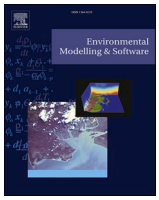
Environmental Modelling and Software 162 (2023) 105634

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



Comparison of machine learning methods emulating process driven crop models

David B. Johnston [a,b,*], Keith G. Pembleton [a,b], Neil I. Huth [d], Ravinesh C. Deo [c]

[a] School of Agriculture and Environmental Sciences, University of Southern Queensland, Toowoomba, Australia
[b] Centre for Sustainable Agricultural Systems, University of Southern Queensland, Toowoomba, Australia
[c] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia
[d] CSIRO, Toowoomba, Australia

ARTICLE INFO

Handling Editor: Daniel P Ames

Keywords:
Metamodels
Surrogates

ABSTRACT

Performing large scale simulation analyses using complex process-driven models can be very time consuming and incur significant computational expense. These analyses involve generating synthetic datasets and include processes such as impacts analysis (IA) and variance-based sensitivity analysis (SA). Machine learning (ML) provides a potential alternative path to reduce computational costs incurred when generating output from large simulation experiments. We assessed the accuracy and computational efficiency of three ML-based emulators (MLEs): artificial neural networks, multivariate adaptive regression splines, and random forest algorithms, to replicate the outputs of the APSIM-NextGen chickpea crop model. The MLEs were trained to predict seven outputs of the process-driven model. All the MLEs performed well ($R^2 > 0.95$) for predicting outputs for the training data set locations but did not perform well for previously unseen test locations. These findings indicate that modellers using process-driven models can benefit from using MLEs for efficient data generation, provided suitable training data is provided.

1. Introduction

The agricultural and environmental science disciplines have long utilised the power of computer modelling for scientific enquiry and knowledge advancement (Jones et al., 2016). Mechanistic models have been developed for many biological and environmental processes, and these models have subsequently been integrated to form whole-of-system simulation computing environments which are complex and computationally expensive to configure, validate and run (Keating et al., 2003; Holzworth et al., 2014). New developments in computer modelling are often driven by the need for cost reduction and improved efficiencies, as these two concepts are integral in the functioning of most modern economies and exist as non-negotiable goals for most projects. As computing costs have progressively reduced over the past few decades, the size and complexity of experiments and analysis based on computer modelling has grown. These simulation experiments can require the running of many thousands, or even millions of model runs, and produce extensive amounts of data (e.g. Phelan et al. (2018) and Casadebaig et al. (2016)). A reduction in the computational costs of producing large amounts of data is one area that is a target of improved efficiency efforts.

Machine learning (ML) approaches for predictive modelling are having a significant impact on many areas of society, including areas of scientific research, not the least of which are agricultural and environmental sciences. Computational efficiency in producing predicted outcomes is one benefit of ML algorithms (Balakrishnan and Muthukumarasamy 2016; Karandish and Šimůnek 2016; Shastry et al., 2016; Singh et al., 2017; Ryan et al., 2018; Feng et al., 2019; Niazian and Niedbała 2020). Much research involving ML technologies revolves around the approaches being able to take diverse data sources, such as remote imaging and multiple sensor inputs, and predict outcomes such as vegetation type, soil water content, biomass and crop health (Shakoor et al., 2017; Prasad et al., 2018; Lawes et al., 2019; Feng et al., 2020; Obsie et al., 2020; Zhang et al., 2020; Fajardo and Whelan 2021; Guo et al., 2021; Paudel et al., 2021), while the potential computational efficiency gains have received much less attention. Systems modelling, be it for weather, environmental or agricultural systems, are undertaken using complex, process driven models. The agricultural production systems simulator (APSIM-NextGen) (Holzworth et al., 2018) is one such modelling system in the agricultural and environmental sciences

* Corresponding author. School of Agriculture and Environmental Sciences, University of Southern Queensland, Toowoomba, Australia.
E-mail address: u1097253@umail.usq.edu.au (D.B. Johnston).

https://doi.org/10.1016/j.envsoft.2023.105634
Received 23 March 2022; Received in revised form 22 November 2022; Accepted 16 January 2023
Available online 26 January 2023

domain. While process driven modelling systems like APSIM-NextGen provide extensive modelling and research opportunities due to their complexity and flexible configuration, they are computationally expensive. This limits experimental designs where resources are insufficient to run large numbers of simulations (e.g. Casadebaig et al. (2016)). Impacts analysis (IA) and sensitivity analysis (SA) are two examples of processes that often requires large numbers of simulations to evaluate the interactions between changes in input factor values and the effects these have on target output values. While the expectations and requirements to validate models using SA continues to grow (Razavi et al., 2021), the ability to undertake thorough SA of complex systems models is compromised by the limitations imposed by computing power. This is just one example of how expanded output from crop models might be used.

Previous studies have utilised ML emulators, or meta-models as they are also referred to, for the computationally efficient expansion of crop modelling outputs for addressing research questions that required analysing very large datasets. For example, Shahhosseini et al. (2019) compared four ML algorithms for the prediction of maize yield and nitrate loss, and generated a simulated dataset of more than three million data points. Results varied between which ML algorithm was best for predicting yield (Extreme Gradient Boosting algorithm) versus nitrate loss (Random Forest), while the ideal size of the training dataset and the sensitivity to different input variables also varied between algorithms. Mandrini et al. (2021) used a large synthetic dataset to compare static nitrogen recommendation tools to ML based dynamic recommendation tools. The dynamic recommendation tools lacked the accuracy in predictions and were therefore found to be of less usefulness in many situations than the static recommendations. There have also been studies which considered the use of emulators to improve the efficiency of performing SA on complex environmental models. For example, Stanfill et al. (2015) and Ryan et al. (2018) both used the statistical approach of generalised additive models to improve computational efficiency of SA applications. Wallach and Thorburn (2017) and Sexton et al. (2017) discuss the relatively new approach, at least in crop modelling research, of utilising machine learning based emulators (MLEs) to improve computational efficiency in uncertainty analysis. These studies highlight the early stage that research into the potential of using ML approaches to improve the computational efficiency of generating expanded synthetic datasets of complex process-driven biophysical models is currently at. More research is required to assess what range of biophysical modelling scenarios and analyses might benefit from expanded crop modelling output using ML techniques. Underlying these questions is the issue of whether any particular ML approach is better able to be trained to predict the outputs of complex systems models.

The objective of this research was to demonstrate that, by using input parameters used to configure and run APSIM-NextGen chickpea crop simulations, MLEs could be developed which are able to predict selected APSIM model outputs. If this is demonstrated, then the use of these MLEs would allow the substitution of the APSIM system model, for the specific and limited purpose of generating synthetic data sets, with a small and efficient predictive model that is effective for the range of input parameter variations used in the training data set. These MLEs could then be used to generate large synthetic data sets. These datasets might be suitable for undertaking a variety of analyses of the underlying modelled relationships, analysis of impacts of varying input settings, and potentially for aiding in developing hybrid modelling approaches which could open new areas of modelling research. A further objective was to test if the MLEs developed were robust enough to be able to accurately predict crop outputs for all locations within the regions covered by the training data set. This required the input parameters used to develop the MLEs to be diverse enough and contain enough variation in values used to cover the expected ranges of values for all locations of interest. These objectives differ from the previous work of Mandrini et al. (2021) by evaluating MLEs ability to predict APSIM outputs rather than comparing the performance of the two modelling approaches against real-world observed data. It also varies from the work of Shahhosseini et al. (2019) by evaluating MLEs predictive ability across different phases of a crops entire lifecycle, and by including metrics to compare the computational costs of developing the MLEs and their statistical accuracy of predicting APSIM outputs. To fulfil these objectives, the APSIM-NextGen chickpea model was configured to simulate crop production over a 120-year period at seven locations throughout the chickpea production regions in Australia. Six model outputs were reported and further used to train emulators based on three ML algorithms: 1) artificial neural network (ANN), 2) multivariate adaptive regression splines (MARS) and 3) a random forest (RF), using 24 input factors from the APSIM simulations. The MLEs were assessed for predictive accuracy, input variable importance and computational effort. The assessments of model performances were conducted for the locations for which the MLEs were trained, as well as two additional locations not included in the training data set to test emulator robustness.

2. Methods

Three MLEs representing different ML algorithmic approaches were developed from data generated from APSIM simulations of chickpea growth, development, and yield for seven locations in the Australian chickpea production regions. The MLEs were trained on a subset of 80% of the randomised generated data and then validated using the remaining 20% of data. A bootstrap process was used to repeat this randomisation and model generation ten times to assess the consistency of the MLEs developed. The workflow of this process is summarised by the flowchart in Fig. 1. Goodness-of-fit of emulator generated data against the original APSIM data for six model outputs were analysed and are presented in the results section. The output targets were as follows: 1) days from sowing to emergence (EmergenceDAS), 2) days from sowing to flowering (FloweringDAS), 3) days from sowing to first fruiting pod (PoddingDAS), 4) days from sowing to crop maturity (MaturityDAS), 5) above ground crop biomass at harvest (kg/ha) (Biomass), and 6) weight of harvested grain (kg/ha) (GrainWt). These results cover some of the more significant chickpea model outputs for monitoring and assessing crop growth from emergence to harvest. Additionally, two test locations within the chickpea production area, but not included in the original seven locations, were used to generate the ML data that was then compared against the APSIM generated outputs for further benchmarking purposes.

2.1. Computing environment

All simulations and data analyses were performed on an Intel Core-*i*7 7600U CPU 2.9 GHz based computer with 16 GB RAM running Microsoft Windows 10 operating system. The APSIM version used was APSIM-NextGen (version 2020.02.05.4679) (Holzworth et al., 2018). The APSIM-NextGen prototype chickpea model was used as the crop model. Built-in features of the APSIM-NextGen User Interface were used to configure and execute factorial simulation experiments which generated the data used for building the MLEs.

