*in silico*
# Plants

OXFORD

# Incorporating a dynamic gene-based process module into a crop simulation model

Fabio A. A. Oliveira[1,2], James W. Jones[1], Willingthon Pavan[2,3],
Mehul Bhakta[4], C. Eduardo Vallejos[5], Melanie J. Correll[1], Kenneth J. Boote[1],
José M. C. Fernandes[2,6], Carlos A. Hölbig[2] and Gerrit Hoogenboom[1,7,*]

[1]Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611, USA
[2]Graduate Program in Applied Computing, University of Passo Fundo, Passo Fundo, RS 99010-970, Brazil
[3]International Fertilizer Development Center, Muscle Shoals, AL 35661, USA
[4]Bayer Crop Science, 700 Chesterfield Parkway, West Chesterfield, MO 63017, USA
[5]Horticultural Sciences Department, University of Florida, Gainesville, FL 32611, USA
[6]EMBRAPA Trigo, Passo Fundo, RS 99010-970, Brazil
[7]Institute for Sustainable Food Systems, University of Florida, Gainesville, FL 32611, USA
*Corresponding author's e-mail address: gerrit@ufl.edu

Handling Editor: Xinyou Yin

## ABSTRACT

Dynamic crop simulation models are tools that predict plant phenotype grown in specific environments for genotypes using genotype-specific parameters (GSPs), often referred to as 'genetic coefficients'. These GSPs are estimated using phenotypic observations and may not represent 'true' genetic information. Instead, estimating GSPs requires experiments to measure phenotypic responses when new cultivars are released. The goal of this study was to evaluate a new approach that incorporates a dynamic gene-based module for simulating time-to-flowering for common bean (*Phaseolus vulgaris* L.) into an existing dynamic crop model. A multi-environment study that included 187 recombinant inbred lines (RILs) from a bi-parental bean family was conducted in 2011 and 2012 to measure the effects of quantitative trait loci (QTLs), environment (E) and QTL × E interactions across five sites. A dynamic mixed linear model was modified in this study to create a dynamic module that was then integrated into the Cropping System Model (CSM)-CROPGRO-Drybean model. This new hybrid crop model, with the gene-based flowering module replacing the original flowering component, requires allelic make-up of each genotype that is simulated and daily E data. The hybrid model was compared to the original CSM model using the same E data and previously estimated GSPs to simulate time-to-flower. The integrated gene-based module simulated days of first flower agreed closely with observed values (root mean square error of 2.73 days and model efficiency of 0.90) across the five locations and 187 genotypes. The hybrid model with its gene-based module also described most of the G, E and G × E effects on time-to-flower and was able to predict final yield and other outputs simulated by the original CSM. These results provide the first evidence that dynamic crop simulation models can be transformed into gene-based models by replacing an existing process module with a gene-based module for simulating the same process.

**KEYWORDS:** CROPGRO-Drybean; DSSAT; mixed-effect model; *Phaseolus Vulgaris* L.; QTL.

## 1. INTRODUCTION

Scientific advances in understanding plant genes combined with advances in technologies for rapidly and inexpensively identifying genetic make-up of plants (Thomson 2014; Rasheed *et al.* 2017) have fuelled considerable interest in using genetic information to predict plant phenotypes. Analytical tools are now available to identify the genes that are associated with the variation in different plant traits. These bioinformatics tools also can identify important gene-by-environment (G × E) interactions that contribute to the observed variation in specific traits (Yin *et al.* 2018). Rapid progress in genome-wide

association studies (GWAS) has enabled researchers to identify genes associated with variation in human diseases (Bush and Moore 2012).

Genome-wide prediction models that use GWAS also have become powerful tools for improving crops such as tropical rice (e.g. Spindel *et al.* 2016). The GWAS approach has been implemented in recent work in other crops (Brown *et al.* 2014; Huang and Han 2014; Cooper *et al.* 2016). Scientists use statistical methods, such as single-locus analysis based on ANOVA, linear regression and mixed linear regression models, to detect a gene or gene combinations associated with variation in a phenotypic trait (White and Hoogenboom 2003; White 2006; Yin and Struik 2010). These tools also assist geneticists and plant breeders for prediction and selection of lines to improve crop yield.

Concepts have been under development since the early 1970s for predicting crop yield variation using dynamic models as affected by environmental conditions and management scenarios, and to some variation among cultivars (Jones *et al.* 2016; Thorburn *et al.* 2018). Differences among cultivars are represented by empirical genotype-specific parameters (GSPs). However, these models do not use information on variation in genes among the cultivars. Instead, the GSPs for each genotype must be estimated using data from laboratory or field studies (Hunt *et al.* 1993; Anothai *et al.* 2008; Buddhaboon *et al.* 2018).

Recognizing the potential for introducing genetic information into crop models, White and Hoogenboom (1996, 2003) showed that some of the BEANGRO model's GSPs (Hoogenboom *et al.* 1992, 1994) could be estimated as linear functions of genetic information. This approach was also used by Messina *et al.* (2006) for the CROPGRO-Soybean model and by other researchers for different crops (Yin *et al.* 2000, 2003; Reymond *et al.* 2003; Hammer *et al.* 2010; Gu *et al.* 2014). Furthermore, this approach of relating existing crop model GSPs to molecular markers was shown to provide better yield predictions than that of a statistical model for maize (Technow *et al.* 2015). More recently, Wallach *et al.* (2018) showed that genetic effects on rate of progress to first flower in common bean can be estimated using field data from a multi-environmental trial containing a large number of genotypes.

Although GSPs can provide high levels of prediction in crop models when they are independently estimated for each genotype, these parameter may not accurately represent the genetic architecture of the associated crop phenotype or process (Hwang *et al.* 2017). Acharya *et al.* (2017) found that commonly used approaches for estimating GSPs for the cropping system model (CSM)-CROPGRO-Drybean model (Boote *et al.* 1998; Jones *et al.* 2003; Hoogenboom *et al.* 2019a) resulted in considerable equifinality among estimated GSPs, which means that multiple sets of possible GSP values produced very similar responses. This was demonstrated by Acharya *et al.* (2017) using a synthetic population based on known GSPs that were used to generate synthetic field data. Then, blind estimates of GSPs using those synthetic data differed from original values. Even though new GSPs reliably predicted crop growth and yield, the procedure was unable to recover the original GSP values.

The previously discussed studies contain relationships and assumptions made by the original crop model developers, including the functional forms used to describe the E and G effects on predicted dynamic rates. As a result, this use of existing relationships makes it difficult to identify G and G × E effects from field data, which can be seen in the expanded original model form published by Wallach *et al.* (2018). Note that this expanded functional form inherently includes many G × E interaction terms that may or may not exist. Incorporating genetic information into an original model's functional form entangles G, E and G × E effects and, thus, does not enable one to study interactive G × E effects on the rate of progress towards flowering. Furthermore, we have learned that there is variation among genotypes that was not captured in the original model formulations and associated assumptions (Boote *et al.* 2003, 2013; Acharya *et al.* 2017). One example is that some combinations of genes may result in different responses to temperature than others, whereas the assumptions imbedded in the existing models mostly assume that temperature responses of all genotypes are the same.

Another issue is that the gene-based approach that thus far has evolved in the crop modelling community has not been widely embraced by the genetics community, nor have the analytical approaches used by geneticists to predict genetic effects on crop traits been adopted by the crop modelling community (Boote *et al.* 2021). There have been limited interactions between these science communities that might lead to more rapid advances in gene-based modelling. Hwang *et al.* (2017) concluded that comprehensive gene-based crop models may be developed using existing crop models by replacing existing component dynamic modules with gene-based modules as they are developed.

Recent progress has helped identify a possible pathway to help converge these communities. A multi-environment trial (MET) that was conducted in 2011 and 2012 included 187 common bean (*Phaseolus vulgaris*) recombinant inbred lines (RILs) from a bi-parental family. As part of this study, significant quantitative trait loci (QTLs) controlling the time-to-flowering in the RIL population were identified (Bhakta *et al.* 2015, 2017), which provided an opportunity to model QTL and environmental effects on the time-to-flowering. This study included geneticists, biostatisticians and crop modellers asking questions about which genes affected different growth and development processes across environments and what G × E interactions were important. One of the outcomes of this study was a QTL-based mixed model to determine the G and G × E interactions in order to build a predictive model for the time-to-flowering trait (Bhakta *et al.* 2017).

Vallejos *et al.* (2021) described how one can use a statistical model to develop a dynamic mixed model that can predict the time to first flower phenotype based on a daily development rates. They developed a model that predicts the daily development rate and discussed its potential integration into an existing dynamic crop model that responds to varying environmental conditions. However, integration of this model was not attempted by Vallejos *et al.* (2021). There could be unknown or implicit assumptions in the original crop model that might lead to erratic responses that would have to be identified and addressed when a gene-based module is integrated into an existing dynamic crop model that does not rely on genetic data inputs.

The goal of this study was, therefore, to address the questions of how this type of integration can be done in a comprehensive dynamic crop model and what complications and limitations are likely to occur. The first objective was to develop and integrate a dynamic statistical gene-based module into the CSM-CROPGRO-Drybean model to predict

the time of first flower appearance using data obtained from the MET bean studies. The second objective was to demonstrate potential applications of the hybrid dynamic model as a breeding tool for studying the G × E interactions using sensitivity analysis and for simulating yield.

## 2. MATERIALS AND METHODS
### 2.1 Genotype population

The bean MET was conducted to collect the time-to-flowering phenotypes of a RIL population from a cross between the Andean bean cultivar, Calima, and a Mesoamerican cultivar, Jamapa (Bhakta *et al.* 2017). The Calima parent is a large-seeded, mottled bean Colombian cultivar with a determinate growth habit, while Jamapa is a small, black-seeded Mexican cultivar with an indeterminate growth habit. The RIL population was developed through single seed descent for 10 generations, followed by bulk propagation for an additional three generations ($F_{11:14}$) giving rise to 187 RILs. Further details for this RIL population

can be found in Bhakta *et al.* (2017), while the QTL-based linkage is described by Bhakta *et al.* (2015).

