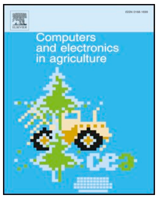
Computers and Electronics in Agriculture 178 (2020) 105778

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms

Efrem Yohannes Obsie[a], Hongchun Qu[b,*], Francis Drummond[c,d]

[a] College of Computer Science, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[c] University of Maine, School of Biology and Ecology, Orono, ME 04469, USA
[d] Cooperative Extension, University of Maine, 5722 Deering, Orono, ME 04469, USA

ARTICLE INFO

Keywords:
Wild blueberry
Yield prediction
Meta-model
Random forest
Boosted decision tree
XGBoost

ABSTRACT

The most challenging task in the agricultural sector is to accurately predict crop yield. A typical machine learning algorithm often uses real data to predict crop yield. In this study, we used data generated by the Wild Blueberry Pollination Model, a spatially explicit simulation model validated by field observation and experimental data collected in Maine USA during the last 30 years. The main aim of this study is to evaluate the relative importance of bee species composition and weather factors in regulating wild blueberry agroecosystems. Specifically, we sought to reveal how bee species composition and weather affect yield and to predict optimal bee species composition and weather conditions that achieve the best yield using computer simulation and machine learning algorithms. Multiple linear regression (MLR), boosted decision trees (BDT), random forest (RF), and extreme gradient boosting (XGBoost) were evaluated as predictive tools. We also performed a predictor selection before submitting our data to the learning algorithms. In this way, we are able to reduce the dimension of the input without a significant drop in prediction accuracy. As a result, clone size, honeybee, bumblebee, *Andrena* bee species, *Osmia* bee species, maximum of upper-temperature ranges, and the number of days with precipitation were chosen as the best predictor variable subset. The results showed that the XGBoost outperformed other algorithms in all measures of model performance for predicting the yield of wild blueberry by achieving a coefficient of determination (R^2) of 0.938, root mean square error (RMSE) of 343.026, mean absolute error (MAE) of 206 and relative root mean square error (RRMSE) of 5.444%. The results are consistent with previous work on predicting wild blueberry fruit yield using digital color photography by (Zaman et al., 2008). This study showed that crop yield predictions can be based on computer simulation modeling datasets. Therefore, if a reasonable prediction can be reached, this study should have a significant impact, especially when data collection in the field is challenging.

1. Introduction

The wild blueberry also known as lowbush blueberry (*Vaccinium angustifolium* Aiton) is the predominant species in a complex of blueberry species cultivated in Maine, which is the largest producer of wild blueberry in the United States, and accounts for about 97% of U.S. wild blueberry total production (Jones et al., 2014; Strik and Yarborough, 2005). The wild blueberry is one of only a few commercially grown crops native to North America. This crop is not planted. It is a complex of five ericaceous plant species that are natural forest understory plants. Forests are harvested and the blueberry plants are then managed to produce fruit (Drummond, 2019a). The productivity of this crop is largely influenced by cross-pollination that requires bees (Asare et al., 2017; Drummond, 2016). The yield of wild blueberry is not always a continuous non-linear relationship to bee density (Asare et al., 2017), it may also be subject to variation of weather, climate warming, soil fertility, pests and disease, and other temporal and spatial abiotic and biotic factors (Tasnim et al., 2020; Aras et al., 1996).

The greatest and the most challenging activities in precision agriculture are accurate predictions of crop yield. Extensive research is underway in agriculture to better predict crop yield using machine learning algorithms (Chlingaryan et al., 2018; Crane-Droesch, 2018; Jeong et al., 2016). Many machine learning algorithms require large amounts of data to provide reliable results (Johnson et al., 2016). One of the major challenges in training and experimenting with machine learning algorithms is the availability of training data in sufficient

[*] Corresponding author.
E-mail address: hcchyu@gmail.com (H. Qu).

https://doi.org/10.1016/j.compag.2020.105778
Received 4 May 2020; Received in revised form 4 September 2020; Accepted 4 September 2020

quality and quantity. For most studies, this remains a limiting factor.

One way to overcome the problem of collecting large training data for machine-learning algorithms is to generate data by using computer simulation modeling techniques. The Naval Postgraduate Schools (NPS) Simulation, Experiments and Efficient Design (SEED) Center for Data Farming defines data farming as the process of using simulations and computer modeling to compile data sets. As such, data farming seeks to provide decision-makers with insights into complex issues by using simulations to produce data (Tolk, 2015). Hence, for this study, we used previously validated simulation-based modeling of wild blueberry pollination (Qu and Drummond, 2018) and designed computational experiments to 'grow' training data, which can then be used to train and validate machine learning algorithms.

The machine learning algorithms trained on a simulation model are called meta-models (Simpson et al., 2001), which possess two obvious advantages over using either machine learning models trained directly on empirical datasets or running computationally intensive simulation models for direct predictions. On one hand, meta-models are surrogates to simulation models which represent detailed causalities of interacting ecological processes that are helpful to precisely predict system behaviors (Drummond et al., 2003; Puntel et al., 2019). Because meta-models can learn patterns of connections among inputs and outputs of the original simulation model and have the ability to extrapolate across varying temporal and spatial scales (Fienen et al., 2015). On the other hand, meta-modeling techniques have been widely accepted by the scientific community in constructing predictive models because in most cases it is more practical to use a real-time meta-model to make predictions than run the often much slower (higher cost) simulation model if the prediction accuracy between the two is within an acceptable range (e.g., less than 20% error) (Fienen et al., 2015).

Simulation data is increasingly being used by researchers and companies for machine learning applications especially in computer vision where a model is trained on a simulation generated dataset (Bohn et al., 2013). Efforts have been made to construct general-purpose simulated data generators to enable data science experiments (Patki et al., 2016). In general, simulated data has several advantages: (1) it is fast and usually inexpensive to produce as much data as needed once the simulation environment is ready; (2) simulation data can have perfectly accurate labels; (3) including labeling that may be very expensive or impossible to obtain by hand, the simulation environment can be modified to improve the model and training; (4) simulation data can be used as a substitute for certain real data segments that contain sensitive information; and (5) it is useful in cases where the generation of training data involves expensive sample acquisition or training data cannot be obtained in sufficient quantity for ethical reasons (Dahmen et al., 2019).

Many researchers have used linear regression models to predict crop yield (Ji et al., 2007; Matsumura et al., 2015; Zaman et al., 2008; Zhang et al., 2019). As far as we know, there is no published study except that by (Shahhosseini et al., 2019) who used simulation data to train machine learning algorithms to predict crop productivity. These authors used the APSIM (Agricultural Production Systems SiMulator; (Holzworth et al., 2014) cropping systems model to generate maize yield and nitrogen loss data for seven locations in the US Midwest. Simulations were conducted for 5–7 years and several management treatments resulting in more than 3 million data points representing maize yields and nitrogen losses.

Our research objective was to develop a predictive model with the assistance of computer simulation and machine learning algorithms. Once we determined the best model to predict yield, our study aimed to address three scientific objectives. First, we conducted simulations and prescreened the simulated data used for developing predictive models of wild blueberry yield. Second, we wanted to determine the important factors that predict yield most effectively. Third, we elucidated, through sensitivity analysis, the optimal bee species composition and weather conditions that predict the best yield estimates compared to the actual simulation derived yields. Furthermore, our goal was to determine the most robust model for yield prediction by comparing traditional and modern machine learning algorithms while at the same time using a minimal number of features.

2. Methodology

2.1. Data generation

The study of predictive models for blueberry yield prediction requires data that sufficiently characterize the influence of plant spatial traits, bee species composition, and weather conditions on production. In a multi-step process, we designed simulation experiments and conducted the runs on the calibrated version of the blueberry simulation model. The simulated dataset was then examined, and critical features were selected to build four machine-learning-based predictive models.

2.1.1. The wild blueberry pollination simulation model

The dataset used for predictive modeling was generated by the Wild Blueberry Pollination Simulation Model (hereafter it is referred to as *the simulation model*), which is an open-source, spatially-explicit computer simulation program (Fig. 1) that enables exploration of how various factors, including plant spatial arrangement, outcrossing (allogamy) and self-pollination (geitonogamy), bee species compositions and weather conditions, in isolation and combination, affect pollination efficiency and yield of the wild blueberry agroecosystem. The simulation model has been validated by the field observation and experimental data collected in Maine USA and Canadian Maritimes during the last 30 years (Qu and Drummond, 2018) and now is a useful tool for hypothesis testing and theory development for wild blueberry pollination researches.

