RESEARCH ARTICLE

Plants People Planet PPP
Open Access

A new framework for predicting and understanding flowering time for crop breeding

Chetan Deva[1] | Laura Dixon[1] | Milan Urban[2] | Julian Ramirez-Villegas[2,3] |
Ioannis Droutsas[1,4] | Andrew Challinor[1]

[1]University of Leeds, Leeds, UK

[2]The International Center for Tropical Agriculture (CIAT), Cali, Colombia

[3]Wageningen University, Wageningen, The Netherlands

[4]Department of Plant Sciences, University of California, Davis, Davis, CA, USA

Correspondence
Chetan Deva, University of Leeds, Leeds, UK.
Email: c.r.deva@leeds.ac.uk

Societal Impact Statement

As the growing season changes, the development of climate resilient crop varieties has emerged as a crucial adaptation in agricultural systems. Breeding new varieties for a changing climate requires enhanced capacity to predict the complex interactions between genotype and environment that determine flowering time. Hundreds of experiments with observations of flowering, the environment and plant genetics were used to build a model that can predict when a variety of common bean is going to flower. This model will help breeders to explore the phenological characteristics of their germplasm, speeding up selection for climate adaptation.

Summary

- There is an urgent need to accelerate crop breeding for adaptation to a changing climate. As the growing season changes, crop improvement programmes must ensure that the phenological characteristics of the varieties they develop remain well suited to their target population of environments.

- Meeting this challenge will require a clear understanding of how existing germplasm behave across Genotype ∗ Environment (G ∗ E) to enhance the efficiency of selection. Recent work calls for the development of simple models that can accurately simulate genotypic variation in key traits across target population of environments.

- Accordingly, we develop a simple machine learning framework for modelling time to flowering across G ∗ E and apply this to common bean in an equatorial target population of environments. Within this framework, we test three machine learning models and find that the best performing models display high levels of accuracy across G ∗ E.

- We advance understanding of the environmental drivers of flowering time in equatorial conditions by showing that thermal time and accumulated evaporation are powerful predictors of flowering time across all three models.

KEYWORDS
ambient temperature, beans, common bean, flowering, machine learning, *Phaseolus vulgaris*, prediction

1 | INTRODUCTION

Plants are particularly vulnerable to abiotic stressors during flowering (Prasad et al., 2017). As global temperatures continue to increase, so does the frequency of extreme weather. This makes the prediction of flowering time crucial for adapting sowing to changing growing seasons. We need to be able to predict and understand the complex interactions between genotype and environment that determine flowering time, so that existing phenological variation in germplasm can be exploited for both farming and breeding. Flowering time is often a target in breeding programmes because it is strongly associated with potential yield and helps breeders match thermal duration with environments.

Temperature and light are the main environmental controls on time to flowering. These act as distinct, but also often reinforcing environmental signals as well as modulators of the other flowering time regulatory pathways (Jung et al., 2017). Both temperature and photoperiod signals have roles in regulating growth, timing developmental transitions with respect to seasons and responding to extreme environmental conditions (Reeves & Coupland, 2000). Temperature effects can be broken down into the effects of extreme temperatures, ambient temperatures and those of extended durations of cold weather (Srikanth & Schmid, 2011). The influence of light can be broken down into two broad categories—photoperiod and quality (Srikanth & Schmid, 2011).

Environmental signals from temperature and light can reinforce each other (Song et al., 2013), as is frequently observed in plants growing in regions with distinct seasons. The sophistication of these complex interactions between environmental variables allows plants to alter the timing of flowering when they experience stress. It has been shown that flowering time is sensitive to extremes in temperature, water availability and soil nutrients (Cho et al., 2017). Plants either flower early in response to stress so that seeds are produced or delay flowering to avoid unfavourable conditions (Cho et al., 2017).

Thus far, research on the pathways through which temperature regulates flowering time have predominantly focused on vernalisation (Pose et al., 2013). The mechanisms through which ambient temperature influences flowering remain less well understood (Melzer, 2017). Research on the influence of ambient temperature on flowering time has tended to explore the impact of daily average temperatures. Much less is known about the impact of sub-daily fluctuations in ambient temperatures (Song, 2016). Recent work by Parent et al. (2019) shows that the impact of environmental drivers on phenological development varies on a sub-daily timescale in field grown maize (*Zea mays*), which aligns with recent findings in field grown *Arabidopsis thaliana* showing that the molecular impacts of changes in ambient temperature are specific to the time of day (McCarthy & Davis, 2018). These results are supported by work in controlled conditions demonstrating that a small variation in thermo cycle can cause significant delays in flowering (Karsai et al., 2008).

To summarise the above, the regulation of flowering time is complex, and our understanding of the interaction between genetic and environmental factors remains partial. It is perhaps for this reason that accurate models predicting flowering time across Genotype * Environment (G * E) have remained elusive. Cutting edge approaches include statistical methods incorporating genetic markers as independent variables (Crossa et al., 2017) and dynamic crop growth models (CGMs) incorporating genetic information through empirical representations of quantitative trait loci (QTLs) (Wallach et al., 2018).

Although both of these approaches are promising, neither are simple to employ for the purposes of simulating phenology across G * E in crop improvement programmes. Statistical models built on high-dimensional marker data quickly encounter the curse of dimensionality, in which the number of predictors exceeds the number of observations (Crossa et al., 2017). Rapid phenotyping in the field would be required to truly grow the power of these approaches, and few crop breeding programmes have the resources to conduct experiments at the required scale.

While less data intensive, approaches built on CGMs require a team of skilled physiologists and modellers. In a recent paper, Wallach et al. (2022) argued that expert knowledge must be combined with sophisticated optimisation routines to calibrate phenological parameters for new genotypes. This process is time consuming and challenging to automate within the context of a fast paced crop breeding pipeline.

By contrast, machine learning (ML) approaches may be well suited to the task of representing complex non-linear relationships between genotype and environment (van Dijk et al., 2021). Unlike crop models, they do not need periodic supervision from a team of physiologists to be re-calibrated and, at least conceptually, are relatively easy to automate within a crop breeding pipeline. Many breeding programmes have collected large numbers of field experiments over the years that may prove sufficient for the development of these models. To date, there have been very few attempts to employ ML to model flowering time across G * E. This likely reflects difficulty in publicly accessing a sufficient volume of data (Yamamoto, 2019).

Because of the difficulty in collecting large volumes of genotypic and phenotypic data, exploiting the power of ML to explore and predict the multidimensionality of G * E interactions in floral development is in its infancy. Chen et al. (2020) developed a range of ML models to predict flowering time (heading) in rice using 112 cultivars tested at seven locations across 14 years. They included temperature data, photoperiod data and genotype marker data as input features and achieved a flowering time prediction with a minimum root mean squared error (RMSE) of 4.4 days using an Extreme Gradient Boosting (XGBoost) algorithm.

Jánosi et al. (2020) further support the argument that ML techniques are well suited to modelling the complex relationship between genotypes and the environment. These authors employed an orthogonal matching pursuit algorithm to predict the first day of flowering of 298 species of bulbous perennials at a site in Budapest (Hungary) over a period of 33 years. Despite the much larger genetic differences between species of bulbous taxa than between contrasting rice genotypes, the orthogonal matching pursuit algorithm was able to achieve an R^2 of .71 over what is also a sizeable data set.

