Identifying causes of crop yield variability with interpretive machine learning

Edward J. Jones [a,*], Thomas F.A. Bishop [a], Brendan P. Malone [b], Patrick J. Hulme [c], Brett M. Whelan [a], Patrick Filippi [a]

[a] *Sydney Institute of Agriculture, School of Life and Environmental Science, Faculty of Science, The University of Sydney, Sydney, NSW, Australia*
[b] *CSIRO Agriculture and Food, Black Mountain, ACT, Australia*
[c] *Sustainable Soils Management, Warren, NSW, Australia*

A B S T R A C T

Machine learning approaches have been widely used for crop yield modelling and yield forecasting but there has been limited application to understanding site-specific yield constraints. Crop yield is driven by a complex interaction of spatial and temporal variables, which makes it challenging to define the exact cause of observed spatial yield variability explicitly. This makes it difficult to design efficient management strategies to address production constraints. There is a need for a more quantitative and systematic approach to identify and understand the causes of variation in crop yield in order to implement appropriate management responses. This study investigated the use of interpretive machine learning (IML) to address this need. The developed methodology was demonstrated on furrow-irrigated cotton fields totalling ~2000 ha in the Condamine-Balonne River catchment, Australia. Digital soil maps of important soil constraints were created at 20 m spatial resolution using 70 soil cores extracted to 1.4 m depth and a combination of on-farm and off-farm spatial data layers. Specifically, the soil constraints represented were exchangeable sodium percentage (ESP – sodicity), pH (alkalinity), and electrical conductivity (EC_e – salinity). Terrain infrastructure variable maps of closed depressions, distance down furrow, and cut and fill (from landforming practices) were also developed. Empirical models of cotton lint yield were created with gradient boosted decision trees (XGBoost) using the digital soil maps and terrain infrastructure data as predictor variables. The models could describe the spatial variation in yield well, with a median Lin's concordance correlation coefficient of 0.67 and root-mean-square error of 0.75b ha^{-1}. SHapley Additive exPlanations (SHAP), an IML approach based on game theory, was then used to identify the contribution of each variable to the modelled yield across the study area. The variable most decreasing yield at each point was identified and mapped across the study area, and the spatial extent represented by each variable quantified. The SHAP values for each predictor variable were also extracted and mapped for a case study field, which demonstrated the magnitude of the impact of each variable on yield with spatial context in easily interpretable units (b ha^{-1}). The presented methodology is promising for cost-benefit analysis of implementing remediation strategies, or where not economically feasible, altering management inputs according to a constrained yield potential.

1. Introduction

Global food, fuel and fibre demand is projected to increase to 11.6 billion tonnes by the year 2050 (Pardey et al., 2014). Current trends in yield produced by innovation in crop breeding and management practices are insufficient to meet this projected demand (Ray et al., 2013). This increased demand must be met by optimising production on existing cropland, as further expansion of arable land by clearing natural

systems comes with externalities of increased carbon emissions and loss of biodiversity and ecosystem services (Tilman et al., 2011). Global yield gap analysis has identified that production could be increased by 30% through improved management of soil constraints and fertilizer application (Pradhan et al., 2015). Remedial agronomic practices are often available to mitigate the negative effects of soil constraints on crop yield potential (Page et al., 2018), but such practices should only be implemented when a positive economic outcome is expected. Crop yields are

---

* Corresponding author.
  E-mail address: edward.jones@sydney.edu.au (E.J. Jones).

affected by diverse spatial and temporal variables, and their complex interactions (Jaynes & Colvin, 1997). While the impact of temporal weather variables, such as precipitation, on crop yield are often clear (French & Schultz, 1984), the impact of spatial factors such as soil and terrain variability are often harder to identify and quantify (Kravchenko & Bullock, 2000). To optimally manage a cropping system both the magnitude of the yield constraint and the lost yield potential must be quantified. Work on estimating yield potential for benchmarking has mainly focused on water-limitations and been assessed at large scales using statistical or crop growth simulation models (e.g. Hochman & Horan, 2018). Finer scale estimates using yield monitor data and/or remote sensing methodologies to quantify local yield potential, accompanied with assessment of local drivers in actual yield variability is required. Developing this knowledge has the potential to inform agronomic decision-making processes and facilitate a determination as to which site-specific remedial activities may be warranted.

There is a growing diversity and availability of on-farm and off-farm spatial and temporal data that can be used to model crop yield. This has led to an increased focus on using empirical models to model crop yield utilising data such as satellite imagery, weather observations, management information, and soil maps as input variables (Chlingaryan et al., 2018; van Klompenburg et al., 2020). Due to the increasingly large volume of data available, and the complex and nonlinear nature of the interactions between variables, there has been a growing trend to use machine learning approaches in crop yield modelling (Liakos et al., 2018). Such yield models have been built for various purposes, including forecasting future crop yields (Filippi et al., 2020), and hindcasting or extrapolating yield historically (Donohue et al., 2018). These models provide management insight into optimising the allocation of agronomic resources, and forecast models may be used for a variety of socio-economic purposes (Müller et al., 2014). However, there is also an opportunity to create empirical yield models for the explicit purpose of understanding the drivers of yield variation.

A common criticism of machine learning models is their 'black box' nature (Rudin, 2019). Interpretive machine learning (IML) describes the collection of techniques developed to identify the importance of individual predictors in the model to discern how a prediction was derived. While many machine learning models have in-built methods for assessing the relative importance of predictor variables, such as feature importance plots, these are quite general and still have many limitations (Strobl et al., 2008; Altmann et al., 2010; Filippi et al., 2020). These methods typically assess variable importance at a global scale, which is a limitation as a variable may not impact yield across most of the dataset, but it is an important factor for a subset of the data. Additionally, no insight is given as to what is driving the prediction for a given observation point. SHapley Additive exPlanations (SHAP) are an opportunity to overcome these limitations and provide interpretation that is more closely aligned to human intuition. SHAP values are derived from Shapley values, which were first utilised in game theory to describe the marginal contribution of a player across all combinations of games (Shapley, 1953). When applied to machine learning, the "game" is reproducing the outcome of the model and the "players" are the predictor variables included in the model. As the number of predictor variables increase the number of combinations of predictor variables increases exponentially and the marginal contribution of predictors becomes computationally intensive to compute. SHAP values overcome this limitation with more efficient computation through the use of sample approximations while conserving the properties of Shapley values (Lundberg & Lee, 2017).

SHAP values have been successfully utilised to understand machine learning models in a diverse variety of fields including medicine (Lundberg et al., 2018), finance (Mokhtari et al., 2019), and digital soil mapping (DSM) (Padarian et al., 2020). SHAP values have also been used to interpret crop forecast models (e.g. Srivastava et al., 2021; Shendryk et al., 2021; Zhu et al., 2021), however these analyses are mostly focussed on meteorological phenomena to explain inter-year crop yield variability, they did not incorporate high-quality, site-specific soil information and there is limited evidence of the application of SHAP values to the interpretation of soil constraints. When applied to yield models, SHAP values have the ability to identify which variables contribute to increasing, or decreasing yield at discrete locations, and quantify their relative contribution. This has the potential to identify the underlying cause of low yielding areas, and can inform decisions on what management interventions are required to overcome these yield constraints.