2.2. Machine learning based emulators

The MLEs were developed and run in an installation of R (version 4.0.3 (2020-10-10)) (R Core Team 2020) in Microsoft Windows. The R environment was also used for data preparation and manipulation, reporting and graphics generation, with the packages ggplot2 (version 3.3.3) (Wickham 2016) and other packages from the tidyverse library (version 1.3.0) (Wickham et al., 2019) primarily used for these functions. The three MLEs, which are detailed below, were: nnet representing an ANN, Earth representing a MARS implementation and a Random Forest representing a decision tree implementation.

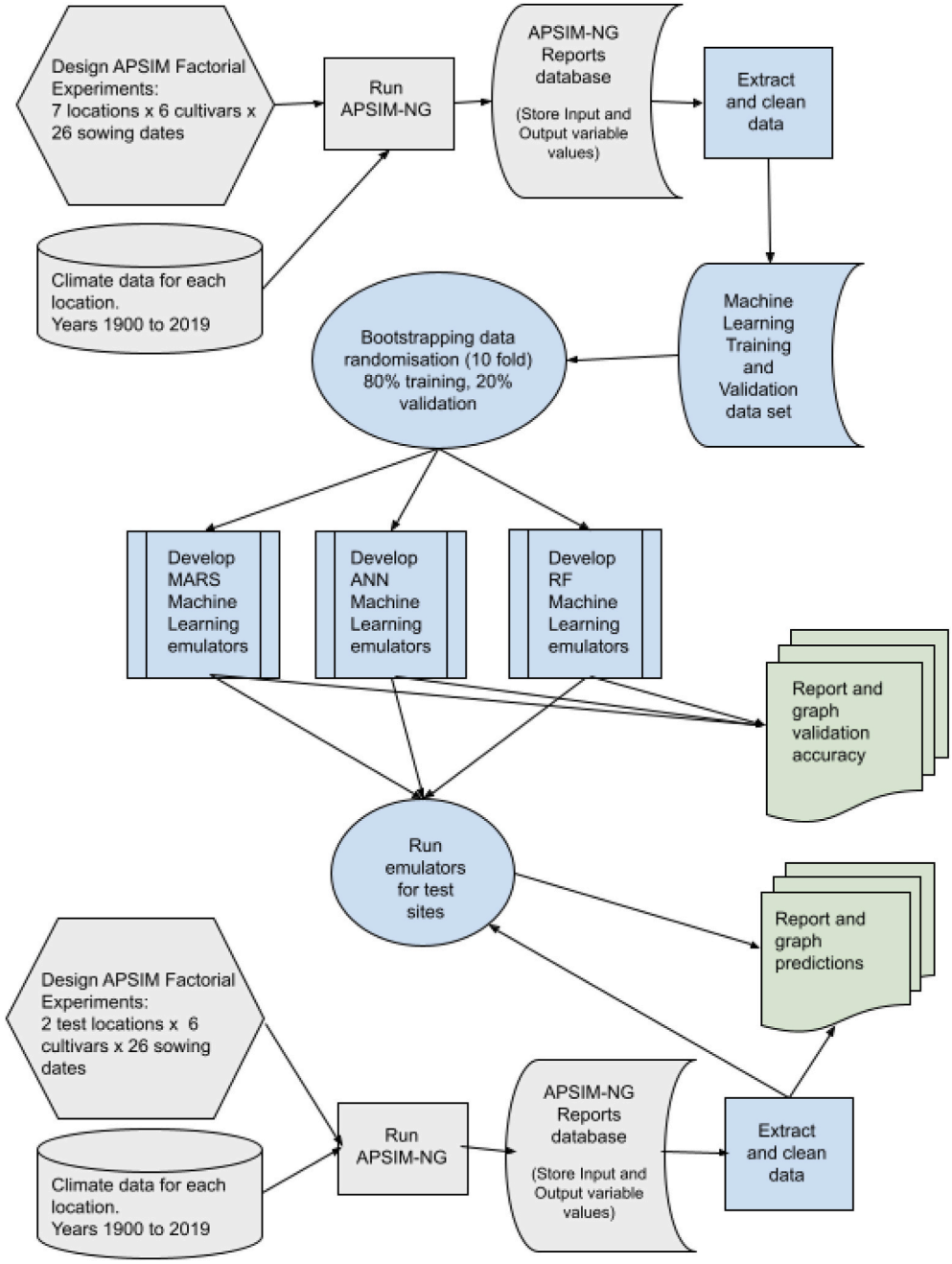


Fig. 1. Flowchart of work design for the generation of the synthetic datasets, and the training, validation, and testing of three machine learning based emulators.

2.2.1. Artificial neural network (ANN)

Artificial neural networks (ANNs) are some of the earliest ML algorithms. They represent a computing paradigm which consists of a massively interconnected network of nodes acting in parallel which simulate the actions of biological neurons. Each network connection is characterised by a weighting factor. Each neuron calculates the sum of its weighted inputs and produces an activation level output value via a generally nonlinear activation function. Models based on ANNs are developed by adjusting the number of neurons, number of layers of neurons (topology), neuron characteristics of activation functions and bias, and the sensitivities to training responses (Lippmann 1987). They are typically characterised by the 'black box' phenomena in ML where networks are trained on input data and automatically self-calibrate to classify or predict output values, the internals of the ML model generally not being able to be observed by a user of the system. Artificial neural networks have been used to predict outputs, such as yield, from biological and environmental systems (Shastry et al., 2016; Ghimire et al., 2018; Sanikhani et al., 2018; Nettleton et al., 2019; Shahhosseini et al., 2021) and were found to be the third most used ML approach in a review of Big Data applications in agriculture (Cravero and Sepúlveda 2021). In this experiment, the standard R library, *nnet* (version 7.3–15, 2021-01-21) based on the work of Venables and Ripley (2002) has been used to implement a feed-forward neural network with 20 nodes in its hidden layer and utilising 100 iterations for self-configuration. These settings were established by trial and error as optimal for predictive accuracy. The number of nodes was tested over the range of 10 nodes to 40 nodes, using increments of 2 nodes. The iterations for self-configuration were tested over a range of 50–200 in increments of 10. Default settings were utilised for all other model parameters. The ANN algorithm has been included in this study because of its general

applicability in environmental and biological studies and its wide use as a baseline for comparative ML studies.

2.2.2. Multivariate adaptive regression splines (MARS)

The Multivariate Adaptive Regression Splines (MARS) method for modelling, developed by Friedman (1991) and further described in Friedman and Roosen (1995), is a flexible regression modelling approach which has its roots in the recursive partitioning approach used in some forms of regression analysis. Continuous models with continuous derivatives are generated by repeatedly splitting product regression splines and introducing new basis functions for additional splines. This continues until the addition of more splines fails to improve the fitting of the response curves to the sampled data (Friedman 1991). The method meets the criteria for a ML data analysis tool as the resulting model is automatically determined by the data used to generate the model and does not require additional programming to address the specific problem that the data relates to. For this study, the *earth* package (version 5.3.0) (Milborrow 2020) in R was used to implement the MARS algorithm. The MARS algorithm has been included in the ML approaches for this study as it provides an interesting comparison for computational performance and predictive accuracy with the other two pure ML based approaches.

2.2.3. Random forest (RF)

Random forests (RF) are a computing paradigm based on an ensemble of decision trees. A random selection of features is used to split each node, with the accuracy of prediction used to weight the strength of each tree. The generalisation error for forests reduces as the number of trees increase and correlation between strong individual trees increases. Random forests have been shown to be quite robust with respect to outlier data points and noise within datasets (Breiman 2001; Sexton and Laake 2009).They are one of the most widely used forms of ML frameworks for both classification and regression, with Cravero and Sepúlveda (2021) finding that they are the second most referenced technique for analysis of big data in agriculture. There are many examples in agriculture of RF models being used for soil models (Gebauer et al., 2019; Hussein et al., 2020), yield forecasting (Kouadio et al., 2018; Feng et al. 2019, 2020; Obsie et al., 2020; Guo et al., 2021) and analysis of remote sensing (Belgiu and Drăguţ 2016; Dahms et al., 2016). The RF algorithm has been included in this study because of its wide applicability and use in agricultural and environmental modelling. The implementation of the RF algorithm used was the *randomForest* package (version 4.6–14 2018-03-22) (Liaw and Wiener 2018) in the R environment. Default values were used for all model settings. The default settings include that the number of features to be included in each decision tree is (p/3), where p is the number of input parameters. The default settings also specify that the algorithm calculates, via its internal code, the number of decision trees that are formed to optimise its predictive accuracy during its learning phase. The RF algorithm has been included in this study because of its wide applicability and use in agricultural and environmental modelling.