One perceived disadvantage of using ML algorithms is that they are harder to interpret than standard parametric tools. This has often led to the mistaken impression that all such techniques are 'black box' methods. However, tree-based techniques are interpretable and can contribute to knowledge of processes. Feature importance methods can rank input features according to their predictive power. Using these techniques, it is possible to find the most important drivers of flowering time and to reveal which drivers of flowering time are required for successful prediction.

These methods can complement detailed experimental approaches by generating hypotheses from large volumes of field data and comparing these with findings from chamber and greenhouse experiments. The non-parametric nature of ML algorithms relaxes distributional requirements for aggregated data sets. This enables use of large volumes of field data that were collated without the imposition of strict aggregate distributional qualities or experimental designs.

In this paper, we develop three ML models that capture the non-linear relationships between environment and genotype to predict flowering time in common bean (*Phaseolus vulgaris* L.). The models are trained and tested on experiments containing large volumes of G ∗ E combinations conducted in three environmentally distinct locations in equatorial Colombia. Common bean is unaffected by photoperiod in Colombia (where day and night are of approximately equal length). Furthermore, it does not undergo the process of vernalisation. Wild-type common bean exhibits an indeterminate growth habit; however, in many environments, a determinate growth habit has been selected for, to shorten the growing season and make mechanical harvesting easier (Repinski et al., [2012]).

Our model and data set allows us to directly address the following research gaps. First, to address the lack of a simple ML framework for simulating flowering time across G ∗ E in common bean. Second, to better understand the meteorological drivers of flowering time in field conditions and third to explore the hypothesis that sub-daily fluctuations in ambient temperature can have a significant impact on flowering time, because little is known about this subject in common bean.

The aims of this paper are therefore as follows: (1) to assess the potential of ML algorithms to predict anthesis across G ∗ E in common bean, (2) to identify the drivers of flowering across G ∗ E in common bean and (3) to test the hypothesis that the influence of environmental drivers of flowering time is mediated by time of day. We hypothesise that a simple ML framework can be developed that will accurately simulate flowering time in common bean, that flowering time in equatorial conditions is primarily driven by temperature in field conditions and that sub-daily weather information will prove to be an important driver of flowering time.

2 | MATERIALS AND METHODS

2.1 | Data

Data for planting dates and days to anthesis were merged from the International Bean Yield and Adaptation Nursery (IBYAN) data set

(International Center for Tropical Agriculture [CIAT], [2016]) and more recent experiments from CIAT's Bean Program database. Details of how the data set was compiled can be found under the metadata tab of the repository linked to in the references. The IBYAN data set has been used to explore genotypic variation in photoperiod response (White & Laing, [1989]) and to explore relationships between seed size and yield (White & Gonzalez, [1990]) in common bean.

In both data sets, anthesis is defined by 50% of plants having at least one flower open. This data set was then subset to only include data from the three locations from which weather data were available—Palmira, Popayan and Quilichao. It was further subset to include only those experiments without missing dates and with less than 5% of missing weather data. Experiments with either planting date, days to flowering or days to maturity missing were also removed. For the Bean Program data, we did not have detailed information on treatments with nutritional deficits, so we subset the data to only include medium fertility treatments. For some experiments, an irrigated and non-irrigated treatment was present. In these cases, treatments shared a weather time series, so it was not possible to include both, as this would result in duplicate weather time series. Where this was the case, we selected only non-irrigated treatments.

This procedure resulted in a data set of 170 unique G ∗ E combinations. These combinations include 57 unique growing seasons, spread across 21 years at three sites with varying climatology. Seven individual genotypes are included in this data set, from seven different families, three from the Andean and four from the Meso-American gene pools. This sample includes black, red and navy seeds and provides variation in seed size, disease resistance and flowering time. It should be noted that not every genotype is present in all of the experiments. For this reason, the number of unique G ∗ E combinations is not equal to the number of genotypes multiplied by the number of unique growing seasons.

Weather data for each growing season were taken from long-term station records at the three sites in Colombia—Palmira, Popayan and Quilichao. Figure 1 provides an annual climatology for the three locations. The progression of daily maximum temperatures through the year is similar at Palmira and Quilichao, while maximum temperatures are much cooler at Popayan. A similar trend emerges from the climatology of minimum temperatures. Although the peaks and troughs of daily rainfall are at similar times of year in all three sites, total daily rainfall is higher and more variable at Popayan and Quilichao than in Palmira. The annual progression of relative humidity in the afternoon is similar at all three sites; however, troughs in relative humidity are much lower in Popayan during the dry months.

At each of these stations, meteorological observations were available at 07:00, 13:00 and 19:00 h each day. Daily meteorological variables were also provided. There is no variation in photoperiod across the three locations, and all three sites are in the same time zone.

The data set includes a wide range of yields, days to flowering and weather conditions. The data set is characterised by a range of humidity and rainfall conditions, with average minimum and maximum
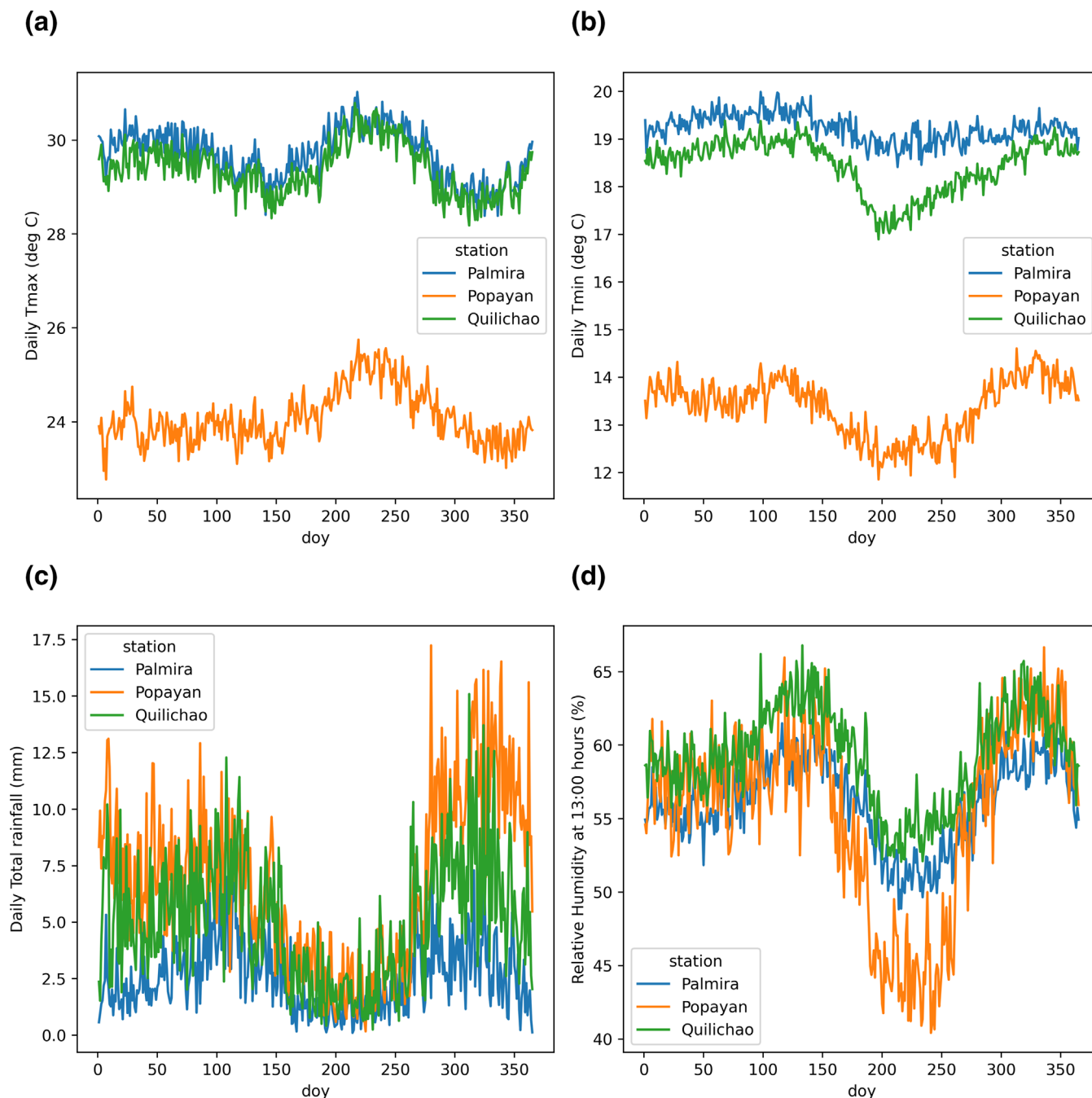
FIGURE 1 Yearly climatology for the study sites at which experiments were conducted. These were calculated from International Center For Tropical Agriculture (CIAT) weather stations: (a) T_{max}, (b) T_{min}, (c) total precipitation and (d) relative humidity at 13:00 h.