In the simulation model, the spatial and genetic structure of wild blueberry plants and bee foraging behaviors for different species are

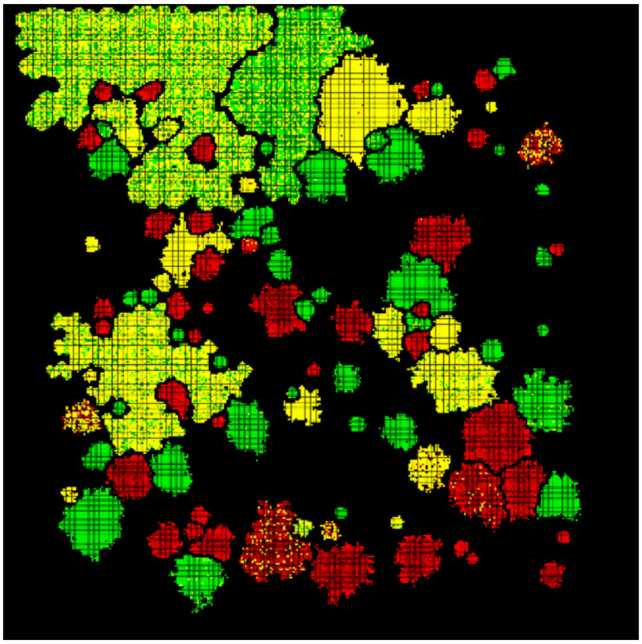


Fig. 1. A simulated wild blueberry field on Julian date 136 of the production season. The green dots are quadrats in which stems are in bud (before bloom) stage, yellow dots are quadrats in which stems are in bloom, red dots are quadrats in which flowers on stems have become fruit (after bloom). Mixed yellow (flower) and green (bud) stem show the pattern of successive waves of flowering within a clone. Red stems with different color saturation indicate the percentage of fruit set, i.e., bright red stems have higher fruit set than the dark red ones. Black areas are bare spots in the field caused by herbicide applications and erosion (Qu and Drummond, 2018).

Table 1
Parameters used to configure the simulation experiments.

Parameter	Unit	Range	Description
Clone size[1]	m^2	10–40	The average blueberry clone size in the field
Honeybee density	bees/m^2/min	0–18.43	Honeybee (*Apis mellifera* (L.)) density in the field
Bumblebee density	bees/m^2/min	0–0.585	Bumblebee (*Bombus* spp.) density in the field
Andrena density	bees/m^2/min	0–0.75	*Andrena* spp. bee density in the field
Osmia density	bees/m^2/min	0–0.75	*Osmia* spp. bee density in the field
Daily air temperature	°F	High, Moderate or Low	The 125%, 100% or 75% of the average daily air temperature from Julian day of 121 and 181 of the past five years (2015–2019)
Daily precipitation	inch	High, Moderate or Low	The 125%, 100% or 75% of the average daily precipitation from Julian day of 121 and 181 of the past five years (2015–2019)

[1] Individual genetically unique blueberry plants are referred to as clones of which approximately 80% require outcrossing during pollination for the fruit to develop (Bell et al., 2010). The larger the clone, the lower the likelihood that a high proportion of flowers will be outcrossed by foraging bees (Qu and Drummond, 2018).

modeled at the individual bee, flower, stem, and clone level. The interactions between individuals are scheduled by blueberry phenology and are linked to environmental conditions. By integrating weather information such as the range of daily air temperature and precipitation during the whole production season, it can simulate and visualize pollination efficiency under varying factor combinations both within a field and at the field level, particularly in the case where logistical, spatial and temporal limitations are encountered in large scale experimentation. As shown in Fig. 1, the simulation model provides a convenient graphic user interface to mimic and visualize the ecological process during a blueberry production season.

2.1.2. Simulation experiments

We used the calibrated version of the simulation model and performed a set of simulation experiments to develop a simulated dataset for machine learning model development and analysis. The simulation experiments aimed to characterize the influence of wild blueberry spatial arrangement, bee species composition in the field, and weather conditions on yield. Therefore, the parameters (i.e., the factors of the simulation model) used to configure the simulation experiments are three-fold (Table 1): (1) the average size of blueberry clones within a field; (2) foraging density of each bee taxon group; and (3) weather information such as temperature, precipitation and wind speed. The range of blueberry clone size has been observed from several to hundreds of square meters, but in most cases, they are smaller than 50 square meters (Drummond, 2016). We set the range of clone sizes between 10 and 40 square meters in the simulation to cover typical scenarios. Native bee density in a blueberry field could be very much different from that of the commercial Honeybee, so we set different ranges of density for different bee taxa (Table 1). As for weather parameters of the simulation model, we collected the maximum and minimum daily air temperature and precipitation from Julian day 121 to 181 between the year 2015 and 2019 of the region of Bangor, Maine of the USA from The Weather Channel (https://weather.com) and averaged the five years data to form the weather input. This weather condition is regarded as the current (or the moderate) climate condition. We then systematically increased or decreased the corresponding daily temperature and precipitation to their 125% and 75% level to create the four climate conditions that are: Warm and Dry, Warm and Wet, Cool and Dry, Cool and Wet, respectively.

Once the factors of the simulation experiments had been determined, we designed experiments and specified the number of levels of each factor for effectively sampling the model space. According to the statistics of field observations (Drummond, 2016), we roughly calculated the levels of (i.e., the number of sampling space within) each factor, which are: 6 levels for clone size; 7 levels for Honeybee density; 10 levels for Bumblebee density; 12 levels for Andrena and Osmia bee density, respectively; and 3 levels for air temperature and precipitation, respectively. If we used the full factorial design and replicated each design entry-point 100 times (Sanchez et al., 2018), the total number of simulation runs would be 54,432,000. This is far beyond any acceptable time cost if one considers that one simulation run for a one-hectare blueberry field takes approximately 2 hours in a blade server with 2X Intel® Xeon® E5-2697 2.7 GHz CPU. Therefore, in this study, we employed the Random Latin Hypercube sampling method (Lucas et al., 2007) to greatly reduce the number of simulations required for our study. However, in addition, we stacked three different RLHs with boundary values to obtain a larger design with better space-filling properties (Sanchez et al., 2018). In this case, finally, we conducted 77,700 simulations in total to achieve both an extensive and intensive sampling effort. The simulation experiments resulted in a dataset consisting of 777 records, each of which is an average of 100 simulation runs conducted with the specific conditions shown in Table 1.

2.1.3. Statistical summary of the simulated dataset

Before predictive model development, an initial investigation of the simulation derived data was conducted to determine distributional patterns described by a statistical summary (Table 2). To build a predictive model, the weather conditions specified in the simulation experiments needed to be converted to features in the simulated dataset. Therefore, in addition to clone size, bee density and yield, we constructed and characterized the weather features of the simulation experiments as follows: MaxOfUpperTRange (MaxUTR), MinOfUpperTRange (MinUTR) and AverageOfUpperTRange (AvUTR), are respectively the highest, lowest record and the average of the upper daily air temperature during the simulated bloom season; MaxOfLowerTRange (MaxLTR), MinOfLowerTRange (MinLTR) and AverageOfLowerTRange (AvLTR) are respectively the highest, lowest record and the average of the lower daily air temperature during the simulated bloom season. RainingDays (RD) is defined as the total number of days during the simulated bloom season, each of which has precipitation larger than zero. AverageRainingDays (AvRD) is the average number of days that rain during the entire simulated bloom season.

2.1.4. Feature selection

In a dataset, there may be features that are not completely relevant or spurious and thus not explanatory of blueberry yield. The contribution of these types of features is often low for predictive modeling compared to the most significant features obtained as a result of feature selection. The purpose of feature selection is to: (1) improve the prediction performance of the predictors; (2) provide faster and more cost-effective predictors; and (3) provide a better understanding of the underlying process that generated the data (Guyon and Elisseeff, 2003). There are various methodologies and techniques that can be used to subset the feature dimensional space and help models perform better and more efficiently such as filters, wrappers, and embedded methods. Filter methods are used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected based on their scores in various statistical tests for their correlation with the outcome variable. Wrapper methods, use a

Table 2
Field spatial traits, bee species composition, and weather variables associated with wild blueberry yield (minimum, maximum, mean, std. deviation, and correlation coefficient r) in the simulated dataset.

Feature (Abbreviation)	N	Min	Max	Mean	Std. Deviation	Yield (r)
Clone size (CS)	777	10.0	40.0	18.768	6.9991	−0.52
Honeybee (HB)	777	0.00	18.43	0.4171	0.97890	0.04
Bumblebee (BB)	777	0.00	0.59	0.2824	0.06634	0.31
Andrena (AD)	777	0.00	0.75	0.4688	0.16105	0.14
Osmia (OS)	777	0.00	0.75	0.5621	0.16912	0.38
MaxOfUpperTRange (MaxUTR)	777	69.7	94.6	82.277	9.1937	−0.19
MinOfUpperTRange (MinUTR)	777	39.0	57.2	49.701	5.5958	−0.18
AverageOfUpperTRange (AvUTR)	777	58.2	79.0	68.723	7.6770	−0.18
MaxOfLowerTRange (MaxLTR)	777	50.2	68.2	59.309	6.6478	−0.19
MinOfLowerTRange (MinLTR)	777	24.3	33.0	28.690	3.2095	−0.18
AverageOfLowerTRange (AvLTR)	777	41.2	55.9	48.613	5.4171	−0.18
RainingDays (RD)	777	1	34	18.31	12.124	−0.54
AverageRainingDays (AvRD)	777	0.06	0.56	0.3200	0.17128	−0.54
Yield	777	1637.70	8969.40	6012.84	1356.95	1

Table 3
Features selected by different techniques. Features in table are abbreviations defined in Table 2.