### 2.2 Experimental sites

We used the data from the MET study that included five locations, 187 RILs and the two parents as reported by Bhakta *et al.* (2017). The five sites had been selected to provide contrasting environmental growing conditions, especially those related to temperature and photoperiod. Three of the five sites are located in the USA: Prosper, North Dakota (ND); Citra, Florida (FL); and Isabela, Puerto Rico (PR), while the other two sites are located in Colombia: Palmira (PA) and Popayan (PO). Figure 1 and Table 1 summarize the seasonal temperature, day length and solar radiation for the five sites in the MET study, which are the main environmental variables that affect the time-to-flowering in common bean. Prosper (ND) has longer days than the other environments, while Palmira and Popayan are close to the equator and have short days. Within Colombia, Popayan,
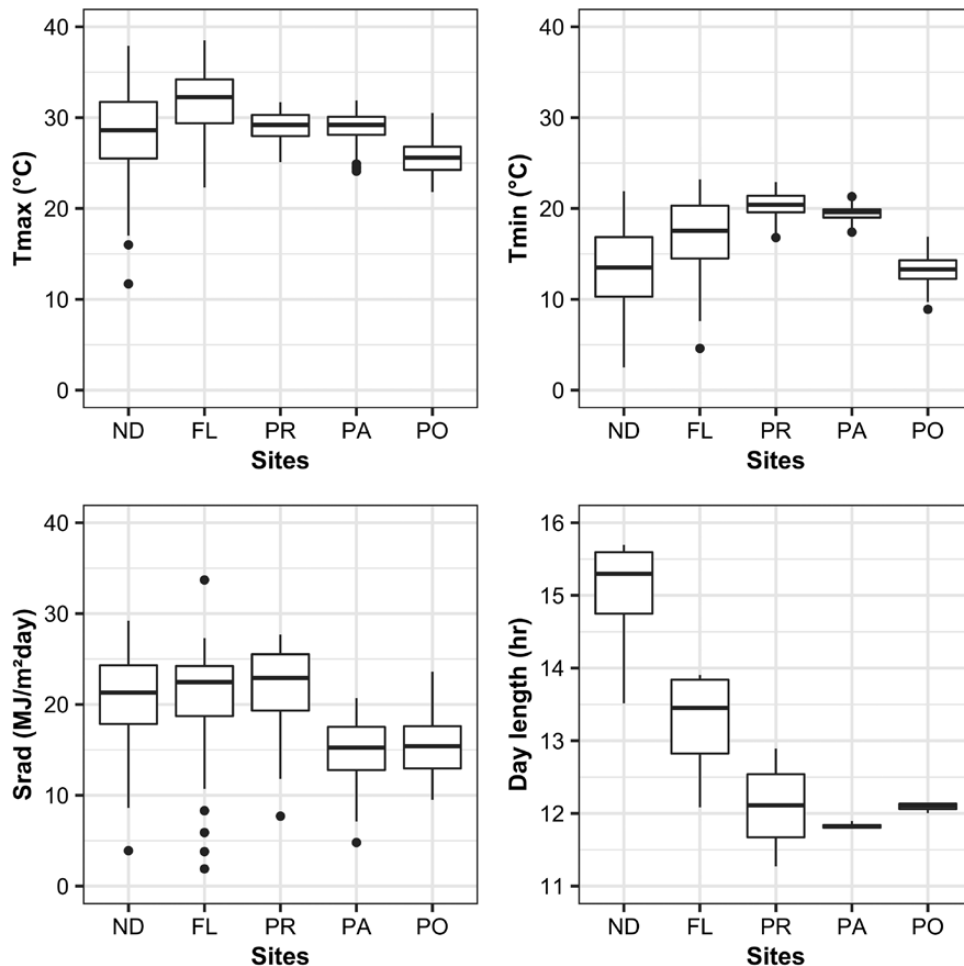


**Figure 1. Boxplots of environmental variables observed for all five sites: Prosper, North Dakota (ND); Citra, Florida (FL); Isabella, Puerto Rico (PR); Palmira, Colombia (PA); Popayan, Colombia (PO). The boxplots show the distribution of daily values of maximum temperature (top left), daily minimum temperature (top right), daily solar radiation (bottom left) and day length (bottom right). The day length for all locations is based on the calculations of the CSM-CROPGRO-Drybean model.**

the coolest site, is located at an elevation of 1800 m, while Palmira, the warmer site, is located at a 1000 m elevation. The experiment was conducted in 2011 and 2012, depending on the site. Each RIL and the two parents were grown in three replicated plots per site, with between 35 and 50 plants per plot. Six individual plants per plot were tagged at the V1 (first trifoliate opening) stage to record the vegetative and reproductive growth stages, resulting in 18 observations per genotype per site for each observation day. The plants were monitored daily to determine the date for each individual plant when first flowering occurred.

### 2.3  Dynamic mixed linear model

Vallejos *et al.* (2021) described procedures used to develop a dynamic mixed linear model to determine the rate of progress towards first flowering. This model was based on earlier work that was conducted by Bhakta *et al.* (2017) who fitted a statistical mixed linear model to predict time-to-flowering of the RILs based on QTL information and the mean environmental variables for each of the five sites. The Bhakta *et al.* (2017) model used a linear function for the effects of maximum and minimum temperature, day length and solar radiation, each averaged over the duration between sowing and first flowering, 12 QTLs, five QTL × E factors and one QTL × QTL factor. This non-dynamic model was able to describe 89 % of the observed variability among the five locations and 187 RILs, with a root mean square error (RMSE) of 2.52 days.

Vallejos *et al.* (2021) used a similar approach to that used by Bhakta *et al.* (2017) to develop their dynamic model [**see Supporting Information—Fig. S1**]. First, the time to first flower data for all RILs and environment combinations were transformed into a development rate towards first flower appearance, calculated as rate = 1/(days to first flower). This approach requires the implementation of a function that predicts the daily development rate towards the time to first flower.

We designed a new module (DMLM; see Table 2 for abbreviations) by converting the Vallejos *et al.* (2021) model into a form that could be integrated with the original CSM-CROPGRO-Drybean model. The dynamic module computes the fraction of daily progress towards flowering based on the developmental rate that is controlled by genotype and daily environmental conditions. The time-to-flowering is determined when

**Table 1.**  Summary of the meteorological and geographical data for each site in the MET as reported by Bhakta *et al.* (2017), except for the computed day length that is based on the Cropping System Model.

|  | ND | FL | PR | PA | PO |
|---|---|---|---|---|---|
| Location | Prosper, North Dakota, USA | Citra, Florida, USA | Isabella, Puerto Rico, USA | Palmira, Colombia | Popayan, Colombia |
| Latitude/longitude | 47°00′N/96°47′W | 29°39′N/82°06′W | 18°28′N/61°02′W | 03°29′N/76°81′W | 02°25′N/76°62′W |
| Elevation (m) | 280 | 31 | 128 | 1000 | 1800 |
| Earliest first flowering date | 30 June 2012 | 26 April 2012 | 6 March 2012 | 9 December 2011 | 28 April 2012 |
| Latest first flowering date | 26 July 2012 | 16 May 2011 | 22 March 2012 | 24 December 2011 | 16 May 2012 |
| Seasonal maximum temperature (°C)[a] | 27 | 32 | 29 | 28 | 25 |
| Seasonal minimum temperature (°C)[a] | 13 | 18 | 19 | 19 | 13 |
| Day length (hh:mm)[b] | 15:29 | 12:41 | 11:33 | 11:49 | 12:03 |
| Solar radiation (MJ m$^{-2}$ day$^{-1}$)[c] | 21.0 | 20.6 | 21.5 | 13.8 | 15.0 |

[a]Growing season average values for maximum and minimum temperature.
[b]Average day length data from sowing to first flower as computed by the CSM of the Decision Support System for Agrotechnology Transfer (DSSAT).
[c]Growing season average daily total solar radiation.

**Table 2.**  Description of model abbreviations.

| Module | Dynamic module description |
|---|---|
| DMLM | Dynamic mixed linear model developed by Vallejos *et al.* (2021). |
| DMLM-DL | Dynamic mixed linear module by Vallejos *et al.* (2021) using day length calculated with the CSM model |
| DPLM | Dynamic piecewise linear module (integrated into CSM-CROPGRO-Drybean model) |
| **Full crop model** | **Full crop model description** |
| DPLM-CSM | DPLM gene-based module integrated into the CSM-CROPGRO-Drybean model using QTL inputs |
| CSMG | CSM-CROPGRO-Drybean model using genetic specific coefficients (GSPs) |

the cumulative addition of the daily progress time steps reaches unity. Equation (1) shows the DMLM module that contains four environmental variables, one QTL × QTL interaction and seven QTL × E interactions.

$$
\begin{aligned}
FR_{s,g}(t) \;=\; & \mu \\
& + a_1 \cdot (DayL_s(t) - DayL_m) \\
& + a_2 \cdot (DayL_s(t) - DayL_m) \cdot QTL_{3,g} \\
& + a_3 \cdot (DayL_s(t) - DayL_m) \cdot QTL_{7,g} \\
& + a_4 \cdot (DayL_s(t) - DayL_m) \cdot QTL_{12,g} \\
& + a_5 \cdot (Srad_s(t) - Srad_m) \\
& + a_6 \cdot (Srad_s(t) - Srad_m) \cdot QTL_{12,g} \\
& + a_7 \cdot (Tmax_s(t) - Tmax_m) \\
& + a_8 \cdot (Tmax_s(t) - Tmax_m) \cdot QTL_{5,g} \\
& + a_9 \cdot (Tmin_s(t) - Tmin_m) \\
& + a_{10} \cdot (Tmin_s(t) - Tmin_m) \cdot QTL_{2,g} \\
& + a_{11} \cdot (Tmin_s(t) - Tmin_m) \cdot QTL_{3,g} \\
& + a_{12} \cdot (QTL_{1,g} \cdot QTL_{12,g}) \\
& + \sum_{q=1}^{12} \beta_q \cdot (QTL_{q,g}) \\
& + \varepsilon_{s,g,t}
\end{aligned}
\tag{1}
$$

where $FR_{s,g}(t)$ is the rate of progress to flowering (1 per day) for the *g*th genotype for the *s*th site at time *t* (in days). $\mu$ represents the overall mean value of the daily development rates across all RILs and sites in the MET data set. In the linear function (Equation (1)), RILs were treated as random effects and all remaining factors were considered as fixed effects. The variance–covariance structure that was used was unstructured, which is the default in the lme4 R package. Note that Equation (1) uses E variables that are centred on the mean values from the MET study for each RIL, based on the Vallejos *et al.* (2021) model. The first terms express the effects of the four environmental variables and the QTL-by-E effects, in which the variables $a_1$ through $a_{11}$ are estimated coefficients that quantify those effects, $DayL_s(t)$ is day length on each day *t* of the experiment at site *s*. Similarly, $Srad_s(t)$ is daily solar radiation (MJ m$^{-2}$), $Tmax_s(t)$ is daily maximum temperature (°C) and $Tmin_s(t)$ is daily minimum temperature (°C) for each day *t* at site *s*. Mean values for each environmental variable (Equation (1)) were used as constants to centre the module calculations within the observed variables. These values were calculated from sowing to first flower for each RIL and all sites in the MET data set, represented by $DayL_m$, $Srad_m$, $Tmax_m$ and $Tmin_m$. Also, $QTL_{2,g}$, $QTL_{3,g}$, $QTL_{5,g}$, $QTL_{7,g}$ and $QTL_{12,g}$ are QTLs that interact with E to affect time to first flower (Bhakta *et al.* 2017). The second part of this equation shows one QTL-by-QTL interaction ($QTL_{1,g}$ interacting with $QTL_{12,g}$); $a_{12}$ is the coefficient for this interaction. The third part of this equation includes the sum of all QTL effects, where $\beta_q$ represents the coefficient for the *q*th QTL allele effect for RIL$_g$ ($QTL_{q,g}$). Each QTL has a marker value numerically assigned according to its allelic identity; Jamapa alleles were assigned as −1 and Calima alleles as +1 values [**see Supporting Information—Table S1**].

The daily rates ($FR_{s,g}(t)$) in Equation (1) are then accumulated or integrated over time to predict day of first flower appearance using Equation (2) and a daily time step ($dt = 1$).

$$
SUMFR_{s,g}(t) = SUMFR_{s,g}(t-1) + FR_{s,g}(t) \cdot dt
\tag{2}
$$

where $SUMFR_{s,g}(t)$ integrates the flowering rate at time *t* (in days) starting on the day of planting. $SUMFR_{s,g}(t)$ is set to 0.0 at the start of the simulation, and when it reaches 1.0, first flowering is simulated to occur on that day *t* for the *g*th RIL at site *s*.

## 2.4 Dynamic piecewise linear module

The CSM-CROPGRO-Drybean model (CSMG; Table 2), which is part of the Decision Support System for Agrotechnology Transfer (DSSAT; Hoogenboom *et al.* 2019b), requires daily weather data, soil surface and profile characteristics, crop management scenarios and cultivar information (GSPs) as input (Jones *et al.* 2003; Hoogenboom *et al.* 2019a). The CSMG crop model uses daily weather variables for maximum and minimum temperature and solar radiation, and these variables have the same units as those used in the DMLM module. However, the day length (h) computed and used in the CSMG model is slightly different from the one used to develop the Bhakta and Vallejos models. The main difference is that the CSMG model accounts for the twilight period at sunrise and sunset, which may affect the photoperiod response of crops. Thus, in this study we used daily day length values computed by the CSMG model as input for the statistical procedures to estimate the numerical coefficients in Equation (1). This was done to make the daily weather and photoperiod variables identical to those used in the CSMG model (Hoogenboom *et al.* 1994; Boote *et al.* 1998) and to allow incorporation of the new dynamic gene-based module.