temperatures of 18.3°C and 29.2°C, respectively (Figure S1). The data set includes a sizeable portion of cold extremes. There is large variation in days to flowering, between 29 and 57 days, with a mean vegetative period of 39 days and many experiments with lengthy vegetative periods longer than 45 days (Figure S2). There is also large variation in yield across the data set, ranging between 463 and 4559 kg ha^{-1}. The mean yield recorded was 2065 kg ha^{-1} with a standard deviation of 786 kg ha^{-1} (Figure S3). This represents common bean yields varying from low to high levels of production.

2.2 | Feature engineering

Feature engineering is ML terminology for describing the process of finding and developing the variables that will be used to predict flowering time. Following Droutsas et al. (2022), in order to ensure that the models could provide dynamic predictions, we converted each day of the growing season to a binary value—0 if the plant was in the vegetative stage and 1 if the plant was in the reproductive stage. This binary feature was our target label (the variable we predicted). This

enabled us to set up the challenge as a classification problem so that given a time series matrices of predictor variables the model can output flowered or not flowered for a particular day. The number of days prior to the first day on which the model outputs 'flowered' gives days to flowering.

In addition to calculating the daily features described above, we also calculated some variables, which were accumulated over the course of the vegetative period. These include a thermal time variable, calculated as the accumulated sum of daily mean temperatures and an accumulated evaporation variable as the accumulated sum of daily pan evaporation. Accumulated evaporation integrates energy received from solar radiation and atmospheric humidity. These calculated meteorological variables were added to the list of existing meteorological features.

Finally, we created a genotype feature by translating the names of the genotypes into dummy variables using pandas 'get.dummies' function, resulting in a set of binary variables describing the lines present in the data set. These were then included alongside the meteorological features. We chose genotype as the hierarchical level at which to include genotypic information as this is the level at which selection and crosses are made.

2.3 | Description of the models

We selected three supervised ML models with the objective of finding out if particular model structures displayed advantages in our use case and to ensure that inference on the importance of different drivers of flowering time were not the result of underlying mathematical structures. All of the models chosen are capable of capturing complex non-linear relationships (Hastie, 2017), which is a requirement of successfully modelling G ∗ E interactions.

The first model we selected was a random forest model. A random forest is a collection of decision trees, each of which is trained on bootstrapped samples of a data set in a process often referred to as bagging. Random forests can maximise the benefits of bagging by ensuring minimum correlation between trees. This is achieved by randomly selecting from the set of predictor features in the development of each tree. By building each tree from a sub-selection of features, trees are less likely to be correlated, and it is possible to grow large numbers of decision trees without overfitting (Kuhn & Johnson, 2013).

The second model we selected was an XGBoost algorithm. While also a tree-based model, gradient boosting models take a different approach to building new trees. Unlike random forests, trees are not generated randomly, instead the structure of new trees depends on the errors generated by previous trees. Stagewise parameterisation of trees are combined with regularisation to slow overfitting (Hastie, 2017). Chen and Guestrin (2016) developed a scalable version for tree boosting systems, which resulted in massive increases in computational efficiency, they called this XGBoost. XGBoost algorithms have won many competitions and are widely considered to be amongst the highest performing ML algorithms available (Chen & Guestrin, 2016).

Although they use very different methods for generating trees, the random forest and XGBoost algorithms are both tree-based algorithms. We selected support vector machines (SVMs) as a third algorithm, because they use a completely different mathematical structure for performing classification tasks. SVMs are based around the concept of setting class boundaries that best separate observations into categories. This is achieved by defining a support vector classifier that sits between two sets of observations. The decision boundary is set based on a cost function that penalises classification errors. When there are many input features, kernel functions are used to find the optimal decision boundaries in higher dimensions (Kuhn & Johnson, 2013). These three modelling structures are sufficiently complex to handle non-linear relationships, while remaining more interpretable than higher dimensional methods such as neural networks (Delerce et al., 2016), which are less suited to the sample size of our data set. All three methods allow direct inclusion of categorical variables, which makes them well suited to our chosen method for incorporating genotypic information through a set of binary genotype features. Our approach is simpler than standard methods of including genotypic information through QTLs or genomic marker data. This approach results in far fewer predictor variables, which allows us to feed the ML model a genotypic signal, without requiring comprehensive data and knowledge of the multi-factorial interactions between genes that influence flowering time.

We include a simple baseline model using the concept of growing degree days as a baseline to compare against our ML models. This modelling approach uses the concept of cardinal temperatures in combination with the temperature on a particular day to determine the rate at which the plant advances towards flowering. We selected the trapezoid function, taking our cardinal temperatures from Boote et al. (2018). We used joint numerical optimisation (scipy.optimize.minimize) to select genotype-specific base temperatures and thermal time requirements. We allowed the base temperature to vary between 5°C and 8°C, choosing the upper limit based upon Kakon et al. (2018). Because our sample of genotypes is phonologically diverse, we allowed thermal time requirements to vary from between 28°C and 38°C.

If the temperature is below a base temperature, no degree days are accumulated. If the temperature is between the base temperature and the first optimum temperature (topt1), then the number of degree days accumulated (between 0 and 1) is determined by a linear equation between the base temperature and topt1. If the temperature falls between topt1 and a second optimum temperature (topt2), then a full growing degree day is accumulated. If the temperature is above topt2 and below T_{max}, then the fraction of a degree days accumulated is determined by the straight line between topt2 and T_{max}. Once the number of growing degree days exceeds a thermal time requirement, flowering occurs in the model.

2.4 | Meteorological feature selection

Section 2.2 discusses the development of features based on current understanding of the environmental drivers of flowering time.

However, many of these meteorological features were correlated; therefore, feature selection was necessary to ensure that the final model included the minimum number of features required to predict flowering time. This was necessary because the inclusion of highly correlated features results in dilution of the importance assigned to a particular feature. Large numbers of highly correlated prediction features would therefore thwart feature interpretation.