2.3. Simulation configuration

Simulations of chickpea crops were configured in APSIM-NextGen for seven locations throughout the chickpea growing regions in Australia (Fig. 2), for six chickpea cultivars, sown on 26 sow dates for each of 120 years (1900–2019). Reports were configured in APSIM to record all relevant input settings, summarise weather details, report the days after sowing of key crop development phases and report final above ground biomass and grain yield. For each combination of year, location, cultivar, and sowing date, one APSIM simulation was performed, with a total of 131,040 simulations. Each of those simulations produced one report with summarised output. Each report is considered one unit of observation in our analysis. In the end, a large database was obtained where each row was one simulation. The input settings and summarised

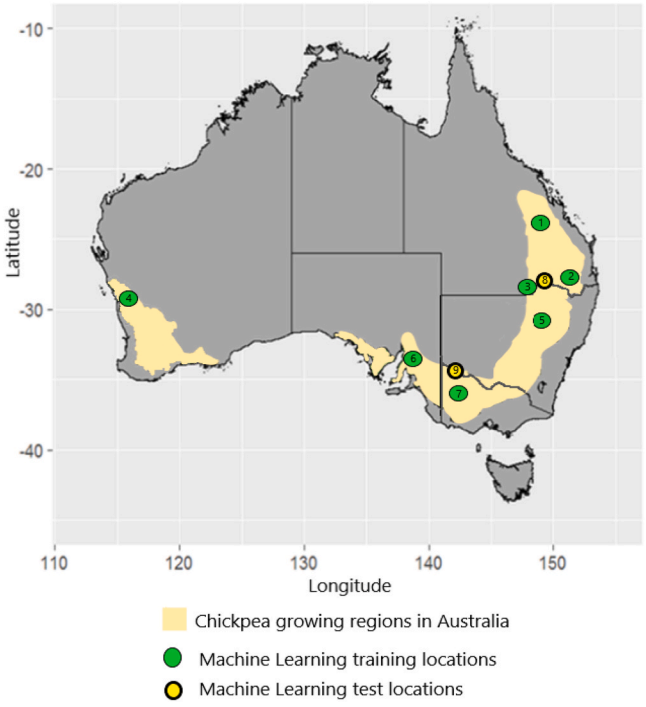


Fig. 2. Chickpea growing regions in Australia with the nine locations used for crop simulations marked by colored dots. Seven locations (green dots) were used to develop (train and validate) the machine learning based emulators. These, ordered by latitude are: 1) Emerald, Qld, 2) Bongeen, Qld, 3) Mungindi, NSW, 4) Mingenew, WA, 5) Gunnedah, NSW, 6) Clare, SA, and 7) Horsham, Vic. Additionally, two unseen test locations (yellow dots) 8) Goondiwindi, Qld, and 9) Mildura, Vic, were included for further model assessment.

weather details were used as the inputs to train the MLEs, with the crop development times, biomass and yield details used as the output targets for training and testing. In addition to the seven locations used to train the MLEs, two extra test locations, not included in the training and testing data sets, were used to test the robustness of the MLEs for locations outside the development data set.

2.3.1. APSIM simulation configuration

A typical soil type for the area was selected for each location. The details of these are shown in Table 2. All simulations had plant available soil water reset to 70% capacity on 1st March in each simulation year. Sowing dates were simulated for each 5-day interval from 30 March until 5 August. Row spacing was consistent at 0.5 m, sowing depth was 50 mm, and plant population was 30 plants/m^2 for northern sites (above 32° S), and 40 plants/m^2 for southern sites (below 32° S). Two chickpea genotypes, Desi and Kabuli, were sown at each location, with three varieties for each genotype; Seamer, HatTrick and CICA1521 for Desi; Monarch, Almaz and Kalkee for Kabuli. The genotypes differed from each other in four phenological parameters, each defined in terms of thermal time; ShootLag, VegTarget, LateVegTarget and FloweringTarget. The cultivar parameters used in APSIM are shown in Table 1.

2.3.2. Machine learning emulator inputs

The MLEs were developed and assessed for six output targets of interest for chickpea production: EmergenceDAS, FloweringDAS, PoddingDAS, MaturityDAS, Biomass and GrainWt. The models were then evaluated using data for seven production locations around Australia, with additional testing of the MLEs undertaken using two additional locations which were not included in the training and validation data set. Input factors (Table 3) used to train the MLEs were sourced from the reports generated by APSIM-NextGen. Weather details were summarised for each simulation for three blocks of time from the day of sowing: 0–30

Table 1
The APSIM NextGen phenological parameters of each chickpea cultivar used in this study. All parameter values are in thermal time units.

	Desi			Kabuli		
	Seamer	HatTrick	CICA1521	Monarch	Almaz	Kalkee
ShootLag	120	120	120	140	140	140
VegTarget	400	400	600	600	500	500
LateVegTarget	200	250	100	0	100	0
FloweringTarget	200	100	100	200	100	200

Table 2
Soil descriptions by location used for chickpea crop simulations.
The soil type descriptions and reference number refer to the APSoil database of soils from which the properties of the modelled soils were sourced.

Location	APSoil description and code	Profile depth (mm)	Plant available water capacity (mm)
1. Emerald	Grey Vertosol (No. 106)	1500	282
2. Bongeen	Black Vertosol (No. 001)	1800	335
3. Mungindi	Grey Vertosol (No. 906)	1800	339
4. Mingenew	Clay (No. 71)	1800	320
5. Gunnedah	Black Vertosol (No. 1174)	1800	285
6. Clare	Clay Loam on Clay Loam over Clay (No. 290)	1500	284
7. Horsham	Grey Cracking Clay (No. 1008)	1300	341
8. Goondiwindi	Grey Vertosol (No. 219)	1800	262
9.Mildura	Sandy Loam over Sandy Clay Loam (No. 332)	1400	142

Table 3
Machine learning input factors used for the development of the MLEs.

Input Factor Name	Description
AvgMaxT0_30	Average maximum temperature for 0–30 days after sowing
AvgMaxT31_60	Average maximum temperature for 31–60 days after sowing
AvgMaxT61_90	Average maximum temperature for 61–90 days after sowing
AvgMinT0_30	Average minimum temperature for 0–30 days after sowing
AvgMinT31_60	Average minimum temperature for 31–60 days after sowing
AvgMinT61_90	Average minimum temperature for 61–90 days after sowing
Cv	Chickpea cultivar (coded as 1 to 6 for the different genotype/cultivar combinations used)
FloweringTarget	Phenological parameter. Differs between genotypes.
FracPAWCmm	Amount of soil water present at sowing. As a fraction of PAWC.
Lat	Latitude of the sowing location.
LateVegTarget	Phenological parameter. Differs between genotypes.
PAWCmm	Soil's plant available water capacity to 1.5m depth (mm)
Population	Sown plant population in plants/m^2
Radn0_30	Sum of solar radiation for 0–30 days after sowing
Radn31_60	Sum of solar radiation for 31–60 days after sowing
Radn61_90	Sum of solar radiation for 61–90 days after sowing
Rain0_30	Sum of rainfall for 0–30 days after sowing
Rain31_60	Sum of rainfall for 31–60 days after sowing
Rain61_90	Sum of rainfall for 61–90 days after sowing
ShootLag	Phenological parameter. Differs between genotypes.
SowDepth	Sowing depth of crop
SowingDOY	Sowing date as Day Of Year
SowingESW	Extractable soil water at sowing
VegTarget	Phenological parameter in thermal time. Differs between genotypes.

days, 31–60 days and 61–90 days. Temperatures, both maximum and minimum, were averaged for each time block, while rain and solar radiation were summed to give totals for each time block. Soil water was represented in two ways. Firstly, a single value of how much plant extractable soil water (mm) was present at sowing (SowingESW) was included. Secondly, the soil's water holding capacity, measured as the plant available water capacity (mm) (PAWCmm) and the sowing water content as a fractional value of this (FracPAWCmm), were included in the input parameters. These two measures are highly correlated within a soil type, but variable between soil types.

2.3.3. Machine learning emulator targets

Six APSIM-NextGen chickpea model outputs were recorded in the APSIM reports, along with their corresponding input factor values, to create 'observed data' sets. Each of the three ML approaches was assessed on how well an emulator could predict the output values generated by the APSIM-NextGen simulation, as well as assessing the time taken, indicating computational effort required, to develop each ML emulator. This was undertaken on a comparative basis to assess differences between the various approaches.

2.4. Statistical measures for 'goodness-of-fit'

The 'goodness-of-fit' between the APSIM generated target values and those generated by the MLEs was assessed using the following statistical measures: (1) mean bias (MB), (2) mean absolute error (MAE), (3) root mean squared error (RMSE), (4) coefficient of determination (R^2), and (5) coefficient of efficiency (COE_{LM}, also known as Legates-McCabe index) (Legates and McCabe Jr 1999). These metrics were used to compare the ML predicted versus APSIM-generated value datasets to determine the degree of match between the tested datasets.

Mean bias (MB) measured in days or kg/ha, depending on the output

$$MB = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n} \quad (1)$$

Mean absolute error (MAE) measured in days or kg/ha, depending on the output

$$\mathrm{MAE} = \frac{\left|\sum_{i=1}^{n}\left|(y_i - x_i)\right|\right|}{n} \quad (2)$$

Root mean squared error (RMSE) measured in days or kg/ha, depending on the output

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n}\left((y_i - x_i)^2\right)}{n}\right)} \quad (3)$$

Coefficient of determination (R^2)

$$R^2 = \left(\frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}\right)^2 \quad (4)$$

Coefficient of efficiency (COE_{LM}: Legates McCabe index)

$$COE_{LM} = 1 - \left[\frac{\sum_{i=1}^{n}|y_i - x_i|}{\sum_{i=1}^{n}|x_i - \bar{x}|}\right] \quad (5)$$

In equation 1 through 5: '*n*' is the number of pairs of (APSIM-generated (*x*), predicted (*y*)) values, where APSIM-generated values are the model output values generated by APSIM; and 'predicted' are the ML emulator 'simulated' value for the model output. '*i*' is the output generated from the *i*th set of input parameters. The six target outputs generated were: EmergenceDAS, FloweringDAS, PoddingDAS, MaturityDAS, Biomass and GrainWt.