A second dynamic module (DMLM-DL) predicts the flowering rate (Equation (1)) on a daily basis using the day lengths from the CSMG and using a linear response to temperature. However, it is well-known that under high-temperature conditions the rate of progression towards flowering does not increase linearly with temperature (White *et al.* 2005). The response is only approximately linear over a specific range of temperatures and it plateaus as an optimum temperature is reached. In fact, the effect of temperature on development rate can be more accurately represented by a beta function (Ritchie and Nesmith 1991). Because the temperature varies considerably within a single season, with location, and over time, plants are frequently exposed to temperatures outside their linear response range. To help account for the non-linearity response of common bean under high temperatures, a third module was created that uses a dynamic piecewise linear function (referred to as the DPLM module) to ensure that the daily simulated development rate is bounded to be within the range of temperatures that were observed in the MET study and used to estimate the coefficients in Equation (1) [**see Supporting Information—Fig. S2**].

$FRMAX_g$ is defined as the rate at which the progress towards flowering proceeds when the environmental conditions, i.e. the daily maximum and minimum temperature, day length and solar radiation, are at 'optimum' values that result in a maximum development rate for the *g*th genotype. The $FRMAX_g$ values were estimated by selecting the maximum rate ($1/DUR_{s,g}$) for each RIL occurring across *s* environments, using the observed duration between planting and first flower across all *s* sites using Equation (3):

$$
FRMAX_g = \max(1/DUR_{s,g}) \; across \; all \; s \; sites
\tag{3}
$$

This resulted in 189 data points (one for each RIL plus the two parental lines). We determined whether the values for $FRMAX_g$ were affected by the same QTLs that significantly affected the time to first flower by estimating a linear relationship shown in Equation (4).

$$FRMAX_g = \gamma_g + \sum_{q=1}^{12} \rho_q \cdot (QTL_{q,g}) \qquad (4)$$

where the variable $\gamma_g$ is the fixed intercept estimated for the $g$th genotype and $\rho_q$ is the coefficient that quantifies the allelic effect of the $q$th QTL on the maximum rate of progress. A linear regression analysis was used to estimate coefficients of Equation (4).

The final daily rate of first flowering was determined using the DPLM module that was integrated into the CSMG model in which a maximum rate of development is limited depending on the genotype. If the daily flowering rate at time $t$ computed by $FR_{s,g}(t)$ exceeds $FRMAX_g$ for any RIL in the DPLM module, $FRD_{s,g}(t)$ limits the maximum rate of flowering to that set by $FRMAX_g$, Equation (5):

$$FRD_{s,g}(t) = \min(FR_{s,g}(t), FRMAX_g) \qquad (5)$$

Finally, $FRD_{s,g}(t)$ is integrated daily to predict the day when first flower occurs, Equation (6), where $SUMFRD_{s,g}(t)$ integrates the flowering rate at time $t$ (in days) in the DPLM module.

$$SUMFRD_{s,g}(t) = SUMFRD_{s,g}(t-1) + FRD_{s,g}(t) \cdot dt \qquad (6)$$

At the start of the simulation, $SUMFRD_{s,g}(t)$ is set to 0.0 and the day when it reaches or exceeds 1.0, flowering is predicted to occur for the $g$th genotype.

## 2.5 Incorporation of a gene-based module into CSM

The CSMG model (Hoogenboom *et al.* 1992, 1994) was developed using a modular structure (Jones *et al.* 2001), where overall development and growth are represented by specific modules, including those for vegetative and reproductive development, photosynthesis, respiration, partitioning, vegetative and reproductive growth, and other soil and crop processes (Hoogenboom *et al.* 2019a). For this study the focus was on the phenology module of CROPGRO where the developmental and phenological phase transitions are implemented. Boote *et al.* (1998) described the physiological development rate in CROPGRO as a function of temperature, photoperiod and water deficit. If these conditions are optimal, one physiological day is accumulated per calendar day. The phenology module in CROPGRO separates the vegetative and reproductive routines that calculate the stages and individual phase durations.

To incorporate the first flower development stage using $SUMFRD_{s,g}(t)$ in the CSMG model, we first developed a new gene-based module to create a link between the DPLM module and the crop model (Fig. 4). This module connects daily input data, the DPLM module and the CSMG phenology module. The inputs for this DPLM module consist of weather data from the crop model and the 12 QTL allelic make-up for each RIL (or genotype). The input QTL data for our study were those for the 187 RILs plus the two parent cultivars, which are processed in a new QTL data subroutine inside the gene-based module. A new input file was created for the CSMG model, named BNGRO047.

GEN that contains QTL data for each of the RILs and their two parent cultivars [see **Supporting Information—Table S1**].

The daily weather and QTL data for a particular site and RIL are inputs for the DPLM module, enabling it to simulate the daily flowering rate as affected by G and E conditions. The integrated development progress to first flower, $SUMFRD_{s,g}(t)$ and the day when first flowering occurs are passed back to the CSMG phenology routine. The outputs from the gene-based module are inputs to the reproductive stage component, where the variables associated with first flowering are calculated. The day when first flowering occurs is set and afterward, progress for subsequent development phases is computed using the original CSMG model.

## 2.6 Sensitivity analysis of simulated variation for G × E

A simulation analysis was performed using the DPLM module to explore all possible combinations of the 12 QTL variation among RILs using the daily weather data across all five sites of the MET study, similar to previous ideotype studies (White *et al.* 2005). This resulted in a total of 4096 ($2^{12}$) RILs. The coefficients estimated using 187 RILs plus the two parents were used to simulate the number of days to flowering for the 4096 RILs. The input file BNGRO047.GEN containing QTL information was revised by adding inputs for each of the 4096 RILs. Crop management including the planting dates and the daily weather data were assumed to be same as for the original five environments of the MET study. The management input file assumes that only the variation in genetics and environments affect the simulated responses, representing potential production for each line. The DPLM module was then used to conduct the 20 480 unique simulations across sites and synthetic RILs.

The time to first flower responses of the DPLM module were compared with those of the DMLM-DL module to determine how the addition of a maximum rate, $FRMAX_g$ affects the simulated results under different high-temperature scenarios in a sensitivity analysis using the 187 RILs and two parents across all five sites of the MET study. The original daily temperature data were used as the base line inputs. Then, both the minimum and maximum temperatures for each day and each site were incremented at a 1 °C increment to create five different temperature scenarios (base, base + 1, base + 2, base + 3 and base + 4 °C), assuming that crop management, daily solar radiation and day length were the same as for the original MET study. The simulated number of days to flowering were analysed and compared for the DPLM and DMLM-DL modules using statistics and a visualization of the distributions of number of days to first. Although we did not have sufficiently high temperatures in the MET study to account for the decrease in development rate as the temperature increases above an optimal threshold, the simple addition of the maximum rate in the piecewise linear module is expected to provide reliable simulations for small increases in temperature.

## 2.7 Yield prediction

An ultimate goal of dynamic crop simulation models is to be able to predict yield. Therefore, we compared the performance of the original CSMG model using GSPs with the DPLM module integrated into CSM-CROPGRO-Drybean model (DPLM-CSM; Table 2). All 18 GSPs were available for 144 genotypes, including 142 RILs and the

parent material, except for PA, for which GSPs for only 143 genotypes were available. The procedures for estimating the GSPs were described by Acharya *et al.* (2017). Daily simulations, starting at planting and continuing until harvest maturity was predicted, were conducted for all five sites for either 2011 or 2012, depending on the MET. Crop management and local weather and soil data based on the original MET study were used as input for the CSMG model (Fig. 1). For the DPLM-CSM hybrid model, the flowering dates were predicted based on the DPLM module using the QTL information as input, rather than the GSPs, while for the other growth and development processes, the GSPs for the individual genotypes were used as input.

## 2.8 Model evaluation

To estimate parameters for the QTL-based modules, we used the *lmer* function of the *lme4* package (Bates *et al.* 2015) of the R programming language (version 3.6.1). To compare the performance of the modules with observed data, we used the estimated parameters for the final QTL-based DPLM module for each site. As a measure of fit of Equation (1) to the data, we used the RMSE, defined as

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (7)$$

where $y_i$ and $\hat{y}_i$ are the $i$th observed and simulated number of days to flowering, respectively, and $n$ is the number of measurements summed for all values for all RILs and for each site and all sites combined. The adjusted $R^2$ was calculated because it indicates module performance adjusted by the number of the terms in the module, defined as

$$R^2_{Adjusted} = 1 - \left[\frac{(1 - R^2)(n-1)}{(n-k-1)}\right] \qquad (8)$$

where $R^2$ represents the coefficient of determination, $n$ is the number of measurements and $k$ is the number of independent variables of the model. A Nash and Sutcliffe (1970) skill score was also used as a measure of model error, referred to as model efficiency (ME) (Wallach *et al.* 2019) and defined as

$$ME = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \qquad (9)$$

If ME = 1.0, the model fits perfectly, and the observed values are equal to the simulated values ($y_i = \hat{y}_i$) for each $i$ and ME = 1. If ME is less than 0.0, the mean of the observed data is a better predictor of the data than the model. If the variance for the observed minus predicted values is equal to the variance of observations from its mean value, then ME = 0.0, which means that the model is not good because it is no better than using the average of observed values to predict responses. For evaluation of the predictive ability of the modules, we also compared the contributions to prediction error caused by model bias and standard deviation. We used the decomposition of the mean square error (MSE) into bias, standard deviation differences and residual errors that was developed by Kobayashi and Salam (2000), Equation (10).

$$MSE = (Bias)^2 + SDSD + LCS \qquad (10)$$

with

$$Bias^2 = \left[\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - \hat{y}_i)\right]^2$$

$$SDSD = (SD_s - SD_m)^2$$

$$LCS = 2SD_s SD_m (1 - r)$$

The first term of Equation (10) is the bias squared, the second term *SDSD* is related to the difference between the simulation standard deviation ($SD_s$) and the standard deviation of the measurements ($SD_m$). The third term, *LCS*, indicates the remaining MSE error that is not accounted for by bias or standard deviation.