We used hierarchical clustering to group meterological variables by similarity. In this approach, variables are agglomerated depending on their distance from each other (Hastie, 2017). In this case, we converted the Spearman correlation matrix to a distance matrix. The results of this clustering procedure were visualised as a dendogram, and from the clusters identified, we chose thermal time, accumulated evaporation, T_{max}, T_{min}, relative humidity at 07:00, relative humidity at 13:00 and relative humidity at 19:00 to include in the model. Table S1 describes each of these variables in detail.

2.5 | Model evaluation

To predict days to flowering by first predicting whether the plant has flowered on each day of the growing season, it must be ensured that data from a single experiment is either assigned to the training data set or the test data set. If data from a single experiment were in both training and test sets, then information leakage would have occurred. As a result, it was necessary to perform the random split of data to the training and test data set by experiment. Because the performance of the model may differ between random splits, we performed cross-validation by training and testing the model on five different random splits. The splits were performed by ensuring that for every fold, 37 unique weather time series were assigned to training data and 20 unique time series were assigned to the test data. For replication purposes, the random seeds used to perform these splits were 1, 4, 20, 30 and 42.

We used four metrics to evaluate the models ability to capture variance and accuracy. The coefficient of determination (R^2) metric was calculated as a test of the model's ability to explain variance in days to flowering, and we also included Pearson's correlation (r) to aid comparability with other studies. The RMSE and mean absolute error (MAE) statistics were calculated to assess the model error.

2.6 | Hyperparameter tuning

A five-fold cross-validation strategy was employed to find the optimal parameters for each of the three models using a grid search method on each of the training sets generated from the random seeds. We used sklearn's GridSearchCV for this procedure. GridSearchCV runs the model with all the possible combinations of hyperparameters and chooses the joint set of parameters that minimises the loss function. In this case, we used accuracy to define the loss function. Full details of the hyperparameters tuned for each model and the exact form of accuracy metric used are provided in Table S2 to ensure reproducibility.

2.7 | Determining feature importance

We employed the permutation method provided by 'sklearn.inspection'. The permutation method determines feature importance by randomly shuffling each feature in turn, so that it no longer matches the corresponding target feature. By assessing the reduction in accuracy from shuffling each feature independently, the permutation importance algorithm assesses the relative importance of each predictor feature (Molnar, 2020). We selected the permutation importance method because it is a model agnostic method for assessing feature importance, which makes it suitable for comparing feature importance across models. Following Molnar (2020), we assessed the permutation importance on the test set generated from each random seed and report the results using variation over the five seeds as confidence intervals.

3 | RESULTS

3.1 | The ML models can predict days to anthesis accurately across G ∗ E

The first aim was to assess the potential of ML to accurately predict flowering time across G ∗ E in common bean. Figure 2 explores variance and bias in the models predictions for the observations in the test set. The results are presented as box plots to show how the models performed in five random splits of the data into train and test sets. Figure 2 shows that the ML models are able to capture a reasonable share of the variance in days to flowering. All models displayed a mean correlation between predictions and observations of .8 or above and a mean coefficient of determination (R^2) above .55. The models also exhibited high accuracy—all models displayed a MAE lower than 3 days and a mean RMSE of less than 4 days.

The tree-based models outperformed the SVM model across all four of the evaluation metrics. Given that uncertainty in observations of days to flowering is approximately 2–3 days, both the XGBoost and random forest models displayed impressively high accuracy. These results suggest that tree-based architectures may be the best ML algorithms for the breeding programme to take forward for predicting phenology. This recommendation is in line with evidence from the literature that XGBoost models are often high performing across a range of classification problems both within, and outside of, the environmental domain (Chen & Guestrin, 2016).

R^2, MAE and RMSE metrics show that performance varied across the five random splits; further, there does exist a random split that reduces model performance significantly for the tree-based algorithms. Variation in performance across splits is not unexpected. In this study, we have compiled a large volume of data to provide the ML algorithms with sufficient numbers of experiments from which to learn the relationships between flowering time and the environment for a selection of genotypes. However, because these data were not originally collected to ensure a balanced sample over time (they were conducted to maximise yield gains through selection), experiments are
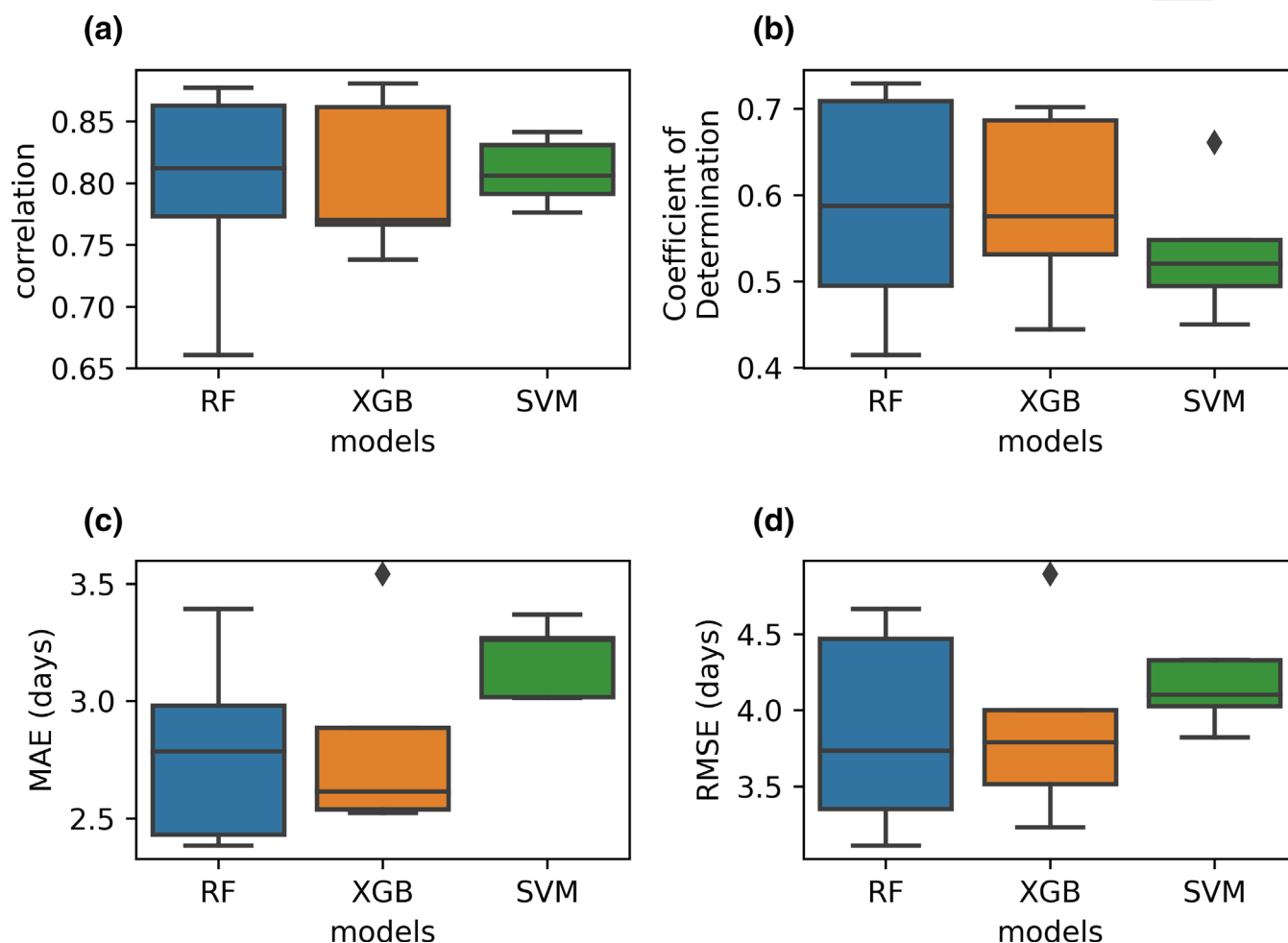
FIGURE 2 Statistics for observed versus predicted days to anthesis for the three machine learning models. (a) Pearson's correlation, (b) coefficient of determination, (c) mean absolute error (days) and (d) root mean squared error (days). RF stands for random forest, XGB stands for Extreme Gradient Boosting and SVM stands for support vector machine.

