UNIVERSIDADE DE PASSO FUNDO

Graduate Program in Applied Computing

Master Dissertation

INCORPORATING A DYNAMIC GENE-BASED PROCESS MODULE INTO A CROP SIMULATION MODEL

FÁBIO AUGUSTO ANTUNES DE OLIVEIRA

UPF

UNIVERSITY OF PASSO FUNDO
INSTITUTE OF EXACT SCIENCES AND GEOSCIENCES GRADUATE PROGRAM IN APPLIED COMPUTING

INCORPORATING A DYNAMIC GENE-BASED MODULE PROCESS MODULE INTO A CROP SIMULATION MODEL

Fábio Augusto Antunes de Oliveira

Thesis submitted to the University of Passo Fundo in partial fulfillment of the requirements for the degree of Master in Applied Computing.

Advisor: Prof. Dr. Willingthon Pavan
Co-Advisor: Prof. Dr. Carlos Amaral Holbig

Passo Fundo
2021

International Publication Cataloging Data (CIP)

CIP – Catalogação na Publicação

O48i Oliveira, Fábio Augusto Antunes de
Incorporating a dynamic gene-based process module into a crop simulatio model [electronic resouce] / Fábio Augusto Antunes de Oliveira. – 2021.
7.8 MB ; PDF.

Advisor: PhD. Willinghton Pavan.
Co-Advisor: Prof. Dr. Carlos Amaral Holbig.
Master's Thesis (Master in Applied Computing) – University of Passo Fundo, 2021.

1. Software engineering. 2. Agricultural informatics. 3. Simulation models. 4. Beans. I. Pavan, Willinghton, advisor. II. Holbig, Carlos Amaral, co-advisor. III. Title.

CDU: 631:004

Catalogação: Bibliotecária Juliana Langaro Silveira – CRB 10/2427

PPGCA
Programa de Pós-Graduação
em Computação Aplicada
Instituto de Ciências Exatas e Geociências | ICEG

MINUTE OF
ACADEMIC COURSE CONCLUSION WORK

FÁBIO AUGUSTO ANTUNES DE OLIVEIRA

On the thirty-one three days of March two thousand and twenty-one, at 09:00 am BRT, by online, through video conference, was held the public defense session of the Course Final Work "Incorporating a dynamic gene-based process module into a crop simulation model" authored by Fábio Augusto Antunes de Oliveira, academic of the Graduate Program in Applied Computing - PPGCA / UPF. According to Postgraduate Council´s information and listed in the PPGCA Secretariat archives, the student fulfilled the requirements to submit his work to be evaluated. The examination committee was composed by Dr. Willingthon Pavan (Advisor/Chair), Dr. Carlos Amaral Hölbig (UPF), PhD. Gerrit Hoogenboom (UF - USA), PhD. José Maurício Cunha Fernandes (Embrapa Trigo) and PhD. James W. Jones (UF - USA). After the presentation and arguing, the examining committee considered the candidate **APPROVED**. A period of up to forty-five (45) days, according to the Rules of the PPGCA, was granted for the academic to submit the final writing to the Postgraduate Council, to make the necessary referrals for the issuance of Master in Applied Computing diploma. To record, this minute was drawn up, signed by the examination committee members and the PPGCA Coordinator.

DocuSigned by:
WILLINGTHON PAVAN
4810E0667FBA4A2...
Prof. Dr. Willingthon Pavan – UPF
Examining Committee President
(Advisor)

DocuSigned by:
26790BBACB944B7...
Prof. Dr. Carlos Amaral Hölbig – UPF
(Internal Examinator)

DocuSigned by:
Gerrit Hoogenboom
5E08EF7B3AF145C...
PhD. Gerrit Hoogenboom – UF (USA)
(External Examinator)

DocuSigned by:
792E0255B0C94A6...
PhD. José Maurício Cunha Fernandes – Embrapa Trigo
(External Examinator)

DocuSigned by:
James W. Jones
9ADF360ADE8B436...
PhD. James W. Jones – UF (USA)
(External Examinator)

DocuSigned by:
26790BBACB944B7...
Prof. Dr. Carlos Amaral Hölbig
PPGCA Coordinator

To my mother Ivania, and in memory of my father Sadi, and my brother Junior.

ACKNOWLEDGMENTS

First, I would like to thank my parents, and my brother for their constant support and their love. They are my motivation for and source of strength.

I would like to especially thank my advisor Dr. Willingthon Pavan, who believe me, and give me the opportunity to join your projects, for your constant support, encouragement, friendship and for teaching how to improve myself professionally and personally.

I would also thank to my co-chair Dr. Carlos Amaral Hölbig, and Dr. Jose Mauricio Cunha Fernandes for their academic assistance throughout my program.

I also am deeply grateful to Dr. Gerrit Hoogenboom, for providing an opportunity to contribute to the DSSAT community, participating in the DSSAT Sprints, and giving me all the support to work, valuable advises and mentorship during this work. I learn a lot and hope to continue contributing.

I am very thankful to Dr. Jones, for guiding me through the genetics and mixed models, for all the patience, assistance, and was always available to answer the questions, and his valuable advises.

The base models for this work and data collection could not have been accomplished without your work Dr. Eduardo Vallejos, Dr. Mehul Bhakta, Dr. Melanie Correll, Dr. Kenneth Boote, thank you for providing all the data, knowledge, and critical reviews and insights during this work.

I would also like to thank the DSSAT Foundation, which I am glad to be part of this group, and especially to Cheryl Porter, who is always happy and open to discuss about DSSAT and working hard to move DSSAT forward.

My thanks go out of the Agricultural and Biological Engineering Department as well. I would like to thank my friends, Vinicius Cerbaro, Vanessa Cerbaro, Mauricio Karrei and Rogério Nóia, for their friendship and support. They made my time in Gainesville very memorable.

I would like to thank my friends, Mari, Marcia, Samuel and Thiago at University of Passo Fundo, for the coffee break time in the morning at the little kitchen. Also, I big thanks to the Mosaico research group, who give me the support in my development as a scientist at University of Passo Fundo.

I would like to thank the Graduate Program in Applied Computing at University of Passo Fundo, and the Agricultural and Biological Engineering Department at University of Florida, for all the assistance.

I would also like to thank all my family, especially my uncle Genuir Marchezi, who is always helping and encouraging to continue working hard.

I sincerely thank God for guiding me and always being with me.

INCORPORANDO UM MÓDULO DINÂMICO BASEADO EM GENES EM UM MODELO DE SIMULAÇÃO DE CULTURAS

RESUMO

Os modelos de culturas usam parâmetros chamados como coeficientes genéticos (CGs) para representar as características da planta (fenótipo) em ambientes específicos. CGs não se referem as informações genéticas “verdadeiras” com base em um nível de gene e são estimados a partir de dados observados no campo, requerendo experimentos para medir a resposta fenotípica quando novos cultivares são lançados. Modelos baseados em genes oferecem o potencial para quantificar e identificar o fenótipo a partir da composição genética da planta (genótipo). Este trabalho propõe uma abordagem para a incorporação de um módulo dinâmico baseado em genes para simular o tempo de florescimento do feijão (*Phaseolus vulgaris* L.). Esta nova abordagem visa trabalhar em um modo híbrido para simular usando locús de característica quantitativa (LCQ) ou CGs para interações genéticas (G), ambientais (E) e G × E e demonstrar aplicações potenciais usando análise de sensibilidade e para simulação de rendimento.

Palavras-chave: DSSAT, CROPGRO-Drybean, Beans, QTLs.

INCORPORATING A DYNAMIC GENE-BASED PROCESS MODULE INTO A CROP SIMULATION MODEL

ABSTRACT

Crop simulation models uses parameters referred as genotype-specific parameters (GSPs) to represent plant characteristics (phenotype) under specific environments. GSPs do not refer to “true” genetic information based on a gene level and are estimated from data observed in the field and require experiments to measure phenotypic response when new cultivars are released. Gene-based models offer the potential to quantify and identify the phenotype from the plant's genetic composition (genotype). This work proposes an approach for incorporates a dynamic gene-based module for simulating time-to-flowering for common bean (*Phaseolus vulgaris* L.). This new approach aims to work in a hybrid mode for simulating using quantitative trait loci (QTLs) or GSPs for genetic (G), environment (E), and G × E interactions, and demonstrate potential applications using sensitivity analysis and for simulating yield.

Keywords: DSSAT, CROPGRO-Drybean, Beans, QTLs.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1. INTRODUCTION

The population growth, expected to reach 9 billion by 2050, will significantly increase consumption and demand for food. To meet the needs of food security and sustainability, new investments and strategies are needed to continue increasing the productivity of agricultural systems. Achieving high yields in low-income countries is of great importance for global demand to be met with minimal environmental impacts [1]–[5]. The use of technologies that combine scenarios with focus on interactions between genetics (G), environment (E) and management (M) practices, could provide more realistic income projections and viable solutions [5].

Since the early 1970s, considerable efforts have been made and continue to be made in the development of crop system models (CSM), to predict the final productivity of agricultural systems and work as support system for decision makers, policy for food security [6]. Crop system models are tools based on processes that dynamically simulate a cropping system affected by environmental conditions, management practices and differences between cultivars, describing the rate of crop development.

The cultivars in the CSMs are represented by parameters called genetic-specific parameters (GSPs) [7]. GSPs are not related to the genetic information and are estimated with data observed in the field. However, GSPs may not accurately represent variation between cultivars and environments, although they can provide good predictions when estimated independently for a given cultivar and specific environment [8]. It was found that the present generation of soil-plant-atmosphere crop models does not support responses to major climatic and cultivar variation in extreme environmental conditions [9]. In parallel, Boote *et al.* [10] and Hwang *et al.* [8] conclude that existing crop models can dynamically replace modules with gene-based processes, as they are developed and made available.

