

ARTICLE

Agronomy, Soils, and Environmental Quality

Predicting within-field cotton yields using publicly available datasets and machine learning

Stephen Leo¹  | Massimiliano De Antoni Migliorati^{1,2}  | Peter R. Grace¹ 

¹ Centre for Agriculture and the Bioeconomy, Queensland Univ. of Technology, 2 George St, Brisbane City QLD 4000, Australia

² Queensland Dep. of Environment and Science, 41 Boggo Rd, Dutton Park QLD 4102, Australia

Correspondence

Stephen Leo, Centre for Agriculture and the Bioeconomy, Queensland Univ. of Technology, 2 George St, Brisbane City, QLD 4000, Australia.

Email: s2.leo@qut.edu.au

Funding information

Cotton Research and Development Corporation, Grant/Award Number: QUT1902

Abstract

Early detection of within-field yield variability for high-value commodity crops, such as cotton (*Gossypium* spp.), offers growers potential to improve decision-making, optimize yields, and increase profits. Over recent years, publicly available datasets have become increasingly available and at a resolution where within-field yield prediction is possible. However, the viability of using these datasets with machine learning to predict within-field cotton lint yield at key growth stages are largely unknown. This study was conducted on two cotton fields, located near Mungindi, New South Wales, Australia. Three years of yield data, soil, elevation, rainfall, and Landsat imagery were collected from each field. A total of 12 models were created using: (a) two machine learning algorithms: random forest (RF) and gradient boosting machines (GBM); (b) three growth stages: squaring, flowering, and boll-fill; and (c) two different amounts of variables: all variables and the optimal variables determined by a recursive feature elimination (RFE). Results showed a strong agreement between predicted and observed yields at flowering and boll-fill when more information was available. At flowering and boll-fill, root mean square error (RMSE) ranged between 0.15 and 0.20 t ha⁻¹ and Lin's concordance correlation coefficient (LCCC) ranged between 0.50 and 0.66, with RF providing superior results in most cases. Models created using the optimal variables determined by the RFE provided similar results compared to using all variables, allowing greater model accuracy and resolution with targeted sampling. Overall, these findings indicate significant potential of publicly available datasets to predict within-field cotton yield and guide decision-making in-season.

Abbreviations: APSIM, Agricultural Production Systems sIMulator; AWC, available water capacity; BOM, Bureau of Meteorology; DEM, digital elevation model; DSSAT, Decision Support System for Agrotechnology Transfer; ECEC, effective cation exchange capacity; EVI, enhanced vegetation index; GBM, gradient boosting machines; GNDVI, green normalized difference vegetation index; LCCC, Lin's concordance correlation coefficient; LiDAR, Light Detection and Ranging; NDVI, normalized difference vegetation index; NIR, near-infrared; RDVI, renormalized difference vegetation index; RF, random forests; RFE, recursive feature elimination; RMSE, root mean square error; SILO, Scientific Information for Land Owners; SLGA, Soil and Landscape Grid of Australia; USGS, U.S. Geological Survey.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Agronomy Journal* published by Wiley Periodicals LLC on behalf of American Society of Agronomy

1 | INTRODUCTION

Cotton (*Gossypium* spp.) is the most economically important fiber crop grown globally. Aside from cereals, soybean [*Glycine max* (L.) Merr.], and rapeseed (*Brassica napus* L.) crops, cotton is the largest crop cultivated in terms of area planted (FAO, 2019). Currently, Australia is one of the world's largest exporters of cotton, with almost 100% of domestic production being exported (ABARES, 2019). Over the past 20 yr, Australia has planted an average 349,000 ha to cotton and produced 651 kilotonnes (kt) per year (ABARES, 2019). The ability to forecast cotton yield can provide growers with invaluable information to make more informed decisions regarding in-season investments and management decisions, such as applying N fertilizer (Horie et al., 1992; Kim et al., 2019). Early prediction of yield can also assist with estimating revenue, gin workload, and market value with greater accuracy.

Forecasting within-field yield, however, is a challenge in many agricultural systems. Within-field cotton yield can vary due to a multitude of factors, including soil properties, localized pests and diseases, and management (Chen et al., 2018; Li et al., 2008; Ping et al., 2007). Yield can also vary year-to-year within the same field due to differences in climate and management; such as fertilizer, irrigation, rotation, and cultivar (Filippi et al., 2019; Wenxuan, 2018). Current in-season yield estimations in cotton are based on boll counting within a small subsection of the field (Sun et al., 2019). However, this is a time-consuming, labor-intensive, and spatially restricted method. By the time stage bolls begin to develop, the implementation of any in-season management, such as N fertilizer applications, is going to be too late to impact yield (Macdonald et al., 2018), particularly when current N fertilizer applications occur between the squaring and flowering growth stages (CRDC & CottonInfo, 2018).

Crop simulation models, such as Decision Support System for Agrotechnology Transfer (DSSAT) (Jones et al., 2003) and Agricultural Production Systems sIMulator (APSIM) (Holzworth et al., 2014), are useful tools for simulating crop development and final yield based on soil, weather, and management conditions (Hoogenboom et al., 2004). However, such models require an extensive number of input parameters and site-specific calibration. These models are also spatially limited, only allowing simulations for the whole paddock or several management units within the paddock (Basso et al., 2001). Empirically derived yield estimates using remote and proximal sensing are practical approaches to assess within-field yield variability (Ballester et al., 2017; Meng et al., 2017; Zarco-Tejada et al., 2005). However, they are constrained when applied to other fields or seasons, as they typically ignore climatic, soil, and management variables (Hatfield et al., 2008; Lobell, 2013).

Core Ideas

- Public datasets were tested to predict within-field cotton yield at key growth stages.
- Flowering provided the highest prediction accuracy compared to squaring and boll-fill.
- Random forests provided slightly greater accuracy than gradient boosting machines.
- Reducing the number of variables via feature elimination maintained model accuracy.
- Contextual information should be incorporated to further understand yield drivers.

Further improvement of empirical approaches have combined proximal or remote sensing with other data sources. For example, Ray et al. (1999) estimated cotton yield using an agrometeorological model with satellite-derived normalized difference vegetation index (NDVI) at the regional level. He and Mostovoy (2019) used Sentinel-2 to derive cotton leaf area index as an input for an ecosystem model. At the sub-paddock scale, Thomasson et al. (2004) predicted cotton yield using NDVI, historical yield, texture, elevation, and slope, and found the addition of more data increased the predictability of yield. Similarly, Huang et al. (2013) improved yield estimates when combining vegetation indices and soil electrical conductivity, compared with vegetation indices alone. Recently, Filippi et al. (2020) assessed the potential of large yield-mapping datasets and diverse spatial covariates to forecast mid- and late-season within-field cotton yield. However, these approaches were limited either by the low spatial resolution, the relatively small dataset of one field and cropping season or the cotton growth stages assessed.

Other data-driven approaches have involved using machine-learning algorithms, which are becoming increasingly popular to deal with more complex problems in agriculture (Liakos et al., 2018). The most common are linear or multiple linear machine-learning algorithms. However, these algorithms have limited capability when handling non-linear relationships (Archontoulis & Miguez, 2015; Chlingaryan et al., 2018). Ensemble-learning methods, such as random forests (RF) and gradient boosting machines (GBM), can capture nonlinear relationships and have shown strong prediction capability or improved prediction of crop yield compared with traditional linear approaches in many cropping systems (Everingham et al., 2016; Filippi, Jones, et al., 2019; Filippi et al., 2020; Kayad et al., 2019; Khanal et al., 2018; Richetti et al., 2018; Shahhosseini et al., 2019).

A limitation of predicting within-field yield, however, is the availability of on-farm data. Growers and consultants are often restricted from obtaining high spatial and

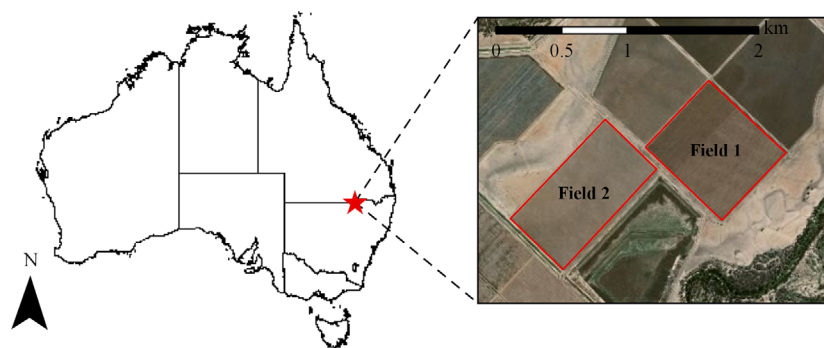


FIGURE 1 Study location at (left) Mungindi in Australia and (right) the two field boundaries used during the 2013/2014, 2015/2016, and 2016/2017 cotton seasons

temporal resolution soil and plant information due to costs and time involved. Over recent years, the quantity and quality of publicly available datasets, that can be freely used and redistributed by anyone, have substantially increased to the point where crop yield can be predicted within-field. For instance, satellites, such as Landsat 7-8, provide multispectral imagery at a 30-m resolution and 8-d frequency. Various soil attributes (90-m resolution), elevation (1–30 m resolution), and climatic conditions (5,000-m resolution) are also publicly available across Australia.