## 3. RESULTS AND DISCUSSION
### 3.1 Dynamic mixed linear module coefficients

The $FR_{s_g}(t)$ function is shown below with estimated parameter values for the DMLM-DL module. The values of coefficients in Equation (11) are based on the influence of day length, solar radiation, temperature and 12 QTL alleles for each RIL, showing G, E, G × E and G × G interaction effects.

$$
\begin{aligned}
FR_{s_g}(t) = {} & 2.35148 \times 10^{-2} \\
& - 1.56357 \times 10^{-3} \cdot (DayL_s(t) - DayL_m) \\
& - 7.66441 \times 10^{-4} \cdot ((DayL_s(t) - DayL_m) \cdot QTL_{3,g}) \\
& - 1.62459 \times 10^{-4} \cdot ((DayL_s(t) - DayL_m) \cdot QTL_{7,g}) \\
& - 1.45956 \times 10^{-4} \cdot ((DayL_s(t) - DayL_m) \cdot QTL_{12,g}) \\
& - 8.50211 \times 10^{-5} \cdot (Srad_s(t) - Srad_m) \\
& - 3.06273 \times 10^{-5} \cdot ((Srad_s(t) - Srad_m) \cdot QTL_{12,g}) \\
& + 5.72311 \times 10^{-4} \cdot (Tmax_s(t) - Tmax_m) \\
& + 8.54093 \times 10^{-5} \cdot ((Tmax_s(t) - Tmax_m) \cdot QTL_{5,g}) \\
& + 5.29789 \times 10^{-4} \cdot (Tmin_s(t) - Tmin_m) \\
& - 2.40896 \times 10^{-5} \cdot ((Tmin_s(t) - Tmin_m) \cdot QTL_{2,g}) \\
& - 8.59042 \times 10^{-5} \cdot ((Tmin_s(t) - Tmin_m) \cdot QTL_{3,g}) \\
& + 3.01579 \times 10^{-4} \cdot (QTL_{1,g} \cdot QTL_{12,g}) \\
& + 9.41278 \times 10^{-4} \cdot (QTL_{1,g}) \\
& + 1.24887 \times 10^{-3} \cdot (QTL_{2,g}) \\
& - 6.08364 \times 10^{-4} \cdot (QTL_{3,g}) \\
& + 2.36803 \times 10^{-4} \cdot (QTL_{4,g}) \\
& + 5.67194 \times 10^{-6} \cdot (QTL_{5,g}) \\
& + 5.27617 \times 10^{-4} \cdot (QTL_{6,g}) \\
& - 4.11459 \times 10^{-4} \cdot (QTL_{7,g}) \\
& - 2.11983 \times 10^{-4} \cdot (QTL_{8,g}) \\
& - 4.42610 \times 10^{-4} \cdot (QTL_{9,g}) \\
& - 2.50138 \times 10^{-4} \cdot (QTL_{10,g}) \\
& + 3.43389 \times 10^{-4} \cdot (QTL_{11,g}) \\
& - 1.53677 \times 10^{-4} \cdot (QTL_{12,g})
\end{aligned}
$$

$$(11)$$

where the first term $(2.35148 \times 10^{-2})$ is the overall average daily rate of progress, indicating that the average time between sowing and appearance of first flowering is 42.5 days $(=1/(2.35148 \times 10^{-2}))$. The first coefficient $(a_1 = -1.56357 \times 10^{-3})$ is the sensitivity to day length, indicating that a 1-h increase in day length would result in a rate of development that is $1.56357 \times 10^{-3}$ below the average rate of $2.35148 \times 10^{-2}$. This 1-h increase in day length simulates that the time to first flower would occur 45.6 days after planting, an increase of 3.1 days compared to the average days to first flower that was observed across the five sites and 187 RILs plus the two parents. This rate of development also varies as a function of QTL alleles, which can increase or decrease the rate resulting in a decrease or increase in the number of days to first flower, respectively. Note that some of the QTL coefficients in Equation (11) have a negative sign while others have a positive sign. This is because each parental genotype has both types of alleles; the allele operator, i.e. Calima = +1 and Jamapa = −1,

will alter the sign of the coefficient accordingly [see **Supporting Information—Table S1**]. The estimated parameter terms with the 2.5 % and 97.5 % confidence intervals, *P*-value and the variance components are shown in the **Supporting Information—Table S2**. The fixed effects variance was $1.80182 \times 10^{-5}$, the random effects variance was $6.34775 \times 10^{-7}$ and the residual variance was $1.3056 \times 10^{-6}$.

Next, we compared the agreement between simulated and observed results for all sites, RILs and parents using the DMLM and DMLM-DL modules (Fig. 2A and B; Table 3). Comparisons of RMSE between simulated and observed values showed that the errors were only slightly different between the two modules (Table 3; DMLM and DMLM-DL modules). When all sites and RILS were included in the comparisons, the RMSE values were 2.73 days and 2.72 days for the DMLM and DMLM-DL, respectively. Similarly, when comparing agreements for each site, the RMSE values using the two module versions were within 0.02 days for ND and 0.04 days for FL. Notably, however, Table 3 shows
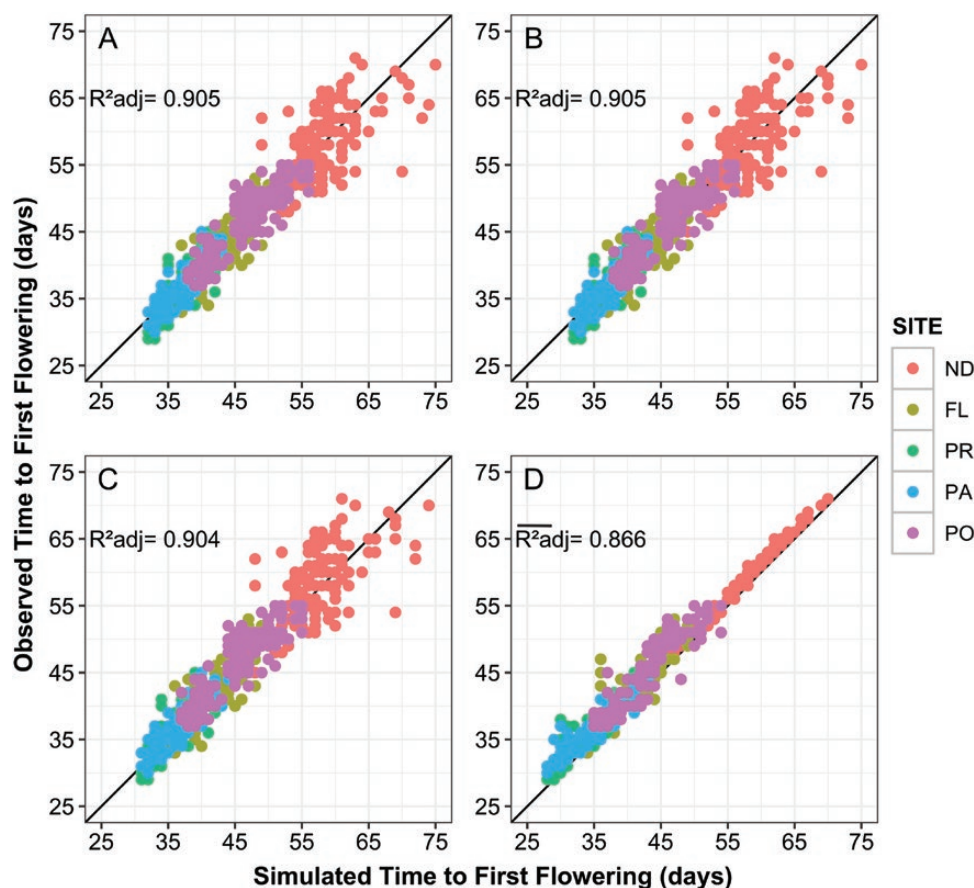


**Figure 2.** Observed versus simulated time to first flower across all five sites for the dynamic mixed linear module (DMLM) (A); the dynamic mixed linear model using the day length computed by the crop module (DMLM-DL) (B); the dynamic piecewise linear module incorporated into CSM-CROPGRO-Drybean (DPLM) (C); and the original CSM-CROPGRO-Drybean model using genetic specific coefficients (CSMG) (D). For A, B and C, the modules simulated for each RIL and for all sites, while for D the simulations were conducted based on the genotype-specific coefficients based on Acharya *et al.* (2017). Each point represents an observed and simulated RIL; the solid 1:1 diagonal line represents equal values for time to first flower. $R^2_{adj}$ for graph D is the average of the values across all five sites for each RIL. The five experimental sites are: Prosper, North Dakota (ND); Citra, Florida (FL); Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).

**Table 3. Measures of agreement between simulated and observed number of days from planting to first flower for all models for each individual site and for all sites combined.**

| Site[a] | Measures of agreement[b] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Genotypes (#) | Observed mean | Simulated mean | Bias | RMSE | ME | MSE | (Bias)$^2$ | SDSD | LCS |
| | | | DMLM[c] | | | | | | | |
| ND | 149 | 57.74 | 58.36 | −0.62 | 4.58 | 0.30 | 21.08 | 0.38 | 0.04746 | 20.65 |
| FL | 170 | 42.46 | 42.96 | −0.50 | 2.51 | 0.72 | 6.32 | 0.25 | 1.02153 | 5.05 |
| PR | 163 | 36.42 | 36.88 | −0.45 | 1.97 | 0.70 | 3.89 | 0.21 | 0.62889 | 3.05 |
| PA | 173 | 36.65 | 37.15 | −0.50 | 1.62 | 0.72 | 2.63 | 0.25 | 0.02306 | 2.36 |
| PO | 173 | 45.96 | 46.15 | −0.19 | 2.26 | 0.81 | 5.16 | 0.04 | 0.09782 | 5.02 |
| All sites | 828 | 43.54 | 43.99 | −0.45 | 2.73 | 0.90 | 7.45 | 0.20 | 0.04168 | 7.20 |
| | | | DMLM-DL | | | | | | | |
| ND | 149 | 57.74 | 58.32 | −0.57 | 4.56 | 0.30 | 20.90 | 0.33 | 0.09217 | 20.48 |
| FL | 170 | 42.46 | 43.03 | −0.56 | 2.55 | 0.71 | 6.53 | 0.32 | 1.11594 | 5.10 |
| PR | 163 | 36.42 | 36.88 | −0.45 | 1.97 | 0.70 | 3.89 | 0.21 | 0.58810 | 3.09 |
| PA | 173 | 36.65 | 37.18 | −0.53 | 1.61 | 0.73 | 2.61 | 0.28 | 0.01977 | 2.31 |
| PO | 173 | 45.96 | 46.08 | −0.12 | 2.23 | 0.81 | 5.02 | 0.01 | 0.10511 | 4.90 |
| All sites | 828 | 43.54 | 43.98 | −0.44 | 2.72 | 0.90 | 7.42 | 0.20 | 0.05778 | 7.17 |
| | | | DPLM | | | | | | | |
| ND | 149 | 57.74 | 57.37 | 0.38 | 4.55 | 0.30 | 20.82 | 0.14 | 0.11116 | 20.57 |
| FL | 170 | 42.46 | 42.14 | 0.32 | 2.49 | 0.72 | 6.22 | 0.10 | 0.96639 | 5.15 |
| PR | 163 | 36.42 | 36.01 | 0.41 | 1.93 | 0.71 | 3.75 | 0.17 | 0.44665 | 3.13 |
| PA | 173 | 36.65 | 36.36 | 0.29 | 1.56 | 0.74 | 2.44 | 0.08 | 0.00026 | 2.36 |
| PO | 173 | 45.96 | 45.07 | 0.89 | 2.41 | 0.78 | 5.84 | 0.79 | 0.10968 | 4.94 |
| All sites | 828 | 43.54 | 43.08 | 0.46 | 2.73 | 0.90 | 7.46 | 0.21 | 0.06855 | 7.17 |
| | | | CSMG | | | | | | | |
| ND | 144 | 57.59 | 56.35 | 1.24 | 1.38 | 0.94 | 1.90 | 1.545 | 0.00366 | 0.35 |
| FL | 144 | 41.85 | 40.47 | 1.38 | 2.20 | 0.76 | 4.85 | 1.891 | 0.41912 | 2.54 |
| PR | 144 | 36.31 | 34.32 | 1.99 | 2.30 | 0.58 | 5.31 | 3.972 | 0.01996 | 1.32 |
| PA | 143 | 36.29 | 34.87 | 1.43 | 1.95 | 0.56 | 3.83 | 2.035 | 0.38940 | 1.41 |
| PO | 144 | 45.54 | 43.27 | 2.27 | 3.00 | 0.66 | 9.03 | 5.157 | 0.47913 | 3.40 |
| All sites | 719 | 43.53 | 41.87 | 1.66 | 2.23 | 0.94 | 4.98 | 2.762 | 0.00011 | 2.21 |

[a]Prosper, North Dakota (ND); Citra, Florida (FL); Isabella, Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).
[b]ME = Nash and Sutcliffe (1970) model efficiency; MSE = mean squared error; (Bias)$^2$, SDSD and LCS present the decomposition of MSE.
[c]DMLM = dynamic mixed linear module; DMLM-DL = dynamic mixed linear module with day length from the CSM model; DPLM = dynamic piecewise linear module as integrated in CSM-CROPGRO-Drybean for flowering prediction; CSMG = original CSM-CROPGRO-Drybean model using GSPs.

relatively large differences in RMSE depending on site, with ND having the largest RMSE or 4.58 days in comparison with the lowest RMSE of 1.61 days for PA. We attributed these differences to the fact that the MET did not include a site with long days and low temperatures to contrast the long days and high temperatures of ND, which did not adequately capture the temperature–day length interactions previously documented by Wallace and Enriquez (1980) and Wallace *et al.* (1991).