Technological advances present a fast and inexpensive way to identify the genetic composition of plants [11], [12]. Analytical tools are available to locate which genes are associated with variation in different cultivar characteristics, increasing interest in identifying the phenotype of plants, through genetic information [13], [14]. Statistical methods are used by scientists to detect a gene or gene combinations associated with a phenotypic trait [15]–[17] and these tools also assist plant breeding

programs for prediction and selection of the cultivar lines to improve crop yield and gene-based models can help simulating interactions between genetics and environment (G × E) that have a major impact on final yield [17].

The time-to-flower trait is one of the major targets in plant breeding programs aiming maximizing the crop yield. The transition from vegetative to reproductive stage which is determined by the first flowering is a key factor for defining the successful reproduction in the ecosystem and depends on the genotype and the interactions with the environment. The major environmental factor that affects the time-to-flower are the photoperiod and temperature [18].

The Common bean (*Phaseolus vulgaris* L.) was the crop used in this work which is a crop of major importance worldwide and source of protein and essential nutrients and most consumed in parts of Africa and the Americas [19]. Time-to-flowering phenotypes of a RIL population are collect from an Andean bean cultivar, Calima, which is photoperiod sensitive and a Mesoamerican cultivar, Jamapa which is less sensitive to photoperiod [18].

The goal of this study is to address the questions of how this type of integration can be done in an existing dynamic crop model for time-to-flower and what are the complications and limitations that can occur. In Chapter 2 are shown the details of the development used for integration an existing dynamic crop model and the gene-based model and demonstrate potential applications of the hybrid dynamic model for the G × E interactions using sensitivity analysis and for simulating yield. On Chapter 3 are presented the final remarks and future work.

2. INCORPORATING A DYNAMIC GENE-BASED PROCESS MODULE INTO A CROP SIMULATION MODEL

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 conducted in 2011 and 2012 included 187 recombinant inbred lines (RILs) from a bi-parental bean family to measure the effects of quantitative trait loci (QTL), environment (E), and QTL×E interactions across five sites. The dynamic mixed linear model from Vallejos et al. (2020) was modified in this study to create a dynamic module that was then integrated into the CSM-CROPGRO-Drybean model. This new hybrid crop model, with the gene-based flowering module replacing the original flowering component, requires allelic makeup of each genotype being 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.

2.1. INTRODUCTION

Scientific advances in understanding plant genes combined with advances in technologies for rapidly and inexpensively identifying genetic makeup of plants [11], [12] have fueled 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 observed variation in specific traits [20]. Rapid progress in genome-wide association studies (GWAS) has enabled researchers to identify genes associated with variation in human diseases [13].

Genome-wide prediction models that use GWAS also have become powerful tools for improving crops such as tropical rice (e.g., Spindel *et al.* [21]). The GWAS approach has been implemented in recent work in other crops [14], [22], [23]. 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 [15]–[17]. 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 [6], [24]. 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 [25]–[27].

Recognizing the potential for introducing genetic information into crop models, White and Hoogenboom [15], [28] showed that some of the BEANGRO model's GSPs [29], [30] could be estimated as linear functions of genetic information. This approach was also used by Messina *et al.* [7] for the CROPGRO-Soybean model and by other researchers for different crops [31]–[35]. 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 [36]. More recently, Wallach *et al.* [37] 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 parameters may not accurately represent the genetic architecture of the associated crop phenotype or process [8]. Acharya *et al.* [38] found that commonly used approaches for estimating GSPs for the Cropping System Model (CSM)-CROPGRO-Drybean model [39]–[41] 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.* [38] 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.* [37]. 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 toward flowering. Furthermore, we have learned that there is variation among genotypes that were not captured in the original model formulations and associated assumptions [10], [38], [42]. 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 modeling 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 modeling community. There have been limited interactions between these science communities that might lead to more rapid advances in gene-based modeling. Hwang *et al.* [8] 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 L.) recombinant inbred lines (RIL) from a bi-parental family. As part of this study, significant QTLs controlling the time to flowering in the RIL population were identified [18], [43], which provided an opportunity to model QTL and environmental effects on the time to flowering. This study included geneticists, biostatisticians, and crop modelers 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 [18].

Vallejos *et al.* [44] 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 rate. 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.* [44]. 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.

2.2. MATERIAL & METHODS

2.2.1. 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 [18]. 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 (F11:14) giving rise to 187 RILs. Further details for this RIL population can be found in Bhakta *et al.* [18], while the QTL-based linkage is described by Bhakta *et al.* [43].

2.2.2. Experimental Sites

We used the data from the MET study that included five locations, 187 RILs, and two parents reported by Bhakta *et al.* [18]. 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, the coolest site, is located at an elevation of 1,800 m, while Palmira, the warmer site, is located at a 1,000 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.

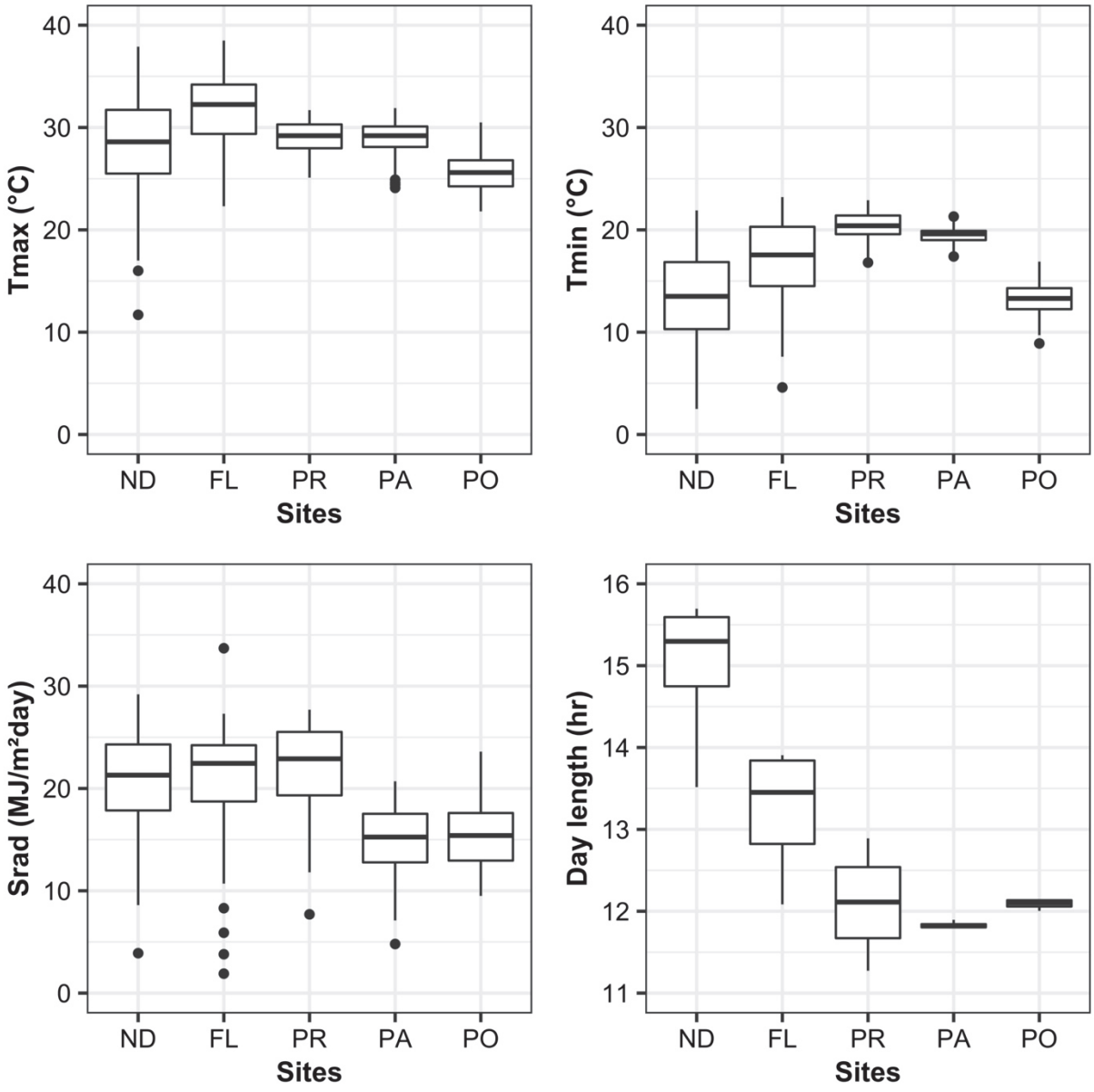

Figure 1. Boxplots of environmental variables observed for all five sites: Prosper, North Dakota (ND); Citra, Florida (FL); Isabella, Puerto Rico (PR); Palmira, Colombia (PA); Popayan, Colombia (PO). The boxplots show the distribution of daily values of maximum temperature (top left), daily minimum temperature (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.

Table 1. Summary of the meteorological and geographical data for each site in the multi-environment trial as reported by Bhakta *et al.* [18], except for the computed day length that is based on the Cropping System Model (CSM).

	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	Jun-30-2012	Apr-26-2012	Mar-6-2012	Dec-9-2011	Apr-28-2012
Latest first flowering date	Jul-26-2012	May-16-2011	Mar-22-2012	Dec-24-2011	May-16-2012
Seasonal Maximum Temperature (ºC)[a]	27	32	29	28	25
Seasonal Minimum Temperature (ºC)[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} d^{-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.

2.2.3. Dynamic mixed linear model

Vallejos *et al.* [44] 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.* [18] 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.* [18] 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, twelve 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.* [44] used a similar approach to that used by Bhakta *et al.* [18] to develop their dynamic model [see Supporting Information Figure S1]. First, the time to first flower data for all RILs and environment combinations were transformed into a development rate toward 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.* [44] 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 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.