To date, little research has evaluated the potential of using only publicly available datasets to predict within-field yield in Australian cotton systems, specifically at key growth stages. Therefore, the primary aim of this study is to predict within-field cotton yield to inform in-season management. The specific objectives are to (a) evaluate the efficacy of publicly available datasets to predict within-field cotton yields at key growth stages, (b) compare the results from two machine-learning algorithms, and (c) identify which key parameters contributed most to each machine-learning model. These findings will enable growers and agronomists without access to high-resolution on-farm data to predict within-field cotton yield.

2 | METHODS AND MATERIALS

2.1 | Study area

The study was conducted on two furrow irrigated paddocks as part of a commercial farm located near Mungindi, New South Wales, Australia (28°58'S, 148°59'E) (Figure 1). The two field sites were of similar size (56 and 57 ha) and cultivated to cotton during the 2013/2014, 2015/2016, and 2016/2017 growing seasons. The 2014/2015 season was not planted to cotton and remained fallow. Sowing occurred in early October of each season and the fields remained fallow in between each cotton crop. The study area displays subtropical climatic conditions, with an average annual rainfall of 504 mm (BOM, 2020a). The mean annual temperature is 20.4 °C, reaching a high of 28.2 °C in January and a low of 11.8 °C in July (BOM,

2020a). The most common soil type in the region, gray vertisol, is characterized by a high clay content and water holding capacity (34–42% v/v) (Dalglish et al., 2012).

2.2 | Datasets

A total of 37 data layers including soil, rainfall, elevation, and remote sensing information were collected from publicly available data sources (Table 1). Soil data was collected from the Soil and Landscape Grid of Australia (SLGA), which provides detailed maps of Australia's soil attributes at a 90-m resolution (Viscarra Rossel et al., 2015). The SLGA provides various soil attributes, such as sand, clay, pH, effective cation exchange capacity (ECEC), and available water capacity (AWC) at six depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm). Gridded monthly rainfall was collected from the Scientific Information for Land Owners (SILO) database of Australian climate data, provided by the Bureau of Meteorology (BOM) (Jeffrey et al., 2001). Monthly rainfall was aggregated into pre-season (1 July–30 September), mid-season (1 October–31 December), and late season (1 January–31 March). Pre-season rainfall was included to indicate potential water storage for irrigation and available water in the soil profile. A 3-mo forecast probability of rainfall above the median was also available (BOM, 2020b). Digital elevation model (DEM) data (derived from Light Detection and Ranging [LiDAR]) was collected from Geoscience Australia at a 5-m resolution (Geoscience Australia, 2015).

Yield was recorded by a cotton yield-monitoring system on a John Deere cotton picker at both fields across the 2013/2014, 2015/2016, and 2016/2017 seasons. Yield data was provided at a 5-m resolution after being quality checked by an agronomist by removing any values outside ± 3 standard deviations of the mean yield. Remote-sensing data was collected from Landsat 7 and 8 (<https://earthexplorer.usgs.gov>) as close as possible to key cotton growth stages (squaring, flowering, and boll-fill) to align with N fertilizer and post-season decisions. Key growth stages were estimated based on sowing in early October. Landsat images (Collection 1 – Level 2) were atmospherically corrected by the U.S.

TABLE 1 Summary of data sources used in the machine learning models to predict within-field cotton lint yield at Mungindi (New South Wales)

Data	Data description	Resolution m	Date source
Yield	2014, 2016, and 2017	5	Farmer
Rainfall	Pre-season (1 July–30 Sept.)	5,000	BOM
	Mid-season (1 Oct.–31 Dec.)	5,000	BOM
	Late-season (1 Jan.–31 Mar.)	5,000	BOM
Soil (0–30 cm)	Available water capacity (AWC), %	90	SLGA
(30–60 cm)	Bulk density, g cm ⁻³	90	SLGA
(60–100 cm)	Clay, %	90	SLGA
	Effective cation exchange capacity (ECEC), meq/100 g	90	SLGA
	Total N, %	90	SLGA
	pH (CaCl ₂)	90	SLGA
	Total P, %	90	SLGA
	Silt, %	90	SLGA
	Sand, %	90	SLGA
	Organic C, %	90	SLGA
DEM	LiDAR	5	Geoscience Australia
Landsat-GNDVI	Squaring	30	NASA/USGS
	Flowering	30	NASA/USGS
	Boll-fill	30	NASA/USGS

Note. BOM, Bureau of Meteorology; SLGA, Soil and Landscape Grid of Australia; DEM, digital elevation model; GNDVI, green normalized difference vegetation index; USGS, U.S. Geological Survey.

Geological Survey (USGS): Earth Resources Observation & Science Center. The green normalized difference vegetation index (GNDVI) (Equation 1) was subsequently calculated using the near-infrared (NIR) and green bands:

$$\text{GNDVI} = \frac{\text{NIR} - \text{Green}}{\text{NIR} + \text{Green}} \quad (1)$$

GNDVI was selected over NDVI because it is less likely to saturate in high leaf area index and biomass conditions (Gitelson et al., 1996).

Yield and DEM data were resampled in ArcGIS Pro 2.3 from a resolution of 5–30 m using the average and projected to WGS-84 UTM Zone 55S. Both rainfall and soil data were not resampled due to the initial coarse spatial resolution and potential greater uncertainty. All data within 20 m of the field boundaries were removed prior to model development to avoid any edge effect.

2.3 | Yield prediction with machine learning

Two machine-learning algorithms, RF and GBM, were used to predict within-field cotton yield. Random forests use an

ensemble-learning method, consisting of many decision trees (Breiman, 2001). Random forests use a bagging technique, which involves splitting the dataset into homogenous subsets (trees) in parallel. When building each tree, RF randomly samples the training data and a random subset of features is used to create a predictive model. Final predictions are made by combining (bagging) all trees/models and using the average predicted results. The RF algorithm was optimized by adjusting the number of variables used at each split (*mtry*) and the number of trees (*ntrees*). The *mtry* values were evaluated using all variables and *ntrees* were tested with 200; 350; 500; 750; 1,000; and 2,000 trees. The *mtry* and *ntrees* with the lowest root mean square error (RMSE) were selected.

Similar to RF, GBM uses an ensemble learning method but uses boosting rather than bagging. In boosting, model predictions are improved by sequentially converting weak learners (variables) that are poorly correlated with the target variable into strong, well-correlated learners (Friedman, 2001). The GBM algorithm was tuned by adjusting the maximum tree depth (interaction depth), the number of boosting iterations or trees (*ntrees*), shrinkage, and minimum terminal node size (*n.minobsinnode*). All machine-learning model

parameters were adjusted using a grid search and 10-fold cross-validation.

The importance of each variable included in the machine-learning algorithms was evaluated with a recursive feature elimination (RFE) technique (Guyon et al., 2002). The RFE uses a backward selection elimination on variables, where a model is fitted using all variables and the least important variables are removed every loop. The model is rebuilt and the variables are ranked by importance. The RFE used the RF algorithm and was applied to all data layers in each model of the training dataset using a 10-fold cross-validation and ranked according to the normalized RMSE.

A total of six different models were created using each machine-learning algorithm. Three of these models included all variables (Table 1) and the other three included the optimal variables determined by the RFE. The six models created using all variables and optimal RFE variables were further characterized based on the data collected at three key cotton growth stages and crop management decisions: squaring, flowering, and boll-fill. The squaring model included all variables except late-season rainfall and GNDVI at flowering and boll fill. The flowering model included all variables except GNDVI at boll-fill. The boll-fill model included all variables. Soil variables and DEM data remained consistent throughout each year of the dataset; whereas yield, rainfall, and GNDVI changed each year.

