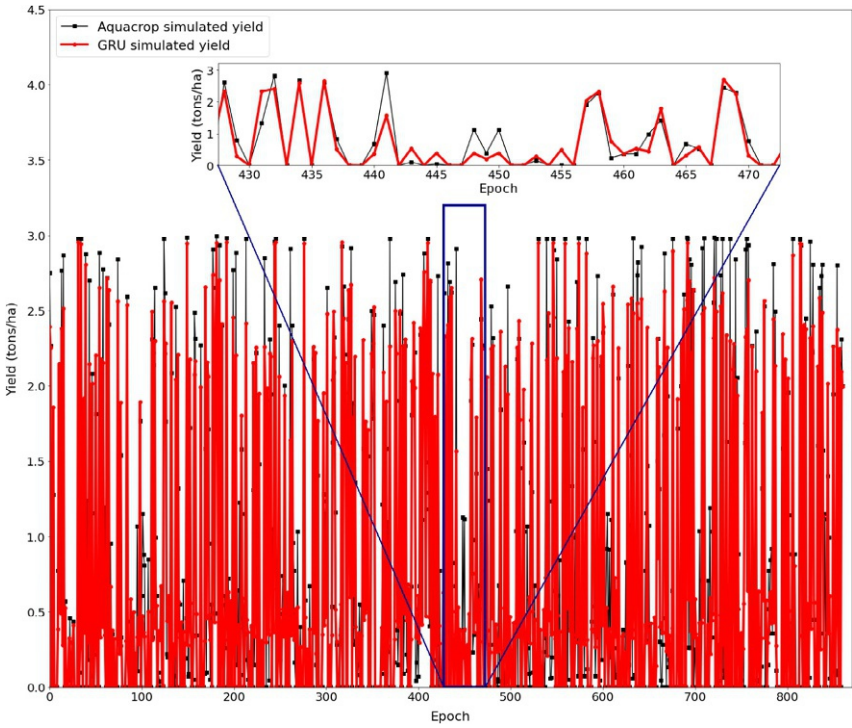


FIG. 11 Simulated yield vs AquaCrop simulation when all variables are included.

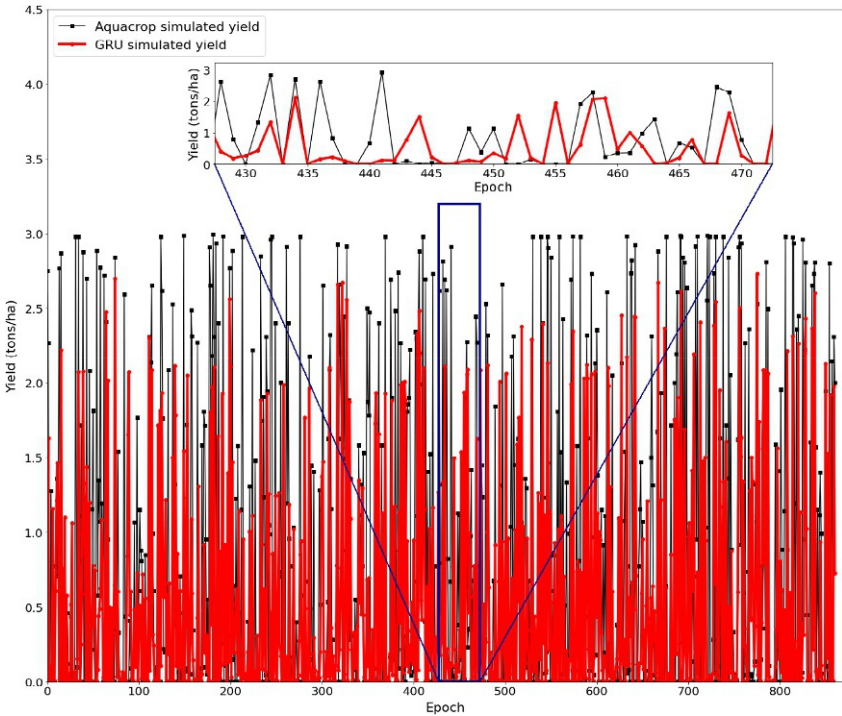


FIG. 12 GRU prediction without incorporating growth stage data.

TABLE 5 GRU performance metrics.