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

Where $FR_{s,g}(t)$ is the rate of progress to flowering (1/d) for the g^{th} genotype for the s^{th} site at time t (in days). μ represents the overall mean value of the daily development rates across all RILs and sites in the MET dataset. 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 centered on the mean values from the MET study for each RIL, based on the Vallejos *et al.* [44] 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/m2), $Tmax_s(t)$ 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 center 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 dataset, 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 [18]. 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 β_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).

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

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.

Table 2. Description of model abbreviations.

Module	Dynamic Module Description
DMLM	Dynamic Mixed Linear Model developed by Vallejos *et al.* [44].
DMLM-DL	Dynamic Mixed Linear Module by Vallejos *et al.* [44] 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)

2.2.4. Dynamic piecewise linear module

The CSM-CROPGRO-Drybean model (CSMG, Table 2), which is part of the Decision Support System for Agrotechnology Transfer (DSSAT; [41]), requires daily weather data, soil surface and profile characteristics, crop management scenarios, and cultivar information (GSPs) as input [40], [41]. 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 [30], [39] 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 daylengths 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 [45]. Instead, the response is only approximately linear over a specific range of temperatures, and response 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 [46]. 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 Figure S2].

$FRMAX_g$ is defined as the rate at which the progress toward 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):

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

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).

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

where the variable γ_g is the fixed intercept estimated for the g^{th} genotype and ρ_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 recombinant inbred line in the DPLM module, $FRD_{s,g}(t)$ limits the maximum rate of flowering to that set by *FRMAXg*, Equation (5):

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

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.

2.2.5. Incorporation of a gene-based module into CSM

The CSMG model [29], [30] 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 [41]. For this study the focus was on the phenology module of CROPGRO where the developmental and phenological phase transitions are implemented.

Boote *et al.* [39] 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 (GBM) to create a link between the DPLM module and the crop model (Figure 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 GBM 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 GBM 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 are computed using the original CSMG model.

2.2.6. 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 [45]. This resulted in a total of 4,096 (212) RILs. The coefficients estimated using 187 RILs plus the two parents were used to simulate the number of days to flowering for the 4,096 RILs. The input file BNGRO047.GEN file containing QTL information was revised by adding inputs for each of the 4,096 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 ºC increment to create five different temperature scenarios (base, base+1, base+2, base+3, and base+4 ºC), 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 analyzed 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.

2.2.7. 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.* [38]. 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 (Figure 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.

2.2.8. Model evaluation

To estimate parameters for the QTL-based modules, we used the *lmer* function of the lme4 package [47] 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 root mean squared error (RMSE), defined as

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

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 R2 was calculated because it indicates module performance adjusted by the number of the terms in the module, defined as

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

where R2 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 [48] skill score was also used as a measure of model error, referred to as model efficiency (ME) [49], and defined as

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

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 [50], Equation (10).

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

With

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

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

$$LCS = 2SD_sSD_m(1 - r)$$

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.

2.3. RESULTS AND DISCUSSION

2.3.1. 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.

$$FR_{s,g}(t) = 2.35148\ x\ 10^{-2} - 1.56357\ x\ 10^{-3} \cdot (DayL_s(t) - DayL_m) \quad (11)$$

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

Where the first term (2.35148 x 10^{-2}) is the overall average rate of progress, indicating that the average time between sowing and appearance of first flowering is 42.5 days (= 1/(2.35148 x 10^{-2})). The first coefficient (α_1 = –1.56357 x 10^{-3}) is the sensitivity to day length, indicating that a one-hour increase in day length would result in a rate of development that is 1.56357 x 10^{-3} below the average rate of

2.35148 x 10^{-2}). This one-hour 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 5 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 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 parameters 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 x 10^{-5}, the random effects variance was 6.34775 x 10^{-7}, and the residual variance was 1.3056 x 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 (Figure 2(A), 2(B) and 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 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 [51] and Wallace *et al.* [52].

The comparison of ME between the two module versions (DMLM and DMLM-DL) showed that both have the same high model efficiency 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 were 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. [18].

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[1]	Measures of Agreement[2]									
	Genotypes(#)	Observed Mean	Simulated Mean	Bias	RMSE	ME	MSE	(Bias)^2	SDSD	LCS
					DMLM[3]					
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

[1]Prosper, North Dakota (ND); Citra, Florida (FL); Isabella, Puerto Rico (PR); Palmira, Colombia (PA), and Popayan, Colombia (PO).

[2]ME = Nash and Sutcliffe [48] model efficiency; MSE = mean squared error; Bias^2, SDSD, and LCS present the decomposition of MSE.

[3]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

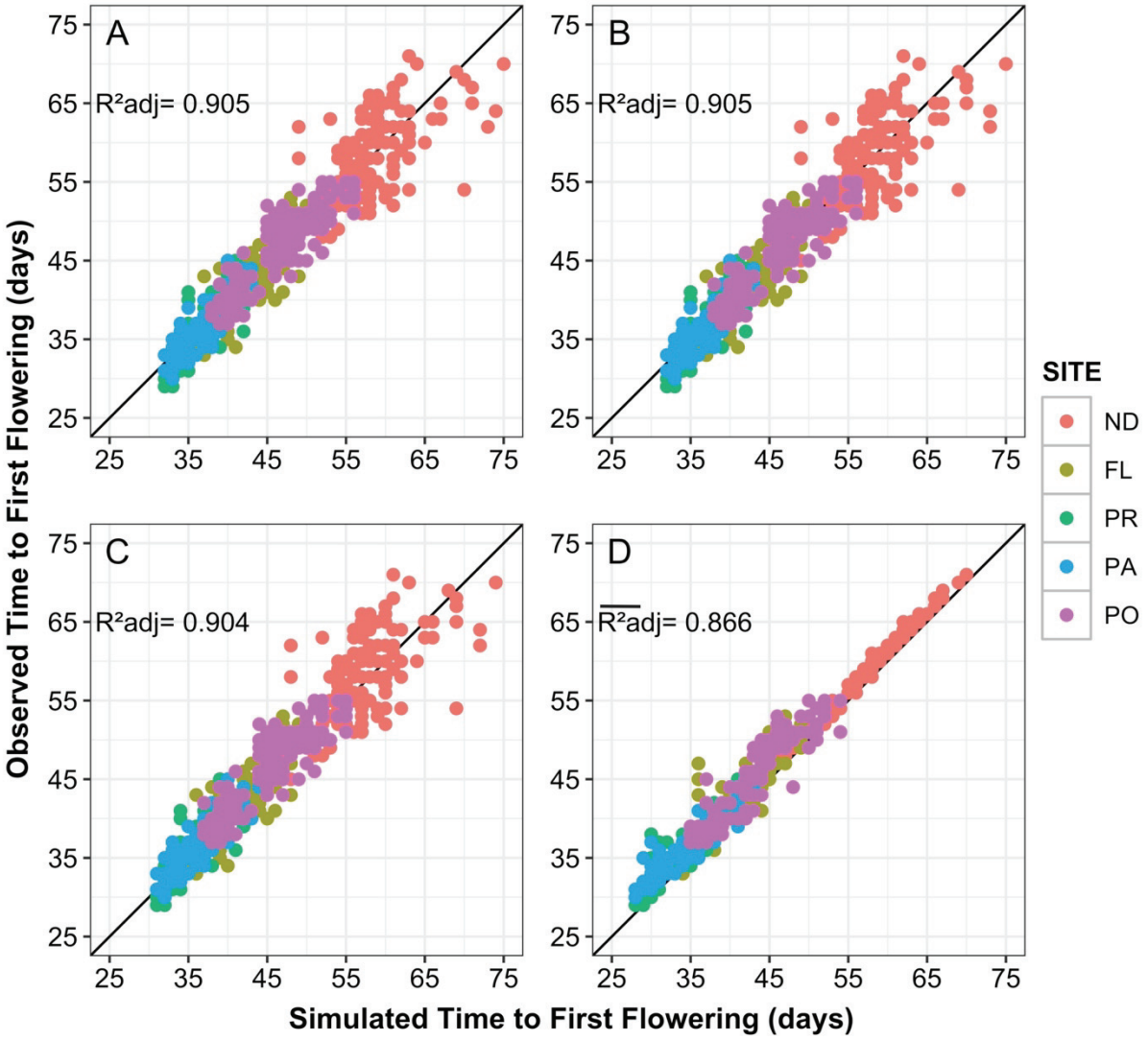

Figure 2. Observed versus simulated time to first flower across all five sites for the dynamic mixed linear module (DMLM) (A); the dynamic mixed linear model using the day length computed by the crop module (DMLM-DL) (B); the dynamic piecewise linear module incorporated into CSM-CROPGRO-Drybean (DPLM) (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 genetic specific coefficients based on Acharya *et al.* [38]. Each point represents an observed & simulated RIL; the solid 1:1 diagonal line represents equal values for time to first flower. R2adj 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).

2.3.2. Dynamic piecewise linear module

The highest maximum rate for any genotype in the MET dataset 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 datasets, ideally under more controlled environmental conditions.

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.

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

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 dataset with a RMSE of 1.66 days, ME of 0.77 days and MSE of 2.78 (Figure 3). These values indicate that the maximum developmental rates were affected by the genetic factors (12 QTLs).