Each model was validated using a “leave one year out” (or temporal) approach, where the model was trained using yield collected across both fields in 2014 and 2016 and tested on both fields the following year (2017). Due to the limited dataset of three seasons, each model was further validated using (a) yield collected in 2016 and 2017 and tested on 2014 yield, and (b) yield collected in 2014 and 2017 and tested on 2016 yield. Each model was optimized using a 10-fold cross-validation prior to validation. Machine-learning models were carried out in R Studio (RStudio Team, 2016) using the “rf” and “gbm” models in the *caret* package (Max, 2008).

Models were evaluated against observed yields using the coefficient of determination (R^2), RMSE, and Lin’s concordance correlation coefficient (LCCC) (Lin, 1989) (Equation 2). The LCCC is a measure of agreement between the predicted data and observed data. The LCCC ranges between -1 to 1 , with 1 being a perfect fit. The LCCC was selected over other indices as it is more appropriate for continuous data (Lin, 1989).

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2)$$

Where ρ_c is the LCCC, ρ is the correlation coefficient between the two variables in question, μ_x and μ_y are the

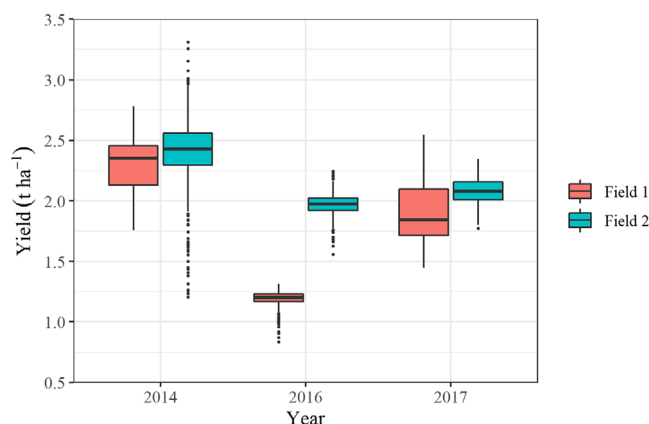


FIGURE 2 Variation in cotton lint yield at field sites in Mungindi (New South Wales) across the 2013/2014, 2015/2016, and 2016/2017 seasons

two variable means, and σ_x and σ_y are the corresponding variances.

3 | RESULTS

3.1 | Yield variation

Yields across both fields and all years displayed either high spatial or temporal variability (Figure 2). Seasonal yields (\pm standard deviation) in Field 1 were particularly variable, averaging 2.29 ± 0.21 t ha $^{-1}$ in 2014, 1.18 ± 0.06 t ha $^{-1}$ in 2016, and 1.91 ± 0.25 t ha $^{-1}$ in 2017. Yields in Field 2 displayed even greater spatial variability, particularly in 2014 ranging from 1.2 to 3.31 t ha $^{-1}$. Field 2 displayed relatively consistent average yields over the 3 yr, with yields averaging 2.41 ± 0.31 , 1.97 ± 0.09 , and 2.09 ± 0.1 t ha $^{-1}$ in 2014, 2016, and 2017, respectively. The average yield differences between fields were largest in 2016 (0.78 t ha $^{-1}$) and relatively small in 2014 (0.11 t ha $^{-1}$) and 2017 (0.17 t ha $^{-1}$).

3.2 | Recursive feature elimination

Figure 3 shows the normalized RMSE between estimated and predicted yields for each variable included in the squaring, flowering, and boll-fill models. To reduce computational time, the top 12 most important variables determined by RFE were selected as they provided sufficient data for accurate predictions. The 12 most important variables were relatively consistent across each growth stage model (Table 2). In each model, the GNDVI sampled during the corresponding growth stage was by far the most important variable. This was followed by either total N in the upper layers or GNDVI sampled during the previous growth stage. For example, in the boll-fill

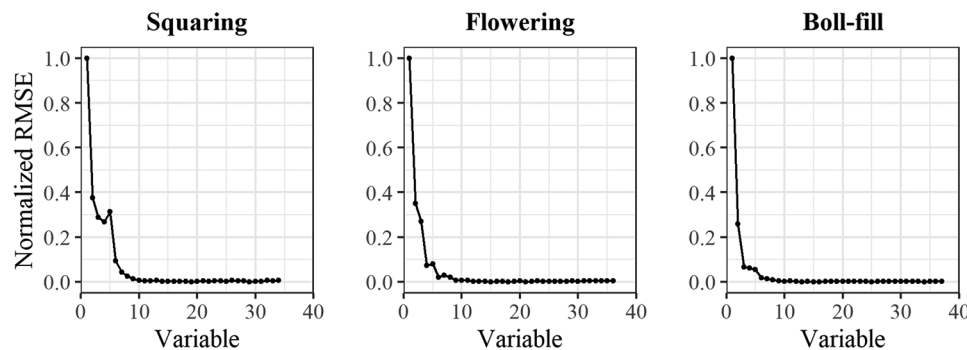


FIGURE 3 The normalized root mean square error (RMSE) (determined by recursive feature elimination, RFE) between observed and predicted cotton lint yields of each variable used in the training dataset for each growth stage model

TABLE 2 The top 12 most important variables to predict cotton lint yield at Mungindi (New South Wales) determined by recursive feature elimination (RFE) for each growth stage model

Variable	Squaring	Flowering	Boll-fill
1	GNDVI (squaring)	GNDVI (flowering)	GNDVI (boll-fill)
2	Total N (0–30 cm)	Total N (0–30 cm)	GNDVI (flowering)
3	Rain (mid-season)	GNDVI (squaring)	Total N (0–30 cm)
4	Silt (60–100 cm)	DEM	Rain (late-season)
5	DEM	Organic C (30–60 cm)	Total N (30–60 cm)
6	Rain (pre-season)	Organic C (0–30 cm)	GNDVI (squaring)
7	Clay (60–100 cm)	Silt (60–100 cm)	Organic C (60–100 cm)
8	Silt (30–60 cm)	Total N (30–60 cm)	Clay (60–100 cm)
9	Organic C (0–30 cm)	Organic C (60–100 cm)	Organic C (30–60 cm)
10	Bulk density (0–30 cm)	Clay (60–100 cm)	DEM
11	Organic C (30–60 cm)	Silt (30–60 cm)	Organic C (0–30 cm)
12	Silt (0–30 cm)	Rain (late-season)	Total P (0–30 cm)

Note. GNDVI, green normalized difference vegetation index; DEM, digital elevation model.

model, GNDVI calculated during the flowering stage was the second most important variable after GNDVI calculated during boll-fill. Silt, clay, organic C, DEM, and rainfall were also identified as significant variables across all models but had a much lower contribution. The AWC, bulk density, ECEC, pH, total P, and sand were the least important predictor variables in all models.

3.3 | Yield prediction

The number and type of parameters adjusted to optimize each machine-learning model varied according to the growth stage and number of variables included in the model (Table 3). In RF, the *mtry* varied from 12 to 18 using all variables and 4 to 6 using the top 12 variables across all growth stage models. The *ntrees* ranged from 200 to 2,000 across all RF models created. In the GBM model, interaction depth ranged from 12 to 34. Shrinkage, *n.minobsinnode*, and *ntrees* were consistent

at 0.1; 10; and 2,000; respectively, across all growth stages and number of variables used. The only exception was GBM at squaring using all variables with 1,000 *ntrees*.

The accuracy of within-field yield prediction in 2017 using all variables varied according to the growth stage data included in the RF and GBM models (Figure 4). Across all growth stages, the R^2 ranged from .15 to .52, RMSE ranged from 0.15 to 0.26 t ha⁻¹, and LCCC ranged from 0.25 to 0.66. At squaring, RF and GBM provided a poor agreement between observed and predicted yields (R^2 = .18 and .15, RMSE = 0.21 and 0.26 t ha⁻¹, LCCC = 0.32 and 0.25, respectively). Model performance was improved at the flowering growth stage, with the highest LCCC by the GBM model of 0.66, followed by a LCCC of 0.50 from the RF model. At the boll-fill growth stage, the RF model had the highest accuracy with a R^2 = .52, RMSE = 0.15 t ha⁻¹ and LCCC = 0.62. The GBM model at boll-fill decreased in accuracy compared with flowering to a R^2 = .40, RMSE = 0.20 t ha⁻¹, and LCCC = 0.55.