Predictor	Performance metrics		
	MAE	RMSE	R^2
Precipitation, maximum and minimum temperature, ET_o , solar radiation, surface temperature, water content, irrigation, and growth stage	0.22	0.34	0.88
Precipitation, maximum and minimum temperature, ET_o , solar radiation, surface temperature, water content, and irrigation	0.64	1	0
Precipitation, maximum and minimum temperature, ET_o , surface temperature, water content, irrigation, and growth stage	0.22	0.35	0.87
Precipitation, maximum and minimum temperature, solar radiation, surface temperature, water content, irrigation, and growth stage	0.22	0.33	0.88
Growth stage, solar radiation, evapotranspiration, and minimum temperature	0.21	0.33	0.89

Note: The boldface numbers indicate the best performing model.

3.4 RF results

A detailed performance of RF in simulating yield using different combinations of inputs is presented in Table 6. For all RF simulations, the results of MAE varied from 0.2 to 0.63, RMSE was between 0.3 and 0.84, the minimum R^2 value obtained was 0.27, while the best model performance resulted in a R^2 value of 0.9. High MAE and RMSE and low R^2 values were found for the model without growth stage data (see Fig. 13). The best performance results were observed when all variables were used as inputs. Graphical results show that RF can simulate yield when trained by a dataset from which it can learn the function that relates soil and climatic variables to yield (Fig. 14).

The assessment approach considers how a varied set of characteristics affects the performance of ML models. Nonetheless, each ML approach uses the same parameters to produce all its models. A comparison of the three algorithms shows that they can be reliable in crop yield estimation even though none of them was able to accurately simulate low yield values. The results from this study support prior findings that ML models can be beneficial in crop yield estimation provided that sufficient data are available to train the models (Fukuda et al., 2013; Prasad et al., 2021; Ren et al., 2023). According to the best performance achieved by each model as presented in Table 7, MLP and GRU can accurately simulate the yield with less inputs compared to RF that achieved the best performance but using more input features.

TABLE 6 RF performance results.

Predictor	Performance metrics		
	MAE	RMSE	R ²
Precipitation, maximum and minimum temperature, ET _o , solar radiation, surface temperature, water content, irrigation, and growth stage	0.19	0.3	0.9
Precipitation, maximum and minimum temperature, ET _o , solar radiation, surface temperature, water content, and irrigation	0.63	0.84	0.27
Precipitation, maximum and minimum temperature, ET _o , surface temperature, water content, irrigation, and growth stage	0.2	0.31	0.9
Precipitation, maximum and minimum temperature, solar radiation, surface temperature, water content, irrigation, and growth stage	0.2	0.3	0.9
Growth stage, solar radiation, evapotranspiration, and minimum temperature	0.2	0.32	0.89

Note: The boldface numbers indicate the best performing model.

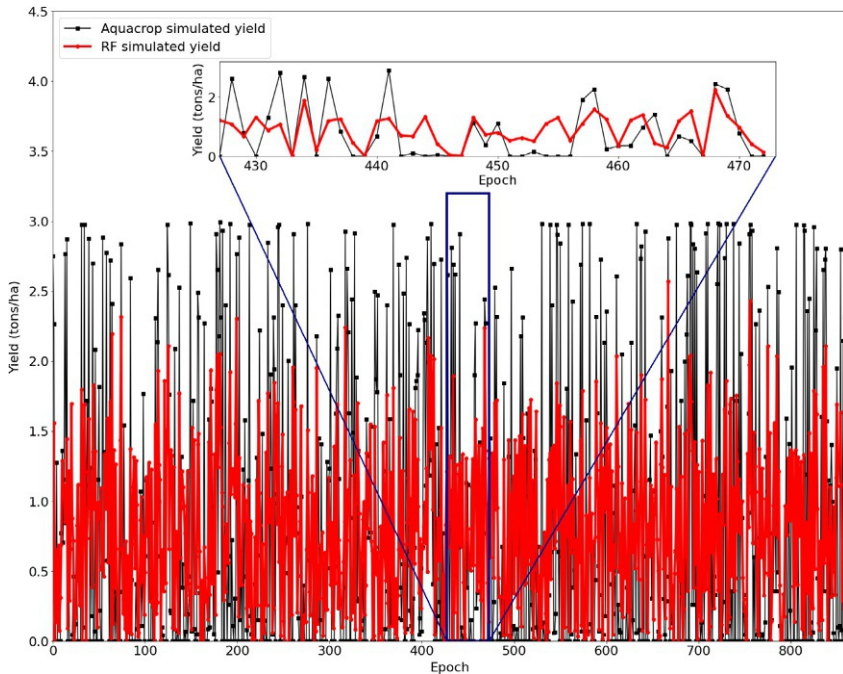


FIG. 13 RF simulated yield when all variables are included except growth stage.

Copyright © 2024. Elsevier. All rights reserved.