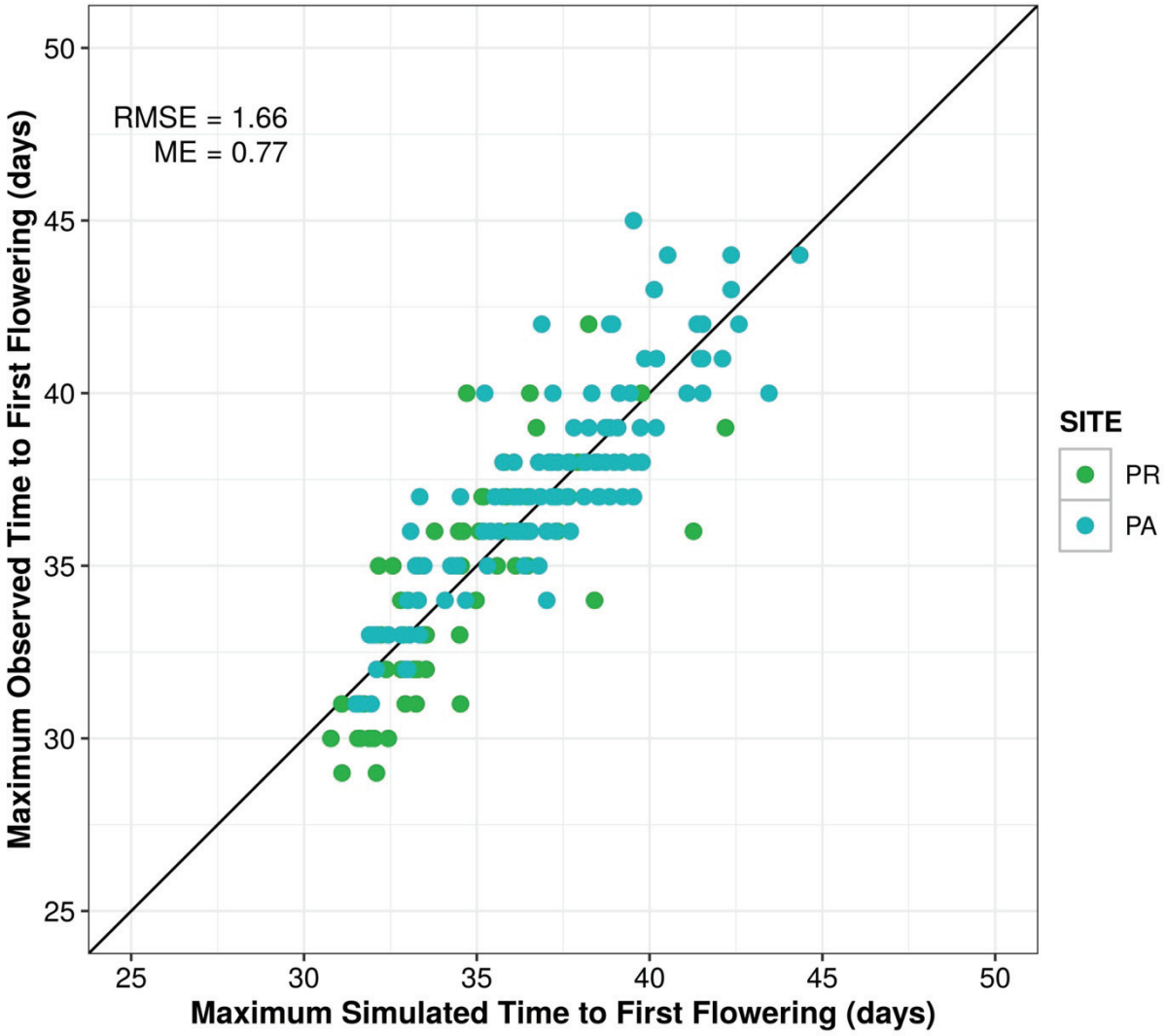

Figure 3. Maximum observed versus simulated time to first flowering for each RIL across all five sites based on a linear model dependent on the 12 QTLs alleles for the 187 RIL plus the two parental lines. RMSE = Root Mean Square Error; ME = Model Efficiency (Nash and Sutcliffe [48]). The solid 1:1 diagonal line represents equal values of maximum simulated/observed time to first flowering.

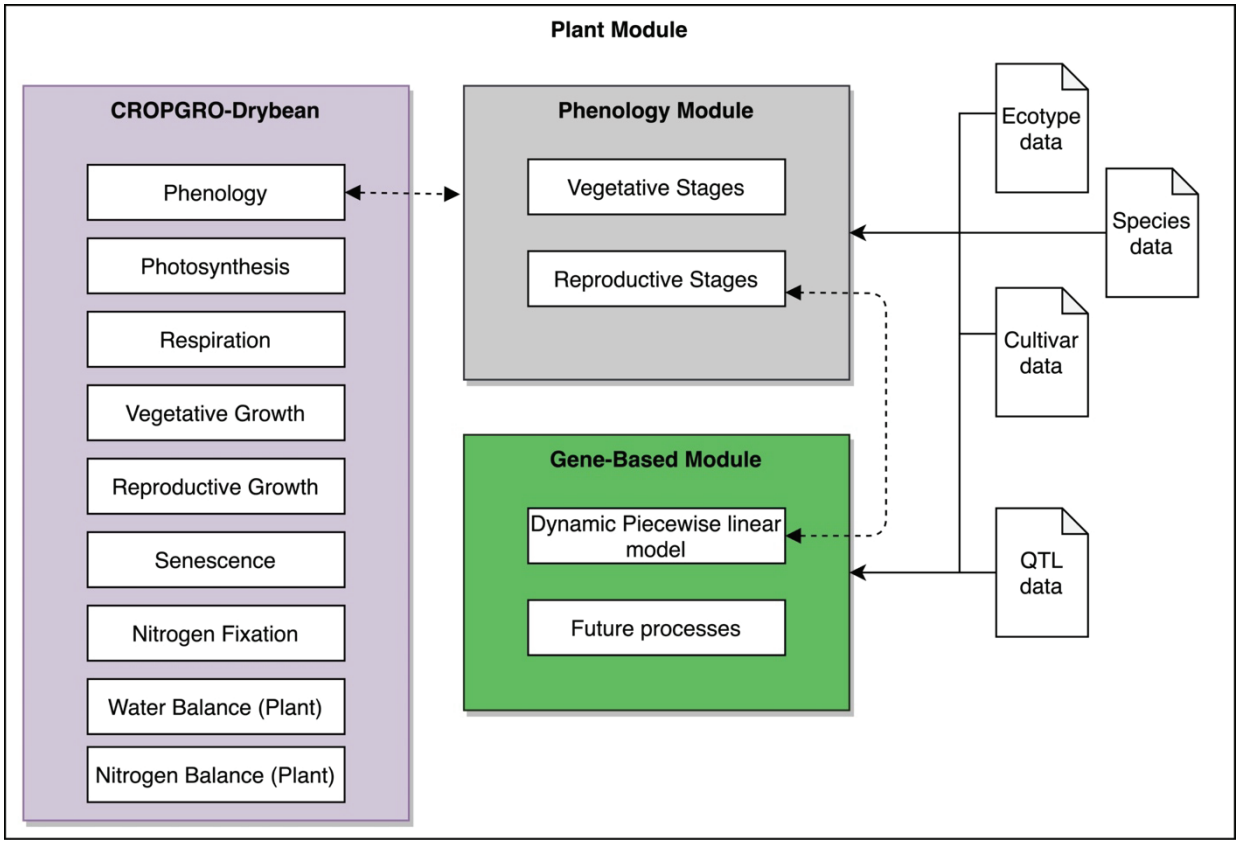

Figure 4. Overview of the DPLM-CSM model developed to integrate the CSM-CROPGRO-Drybean model (CSMG) with the Dynamic Piecewise Linear Module (DPLM) using a new gene-based module (GBM). The DPLM simulates the first time of first flowering module developed from the dynamic mixed linear model first developed by Vallejos *et al.* [44]. 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.

2.3.3. Structural changes of the CSM-CROPGRO-Drybean model

The new gene-based module (GBM) 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 Figure 4. This GBM 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). 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 GBM to simulate other vegetative and reproductive processes.

2.3.4. 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 Figure 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. QTL_2, which is associated with the growth habit gene *Fin*, shows interaction with *Tmin* and is likely responsible for this bimodality [18]. 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. [18] detected this transgressive behavior 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 (Figure 2(C) and 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 R2 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 and Figure 2(D)). However, these results are misleading in that the GSPs that produced these results were estimated for each individual RIL, which means that only 5 data points were used to estimate 3 GSPs for each RIL, and thus the agreements were forced in the GSP estimation process. The adjusted R2 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 R2 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 R2 values were lower than all of those for the gene-based modules for each site except at ND. As Acharya *et al.* [38] 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 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 Figure 6. The comparisons of the means and standard deviations of the populations for each 4-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 Tmin and Tmax 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 DMLM-DL module. In contrast, results for the cooler sites (PO and ND) were similar for both module versions. The frequency distributions (Figure 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 nonlinear 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 nonlinear functions to represent the full range of environmental responses in this crop species.

2.3.5. 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 twelve 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 (Figure 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 behavior 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.