Another limitation to translating empirical yield models into agronomic decisions is the use of data that are not directly interpretable or actionable as input variables. For example, many studies have modelled crop yield using apparent electrical conductivity (EC_a) data from an electromagnetic induction (EMI) survey and crop or bare soil satellite indices (e.g. Robinson et al., 2009; Stadler et al., 2015; Guo, 2018; Filippi et al., 2019a). While such analyses are useful at identifying that a coherent spatial driver of yield variability exists, they provide limited understanding of the underlying cause of yield variability. This is because the input data can be affected by multiple properties simultaneously. For example, EC_a readings are affected by multiple soil properties, including moisture and clay content, organic matter, and salinity (Friedman 2005). As such, it is difficult to make a management decision based on observed relationships between EC_a readings and crop yield without first performing additional investigation into the underlying cause of the observed variation in the EC_a readings. To make the outputs of yield models directly actionable for remediation purposes, the inputs of the yield model must adequately represent the physical and chemical agronomic constraints of the study site and include properties that may be modified through management. This may be achieved by first transforming EC_a, remote sensing data and other observations into maps of soil properties using DSM techniques (McBratney et al., 2003; Triantafilis & Lesch, 2005), and then using these maps representing soil properties as input variables to the yield model.

Digital soil mapping is the process of building statistical relationships between point-based soil observations and covariate data to infer soil phenomena in a spatially consistent and explicit manner. The resulting product is usually a gridded or raster-based map for the defined spatial extent. This procedure may be used to create a three-dimensional (i.e. lateral and vertical) representations of the soil environment. Digital soil mapping provides a cost-efficient solution to providing soil data across an entire study area using a limited number of soil observation points. Digital soil mapping has been implemented at local (Filippi et al., 2019b), national (Grundy et al., 2015) and global scale (Poggio et al., 2021), and provides an objective and statistically rigorous framework that may be updated as additional observations are made available. Digital soil mapping products have great potential to provide soil information for agricultural purposes (Searle et al., 2021). Orton et al. (2018) used DSM products to quantify the yield loss incurred by soil constraints for the entire cropping region of Australia, identifying an annual loss of wheat production equivalent to $1.9 billion (AUD). This analysis is commendable for quantifying the impact of soil constraint on crop production at a high level but the low accuracy of the underlying maps means the analysis is not appropriate for within field management (Kidd et al., 2020). For precision agriculture purposes the only current solution to provide soil maps of utilisable accuracy is the incorporation of on-farm observations during downscaling of existing coarse-scale maps (Malone et al., 2017), or creating new independent farm-scale maps. Digital soil mapping can produce a large number of data layers when multiple soil properties and multiple depths are investigated, which can lead to difficulty in interpretation. A method to reduce the dimensionality of DSM data, while retaining agronomically relevant information, is depth to constraint analysis (Filippi et al., 2019b). Depth to constraint analysis can transform multiple soil property maps into one map by identifying the depth at which an agronomically relevant threshold value has been reached. Such constraints represent physically or chemically hostile environments for crop root growth, such as

extreme pH, salinity or sodicity. When constraints are present the exploitable depth of plant roots in the soil is reduced. Depth to constraint maps are usually based on a threshold at which a yield cost start to be incurred or yield potential decreases by a given percentage of its unconstrained value. Yield models developed using depth to constraint products combined with IML has potential to directly translate into agronomic decisions as yield constraints are identified, the yield loss quantified, and targeted amelioration strategies can be readily implemented as the products are agronomically interpretable.

The utility of using IML techniques to describe crop yield, map the primary yield constraints and quantify their impact appears promising but it has not yet been attempted. The current study investigated a workflow incorporating DSM techniques, machine learning based yield modelling, and IML to describe irrigated cotton yield on a commercial property, and quantify and map the yield limiting factors.

2. Methods

The methodology used in this investigation followed a process that included: collation and processing of proximal and remotely sensed data; identification of soil sampling sites; soil sampling and laboratory analysis; production of digital soil maps and depth to constraint analysis; collation, cleaning and processing of cotton yield data; cotton yield modelling using machine learning; and the interpretation of cotton lint yield and identification of limiting factors using IML. All data processing and analysis was performed in the R platform for statistical computing (R Core Team, 2020).

2.1. Study area

The study area comprised eight irrigated cotton fields (~2000 ha) from a farm within the Condamine-Balonne River Catchment, Queensland, Australia (28°42′21″ S, 147°52′28″ E). Climate at the site is classified as hot semi-arid (BSh) under the Köppen–Geiger system (Peel et al., 2007), and is characterised by hot summers and warm to cool winters. The long-term average annual precipitation observed at the nearby Yamburgan weather station was 448 mm between the years 1898 to 2020, and is summer-dominant (BOM, 2021). Native vegetation at the site was comprised of *Eucalyptus coolabah* open woodland with grass and forb understorey. This vegetation was cleared when the site was landformed for furrow irrigation in 2007. The site experiences an ustic to aridic soil moisture regime and a thermic soil temperature regime (Soil Survey Staff, 2014). The soils are primarily cracking clays, i.e. Vertisols according to Soil Taxonomy (Soil Survey Staff, 2014), which are well suited to irrigated cotton production.

2.2. Environmental covariates

A suite of environmental covariates were used to identify soil sampling locations and map soil properties across the study area. A proximal sensing survey was conducted to collect on-farm data across the study area on a 48 m swathe width. As part of this survey, apparent electrical conductivity (EC$_a$) was obtained using a DUALEM-21S (Dualem Inc., Milton, Canada), gamma radiometric data was obtained using an RSX-1 gamma-ray spectrometer (Radiation Solutions Inc., Mississauga, Canada), and elevation data was recorded using a NovAtel Smart6L differential GPS (DGPS) (NovAtel Inc., Alberta, Canada) with TerraStar-C precise point positioning. Observations were smoothed and interpolated using local block-kriging on a 20 m standard grid to which all future observations were referenced. The local block kriging was implemented using the Vesper Software (Minasny et al., 2005). The block size was set equivalent to the grid size, and variograms were fitted (automatically) for each prediction block from proximally sensed data within a 250 m radius of the block location. A cut and fill map was constructed based on elevation maps produced prior to landforming. Cut and fill values ranged from −0.62 to 0.94 m, representing the depth of soil that was removed (negative values) or deposited (positive values) as the site was landformed to produce a uniform gradient for furrow irrigation purposes. A bare soil redness index was constructed from Landsat 5 Tier 1 surface reflectance satellite imagery using Google Earth Engine (Gorelick et al., 2017). This index represented the 95th percentile of the red band after cloud masking and stacking images from the period January 2008 to April 2012. The 95th percentile is used to represent the soil surface under fallow conditions and contains information related to topsoil variability.