TABLE 3 Optimal parameters selected for each model to predict cotton lint yield at Mungindi (New South Wales) based on the lowest root mean square error (RMSE)

Parameter	RF (all variables)			RF (top 12 variables)		
	Squaring	Flowering	Boll-fill	Squaring	Flowering	Boll-fill
<i>mtry</i>	13	18	12	4	6	5
<i>ntrees</i>	200	1000	750	200	1000	2000
Parameter	GBM (all variables)			GBM (top 12 variables)		
	Squaring	Flowering	Boll-fill	Squaring	Flowering	Boll-fill
Interaction depth	26	34	22	12	12	12
<i>ntrees</i>	1000	2000	2000	2000	2000	2000
Shrinkage	0.01	0.01	0.01	0.01	0.01	0.01
<i>n.minobsinnode</i>	10	10	10	10	10	10

Note. RF, random forests; *mtry*, number of variables used at each split; *ntrees*, number of trees; GBM, gradient boosting machines.

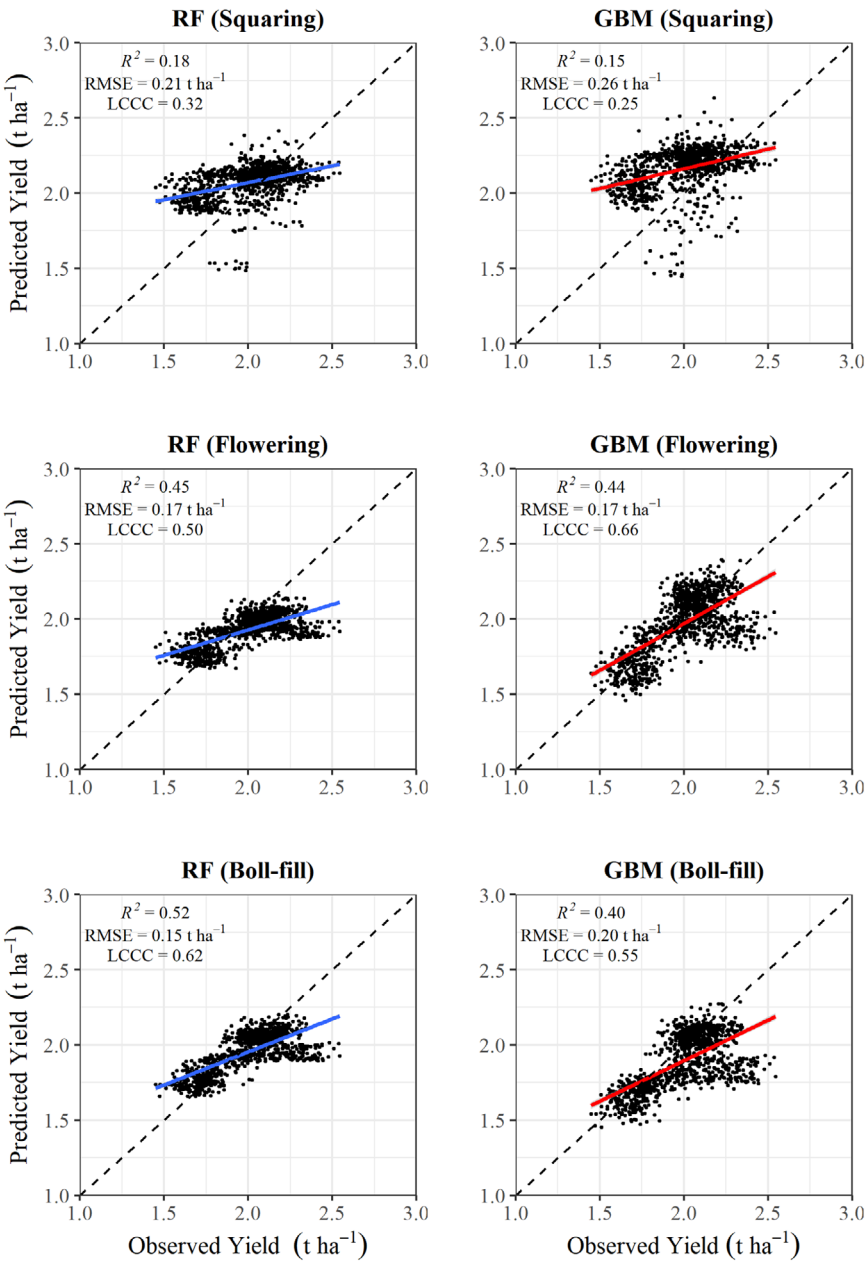


FIGURE 4 Observed cotton lint yields compared with predicted yields across both fields at Mungindi (New South Wales) in 2016/2017 using all variables in the random forests (RF) and gradient boosting machine (GBM) models

TABLE 4 Model accuracy (R^2 , root mean square error, RMSE [t ha^{-1}] and Lin's concordance correlation coefficient, LCCC) of observed cotton lint yields compared with predict yields across both fields at Mungindi (New South Wales) using all variables in the random forests (RF) and gradient boosting machines (GBM) models

Predicted year	Model	Squaring			Flowering			Boll-fill		
		R^2	RMSE	LCCC	R^2	RMSE	LCCC	R^2	RMSE	LCCC
			t ha^{-1}			t ha^{-1}			t ha^{-1}	
2014	RF	.31	0.86	0.03	.32	0.77	0.03	.48	0.73	0.06
	GBM	.27	0.80	0.04	.22	0.81	0.03	.46	0.83	0.05
2016	RF	.03	0.84	0.03	.06	0.81	-0.06	.30	0.80	0.08
	GBM	.08	0.82	0.06	.03	0.81	-0.04	.70	0.80	0.16

The further model validation using different combinations of years in the training and testing dataset provided poor results (Table 4). The prediction of 2014 yield using 2016 and 2017 as the training data obtained a $R^2 = .31$ –.48, RMSE = 0.73–0.86 t ha^{-1} and LCCC = 0.03 to 0.06 using the RF and GBM models across all growth stages. Similar results were obtained when predicting 2016 yield using 2014 and 2017 as the training dataset. This provided a $R^2 = .03$ –.70, RMSE = 0.80 to 0.84 t ha^{-1} , and LCCC = -0.06 to 0.16 using the RF and GBM models across all growth stages.

Within-field yield predictions in 2017 using the top 12 most important variables provided very similar and, in some cases, better results compared with models using all variables (Figure 5). At squaring, both RF and GBM models provided a poor agreement between observed and predict yields with a maximum LCCC of 0.29. At flowering, RF and GBM provided similar results with a $R^2 = .39$, RMSE = 0.17 and 0.19 t ha^{-1} and LCCC = 0.55 and 0.60, respectively. The RF model RMSE and LCCC at flowering improved compared with using all variables. At boll-fill, both RF and GBM models tended to slightly underestimate yields but provided a moderate LCCC of 0.57 and 0.55 and RMSE of 0.17 and 0.19 t ha^{-1} , respectively.

The predicted yield maps of the two models with the highest LCCC using all variables (GBM at flowering and RF at boll-fill) provided similar results to observed yields in both fields (Figure 6). Observed yields in Field 1 ranged from 1.45–2.53 t ha^{-1} , with the lowest yields occurring in the Northeast section closest to the head ditch. Predicted yield in Field 1 by RF at boll-fill and GBM at flowering ranged from 1.65 to 2.02 t ha^{-1} and 1.47 to 2.09 t ha^{-1} , respectively, with similar yields reported near the head ditch compared to observed yields but much lower yields near the tail ditch (Southwest section). Observed and predicted yield in Field 2 had a similar range of approximately 1.80–2.40 t ha^{-1} . In Field 2, the lowest yields were observed and predicted near the tail ditch in the Northwest section of the field and a strip running from Northeast to Southwest through the field.

4 | DISCUSSIONS

4.1 | Model performance

This is the first study to assess the capability of using publicly available datasets and machine learning to predict within-field yield at key growth stages in Australian cotton production systems. Previous studies have mostly focused on the use of on-farm or remote-sensing data sources (Feng, Zhou et al., 2020; Haghverdi et al., 2018; Huang et al., 2016; Huang et al., 2013; Meng et al., 2019; Thomasson et al., 2004). Other studies have focused on publicly available datasets, but were either conducted on one field site or neglected the squaring growth stage when N fertilizer is commonly applied (Filippi et al., 2020; Nguyen et al., 2019). In the current study, the use of publicly available datasets to predict lint yield across two field sites provided poorer predictions at squaring (RMSE = 0.21–26 t ha^{-1} , LCCC = 0.25–0.32) compared with later in the season at flowering (RMSE = 0.17 t ha^{-1} , LCCC = 0.50–0.66) and boll-fill (RMSE = 0.15–0.20 t ha^{-1} , LCCC = 0.55–0.62). These findings are in agreement with other studies where the addition of more data improved yield predictions as the season progressed (Filippi, Jones, et al., 2019; Filippi et al., 2020; Huang et al., 2013; Nguyen et al., 2019; Thomasson et al., 2004). The improved predictions in the current study later in the season were primarily attributed to stronger relationships between GNDVI and yield. Several authors have also found higher R^2 values between cotton yield and vegetation indices later in the season at flowering compared with squaring (Ballester et al., 2017; Bronson et al., 2003; Yang et al., 2007; Zhao et al., 2007). This is due to minimal within-field spatial variability during early growth stages and the potential negative impact of soil background reflectance.