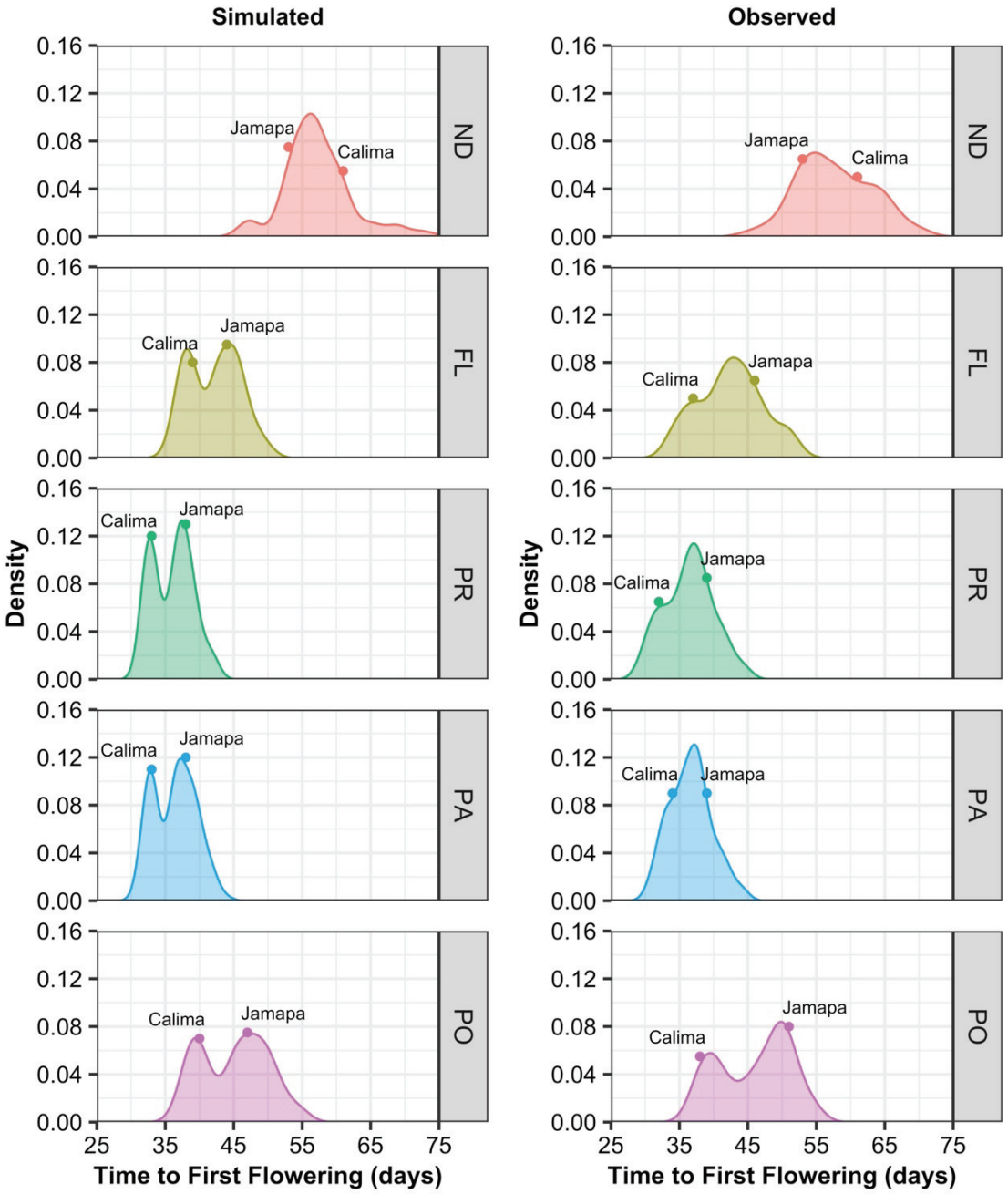

Figure 5. Density plots of time to first flower in days across five sites. Distribution of simulated time to first flower using the dynamic piecewise linear module (DPLM) (left panel) and the distribution of observed time to first flower (right panel). The parental lines Jamapa and Calima are highlighted 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).

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[1]	DPLM				DMLM-DL			
	Simulated Mean	Min[2]	Max[2]	Standard Deviation	Simulated Mean	Min	Max	Standard Deviation
	Base Temperature							
ND[1]	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 ºC							
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 ºC							
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 ºC							
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 ºC							
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

[1]Prosper, North Dakota (ND); Citra, Florida (FL); Isabella, Puerto Rico (PR); Palmira, Colombia (PA), and Popayan, Colombia (PO).

[2] Minimum/Maximum simulated number of days from sowing to first flower.

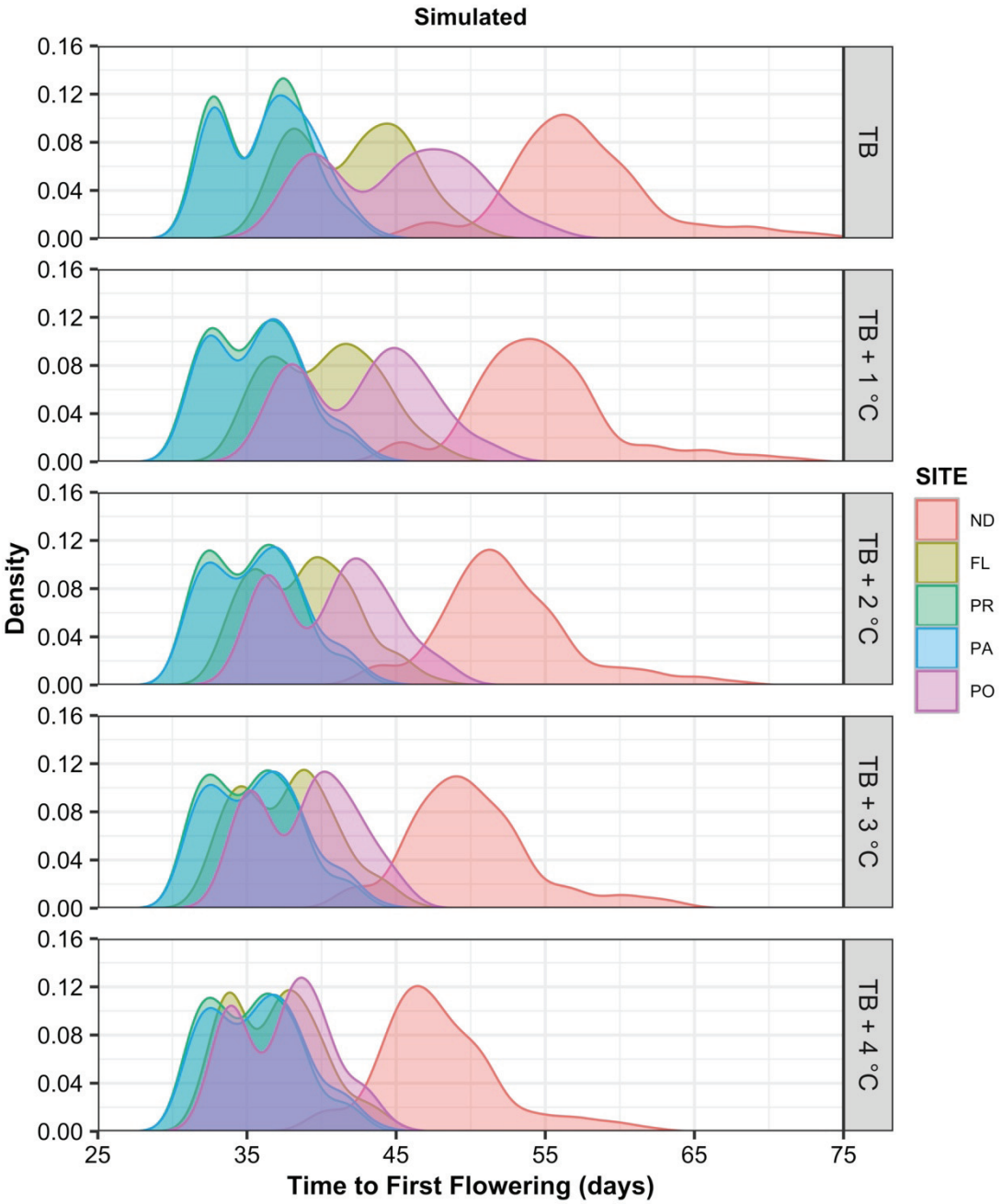

Figure 6. Density plots of distributions for simulated time to first flower (in days) using the dynamic piecewise linear module (DPLM). Simulated days between planting to first flower shows the responses to increasing the base maximum and minimum temperature from 1 °C through 4 °C. The five experimental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