Features	Feature selection techniques				
	Forward	Backward	VIF	XGBoost	Random Forest
CS	✓	✓	✓	✓	✓
HB	✓	✓	✓	✓	
BB	✓	✓	✓	✓	✓
AD	✓	✓	✓	✓	
OS	✓	✓	✓	✓	✓
MaxUTR	✓	✓		✓	✓
MinUTR	✓	✓	✓		✓
AvUTR					
MaxLTR	✓		✓		
MinLTR		✓			
AvLTR					✓
RD	✓	✓	✓	✓	✓
AvRD	✓	✓	✓		✓
R^2	*0.894*	*0.894*	*0.813*	*0.918*	*0.915*

subset of features and train a model with them. Based on the inferences drawn from the previous model, the algorithm adds or removes features from the subset. Embedded methods perform variable selection during the process of training, accomplished by algorithms that have their built-in feature selection methods. The Variance Inflation Factor (VIF) is a filter method, while Sequential Forward Feature Selection (SFFS), Sequential Backward Elimination Feature Selection (SBEFS) are wrapper methods. Extreme Gradient Boosting based upon *feature_importance* and Random Forest based upon *feature_importance* are embedded methods. The selected features that are implemented in our study using the five feature selection techniques are shown in Table 3.

Table 3 contains the features selected by each of the feature selection techniques and their R^2 values (coefficient of determination). The R^2 value explains the proportion of variance in the dependent variable that is explained by the independent predictor variables. It has been observed that the R^2 for VIF was the lowest with a value of 0.813, SFFS, and SBEFS had an equal value of 0.894, XGBoost *feature_importance*_and Random Forest *feature_importance*_had R^2 values of 0.918 and 0.915, respectively. The forward and backward feature selection resulted in the same features and also the same R^2 value. However, when computational time is considered, forward feature selection performs better than backward feature selection (Gopal and Bhargavi, 2019). The performance of all feature selection algorithms gave very similar results except for the VIF method. Each model was developed with one, two, three, and so on up to the total number of predictors (*n* = *13*) in our dataset and compared (Harteveld et al., 2017). There were significant differences among models run with varying numbers of predictors (Fig. 2). For instance, with Random Forest when the numbers of predictors were *n* = *13, n* = *12, n* = *11*, *n* = *10, n* = *9*, *n* = *8*, *n* = *7,* *n* = *6*, *n* = *5*, *n* = *4*, *n* = *3–2*, and *n* = *1,* model accuracies (expressed as coefficients of determination) were 0.894, 0.910, 0.914, 0.912, 0.913, 0.915, 0.902, 0.903, 0.745, 0.713, 0.684 and 0.350; respectively. Similarly, with XGBoost, with a sequential decrease in the number of predictors, model accuracies dropped from 0.918 to 0.345. In both cases, the decrease or asymptotic leveling off of inaccuracy was generally observed with less than *n* = *7* predictors. From this, we hypothesized that our algorithm needs an optimal number of features less than the maximum number of features to make improved predictions. Therefore, we chose to compare the five feature selection algorithms based on their R^2 value and the number of predictors they use. The feature composition produced by the feature selection method with the highest R^2 is chosen as the best set. Accordingly, the XGBoost algorithm outperformed the others by selecting a seven-feature subset (clone size, Honeybee, Bumblebee, Andrena, Osmia, maximum of upper-temperature ranges, and raining days) with a model accuracy of 0.918, a slight difference from the Random Forest technique. The selected seven features were later applied to train the machine learning algorithms.

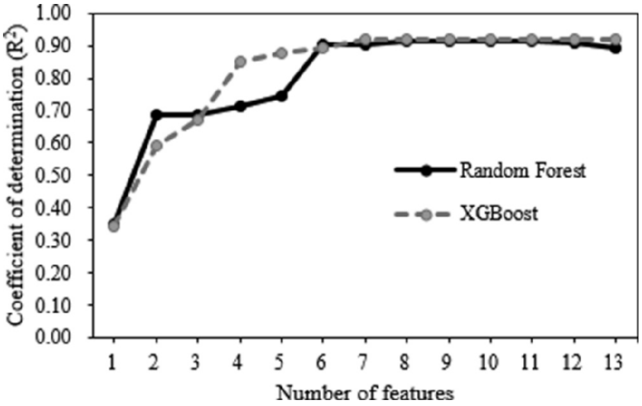


Fig. 2. The number of features used in models (RF and XGBoost) and their associated accuracies (coefficient of determination, R^2).

2.2. Predictive model development

In a multi-step process, we developed predictive models of wild blueberry yield using four machine learning algorithms from the dataset generated by the simulation model. The data used for model development included wild blueberry yield as the dependent variable, bee species composition, and weather conditions data as independent predictor variables. The dataset was split into training data included 621 randomly selected records, comprising 80% of the total dataset, and testing data included 156 records, comprising the remaining 20% using the “traintestsplit” function, part of the *sklearn* package. Each model was trained using the training set and validated with a test set. The four predictive models were developed in a Python IDE (integrated development environment) platform (version 3.7.4 for Windows) using available functions in Python *scikit-learn* which is an efficient tool for predictive modeling (Pedregosa et al., 2011).

2.2.1. Multiple linear regression (MLR)

MLR is a statistical modeling approach that has been used extensively in research to develop predictive models that involve more than one independent variable. It allows a response variable Y to be modeled as a linear function of a multidimensional feature vector. In the present work, clone size, Honeybee, Bumblebee, Andrena, Osmia, maximum upper-temperature ranges, and raining days are the independent variables used in MLR and yield the dependent variable. Multiple linear regression models (or a regression equation) based on several independent (or predictor) variables can be given by the general additive equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_k X_k + \varepsilon \tag{1}$$

where Y = dependent variable, X_k = independent variables, β_k = the coefficients of the independent variables and ε = random error. The coefficients are determined by training the samples.

2.2.2. Boosted decision tree (BDT)

We also used a relatively new machine learning technique, Boosted Decision Tree regression. These models are a combination of two techniques: decision tree algorithms and boosting methods. This method improves the performance of a single model by fitting several models and combining them for prediction. The boosting methods weight the input data in subsequent trees. The weights are applied in such a way that data poorly modeled by previous trees has a higher probability of being selected in the new tree. This means that after each tree is fit, the model takes into account the error in the prediction of the previously fit tree to fit the next tree (Elith et al., 2008). Taking into account fits of previous trees, the algorithm continuously tries to improve model accuracy. Boosted Decision Tree regression is widely used because it can model nonlinear relationships but often requires data transformation and outlier elimination to produce desired results (Dube and Mutanga, 2015; Elith et al., 2008). In our study, a decision tree was boosted using AdaBoost (Drucker, 1997).

2.2.3. Random forests (RF)

The RF algorithm is a type of ensemble method that makes predictions by averaging over predictions of several independent base models. This framework has been extremely successful as a general-purpose classification and regression method since its introduction by Breiman (2001). The RF algorithm makes use of bootstrap aggregating (i.e. bagging), which reduces the variance of a statistical learning model (Friedman et al., 2001). Briefly, many bootstrapped samples of the training data are taken and trees are constructed from this data. For each tree, a majority vote is taken for the predicted class of that tree, and an average prediction is returned. This process greatly increases the overall predictive ability of the model. Moreover, bootstrap aggregating allows for an out-of-bag error estimate which is a valid estimate of the test error (Breiman, 2001). In our study, we trained and applied RF to predict yields of wild blueberry. Since RF can be used for classification and regression purposes. The scope of this study was to use it as a regression tool.