The comparison of ME between the two module versions (DMLM and DMLM-DL) showed that both have the same high ME value of 0.90. Although ME values for the ND site were lower (0.30 for both modules), these positive numbers indicate that the modules are more effective than using the mean value of the observations. Table 3 also shows that the bias in simulating time-to-flowering was low across all sites (less than 0.5 days), and MSE values were low except for ND. The remaining error after accounting for bias and standard deviation differences was much larger for both module versions at ND than for any of the other sites. Overall, the module implementation using the CSMG-computed day lengths (DMLM-DL) showed that the agreement indicators were only slightly different from the DMLM module using the day lengths from Bhakta *et al.* (2017).

### 3.2 Dynamic piecewise linear module

The highest maximum rate for any genotype in the MET data set was 0.0345 and the lowest observed maximum rate for any genotype was 0.0222. This means that the duration from planting to first flower varied from 29 to 45 days among genotypes under optimal environmental conditions. These maximum rates of development occurred at the tropical PA and PR locations where temperatures were warm and day lengths were relatively short. Although environmental conditions may not have been optimal at these locations, most of the maximum rates across locations for any RIL occurred in PA; only a few occurred in PR where the maximum rates for some RILs were only slightly higher than in PA. These results could likely be improved by using other data sets, ideally under more controlled environmental conditions.
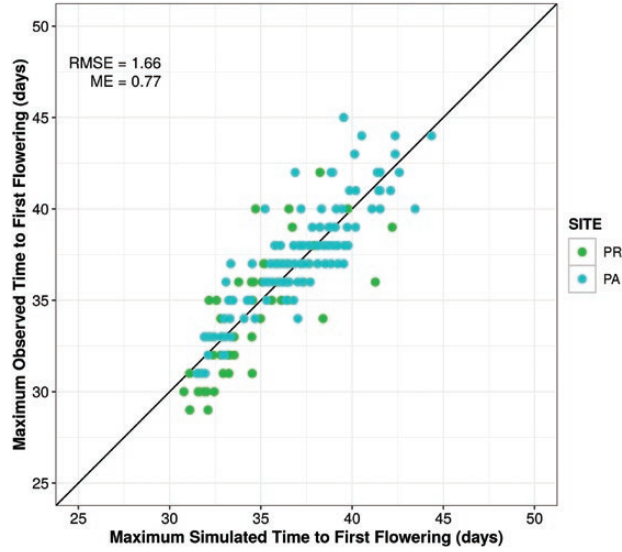
**Figure 3. Maximum observed versus simulated time to first flowering for each RIL across all five sites based on a linear model of maximum development rates dependent on the 12 QTLs alleles for the 187 RIL plus the two parental lines. RMSE = root mean square error; ME = model efficiency (Nash and Sutcliffe 1970). The solid 1:1 diagonal line represents equal values of maximum simulated/observed time to first flowering.**

The main purpose of Equation (12) is to prevent predictions of excessively high values for the development rate that could lead to unrealistically low predictions for the number of days to first flower appearance under environmental conditions with a high temperature, a short day length and a high solar radiation values that are likely to occur in many environments. This equation was incorporated in the DPLM module and integrated into the full DPLM-CSM hybrid model.

$$
\begin{aligned}
FRMAX_g \;=\; & 2.79856 \times 10^{-2} \\
+ \; & 1.07126 \times 10^{-3} \cdot QTL_{1,g} \\
+ \; & 1.24937 \times 10^{-3} \cdot QTL_{2,g} \\
- \; & 3.53505 \times 10^{-4} \cdot QTL_{3,g} \\
+ \; & 3.99455 \times 10^{-4} \cdot QTL_{4,g} \\
+ \; & 7.08516 \times 10^{-5} \cdot QTL_{5,g} \\
+ \; & 5.63062 \times 10^{-4} \cdot QTL_{6,g} \\
- \; & 4.28549 \times 10^{-4} \cdot QTL_{7,g} \\
- \; & 2.35099 \times 10^{-4} \cdot QTL_{8,g} \\
- \; & 6.09052 \times 10^{-4} \cdot QTL_{9,g} \\
- \; & 2.97020 \times 10^{-4} \cdot QTL_{10,g} \\
+ \; & 6.35384 \times 10^{-4} \cdot QTL_{11,g} \\
- \; & 2.31698 \times 10^{-4} \cdot QTL_{12,g}
\end{aligned} \tag{12}
$$

The fitting of $FRMAX_g$ using Equation (12) resulted in predicted caps on the rate of progress for the 187 RILs plus the two parents of our data set with a RMSE of 1.66 days, ME of 0.77 days and MSE of 2.78 (Fig.
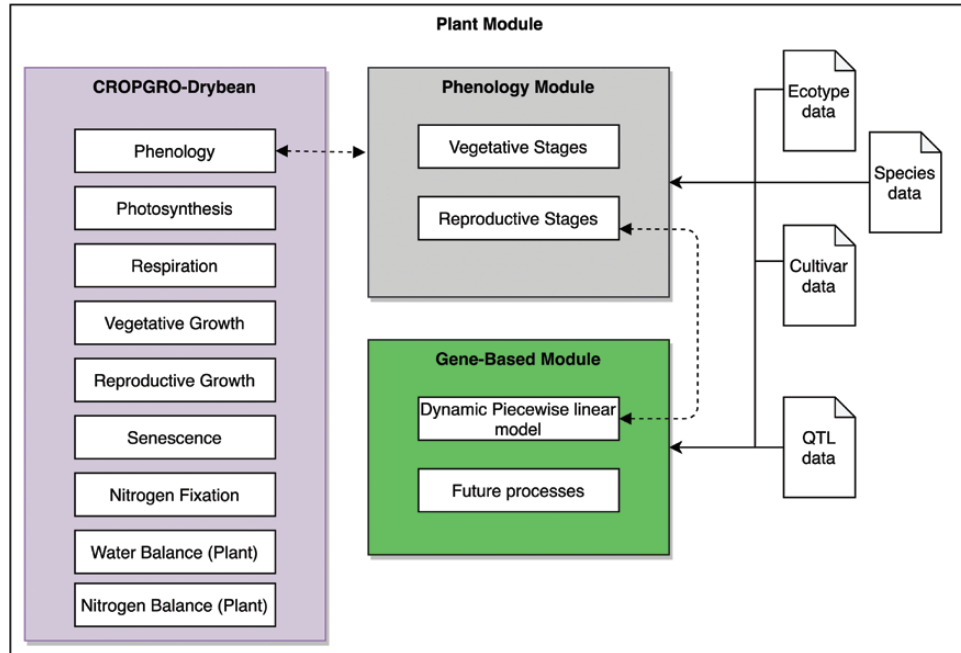


**Figure 4. Overview of the DPLM-CSM model developed to integrate the CSM-CROPGRO-Drybean model (CSMG) with the dynamic piecewise linear module (DPLM) using a new gene-based module. The DPLM simulates the first time of first flowering module developed from the dynamic mixed linear model first developed by Vallejos *et al.* (2021). The integrated model uses QTL data, which contains the 12 QTL allele information to simulate the daily rate of development towards first flowering, in addition to the other input data used by the original CSMG.**

3). These values indicate that the maximum developmental rates were affected by the genetic factors (12 QTLs).

### 3.3 Structural changes of the CSM-CROPGRO-Drybean model

The new gene-based module operates on a daily time step in the DPLM-CSM phenology module to simulate the rate of development towards first flowering for a particular RIL or cultivar and for a specific site, as shown in Fig. 4. This gene-based module incorporates the DPLM module and processes the QTL data obtained from the revised BNGRO047.GEN file, while weather data are passed to the DPLM

module from the CSMG routines. When the value of $SUMFRD_{s_g}(t)$ reaches 1.0 (Equation (6)), the day of first flower is simulated to occur and this date is passed back to the phenology module for its use in updating first flowering in the reproductive development module.

The DPLM module was designed to be flexible, operating in parallel with the original CSMG using GSPs. This allows the CSMG model to work in a hybrid mode using either the original cultivar coefficients or the QTL input data to simulate the development of first flowering. Regardless, all other stages in the DPLM-CSM model are simulated using inputs from the original cultivar coefficient file. This option was added as a new switch in the crop management input file (FileX).
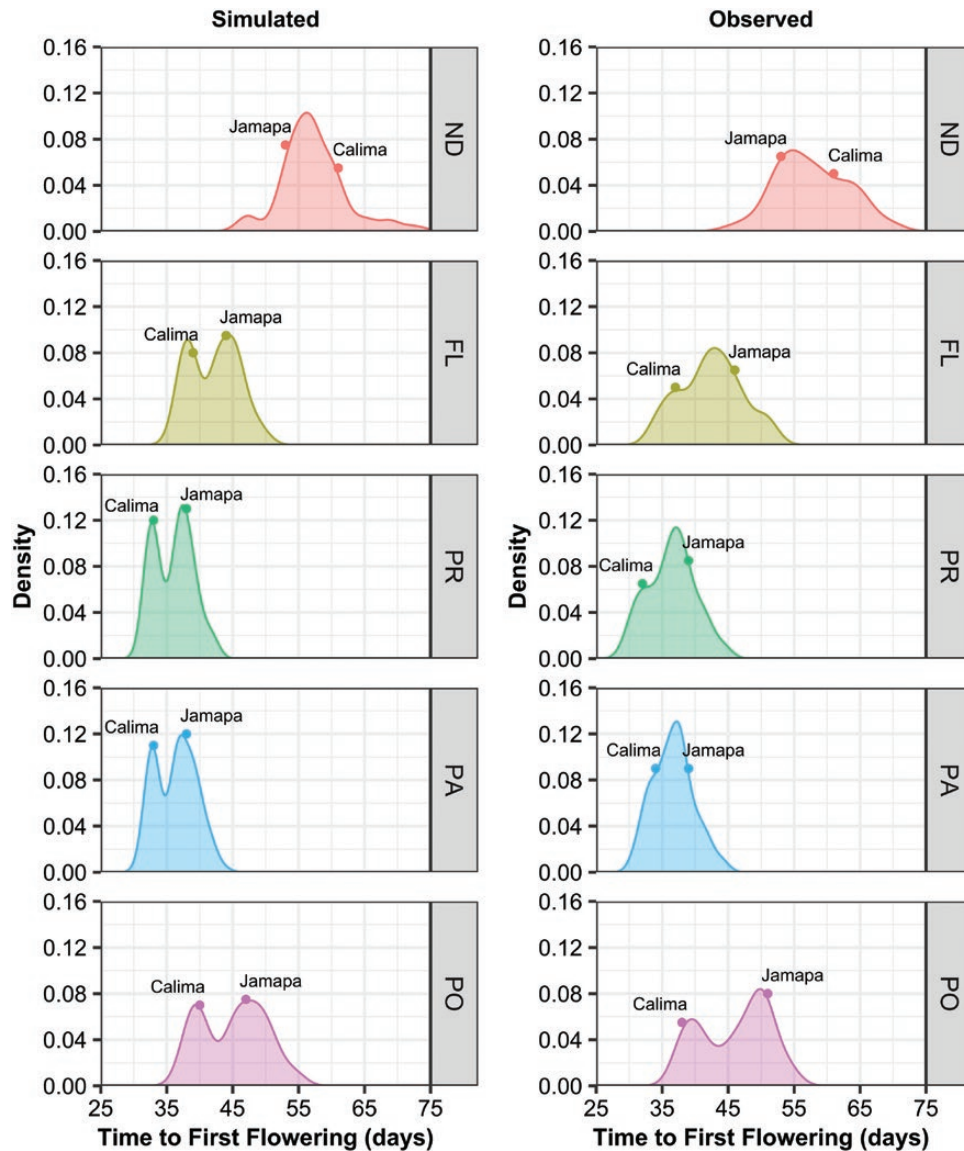


**Figure 5. Density plots of time to first flower in days across five sites.** Distribution of simulated time to first flower using the dynamic piecewise linear module (DPLM) (left panel) and the distribution of observed time to first flower (right panel). The parental lines Jamapa and Calima are highlighted at the top of each distribution. The five experimental sites are: Prosper, North Dakota (ND); Citra, Florida (FL); Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).