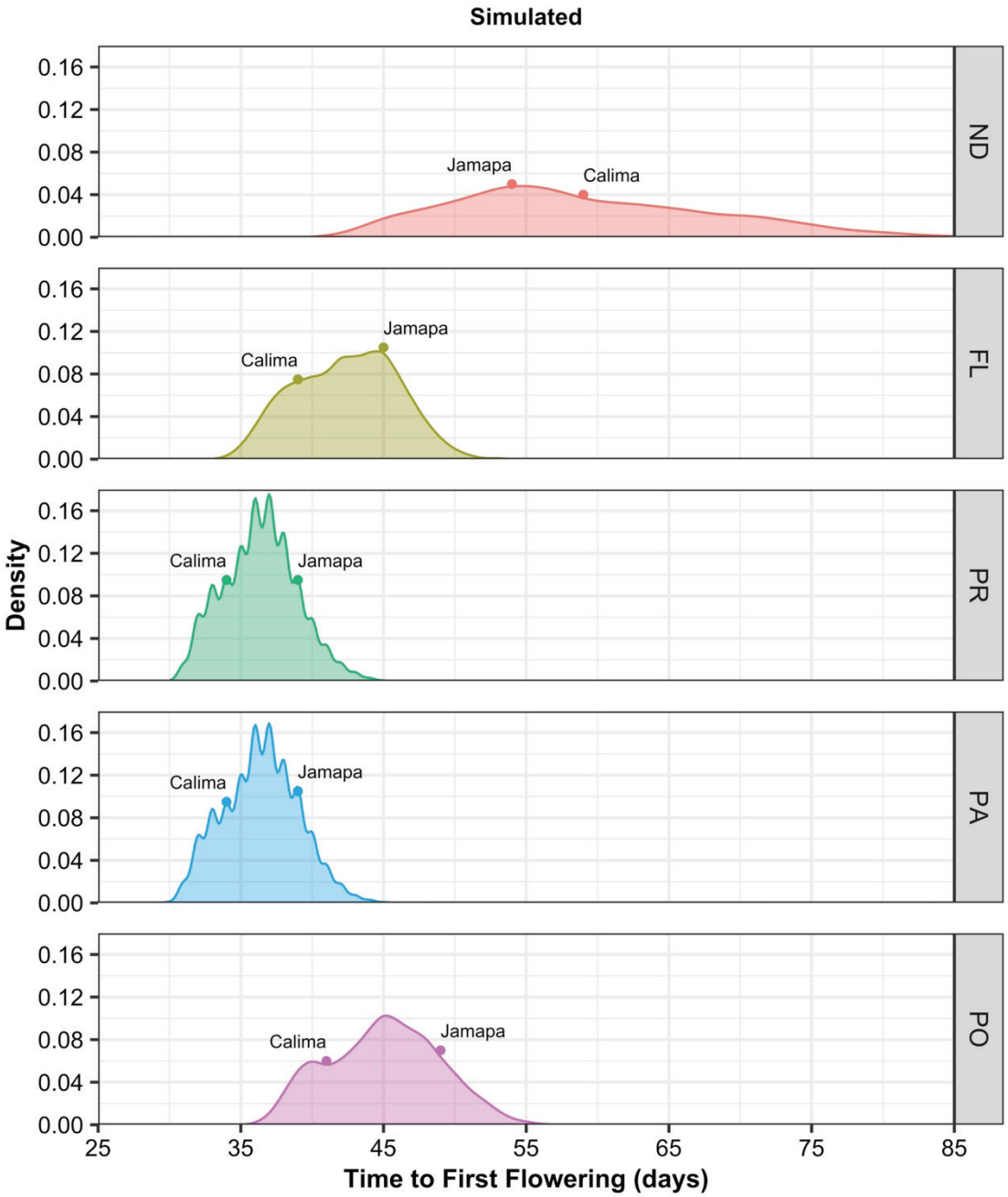

Figure 7. Density plots for simulated time to first flower (days) across the five sites showing all possible genetic combinations. The distribution of simulated days to flower by site includes all recombinant inbred line combinations (212 = 4096) as simulated by the dynamic piecewise linear module (DPLM), 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).

2.3.6. 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.* [38] for each individual RIL. For the MET crop management practices and one year of environmental conditions (Figure 1), simulated mean yield by CSMG was lowest for PA (190.0 ± 89 kg/ha) and highest for ND (637.3 ± 242 kg/ha) while for the other three sites mean yield ranged from 304.7 ± 135 kg/ha for FL, 505.0 ± 270.0 kg/ha for PO, and 540.0 ± 244.8 kg/ha for PR (Figure 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), while the lowest mean yield was obtained for PA (234.4 ± 104.2 kg/ha). Mean yield for FL was 354.8 ± 156.4 kg/ha, for PO was 625.0 ± 311.2 kg/ha, and for PR was 673.3 ± 264.7 kg/ha (Figure 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 (Figure 2 and Figure 5), while all other growth and development processes were simulated exactly the same as the CSMG model using the same inputs (Figure 4).

2.3.7. Further advancement in gene-based modeling

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.* [44]. Only minor modifications were needed to enable their dynamic model to be integrated as a module into the existing CSMG model [30], [40]. 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.* [44]. 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 dataset.

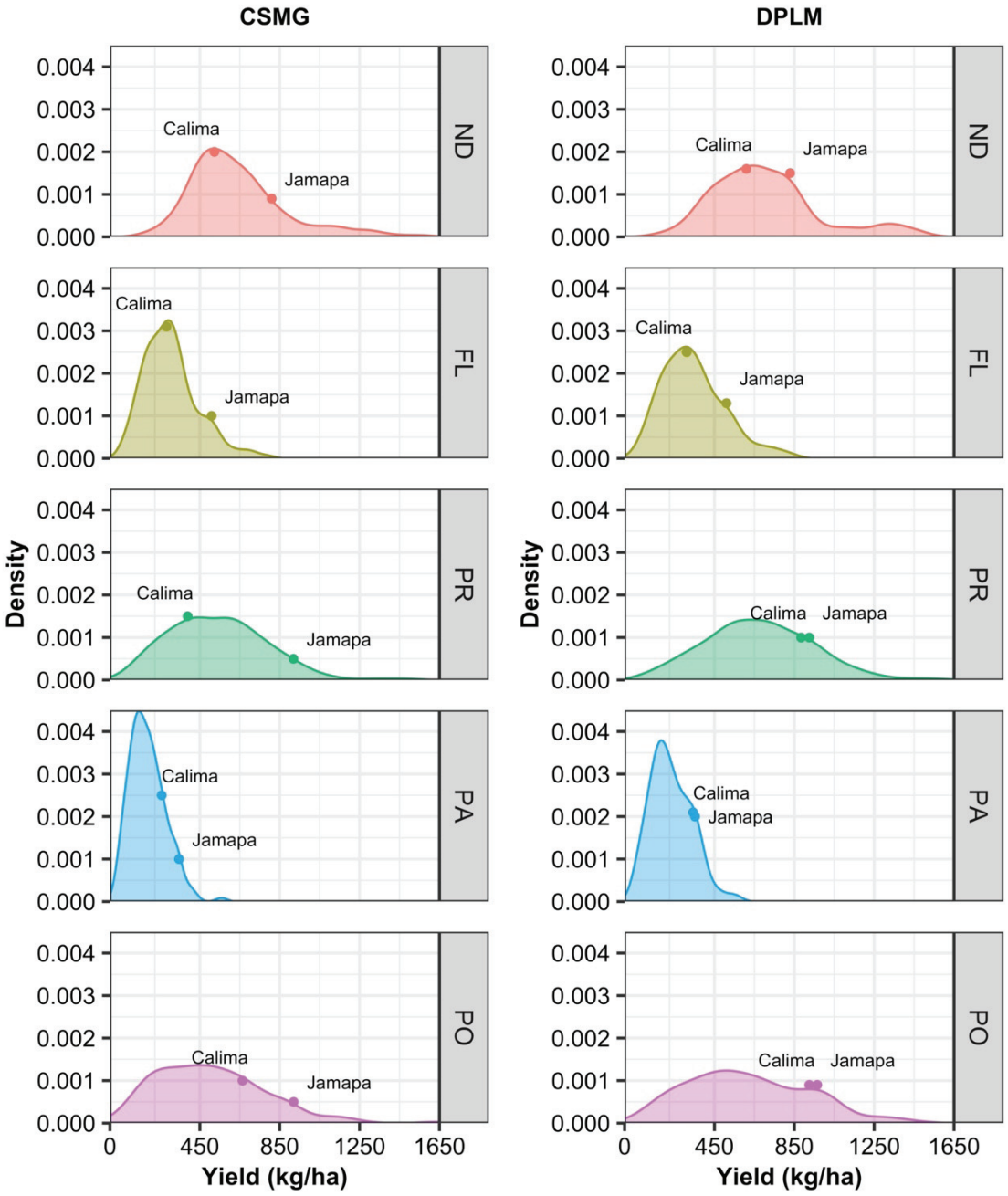

Figure 8. Density plot for simulated yield using the original CSM-CROPGRO-Drybean model and genetic specific coefficients (CSMG) (left panel) and the dynamic piecewise linear module (DPLM) integrated with the CSM-CROPGRO-Drybean model (right panel). The five experimental sites are: Prosper, North Dakota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

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 [53]. 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 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. Further work is required to improve the gene-based module and to add other processes that are linked dynamically with the crop model.

2.4. CONCLUSION

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 for which QTL information is readily available.

3. FINAL REMARKS

The extension of the gene-based module into the crop growth model for future processes are important to keep improving this new approach by connecting genetic mechanisms that increase the granularity of the model. Based on robust the structure of the CSM-CROPGRO-Drybean for phenotypic responses, and availability of the multi-environment trial in this work for further scientific advances a major target is to simulate main stem node number which can be determined by the time between two successive leaves.

Including this development phase, we aim to simulate main stem node number over time with an approach to evaluate the crop model response due to growth rate and duration of vegetative phase of development that is important for the success of the reproductive phase and affect the crop yield.

Further advancements are needed in the core of the gene-based module where development of loosely coupled components can be an option to extend and provide an easily dynamic manipulation of new equations and QTL information. These advances aim to continue extend and fully explore the existing crop growth model, and the gene-based module simulating and predicting yield for multi-environments and a wide range of cultivars accelerating the development of the next generation of gene-based crop models.

Also, the multi-disciplinary collaboration among crop modelers, plant breeders, geneticists, and physiologists are important to lead further advancing linking QTLs to markers for new cultivars and different crops the data availability to link other processes dynamically.

REFERENCES

[1] GODFRAY, H. C. J. *et al.* Food Security: The Challenge of Feeding 9 Billion People. *Science*. vol. 327, no. 5967. p. 812 LP – 818. 2010.

[2] FOLEY, J. A. *et al.* Solutions for a cultivated planet. *Nature*. vol. 478, no. 7369. p. 337–342. 2011.

[3] RAY, D. K. *et al.* Recent patterns of crop yield growth and stagnation. *Nature Communications*. vol. 3, no. 1. p. 1293. 2012.

[4] RAY, D. K. *et al.* Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLOS ONE*. vol. 8, no. 6. p. e66428. 2013.

[5] HATFIELD, J. L.; WALTHALL, C. L. Meeting Global Food Needs: Realizing the Potential via Genetics × Environment × Management Interactions. *Agronomy Journal*. vol. 107, no. 4. p. 1215–1226. 2015.

[6] JONES, J. W. *et al.* Brief history of agricultural systems modeling. *Agricultural Systems*. vol. 155. p. 240–254. 2016.

[7] MESSINA, C. D. *et al.* A Gene-Based Model to Simulate Soybean Development and Yield Responses to Environment. *Crop Science*. vol. 46, no. 1. p. 456–466. 2006.

[8] HWANG, C. *et al.* Next generation crop models: A modular approach to model early vegetative and reproductive development of the common bean (Phaseolus vulgaris L). *Agricultural Systems*. vol. 155. p. 225–239. 2017.

[9] ROSENZWEIG, C. *et al.* The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. *Agricultural and Forest Meteorology*. vol. 170. p. 166–182. 2013.

[10] BOOTE, K. J. *et al.* Putting mechanisms into crop production models. *Plant, Cell & Environment*. vol. 36, no. 9. p. 1658–1672. 2013.