2.3. Digital soil mapping (DSM)

Conditioned Latin hypercube sampling (Minasny & McBratney, 2006) was utilised to identify 70 physical soil sampling locations using the 3.2 m EC$_a$ reading, total count from gamma radiometrics, elevation, cut and fill, and bare soil redness index as input variables. At each sampling location, soil cores to a depth of 1.4 m were extracted in triplicate with 5 m spacing between each core (Fig. 1). Composite samples of the three triplicates were taken at depths of 0–30, 30–60, 60–100 and 100–140 cm. Particle size analysis was performed on each sample using the hydrometer method (Gee & Bauder, 1986). Electrical conductivity was measured in 1:5 soil:water solution (Rayment & Lyons, 2010, p. 20) and used in conjunction with the particle size analysis data to estimate the electrical conductivity of saturated extract (EC$_e$) (Slavich & Petterson, 1993). pH was measured in 1:5 soil:CaCl$_2$ solution (Rayment & Lyons, 2012, p. 40). Exchangeable cations were quantified following extraction with 1 M ammonium acetate at pH 7.0 (Rayment & Lyons, 2010, p. 520). Laboratory observations were co-located with environmental variables and used to map the soil properties across the study area utilising a regression kriging approach (Odeh et al., 1995). The regression component was provided by fitting a Cubist model to the observed laboratory data and environmental covariates (Quinlan, 1992). Residuals at observation points were calculated, followed by automatic variogram fitting (Hiemstra et al., 2009) to parameterise their spatial properties. The fitted variogram facilitated ordinary kriging to interpolate the residuals across the study area, which was added to the regression output to produce the final prediction. The regression kriging procedure was repeated 100 times using a subset of 80% (repeated cross-validation) of the available data in each iteration to quantify the prediction uncertainty. This provided a better representation of prediction quality as compared to doing the split one time only. Mean predictions
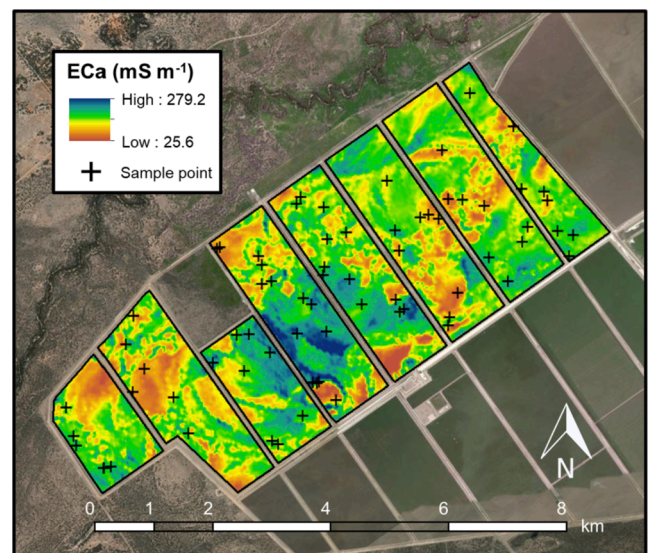

Fig. 1. Satellite image of the study area and its surrounds overlayed with a map of apparent electrical conductivity (EC$_a$ 3.0 m) and soil sampling points (n = 70).

for each property at each depth were used for further analysis.

2.4. Depth to constraint analysis

Three soil constraints representing alkaline, saline and sodic soil conditions were utilised for depth to constraint analysis. Critical values were selected heuristically based on their negative effects on cotton yield. The critical values utilised were pH(1:5 CaCl$_2$) of 8, EC$_e$ of 10 dS m^{-1}, and ESP of 10%, which individually represent moderate to severe constraints to cotton production (McKenzie, 1998; Chinnusamy & Zhu, 2005).

For each property, the modelled value of a map layer was assigned to the mid-depth represented by that layer (i.e. 15, 45, 80 or 120 cm) and values for intermediate depths were obtained by interpolating between the mid-depths at 1 cm increments. The depth at which a critical value was first reached was identified from the interpolated data and recorded (Fig. 2). If the critical value was reached in the topsoil layer, a depth to constraint of 15 cm was recorded which is the mid-depth of the 0–30 cm layer. If the critical value was not reached at any depth a value of 141 cm was recorded. The subscript "crit" has been used to denote the depth at which the critical value was first reached, e.g. ESP$_{crit}$ = 50 indicates that the modelled ESP first exceeded 10% at 50 cm depth.

2.5. Terrain infrastructure variables

Two additional terrain infrastructure variables were developed that describe the influence of irrigation flow and hydrology on observed cotton yield (Fig. 2). A distance down furrow metric was calculated as the distance in metres from the head ditch that supplied irrigation water. The distance down furrow can affect the run-on and run-off rate of irrigation water, with resulting implications for infiltration. The presence of closed depressions within each field was also investigated, and the depth to which standing water would pool was calculated. Closed depressions represent local areas of low relief, as opposed to the uniform gradient expected in a field landformed for furrow irrigation. Closed depressions may form due to settling of soil in areas of fill and also pedoturbation of the shrink-swell soils resulting in the natural reformation of gilgai features (Edelman & Brinkman, 1962). In agronomic terms, closed depressions represent areas where water may pond following irrigation events, which may lead to root suffocation and reduced crop yield. The cut and fill map, as described in the *Environmental covariates* section, was also included as a terrain infrastructure variable. While areas of deep cuts are often associated with exposing saline and sodic conditions at the soil surface, these variables are already represented in the depth to constraint analysis. The inclusion of cut and fill maps in the yield analysis acts as a surrogate for loss of organic matter, soil structure decline, reduced soil depth, and other implications incurred from the soil disturbance that may affect soil hydrology or furrow and seedbed preparation.

2.6. Crop yield modelling

Cotton lint yield monitor data for the 2016/17 season for the eight fields was cleaned using Yield Editor software (version 2.0; Sudduth et al., 2012) and interpolated using block-kriging to the existing 20 m grid. As the yield monitor data was derived from six individual pickers,
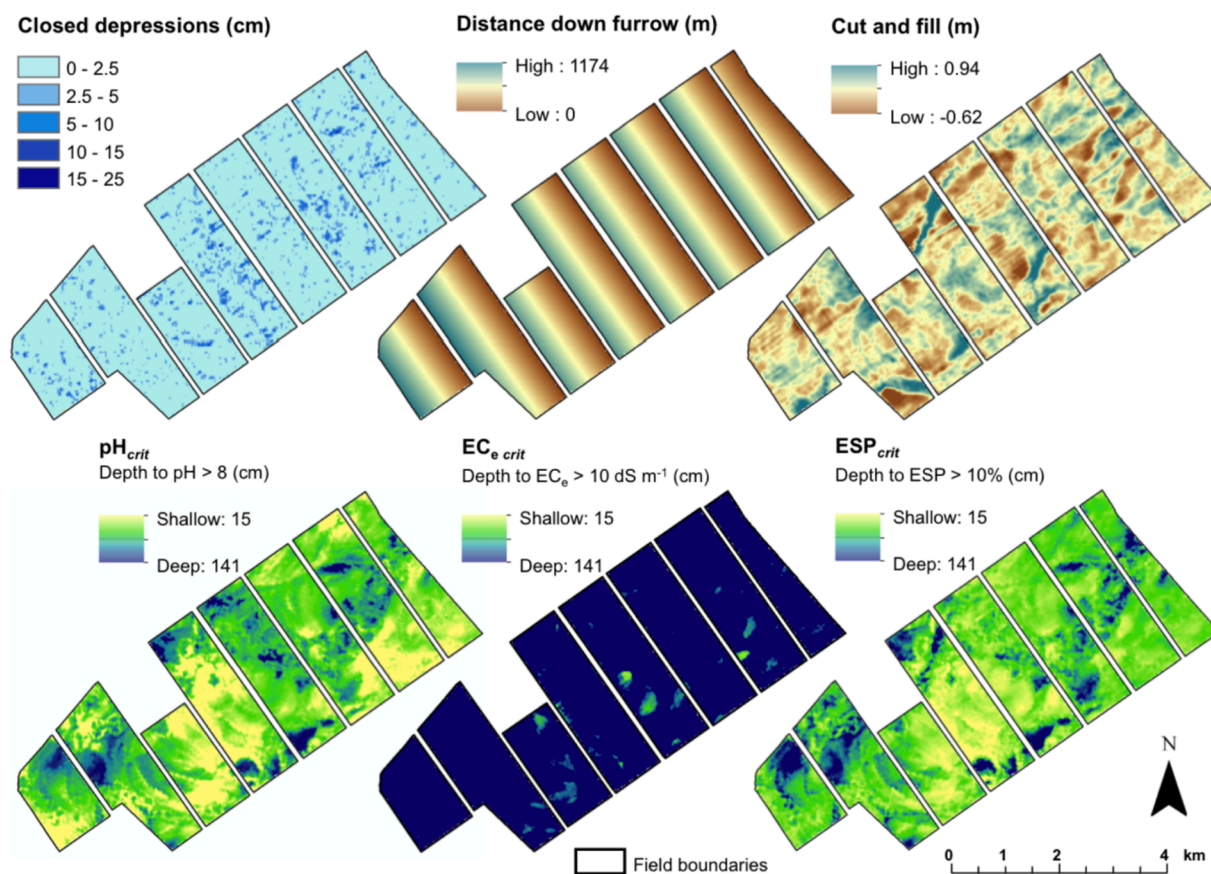