2.5. Variable importance

The contribution that each input factor (Table 3) has towards the value of the output target (EmergenceDAS, FloweringDAS, PoddingDAS, MaturityDAS, Biomass or GrainWt) is calculated by each of the ML algorithms. The values reported and presented as a heat-map (Fig. 4) have been standardised so that the most significant input is assigned an importance index value of 100, non-contributing inputs are given a value of zero (0) and all other inputs are rated with index values proportionate to the most influential input. Each of these routines was configured to report index values rated on the reduction in the residual sum of squares (RSS) value of generated predictions versus the actual target values when the input parameter being assessed was included in the model. That is, the input that resulted in the greatest reduction in the RSS when it was added to the algorithm was assigned an importance index of 100.

3. Results

3.1. Performance based on training data set

Results from the training data set, where the MLEs were trained on a random subset of 80% of the data and then validated on the unused 20% of data, showed that each of the three ML approaches, ANN, MARS and RF algorithms, can produce MLEs with significant predictive accuracy for each of the six crop output targets (Table 4). There were no observed occurrences of any model encountering overfitting issues, which would have been evidenced by the accuracy of the predictions of the validation data set being significantly lower than the accuracy for the training data sets. All reported values are those for the validation data sets for each MLE. The accuracy of prediction, the importance of input variables used to achieve these predictions, and the computational effort required to develop the MLEs, did vary between the approaches. Across all outputs, the RF emulators showed the best and most consistent accuracy at prediction. This, however, come at significant computational investment.

3.1.1. Graphical and statistical analysis of ML approaches

A visual inspection of the plots of ML predicted versus APSIM generated data as Hexbin plots (Fig. 3) confirmed the accuracy of the dataset of validation predictions for the six target outputs (EmergenceDAS, FloweringDAS, PoddingDAS, MaturityDAS, Biomass and GrainWt). The corresponding values from the statistical analyses of the data of these graphs is presented in Table 4. Of note is the superiority of the RF emulators' predictions for each output target. All three MLEs produced exceptional results for predicting the start of flowering (FloweringDAS). Regional variations are evident for each ML emulator with northern locations flowering after a shorter duration than locations with more southern latitudes (cooler climates). Predictions of podding date were much less precise for each of the MLEs, with noticeably wider variations occurring at Mingenew. This indicates that some crop growth factor(s) used within APSIM which affected early pod development were possibly not included in the input parameter details. One APSIM parameter that was omitted and fits this profile is the "Phenology.Budding.Target.XYPairs" values. These values represent the budding target response curve measured in thermal time. This response relationship was not converted into an equivalent single response numeric suitable of inclusion as an ML predictor variable, and so was not included in the input parameter list for developing the MLEs. While producing the most accurate predictions of podding date for most locations, most noticeably for Horsham, the RF emulator's predictions for Bongeen were slightly less accuracy than other MLEs. There was no clear indication as to why this was the case. The 'black-box' nature of ML models makes the analysis of outputs and explanation of model performance challenging. The above ground crop biomass and the crop yield, reported as GrainWt, were the least predictable outputs for each ML algorithm. The MARS emulators, on average, had the greatest tendency to under-predict the output values, as indicated by the negative mean bias values (Table 4). The RF emulators had about one quarter the amount of variance of the other two MLEs, as shown by the mean absolute error (MAE) values (Table 4). The ANN and MARS emulators each produced predictions

Table 4
A summary of the predictive ability of the MLEs against outputs generated by the APSIM-NextGen chickpea crop model.
The mean and standard deviations (sd) of the statistical measures for goodness-of-fit analysis for the training validation of the machine learning emulators (MLEs) using 10 fold repetition. The statistics shown are, MB: mean bias reported in days or kg/ha, depending upon the output variable; MAE: mean absolute error reported in days or kg/ha, depending upon the output variable; RMSE: root mean squared error reported in days or kg/ha, depending upon the output variable; R^2: coefficient of determination; COE_{LM}: coefficient of efficiency (Legates McCabe index). The three MLEs are Artificial Neural Networks (ANN), Multivariate Adaptive Regression Spline (MARS) and Random Forest (RF).

Development validation statistics of accuracy						
Emulator/Target		MB	MAE	RMSE	R^2	COE_{LM}
ANN						
EmergenceDAS	mean	0.000	0.683	0.884	0.950	0.791
(days)	sd	0.006	0.003	0.007	0.001	0.002
FloweringDAS	mean	−0.012	1.119	1.525	0.995	0.933
(days)	sd	0.035	0.060	0.076	0.001	0.004
PoddingDAS (days)	mean	−0.026	5.152	7.838	0.949	0.817
	sd	0.080	0.156	0.130	0.002	0.005
MaturityDAS (days)	mean	−0.013	3.255	4.665	0.980	0.880
	sd	0.050	0.026	0.042	0.000	0.001
Biomass (kg/ha)	mean	−0.279	76.030	102.031	0.923	0.757
	sd	1.225	1.055	1.227	0.002	0.004
GrainWt (kg/ha)	mean	0.076	35.291	47.784	0.910	0.740
	sd	0.876	0.418	0.512	0.002	0.003
MARS						
EmergenceDAS	mean	−0.002	0.695	0.899	0.948	0.787
(days)	sd	0.004	0.003	0.005	0.001	0.001
FloweringDAS	mean	0.002	2.031	2.658	0.984	0.879
(days)	sd	0.019	0.143	0.187	0.002	0.009
PoddingDAS (days)	mean	−0.013	5.968	8.891	0.934	0.788
	sd	0.051	0.030	0.050	0.001	0.001
MaturityDAS (days)	mean	−0.018	4.009	5.600	0.971	0.851
	sd	0.030	0.041	0.058	0.001	0.002
Biomass (kg/ha)	mean	−0.103	87.392	115.039	0.902	0.721
	sd	0.867	0.494	0.859	0.001	0.002
GrainWt (kg/ha)	mean	−0.366	40.606	53.872	0.886	0.701
	sd	0.315	0.336	0.558	0.002	0.002
RF						
EmergenceDAS	mean	0.001	0.221	0.326	0.993	0.932
(days)	sd	0.002	0.003	0.010	0.000	0.001
FloweringDAS	mean	0.012	1.000	1.355	0.996	0.940
(days)	sd	0.011	0.008	0.012	0.000	0.001
PoddingDAS (days)	mean	0.029	2.507	4.356	0.984	0.911
	sd	0.026	0.021	0.057	0.000	0.001
MaturityDAS (days)	mean	0.018	1.808	2.830	0.993	0.933
	sd	0.018	0.014	0.047	0.000	0.001
Biomass (kg/ha)	mean	−0.037	16.088	26.956	0.995	0.949
	sd	0.254	0.138	0.761	0.000	0.000
GrainWt (kg/ha)	mean	0.007	12.272	20.241	0.984	0.910
	sd	0.133	0.094	0.313	0.001	0.001

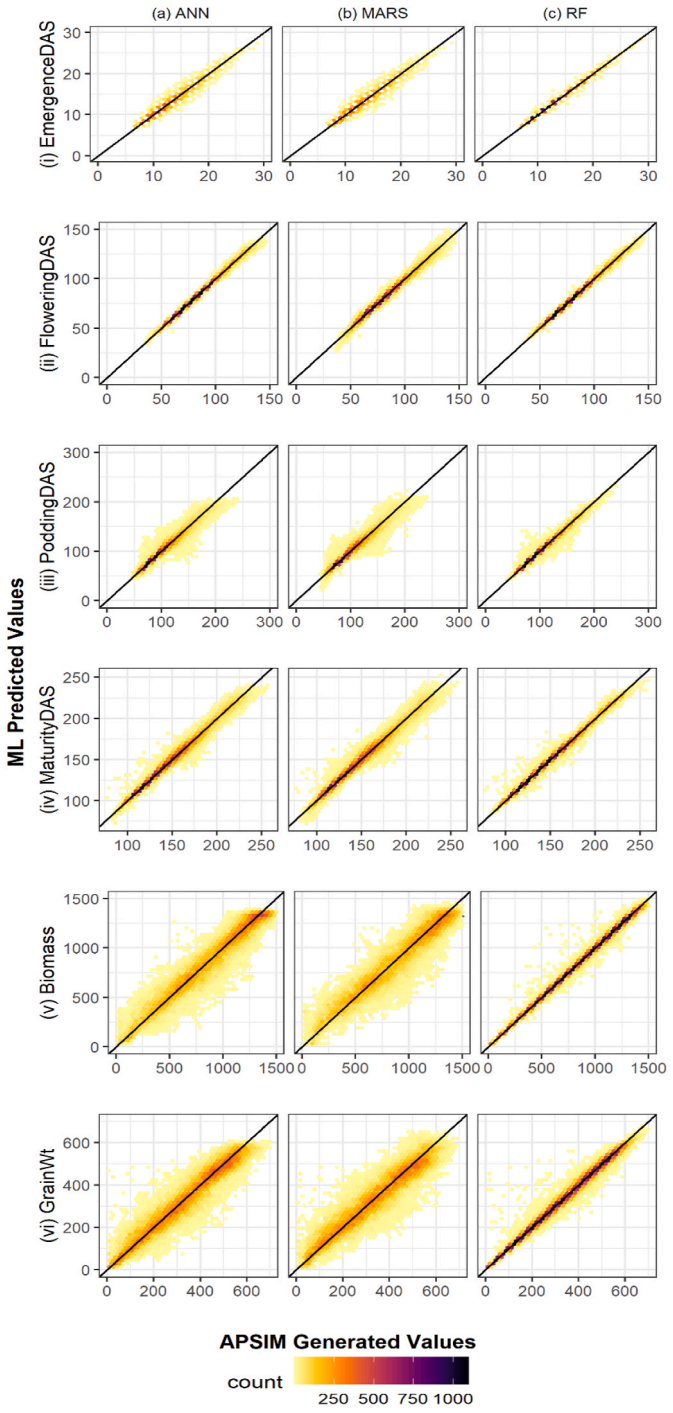


Fig. 3. HexBin plot of the distribution density of data points for the emulator development validation data tests.
Each panel shows the summary of 26,185 data points, being the 20% validation portion of the full dataset of 130,928 data points. These consisted of 26 sowing dates for each of six cultivars at seven locations for each of 120 years, less crops that did not produce a yield.

with a wider distribution around the APSIM predicted values than the predictions of the RF emulators. The data points, however, are still most densely clustered along the one-to-one line, as Fig. 3 shows. Again, RF emulators did a noticeably better job of predicting each of these outputs than emulators based on the other ML algorithms.