When this switch is set to 'Y' the DPLM-CSM model uses the DPLM module and the QTL input data to simulate the time of first flower. Otherwise if the switch is set to 'N', the DPLM-CSM model uses the original GSPs for all phenological development stages, including the prediction of flowering, and the DPLM module is ignored. These changes do not affect any other phenological processes in the crop growth model. In this way, additional dynamic gene-based modules can be added to the gene-based module to simulate other vegetative and reproductive processes.

## 3.4 Comparing simulated and observed frequency distributions of time-to-flower

The simulated and observed frequency distributions of days between sowing and first flower are presented in Fig. 5 for the DPLM module simulations (left panel) and for observed data from the MET study (right panel). The shapes of the simulated distributions appeared to be bimodal for all locations except for ND where it showed a distribution close to normal. The distributions for the observed data did not exhibit bimodal characteristics, except for the PO site, which had cooler temperatures than the other sites. QTL2, which is associated with the growth habit gene *Fin*, shows interaction with $T_{min}$ and is likely responsible for this bimodality (Bhakta *et al.* 2017). Also, on average, the indeterminate growth habit RILs generally flowered later than the determinate growth habit RILs. These graphs showed that Calima flowered earlier than Jamapa except for the ND site. The time-to-flowering pattern of the two parents was captured by the module. Bhakta *et al.* (2017) detected this transgressive behaviour of some RILs, those flowering earlier or later than the parents, a phenomenon explained by the presence of genes that accelerate development and others that retard development in both parents.

Table 3 shows a comparison using various measures of agreement between the simulated and observed data for DMLM, DMLM-DL, DPLM and CSMG. The simulation results of the DPLM module displayed a strong agreement between the simulated and observed time to first flower (Fig. 2C; Table 3). Simulated results showed an average bias of 0.55, a RMSE of 2.73 days, a ME of 0.90, a MSE of 7.46 and an adjusted $R^2$ of 0.905. The differences between the simulated and observed values were larger for ND than for the other sites. The average bias for ND was 0.38, the RMSE was 4.55 days and the ME was 0.30, whereas the corresponding values for the other four sites showed a much closer agreement between the simulated and observed days to first flower. Comparisons of these agreement indicators with those for the DMLM and DMLM-DL modules showed nearly identical bias, RMSE and ME values, demonstrating that the implementation of the DPLM module provided simulated results that were nearly identical to the other two module versions listed in Table 3.

The original CSMG mostly produced simulated days to first flower that were in closer agreement with observed results across all sites than the other modules (Table 3; Fig. 2D). However, these results are misleading in that the GSPs that produced these results were estimated for each individual RIL, which means that only five data points were used to estimate three GSPs for each RIL, and thus the agreements were forced in the GSP estimation process. The adjusted $R^2$ was calculated using five parameters for each RIL; the 113 RILs that had

observations for all five sites resulted in estimating 339 GSP parameters. The adjusted $R^2$ averaged for the RILs was 0.866, ranging from 0.321 to 0.997 with a standard deviation of 0.129. However, note that the adjusted $R^2$ values were lower than all of those for the gene-based modules for each site except at ND. As Acharya *et al.* (2017) point out, the estimated GSPs were highly uncertain and that different combinations of the GSPs could provide the same fit to observed data (showing equifinality in the estimation process) such that the GSP estimates are not reliable even though they can nearly reproduce the data. Estimating three parameters with only five data points, then repeating this process for each of the RILs, results in estimates that reliably reproduce the data used to estimate them but should not be interpreted as values that can be used for other environments or genotypes. Estimation of these coefficients also requires considerable effort and resources, which has to be repeated every time a new cultivar is released. Instead statistical gene-based modules can estimate independently phenotypic traits using as input G, E and G × E interactions data. By contrast, estimating the 25 coefficients in the dynamic linear module (Equation (11)) used all data across the five sites and 189 (RILs plus parents); thus, 945 observations were used to estimate 25 coefficients. Therefore, using
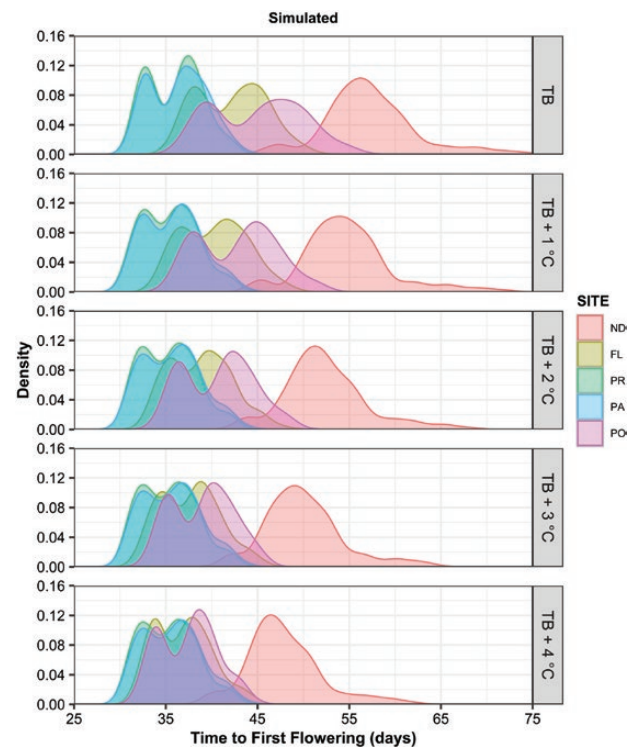


Figure 6. Density plots of distributions for simulated time to first flower (in days) using the dynamic piecewise linear module (DPLM). Simulated days between planting to first flower shows the responses to increasing the base maximum and minimum temperature from 1 °C through 4 °C. The five experimental sites are: Prosper, North Dakota (ND); Citra, Florida (FL); Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).

the dynamic mixed linear module estimation process has potential for a more robust use of the module across environments and genotypes, especially for a new genotype that has QTL information but does not have field phenotype data.

The frequency distributions associated with genetic variation in the RIL population for simulated time to first flower at each site and for the five temperature scenarios are shown in Fig. 6. The comparisons of the means and standard deviations of the populations for each four-temperature/site combination are summarized in Table 4. These

results demonstrate the effects of including the maximum rate of development ($FRMAX_g$) for each genotype in the DPLM module for comparison with the module without this upper limit. The largest differences in simulated days to first flower occurred when the temperature was increased by 4 °C (Table 4) at sites with higher temperatures. For example, for the PR site, increasing the daily $T_{min}$ and $T_{max}$ values by 4 °C only decreased the mean days to first flower by 0.6 days for the DPLM module, whereas the increase in temperature by 4 °C unrealistically decreased the mean days to first flower by 5.1 days for the

**Table 4. Temperature sensitivity analysis for the simulated number of days to first flower for the dynamic piecewise linear module (DPLM) and the dynamic mixed linear module with CSM-CROPGRO-Drybean day length (DMLM-DL) using the original weather data from the five sites.**

| Site[a] | DPLM | | | | DMLM-DL | | | |
|---|---|---|---|---|---|---|---|---|
| | Simulated mean | Min[b] | Max[b] | Standard deviation | Simulated mean | Min | Max | Standard deviation |
| | | | | Base temperature | | | | |
| ND[a] | 57.37 | 46 | 76 | 5.14 | 58.32 | 47 | 77 | 5.17 |
| FL | 42.14 | 35 | 51 | 3.74 | 43.03 | 36 | 51 | 3.67 |
| PR | 36.01 | 31 | 43 | 2.94 | 36.88 | 32 | 43 | 2.85 |
| PA | 36.36 | 31 | 44 | 3.07 | 37.18 | 32 | 43 | 2.94 |
| PO | 45.07 | 37 | 55 | 4.87 | 46.08 | 38 | 56 | 4.88 |
| All sites | 43.08 | 31 | 76 | 8.56 | 43.98 | 32 | 77 | 8.58 |
| | | | | Base temperature + 1 °C | | | | |
| ND | 54.74 | 45 | 72 | 4.81 | 55.66 | 46 | 73 | 4.83 |
| FL | 40.29 | 34 | 49 | 3.53 | 41.04 | 35 | 49 | 3.38 |
| PR | 35.48 | 30 | 43 | 2.97 | 35.40 | 31 | 41 | 2.55 |
| PA | 35.77 | 30 | 44 | 3.10 | 35.69 | 31 | 41 | 2.65 |
| PO | 42.87 | 36 | 52 | 4.20 | 43.86 | 37 | 53 | 4.21 |
| All sites | 41.54 | 30 | 72 | 7.77 | 42.03 | 31 | 73 | 8.02 |
| | | | | Base temperature + 2 °C | | | | |
| ND | 52.37 | 43 | 68 | 4.49 | 53.13 | 44 | 69 | 4.58 |
| FL | 38.81 | 33 | 48 | 3.36 | 39.29 | 34 | 47 | 3.07 |
| PR | 35.39 | 30 | 43 | 2.99 | 34.09 | 30 | 39 | 2.33 |
| PA | 35.69 | 30 | 44 | 3.13 | 34.29 | 30 | 40 | 2.41 |
| PO | 40.81 | 34 | 49 | 3.78 | 41.80 | 35 | 50 | 3.78 |
| All sites | 40.34 | 30 | 68 | 6.97 | 40.24 | 30 | 69 | 7.51 |
| | | | | Base temperature + 3 °C | | | | |
| ND | 50.03 | 42 | 64 | 4.26 | 50.70 | 42 | 65 | 4.39 |
| FL | 37.67 | 32 | 46 | 3.21 | 37.68 | 33 | 45 | 2.71 |
| PR | 35.36 | 30 | 43 | 3.00 | 32.85 | 29 | 38 | 2.14 |
| PA | 35.68 | 30 | 44 | 3.13 | 33.07 | 29 | 38 | 2.20 |
| PO | 39.03 | 33 | 46 | 3.31 | 40.01 | 34 | 47 | 3.30 |
| All sites | 39.31 | 30 | 64 | 6.21 | 38.60 | 29 | 65 | 7.00 |
| | | | | Base temperature + 4 °C | | | | |
| ND | 47.91 | 40 | 62 | 4.08 | 48.48 | 41 | 63 | 4.14 |
| FL | 36.88 | 32 | 45 | 3.06 | 36.21 | 32 | 43 | 2.46 |
| PR | 35.36 | 30 | 43 | 3.00 | 31.77 | 28 | 36 | 1.96 |
| PA | 35.68 | 30 | 44 | 3.13 | 31.95 | 28 | 37 | 2.01 |
| PO | 37.50 | 32 | 45 | 3.12 | 38.34 | 33 | 45 | 3.01 |
| All sites | 38.44 | 30 | 62 | 5.57 | 37.10 | 28 | 63 | 6.54 |

[a]Prosper, North Dakota (ND); Citra, Florida (FL); Isabella, Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).
[b]Minimum/maximum simulated number of days from sowing to first flower.