Fig. 2. Predictor variables used in the yield modelling representing three terrain infrastructure variables (closed depressions, distance down furrow, and cut and fill) and three depth to constraint variables (pH$_{crit}$, EC$_{e\ crit}$ and ESP$_{crit}$).

the observations were harmonised so that the average monitor observed yield for each field was within 5% of the field averaged observed ginned yield by applying a constant factor to the data from each of the individual pickers across all fields. This yield data was then joined with the co-located soil and terrain variables to form a single dataset. An XGBoost model was then used to create predictive models of crop yield using the 'xgboost' package (Chen et al., 2020). XGBoost models are an implementation of gradient boosted decision trees. Rather than creating one model across all fields, individual models were created on a per field basis. The reasoning for this was that the aim of the modelling was to understand the primary spatial drivers of historical yield locally and thus information from outside of fields would provide minimal value. In addition, each field had differences in management, notably differential allocation of irrigation resources and sowing date, which impacted observed yield and reduced correlations between yield and the covariates used for modelling when assessing the whole study area simultaneously (Table 1).

The ability of the XGBoost models to predict yield was assessed using 10-fold cross-validation for each field. This approach involved randomly splitting the data for a field into ten equally sized groups, building a model with nine groups, applying that model to predict yield for the left-out group, and then repeating this sequence each of the ten groups. Lin's concordance correlation coefficient (LCCC) (Lin, 1989), the coefficient of determination (R^2) and the root-mean-square error (RMSE) were calculated using the observed and predicted data from the 10-fold cross-validation process. All three metrics are useful to assess the prediction quality of the models: the LCCC is a measure of the agreement between observed and predicted values in relation to the 1:1 line; R^2 is a measure of the strength of the linear association between observed and predicted values, but does not indicate if predictions fall on the 1:1 line; and the RMSE is a measure of the differences between the observed and predicted values. The use of R^2 for model validation has been debated, nonetheless, it has been included to facilitate comparison to previous studies, which have frequently utilised the metric.

2.7. Interpretive machine learning (IML)

After the yield models were created for each field, IML techniques were then used to identify the driving factors of yield variability for each observation point. More specifically, SHapley Additive exPlanations (SHAP) values were calculated using the 'SHAPforxgboost' package (Liu & Just, 2020) on a per field basis. SHAP values provide local model explanation and the contribution of each predictor variable for each data point. It is an advanced method of interpreting results from machine learning models and presents variable importance based on the marginal contribution to the model outcome. SHAP values are calculated for each observation in the training dataset, the sum of SHAP values provides the deviation of the modelled output at each observation point from the field average. For the model parameters, the maximum number of boosting iterations was set to be 500, and the learning rate ('eta') was set at 0.01. A low learning rate, such as the one selected, prevents overfitting and creates a more robust model (Liu & Just, 2020). The variable that contributed most to decreasing yield (lowest SHAP value) at each location was identified and then mapped across the study area to visualise the distribution of each yield constraint. In addition, the SHAP

value of each individual predictor variable was mapped for a case study field to visualise the contribution of each variable to the observed yield.

3. Results and discussion

3.1. Yield model quality

The XGBoost models described the observed variation in cotton yield well with 10-fold cross-validation producing an RMSE of 0.78b ha^{-1}, 0.90 LCCC and 0.83 R^2 when all fields in the study area were compiled (Fig. 3a). When considering individual fields, the distribution of RMSE values ranged from 0.65 to 0.90b ha^{-1}, LCCC values ranged from 0.50 to 0.74, and R^2 values ranged from 0.44 to 0.63 (Fig. 4). An RMSE of 0.83b ha^{-1}, 0.68 LCCC and 0.59 R^2 were observed for the case study field (Fig. 3b). These validation statistics demonstrated that a machine learning approach could effectively describe the observed spatial cotton yield variability using the selected suite of terrain infrastructure and depth to constraint input variables. The validation statistics were comparable to Corwin et al. (2003) who found that ~60% of observed cotton yield variability could be accounted for when using soil properties from 59 sites selected based on an EC$_a$ map in a 32.4 ha field in central California. This comparison is notable as Corwin et al. (2003) used point-based soil observations as input variables for their model, while DSM products were used for this study. The benefits of DSM are evident as the utilisation of DSM in this study produced a yield model with a similar level of accuracy from a sampling density of ~4 samples 100 ha^{-1} compared to ~182 samples 100 ha^{-1} utilised by Corwin et al. (2003). Although this is inherent on the provision of high-quality soil data from on-farm sampling, similar levels of accuracy are not expected when using national-scale DSM products.

3.2. SHAP summary plots

The SHAP summary plots efficiently conveyed the feature importance, feature value and feature effects (Fig. 5). Features are ranked in descending importance based on mean absolute SHAP value. When considering the entire study area, the distance down furrow had the highest mean feature importance, followed by pH$_{crit}$, cut and fill, ESP$_{crit}$, closed depressions, and then EC$_{e\ crit}$ (Fig. 5a). Each point on the SHAP plot represents a data point in the model. In this case, a data point is a 20 m point on the standard grid. The SHAP values on these plots are in bales per hectare, making it easy to interpret and understand the magnitude of the impact of each variable on yield. A benefit of SHAP plots is the richness of information conveyed, and that representation of the strong drivers at discrete locations is conserved. For example, closed depressions were not ranked highly in terms of feature importance as it did not affect the majority of the field, but deep closed depressions produced the greatest modelled decrease in yield for a set of affected points. Global feature importance plots do not have the capability to represent such phenomena. This could result in overlooking the effect of some variables, as their mean feature importance is low, however, they are actually strong drivers of yield for a subset of the dataset. The SHAP plot for the case study field (Fig. 5b) demonstrated that the predictor variables and their values have a similar impact on driving yield variability as the whole study area.

Table 1

Spearman's rank order correlation (ρ) between covariates used in crop yield modelling and cotton lint yield for the entire study area.