The capability of both machine-learning models to predict yields as early as flowering allows the potential for growers to inform management decisions in-season, such as N fertilizer applications. Growers typically apply N fertilizer in-season between the squaring to flowering growth stages (CRDC & CottonInfo, 2018). The yield prediction models created using

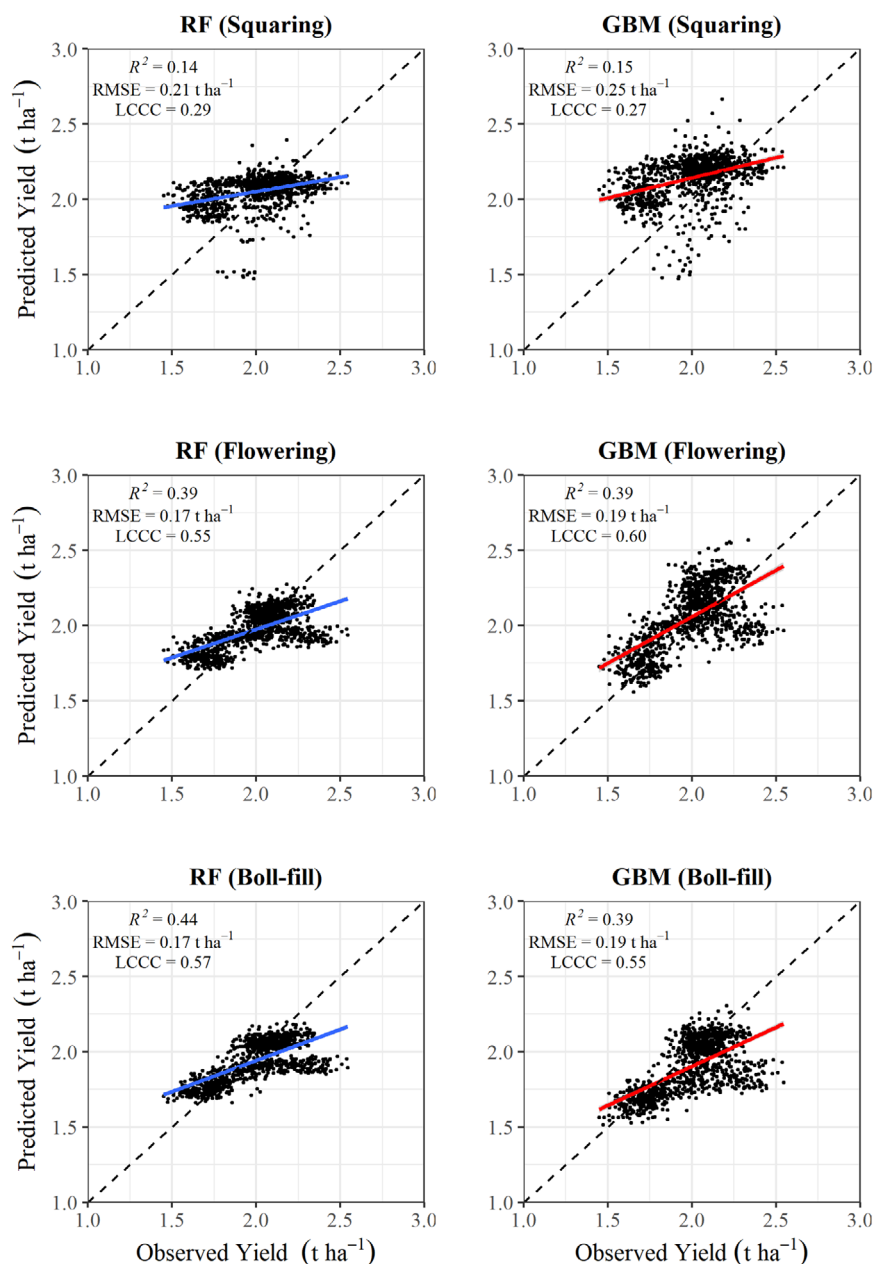


FIGURE 5 Observed cotton lint yields compared with predicted yields across both fields at Mungindi (New South Wales) in 2016/2017 using the top 12 most important variables, determined by the recursive feature elimination (RFE), in the random forests (RF) and gradient boosting machine (GBM) models

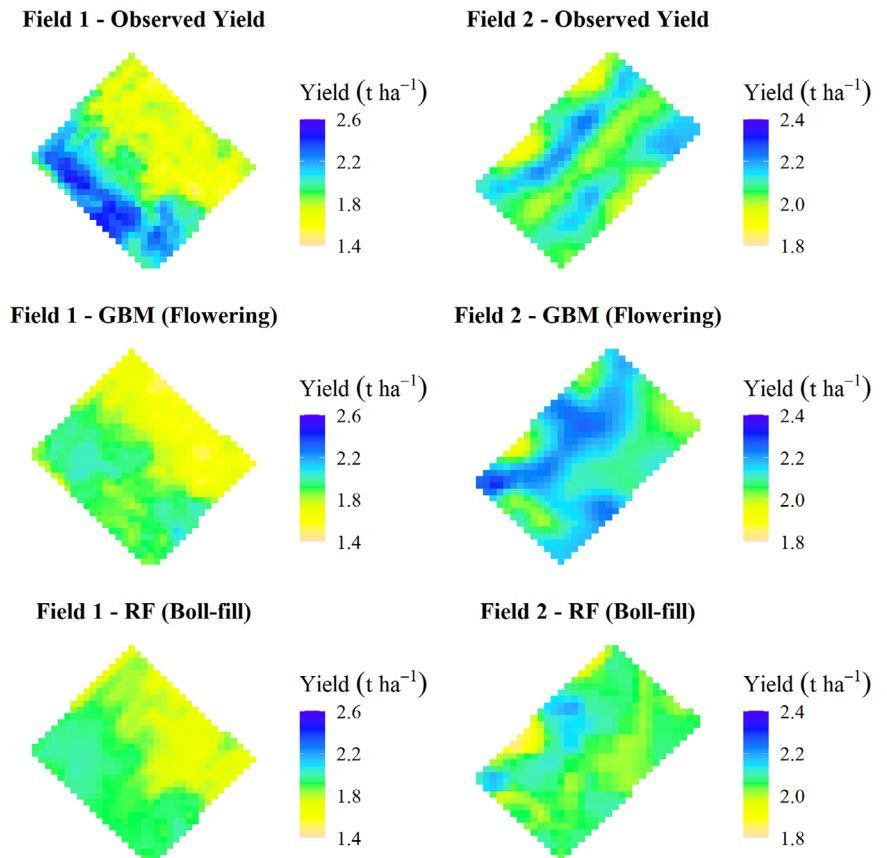
data up to squaring provided poor results and is, therefore, too early in the season to inform variable N applications. In contrast, model accuracy was improved at the flowering growth stage and can be used to inform N management. Although the boll-fill models performed equally with the flowering models, N management is too late to impact yield. However, it can provide information as an aid to estimate revenue and guide investments for the following season.

The flowering and boll-fill models, however, were negatively impacted by the saturation of GNDVI. This observation was particularly apparent in Field 1, where predicted yields were underestimated near the tail ditch by ~ 0.5 t ha⁻¹ (Figure 6). The models developed at time of flowering were less impacted by GNDVI saturation compared with the models developed at boll-fill. This is likely due to a higher leaf area

index from full canopy closure during boll-fill. The GNDVI was initially selected over NDVI because it is less likely to saturate (Gitelson et al., 1996). However, GNDVI saturation has still been reported in other studies during high leaf area index conditions (Krishna, 2018; Kross et al., 2015; Tan et al., 2013). Other vegetation indices less prone to saturation such as the enhanced vegetation index (EVI) and renormalized difference vegetation index (RDVI) were also tested but displayed similar results, with saturation at higher values (data not shown).