not split equally across locations and genotypes. For this reason, there will be random splits in which the statistical characteristics of the training set do not perfectly match those of the test set, and model performance will be lower in such cases.

All models far outperformed the baseline GDD model. We developed a GDD model for one of the five cross-validation folds (fold1). The GDD model achieved a correlation of .65, a coefficient of determination of .12 and an RMSE of 5.44 days. This compares with correlations of above .75, coefficient of determinations of above .51 and RMSEs of below 4.1 days for all the ML models. The basic GDD model achieved a reasonable accuracy and correlation but explained far less of the variance in the test set than the ML models.

3.2 | Thermal time and accumulated evaporation are the most important drivers of days to anthesis across G * E

The second aim was to identify the main drivers of flowering time across G * E for common bean. Figure 3 presents the prediction

features ranked by permutation importance over the test set. The error bars represent the variation in permutation importance over cross-validation folds. All three models find that thermal time and accumulated evaporation are the most important drivers of whether the plant has flowered or not. Although there is clear agreement on this hypothesis across models, there is divergence between the relative importance across the tree-based models and the SVM. In the tree-based models, thermal time is a far more important driver of flowering than accumulated evaporation, whereas in the SVM, they are of similar importance.

Beyond agreement on these two central drivers, feature importance differed between models. Relative humidity in the afternoon and evening was found to be important in the SVM, but not in the tree-based models. Conversely, the tree-based models made more use of genotypic information than the SVM. Interestingly, although the models did not agree on the relative importance of genotypic information, all three models found the categorical representation of genotype A21 to be of greater importance than the other genotypes. This suggests that A21 may exhibit variation across G * E that is worth exploring further. The extent to which feature importance
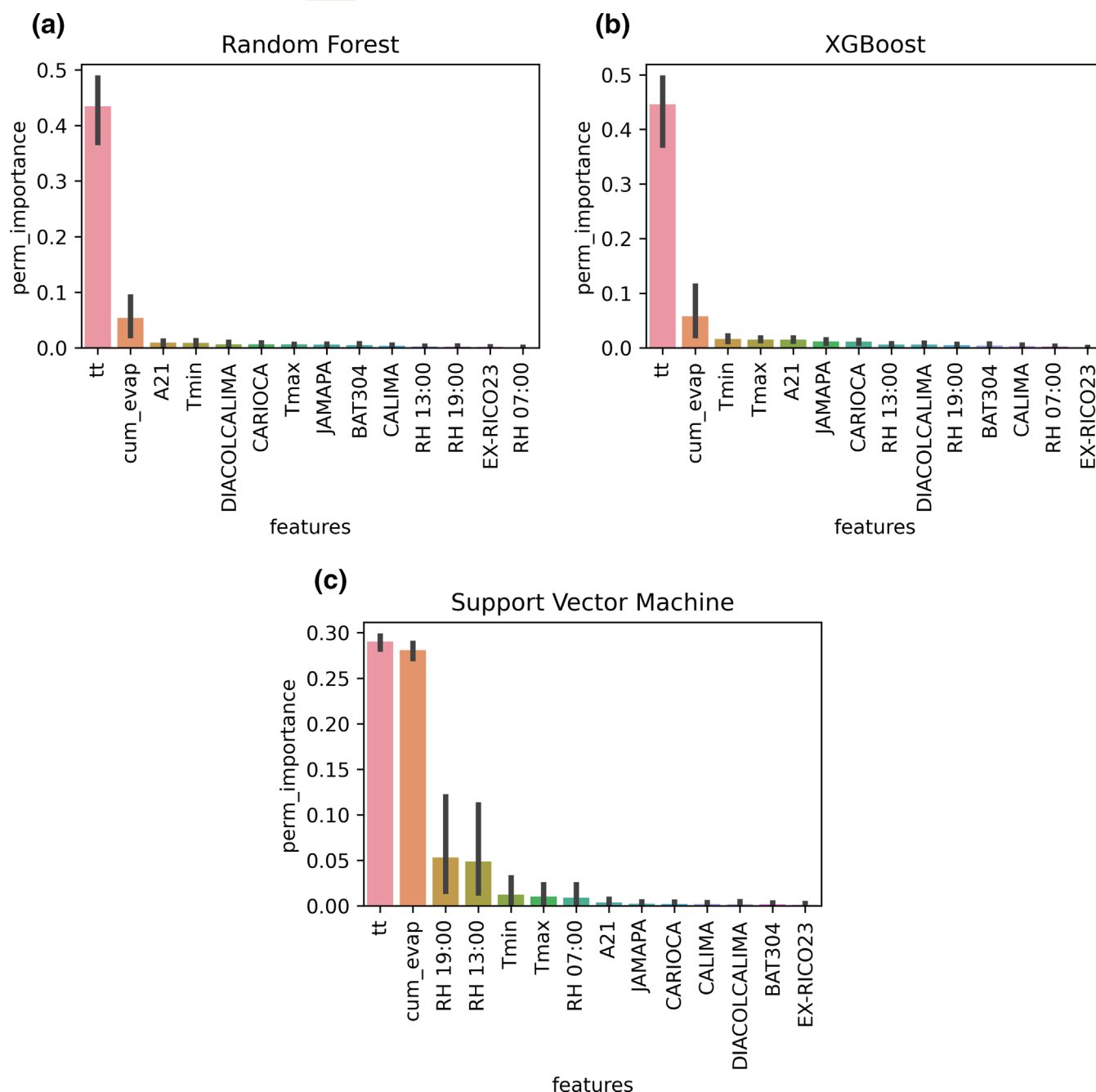
FIGURE 3 The permutation feature importance describes the importance of each feature in the model to the prediction of flowering time. Here, it is presented for all three machine learning models, where (a) refers to the random forest model, (b) refers to the XGBoost model and (c) refers to the support vector machine model. 'tt' is the time series of thermal time as defined in our modelling and 'cum evap' is the time series of accumulated evaporation. The x-axis labels in capitals are the names of the genotypes included in the model as categorical features. The error bars represent the 95% confidence intervals.

varies between models highlights the importance of using multiple ML techniques for physiological hypothesis generation.