2.2.4. Extreme gradient boosting (XGBoost)

Lately, many agricultural researchers have been using XGBoost to predict crop yields. XGBoost is an open-source library that implements gradient boosted decision trees that are efficient and highly optimized. In gradient tree boosting, models are not trained in isolation of each other, but rather in succession, where each model iteratively reduces errors made by previous models. Instead of assigning different weights to the classifiers after every iteration, this method fits a new model to new residuals from the previous prediction and then minimizes the loss when adding the latest prediction (Chen and Guestrin, 2016). The regular updating of the weights of leaves within a tree ensemble model allows for deriving an optimal model that minimizes the evaluation formula. Predicted values of the tree ensemble model are calculated as follows:

$$\begin{aligned} &\hat{y}_i = \varnothing(x_i) = \textstyle\sum_{k=1}^{n} k = f_k(x_i), f_k \in F\# \\ &F = f(x) = w_{q_{(x)}}(q\colon R^m \to T, w \in R^t) \end{aligned} \tag{2}$$

Eq. (2) represents the regression tree space and x_i is the input, $\hat{y}_i$is the output, q the tree structure, and the number of leaves in the tree are identified as T. Each f_k matches the independent tree structure q and the weight w, where w_i represents the score of the i^{th}leaf. This predicted value can be evaluated by:

$$\begin{aligned} &L(\varnothing) = \textstyle\sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ &\quad where, \Omega(f) = y^T + \tfrac{1}{z}\lambda\|w\|^2 \end{aligned} \tag{3}$$

and L is the loss function being the difference between the predicted value $\hat{y}_i$ and the target value y_i. The entity Ω represents the complexity of the model and is a regularization term that has the function of smoothing the weight to avoid overfitting.

2.2.5. Model evaluation

Four measures were used to evaluate the performances of the predictive models. First, we used the coefficient of determination (R^2), defined as the proportion of the variance in the response variable that is explained by independent variables. Second, we used the root mean squared error (RMSE), a measure of the difference between predicted and actual values. Third, the mean absolute error (MAE)was used and is defined as the absolute mean difference between actual and predicted yield values. Fourth, the relative root mean square error (RRMSE) was used and it is calculated by dividing RMSE by the mean of observed/actual values. The RMSE and MAE were determined using the formulae described by (Elminir and Abdel-Galil, 2006) while RRMSE is determined according to (Li et al., 2013):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(Y_i - X_i)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{\sum_{i=2}^{n} |Y_i - X_i|}{n} \tag{5}$$

where n is the number of observations, Y_i is the actual yield, and X_i is the model-predicted yield of wild blueberry. The coefficient of determination (R^2) was computed according to (Montgomery et al., 2012):

$$R^2 = \sqrt{1 - \frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{6}$$

$$RRMSE = \frac{RMSE}{\bar{Y}} x^{100} \tag{7}$$

where X_i Y_i and n are as defined above, and $\bar{Y}$ is the mean actual yield. An RMSE value of zero indicates that all output yield computed by the model perfectly matches the corresponding actual yield. The closer the R^2 value is to 1.0, the more linear the relationship is between model-predicted and actual yield. Therefore, as the metrics, MAE and RMSE approach zero and R^2 approaches 1.0, predictions of yield increase in accuracy.

2.3. Predictor importance evaluation

Once the predictive models were built and evaluated, the next step was to test the potential of the four predictive models to identify important predictors that influence wild blueberry yield. There are many ways to calculate predictor importance scores of a predictive model and the simplest way is to calculate coefficient statistics between each predictor and the target variable. For example, if a machine learning algorithm finds a set of coefficients to use in the weighted sum to make a prediction, these coefficients can be used directly as a crude type of predictor importance score. In this study, we applied the Python *scikit-learn* library version 0.21.3 (Pedregosa et al., 2011) to calculate the predictor's relative importance of the four predictive models, which were built based on the seven features chosen under Section 2.1.4. The predictor importance property available in this library assigns a score to the input predictors based on how they are important in predicting the target variable in the predictive models; the higher the score, the more important or relevant the predictor is in relation to the output. As for the Multiple Linear Regression technique, a *LinearRegression()* model was fit on the dataset to find the coefficients, then the scores for each input was summarized to provide predictor importance. In the same way, or the Boosted Decision Tree, Random Forest, and Extreme Gradient Boosting techniques, after being fit, the models provided a *feature_importances_* function. This is an inbuilt class that comes with tree-based machine learning algorithms implemented in a *scikit-learn* library that can be used to retrieve and report the coefficient score for each predictor.

2.4. Sensitivity analysis

In a sensitivity analysis, we aimed at evaluating how bee species composition and different weather patterns affect blueberry yield. Specifically, we tested which predictor or combination of predictors (parameters) influences the model output of yield. To evaluate the effects, we developed predictions using the best-trained model XGBoost. A fully crossed factorial experiment was designed among 4 bee species and 4 weather scenarios for a total of 16 scenarios.

To make predictions for the 16 combinations, bee density and weather parameters were adjusted in the following way. For bee species dominance, the foraging density of the species of interest was set to 1.00 bees/m^2 and the other bee taxa were set to densities of 0.25 bees/m^2. As an example, for Honeybee dominance, pollination was performed by Honeybees at a density of 1.00 bees/m^2; while simultaneously Bumblebees, Andrena, and Osmia bees foraged at densities of 0.25 bees/m^2 under varying weather conditions. For the weather condition *warm and wet*, we kept all temperatures and rain factors to the maximum of their ranges (see Table 2). In contrast, for *warm and dry* conditions, all temperature factors were set to the maximum of their range while rain factors set to the minimum of their range. For the *cold and wet* weather conditions, all temperature factors were set to the minimum of their range and all rain factors were set to the maximum of their range. Lastly, for *cold and dry* weather conditions, all temperatures and rain factors were simultaneously set to the minimum of their ranges. In order to evaluate the relative impact on yield due to changes of bee species composition and weather conditions, we compared yield in the 16 combinations to the baseline scenario where all weather parameters were set as the mean of their range and all bee species densities were equally set at 0.25 bees/m^2. The baseline scenario is regarded as the yield expectation produced by an equally balanced bee community and moderate weather conditions under the current climate. Graphic visualization was used to assess the results of the 16 combinations of bee species composition and weather on the range of expected effects on yield.

Table 4
The evaluation metrics: R^2, RMSE, and MAE for the four predictive models: MLR, BDT, RF, and XGBoost.

Model	Metrics			
	R^2	RMSE	MAE	RRMSE
MLR	0.776	661.923	551.863	11.008%
BDT	0.823	489.659	372.772	8.181%
RF	0.902	430.852	221.890	7.115%
XGBoost[1]	**0.938**	**343.026**	**206.239**	**5.444%**

[1] Bold type reflects the best overall model.

3. Results

We carried out several experiments to: (1) evaluate the strength of predictive models comparing traditional and modern machine learning techniques; (2) identify important factors that affect yield most; (3) assess the effects of bee composition and weather conditions on yield; and (4) seek optimal bee composition and weather conditions that achieve the highest predicted yield.

3.1. Model evaluation

Our first experiment illustrates the potential of the models for predicting blueberry yield. A comparison of the models was conducted (for details see Table 4). The models were compared based on the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and relative root mean square error (RRMSE) evaluation metrics. Table 4 reports that the models using modern machine learning techniques BDT, RF, and XGBoost resulted in the lowest RMSE values: 489.659 kg/ha, 430.852 kg/ha, and 343.026 kg/ha; respectively (Fig. 3B-D). These values are less than 10% of the average yield (5897.134 kg/ha). This shows that the three models can predict yield effectively. However, MLR, the model using the traditional machine learning technique was characterized by the highest RMSE, 661.923 kg/ha, which was greater than 10% of the average yield value. This illustrates that the MLR model was not accurate as the other models but still resulted in reasonably good predictions. The MAE score was also the highest for MLR, 551.863 kg/ha. The remaining three models ranged from 206.239 to 372.772 kg/ha. When we compared the models' coefficients of determination (R^2), XGBoost had the highest value, 0.938. The second-best performing model, in terms of R^2, is RF and had an R^2 of 0.902. The MLR and BDT models had R^2's of 0.776 and 0.823, respectively. Thus, among the four predictive models, XGBoost possessed the highest R^2 and lowest RMSE and MAE compared to BDT and RF (Table 4). In terms of RRMSE, XGBoost was better than other models. The model accuracy is considered excellent if the RRMSE is $< 10\%$, good if the RRMSE is $\geq 10\%$ and $< 20\%$, fair if the RRMSE is $\geq 20\%$ and $< 30\%$, and poor if the RRMSE $\geq 30\%$ (Li et al., 2013). Therefore, based on RRMSE; XGBoost, RF, and BDT are classified as excellent while MLR is classified as good. In summary, the XGBoost model outperformed the other three models examined in this study.

Scatter plots of the test yield data versus the predicted yield estimates for the different predictive models are presented in Fig. 3. It was shown that the error for the MLR model built by the traditional machine learning technique (Fig. 3A) was higher than the models built by modern machine learning techniques (Fig. 3B–D). Furthermore, the unexplained variance in yield for the XGBoost model was much less compared to others in almost all test samples ($R^2 = 0.938$).