Red-edge based vegetation indices are also less prone to saturation and have provided stronger relationships with lint yield compared with red or green-based vegetation indices in previous studies (Ballester et al., 2017; Bronson et al., 2020; Raper & Varco, 2015). However, Landsat 7–8 (used

FIGURE 6 Observed cotton lint yields and predicted yield maps for Fields 1 and 2 at Mungindi (New South Wales) using the gradient boosting machine (GBM) model at the flowering stage and the random forests (RF) model at the boll-fill stage using all variables



in the current study) was not equipped with red-edge and was the only publicly available multispectral satellite imagery available during all seasons and at a within-field resolution. The recently launched Sentinel-2A in 2015 and Sentinel-2B in 2017 are equipped with higher resolution multispectral cameras that include the red-edge band and have potential to provide greater prediction accuracy. Sentinel-2 was not used in the current study due to the lack of data for the 2013/2014 season. The effectiveness of red-edge and other satellite imagery need to be validated in further research.

Aside from the negative impact of GNDVI saturation, the prediction accuracy of the models were largely attributed to the inclusion of historical yields in the training dataset. This study focused on the “leave one year out” (temporal) cross-validation rather than the “leave one field out” (spatial) approach because the latter has provided poor results in other studies (Fajardo et al., 2019). Similarly, Filippi, Jones, et al. (2019) compared both approaches and found the temporal approach provided much better results when predicting grain yield at field level. However, as the current study was conducted over a small-term scale of 3 yr and two fields, this creates uncertainty regarding the validity of the models on a medium to long-term scale and transferability to other fields and regions. The limitations of the current limited dataset were particularly apparent when further validating

the models using other years of yield data as the training and testing dataset, where low accuracy was observed (Table 4). The model accuracy should, therefore, be further explored with a greater timeseries of yield data and field sites.

In addition to a greater timeseries of data, the inclusion of site-specific contextual information provided by the grower, such as N fertilizer rates, irrigation, sowing dates, presence of pests or diseases, cultivar, available N etc. could also provide a better understanding of the drivers of seasonal variability but also determine whether additional grower information or detailed sampling can further improve model accuracy. If additional data is collected from other fields and more seasons, management decisions such as N fertilizer rates could be adjusted in the model to optimize yield; however, this needs to be further explored.

Another important factor to improve model accuracy is to determine at what spatial scale these models provide the greatest accuracy. This study explored a small area of two fields and provided a maximum LCCC of 0.66, indicating that a fine-scale model can potentially provide greater accuracy when predicting within-field yield compared to large-scale models at the regional level (Filippi et al., 2020). However, this needs to be further explored and validated with the inclusion of more years of yield data, field sites, and contextual information.

TABLE 5 Model accuracy (R^2 , root mean square error, RMSE [t ha^{-1}] and Lin's concordance correlation coefficient, LCCC) of observed cotton lint yields compared with predict yields across both fields at Mungindi (New South Wales) using the top 5, 7, and 10 most important variables, determined by the recursive feature elimination (RFE), in the random forests (RF) and gradient boosting machines (GBM) models

No. of variables	Model	Squaring			Flowering			Boll-fill		
		R^2	RMSE	LCCC	R^2	RMSE	LCCC	R^2	RMSE	LCCC
			t ha^{-1}			t ha^{-1}			t ha^{-1}	
5	RF	.14	0.36	0.18	.40	0.49	0.21	.45	0.25	0.37
	GBM	.13	0.35	0.18	.42	0.53	0.22	.34	0.30	0.29
7	RF	.16	0.20	0.28	.44	0.46	0.21	.44	0.17	0.53
	GBM	.14	0.24	0.27	.42	0.54	0.21	.39	0.19	0.61
10	RF	.19	0.20	0.33	.44	0.45	0.22	.44	0.16	0.60
	GBM	.16	0.24	0.28	.41	0.55	0.21	.38	0.18	0.61

4.2 | Model comparison

The comparison of RF and GBM algorithms provided similar results across all growth stages and the number of variables included (Figures 4 and 5). In most cases, RF provided a stronger agreement between observed and predicted yield compared with GBM. Gradient boosting machines provided stronger predictions in terms of LCCC and RMSE only once, in the flowering model using all variables. The superior results of RF are in agreement with other studies where RF provided higher predictive accuracy compared with other machine-learning algorithms for predicting yield or other crop parameters (Kayad et al., 2019; Liang et al., 2015; Richetti et al., 2018; Wu et al., 2019). In the current study, the superior performance of RF compared with GBM is attributed to how each machine-learning algorithm builds the optimal model. Both RF and GBM use ensemble-learning methods, where several models are combined to improve overall model performance (Okun et al., 2011). In GBM, weak learners (variables) are converted to strong learners sequentially in order to decrease bias from highly correlated variables. In contrast, RF generates models (trees) from a random subset of the training data and then averages all trees in order to decrease variance. The dataset used in this study displayed high spatial and temporal variability and therefore, RF proved more effective than GBM due its greater robustness to variable or noisy data.

4.3 | Variable importance

The application of a feature elimination method, such as an RFE, prior to model development indicated potential to reduce the number of variables without comprising model accuracy. Identifying key drivers of yield variability allows growers the opportunity to target these specific variables on-farm to save sample analysis costs and provide greater accuracy. In the current study, the top 12 variables determined by the RFE best replicated observed results compared with using

all variables across all growth stages. Initially, the top 5, 7, and 10 most important variables were tested (Table 5). The squaring and boll-fill models did not substantially improve after the top 7 most important variables were included in the model. Boll-fill model accuracy improved once GNDVI at squaring was included in model. At squaring, GNDVI averaged 0.38 in 2014, 0.30 in 2016, and 0.37 in 2017 across both fields. The higher GNDVI reflectance in 2017 alerted the model that it was underestimating yield based on the GNDVI reflectance recorded in the training dataset.

A similar situation occurred in the flowering growth stage models, which improved once the top 12 most important variables were included from the RFE. In this case, predicted yield at flowering was initially overestimated prior to the inclusion of forecasted rain late season. Forecasted rain late season in 2017 was substantially lower (123.6 mm) compared with 2014 (220 mm) and 2016 (193 mm). Cotton is a water-intensive crop and the lower rainfall received in 2017 negatively impacted yield, particularly in Field 1.

The RFE also determined soil texture (clay and silt), total N, DEM, and organic C were other key predictor variables in all models. Similar to rainfall, soil texture played a large role because of its influence on water availability. In addition to water, N played a large role as it is required in large quantities to maximize yields (Rochester & Bange, 2016). However, N included in the current study was total N and not available N that typically informs N fertilizer strategies. This provides further scope for available N levels, particularly prior to sowing each season, to be included in the yield prediction models to inform N management.

Furthermore, since both fields were furrow irrigated, slope (DEM) played a crucial role in yield spatial patterns as the transport of nutrients and water down the paddock resulted in higher yields near the tail ditch (McHugh et al., 2008). However, slope may not be an important predictor variable in other irrigation systems, such as overhead irrigation. Other factors, including irrigation amounts and growing degree days, are important factors influencing cotton growth and

yield and should be considered as variables in future models. It is recommended to conduct an RFE when more data is added to the model to understand which predictor variables are key drivers under different management strategies and environmental conditions. Once a thorough understanding of which variables are the key drivers of yield, the grower or agronomist can target these variables with precision sampling to provide a greater accuracy and resolution for predicting yield. This would provide a more economically efficient sampling scheme compared with the traditional approach of grid sampling.

5 | CONCLUSIONS

This study demonstrated the potential of using publicly available datasets and machine-learning algorithms to predict within-field cotton yield. The results showed strong agreement between predicted and observed yields in both RF and GBM, particularly at the flowering growth stage. This shows great potential for growers to make more informed decisions regarding in-season investments, such as N fertilizer applications and post-season decision-making. Furthermore, reducing the number of variables in the machine-learning models did not negatively impact the prediction accuracy. This allows growers the potential to target these specific variables with greater accuracy and resolution to save costs. The inclusion of more fields and years of data, contextual information, and higher resolution satellite imagery equipped with the red-edge band have potential to further improve model accuracy and better inform management decisions.