The third aim was to test the hypothesis that the influence of environmental drivers on flowering time is mediated by time of day. If this were the case, then we would expect the time series of temperatures at particular times of the day to be more powerful predictors than summary measures like the maximum or minimum temperature. We perform a test of this hypothesis by replacing maximum and

minimum temperatures with the 07:00, 13:00 and 19:00 temperature time series. Table 1 compares the performance of these models and shows that the inclusion of time of day information does not improve the models, if anything the performance of the models is slightly reduced. On the basis of these results, we cannot reject the null hypothesis that the daily maximum and minimum temperatures adequately provide the information required for phenological development towards flowering.

TABLE 1 Mean statistics across folds for the random forest (RF), XGBoost (XGB) and support vector machine (SVM) models of flowering time.

ML type	Input features	R^2	RMSE	MAE
RF	Original model	.59	3.87	2.80
RF	Sub-daily model	.56	3.97	2.84
XGB	Original model	.59	3.89	2.82
XGB	Sub-daily model	.57	3.92	2.90
SVM	Original model	.53	3.89	3.18
SVM	Sub-daily model	.54	3.92	3.22

Note: For each of the three models, original and sub-daily versions of the model are presented. The sub-daily model differs from the original model by replacing the daily time series of T_{min} and T_{max}, with the 07:00, 13:00 and 19:00 daily temperature time series. The root mean squared error (RMSE) and the mean absolute error (MAE) are provided in days.

4 | DISCUSSION

4.1 | Modelling flowering time

We have presented an ML-based approach to modelling flowering time in common bean across G ∗ E for a Colombian target population of environments. We have demonstrated that within our ML framework, the simple approach of incorporating genotypes as categorical variables results in an accurate model of days to flowering. The models presented in this paper are easy to build upon. To add data for new genotypes, all that is required is days to flowering and maturity, the name of the genotype and a weather time series for the duration of the experiment. These variables will be collected as a matter of course in breeding programmes, so no additional data collection efforts linked to phenotyping or genotyping are required. Because hyperparameter tuning occurs simultaneously for all genotypes, updating the model can be automated so long as an excel spreadsheet containing the simple phenological and weather information described above is maintained and added to (see exact list of variables in the Supporting Information).

This approach offers a number of advantages versus state of the art statistical models which incorporate genetic information through genomic markers. The use of high dimensionality marker data encounters the 'curse of dimensionality', in which the number of markers exceeds the number of observations (Crossa et al., 2017). In addition, many of these predictor variables are highly correlated, and so such models must carefully guard against overfitting (Crossa et al., 2017). This kind of statistical modelling requires careful data collection strategies to ensure input data meets distributional assumptions and needs intensive phenotyping and genotyping. While highly useful and a proven source of genetic gain in breeding programmes (Crossa et al., 2017), these models are less suited to real-time simulation of phenology across target population of environments.

Our approach also offers advantages versus dynamic CGMs, which are increasingly being used for characterising genotypic responses across target populations of environments (Heinemann et al., 2015). CGMs employ process-based equations to simulate the impact of the environment on crop growth and development.

Genotypic information is incorporated through empirical coefficients that mediate the relationship between the environment and the plant. These genetic coefficients must be parameterised, and doing so is a time consuming and skill intensive process (Wallach et al., 2022). Indeed, the latest thinking on the calibration of phenology modules in CGMs develops a detailed protocol that combines sophisticated calibration algorithms with expert knowledge (Wallach et al., 2022). Because (in most current approaches) genotypes must be calibrated one by one, the calibration of multiple genotypes in CGMs is by nature much more involved than using the methods presented in this paper.

It is very difficult to compare the performance of new models with those in the published literature. This is because existing studies tend to use data sets of different sizes—both in terms of the number of experiments they model and the genetic diversity within those experiments. What is possible is to assess whether studies that compare the performance of both ML approaches and traditional crop modelling approaches find one to be better than the other. In this study, we showed that ML-based approaches clearly outperform simple GDD approaches in a tropical bean growing target population of environments. Our results support Chen et al. (2020), who found that ML techniques outperformed traditional methods of simulating flowering time in rice and McCormick et al. (2021), who found the same for soybean.

We hypothesise that ML models are more accurate because their more flexible structure allows them to more easily capture non-linear interactions between genotype and environment. Tree-based ML models such as random forests and XGBoost algorithms are also flexible in the structure of these interactions across different sub-samples of the data. This allows for changes in the structure of non-linear interactions between input features in varying environmental conditions and may explain why they outperformed both SVMs and traditional crop models.

The inherent flexibility of ML architectures contrasts with the structure of phenology algorithms in most process-based crop models. These models assume a fixed structure of thermal time accumulation, in which the rate of phenological development progresses above a base temperature, is maximised at an optimal temperature and slows as temperatures become sub-optimally warm (Wallach et al., 2019).

Although temperature response functions can vary between process-based crop models (Wang et al., 2017), the assumption that there is a stable shape to this function across multidimensional environmental space remains implicit in these models. For example, in process-based crop models, the phenological response to temperature does not change with realtive humidity—even though this would clearly change the tissue temperature the plant experiences (Parent et al., 2019), as has been shown for common bean (Deva et al., 2020; Suárez et al., 2021). Tree-based ML algorithms do not make these assumptions and instead learn how thresholds change in different growing conditions. We hypothesise that this facilitates emergent phenological responses across G * E, leading to higher accuracy.

While the models presented here are easy to update, re-calibrate and perform large volumes of simulations with, it is important to note that they cannot estimate the phenological behaviour of new genotypes for which no experimental information is yet available. This challenge will need to be answered by different kinds of models that directly incorporate the genetic drivers of flowering. The more complex statistical models above or perhaps new methods building on the pioneering work of Wallach et al. (2018) will be needed.

4.2 | The determinants of flowering time in common bean

Our findings support the consensus in the literature that temperature is the main determinant of flowering in the absence of variation in photoperiod or vernalisation (Craufurd & Wheeler, 2009; Pose et al., 2013). Figure 3 shows that all three of the models agree with the notion that thermal time is the most important determinant of flowering time in equatorial conditions. We employed a very simple definition of thermal time in our ML models (simply accumulating temperature day by day), compared with the more complex definition we used in our GDD model. Because common bean is not a vernalising plant, the importance of thermal time in the ML models highlights the role of ambient temperature as an important development signal.

All three of the models also agree that accumulated evaporation is the second most important determinant of whether the plant has flowered or not on a given day. These results are in agreement with Parent et al. (2019) who hypothesise that it is the combination of temperature and VPD that drive phenological progression in maize plants. Whether this is because of the direct effects of changes in water status on the growth of leaves (Lacube et al., 2017) or because increased evaporative demand impacts the temperature response curve by cooling the plant, remains unclear. Our findings are consistent with both hypotheses.

In addition to generating hypothesis on the determinants of flowering time, our results also have implications for the timescales at which common bean responds to temperature signals. Interpretation of feature importance shows that it is the variables that accumulate over time that are the most powerful determinants of whether the plant has flowered or not. This supports the prevailing view that plants integrate information from environmental signals over long time periods to ensure that phenological changes are responding to seasonal progression (Hepworth et al., 2018). Experiments in *A. thaliana* and *Arabidopsis halleri* suggest a memory for temperature signals of between 4 and 6 weeks (Kudoh, 2016), and our results are consistent with the magnitude of these estimates.

Our finding that the two time integrated features are the most important drivers of flowering differs from those of Jánosi et al. (2020), who found that the most powerful predictors in their model of flowering time in bulbous taxa were daily snow depth anomalies on 1–3 days of the growing season. It is possible however as Wesselingh (2020) suggest that these few days acted as a proxy for a winter signal, which would reconcile their findings with the prevailing view that phenological development is driven by the integration of information over longer periods of time.