Further analysis of the least accurate ten percent of predictions for each MLE for the outputs biomass and crop yield, showed highly variable results between the three MLEs. For the ANN emulators, the least accurate predictions generally resulted in significant under-predictions of biomass and crop yield. These results were strongly associated with late maturing crops, with a mean MaturityDAS value of 172 days compared to an average for the rest of the simulations of 148 days. A likely cause of such errors is that environmental factors that caused a decrease in the above ground crop biomass and yield in the APSIM simulations occurred late in the crop lifecycle. With ML weather inputs only recording meteorological data up to 90 days after sowing, weather events or dry conditions late in the crop cycle would not have been considered by the ANN emulators. For the MARS emulators, the least accurate predictions also tended to result in under-prediction of biomass and crop yield, but these were not biased towards late maturing crops. Instead, these simulations tended to have drier soil conditions at sowing (low SowingESW) and lower solar radiation levels later in the crop's life. The RF emulators showed a very different pattern again, with the least accurate ten percent of predicted biomass and crop yield values generally being associated with over-prediction of values. For the RF emulators, the poor predictions were more strongly associated with elevated soil water at sowing (high SowingESW), higher than average rainfall beyond 60 days and lower solar radiation during the same period. Poor prediction of biomass was also associated with earlier sowing dates and small LateVegTarget parameter values. The 'black-box' nature of ML models makes detailed and accurate investigation of underlying model issues impossible.

In the case of the RF emulators, it is worth noting that the outlier values only represent between 20 and 40 data points out of a set of 26,185 data points, indicating that any visual impact of these points might have in data plots is overstating their importance. This is confirmed by considering the hexbin plot of the distribution density of the data points (Fig. 3). Low numbers of data points are seen in a clearer perspective of their importance. One interesting aspect to note that differs between the MLEs is the generation of a small number of erroneous negative values for GrainWt. RF did not suffer from this feature, while the MARS emulators showed this feature for both the crop yield and above ground biomass predictions. One of the noted strengths of the RF algorithm is bootstrap aggregation, also known as 'bagging', which results in an ensemble of RF models. This approach has the benefits of reducing bias and variance in the resulting prediction model and producing a more representative outcome for variable data (Sexton and Laake 2009; Biau and Scornet 2016). A disadvantage of this ensemble approach is the increased computational effort required. The pattern of fast emulators being the least accurate in both bias and error statistics calculated, as well as the accuracy of predicted target values, is observed in the data presented in Table 4. This is most likely a reflection of the fact that accurate predictions are more consistently produced when greater numbers of values are processed and averaged. There appears to be a generalised inverse relationship between emulator speed and accuracy of prediction.

3.1.2. *Variable importance*

By comparing the influence that the input factors have on the outputs across each of the MLEs, patterns and variations can be observed in what is driving each emulator. Fig. 4 highlights the patterns of the index values. For EmergenceDAS, all three MLEs were strongly influenced by the maximum and minimum temperatures during the first 30 days after sowing. This is expected as emergence is primarily a temperature driven response in the chickpea model, and it occurs in the first 30 days of the crop simulation. The MARS algorithm was shown to be significantly more sensitive to the input variable ShootLag than were the other two MLEs for the prediction of EmergenceDAS, with the ShootLag input having an input variable importance (Fig. 4) of 93 for the MARS emulator, but values of only nine and six for the other emulators. Interestingly the ANN and MARS emulators had consistent values for the accuracy of EmergenceDAS predictions (Table 4), with R^2 values of 0.95, and COE_{LM} values of 0.79, which is a clear demonstration that different ML algorithms can use different input information to achieve similarly

Output Target	EmergenceDAS			FloweringDAS			PoddingDAS			MaturityDAS			Biomass			GrainWt		
Emulator	ANN	MARS	RF	ANN	MARS	RF	ANN	MARS	RF	ANN	MARS	RF	ANN	MARS	RF	ANN	MARS	RF
Input Factor																		
AvgMaxT0_30	97	100	89	27	16	13	9	14	4	6	8	4	9	0	6	8	7	7
AvgMaxT31_60	4	0	7	18	100	93	9	11	14	4	0	13	5	0	46	11	0	42
AvgMaxT61_90	0	0	3	19	12	100	17	100	100	16	100	100	26	86	33	36	100	72
AvgMinT0_30	100	40	100	18	24	9	8	18	5	4	12	3	4	23	9	5	32	12
AvgMinT31_60	2	0	7	9	33	34	3	35	6	3	33	3	1	19	5	3	0	6
AvgMinT61_90	0	0	2	10	10	45	19	42	55	10	33	61	6	0	9	19	15	16
Cv	0	0	6	32	39	11	6	7	2	0	0	1	0	0	1	1	0	2
FloweringTarget	0	0	0	2	0	1	2	0	0	1	0	0	1	0	0	2	0	0
FracPAWCmm	2	0	2	3	0	1	19	0	1	13	0	1	75	0	100	17	24	74
Lat	0	0	22	2	0	22	12	9	17	10	18	20	30	46	97	10	28	100
LateVegTarget	4	0	1	17	0	3	4	24	2	2	0	1	16	0	1	23	0	1
PAWCmm	0	0	2	0	0	3	2	9	4	1	0	3	21	58	53	31	49	40
Population	0	0	3	1	0	1	3	0	2	3	33	2	66	0	17	50	0	11
Radn0_30	2	0	11	4	0	14	2	0	4	1	0	3	8	0	7	3	0	6
Radn31_60	6	0	3	2	18	51	1	27	38	0	33	39	3	0	12	6	0	22
Radn61_90	0	0	7	4	6	13	7	0	46	3	0	46	17	6	9	8	18	13
Rain0_30	0	0	2	0	0	0	1	0	1	0	0	1	11	35	23	15	39	20
Rain31_60	0	0	1	0	0	0	1	0	1	0	0	1	18	100	35	23	78	28
Rain61_90	0	0	1	0	0	0	1	0	1	0	0	1	19	58	26	23	49	21
ShootLag	9	93	6	7	0	8	7	0	1	4	0	1	2	0	0	10	0	1
SowDepth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SowingDOY	1	17	29	100	53	14	100	43	17	100	33	14	19	0	4	26	0	6
sowingESW	2	0	1	3	0	1	24	0	2	16	0	1	100	100	75	100	78	54
VegTarget	2	0	1	7	0	3	4	0	1	3	0	0	10	0	0	18	0	1

Fig. 4. Heat maps of input variable importance for three MLEs.
Results are for six output parameters. Importance indices are rated from zero (0) for no effect on the output value, to 100 being the input with the most significant effect on the output values. Index values are relative to the most significant input rated at 100.

accurate predictions. This finding is consistent with the findings of Shahhosseini et al. (2019) who also found that ML models differed in their sensitivities to input variables.

Other output targets showed greater diversity in the input variables identified as most important. For the ANN emulators, the input SowingDOY was very significant for predicting the output target FloweringDAS, while the MARS and RF emulators rated average maximum temperatures between 31 and 60 days after sowing as highly influential. PoddingDAS and MaturityDAS showed something of a consist pattern between MLEs, with SowingDOY being most important for the ANN emulator, while average maximum temperature for 61 to 90 DAS was the most significant for the MARS and RF emulators. The RF emulator was the only one to have an additional value over 50, that of AvgMinT61_90.

The patterns of rating significance for both biomass and crop yield were similar in each of the MLEs. Above ground biomass and crop yield were both strongly influenced by SowingESW by all ML algorithms, although RF emulators used the closely correlated FracPAWCmm input instead. Only the RF emulator rated the latitude (Lat) variable as a significantly important input, which it did for both above ground biomass and crop yield. Both the MARS and the RF emulators used the AvgMaxT61_90 for crop yield prediction, while no temperature, rainfall or radiation inputs were rated above an importance of 36 by the ANN emulator for crop yield.

3.1.3. Computational requirements

The time taken to train the MLEs is an indicator of the computational costs associated with developing each emulator system. Table 5 shows that there was a great spread in the computational requirements needed to develop each type of emulator. Times ranged from 12.1 s for the MARS algorithm to develop a predictive emulator for the output EmergenceDAS, to a high of 17,644.8 s (4hrs 54mins) for the RF algorithm to produce a predictive emulator for the same output. On average, MARS emulators were developed with least computational effort, ANN emulators were almost three times more costly, and RF emulators were approximately 500 times more costly. This observation is based solely on the performance measured for the code libraries and computing environment used for this study.

Table 5
Time (in seconds) taken to train each MLE.
Training data sets used 26,185 data points for each target output. The times are representative only and were obtained from developing the MLEs in an R environment on an Intel core-*i7* laptop computer.