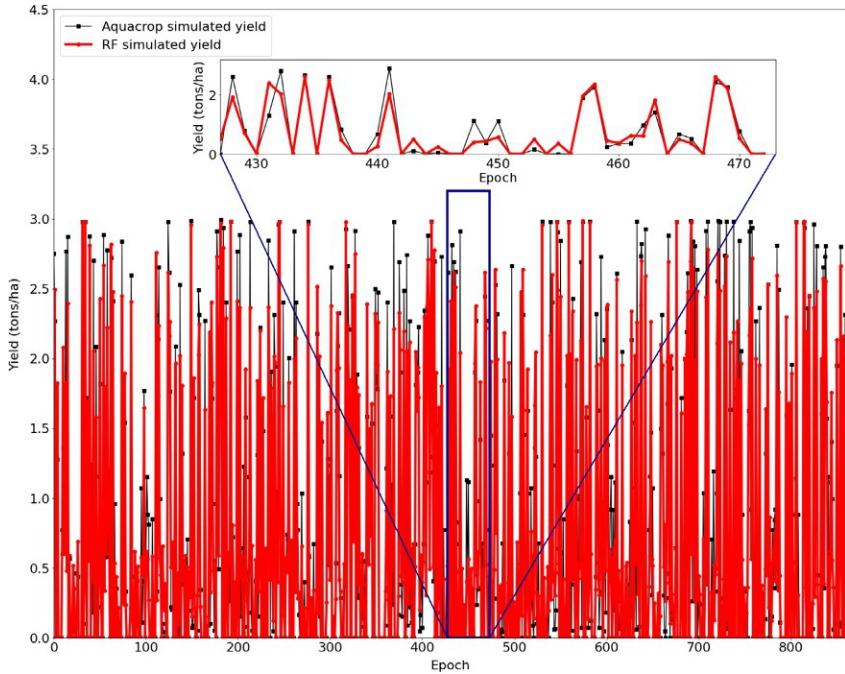


FIG. 14 RF yield simulation results when all variables are included.

TABLE 7 MLP, GRU, and RF best performances.

Model	Input features	Performance metrics		
		MAE	RMSE	R^2
MLP	Growth stage, solar radiation, evapotranspiration, and minimum temperature	0.22	0.32	0.89
GRU	Growth stage, solar radiation, evapotranspiration, and minimum temperature	0.21	0.33	0.89
RF	Precipitation, maximum and minimum temperature, ET_o , solar radiation, surface temperature, water content, irrigation, and growth stage	0.19	0.3	0.9

4 Conclusion

The purpose of this study was to determine whether MLP, GRU, and RF could accurately estimate the daily cotton yield during a growing season by learning from data generated by AquaCrop. We demonstrated how the performance of these three models varies with changing inputs and presented the most significant variables required for an accurate prediction of daily yield. Analysis suggested that the most important input was growth stage as it helps in predicting a crop's vegetative development during its life cycle. We conclude that using ML algorithms in crop yield estimation can be another approach of predicting yield when sufficient data are available. An advantage of ML-based crop yield estimation models over traditional models is that the number of parameters to be calibrated is minimum for the former; hence, a well-trained model can be used and applied to simulate yield of other crops in regions. Such models can be useful in irrigation planning and facilitate better management of agricultural activities to increase the productivity.

In terms of simulation performance, MLP, GRU, and RF models employed in this research simulated yield values accurately. In this case, the three algorithms demonstrated the potential to simulate complex yet nonlinear cotton yield data at a farm scale. We were able to improve the model performance in terms of including additional climatic data and features, although we were less successful at improving the estimation of low yield values. There is some evidence that this failure may be related to fluctuations in field conditions and nutrient applications that were not considered in the simulation. This is a substantial challenge as this type of simulation is tempered by the heterogeneity in climatic attributes, soil (soil moisture, drainage and evapotranspiration, etc.), crop characteristics, complexities in soil water storage capacity due to rapid and active evapotranspiration process, and difficulties in computing subsurface/soil moisture contribution to the yield estimation.

Based on the results, we argue that the potential benefits of ML models for crop yield simulation can only be realized via the formulation of a robust and rigorous GRU model that is well suited to evaluate input-output behavior (behavioral consistency) for sequential climatic and yield time-series data. One obvious way to reduce the obstacles and improve ML fidelity is to use more advanced ML techniques such as Transformers (Lim et al., 2021) algorithms to overcome the problem of sequence transduction in simulation. Transformers algorithms are faster than recurrent layers for shorter sequence lengths and can be restricted to consider only a neighborhood in the input sequence for long seasonal sequence lengths. Enhancing the ability of data-driven models to simulate extreme fluctuations within yield time-series records is also important, which needs further attention.