ACKNOWLEDGMENTS


This research was funded by the Australian Cotton Research Development Corporation as part of the “Future Farm” project (QUT1902). We would like to acknowledge the assistance from the growers in Mungindi for providing the yield data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Stephen Leo  <https://orcid.org/0000-0002-7825-0662>

Massimiliano De Antoni Migliorati  <https://orcid.org/0000-0002-5651-5834>

Peter R. Grace  <https://orcid.org/0000-0003-4136-4129>

REFERENCES

- Archontoulis, S. V., & Miguez, F. E. (2015). Nonlinear regression models and applications in agricultural research. *Agronomy Journal*, 107(2), 786–798. <https://doi.org/10.2134/agronj2012.0506>
- Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES). (2019). *Agricultural commodities and trade data*. Retrieved from <https://www.agriculture.gov.au/abares/research-topics/agricultural-commodities/agricultural-commodities-trade-data#2019>
- Ballester, C., Hornbuckle, J., Brinkhoff, J., Smith, J., & Quayle, W. (2017). Assessment of in-season cotton nitrogen status and lint yield prediction from unmanned aerial system imagery. *Remote Sensing*, 9(11), 1149. <https://doi.org/10.3390/rs9111149>
- Basso, B., Ritchie, J. T., Pierce, F. J., Braga, R. P., & Jones, J. W. (2001). Spatial validation of crop models for precision agriculture. *Agricultural Systems*, 68(2), 97–112. [https://doi.org/10.1016/S0308-521X\(00\)00063-9](https://doi.org/10.1016/S0308-521X(00)00063-9)
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bronson, K. F., Chua, T. T., Booker, J. D., Keeling, J. W., & Lascano, R. J. (2003). In-season nitrogen status sensing in irrigated cotton: II. Leaf nitrogen and biomass. *Soil Science Society of America Journal*, 67(5), 1439–1448.
- Bronson, K., Conley, M., French, A., Hunsaker, D., Thorp, K., & Barnes, E. (2020). Which active optical sensor vegetation index is best for nitrogen assessment in irrigated cotton? *Agronomy Journal*, 112, 2205–2218. <https://doi.org/10.1002/agj2.20120>
- Bureau of Meteorology (BOM). (2020a). *Climate data online*. Retrieved from <http://www.bom.gov.au/climate/data/>
- Bureau of Meteorology (BOM). (2020b). *Climate outlooks—Weeks, months and seasons*. Retrieved from <http://www.bom.gov.au/climate/outlooks/#/rainfall/median/seasonal/0>
- Chen, T., Zeng, R., Guo, W., Hou, X., Lan, Y., & Zhang, L. (2018). Detection of stress in cotton (*Gossypium hirsutum* L.) caused by aphids using leaf level hyperspectral measurements. *Sensors*, 18(9), 2798. <https://doi.org/10.3390/s18092798>
- Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
- Cotton Research Development Corporation (CRDC), & CottonInfo. (2018). *NUTRIpak: A practical guide to cotton nutrition*. Retrieved from <https://cottoninfo.com.au/publications/nutripak>
- Dalgliesh, N., Cocks, B., & Horan, H. (2012). *APSoil-providing soils information to consultants, farmers and researchers*. Australian Agronomy Conference, 14–18th October 2012. Armidale, Australia: Australian Society of Agronomy. Retrieved from http://www.regional.org.au/au/asa/2012/soil-water-management/7993_dalglieshnp.htm#TopOfPage
- Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, 36(2), 1–9. <https://doi.org/10.1007/s13593-016-0364-z>
- Fajardo, M., Whelan, B., Filippi, P., & Bishop, T. (2019). Wheat yield forecast using contextual spatial information. In J. V. Stafford (Ed.), *Precision agriculture '19* (pp. 713–719). Wageningen, the Netherlands: Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-888-9_88
- Feng, A., Zhou, J., Vories, E. D., Sudduth, K. A., & Zhang, M. (2020). Yield estimation in cotton using UAV-based multi-sensor imagery. *Biosystems Engineering*, 193, 101–114. <https://doi.org/10.1016/j.biosystemseng.2020.02.014>

- Filippi, P., Bishop, T. F. A., & Whelan, B. (2019). Identifying yield stability and drivers of yield variability in cotton using multi-layered, whole-farm datasets. In J. V. Stafford (Ed.), *Precision agriculture '19* (pp. 45–52). Wageningen, the Netherlands: Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-888-9_4
- Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D. S. N., Pozza, L. E., Ugbaje, S. U., Jephcott, T. G., Paterson, S. E., Whelan, B. M., & Bishop, T. F. A. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20(5), 1015–1029. <https://doi.org/10.1007/s11119-018-09628-4>
- Filippi, P., Whelan, B. M., Vervoort, R. W., & Bishop, T. F. A. (2020). Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. *Agricultural Systems*, 184, 102894. <https://doi.org/10.1016/j.agsy.2020.102894>
- Food and Agriculture Organization of the United Nations (FAO). (2019). *FAO Global Statistical Yearbook, FAO Regional Statistical Yearbooks*. Retrieved from <http://fao.org/faostat/en/#data/QC>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Geoscience Australia. (2015). *Digital elevation model (DEM) of Australia derived from LiDAR 5 metre grid*. Canberra, Australia: Geoscience Australia. Retrieved from <http://pid.geoscience.gov.au/dataset/ga/89644>
- Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58(3), 289–298. [https://doi.org/10.1016/S0034-4257\(96\)00072-7](https://doi.org/10.1016/S0034-4257(96)00072-7)
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1), 389–422. <https://doi.org/10.1023/A:1012487302797>
- Haghverdi, A., Washington-Allen, R. A., & Leib, B. G. (2018). Prediction of cotton lint yield from phenology of crop indices using artificial neural networks. *Computers and Electronics in Agriculture*, 152, 186–197. <https://doi.org/10.1016/j.compag.2018.07.021>
- Hatfield, J. L., Gitelson, A. A., Schepers, J. S., & Walthall, C. L. (2008). Application of spectral remote sensing for agronomic decisions. *Agronomy Journal*, 100(S3), S-117–S-131. <https://doi.org/10.2134/agronj2006.0370c>
- He, L., & Mostovoy, G. (2019). Cotton yield estimate using Sentinel-2 data and an ecosystem model over the southern US. *Remote Sensing*, 11(17), 2000. <https://doi.org/10.3390/rs11172000>
- Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., ... Keating, B. A. (2014). APSIM – Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62, 327–350. <https://doi.org/10.1016/j.envsoft.2014.07.009>
- Hoogenboom, G., White, J. W., & Messina, C. D. (2004). From genome to crop: Integration through simulation modeling. *Field Crops Research*, 90(1), 145–163. <https://doi.org/10.1016/j.fcr.2004.07.014>
- Horie, T., Yajima, M., & Nakagawa, H. (1992). Yield forecasting. *Agricultural Systems*, 40(1), 211–236. [https://doi.org/10.1016/0308-521X\(92\)90022-G](https://doi.org/10.1016/0308-521X(92)90022-G)
- Huang, Y., Brand, H. J., Sui, R., Thomson, S., Furukawa, T., & Ebelhar, M. W. (2016). Cotton yield estimation using very high-resolution digital images acquired with a low-cost small unmanned aerial vehicle. *Transactions of the ASABE*, 59, 1563–1574. <https://doi.org/10.13031/trans.59.11831>
- Huang, Y., Sui, R. X., Thomson, S. J., & Fisher, D. K. (2013). Estimation of cotton yield with varied irrigation and nitrogen treatments using aerial multispectral imagery. *International Journal of Agricultural and Biological Engineering*, 6(2), 37–41. <https://doi.org/10.3965/j.ijabe.20130602.005>
- Jeffrey, S. J., Carter, J. O., Moodie, K. B., & Beswick, A. R. (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, 16(4), 309–330. [https://doi.org/10.1016/S1364-8152\(01\)00008-1](https://doi.org/10.1016/S1364-8152(01)00008-1)
- Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., ... Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, 18(3), 235–265. [https://doi.org/10.1016/S1161-0301\(02\)00107-7](https://doi.org/10.1016/S1161-0301(02)00107-7)
- Kayad, A., Sozzi, M., Gatto, S., Marinello, F., & Pirotti, F. (2019). Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. *Remote Sensing*, 11(23), 2873. <https://doi.org/10.3390/rs11232873>
- Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., & Shearer, S. (2018). Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and Electronics in Agriculture*, 153, 213–225. <https://doi.org/10.1016/j.compag.2018.07.016>
- Kim, N., Ha, K., Park, N., Cho, J., Hong, S. Y., & Lee, Y. (2019). A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern United States, 2006–2015. *ISPRS International Journal of Geo-Information*, 8, 240. <https://doi.org/10.3390/ijgi8050240>
- Krishna, K. R. (2018). Drones in production agronomy. In K. R. Krishna (Ed.), *Agricultural drones: A peaceful pursuit* (pp. 153–220). New York: Apple Academic Press.
- Kross, A., McNairn, H., Lapen, D., Sunohara, M., & Champagne, C. (2015). Assessment of RapidEye vegetation indices for estimation of leaf area index and biomass in corn and soybean crops. *International Journal of Applied Earth Observation and Geoinformation*, 34, 235–248. <https://doi.org/10.1016/j.jag.2014.08.002>
- Li, Y., Shi, Z., Wu, C., Li, H., & Li, F. (2008). Determination of potential management zones from soil electrical conductivity, yield and crop data. *Journal of Zhejiang University SCIENCE B*, 9(1), 68–76. <https://doi.org/10.1631/jzus.B071379>
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674. <https://doi.org/10.3390/s18082674>
- Liang, L., Di, L., Zhang, L., Deng, M., Qin, Z., Zhao, S., & Lin, H. (2015). Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sensing of Environment*, 165, 123–134. <https://doi.org/10.1016/j.rse.2015.04.032>
- Lin, L. I. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268. <https://doi.org/10.2307/2532051>
- Lobell, D. B. (2013). The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143, 56–64. <https://doi.org/10.1016/j.fcr.2012.08.008>
- Macdonald, B., Latimer, J., Schwenke, G., Nachimuthu, G., & Baird, J. (2018). The current status of nitrogen fertiliser use efficiency and future research directions for the Australian cotton industry. *Cotton Research*, 1(1), 15. <https://doi.org/10.1186/s42397-018-0015-9>