[11] RASHEED, A. *et al.* Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Molecular Plant*. vol. 10, no. 8. p. 1047–1064. 2017.

[12] THOMSON, M. J. High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breeding and Biotechnology*. vol. 2, no. 3. p. 195–212. 2014.

[13] BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*. vol. 8, no. 12. p. e1002822. dez. 2012.

[14] HUANG, X.; HAN, B. Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annual Review of Plant Biology*. vol. 65, no. 1. p. 531–551. 2014.

[15] WHITE, J. W.; HOOGENBOOM, G. Gene-Based Approaches to Crop Simulation. *Agronomy Journal*. vol. 95, no. 1. p. 52–64. jan. 2003.

[16] WHITE, J. W. From genome to wheat: Emerging opportunities for modelling wheat growth and development. *European Journal of Agronomy*. vol. 25, no. 2. p. 79–88. 2006.

[17] YIN, X.; STRUIK, P. C. Modelling the crop: from system dynamics to systems biology. *Journal of Experimental Botany*. vol. 61, no. 8. p. 2171–2183. 2010.

[18] BHAKTA, M. S. *et al.* A predictive model for time-to-flowering in the common bean based on QTL and environmental variables. *G3: Genes, Genomes, Genetics*. vol. 7, no. 12. p. 3901–3912. 2017.

[19] SCHMUTZ, J. *et al.* A reference genome for common bean and genome-wide

analysis of dual domestications. *Nature Genetics*. vol. 46, no. 7. p. 707–713. 2014.
[20] YIN, X.; VAN DER LINDEN, C. G.; STRUIK, P. C. Bringing genetics and biochemistry to crop modelling, and vice versa. *European Journal of Agronomy*. vol. 100. p. 132–140. 2018.
[21] SPINDEL, J. E. *et al.* Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*. vol. 116, no. 4. p. 395–408. 2016.
[22] BROWN, T. B. *et al.* TraitCapture: genomic and environment modelling of plant phenomic data. *Current Opinion in Plant Biology*. vol. 18. p. 73–79. 2014.
[23] COOPER, M. *et al.* Use of Crop Growth Models with Whole-Genome Prediction: Application to a Maize Multienvironment Trial. *Crop Science*. vol. 56, no. 5. p. 2141–2156. 2016.
[24] THORBURN, P. J. *et al.* Recent advances in crop modelling to support sustainable agricultural production and food security under global change. *European Journal of Agronomy*. vol. 100. p. 1–3. 2018.
[25] HUNT, L. A. *et al.* GENCALC: Software to Facilitate the Use of Crop Models for Analyzing Field Experiments. *Agronomy Journal*. vol. 85, no. 5. p. 1090–1094. 1993.
[26] ANOTHAI, J. *et al.* A sequential approach for determining the cultivar coefficients of peanut lines using end-of-season data of crop performance trials. *Field Crops Research*. vol. 108, no. 2. p. 169–178. 2008.
[27] BUDDHABOON, C.; JINTRAWET, A.; HOOGENBOOM, G. Methodology to estimate rice genetic coefficients for the CSM-CERES-Rice model using GENCALC and GLUE genetic coefficient estimators. *The Journal of Agricultural Science*. vol. 156, no. 4. p. 482–492. 2018.
[28] WHITE, J. W.; HOOGENBOOM, G. Simulating Effects of Genes for Physiological Traits in a Process-Oriented Crop Model. *Agronomy Journal*. vol. 88, no. 3. p. 416–422. 1996.
[29] HOOGENBOOM, G.; JONES, J. W.; BOOTE, K. J. Modeling growth, development, and yield of grain legumes using SOYGRO, PNUTGRO, and BEANGRO: a review. vol. 35, no. 6. p. 2043–2056. 1992.
[30] HOOGENBOOM, G. *et al.* BEANGRO: A Process-Oriented Dry Bean Model with a Versatile User Interface. *Agronomy Journal*. vol. 86, no. 1. p. 182–190. 1994.
[31] YIN, X. *et al.* Coupling estimated effects of QTLs for physiological traits to a crop growth model: predicting yield variation among recombinant inbred lines in barley. *Heredity*. vol. 85, no. 6. p. 539–549. 2000.
[32] YIN, X. *et al.* Crop Modeling, QTL Mapping, and Their Complementary Role in Plant Breeding. *Agronomy Journal*. vol. 95, no. 1. p. 90–98. 2003.
[33] REYMOND, M. *et al.* 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*. vol. 131, no. 2. p. 664 LP – 675. 2003.
[34] HAMMER, G. L. *et al.* Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. *Journal of Experimental Botany*. vol. 61, no. 8. p. 2185–2202. 2010.
[35] GU, J. *et al.* 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*. vol. 114, no. 3. p. 499–511. 2014.

[36] TECHNOW, F. *et al.* Integrating Crop Growth Models with Whole Genome Prediction through Approximate Bayesian Computation. *PLOS ONE*. vol. 10, no. 6. p. e0130855. 2015.
[37] WALLACH, D. *et al.* A dynamic model with QTL covariables for predicting flowering time of common bean (Phaseolus vulgaris) genotypes. *European Journal of Agronomy*. vol. 101. p. 200–209. 2018.
[38] ACHARYA, S. *et al.* Reliability of genotype-specific parameter estimation for crop models: Insights from a Markov chain Monte-Carlo estimation approach. *Transactions of the ASABE*. vol. 60, no. 5. p. 1699–1712. 2017.
[39] BOOTE, K. J. *et al.* The CROPGRO model for grain legumes. in *Understanding Options for Agricultural Production*. 7th ed. G. Y. Tsuji, G. Hoogenboom, and P. K. Thornton, Eds. Dordrecht: Springer Netherlands. 1998. p. 99–128.
[40] JONES, J. W. *et al.* The DSSAT cropping system model. *European Journal of Agronomy*. vol. 18, no. 3. p. 235–265. 2003.
[41] HOOGENBOOM, G. *et al.* The DSSAT crop modeling ecosystem. in *Advances in crop modelling for a sustainable agriculture*. Burleigh Dodds Science Publishing. 2019. p. 173–216.
[42] BOOTE, K. J. *et al.* Genetic Coefficients in the CROPGRO–Soybean Model Florida Agric. Exp. Stn. Journal Series no. R-08652. Supported in part by grants from the United Soybean Board and the Soybean Research and Development Council. *Agronomy Journal*. vol. 95, no. 1. p. 32–51. 2003.
[43] BHAKTA, M. S.; JONES, V. A.; VALLEJOS, C. E. Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in Phaseolus vulgaris L. *PLoS ONE*. vol. 10, no. 1. p. 1–20. 2015.
[44] VALLEJOS, C. E. *et al.* Dynamic Gene-Based Ecophysiological Model to Predict Phenotype from Genotype and Environment Data. *Submitted in 2021*. 2021.
[45] WHITE, J. W.; HOOGENBOOM, G.; HUNT, L. A. A Structured Procedure for Assessing How Crop Models Respond to Temperature. *Agronomy Journal*. vol. 97, no. 2. p. 426–439. 2005.
[46] RITCHIE, J. T.; NESMITH, D. S. Temperature and Crop Development. *Modeling Plant and Soil Systems*. p. 5–29. 1991.
[47] BATES, D. *et al.* Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software; Vol 1, Issue 1 (2015)*. 2015.
[48] NASH, J. E.; SUTCLIFFE, J. V. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*. vol. 10, no. 3. p. 282–290. 1970.
[49] WALLACH, D. *et al.* Chapter 9 - Model Evaluation. in *Working with Dynamic Crop Models*. 3rd ed. D. Wallach, D. Makowski, J. W. Jones, and F. Brun, Eds. Academic Press. 2019. p. 311–373.
[50] KOBAYASHI, K.; SALAM, M. U. Comparing Simulated and Measured Values Using Mean Squared Deviation and its Components. *Agronomy Journal*. vol. 92, no. 2. p. 345–352. 2000.
[51] WALLACE, D. H.; ENRIQUEZ, G. A. Daylength and temperature effects on days to flowering of early and late maturing beans (Phaseolus vulgaris L.). *Journal of American Society for Horticultural Science*. vol. 105. p. 583–591. 1980.
[52] WALLACE, D. H. *et al.* Photoperiod, Temperature, and Interaction Effects on Days and Nodes Required for Flowering of Bean. *Journal of the American Society for Horticultural Science*. vol. 116, no. 3. p. 534–543.

[53] HOOGENBOOM, G.; WHITE, J. W.; MESSINA, C. D. From genome to crop: integration through simulation modeling. *Field Crops Research*. vol. 90, no. 1. p. 145–163. 2004.