3.2. Predictor importance

The second experiment involved testing the potential of the four predictive models to identify important factors that influence wild blueberry yield. Each model was developed with the seven features

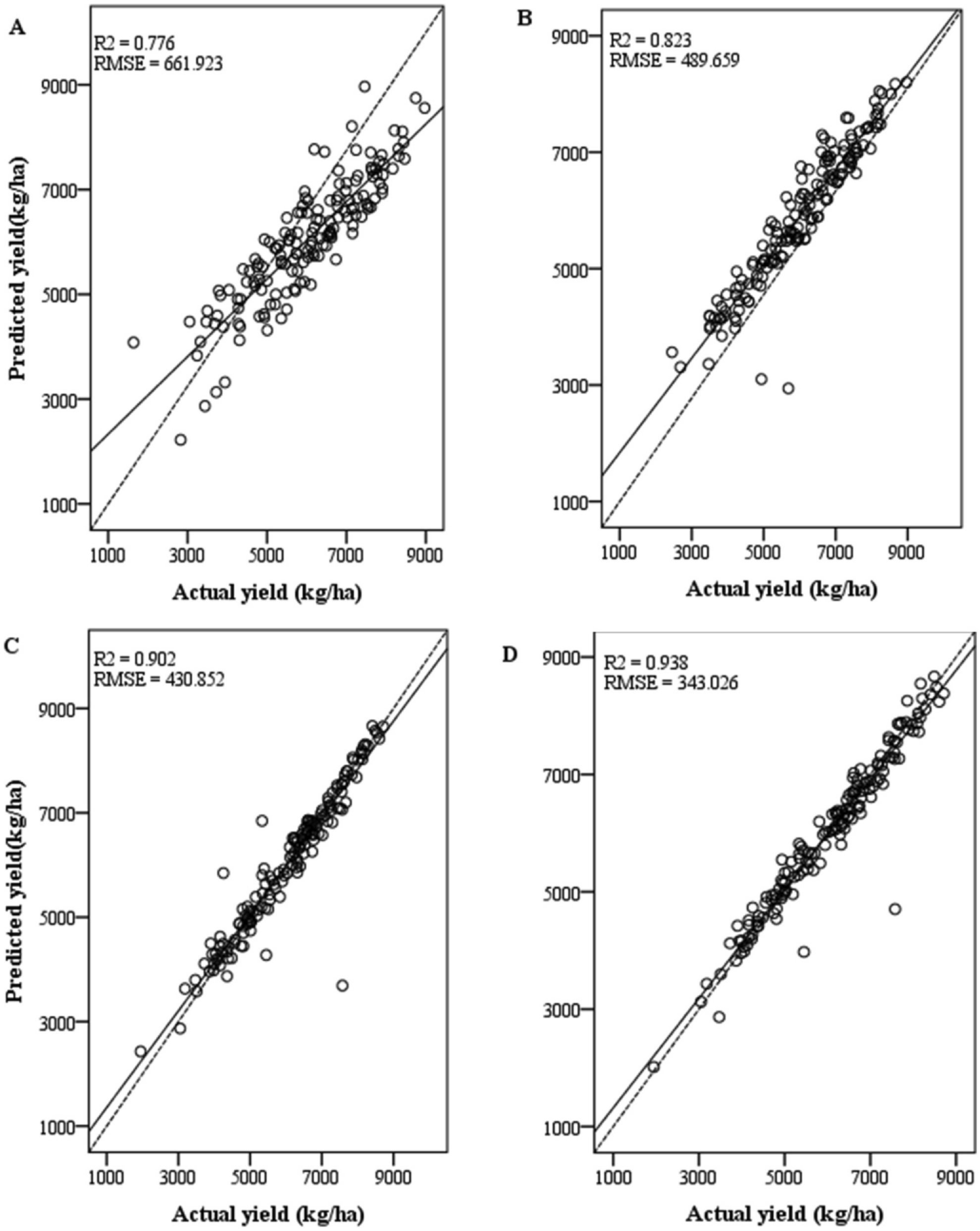


Fig. 3. Model performance for test dataset samples of yield (actual simulated) vs predicted yield from the following models: MLR (A), BDT (B), RF (C) and XGBoost (D).

Table 5
Summary of predictor relative importance[1] by MLR, BDT, RF, and XGBoost models.

Predictor	MLR	BDT	RF	XGBoost
Clone size (CS)	−1	19	22	23
Honeybee (HB)	2	7	2	2
Bumblebee (BB)	71	17	5	10
Andrena (AD)	5	4	2	3
Osmia (OS)	24	21	16	20
MaxOfUpperTRange (MaxUTR)	−0.3	12	18	11
RainingDays (RD)	−0.8	19	34	33

[1] Values in the table are percentages (%).

chosen under Section 2.1.4, *Feature selection.* Table 5 was created to show predictor importance scores for each of the predictive models.

The MLR equation for yield (y) estimation:

$$Yield = 7830.06 - 98.16\ x\ CS + 124.59\ x\ HB + 5836.87\ x\ BB + 411.85\ x\ AD + 2005.38\ x\ OS - 21.97\ x\ MaxUTR - 64.55\ x\ RD$$

The effect of predictors on yield variability has been demonstrated by the MLR coefficients. Although the model is less effective for further use, the predictor Bumblebee (*Bombus* spp.) was found to be the most important factor affecting yield as highlighted in Table 6 with a relative importance of 71%. In addition, Osmia (*Osmia* spp.) and Andrena (*Andrena* spp.) bees were highly significant predictors of yield in descending order. They achieved relative importance values of 24% and 5%. Honeybees (*A. melifera*) at 2% was the least important bee predictor. The negative coefficient was more evident for clone size than maximum upper-temperature ranges and raining days with a value of −1%, −0.3%, and −0.8%; respectively (Table 5). Moreover, one can see that for a per unit increase in Clone size, Maximum upper-temperature ranges, and Raining days, wild blueberry yield decreased by 98.16, 21.97, and 64.55 kg/ha; respectively. On the other hand, per

Table 6
Multiple Linear Regression intercept and coefficient values ($P < 0.05$).

Predictor[1]	Coefficient values[2]	Standard Error	Standardized Coefficient (Beta)	p-value
Intercept	7830.06	270.792	–	< 0.001
Clone size (CS)	−98.16	3.536	0.018	< 0.001
Honeybee (HB)	124.59	24.559	0.019	< 0.001
Bumblebee (BB)	**5836.87**	**403.505**	**0.019**	**< 0.001**
Andrena (AD)	411.85	163.888	0.020	0.012
Osmia (OS)	2005.38	167.412	0.021	< 0.001
MaxOfUpperTRange (MaxUTR)	−21.97	2.643	0.018	< 0.001
RainingDays (RD)	−64.55	2.013	0.018	< 0.001
R^2	*0.776*			
Adjusted R^2	*0.772*			

[1] Letters within parentheses are predictor abbreviations depicted in the MLR equation for yield.
[2] Bold value denotes the most important predictor.

unit increases in bee species densities of Honeybees, Bumblebees, Andrena, and Osmia, increased wild blueberry yield by 124.59, 5836.87, 411.85, and 2005.38 kg/ha; respectively (Table 6).

Using the Boosted Decision Tree (BDT) model, analysis of predictor importance was performed to evaluate the relative importance of each predictor contribution to wild blueberry yield. Accordingly, Osmia, Clone size, Raining days, Bumblebees, and Maximum upper-temperature ranges were identified as the most significant predictors affecting yield in the BDT model and their relative predictor importance to the prediction of yield were 21%, 19%,19%, 17%, and 12%; respectively (Table 5). The contribution of Honeybee and Andrena is lower compared to other predictors with a relative importance value of 7% and 4%.

Important predictors were also selected by applying the Random Forest (RF) model. The selected predictors contribute from 2% to 34% of the variability in yield. The most relevant predictors in the RF model were: Raining days, Clone size, Maximum upper-temperature ranges, and Osmia, in descending order, had relative importance values > 10% (Table 5). The impact of the remaining three predictors was less (< 10%) with Bumblebees, 5%; and Honeybees, and Andrena both with a contribution of 2%.

Finally, after applying the XGBoost modeling *feature_importance* function; Raining days, Clone size, Osmia, Maximum upper-temperature ranges, and Bumblebees were chosen as the most significant predictors with relative importance values of 33%, 23%, 20%, 11%, and 10%; respectively (Table 5). The remaining predictors contributed only 3% (Andrena) and 2% (Honeybees) to the prediction of yield.

The discrepancy in the ranking of predictor importance identified by the MLR model compared with the machine-learning models could be due to the difference in algorithms, where the machine-learning algorithms are more advanced in their ability to fit nonlinear data in comparison to MLR (Harteveld et al., 2017).

3.3. Sensitivity analysis

Results of the sensitivity analysis provided insight into how bee density and weather conditions influence blueberry yield. Fig. 4 shows that yield increased with an increase in the density of Honeybees, Bumblebees, Andrena bees and Osmia bees under cool dry and warm dry conditions. In contrast, yield was decreased with an increase in the density of Honeybees, Bumblebees, Andrena bees and Osmia bees under cool or warm wet conditions (Fig. 4A). Honeybees did not produce as high a yield as Osmia bees, Bumblebees, and Andrena bees in decreasing order (Fig. 4B). The overall results summarized in Fig. 4 suggest that bee composition and weather are equally important in the yield prediction and explain much of the variation in yield.