Furthermore, other parameters indicating the growth of plants, including leaf area index, could be mathematically formulated and incorporated into the ML models. Our research group has sought, therefore, to develop appropriate ML

models that are comparable with a farm-scale crop simulation model. The main limitation of data-driven approaches, however, is their overfitting problems that make them somewhat poor predictors. Our ongoing work explores the development of more advanced and computationally cost-effective data-driven algorithms. MLP, GRU, and RF models could be extended to similar climate regions or could be developed for other regions with available field data for yield simulation. Moreover, this study validated the usefulness of coupling process-based models with ML in yield estimation to help growers in applying water and nutrients more efficiently to optimize yield and for better farm management. Although incorporating field-based yield data may challenge data-driven simulations due in part to the higher variability of yields from year to year as well as a lack of adequate field-based yield datasets. We acknowledge that training the models with observed data would render the model more robust to handle the high variability in soil and crop features and encourage future work to identify if alternative data sources such as remote sensing could lead to improved model accuracy.

Acknowledgment

The authors appreciate the funding support from the Southern Sustainable Agriculture Research and Education (SARE) graduate fellowship program as well as USDA-NIFA (grant # 2023000603). Any opinions, findings, and discussions expressed in this study are those of the authors and do not necessarily reflect the views of the SARE and USDA-NIFA.

References

- Ahmed, M., Akram, M.N., Asim, M., Aslam, M., Hassan, F., Higgins, S., Stöckle, C.O., Hoogenboom, G., 2016. Calibration and validation of APSIM-Wheat and CERES-Wheat for spring wheat under rainfed conditions: Models evaluation and application. *Comput. Electron. Agric.* 123. <https://doi.org/10.1016/j.compag.2016.03.015>.
- Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration—guidelines for computing crop water requirements—FAO irrigation and drainage paper 56. *Irrig. Drain.* <https://doi.org/10.1016/j.eja.2010.12.001>.
- Archontoulis, S.V., Castellano, M.J., Licht, M.A., Nichols, V., Baum, M., Huber, I., Martinez-Feria, R., Puntel, L., Ordóñez, R.A., Iqbal, J., Wright, E.E., Dietzel, R.N., Helmers, M., Vanloocke, A., Liebman, M., Hatfield, J.L., Herzmann, D., Córdova, S.C., Edmonds, P., et al., 2020. Predicting crop yields and soil-plant nitrogen dynamics in the US Corn Belt. *Crop Sci.* 60 (2). <https://doi.org/10.1002/csc2.20039>.
- Asseng, S., Zhu, Y., Basso, B., Wilson, T., Cammarano, D., 2014. Simulation modeling: applications in cropping systems. In: *Encyclopedia of Agriculture and Food Systems.*, <https://doi.org/10.1016/B978-0-444-52512-3.00233-3>.
- Bali, N., Singla, A., 2022. Emerging trends in machine learning to predict crop yield and study its influential factors: a survey. *Arch. Comput. Methods Eng.* 29 (1). <https://doi.org/10.1007/s11831-021-09569-8>.
- Basso, B., Liu, L., 2019. Seasonal crop yield forecast: methods, applications, and accuracies. *Adv. Agron.* 154. <https://doi.org/10.1016/bs.agron.2018.11.002>.