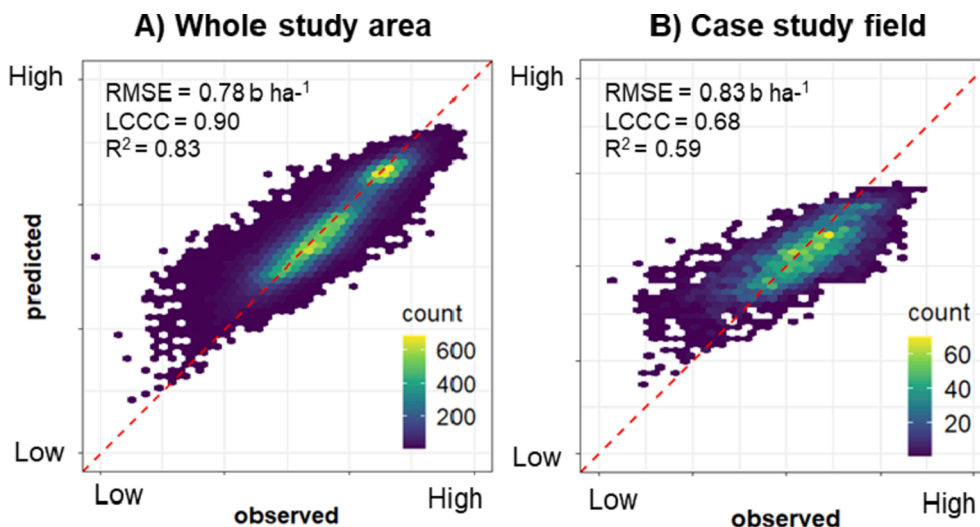|	Yield	Closed depressions	Distance down furrow	Cut and fill	pH$_{crit}$	EC$_{e\ crit}$
Closed depressions	−0.09					
Distance down furrow	−0.04	−0.04				
Cut and fill	−0.03	0.05	−0.23			
pH$_{crit}$	0.02	0.09	0.01	0.05		
EC$_{e\ crit}$	0.07	0.04	−0.04	0.25	0.23	
ESP$_{crit}$	0.17	0.06	−0.04	0.18	0.75	0.28

Fig. 3. Density plot of observed and predicted values for all fields, within: a) the whole study area; and b) the case study field. Predicted values for the whole study area represent a composite of predictions derived from individual models fitted separately to each field. Yield values have been anonymised at the request of the farm manager.
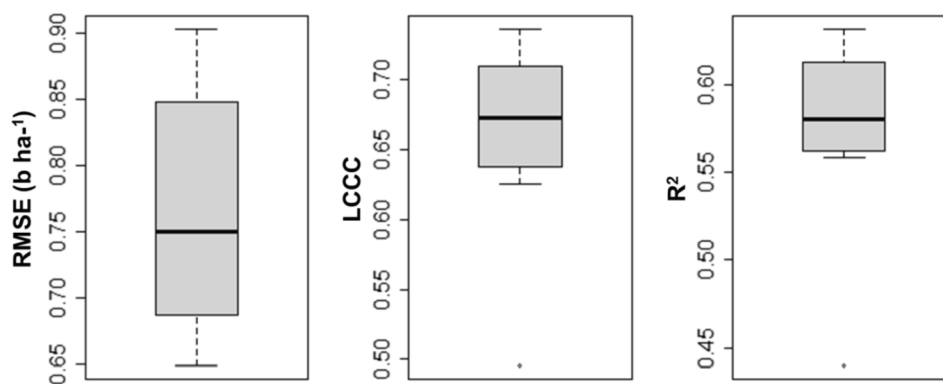

Fig 4. Boxplot showing distribution of 10-fold cross-validated RMSE, LCCC and R^2 of each field within the study area.

3.3. Mapping the feature most limiting yield

To visualise the spatial distribution of yield constraints, the feature with the greatest negative effect for each data point was identified and mapped across the entire study area (Fig. 6). Visual assessment revealed distinct spatial patterns for each feature and demonstrated that no particular variable was dominant across the whole study area. The relative area where each feature had the dominant negative yield effect was quantified for the entire study area and the case study field (Table 2). Distance down furrow was the most frequent feature, representing 26% of the entire study area. Distance down furrow presented as wide bands at the mid-length of most fields. Cut and fill was the second most frequent feature, representing 24% of the study area. Areas where cut and fill was the most limiting feature presented as large, spatially coherent regions and it occupied more than 50% of the area for one field. Closed depressions represented 12% of the study area, with morphology characterised by clusters of small and distinct patches. ESP_{crit} and pH_{crit} represented 15% and 17% of the study area respectively, with morphology characterised by medium to large sized patches. $EC_{e\ crit}$ was the least frequent at 6%, with a low number of small patches present in some fields. Overall, the map demonstrated that drivers of yield were highly spatially variable within fields and that there is a changing importance of these influences on yield.

3.4. Mapping feature effect on yield

It is important to assess into the magnitude of the modelled yield decrease imparted by each feature and whether this effect is logical. This can help inform decisions on managing the factors controlling yield and identify areas where remedial activities may be warranted. This analysis has been performed on the case study field to showcase what would be done realistically in the industry to understand causes of yield variability. Fig. 7 shows the maps of the predictor variables used in the modelling process for the case study field, and the SHAP value maps showing how each variable impacts on the yield prediction for that field in bales per hectare (b ha^{-1}) relative to the field average. The SHAP value maps are limited to −1.5 (red areas) and +1.5 (blue areas) to improve the interpretation of the magnitude to which each variable is limiting yield. These maps allow the causes of spatial yield variation to be better understood, and help to identify the magnitude that each variable negatively influences yield at different locations within fields in units that are directly interpretable.

Small pockets of the field were highly negatively impacted by closed depressions (Fig. 7a). The corresponding map of closed depression depth corroborates that yield affected areas occur where deeper closed depressions were observed, whereas the majority of the field is largely unaffected by closed depressions. The map of SHAP value for distance to head ditch displayed a number of distinct banded features. Modelled yields were positively, or negligibly, affected in areas close to the head
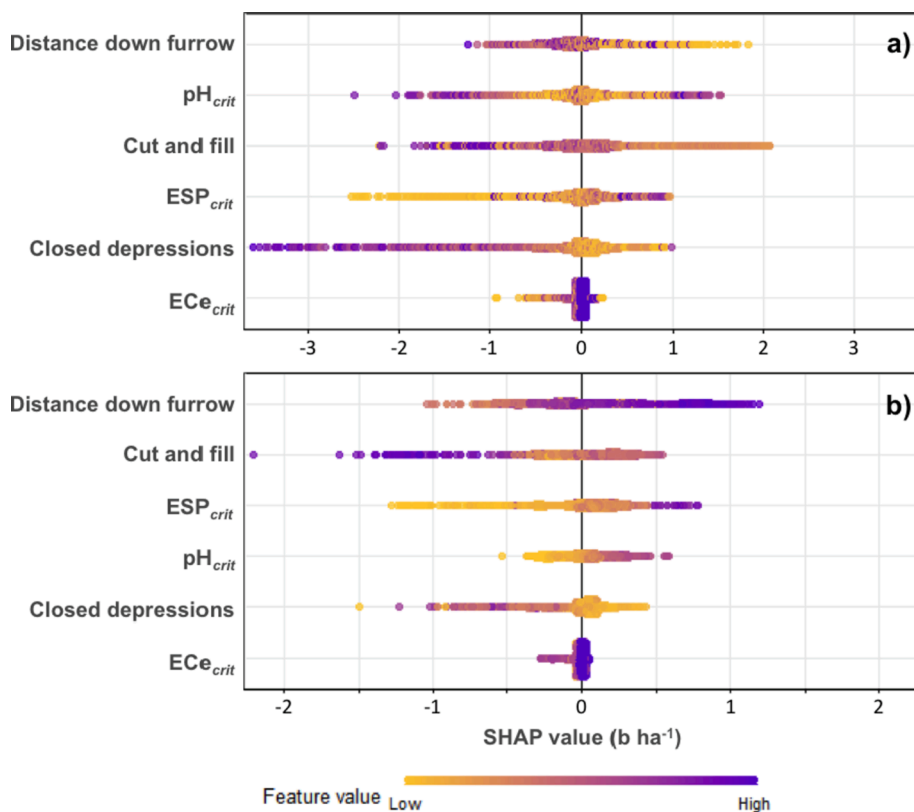
Fig. 5. SHAP summary plots from the XGBoost yield model for: a) the whole study area; and b) the case study field. The position on the x-axis is determined by the SHAP value, which represents the feature effect on yield at the discrete location. The colour indicates the feature value from low to high. Position on the y-axis is ordered by decreasing mean absolute SHAP value for each feature.
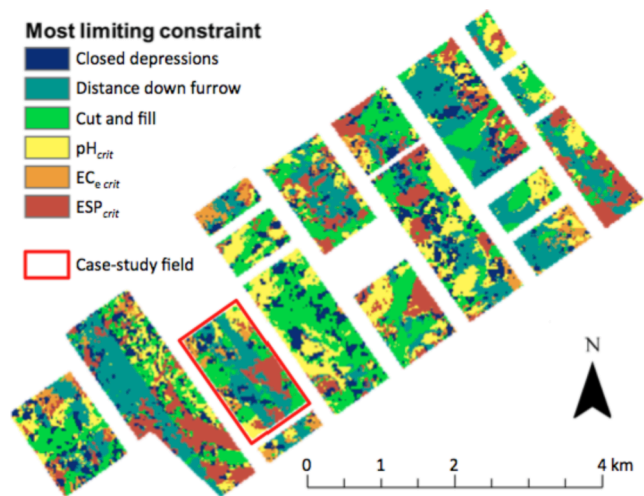