- Max, K. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 143147. <https://doi.org/10.18637/jss.v028.i05>
- McHugh, A. D., Bhattarai, S., Lotz, G., & Midmore, D. J. (2008). Effects of subsurface drip irrigation rates and furrow irrigation for cotton grown on a vertisol on off-site movement of sediments, nutrients and pesticides. *Agronomy for Sustainable Development*, 28(4), 507–519. <https://doi.org/10.1051/agro:2008034>
- Meng, L., Liu, H., Zhang, X., Ren, C., Ustin, S. L., Qiu, Z., Xu, M., & Guo, D. (2019). Assessment of the effectiveness of spatiotemporal fusion of multi-source satellite images for cotton yield estimation. *Computers and Electronics in Agriculture*, 162, 44–52. <https://doi.org/10.1016/j.compag.2019.04.001>
- Meng, L., Zhang, X., Liu, H., Guo, D., Yan, Y., Qin, L., & Pan, Y. (2017). Estimation of cotton yield using the reconstructed time-series vegetation index of Landsat data. *Canadian Journal of Remote Sensing*, 43, 244–255. <https://doi.org/10.1080/07038992.2017.1317206>
- Nguyen, L. H., Zhu, J., Lin, Z., Du, H., Yang, Z., Guo, W., & Jin, F. (2019). Spatial-temporal multi-task learning for within-field cotton yield prediction. In Q. Yang, Z. Zhou, Z. Gong, M. Zhang, & S. Huang (Eds.), *Advances in knowledge discovery and data mining* (pp. 343–354). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-16148-4_27
- Okun, O., Valentini, G., & Re, M. (2011). *Ensembles in machine learning applications*. Heidelberg, Germany: Springer. <https://doi.org/10.1007/978-3-642-22910-7>
- Ping, J. L., Green, C. J., Zartman, R. E., Bronson, K. F., & Morris, T. F. (2007). Spatial variability of soil properties, cotton yield, and quality in a production field. *Communications in Soil Science and Plant Analysis*, 39(1–2), 1–16. <https://doi.org/10.1080/00103620701758840>
- Raper, T. B., & Varco, J. J. (2015). Canopy-scale wavelength and vegetative index sensitivities to cotton growth parameters and nitrogen status. *Precision Agriculture*, 16(1), 62–76. <https://doi.org/10.1007/s11119-014-9383-4>
- Ray, S. S., Pokharna, S. S., & Ajai (1999). Cotton yield estimation using agrometeorological model and satellite-derived spectral profile. *International Journal of Remote Sensing*, 20(14), 2693–2702. <https://doi.org/10.1080/014311699211741>
- Richetti, J., Judge, J., Boote, K. J., Johann, J. A., Uribe-Opazo, M. A., Becker, W. R., ... Silva, L. C. D. A. (2018). Using phenology-based enhanced vegetation index and machine learning for soybean yield estimation in Paraná State, Brazil. *Journal of Applied Remote Sensing*, 12(2), 026029. <https://doi.org/10.1117/1.JRS.12.026029>
- Rochester, I. J., & Bange, M. (2016). Nitrogen fertiliser requirements of high-yielding irrigated transgenic cotton. *Crop and Pasture Science*, 67(6), 641–648. <https://doi.org/10.1071/CP15278>
- RStudio Team. (2016). *RStudio: Integrated development for R*. Boston, MA: RStudio, Inc.
- Shahhosseini, M., Martinez-Feria, R., Hu, G., & Archontoulis, S. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters*, 14(12), 124026. <https://doi.org/10.1088/1748-9326/ab5268>
- Sun, S., Li, C., Paterson, A. H., Chee, P. W., & Robertson, J. S. (2019). Image processing algorithms for infield single cotton boll counting and yield prediction. *Computers and Electronics in Agriculture*, 166, 104976. <https://doi.org/10.1016/j.compag.2019.104976>
- Tan, C., Samanta, A., Jin, X., Tong, L., Ma, C., Guo, W., Knyazikhin, Y., & Myneni, R. B. (2013). Using hyperspectral vegetation indices to estimate the fraction of photosynthetically active radiation absorbed by corn canopies. *International Journal of Remote Sensing*, 34(24), 8789–8802. <https://doi.org/10.1080/01431161.2013.853143>
- Thomasson, J. A., Wooten, J. R., Gogineni, S., Sui, R., & Kolla, B. M. (2004). Remote sensing and weather information in cotton yield prediction. In W. Gao & D. R. Shaw (Eds.), *Ecosystems' Dynamics, Agricultural Remote Sensing and Modeling, and Site-Specific Agriculture*. Society of Photo-Optical Instrumentation Engineers (SPIE) 48th Annual Meeting, San Diego, CA. *Proceedings of the optical science and technology, SPIE's 48th Annual Meeting, San Diego, CA* (Vol. 5153, pp. 127–135). 3–8 August 2003. Bellingham, USA: The International Society for Optics and Photonics. <https://doi.org/10.1117/12.506984>
- Viscarra Rossel, R. A., Chen, C., Grundy, M. J., Searle, R., Clifford, D., & Campbell, P. H. (2015). The Australian three-dimensional soil grid: Australia's contribution to the *GlobalSoilMap* project. *Soil Research*, 53(8), 845–864. <https://doi.org/10.1071/SR14366>
- Wenxuan, G. (2018). Spatial and temporal trends of irrigated cotton yield in the southern high plains. *Agronomy*, 8(12), 298. <https://doi.org/10.3390/agronomy8120298>
- Wu, L., Zhu, X., Lawes, R., Dunkerley, D., & Zhang, H. (2019). Comparison of machine learning algorithms for classification of LiDAR points for characterization of canola canopy structure. *International Journal of Remote Sensing*, 40(15), 5973–5991. <https://doi.org/10.1080/01431161.2019.1584929>
- Yang, C., Everitt, J. H., & Bradford, J. M. (2007). Airborne hyperspectral imagery and linear spectral unmixing for mapping variation in crop yield. *Precision Agriculture*, 8(6), 279–296. <https://doi.org/10.1007/s11119-007-9045-x>
- Zarco-Tejada, P. J., Ustin, S. L., & Whiting, M. L. (2005). Temporal and spatial relationships between within-field yield variability in cotton and high-spatial hyperspectral remote sensing imagery. *Agronomy Journal*, 97(3), 641–653. <https://doi.org/10.2134/agronj2003.0257>
- Zhao, D., Reddy, K. R., Kakani, V. G., Read, J. J., & Koti, S. (2007). Canopy reflectance in cotton for growth assessment and lint yield prediction. *European Journal of Agronomy*, 26(3), 335–344. <https://doi.org/10.1016/j.eja.2006.12.001>

How to cite this article: Leo S, Migliorati MDA, Grace PR. Predicting within-field cotton yields using publicly available datasets and machine learning. *Agronomy Journal*. 2021;113:1–14. <https://doi.org/10.1002/agj2.20543>