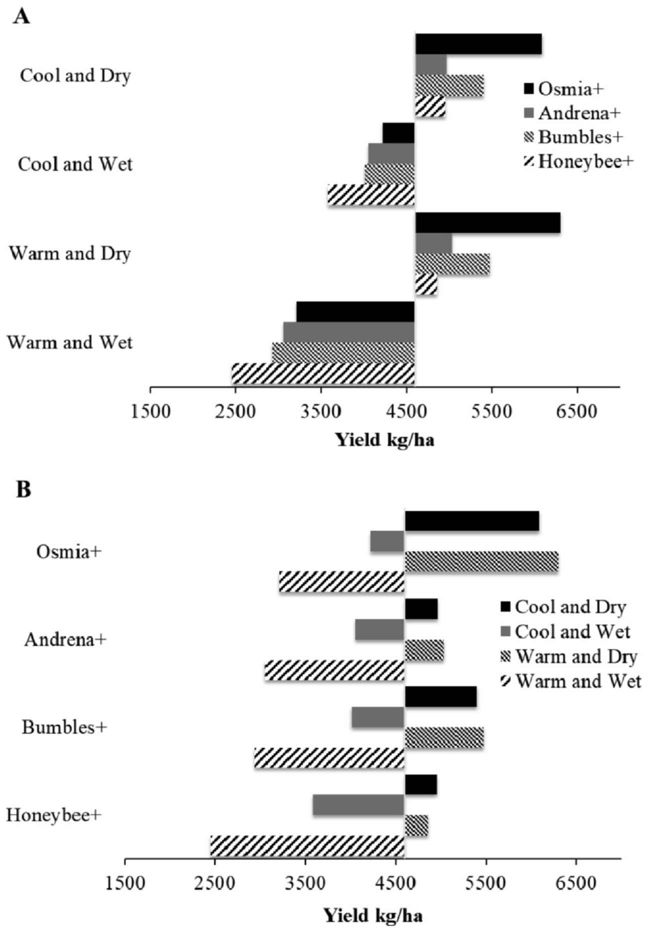


Fig. 4. Sensitivity analysis showing the bee community dominance and weather conditions interaction effect on wild blueberry yield. Panel A is grouped by the four weather conditions for each dominant bee community and panel B is grouped by bee species community for each weather condition. The baseline condition yield for the graphs is 4603.64 kg/ha. The bar in the graphs indicates the values of the predicted yield.

Quantitatively, bee species composition is one of the important driving forces to affect blueberry production as shown in Fig. 4A. Our predictions showed that the highest efficiency bee composition, i.e., Osmia dominance results in up to 37% increase in yield, almost 20% higher than the next most efficient bee group, Bumblebees. Bumblebee dominance was followed by Andrena with a 9% increase in yield. Honeybee dominance resulted in a 6% increase in yield compared to a balanced bee species community. When weather conditions were wet, Honeybee dominance posed the strongest negative impact on yield, 2455.25 kg/ha. This yield level was 53% of the baseline scenario (a balanced bee species community with average weather conditions).

Our sensitivity analysis also showed that weather conditions (especially precipitation rate) had strong impacts on yield (Fig. 4B). In contrast with the baseline situation, dry weather contributed to yield increases of up to 37% where the bee composition was dominated by native Osmia bees. However, dry weather did not achieve high yield increases in association with other bee species dominances. Bumblebee, Andrena, and Honeybee dominance only increased blueberry yield compared to the baseline by 19%, 9%, and 5%; respectively. The predictive model (XGBoost) also predicted that cool weather in dry conditions neutralized these positive effects. However, wet warm weather had a negative impact on yield. Surprisingly, Honeybee dominance in wet weather conditions decreased predicted yield by almost half (47%) compared to the baseline situation. Under these same wet warm conditions, Bumblebee, Andrena, and Osmia did not result in high losses but still resulted in yield declines of 36%, 34%, and 30%; respectively. These negative impacts of wet weather on yield can be mitigated when the weather became cool. As shown in Fig. 4B, yield losses in wet cool weather conditions were reduced by 50% when air temperatures dropped from warm to cool when the bee composition was held constant.

4. Discussion

4.1. Model development and interpretation

This research has investigated an integrative method for crop yield prediction, which overcomes the limitations of approaches using either empirical modeling or computer simulation. Empirical modeling enables quick prediction. However, when the predicted values of important predictors go beyond the range that was used for model calibration, the validity of future extrapolation is questionable (Soltani et al., 2016). In contrast with empirical models, computer simulations which are based on mechanistic relationships, have been useful tools that help researchers to reliably predict future behaviors of a variety of ecological and agricultural complex systems (Qu et al., 2013, 2017; Qu and Drummond, 2018) in the context of climate change. Nevertheless, effectively using simulation models to analyze the target system and make an accurate prediction require a thorough exploration of parameter space, which is both theoretically and practically difficult due to the “curse” of high dimensionality (Sanchez et al., 2018). Alternatively, meta-models, i.e., *models of the model*, are deterministic proxies to the understanding of input–output relationships of the stochastic simulation model. Developing a meta-model involves running the complex simulation model for many combinations and then using the information contained within these “runs” to create a simpler statistical model that is an acceptable approximation of the simulation model. Once the meta-model has been developed, it can then be evaluated and make a prediction very quickly for different input values. Meta-modeling is not a novel concept in the simulation community (Sanchez et al., 2018), however, only a few agricultural research studies (Shahhosseini et al., 2019; Soltani et al., 2016) have taken advantage of its potential and flexibilities. This suggests that it would be valuable to shed light into sufficient investigation and communication of the detailed procedures of meta-modeling for crop yield prediction, which are, but not limited to simulation experimental design, feature engineering, important factor identification, predictive model construction, as well as, model evaluation and sensitivity analysis.

Wild blueberry yield is determined by a multitude of factors. Bell et al. (2010), Yarborough et al. (2017), Drummond (2019a), and Drummond (2019b) have previously shown that density and species diversity of bee pollinators, insect pests, weeds, plant disease, soil fertility, pruning practice, clone genetic composition, and spatial pattern, weather factors such as winter damage, spring frosts, and summer drought; all affect yield. Yield has increased in a near-linear manner over the past five decades due to advancements in production practices, but this increase was just as much due to increased bee density in fields during bloom by supplementation of native bees with the commercial honeybee. The average yield obtained from field surveys is now approximately 5000 kgs/ha in Maine, USA. Asare et al. (2017) showed that in almost 25 years of field survey, 67.3% of the variation in yield was due to fruit set, a direct result of bee diversity and density, and production intensity. We focused on this one very important driver of yield, pollination. Our results demonstrate that simulation modeling in combination with modern machine-learning modeling approaches can produce predictive models with high accuracy of estimated crop yield compared with the traditional MLR model. This has also been demonstrated in other studies when comparing traditional and modern machine learning methods (Chakraborty et al., 2004; De Wolf and Francl, 2000; Mehra et al., 2016) and ensures implementation of such methods to model nonlinear complex data.

Model development using the four different machine learning algorithms resulted in not only different levels of prediction accuracy, but also different weightings of the significant predictors within and between models. The three most important predictors in the four models provide a good example. The top three predictors (in descending order) in the MLR model were: Bumblebees, Osmia, and Andrena; for the RF model: RainingDays, Clone size, and MaxOfUpperTRange; for the BDT model: Osmia, RainingDays and Clone size (tied), and Bumblebees; and for the XGBoost model: RainingDays, Clone size, and Osmia. This suggests that each algorithm explains feature weights differently through different mathematical techniques employed interacting with the specific training data structure used. For example, the RF model showed different ranks of predictor importance and performed better than MLR and BDT. This is likely due to its ability to reduce the correlation between trees by selecting a random subset of the input variables in the tree-building process (Breiman, 2001). This capability mitigates bias by having a single predictor that is more important in each iteration of the model-building process, something the other techniques lack. Furthermore, the disparity of handling non-linearly correlated data, robustness to noise, and being fast and scalable among these machine learning techniques also leads to the variability of ranking predictors. One should exercise caution in providing biological or agricultural production interpretation to the order of significant predictors. A more holistic interpretation of the different ordering of importance of predictors is that across all the different machine learning models: the bee community and weather conditions are the significant subsets of features that provide accurate predictions. This is borne out by our sensitivity analysis which is discussed below.