Output	ANN	MARS	RF
EmergenceDAS	77.8	12.1	17644.8
FloweringDAS	77.9	37.4	10930.8
PoddingDAS	86.5	34.8	16149.9
MaturityDAS	86.3	31.5	17191.2
Biomass	76.5	35.0	13544.1
GrainWt	77.5	34.9	14530.5
Average: (all times in seconds)	80.4	30.9	14998.6

3.2. Performance at test locations

The MLEs developed using data from seven locations within the Australian chickpea production regions, were tested using data from two additional locations, also within the same production regions. Hexbin plots of the APSIM generated values plotted against the values generated by the predictive MLEs are shown in Fig. 5, with the statistical analyses of the 'goodness-of-fit' of the data values provided in Table 6. For predictions of EmergenceDAS and FloweringDAS, the three ML algorithms, ANN, MARS and RF, all performed well with consistent R^2 values of 0.91 for EmergenceDAS and of 0.98 for FloweringDAS. The corresponding values of COE_{LM} ranged between 0.72 and 0.73 for EmergenceDAS and between 0.86 and 0.88 for FloweringDAS. Values for each test location were equally well predicted. For the three ML emulators, ANN, MARS and RF, the prediction of MaturityDAS was the next most accurate output with R^2 values of 0.95 and 0.96 and COE_{LM} values ranging from

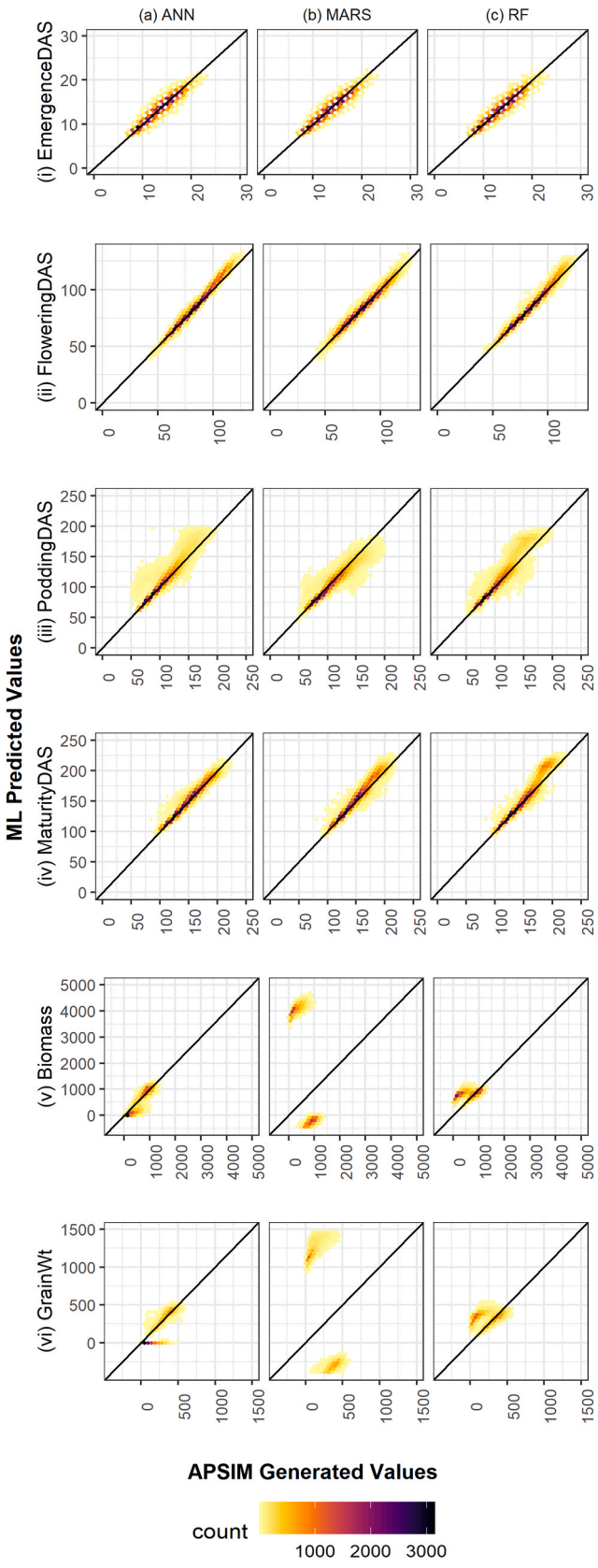


Fig. 5. HexBin plot of the distribution density of data points for the test location data sets.
Each panel shows the summary of 37,395 data points, being the full dataset consisting of 26 sowing dates for each of six cultivars at two locations for each of 120 years, less crops that failed to emerge.

Table 6
The predictive ability of the MLEs for two unseen test locations against outputs generated by the APSIM-NextGen chickpea crop model.
Statistical measures for goodness-of-fit performance analysis for two test locations. The statistics shown are, mean bias (MB), mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), coefficient of efficiency (Legates McCabe index) (COE_{LM}). The three machine learning emulators (MLEs) are Artificial Neural Net (ANN), Multivariate Adaptive Regression Spline (MARS) and Random Forest (RF).

Statistics of accuracy for unseen test locations					
Emulator/Target	MB	MAE	RMSE	R^2	COE_{LM}
ANN					
EmergenceDAS (days)	0.07	0.74	0.95	0.91	0.72
FloweringDAS (days)	0.61	1.58	2.10	0.98	0.88
PoddingDAS (days)	3.32	8.97	13.18	0.87	0.60
MaturityDAS (days)	2.04	3.84	5.42	0.96	0.82
Biomass (kg/ha)	−136.26	193.16	260.08	0.76	0.35
GrainWt (kg/ha)	−21.50	74.85	95.84	0.79	0.39
MARS					
EmergenceDAS (days)	−0.08	0.75	0.95	0.91	0.72
FloweringDAS (days)	−0.04	1.77	2.25	0.98	0.86
PoddingDAS (days)	−1.96	6.26	9.17	0.89	0.72
MaturityDAS (days)	3.68	5.41	7.32	0.95	0.75
Biomass (kg/ha)	1315.60	2437.08	2772.21	0.62	−7.25
GrainWt (kg/ha)	227.21	901.09	931.87	0.48	−6.28
RF					
EmergenceDAS (days)	0.02	0.73	0.93	0.91	0.73
FloweringDAS (days)	0.70	1.74	2.47	0.98	0.87
PoddingDAS (days)	4.40	7.23	11.43	0.90	0.68
MaturityDAS (days)	3.88	6.12	9.18	0.95	0.72
Biomass (kg/ha)	224.23	290.65	374.34	0.20	0.02
GrainWt (kg/ha)	105.06	134.92	173.40	0.10	−0.09

0.72 to 0.82. This was followed by the predictions for PoddingDAS with R^2 values ranging from 0.87 to 0.90 and COE_{LM} values ranging from 0.60 to 0.72.

All three ML approaches, however, failed to accurately predict above ground biomass and crop yield at the unseen test locations, although the ANN models did come close to being acceptable. The MLEs were incapable of making accurate predictions for the test locations based on the data from the training locations. Given that biomass and crop yield were both strongly influenced by soil water holding capacity and soil water content at sowing (Fig. 4), it is most likely that insufficient soil types and soil water conditions were included in the original data set to allow the test locations to be accurately modelled. The test locations effectively fell outside the parameter value ranges and effects observed at the training locations and so predicted values were nonsense. This highlights a failing of the input data used, not of the MLEs or the modelling approach.

4. Discussion

4.1. Performance with training data set

The focus of this research is on the accuracy of MLEs to predict APSIM generated outputs, and the computational costs associated with developing the MLEs. The discussion that follows concentrates on these issues and does not attempt to review or discuss the implications of the agronomic or environmental results which would involve shifting the focus to a review of the APSIM-NextGen chickpea model itself. The results of this study have shown that MLEs can be developed that can aid in expanding biophysical crop modelling systems, such as APSIM, by providing a computationally efficient approach for the generation of very large synthetic datasets, such as are required for IA and variance-based SA. They show that all three ML approaches reviewed are

capable of being used to generate predictive regression MLEs for the crop model outputs tested. The FloweringDAS prediction was the most accurate output for each of the MLEs, indicating that the input factors included did cover all the important driving variables for this output. It is revealing that the importance of the input variables (Fig. 4, panel 'FloweringDAS') was not consistent between the different algorithms. For FloweringDAS, the ANN emulator was heavily reliant upon the time of sowing, with no other input coming close to having as significant an impact. The MARS emulator relied almost entirely on mid-season maximum temperatures, with its next most important input, time of sowing, rated as only half as important. The RF emulator was most strongly influenced by mid and late-season maximum temperature. This shows clearly that great care must be taken if interpreting the input importance values for MLEs as being an accurate predictor of the importance of input factors for an underlying model. Different algorithms can, and do, predict the correct answer in the majority of instances, using significantly different importance weightings of input values. Boehmke and Greenwell (2019) have previously warned that algorithms, like that used in the MARS approach, can give misleading results for variable importance where there are closely correlated input factors. This is due to the algorithms approach of selecting input factors based on their contribution to an output value and discarding additional inputs if they do not improve the prediction by some given marginal amount. This can result in only one of a closely correlated set of inputs being used to predict output values, with the other inputs, although equally as influential on the output, rated as not used or of low importance. Breiman (2001) and Dumancas and Bello (2015) indicate that the RF algorithm is well suited to cope with multico-linearity of inputs, and so is not subject to this limitation to the same degree as the MARS algorithm. For neural networks, which is represented by the ANN algorithm, the robustness and accuracy of their predictions have been found to be adversely affected by co-linearity between input factors (Dumancas and Bello 2015; Samarasinghe 2016). These authors advise that feature selection needs to be undertaken in order to remove non-influential inputs and inputs that exhibit co-linearity from the data set, before reliable neural net models can be built. For the purpose of comparing the ML algorithms based on a consistent approach, this step was not undertaken in this study.