DMLM-DL module. In contrast, results for the cooler sites (PO and ND) were similar for both module versions. The frequency distributions (Fig. 6) visually demonstrate the effect of the $FRMAX_g$ on days to first flower. The distributions for PO and ND shifted to the left for each temperature increase of 1 °C, indicating a more rapid rate of development for each site, whereas the distributions of responses of the same populations at the warm sites (PR and PA) changed very little even for the 4 °C temperature increase.

We are not suggesting that use of the upper limit on development rate is robust for broad use, but instead that the MET should include more sites that have a wider range of temperatures and day lengths to enable non-linear responses to be estimated. For example, improvements could be attained using a beta function for temperature response, in addition to controlled environment experiments with a wider range of genetic material to develop non-linear functions to represent the full range of environmental responses in this crop species.

### 3.5 Simulating response distributions for all potential genotype combinations

The performance of DPLM module across the five experimental sites was simulated for RILs with all possible allelic combination (4096) of the 12 QTLs used by the dynamic time-to-flower module. Frequency distributions for the number of days from planting to first flower were produced for each location (Fig. 7). The dots in the figure highlight the number of days required for the Jamapa and Calima parents to reach the stage of first flowering. The simulated first flowering dates at the ND site were later (mean of 59.1 days) and had a larger standard deviation (9.10 days) compared to the other sites. The spread of simulated days to first flower ranged between 41 and 94 days at ND due to its longer day lengths and some days with cooler temperatures than other sites. The smallest average number of days to first flower was for sites with high-temperature conditions and short day lengths (PR and PA), where simulated means were about 36 days for both locations, and standard deviations of 2.6 and 2.6 days, respectively, with response ranges varying from 34 to 39 days for each location. The FL and PO sites with their warm conditions showed simulated means of 42 days and 45 days, respectively, and standard deviations of 3.5 days and 3.9 days, respectively. The shapes of the distributions were bell-shaped across all sites except for ND which showed the flattest shape due to the G, E, G × E and G × G interactions in the mixed piecewise dynamic module. The altered behaviour of the parental lines was also found in these simulations, where Jamapa flowered earlier than Calima for the FL, PA, PO and PR sites and Calima flowered earlier than Jamapa for the ND site.

### 3.6 Yield prediction

For each of the five sites, we simulated yield using the original CSMG model and the original GSPs that were calibrated by Acharya et al. (2017) for each individual RIL. For the MET crop management practices and 1 year of environmental conditions (Fig. 1), simulated mean yield by CSMG was lowest for PA (190.0 ± 89 kg ha⁻¹) and highest for ND (637.3 ± 242 kg ha⁻¹) while for the other three sites mean yield ranged from 304.7 ± 135 kg ha⁻¹ for FL, 505.0 ± 270.0 kg ha⁻¹
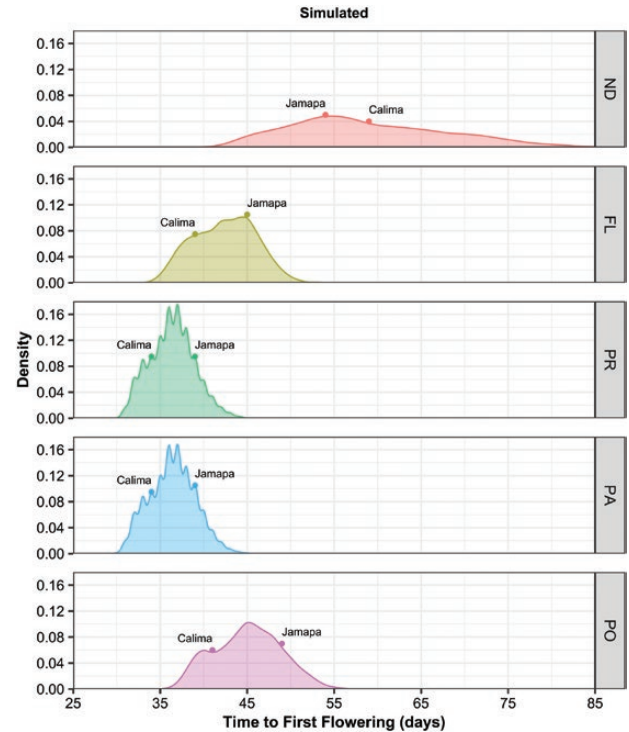


**Figure 7. Density plots for simulated time to first flower (days) across the five sites showing all possible genetic combinations. The distribution of simulated days to flower by site includes all RIL combinations ($2^{12}$ = 4096) as simulated by the dynamic piecewise linear module (DPLM), while dots at the top of each distribution represent the simulated parental lines Jamapa and Calima. The five experimental sites are: Prosper, North Dakota (ND); Citra, Florida (FL); Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).**

for PO and 540.0 ± 244.8 kg ha⁻¹ for PR (Fig. 8). For the DPLM-CSM hybrid model, simulated yield ranking among the five sites was similar. The highest mean yield was obtained for ND (720.0 ± 291.0 kg ha⁻¹), while the lowest mean yield was obtained for PA (234.4 ± 104.2 kg ha⁻¹). Mean yield for FL was 354.8 ± 156.4 kg ha⁻¹, for PO was 625.0 ± 311.2 kg ha⁻¹ and for PR was 673.3 ± 264.7 kg ha⁻¹ (Fig. 8). The differences in yield were due to the slight differences in simulated flowering dates between the original CSMG model and the DPLM-CSM hybrid model (Figs 2 and 5), while all other growth and development processes were simulated exactly the same as the CSMG model using the same inputs (Fig. 4).

### 3.7 Further advancement in gene-based modelling

This work presents an approach for incorporating gene-based modules into an existing crop growth model for simulating days to first flower (by the DPLM module) and simulating all other processes and final yield using original components of the CSMG. It builds on the approach discussed by Vallejos et al. (2021). Only minor modifications were needed to enable their dynamic

model to be integrated as a module into the existing CSMG model (Hoogenboom *et al.* 1994; Jones *et al.* 2003). There were only small differences between results from the dynamic piecewise linear module integrated into the CSMG model from our work and model results published by Vallejos *et al.* (2021). We recognize the need for use of independent data to evaluate the predictive capabilities in other environments and are working on that. In addition, there is a need to use a more diverse population to evaluate the ability of the model to predict first flower occurrence across genetic variation that may not be in the population used in the MET data set.

The model integration approach used here is different from previously published approaches because it incorporates a gene-based dynamic model to replace an existing dynamic component in a comprehensive crop growth simulation model. The approach used for integrating the gene-based first flower module into the CSMG model possibly can be used to incorporate other gene-based modules to systematically transition from a GSP-based model to a gene-based model (Hoogenboom *et al.* 2004). In this work, we only added one type of input data, the genetic information for each RIL and parent. All of the other inputs in the original crop model were not modified, and information on planting date and daily weather data were used by the new



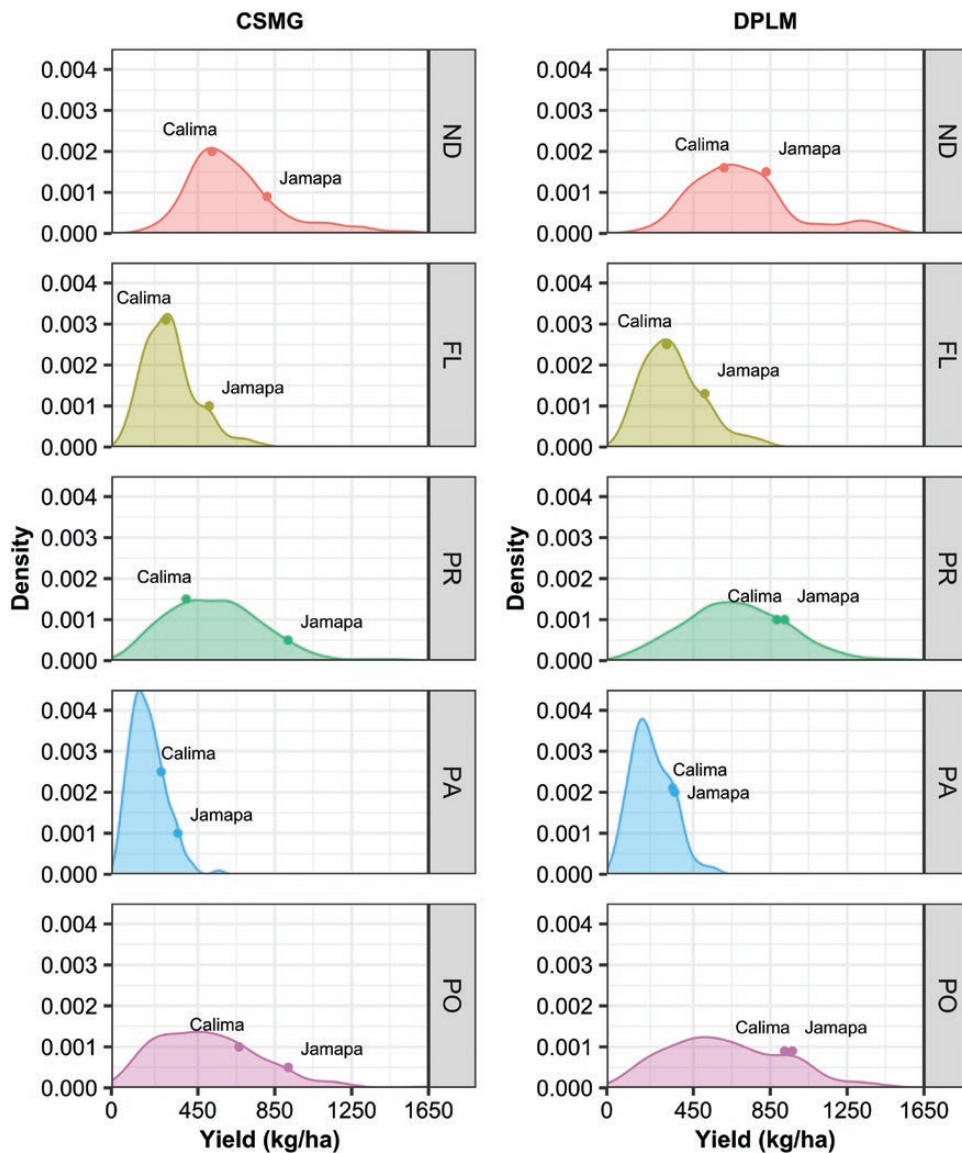**Figure 8. Density plot for simulated yield using the original CSM-CROPGRO-Drybean model and genetic specific coefficients (CSMG) (left panel) and the dynamic piecewise linear module (DPLM) integrated with the CSM-CROPGRO-Drybean model (right panel). The five experimental sites are: Prosper, North Dakota (ND); Citra, Florida (FL); Puerto Rico (PR); Palmira, Colombia (PA); and Popayan, Colombia (PO).**

gene-based time to flower module, ensuring consistency in inputs across all existing and new components of the model.