The hypotheses above should be viewed in the context of an equatorial target population of environments. While the data set contains a large number of treatment * genotype combinations and includes experiments in cold, hot, dry and wet conditions, the majority of experiments were conducted at sites that are well adapted to breeding beans. To date, a comparably sized number of experiments have not been conducted in more extreme conditions. Future work could repeat this process once a suitably large tranche of field experiments from more extreme environments becomes available.

4.3 | Societal impact

There is an urgent need to accelerate crop breeding for adaptation to a changing climate (Ramirez-Villegas et al., 2020). To do so, we must be able to conduct accession selection experiments with a clear understanding of how available germplasm behave across environments. In recent years, breeding programmes have begun incorporating dynamic CGMs to their teams, as it is much faster and cheaper to simulate the behaviour of germplasm across environments than it is to perform the hundreds of experiments that would be required to truly sample climate variability.

In addition, the careful calibration and precise evaluation required to ensure that CGMs accurately capture a genotypes behaviour across environments is skill intensive and time consuming (McCormick et al., 2021). These labour-intensive steps preclude rapid and real-time input to the selection process. For this reason, Ramirez-Villegas et al. (2020) recommended that CGIAR breeding institutes also consider the development of simpler models that can simulate key traits across G * E with high levels of accuracy. The modelling approach for flowering time in this paper meets that recommendation directly.

We have created a tool that allows CIAT common bean breeders to simulate flowering time across G * E in their Colombian bean growing TPE more accurately and easily than was previously possible. The ML pipeline only requires the user to be able to run scripts in Python, and re-calibration of the model as more experiments become available is an automated objective procedure that does not require more programming or a detailed understanding of the underlying algorithms.

Computing requirements are minimal, and the simulation of hundreds of experiments takes minutes rather than days.

In addition to being an important trait in itself, days to anthesis is also an essential component of the length of the growing season. The enhanced capacity to conduct accurate simulations of flowering time across G $*$ E in real time can help to deepen the information available to breeders about the phenological characteristics of their germplasm and can aid the design of experiments to ensure targeted inclusion of early and late flowering genotypes.

This can be useful both to increase the efficiency of the selection process and to ensure that the design of heat and drought stress experiments take account of the probability of escape through phenological variation. For example, an a priori classification of genotypes into those likely to flower early under heat can help breeders to understand the contribution of avoidance vs. true tolerance in their germplasm. Further, combining the ability to predict flowering time with an understanding of pest dynamics can help breeders to ensure that experiments are sown to avoid flowering coinciding with severe attacks of insect or fungi.

The methodology demonstrated in this paper is modular and could be imported into existing crop models to improve their ability to simulate G $*$ E. The inclusion of modular ML classification modules into existing crop models has been successfully demonstrated by Droutsas et al. ([2022](#)), and a similar approach could be taken for the genotype-specific classification algorithm presented in this paper. As part of rapid modernisation of breeding programmes targeted at increasing genetic gain, software is being built to simulate the breeding pipelines themselves, and a modular phenological model such as this one could also feed into these optimisation procedures.

4.4 | Conclusion

We have shown that using ML, it is possible to build an accurate model of days to anthesis across G $*$ E for common bean with only a simple representation of genetic information as categories. The model we have developed can be used by breeders at CIAT to explore the phenological variation of existing germplasm in their Colombian target population of environments. It is easy to update the underlying data set on which the model is built to include more genotypes. Having demonstrated proof of concept, it would be useful to begin incorporating heat tolerant germplasm of interest. In particular, experiments with inter-specific crosses including *Phaseolus acutifolius* (tepary) and Andean genotypes exhibiting drought tolerance could be introduced to the training data set and the extent to which they utilise phenological heat avoidance strategies compared.

Our equatorial common bean data set has provided a useful set of environments in which to test this approach; however, it did not include changes in photoperiod which is a feature of bean growing environments from the outer edge of the subtropics to the poles. It would be interesting to see if this modelling framework can be extended to simulate G $*$ E in common bean grown under conditions of changing photoperiod or to simulate different crop species for

which larger volumes of data are available. All that would be required is the addition of day length as a prediction feature.

In addition to their utility in prediction, ML techniques can be useful in the plant sciences to help ask interesting questions of large collections of unstructured field experiments. By identifying patterns across G $*$ E, these techniques provide a new source of hypothesis generation about environmentally driven plant processes. For the equatorial population of environments studied in this paper, we found that thermal time and accumulated evapotranspiration were powerful drivers of flowering in all three ML models.

While hypotheses generated by ML models offers an exciting new path for modellers to contribute towards understanding, our work also suggests a note of caution. Different families of ML models used information in different ways, resulting in some variation in the environmental variables they determined to be important. It is therefore essential to use multiple ML models with assorted mathematical structures to generate physiological or genetic hypotheses. More work is needed to explore how the mathematical underpinnings of different models impacts feature importance in this context.

It is only through the experimental work of physiologists, biologists and geneticists that ML-generated hypotheses can be definitively tested and a deeper understanding of the causal mechanisms driving these empirical signals revealed. For example, this ML framework could identify the categorical variable associated with a particular genotype as being an important determinant of flowering. This information could provide a starting point for allocating some resources towards increased phenotyping and genotyping of a particular genotype. In this context, ML-based models can utilise existing experimental data to help breeding programmes to coordinate phenotyping and genotyping with information from experimental work that has already been conducted.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data are available in the Supporting Information.

ORCID

Chetan Deva https://orcid.org/0000-0001-5434-5416

Ioannis Droutsas https://orcid.org/0000-0002-5123-3379

REFERENCES

Boote, K. J., Prasad, V., Allen Jr, L. H., Singh, P., & Jones, J. W. (2018). Modeling sensitivity of grain yield to elevated temperature in the DSSAT crop models for peanut, soybean, dry bean, chickpea, sorghum, and millet. *European Journal of Agronomy*, *100*, 99–109.

Chen, T.-S., Aoike, T., Yamasaki, M., Kajiya-Kanegae, H., & Iwata, H. (2020). Predicting rice heading date using an integrated approach combining a machine learning method and a crop growth model. *Frontiers in Genetics*, *11*, 1643.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp. 785–794.

Cho, L.-H., Yoon, J., & An, G. (2017). The control of flowering time by environmental factors. *The Plant Journal*, *90*(4), 708–719.

Craufurd, P. Q., & Wheeler, T. R. (2009). Climate change and the flowering time of annual crops. *Journal of Experimental Botany*, *60*(9), 2529–2539.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., & Dreisigacker, S. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975.

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M. C., Prager, S. D., Patiño, V. H., Garcés Varón, G., & Jiménez, D. (2016). Assessing weather-yield relationships in rice at local scale using data mining approaches. *PloS One*, *11*(8), e0161620.

Deva, C. R., Urban, M. O., Challinor, A. J., Falloon, P., & Svitákova, L. (2020). Enhanced leaf cooling is a pathway to heat tolerance in common bean. *Frontiers in Plant Science*, *11*, 19.

van Dijk, A. D. J., Kootstra, G., Kruijer, W., & de Ridder, D. (2021). Machine learning in plant science and plant breeding. *Iscience*, *24*(1), 101890.