The greatest differences between the accuracy of predictions of the MLEs was for the outputs of above ground biomass and grain weight (yield). These two outputs are the ones in the output set most influenced by a wide range of crop, environmental and management factors, and represent the sum of everything the crop has experienced. They are key outputs for most crop models (Stöckle et al. 1994, 2003; Jones et al., 2003; Keating et al., 2003). For these two outputs, the RF emulator was clearly a superior predictor than the emulators produced by the other two ML algorithms. The reasons for this difference in accuracy are not easily determined. Contributing factors are likely to include the inherent suitability of the underlying ML algorithm for the data being analysed, and the extent to which the data set has been optimised for the ML approach. One factor that was identified during analysis of this data was that the summary climate details were only recorded until 90 days after sowing, while many of the crops with the poorest predictions of biomass and yield reached maturity (as shown by harvest date) well beyond this cut-off. It is probable that adverse weather conditions during the final stages of crop growth and crop maturation resulted in unpredictable crop vigour and yield loss. Extended periods of weather details in the input parameters may have aided in more accurate predictions of biomass and yield. While feature selection and dimensionality reduction steps are warranted for the neural net based algorithms (Samarasinghe 2016), the purpose of this study was to compare the performance of the core approaches. The investigation of optimal feature selection algorithms would constitute a research study in its own right. It is worth noting that, under the constructs of this study, where the outputs of the simulation model are being predicted rather than trying to match real world observations, all potential input factors for the MLEs are known, albeit a very large number of them. This makes the possibility of identifying a complete set of driving input factors a feasible objective.

Based on the accuracy of predicted values, the RF algorithm is the best of the three algorithms tested. The accuracy of predicted output values produced by the RF emulators for the locations on which it was trained are good, with the lowest accuracy being for both PoddingDAS and GrainWt at $R^2 = 0.98$ and $COE_{LM} = 0.91$. With this level of accuracy, the RF emulators could be used to predict with a high degree of confidence, any of the six model outputs for any of the seven training locations for input values within the range of values observed in the training set. The design of this experiment meant that one set of input factors was tested for their ability to be used to predict each of the six outputs. With careful review and iterative testing, it should be possible to improve the predictive accuracy for any chosen output.

The computational costs involved in developing, or training, the MLEs (Table 5) varied widely between the different algorithms. One potential application of using MLEs to efficiently expand the output of process-driven crop models is that of running SA. Studies, such as that by Zhao et al. (2014), which looked at the SA of the APSIM wheat model, focused considerable effort on identifying a data efficient analysis method to maximise their research outcome and minimise their cost of running APSIM simulations. The use of MLEs could provide an alternative method for generating such datasets. To be useful as a tool to run IA or SA as a background process on a systems model, such as APSIM, an MLE needs to be able to be rapidly developed, used, and discarded rather than having an iterative development and retention lifecycle. This is because each analysis will be based on a different scenario and designed to test different input parameters or different ranges for input parameter values each time they are run. As the MLEs are generated for specific sets of inputs and can only be used to predict outputs for input settings within the value ranges with which they were developed, reuse of MLEs may be limited. This would depend upon the design of the experiment at development time. Even where MLEs can be reused, great care would be required to ensure that the value ranges of all input parameters were within the development limits of the MLE, thus avoiding covariate shift, and that the mix of those inputs was of a pattern that was not dissimilar to patterns used to develop the MLEs. While broadly applicable MLEs might be possible to produce, a narrowly applicable MLE developed for a specific application is a safer option if unpredictable outcomes are to be avoided. The focus on the expected use and life timeframe of the emulator is a key feature of this study that differs from many studies into the development of ML models. Comparisons of development times of ML models are not readily available in the literature. In this study the MARS algorithm was, on average, almost 500 times faster to train than the RF algorithm, with the ANN algorithm being approximately 200 times faster than the RF algorithm. It must be noted that this represents just one snapshot of specific implementations of three algorithms out of potentially dozens of alternative algorithms. The code used to implement the algorithm solution, the computing environment utilised to run the code and the computing hardware that the ML was run on, all have the potential to significantly affect the outcomes of such a comparison. Advances in, or reimplementation of, any of these factors, or the selection of alternative algorithms or environments, will have effects on the outcomes. For this study, the outcome is clear; the RF algorithm was the most accurate of the ML approaches, but it came at a significant computational cost. The superior results from the RF emulator are in contrast to Kouadio et al. (2018) who found an extreme learning machine, which is an advanced form of ANN algorithm, superior at forecasting coffee yield. Obsie et al. (2020) reported an extreme gradient boosting model produced better results than a RF model for blueberry yield prediction, although both the gradient boosting model and the RF model performed better than a multi linear regression approach. Other researchers (Jeong et al., 2016; Dayal et al., 2019; Feng et al. 2019, 2020; Lawes et al., 2019) have chosen RF models as their preferred ML approach in studies predicting crop growth.

4.2. Performance with test locations

A second part of this study assessed the robustness of the MLE solutions by generating prediction for output values at locations which were not included in the development (training and validation) data sets. The three ML algorithms were not as accurate in predicting the chronological development of the crop, that is the EmergenceDAS, FloweringDAS, PoddingDAS and MaturityDAS, as when predicting values for locations in the training data set, but predictions were not unrealistic for the ANN, MARS and RF emulators, as shown in Fig. 5 and associated statistical values in Table 6. This demonstrates that the MLEs, if developed with sufficiently diverse data sets, are robust enough to predict outputs for any location in the production region, regardless of whether that location was used in the training data set or not. The FloweringDAS predictions, with R^2 values 0.98 and COE_{LM} values ranging from 0.86 to 0.88 for each of the algorithms, were the most accurate of the predictions for the test locations. The other statistical measures generated to test the accuracy of the emulators, MB, MAE and RMSE, all followed the same relative patterns of which was the most to least accurate emulator, with RF being the most accurate, ANN being next, and MARS being the least accurate. With this level of accuracy, the use of any of these three MLEs to predict flowering date as days after sowing, for any location within the Australian chickpea production regions, would be justifiable.

By using test locations, most of the input factors used to train the MLEs were able to be controlled and ensure that they fell within the ranges used to develop the MLEs. Factors that were not controlled and had the potential to fall outside the development dataset boundaries were related to the soil, specifically the water holding capacity of the soil and starting soil moisture levels. The predictions for above ground biomass and grain weight (Fig. 5 and Table 6) are shown to be erroneous for all three ML algorithms. As noted previously, these outputs reflect the sum of all the factors that influence crop growth. Consequently, their predicted values are most likely to reveal any weakness in the robustness of the MLEs. Even though the management and genomic factors were consistent with the training data, the test locations introduced different soils to the simulations. For example, the predictions of biomass and yield for Mildura were the least accurate and most varied between the different MLEs. Mildura soil was the only sandy loam in the data set and had the lowest water holding capacity of any of the soils. This soil was the most contrasting soil, and the emulators performed most poorly with it. This is consistent with the findings of Shahhosseini et al. (2021) who identified soil water parameters as key drivers of ML models used to predict corn yields. Some of the patterns that define the relationships between input factors and output values observed at the test locations in our study were not present in the training data, so none of the ML algorithms could predict them for the new locations. The situation where input data values fall outside the range of the training dataset is referred to as covariate shift, and it is a known limitation of ML that predictive models are unable to handle such data variations. This clearly stands as a warning about the potential use of MLEs for generating synthetic dataset by expanding the outputs of process driven models. All patterns of input factors affecting output values must be included in the training data to develop an ML emulator that is capable of robustly predicting outputs. Other recent research integrating process-driven models with ML has focused on the effects of climate change on crop yields (Feng et al., 2019; Leng and Hall 2020). Both studies have reported significant benefits in integrating the two modelling approaches but have not highlighted the dangers and limitations of supplying incomplete data sets to the ML models during development. In this study, the training data included all required patterns for predicting FloweringDAS but lacked details which determine above ground biomass and grain weight. As a result, the FloweringDAS predictors are more robust than the above ground biomass and yield predictors.

5. Conclusion

This study has shown that emulators of crop models, built on ML algorithms, can be developed to predict a range of simulated crop outputs. The accuracy of predictions varies based the algorithm used and the output being predicted, with the RF emulator being the most consistently accurate emulator used in this study. Computational costs, measured as the time taken to train the MLEs, also varied by algorithm. The MARS emulators were the fastest emulators to be trained in this study, with the RF emulators having the longest training times. These findings will have implications for the choice of algorithm if this approach of utilising MLEs were to be used to improve the time efficiency of running very large numbers of model simulations. Additionally, the robustness of the emulator needs to be tested for each output variable. There is no set of input factors that will be suitable for predicting all outputs in all situations. It is, however, reasonable to assume that it is possible to develop accurate predictive MLEs for any output as all input factors for the process driven simulation model are known, so it should be possible to generate training data sets with all input factors required for the prediction of the target output. A potential disadvantage of the MARS algorithm is that it discards input parameters if they are found to be unimportant during its development. This could limit its usefulness as a generation tool for datasets intended for IA or SA as parameters of low importance within one scope may become more important if the scope is altered by fixing some of the more influential parameters.

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.

Data availability

Data will be made available on request.

Acknowledgements

This work has been supported by the Australian Government via a Research Training Place scholarship and by the Queensland Government via an Advance Queensland PhD scholarship.

Acknowledgment is made to the APSIM Initiative which takes responsibility for quality assurance and a structured innovation programme for APSIM's modelling software, which is provided free for research and development use (see www.apsim.info for details).

The author wishes to thank Dr Allan Peake of CSIRO Toowoomba, for providing the base APSIM-NextGen chickpea simulation configuration and the guidelines for simulating chickpea production in each of the Australian chickpea growing regions. These guidelines and concepts formed the basis of the APSIM experimental design used in this study.