4.2. Insights gained from model prediction

We found in our modeling approach that prediction of yield is extremely complex. This resulted in different predictors for the different models and evidence of non-linearity when we performed a sensitivity analysis. One aspect that is universal across the predictive models is that honeybees while significant predictors in some models, they are never ranked as one of the top three predictors. This reflects the current knowledge regarding honeybee pollination of wild blueberry. Drummond (2016) showed that in controlled field cage experiments honeybees were 6–8 times less efficient at placing wild blueberry pollen tetrads on blueberry floral stigmas on a per flower visit basis than the bumblebee species, *Bombus impatiens,* and the mason bee, *Osmia atriventris* and 5 times less efficient than the Adrena bee, *Andrena carlini* bees. In addition, he also showed that honey bees took 1.8–3.5 times longer to handle wild blueberry flowers than these other bee species (14 s/flower for honey bee compared to 6–8 s per flower for the other bees). Field studies conducted by (Asare et al., 2017) showed that native bees (Bumblebees, Osmia, and Andrena) were 1.6 times more efficient on a per bee basis in the field than honeybees at pollinating wild blueberry flowers to produce fruit. This dynamic with honeybees is also reflected in our computer simulation model (Qu and Drummond, 2018). Honeybees, on a per bee basis, do not produce as high a yield as *Osmia* spp. bees, bumblebees, and *Andrena* spp. bees. However, it must be stated that honeybees are one of the single most expensive input practices in wild blueberry production. This reflects that it is not the per bee efficiency of pollination that is of consequence to farmers. Farmers rent honeybees for their wild blueberry fields during pollination. They can deploy hundreds of thousands of foraging honeybees per ha and high population levels of honeybees can result in adequate yields.

Another aspect of predicting yield that was characteristic of all the machine learning models is that plant or clone size (area) is extremely important. The impact is negative, meaning that as clone sizes increase in fields, yields decrease. However, clone sizes are highly variable in a field and random in distribution. Because fields are not planted but the result of clearcutting a forest (Jones et al., 2014), farmers have no control over this field feature. This relationship of clone size is present in the simulation model (Qu and Drummond, 2018), but it is an emergent property (Gilbert, 1996). The dynamics that give rise to this property are: (1) pollination efficiency of individual bee species (discussed above), (2) continuous angular and linear movement combined with jump dispersal of spatial bee foraging patterns within a blueberry field as described by (Drummond, 2016) and (Rowland et al., 2019); (3) differential outcrossing success depending upon clone genetics of sire and recipient (Bell et al., 2010); and (4) clone size distribution in a wild blueberry field (Bajcz et al., 2017). It is hypothesized that bees lose pollen picked up at one clone when visiting other clones at a rate proportional to the number of flower visits in subsequent clones (Qu and Drummond, 2018). Thus, as a bee visits a flower from outside a focal clone and picks up pollen and then flies into the focal clone, outcrossing (depositing pollen of one genetically distinct clone onto the stigma of another genetically distinct clone) declines as more and more of the pollen on the visiting bee in the focal clone becomes that of the focal clone resulting in self-pollination. This results in poorer and poorer fruit set unless the clone is self-compatible with its own pollen, a likelihood of only 20% distributed to a greater extent in the early part of the bloom season (Bell et al., 2010; Qu and Drummond, 2018; Drummond and Rowland, 2020). This emergent property of the Clone size feature would be difficult if not impossible to collect data in the field for predicting yield since it integrates bee species pollination efficiency, bee pollen foraging movement patterns, clone self-compatibility status, and clone size.

Finally, our sensitivity analysis showed that predictions of yield are dependent upon the interaction between the dominance of the bee species bee community and weather conditions as defined by four general spring or bloom period weather conditions. This interaction is non-linear and dependent upon the complex interaction of bee species response to air temperature and precipitation (Qu and Drummond, 2018), although wind is also an important factor that affects bee foraging (Drummond, 2016). Our sensitivity analysis showed that bees resulted in the most yield in dry warm weather. This is not expected since observations in the field have shown that bees do not forage to any measurable extent during rain (Drummond et al., 2017). Native bees forage at cooler air temperatures compared to honeybees (Qu and Drummond, 2018). Our sensitivity analysis reflects this, showing the lowest yields during cool wet weather for honeybees (Fig. 4B). Warm wet weather resulted in the worst predictions of yield. This may be due to the fact that during warm weather, even if wet, bloom progresses quickly in the field and so the duration of days that bees have to pollinate the crop, the days with no rain, are few, resulting in fewer flower visits and thus less fruit set and yield compared to dry conditions. This is shown to be a particularly bad situation for honeybees (Fig. 4B) because, within the reduced number of pollination days after rainy weather, honey bees being highly inefficient bees (on a per bee basis) can't successfully set flowers to fruit because each floral visit by a honey bee results in a low number of pollen grains deposited on floral stigmas. An average of 12 pollen grains per stigma results in a 50% chance of a fruit developing from a flower (Drummond, 2019b). Three to four pollen grains on a stigma (honeybee average deposition) results in only a 10% likelihood of a fruit developing (Drummond, 2019b). The anomaly appears to be the bumblebee. They were the second-best performer in dry weather conditions. However, they were the second-worst performer (producing 2935.22 kg/ha, 64% of the baseline scenario productivity of 4586.25 kg/ha) in the wet warm weather conditions. This was unexpected due to bumblebees being the most efficient pollinators in wild blueberry (on a per flower visit, Drummond, 2016), but is most likely due to the same phenomenon as was previously discussed with the honeybee, rapid progression of bloom during warm weather conditions that reduce the number of days for pollination. But, in addition, exacerbating this situation is the pollen foraging pattern of bumblebees. They forage in wild blueberry fields in a highly reticulated manner making random turns when moving from stem to stem (Drummond, 2016) and they visit many flowers upon a stem (up to 120 flowers on a stem (Drummond, 2019b)). This pattern of pollen foraging is characteristic of bumblebee optimal foraging leading to minimizing energy output while maximizing food collection (Hodges, 1981). For the blueberry plant, this pattern of foraging leads to poor pollination as it results in high numbers of consecutive flowers visited within a clone. This reduces outcrossing and except for self-compatible clones, results in incompatible "self" pollen received on floral stigmas. However, countering this hyper-foraging pattern within clones is jump dispersal. Bumblebees forage within clones and then periodically move to new locations (new clones). This foraging behavior is incorporated in our simulation model (Qu and Drummond, 2018) and has been documented to result in higher levels of pollination within large clones (Rowland et al., 2019). So, why the low pollination and low yields by bumblebees with jump dispersal countering high levels of within clone foraging? The reduced pollination period by wet weather limits not only the overall total flower visitation and resulting pollination, but also the frequency of jump dispersal from clone to clone which enhances outcrossing.

Lower yields as a result of wet weather impacts on bee foraging not only is important to understand as a basis for predicting yields in this unique native plant agricultural system. It also has ramifications in assessing the long-term implications of climate change. Most of the wild blueberry production in Maine and Maritime Canada occurs along the North Atlantic coast. Drummond et al. (2017) have documented the decline in the number of pollinator days since the early 1990s. Using a simulation model of bloom phenology and historical weather data since 1960, they showed that the linear decline in good pollination days for bees has been due to increasingly wet springs. Currently, in Maine, the number of pollination days is only half of what was experienced from 1960 to 1990. Our yield prediction model demonstrates that wet rainy springs will greatly reduce yield for farmers in the future, if the trend continues.

5. Conclusions

We investigated four machine learning algorithms: multiple linear regression, boosted decision tree, random forest, and extreme gradient boosting algorithms to develop predictive models for wild blueberry yield. The input dataset was developed from a simulation model of wild blueberry pollination. This model takes advantage of field data accumulated over 30 years of wild blueberry pollination research in Maine and the Canadian Maritimes (Qu and Drummond, 2018). The performances of the predictive models were relatively high but varied (R^2: 0.776 – 0.938). Extreme Gradient Boosting achieved the highest prediction accuracy, Random Forest was the second best-performing model followed by the Boosted Decision Tree model. The performance of the multiple linear regression model was much less accurate compared to the other models investigated in this study. It is also clear from our study that feature selection plays an important role in the development of crop yield prediction models. Therefore, when all features are provided as input to the model without performing feature selection, the accuracy of the model is reduced to a predictive R^2 of 0.813. The sensitivity analysis indicates that the interaction among bee density, species composition, and climatic conditions is significant for determining the variability in blueberry production. In general, our study demonstrated that crop yields can be predicted effectively by using data generated with a validated simulation model. Thus, modeling of agricultural production systems is of paramount importance, especially when the collection of temporal and spatial large-scale datasets is extremely costly, difficult, or in some cases not practically feasible.

CRediT authorship contribution statement

Efrem Yohannes Obsie: Methodology, Software, Formal analysis, Writing - original draft. **Hongchun Qu:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review

& editing, Supervision, Project administration, Funding acquisition. **Francis Drummond:** Conceptualization, Investigation, Supervision, Resources, Formal analysis, 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.

Acknowledgments

We acknowledge funding by the National Natural Science Foundation of China (61871061) and Program of Basic Research and Frontier Technology of Chongqing Municipal Science Commission (cstc2017jcyjAX0453) to HQ.