This work also shows that integrating the genetic information is a promising approach to predict plant development stages of new genotypes and new environments. Instead of estimating GSPs for a specific trait, it requires less effort when a new cultivar is released in that only QTL information is required, saving time and resources that would be otherwise needed for phenotyping. The long-term expectation associated with most QTL studies is the replacement of each QTL linked marker with the gene responsible for that particular QTL effect. This work further shows that genetic modules for other processes can be based on statistical methods that are routinely used by geneticists, if they are developed to replace equivalent modules in existing dynamic crop models.

However, it is clear that considerably more progress is needed to identify other issues that might occur by combining these two types of models. There is a need to extend gene-based modules to cover the full genetic variability of a crop and to introduce other process modules into existing models (Boote *et al.* 2021). Further work is required to improve the gene-based module and to add other processes that are linked dynamically with the crop model.

## 4. CONCLUSIONS

This study showed the potential for integrating a process-oriented gene-based module that only requires genetic input information into an existing comprehensive crop model with its empirical cultivar inputs without changing other modules or inputs. The CSM-CROPGRO-Drybean model with the integrated gene-based module was able to not only predict flowering date using only QTL and weather information, but also final yield using the original GSPs for all processes except rate of progression to first flower. This approach can be extended to other processes by using procedures that are similar to those described in this study to obtain genetic and phenotypic information and to estimate dynamic model parameters as a function of QTL and environmental data.

## SUPPORTING INFORMATION

The following additional information is available in the online version of this article—

**Table S1.** Recombinant inbred lines for common bean (*Phaseolus vulgaris*). Each Quantitative trait loci (QTL) has a marker value according to its allelic identity, assigned as '+1' for the Calima alleles and '−1' for the Jamapa alleles. This information was used as input for the dynamic piecewise linear module (DPLM) coupled with the CSM-CROPGRO-Drybean model.

**Table S2.** Estimated values in the dynamic QTL effect module showing the estimated parameters with confidence intervals and p-values for the rate of progress from planting to flowering.

**Figure S1.** Parameter estimation process for predicting first flowering across all sites using the dynamic mixed linear module (DMLM).

**Figure S2.** Computer code for the dynamic piecewise linear module (DPLM) coupled with the CSM-CROPGRO-Drybean model.

## LITERATURE CITED

Acharya S, Correll M, Jones JW, Boote KJ, Alderman PD, Hu Z, Vallejos CE. 2017. Reliability of genotype-specific parameter estimation for crop models: insights from a Markov chain Monte-Carlo estimation approach. *Transactions of the ASABE* **60**:1699–1712.

Anothai J, Patanothai A, Jogloy S, Pannangpetch K, Boote KJ, Hoogenboom G. 2008. A sequential approach for determining the cultivar coefficients of peanut lines using end-of-season data of crop performance trials. *Field Crops Research* **108**:169–178.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**:1–48.

Bhakta MS, Gezan SA, Clavijo Michelangeli JA, Carvalho M, Zhang L, Jones JW, Boote KJ, Correll MJ, Beaver J, Osorno JM, Colbert R, Rao I, Beebe S, Gonzalez A, Ricaurte J, Vallejos CE. 2017. A predictive model for time-to-flowering in the common bean based on QTL and environmental variables. *G3 (Bethesda, Md.)* **7**:3901–3912.

Bhakta MS, Jones VA, Vallejos CE. 2015. Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris* L. *PLoS One* **10**:e0116822.

Boote KJ, Jones JW, Batchelor WD, Nafziger ED, Myers O. 2003. Genetic coefficients in the CROPGRO-Soybean model: links to field performance and genomics. *Agronomy Journal* **95**:32–51.

Boote KJ, Jones JW, Hoogenboom G. 2021. Incorporating realistic trait physiology into crop growth models to support genetic improvement. *In Silico Plants.* **2021**:diab002; doi:10.1093/insilicoplants/diab002.

Boote KJ, Jones JW, Hoogenboom G, Pickering NB. 1998. The CROPGRO model for grain legumes. In: Tsuji GY, Hoogenboom G, Thornton PK, eds. *Understanding options for agricultural production.* Dordrecht, The Netherlands: Springer, 99–128.

Boote KJ, Jones JW, White JW, Asseng S, Lizaso JI. 2013. Putting mechanisms into crop production models. *Plant, Cell & Environment* **36**:1658–1672.

Brown TB, Cheng R, Sirault XR, Rungrat T, Murray KD, Trtilek M, Furbank RT, Badger M, Pogson BJ, Borevitz JO. 2014. TraitCapture: genomic and environment modelling of plant phenomic data. *Current Opinion in Plant Biology* **18**:73–79.

Buddhaboon C, Jintrawet A, Hoogenboom G. 2018. Methodology to estimate rice genetic coefficients for the CSM-CERES-Rice model using GENCALC and GLUE genetic coefficient estimators. *The Journal of Agricultural Science* **156**:482–492.

Bush WS, Moore JH. 2012. Chapter 11: genome-wide association studies. *PLoS Computational Biology* **8**:e1002822.

Cooper M, Technow F, Messina C, Gho C, Totir LR. 2016. Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. *Crop Science* **56**:2141–2156.

Gu J, Yin X, Zhang C, Wang H, Struik PC. 2014. Linking ecophysiological modelling with quantitative genetics to support marker-assisted crop design for improved yields of rice (*Oryza sativa*) under drought stress. *Annals of Botany* **114**:499–511.

Hammer GL, van Oosterom E, McLean G, Chapman SC, Broad I, Harland P, Muchow RC. 2010. Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. *Journal of Experimental Botany* **61**:2185–2202.

Hoogenboom G, Jones JW, Boote KJ. 1992. Modeling growth, development, and yield of grain legumes using SOYGRO, PNUTGRO, and BEANGRO: a review. *Transactions of ASAE* **35**:2043–2056.

Hoogenboom G, Porter CH, Boote KJ, Shelia V, Wilkens PW, Singh U, White JW, Asseng S, Lizaso JI, Moreno LP, Pavan W, Ogoshi R, Hunt LA, Tsuji GY, Jones JW. 2019a. The DSSAT crop modeling ecosystem. In: Boote KJ, eds. *Advances in crop modelling for a sustainable agriculture*. Cambridge: Burleigh Dodds Science Publishing, 173–216.

Hoogenboom G, Porter CH, Shelia V, Boote KJ, Singh U, White JW, Hunt LA, Ogoshi R, Lizaso JI, Koo J, Asseng S, Singels A, Moreno LP, Jones JW. 2019b. *Decision Support System for Agrotechnology Transfer (DSSAT) version 4.7.5*. Gainesville, FL: DSSAT Foundation. https://DSSAT.net (20 December 2020).

Hoogenboom G, White JW, Jones JW, Boote KJ. 1994. BEANGRO: a process-oriented dry bean model with a versatile user interface. *Agronomy Journal* **86**:182–190.

Hoogenboom G, White JW, Messina CD. 2004. From genome to crop: integration through simulation modeling. *Field Crops Research* **90**:145–163.

Huang X, Han B. 2014. Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology* **65**:531–551.

Hunt LA, Pararajasingham S, Jones JW, Hoogenboom G, Imamura DT, Ogoshi RM. 1993. GENCALC: software to facilitate the use of crop models for analyzing field experiments. *Agronomy Journal* **85**:1090–1094.

Hwang C, Correll MJ, Gezan SA, Zhang L, Bhakta MS, Vallejos CE, Boote KJ, Clavijo-Michelangeli JA, Jones JW. 2017. Next generation crop models: a modular approach to model early vegetative and reproductive development of the common bean (*Phaseolus vulgaris* L). *Agricultural Systems* **155**:225–239.

Jones JW, Antle JM, Basso B, Boote KJ, Conant RT, Foster I, Godfray HCJ, Herrero M, Howitt RE, Janssen S, Keating BA, Munoz-Carpena R, Porter CH, Rosenzweig C, Wheeler TR. 2016. Brief history of agricultural systems modeling. *Agricultural Systems* **155**:240–254.

Jones JW, Hoogenboom G, Porter CH, Boote KJ, Batchelor WD, Hunt LA, Wilkens PW, Singh U, Gijsman AJ, Ritchie JT. 2003. The DSSAT cropping system model. *European Journal of Agronomy* **18**:235–265.

Jones JW, Keating BA, Porter CH. 2001. Approaches to modular model development. *Agricultural Systems* **70**:421–443.

Kobayashi K, Salam MU. 2000. Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal* **92**:345–352.

Messina CD, Jones JW, Boote KJ, Vallejos CE. 2006. A gene-based model to simulate soybean development and yield responses to environment. *Crop Science* **46**:456–466.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I—a discussion of principles. *Journal of Hydrology* **10**:282–290.

Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z. 2017. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Molecular Plant* **10**:1047–1064.

Reymond M, Muller B, Leonardi A, Charcosset A, Tardieu F. 2003. Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. *Plant Physiology* **131**:664–675.

Ritchie JT, Nesmith DS. 1991. Temperature and crop development. In: Hanks J, Ritchie JT, eds. *Modeling plant and soil systems*. Madison, WI: American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, 5–29.

Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, Jannink JL, McCouch S. 2016. Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**:395–408.

Technow F, Messina CD, Totir LR, Cooper M. 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PLoS One* **10**:e0130855.

Thomson MJ. 2014. High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology* **2**:195–212.

Thorburn PJ, Boote KJ, Nendel C, Rötter RP, Ewert F. 2018. Recent advances in crop modelling to support sustainable agricultural production and food security under global change. *European Journal of Agronomy* **100**:1–3.

Vallejos CE, Jones JW, Bhakta MS, Gezan SA, Correll MJ. 2021. Dynamic gene-based ecophysiological models to predict phenotype from genotype and environment data. *In Silico Plants* (submitted). doi:10.1101/2021.02.07.429927

Wallace DH, Enriquez GA. 1980. Day length and temperature effects on days to flowering of early and late maturing beans (*Phaseolus vulgaris* L.). *Journal American Society For Horticultural Science* **105**:583–591.

Wallace DH, Gniffke PA, Masaya PN, Zobel RW. 1991. Photoperiod, temperature, and interaction effects on days and nodes required for flowering of bean. *Journal American Society for Horticultural Science* **116**:534–543.

Wallach D, Hwang C, Correll MJ, Jones JW, Boote K, Hoogenboom G, Gezan S, Bhakta M, Vallejos CE. 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, Makowski D, Jones JW, Brun F. 2019. Model evaluation. In: Wallach D, Makowski D, Jones JW, Brun F, eds. *Working with dynamic crop models*. 3rd edn. London, UK: Academic Press, 311–373.

White JW. 2006. From genome to wheat: emerging opportunities for modelling wheat growth and development. *European Journal of Agronomy* **25**:79–88.

White JW, Hoogenboom G. 1996. Simulating effects of genes for physiological traits in a process-oriented crop model. *Agronomy Journal* **88**:416–422.

White JW, Hoogenboom G. 2003. Gene-based approaches to crop simulation. *Agronomy Journal* **95**:52–64.

White JW, Hoogenboom G, Hunt LA. 2005. A structured procedure for assessing how crop models respond to temperature. *Agronomy Journal* **97**:426–439.

Yin X, Chasalow SD, Dourleijn CJ, Stam P, Kropff MJ. 2000. Coupling estimated effects of QTLs for physiological traits to a crop growth model: predicting yield variation among recombinant inbred lines in barley. *Heredity* **85**:539–549.

Yin X, Stam P, Kropff MJ, Schapendonk AHCM. 2003. Crop modeling, QTL mapping, and their complementary role in plant breeding. *Agronomy Journal* **95**:90–98.

Yin X, Struik PC. 2010. Modelling the crop: from system dynamics to systems biology. *Journal of Experimental Botany* **61**:2171–2183.

Yin X, van der Linden CG, Struik PC. 2018. Bringing genetics and biochemistry to crop modelling, and vice versa. *European Journal of Agronomy* **100**:132–140.