References

Balakrishnan, N., Muthukumarasamy, G., 2016. Crop production-ensemble machine learning model for prediction. Int. J. Comput. Syst. Sci. Eng. 5 (7), 148–153.
Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogrammetry Remote Sens. 114, 24–31.
Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25 (2), 197–227.
Boehmke, B., Greenwell, B.M., 2019. Hands-On Machine Learning with R. CRC Press.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Casadebaig, P., Zheng, B., Chapman, S., Huth, N., Faivre, R., Chenu, K., 2016. Assessment of the potential impacts of wheat plant traits across environments by combining crop modeling and global sensitivity analysis. PLoS One 11 (1), e0146385.
Cravero, A., Sepúlveda, S., 2021. Use and adaptations of machine learning in big data—applications in real cases in agriculture. Electronics 10 (5), 552.
Dahms, T., Seissiger, S., Conrad, C., Borg, E., 2016. Modelling biophysical parameters of maize using LandSat 8 times series. In: *XXIII ISPRS Congress: Proceedings of the XXIII*

ISPRS Congress the International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences, Prague, Czech Republic.
Dayal, K., Weaver, T., Bange, M., CSD Ltd. Extension & Development Team, 2019. Using machine learning to sharpen agronomic insights to improve decision making in Australian cotton systems. In: Pratley, J. (Ed.), 19th Australian Society of Agronomy Conference: Proceedings of the 19th Australian Society of Agronomy Conference (Wagga Wagga New South Wales).
Dumancas, G.G., Bello, G.A., 2015. Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high-performance data mining. In: SC15: the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 41–42.
Fajardo, M., Whelan, B.M., 2021. Within-farm wheat yield forecasting incorporating off-farm information. Precis. Agric. 22 (2), 569–585.
Feng, P., Wang, B., Liu, D.L., Waters, C., Yu, Q., 2019. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. Agric. For. Meteorol. 275, 100–113.
Feng, P., Wang, B., Liu, D.L., Waters, C., Xiao, D., Shi, L., Yu, Q., 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric. For. Meteorol. 285, 107922.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.
Friedman, J.H., Roosen, C.B., 1995. An introduction to multivariate adaptive regression splines. Stat. Methods Med. Res. 4 (3), 197–217.
Gebauer, A., Brito Gómez, V.M., Ließ, M., 2019. Optimisation in machine learning: an application to topsoil organic stocks prediction in a dry forest ecosystem. Geoderma 354, 113846.
Ghimire, S., Deo, R.C., Downs, N.J., Raj, N., 2018. Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities. Remote Sens. Environ. 212, 176–198.
Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Robin Bryant, C., Senthilnath, J., 2021. Integrated phenology and climate in rice yields prediction using machine learning methods. Ecol. Indicat. 120, 106935.
Holzworth, D., Huth, N.I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N.I., Zheng, B., Snow, V., 2018. APSIM Next Generation: overcoming challenges in modernising a farming systems model. Environ. Model. Software 103, 43–51.
Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM – evolution towards a new generation of agricultural systems simulation. Environ. Model. Software 62, 327–350.
Hussein, E.A., Thron, C., Ghaziasgar, M., Bagula, A., Vaccari, M., 2020. Groundwater prediction using machine-learning tools. Algorithms 13, 11.
Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.-M., Gerber, J.S., Reddy, V.R., 2016. Random forests for global and regional crop yield predictions. PLoS One 11 (6), e0156571.
Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18 (3), 235–265.
Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C. H., Rosenzweig, C., Wheeler, T.R., 2016. Brief history of agricultural systems modeling. Agric. Syst. 155, 240–254.
Karandish, F., Šimůnek, J., 2016. A comparison of numerical and machine-learning modeling of soil water content with limited input data. J. Hydrol. 543, 892–909.
Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18 (3), 267–288.
Kouadio, L., Deo, R.C., Byrareddy, V., Adamowski, J.F., Mushtaq, S., Phuong Nguyen, V., 2018. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. Comput. Electron. Agric. 155, 324–338.
Lawes, R.A., Oliver, Y.M., Huth, N.I., 2019. Optimal nitrogen rate can Be predicted using average yield and estimates of soil water and leaf nitrogen with infield experimentation. Agron. J. 111 (3), 1155–1164.
Legates, D.R., McCabe Jr., G.J., 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Leng, G., Hall, J.W., 2020. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environ. Res. Lett. 15 (4), 044027.
Liaw, A., Wiener, M., 2018. Random Forests for Classification and Regression - Breiman and Cutler's Implementation, 4.6-14, R. https://www.stat.berkeley.edu/~breiman/RandomForests/.
Lippmann, R., 1987. An introduction to computing with neural nets. IEEE ASSP Mag. 4 (2), 4–22.
Mandrini, G., Pittelkow, C.M., Archontoulis, S.V., Mieno, T., Martin, N.F., 2021. Understanding differences between static and dynamic nitrogen fertilizer tools using simulation modeling. Agric. Syst. 194, 103275.
Milborrow, S., 2020. Earth: multivariate adaptive regression splines. Derived from mda: mars by trevor hastie and rob tibshirani. Uses alan miller's fortran utilities with thomas lumley's leaps wrapper. R package version 5.3.0. https://CRAN.R-project.org/package=earth.
Nettleton, D.F., Katsantonis, D., Kalaitzidis, A., Sarafijanovic-Djukic, N., Puigdollers, P., Confalonieri, R., 2019. Predicting rice blast disease: machine learning versus process-based models. BMC Bioinf. 20 (1), 514.
Niazian, M., Niedbała, G., 2020. Machine learning for plant breeding and biotechnology. Agriculture 10 (10), 436.
Obsie, E.Y., Qu, H., Drummond, F., 2020. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Comput. Electron. Agric. 178, 105778.
Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016.
Phelan, D.C., Harrison, M.T., McLean, G., Cox, H., Pembleton, K.G., Dean, G.J., Parsons, D., do Amaral Richter, M.E., Pengilley, G., Hinton, S.J., 2018. Advancing a farmer decision support tool for agronomic decisions on rainfed and irrigated wheat cropping in Tasmania. Agric. Syst. 167, 113–124.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 330, 136–161.
R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.
Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J.H.A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., Smith, S., Sheikholeslami, R., Hosseini, N., Asadzadeh, M., Puy, A., Kucherenko, S., Maier, H. R., 2021. The Future of Sensitivity Analysis: an Essential Discipline for Systems Modeling and Policy Support, vol. 137. Environmental Modelling & Software, 104954.
Ryan, E., Wild, O., Voulgarakis, A., Lee, L., 2018. Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output. Geosci. Model Dev. (GMD) 11 (8), 3131–3146.
Samarasinghe, S., 2016. Neural Networks for Applied Sciences and Engineering: from Fundamentals to Complex Pattern Recognition. Crc Press.
Sanikhani, H., Deo, R.C., Yaseen, Z.M., Eray, O., Kisi, O., 2018. Non-tuned data intelligent model for soil temperature estimation: a new approach. Geoderma 330, 52–64.
Sexton, J., Laake, P., 2009. Standard errors for bagged and random forest estimators. Comput. Stat. Data Anal. 53 (3), 801–811.
Sexton, J., Everingham, Y.L., Inman-Bamber, G., 2017. A global sensitivity analysis of cultivar trait parameters in a sugarcane growth model for contrasting production environments in Queensland, Australia. Eur. J. Agron. 88, 96–105.
Shahhosseini, M., Martinez-Feria, R.A., Hu, G., Archontoulis, S.V., 2019. Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. 14 (12), 124026.
Shahhosseini, M., Hu, G., Archontoulis, S.V., Huber, I., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt. Sci. Rep. 11 (1), 1–15.
Shakoor, M.T., Rahman, K., Rayta, S.N., Chakrabarty, A., 2017. Agricultural production output prediction using Supervised Machine Learning techniques. In: 2017 1st International Conference on Next Generation Computing Applications (NextComp), pp. 182–187. http://ieeexplore.ieee.org/document/8016196/.
Shastry, K.A., Sanjay, H.A., Deshmukh, A., 2016. A parameter based customized artificial neural network model for crop yield prediction. J. Artif. Intell. 9, 23–32.
Singh, V., Sarwar, A., Sharma, V., 2017. Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. Int. J. Adv. Res. Comput. Sci. 8 (5), 1254.
Stanfill, B., Mielenz, H., Clifford, D., Thorburn, P., 2015. Simple approach to emulating complex computer models for global sensitivity analysis. Environ. Model. Software 74, 140–155.
Stöckle, C.O., Martin, S.A., Campbell, G.S., 1994. CropSyst, a cropping systems simulation model: water/nitrogen budgets and crop yield. Agric. Syst. 46 (3), 335–359.
Stöckle, C.O., Donatelli, M., Nelson, R., 2003. CropSyst, a cropping systems simulation model. Eur. J. Agron. 18 (3), 289–307.
Venables, W.N., Ripley, B.D., 2002. In: Modern Applied Statistics with S, fourth ed. Springer.
Wallach, D., Thorburn, P.J., 2017. Estimating uncertainty in crop model predictions: current situation and future prospects. Eur. J. Agron. 88, 1–7.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino-McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Milton Bache, S., Müller, K., Ooms, J., Robinson, D., Paige Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Software 4 (43), 1686.
Zhang, J., Chen, Y., Zhang, Z., 2020. A remote sensing-based scheme to improve regional crop model calibration at sub-model component level. Agric. Syst. 181, 102814.
Zhao, G., Bryan, B.A., Song, X., 2014. Sensitivity and uncertainty analysis of the APSIM-wheat model: interactions between cultivar, environmental, and management parameters. Ecol. Model. 279, 1–11.