Fig. 6. The feature with the greatest negative yield effect for the whole study area. The boundary of the case study field is highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Variables most negatively contributing to yield and the proportion area that they cover for the whole study area and the case study field.

Variable	Study area	Case study field
Closed depressions	12%	11%
Distance down furrow	26%	35%
Cut and fill	24%	19%
pH_{crit}	17%	11%
$EC_{e\ crit}$	5%	10%
ESP_{crit}	15%	14%

distance to head ditch variable. The impact of the cut and fill terrain infrastructure variable on yield was complex to interpret. Areas of deep cuts were clearly associated with a negative yield effect, as well as areas of deep fill, most notably in the eastern corner of the field. Meanwhile, areas that were minimally disturbed during landforming and areas of shallow fill produced positive yield effects.

The SHAP value map of ESP_{crit} revealed that it was a strong driver of yield in the case study field. Two large areas in the south and northeast edge of the field were negatively impacted by ESP_{crit}, which corresponded to an ESP constraint being reached either in the topsoil layer or at a shallow depth in the soil profile. In contrast, a large area in the northwest of the field was positively impacted, which corresponded to ESP_{crit} being reached deeper in the soil profile. The pH_{crit} variable did not have a large impact on yield, however, when a pH_{crit} was reached deeper in the profile there was a small positive impact on yield, and when pH_{crit} was observed in the topsoil, there was a small negative impact on yield. This impact of depth to pH and ESP constraints on crop yield has also been found in other studies (Filippi et al., 2019b; Filippi et al., 2020). The subdued impact of the pH_{crit} variable on yield could be due its correlation with ESP_{crit} (Table 1), as the ESP map could be masking the

ditch and tail drain. These patterns are logical as these areas both have a high run-on and run-off rate that facilitates the provision of adequate soil moisture to the plant, followed by adequate drainage to prevent waterlogging. Two bands in the middle of the field displayed negative yield effects; this may be due to imperfect edge matching during landforming, although further investigation is required to definitively identify the causal factor. It is clear that there are interactions with other predictor variables, and this impacts the relative contribution for the
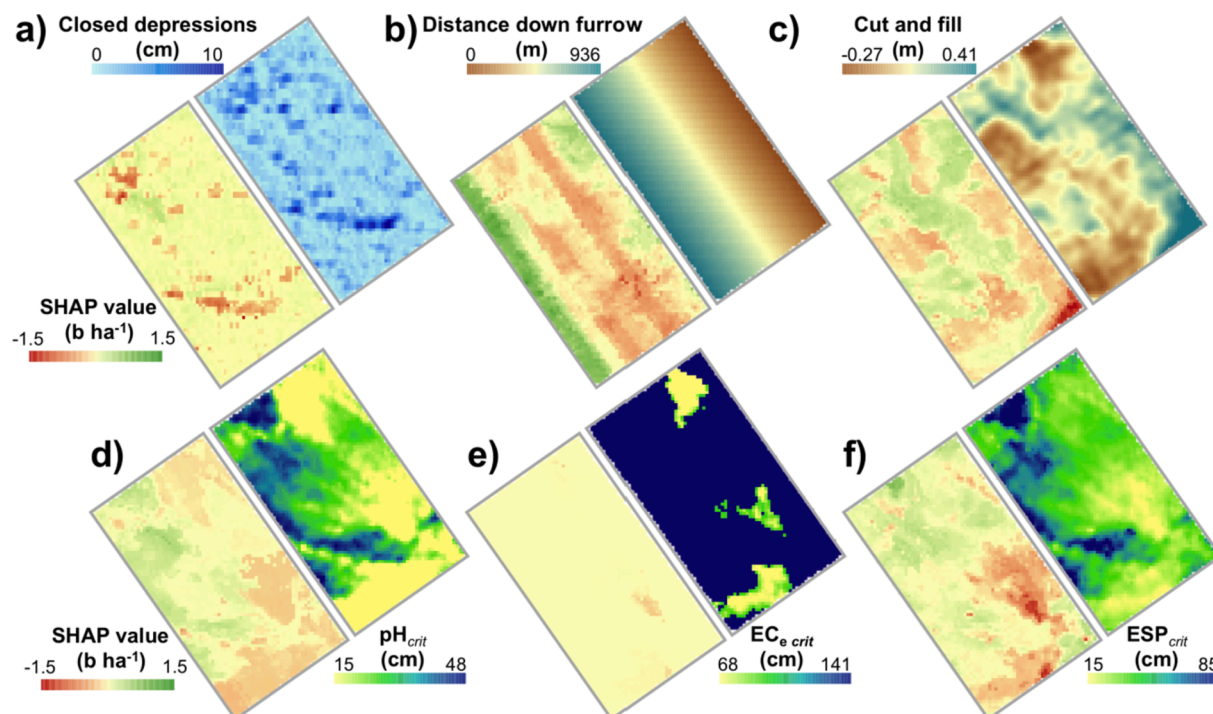
Fig. 7. Maps of predictor variables used for yield modelling of the case study field and corresponding maps of SHAP values, which represent the modelled impact of that variable on yield in b ha^{-1}.

importance of the pH map. Alternatively, the reduced range of pH_{crit} (15–48 cm) may limit the identification of yield effects. The analysis found minimal impact of $EC_{e\ crit}$ on yield. This may be due to critical EC_e values only being reached for a small subset of the field, and when this did occur it was deeper in the subsoil, i.e. the shallowest $EC_{e\ crit}$ was observed at 68 cm compared to ESP_{crit} which was observed in the topsoil layer.

3.5. General discussion

The impact of spatial factors such as soil and terrain have a large impact on yield variability, but the interaction between all of these variables makes it difficult to pinpoint the primary constraint, as well as the magnitude of the impact on yield. This study demonstrated that creating empirical models of crop yield with important spatial drivers, and combining this with IML is a promising avenue to fill this gap in the agricultural sector.

The IML approach used in this investigation extracted information from the XGBoost machine learning yield model that provided insight into the variables most limiting yield, which may prove useful for on-farm decision-making processes. Using the maps constructed, growers and agronomists can quantify the yield cost of constraints, and calculate the cost-benefit of an amelioration strategy. Additionally, they can identify areas of high chance of success to trial amelioration strategies before implementing a full strategy. In the event that a remediation procedure is not implemented, understanding of the limitation can still facilitate more efficient use of agricultural inputs, for example, reducing fertiliser applications in constrained areas to match the lower yield potential.