Droutsas, I., Challinor, A. J., Deva, C. R., & Wang, E. (2022). Integration of machine learning into process-based modelling to improve simulation of complex crop responses. *In Silico Plants*, *4*(2), diac017.

Hastie, T. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*: Springer Open.

Heinemann, A. B., Barrios-Perez, C., Ramirez-Villegas, J., Arango-Londoño, D., Bonilla-Findji, O., Medeiros, J. C., & Jarvis, A. (2015). Variation and impact of drought-stress patterns across upland rice target population of environments in brazil. *Journal of Experimental Botany*, *66*(12), 3625–3638.

Hepworth, J., Antoniou-Kourounioti, R. L., Bloomer, R. H., Selga, C., Berggren, K., Cox, D., Harris, B. R. C., Irwin, J. A., Holm, S., Säll, T., & Howard, M. (2018). Absence of warmth permits epigenetic memory of winter in Arabidopsis. *Nature Communications*, *9*(1), 1–8.

International Center for Tropical Agriculture (2016). Bean Program: The VEF(IBN)-EP-IBYAN scheme. https://doi.org/10.7910/DVN/S7LJBI. Accessed: 2023-03-25.

Jánosi, I. M., Silhavy, D., Tamás, J., & Csontos, P. (2020). Bulbous perennials precisely detect the length of winter and adjust flowering dates. *New Phytologist*, *228*(5), 1535–1547.

Jung, C., Pillen, K., Staiger, D., Coupland, G., & von Korff, M. (2017). Recent advances in flowering time control. *Frontiers in Plant Science*, *7*, 2011.

Kakon, S. S., Mian, M. A. K., Saha, R. R., Nasreen, S. S., Naher, Q., & Hossain, M. A. (2018). Phenology, growth and pod yield of French bean (Phaseolus vulgaris L.) as influenced by temperature variations under different sowing dates. *Bangladesh Agronomy Journal*, *21*(2), 25–32.

Karsai, I., Szűcs, P., Kőszegi, B., Hayes, P. M., Casas, A., Bedő, Z., & Veisz, O. (2008). Effects of photo and thermo cycles on flowering time in barley: A genetical phenomics approach. *Journal of Experimental Botany*, *59*(10), 2707–2715.

Kudoh, H. (2016). Molecular phenology in plants: In natura systems biology for the comprehensive understanding of seasonal responses under natural environments. *New Phytologist*, *210*(2), 399–412.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*, 26. Springer.

Lacube, S., Fournier, C., Palaffre, C., Millet, E. J., Tardieu, F., & Parent, B. (2017). Distinct controls of leaf widening and elongation by light and evaporative demand in maize. *Plant, Cell & Environment*, *40*(9), 2017–2028.

McCarthy, K., & Davis, S. J. (2018). Rediscovering natural flowering. *Nature Plants*, *4*(10), 750–751.

McCormick, R. F., Truong, S. K., Rotundo, J., Gaspar, A. P., Kyle, D., Van Eeuwijk, F., & Messina, C. D. (2021). Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. *In Silico Plants*, *3*(1), diab004.

Melzer, R. (2017). Regulation of flowering time: A splicy business. *Journal of Experimental Botany*, *68*(18), 5017.

Molnar, C. (2020). *Interpretable machine learning*: Lulu.com.

Parent, B., Millet, E. J., & Tardieu, F. (2019). The use of thermal time in plant studies has a sound theoretical basis provided that confounding effects are avoided. *Journal of Experimental Botany*, *70*(9), 2359–2370.

Pose, D., Verhage, L., Ott, F., Yant, L., Mathieu, J., Angenent, G. C., Immink, R. G. H., & Schmid, M. (2013). Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature*, *503*(7476), 414–417.

Prasad, P. V. V., Bheemanahalli, R., & Jagadish, S. V. K. (2017). Field crops and the fear of heat stress opportunities, challenges and future directions. *Field Crops Research*, *200*, 114–121.

Ramirez-Villegas, J., Molero Milan, A., Alexandrov, N., Asseng, S., Challinor, A. J., Crossa, J., van Eeuwijk, F., Ghanem, M. E., Grenier, C., Heinemann, A. B., & Wang, J. (2020). CGIAR modeling approaches for resource-constrained scenarios: I. Accelerating crop breeding for a changing climate. *Crop Science*, *60*(2), 547–567.

Reeves, P. H., & Coupland, G. (2000). Response of plant development to environment: Control of flowering by daylength and temperature. *Current Opinion in Plant Biology*, *3*(1), 37–42.

Repinski, S. L., Kwak, M., & Gepts, P. (2012). The common bean growth habit gene PvTFL1y is a functional homolog of Arabidopsis TFL1. *Theoretical and Applied Genetics*, *124*(8), 1539–1547.

Song, Y. H. (2016). The effect of fluctuations in photoperiod and ambient temperature on the timing of flowering: Time to move on natural environmental conditions. *Molecules and Cells*, *39*(10), 715.

Song, Y. H., Ito, S., & Imaizumi, T. (2013). Flowering time regulation: Photoperiod- and temperature-sensing in leaves. *Trends in Plant Science*, *18*(10), 575–583.

Srikanth, A., & Schmid, M. (2011). Regulation of flowering time: All roads lead to Rome. *Cellular and Molecular Life Sciences*, *68*(12), 2013–2037.

Suárez, J. C., Urban, M. O., Contreras, A. T., Noriega, J. E., Deva, C., Beebe, S. E., Polanía, J. A., Casanoves, F., & Rao, I. M. (2021). Water use, leaf cooling and carbon assimilation efficiency of heat resistant common beans evaluated in Western Amazonia. *Frontiers in Plant Science*, *12*, 644010.

Wallach, D., Hwang, C., Correll, M. J., Jones, J. W., Boote, K., Hoogenboom, G., Gezan, S., Bhakta, M., & Vallejos, C. E. (2018). A dynamic model with QTL covariables for predicting flowering time of common bean (Phaseolus vulgaris) genotypes. *European Journal of Agronomy*, *101*, 200–209.

Wallach, D., Palosuo, T., Thorburn, P., Mielenz, H., Buis, S., Hochman, Z., Gourdain, E., Andrianasolo, F., Dumont, B., Ferrise, R., & Gaiser, T. (2022). Proposal and extensive test of a calibration protocol for crop phenology models. *bioRxiv*.

Wallach, D., Palosuo, T., Thorburn, P., Seidel, S. J., Gourdain, E., Asseng, S., Basso, B., Buis, S., Crout, N. M. J., Dibari, C., & Dumont, B. (2019). How well do crop models predict phenology, with emphasis on the effect of calibration? *bioRxiv*, 708578.

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R. P., Kimball, B. A., Ottman, M. J., Wall, G. W., White, J. W., & Reynolds, M. P. (2017). The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nature Plants*, *3*(8), 1–13.

Wesselingh, R. A. (2020). Winter is coming and the clock starts ticking. *New Phytologist*, *228*(5), 1458–1459.

White, J. W., & Gonzalez, A. (1990). Characterization of the negative association between seed yield and seed size among genotypes of common bean. *Field Crops Research*, *23*(3-4), 159–175.

White, J. W., & Laing, D. R. (1989). Photoperiod response of flowering in diverse genotypes of common bean (Phaseolus vulgaris). *Field Crops Research*, *22*(2), 113–128.

Yamamoto, K. (2019). Distillation of crop models to learn plant physiology theories using machine learning. *PloS One*, *14*(5), e0217075.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.