References

Aras, P., De Oliveira, D., Savoie, L., 1996. Effect of a honey bee (Hymenoptera: Apidae) gradient on the pollination and yield of lowbush blueberry. J. Econ. Entomol. 89, 1080–1083.

Asare, E., Hoshide, A.K., Drummond, F.A., Criner, G.K., Chen, X., 2017. Economic risk of bee pollination in Maine wild blueberry, Vaccinium angustifolium. J. Econ. Entomol. 110, 1980–1992.

Bajcz, A.W., Hiebeler, D., Drummond, F.A., 2017. Grid-Set-Match, an agent-based simulation model, predicts fruit set for the lowbush blueberry (Vaccinium angustifolium) agroecosystem. Ecol. Modell. 361, 80–94.

Bell, D.J., Rowland, L.J., Stommel, J., Drummond, F.A., 2010. Yield variation among clones of lowbush blueberry as a function of genetic similarity and self-compatibility. J. Am. Soc. Hortic. Sci. 135, 259–270.

Bohn, B., Garcke, J., Iza-Teran, R., Paprotny, A., Peherstorfer, B., Schepsmeier, U., Thole, C.-A., 2013. Analysis of car crash simulation data with nonlinear machine learning methods. Procedia Comput. Sci. 18, 621–630.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Chakraborty, S., Ghosh, R., Ghosh, M., Fernandes, C.D., Charchar, M.J., Kelemu, S., 2004. Weather-based prediction of anthracnose severity using artificial neural network models. Plant Pathol. 53, 375–386.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

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

Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13, 114003.

Dahmen, T., Trampert, P., Boughorbel, F., Sprenger, J., Klusch, M., Fischer, K., Kübel, C., Slusallek, P., 2019. Digital reality: a model-based approach to supervised learning from synthetic data. AI Perspect. 1, 2. https://doi.org/10.1186/s42467-019-0002-0.

De Wolf, E.D., Francl, L.J., 2000. Neural network classification of tan spot and Stagonospora blotch infection periods in a wheat field environment. Phytopathology 90, 108–113.

Drucker, H., 1997. Improving regressors using boosting techniques. ICML. 107–115.

Drummond, F., 2019a. Reproductive biology of wild blueberry (Vaccinium angustifolium Aiton). Agriculture 9, 69–80.

Drummond, F., 2019b. Factors That Affect Yield in Wild Blueberry, (Vaccinium Angustifolium Aiton). Agric. Res. Technol. Open Access J. 5, 556212. https://doi.org/10.19080/ARTOAJ.2019.22.556212.

Drummond, F.A., 2016. Behavior of bees associated with the wild blueberry agro-ecosystem in the USA. Int. J. Entomol. Nematol. 2, 27–41.

Drummond, F.A., Dibble, A.C., Stubbs, C., Bushmann, S.L., Ascher, J.S., Ryan, J., 2017. A natural history of change in native bees associated with lowbush blueberry in Maine. Northeast. Nat. 24, 49–68.

Drummond, F.A., Rowland, L.J., 2020. The ecology of autogamy in wild blueberry (*Vaccinium angustifolium* Aiton): Does the early clone get the bee? Agron (in press).

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and neural methods for site–specific yield prediction. Trans. ASAE 46, 5.

Dube, T., Mutanga, O., 2015. Evaluating the utility of the medium-spatial resolution Landsat 8 multispectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa. ISPRS J. Photogramm. Remote Sens. 101, 36–46.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.

Elminir, H.K., Abdel-Galil, H., 2006. Estimation of air pollutant concentrations from meteorological parameters using artificial neural network. J. Electr. Eng. 57, 105–110.

Fienen, M.N., Nolan, B.T., Feinstein, D.T., Starn, J.J., 2015. Metamodels to bridge the gap between modeling and decision support.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Springer series in statistics New York.

Gilbert, G.N., 1996. Holism, individualism and emergent properties. In: Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Springer, pp. 1–12.

Gopal, P.S.M., Bhargavi, R., 2019. Optimum Feature Subset for Optimizing Crop Yield Prediction Using Filter and Wrapper Approaches. Appl. Eng. Agric. 35, 9–14.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Harteveld, D.O.C., Grant, M.R., Pscheidt, J.W., Peever, T.L., 2017. Predicting Ascospore release of Monilinia vaccinii-corymbosi of blueberry with machine learning. Phytopathology 107, 1364–1371.

Hodges, C.M., 1981. Optimal foraging in bumblebees: hunting by expectation. Anim. Behav. 29, 1166–1171.

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., 2014. APSIM–evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350.

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.

Ji, B., Sun, Y., Yang, S., Wan, J., 2007. Artificial neural networks for rice yield prediction in mountainous regions. J. Agric. Sci. 145, 249–261.

Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. 218, 74–84.

Jones, M.S., Vanhanen, H., Peltola, R., Drummond, F., 2014. A global review of arthropod-mediated ecosystem-services in Vaccinium berry agroecosystems. Terr. Arthropod Rev. 7, 41–78.

Li, M.-F., Tang, X.-P., Wu, W., Liu, H.-B., 2013. General models for estimating daily global solar radiation for different solar radiation zones in mainland China. Energy Convers. Manag. 70, 139–148.

Lucas, T.W., Sanchez, S.M., Sickinger, L.R., Martinez, F., Roginski, J.W., 2007. Defense and homeland security applications of multi-agent simulations. In: 2007 Winter Simulation Conference. IEEE, pp. 138–149.

Matsumura, K., Gaitan, C.F., Sugimoto, K., Cannon, A.J., Hsieh, W.W., 2015. Maize yield forecasting by linear regression and artificial neural networks in Jilin. China. J. Agric. Sci. 153, 399–410.

Mehra, L.K., Cowger, C., Gross, K., Ojiambo, P.S., 2016. Predicting pre-planting risk of Stagonospora nodorum blotch in winter wheat using machine learning models. Front. Plant Sci. 7, 390.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to linear regression analysis. John Wiley & Sons.

Patki, N., Wedge, R., Veeramachaneni, K., 2016. The synthetic data vault. In: in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 399–410.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Puntel, L.A., Pagani, A., Archontoulis, S.V., 2019. Development of a nitrogen recommendation tool for corn considering static and dynamic variables. Eur. J. Agron. 105, 189–199.

Qu, H., Drummond, F., 2018. Simulation-based modeling of wild blueberry pollination. Comput. Electron. Agric. 144, 94–101.

Qu, H., Seifan, T., Seifan, M., 2017. Effects of plant and pollinator traits on the maintenance of a food deceptive species within a plant community. Oikos 126, 1815–1826.

Qu, H., Seifan, T., Tielbörger, K., Seifan, M., 2013. A spatially explicit agent-based simulation platform for investigating effects of shared pollination service on ecological communities. Simul. Model. Pract. Theory 37, 107–124.

Rowland, L.J., Ogden, E.L., Bell, D.J., Drummond, F.A., 2019. Pollen-mediated gene flow in managed fields of lowbush blueberry. Can. J. Plant Sci. 100, 95–102.

Sanchez, S.M., Sánchez, P.J., Wan, H., 2018. Work smarter, not harder: a tutorial on designing and conducting simulation experiments. In: 2018 Winter Simulation Conference (WSC). IEEE, pp. 237–251.

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, 124026.

Simpson, T.W., Poplinski, J.D., Koch, P.N., Allen, J.K., 2001. Metamodels for computer-based engineering design: survey and recommendations. Eng. Comput. 17, 129–150.

Soltani, A., Bakker, M., Veldkamp, A., Stoorvogel, J., 2016. Comparison of three modelling approaches to simulate regional crop yield: a case study of winter wheat yield in Western Germany. J. Agric. Sci. Technol. 18, 191–206.

Strik, B.C., Yarborough, D., 2005. Blueberry production trends in North America, 1992 to 2003, and predictions for growth. Horttechnology 15, 391–398.

Tasnim, R., Calderwood, L., Annis, S., Drummond, F.A., Zhang, Y. J., 2020. The future of wild blueberries: Testing warming impacts using open-top chambers. Spire.

Tolk, A., 2015. The next generation of modeling & simulation: integrating big data and deep learning. In: Proceedings of the Conference on Summer Computer Simulation, pp. 1–8.

Yarborough, D., Drummond, F., Annis, S., D¿ Appollonio, J., 2017. Maine wild blueberry systems analysis. In: XI International Vaccinium Symposium 1180, pp. 151–160.

Zaman, Q.U., Schumann, A.W., Percival, D.C., Gordon, R.J., 2008. Estimation of wild blueberry fruit yield using digital color photography. Trans. ASABE 51, 1539–1544.

Zhang, L., Traore, S., Ge, J., Li, Y., Wang, S., Zhu, G., Cui, Y., Fipps, G., 2019. Using boosted tree regression and artificial neural networks to forecast upland rice yield under climate change in Sahel. Comput. Electron. Agric. 166, 105031.