Salinity was not observed to be a significant yield limitation across the study area, however, sodicity and alkalinity relatively shallow in the soil profile were widespread. Sodicity and alkalinity are a particularly common in the irrigated cropping soils of northern New South Wales and southern Queensland (Orton et al., 2018; Filippi et al., 2020). When considering remediation approaches, soil is expected to gradually acidify under agricultural production (Moody & Aitken, 1997), and this

yield limitation will gradually decrease the alkalinity constraint over time without further intervention (Filippi et al., 2018). In contrast, sodicity will not improve and is likely to gradually worsen without intervention. High sodicity can be addressed through the application of gypsum (Roberton et al., 2020). The areas identified in this analysis where a shallow ESP_{crit} was associated with a negative SHAP value represent an ideal location where application of gypsum could be trialled and monitored for improvements in crop production to better understand the cost-benefit of the amelioration strategy.

The terrain infrastructure variables were useful to include in the study as they accounted for a large degree of the observed yield variability. The presence of closed depressions was revealed to have a large negative effect on yield, however, this effect was shown to only occur in small discrete patches. High-resolution elevation data is routinely captured during on-farm operations. This data could be used to monitor the evolution of closed depressions and their effect on yield through time to determine when the cost-benefit threshold of landforming to remove the closed depressions makes economic sense. Distance down furrow and cut and fill are difficult to translate into direct management decisions for the existing study area. However, they may provide useful information for future developments. Negative yield effects were observed for deep cut and deep fill areas indicating that these should be avoided when landforming future sites. Additionally, areas of shallow $EC_{e\ crit}$ were correlated with deep cut areas, indicating that the landforming event may have brought these areas closer to the soil surface, further demonstrating that this should be avoided. Distance down furrow was found to positively affect cotton yield close to the head ditch and tail drain, and negatively affect yield in the middle of the field. Distance down furrow impacts soil moisture, and drainage, and it is likely that the importance of this on yield would fluctuate seasonally depending on precipitation events and irrigation management. This relationship could be investigated further to optimise irrigation development or investigate the benefits of site-specific variety and seeding rates to efficiently utilise these different moisture environments (Matcham et al., 2020).

Feature extraction and the transformation of datasets into

agronomically interpretable variables is an important step in pinpointing the driving factors of yield variability in this study. For machine learning yield models and IML to be applied effectively, it is vital that the variables used in the yield model can be directly related to yield, and translated into management decisions. This study achieved this by taking proximally and remotely sensed spatial data, and combining them with point-based soil observations to create digital soil maps. The approach used in this study could be further improved and modified for other agricultural environments. The current study only looked at one season of yield data, but the impact of spatial variables on yield may also shift from season to season. This is due to interactions with temporal variables, most-notably the interaction between landscape position and precipitation (Kravchenko & Bullock, 2000; Kaspar et al., 2004). These interactions would be particularly import in rain-fed cropping systems that are not landformed. The approach implemented in this study could be expanded and up-scaled across years by including further information on management (e.g. crop variety, sowing date, irrigation), abiotic variables (e.g. precipitation, temperature, cloud-cover), and biotic pressures (e.g. weed competition, presence of disease and animal pests).

4. Conclusion

This study modelled cotton lint yield using machine learning, yield monitor datasets, digital soil maps of important constraints (pH, ESP, EC_e), and terrain infrastructure attributes for furrow-irrigated fields. The models could describe yield variability within the fields well, with a median LCCC of 0.67. SHAP values were used to identify the primary spatial drivers of yield and proved useful for quantifying the magnitude of impact on yield imparted by each variable. Visualising these SHAP values as maps showing the negative impact of constraints on yield in bales per hectare (b ha^{-1}), improved the interpretation of the results. Overall, the approach implemented in this study provided a quantitative and robust method for growers and agronomists to implement management strategies to manage and overcome the primary limitations of yield. It also allows the further quantification of the economic cost of yield losses, and the economic benefit of amelioration. This work may be developed further by including multiple seasons of yield observation, and incorporating both spatial and temporal variables during yield modelling.

CRediT authorship contribution statement

Edward J. Jones: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Thomas F.A. Bishop:** Conceptualization, Supervision, Writing – review & editing. **Brendan P. Malone:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Patrick J. Hulme:** Conceptualization, Data curation, Funding acquisition, Investigation, Writing – review & editing. **Brett M. Whelan:** Methodology, Supervision, Writing – review & editing. **Patrick Filippi:** Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Altmann, A., Tološi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinform. 26 (10), 1340–1347.

BOM, 2021. Bureau of Meteorology: Monthly rainfall – Yamburgan (044166), available at: http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=044166. [Accessed: 22 June 2021]. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., 2020. xgboost: Extreme Gradient Boosting. R package version 1.1.1.1. URL: https://CRAN.R-project.org/package=xgboost.

Chinnusamy, V., Zhu, J.K., 2005. Understanding and improving salt tolerance in plants. Crop Sci. 45, 437–448.

Chlingaryan, A., Sukkarieh, S., Whelan, B.M., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69.

Corwin, D.L., Lesch, S.M., Shouse, P.J., Soppe, R., Ayars, J.E., 2003. Identifying soil properties that influence cotton yield using soil sampling directed by apparent soil electrical conductivity. Agronomy 95 (2), 352–364.

Donohue, R.J., Lawes, R.A., Mata, G., Gobbett, D., Ouzman, J., 2018. Towards a national, remote-sensing-based model for predicting field-scale crop yield. Field Crops Res. 227, 79–90.

Edelman, C.H., Brinkman, R., 1962. Physiography of gilgai soils. Soil Sci. 94 (6), 366–370.

Filippi, P., Cattle, S.R., Bishop, T.F., Odeh, I.O., Pringle, M.J., 2018. Digital soil monitoring of top-and sub-soil pH with bivariate linear mixed models. Geoderma 322, 149–162.

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., 2019a. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. 20, 1015–1028.

Filippi, P., Jones, E.J., Ginns, B.J., Whelan, B.M., Roth, G.W., Bishop, T.F.A., 2019b. Mapping the depth-to-soil pH constraint, and the relationship with cotton and grain yield at the within-field scale. Agronomy 9 (5), 251.

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. Agric. Syst. 184, 102894.

French, R.J., Schultz, J.E., 1984. Water use efficiency of wheat in a Mediterranean-type environment. I. The relation between yield, water use and climate. Aust. J. Agric. Res. 35, 743–764.

Friedman, S.P., 2005. Soil properties influencing apparent electrical conductivity: a review. Comput. Electron. Agric. 46 (1–3), 45–70.

Gee, G., Bauder, J., 1986. Particle size analysis. In: Klute, A. (Ed.) Methods of soil analysis. Part 1: Physical and mineralogical methods. Soil Science Society of America and American Society of Agronomy, Madison, WI, USA, pp. 383–411.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Grundy, M.J., Viscarra Rossel, R., Searle, R.D., Wilson, P.L., Chen, C., Gregory, L.J., 2015. Soil and landscape grid of Australia. Soil Res. 53 (8), 835–844.

Guo, W., 2018. Spatial and temporal trends of irrigated cotton yield in the Southern High Plains. Agronomy 8, 298.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J., Heuvelink, G.B., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Comput. Geosci. 35 (8), 1711–1721.

Hochman, Z., Horan, H., 2018. Causes of wheat yield gaps and opportunities to advance the water-limited yield frontier in Australia. Field Crops Res. 228, 20–30.