- Bengio, Y., Frasconi, P., Simard, P., 1993. Problem of learning long-term dependencies in recurrent networks. In: 1993 IEEE International Conference on Neural Networks., <https://doi.org/10.1109/icnn.1993.298725>.
- Bogard, M., Biddulph, B., Zheng, B., Hayden, M., Kuchel, H., Mullan, D., Allard, V., Le Gouis, J., Chapman, S.C., 2020. Linking genetic maps and simulation to optimize breeding for wheat flowering time in current and future climates. *Crop Sci.* 60 (2). <https://doi.org/10.1002/csc2.20113>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014—2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference., <https://doi.org/10.3115/v1/d14-1179>.
- Daw, A., Karpatne, A., Watkins, W.D., Read, J.S., Kumar, V., 2022. Physics-guided neural networks (PGNN): an application in lake temperature modeling. In: Knowledge-Guided Machine Learning., <https://doi.org/10.1201/9781003143376-15>.
- Ersoz, E.S., Martin, N.F., Stapleton, A.E., 2020. On to the next chapter for crop breeding: convergence with data science. *Crop Sci.* 60 (2). <https://doi.org/10.1002/csc2.20054>.
- Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardud, V., Müller, J., 2013. Random Forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agric. Water Manag.* 116. <https://doi.org/10.1016/j.agwat.2012.07.003>.
- Gaydon, D.S., Balwinder, S., Wang, E., Poulton, P.L., Ahmad, B., Ahmed, F., Akhter, S., Ali, I., Amarasingha, R., Chaki, A.K., Chen, C., Choudhury, B.U., Darai, R., Das, A., Hochman, Z., Horan, H., Hosang, E.Y., Kumar, P.V., Khan, A.S.M.M.R., Laing, A.M., Liu, L., Malaviachi, M.A.P.W.K., et al., 2017. Evaluation of the APSIM model in cropping systems of Asia. *Field Crops Res.* 204. <https://doi.org/10.1016/j.fcr.2016.12.015>.
- Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 6 (2). <https://doi.org/10.1142/S0218488598000094>.
- Hsiao, T.C., Heng, L., Steduto, P., Rojas-Lara, B., Raes, D., Fereres, E., 2009. AquaCrop—the FAO crop model to simulate yield response to water: III. Parameterization and testing for maize. *Agron. J.* 101 (3), 448–459. <https://doi.org/10.2134/AGRONJ2008.0218S>.
- Lim, B., Arnk, S., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* 37. <https://doi.org/10.1016/j.ijforecast.2021.03.012>.
- Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N., 2022. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sens.* 14 (9). <https://doi.org/10.3390/rs14091990>.
- Patterson, J., Gibson, A., 2017. Deep Learning: A Practitioner’s Approach. https://books.google.com/books?hl=en&lr=&id=qrcuDwAAQBAJ&oi=fnd&pg=PR2&dq=Patterson,+J.%3B+Gibson,+A.+Deep+Learning:+A+Practitioner%E2%80%99s+Approach%3B+O%E2%80%99Reilly:+Beijing,+China,+2017.&ots=6nrJwzcpLP&sig=yiStGvUM674t9o7Br_hMTb8rGz4.
- Prasad, N.R., Patel, N.R., Danodia, A., 2021. Crop yield prediction in cotton for regional level using random forest approach. *Spat. Inf. Res.* 29 (2). <https://doi.org/10.1007/s41324-020-00346-6>.
- Qiao, X., 2012. Parameterization of FAO AquaCrop Model for Irrigated Cotton in the Humid South-east USA. Clemson University (Doctoral dissertation).

- Raes, D., Steduto, P., Hsiao, T.C., Fereres, E., 2009. AquaCrop—the FAO crop model to simulate yield response to water: II. Main algorithms and software description. *Agron. J.* 101 (3). <https://doi.org/10.2134/agronj2008.0140s>.
- Reddy, K.R., Hodges, H.F., McCarty, W.H., McKinion, J.M., 1996. *Weather and cotton growth: present and future*. MSU-MAFES 1061.
- Ren, Y., Li, Q., Du, X., Zhang, Y., Wang, H., Shi, G., Wei, M., 2023. Analysis of corn yield prediction potential at various growth phases using a process-based model and deep learning. *Plants* 12 (3). <https://doi.org/10.3390/plants12030446>.
- Shahhosseini, M., Martinez-Feria, R.A., Hu, G., Archontoulis, S.V., 2019. Maize yield and nitrate loss prediction with machine learning algorithms. *Environ. Res. Lett.* 14 (12). <https://doi.org/10.1088/1748-9326/ab5268>.
- Steduto, P., Hsiao, T.C., Raes, D., Fereres, E., 2009. AquaCrop—the FAO crop model to simulate yield response to water: I. Concepts and underlying principles. *Agron. J.* 101 (3). <https://doi.org/10.2134/agronj2008.0139s>.
- Vanuytrecht, E., Raes, D., Steduto, P., Software, T.H., 2014. AquaCrop: FAO's crop water productivity and yield response model. *Environ. Model Softw.* 62, 351–360. <https://www.sciencedirect.com/science/article/pii/S136481521400228X>.
- Washburn, J.D., Burch, M.B., Franco, J.A.V., 2020. Predictive breeding for maize: making use of molecular phenotypes, machine learning, and physiological crop models. *Crop Sci.* 60 (2). <https://doi.org/10.1002/csc2.20052>.
- Xu, X., Gao, P., Zhu, X., Guo, W., Ding, J., Li, C., Zhu, M., Wu, X., 2019. Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China. *Ecol. Indic.* 101. <https://doi.org/10.1016/j.ecolind.2019.01.059>.

This page intentionally left blank