Jaynes, D.B., Colvin, T.S., 1997. Spatiotemporal variability of corn and soybean yield. Agronomy 89, 30-37.

Kaspar, T.C., Pulido, D.J., Fenton, T.E., Colvin, T.S., Karlen, D.L., Jaynes, D.B., Meek, D.W., 2004. Relationship of corn and soybean yield to soil and terrain properties. Agronomy 96 (3), 700–709.

Kidd, D., Searle, R., Grundy, M., McBratney, A., Robinson, N., O'Brien, L., Zund, P., Arrouays, D., Thomas, M., Padarian, J., Jones, E., Bennett, J.M., Minasny, B., Holmes, K., Malone, B.P., Liddicoat, C., Meier, E.A., Stockmann, U., Wilson, P., Wilford, J., Payne, J., Ringrose-Voase, A., Slater, B., Odgers, N., Gray, J., van Gool, D., Andrews, K., Harms, B., Stower, L., Triantafilis, J., 2020. Operationalising digital soil mapping–Lessons from Australia. Geoderma Reg. e00335.

Kravchenko, A.N., Bullock, D.G., 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agronomy 92 (1), 75–83.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: A review. Sensors 18 (8), 2674.

Liu, Y., Just, A., 2020 SHAPforxgboost: SHAP Plots for 'XGBoost'. R package version 0.0.4. URL: https://CRAN.R-project.org/package=SHAPforxgboost.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, pp. 4768-4777.

Lundberg, S.M., Erion, G.G., Lee, S.I. 2018. Consistent individualized feature attribution for tree ensembles. arXiv 1802.03888.

Malone, B.P., Styc, Q., Minasny, B., McBratney, A.B., 2017. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91-99.

Matcham, E.G., Hamman, W.P., Hawkins, E.M., Fulton, J.P., Subburayalu, S., Lindsey, L.E., 2020. Soil and terrain properties that predict differences in local ideal seeding rate for soybean. Agronomy 112 (3), 1981–1991.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

McKenzie, D.C., (Ed.), 1998. SOILpak for cotton growers, 3rd edn. NSW Agriculture, Orange, New South Wales.

Minasny, B., McBratney, A.B., and Whelan, B.M., 2005. VESPER version 1.62. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. geosci. 32 (9), 1378–1388.

Mokhtari, K.E., Higdon, B.P., Başar, A., 2019. Interpreting financial time series with SHAP values. In: In: Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering, pp. 166–172.

Moody, P.W., Aitken, R.L., 1997. Soil acidification under some tropical agricultural systems. 1. Rates of acidification and contributing factors. Soil Res. 35 (1), 163–174.

Müller, C., Robertson, R.D., 2014. Projecting future crop productivity for global economic modeling. Agric. Econ. 45 (1), 37–50.

Odeh, I.O., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67 (3–4), 215–226.

Orton, T.G., Mallawaarachchi, T., Pringle, M.J., Menzies, N.W., Dalal, R.C., Kopittke, P. M., Searle, R., Hochmann, Z., Dang, Y.P., 2018. Quantifying the economic impact of soil constraints on Australian agriculture: A case-study of wheat. Land Degrad. Dev. 29 (11), 3866–3875.

Padarian, J., McBratney, A.B., Minasny, B., 2020. Game theory interpretation of digital soil mapping convolutional neural networks. Soil 6 (2), 389–397.

Page, K.L., Dalal, R.C., Wehr, J.B., Dang, Y.P., Kopittke, P.M., Kirchhof, G., Fujinuma, R., Menzies, N.W., 2018. Management of the major chemical soil constraints affecting yields in the grain growing region of Queensland and New South Wales, Australia–a review. Soil Res. 56, 765–779.

Pardey, P.G., Beddow, J.M., Hurley, T.M., Beatty, T.K., Eidman, V.R., 2014. A bounds analysis of world food futures: Global agriculture through to 2050. Aust. J. Agric. Res. Econ. 58 (4), 571–589.

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240.

Pradhan, P., Fischer, G., van Velthuizen, H., Reusser, D.E., Kropp, J.P., 2015. Closing yield gaps: how sustainable can we be? PLoS ONE 10 (6), e0129487.

Quinlan, J.R., 1992. Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, 92, pp. 343-348.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Ray, D.K., Mueller, N.D., West, P.C., Foley, J.A., 2013. Yield trends are insufficient to double global crop production by 2050. PLoS ONE 8 (6), e66428.

Rayment, G., Lyons, D.J. (Eds.), 2010. Soil Chemical Methods - Australasia. CSIRO Publishing, Collingwood, Australia.

Roberton, S.D., Bennett, J.M., Lobsey, C.R., Bishop, T.F.A., 2020. Assessing the Sensitivity of Site-Specific Lime and Gypsum Recommendations to Soil Sampling Techniques and Spatial Density of Data Collection in Australian Agriculture: A Pedometric Approach. Agronomy 10 (11), 1676.

Robinson, N.J., Rampant, P.C., Callinan, A.P.L., Rab, M.A., Fisher, P.D., 2009. Advances in precision agriculture in south-eastern Australia. II. Spatio-temporal prediction of crop yield using terrain derivatives and proximally sensed data. Crop Pasture Sci. 60 (9), 859–869.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.

Searle, R., McBratney, A., Grundy, M., Kidd, D., Malone, B., Arrouays, D., Stockman, U., Zund, P., Wilson, P., Wilford, J., Van Gool, D., Triantafilis, J., Thomas, M., Stower, L., Slater, B., Robinson, N., Ringrose-Voase, A., Padarian, J., Payne, J., Orton, T., Odgers, N., O'Brien, L., Minasny, B., Bennett, J.M., Liddicoat, C., Jones, E., Holmes, K., Harms, B., Gray, J., Bui, E., Andrews, K., 2021. Digital soil mapping and assessment for Australia and beyond: A propitious future. Geoderma Regional, e00359.

Shapley, L.S., 1953. A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games, 2. Princeton University Press, pp. 307–317.

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crops Res. 260, 107984.

Slavich, P.G., Petterson, G.H., 1993. Estimating the electrical conductivity of saturated paste extracts from 1: 5 soil, water suspensions and texture. Soil Res. 31 (1), 73–81.

Soil Survey Staff, 2014. Keys to soil taxonomy. United States Department of Agriculture and Natural Resources Conservation Service, 12, p 372.

Srivastava, A.K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., Rahimi, J., 2021. Comparison of Machine Learning Methods for Predicting Winter Wheat Yield in Germany. arXiv preprint arXiv:2105.01282.

Stadler, A., Rudolph, S., Kupisch, M., Langensiepen, M., van der Kruk, J., Ewert, F., 2015. Quantifying the effects of soil variability on crop growth using apparent soil electrical conductivity measurements. Eur. J. Agron. 64, 8–20.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC bioinform. 9 (1), 1–11.

Sudduth, K.A., Drummond, S.T., Myers, D.B., 2012. Yield Editor 2.0: Software for automated removal of yield map errors. Paper presented at the 2012 ASABE Annual International Meeting, Dallas, Texas.

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. PNAS 108 (50), 20260–20264.

Triantafilis, J., Lesch, S.M., 2005. Mapping clay content variation using electromagnetic induction techniques. Comput. Electron. Agric. 46 (1–3), 203–237.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709.

Zhu, P., Abramoff, R., Makowski, D., Ciais, P., 2021. Uncovering the past and future climate drivers of wheat yield shocks in Europe with machine learning. Earth's Future, 9(5), e2020EF001815.