SUPPLEMENTARY MATERIAL

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

RIL	QTL1	QTL2	QTL3	QTL4	QTL5	QTL6	QTL7	QTL8	QTL9	QTL10	QTL11	QTL12
Calima	1	1	1	1	1	1	1	1	1	1	1	1
Jamapa	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
RIJC001	-1	-1	-1	1	-1	-1	-1	-1	1	1	1	1
RIJC002	1	-1	1	1	1	-1	-1	-1	-1	1	-1	-1
RIJC003	1	1	1	1	1	1	-1	1	-1	-1	1	1
RIJC004	-1	-1	-1	-1	1	1	-1	1	1	1	-1	-1
RIJC005	1	1	1	-1	1	-1	-1	-1	-1	-1	1	1
RIJC006	1	1	1	1	1	1	-1	-1	1	-1	1	1
RIJC007	1	-1	1	1	1	1	-1	1	1	-1	-1	-1
RIJC008	1	1	1	1	-1	-1	-1	1	-1	-1	1	1
RIJC009	-1	-1	-1	1	1	-1	-1	1	-1	1	1	1
RIJC011	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1
RIJC012	1	1	1	1	-1	-1	-1	1	-1	-1	1	1
RIJC013	1	1	1	1	-1	-1	-1	-1	1	-1	1	1
RIJC014	1	1	1	1	1	1	1	-1	1	1	1	1
RIJC015	1	1	1	1	-1	1	-1	-1	1	-1	1	1
RIJC016	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1
RIJC017	-1	-1	1	1	-1	-1	-1	1	-1	-1	1	1
RIJC018	1	1	1	1	1	1	1	-1	1	1	-1	1
RIJC019	1	1	1	1	-1	-1	1	1	-1	1	-1	-1
RIJC020	1	1	-1	-1	-1	-1	-1	-1	1	1	-1	-1
RIJC021	1	1	1	-1	-1	1	1	1	-1	-1	-1	-1
RIJC022	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1
RIJC024	-1	-1	1	1	1	1	-1	1	-1	1	-1	-1
RIJC025	-1	1	1	-1	-1	-1	-1	-1	1	-1	1	1
RIJC026	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1
RIJC027	1	-1	1	1	1	1	-1	1	-1	1	-1	-1
RIJC029	1	1	1	1	-1	-1	1	1	-1	1	-1	-1
RIJC030	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	1
RIJC031	-1	-1	-1	-1	1	1	1	-1	-1	1	1	1
RIJC032	-1	-1	-1	-1	1	1	-1	-1	1	-1	-1	-1
RIJC045	1	1	1	1	-1	-1	-1	1	1	-1	-1	-1
RIJC046	1	-1	-1	-1	1	1	1	1	-1	1	-1	1
RIJC047	1	1	1	1	1	1	1	1	-1	-1	-1	-1
RIJC048	-1	-1	-1	-1	1	1	1	-1	-1	1	-1	-1
RIJC049	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
RIJC059	-1	-1	1	1	1	1	-1	-1	-1	1	1	1
RIJC061	-1	1	1	1	-1	-1	1	-1	1	1	1	1
RIJC062	1	1	1	-1	1	-1	-1	-1	1	1	-1	-1
RIJC064	1	1	-1	1	1	1	1	1	1	-1	-1	-1
RIJC065	1	-1	-1	1	1	1	1	1	-1	-1	-1	-1
RIJC066	1	1	1	1	1	1	1	1	1	1	1	1
RIJC067	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1	1
RIJC069	-1	-1	-1	-1	1	1	-1	-1	-1	1	1	1
RIJC070	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1
RIJC071	1	1	1	-1	1	1	1	1	-1	1	1	-1
RIJC072	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1	-1
RIJC073	-1	-1	-1	-1	1	1	-1	1	-1	-1	1	1
RIJC074	1	1	1	1	1	1	1	-1	1	-1	1	-1
RIJC075	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	1
RIJC076	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1
RIJC078	1	1	-1	1	1	-1	1	-1	-1	-1	1	1

RIJC079	-1	1	1	1	1	1	-1	-1	-1	1	1	1
RIJC080	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1
RIJC081	-1	1	-1	-1	1	1	1	-1	1	1	1	1
RIJC082	-1	-1	-1	-1	1	1	1	-1	-1	-1	1	1
RIJC129	1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1
RIJC130	1	1	-1	-1	-1	-1	1	1	-1	1	-1	1
RIJC131	1	1	1	1	1	1	-1	-1	1	1	1	-1
RIJC133	-1	1	1	1	1	1	1	-1	1	1	1	1
RIJC135	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1
RIJC136	-1	-1	-1	-1	1	1	-1	-1	-1	1	-1	-1
RIJC137	1	1	1	1	-1	-1	1	-1	-1	-1	1	1
RIJC138	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1
RIJC139	-1	-1	-1	-1	1	1	1	-1	-1	1	-1	-1
RIJC140	1	1	1	1	-1	-1	1	1	-1	-1	1	1
RIJC141	-1	-1	-1	-1	-1	-1	-1	1	1	-1	1	1
RIJC142	-1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1
RIJC144	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1
RIJC145	-1	-1	-1	1	1	1	-1	1	1	-1	-1	-1
RIJC146	-1	-1	-1	-1	1	1	1	1	-1	-1	1	1
RIJC147	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
RIJC148	1	1	1	1	-1	1	1	1	-1	1	-1	-1
RIJC149	-1	-1	-1	-1	1	1	1	-1	-1	-1	1	1
RIJC151	-1	-1	-1	-1	1	1	-1	1	1	1	1	1
RIJC201	1	1	1	1	-1	-1	1	-1	-1	-1	-1	-1
RIJC202	-1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1
RIJC203	-1	1	-1	-1	-1	-1	1	-1	1	1	1	1
RIJC204	-1	1	-1	-1	-1	-1	1	-1	1	1	1	1
RIJC205	1	-1	-1	-1	1	1	-1	-1	1	-1	1	1
RIJC206	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1
RIJC207	-1	-1	-1	-1	1	1	-1	-1	-1	1	1	1
RIJC208	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
RIJC209	-1	-1	-1	-1	1	1	1	-1	1	-1	1	1
RIJC210	1	-1	-1	1	-1	-1	1	-1	-1	-1	1	1
RIJC212	1	-1	1	1	1	1	1	-1	-1	-1	1	1
RIJC213	1	1	1	1	1	1	1	1	1	-1	1	1
RIJC214	1	1	-1	1	1	1	1	1	1	-1	1	1
RIJC216	1	-1	-1	1	-1	-1	1	-1	1	-1	1	1
RIJC217	1	1	1	-1	-1	-1	-1	-1	-1	-1	1	-1
RIJC218	1	-1	1	-1	1	1	-1	-1	1	-1	1	1
RIJC219	-1	-1	-1	-1	1	1	1	1	-1	-1	1	1
RIJC220	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1
RIJC221	1	-1	1	-1	1	-1	1	1	1	1	-1	1
RIJC223	1	1	1	-1	-1	-1	-1	1	-1	-1	1	1
RIJC224	1	-1	-1	1	-1	-1	-1	1	-1	1	1	1
RIJC225	-1	1	1	1	-1	1	-1	-1	-1	1	1	1
RIJC226	-1	-1	1	1	1	1	-1	1	1	1	1	1
RIJC229	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1	1
RIJC230	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1
RIJC231	1	1	1	1	1	-1	-1	-1	1	1	1	-1
RIJC232	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1
RIJC233	1	1	1	-1	1	1	1	1	-1	-1	-1	-1
RIJC234	1	1	1	1	-1	1	1	1	1	-1	-1	1
RIJC235	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1
RIJC236	-1	-1	-1	-1	-1	-1	1	1	-1	1	-1	-1
RIJC237	1	-1	-1	-1	1	1	1	1	-1	1	-1	-1
RIJC238	-1	-1	-1	-1	1	1	1	1	-1	1	-1	-1
RIJC242	-1	-1	-1	-1	-1	1	1	-1	-1	1	-1	-1
RIJC243	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1	-1
RIJC244	1	1	1	-1	1	1	1	-1	-1	-1	1	1
RIJC245	1	1	1	-1	1	1	-1	-1	-1	-1	1	1
RIJC247	-1	1	-1	1	-1	-1	1	1	-1	-1	-1	1
RIJC248	1	1	1	1	-1	-1	1	1	-1	-1	1	1
RIJC249	1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1
RIJC250	1	1	1	1	1	1	1	1	-1	-1	1	1
RIJC251	1	1	-1	-1	-1	1	1	-1	-1	1	1	1
RIJC252	1	1	1	1	-1	-1	1	-1	1	1	1	1
RIJC253	1	1	1	1	-1	-1	-1	-1	1	1	1	1

RIJC254	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
RIJC255	1	1	1	1	-1	-1	-1	-1	-1	1	-1	-1
RIJC256	-1	-1	1	1	1	1	1	-1	1	-1	1	1
RIJC257	1	1	1	1	-1	1	1	-1	-1	-1	-1	-1
RIJC259	-1	-1	-1	-1	1	1	1	1	-1	-1	1	1
RIJC261	-1	1	1	1	-1	-1	1	1	1	-1	-1	-1
RIJC262	-1	-1	-1	1	-1	-1	1	-1	1	-1	-1	-1
RIJC264	-1	-1	-1	-1	1	1	1	-1	1	1	1	1
RIJC301	1	-1	-1	-1	1	-1	1	-1	-1	1	-1	1
RIJC302	-1	1	1	1	1	1	-1	1	1	-1	-1	-1
RIJC303	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1
RIJC305	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1
RIJC306	-1	-1	-1	-1	1	1	-1	-1	1	-1	1	1
RIJC307	1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1
RIJC309	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1
RIJC310	1	1	1	1	1	1	-1	1	-1	-1	-1	-1
RIJC311	1	1	1	1	-1	-1	1	-1	1	-1	1	-1
RIJC312	-1	-1	-1	-1	1	1	1	-1	-1	1	1	1
RIJC313	1	1	1	1	-1	-1	-1	-1	1	-1	1	1
RIJC314	-1	1	1	1	-1	-1	1	-1	-1	1	1	1
RIJC316	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
RIJC317	1	-1	-1	-1	-1	-1	1	1	-1	1	1	1
RIJC318	-1	-1	1	1	-1	-1	1	1	1	-1	-1	-1
RIJC319	-1	1	-1	1	-1	-1	-1	1	-1	1	-1	-1
RIJC320	-1	-1	1	1	-1	-1	-1	1	-1	-1	1	1
RIJC321	-1	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1
RIJC322	1	1	1	1	-1	1	1	-1	1	1	1	1
RIJC325	-1	-1	-1	-1	1	1	1	1	1	1	1	1
RIJC326	1	1	-1	-1	1	1	1	-1	1	-1	-1	-1
RIJC327	1	1	1	1	-1	-1	1	-1	1	-1	1	-1
RIJC328	1	1	1	1	-1	-1	1	-1	1	-1	-1	-1
RIJC330	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1
RIJC332	-1	1	1	-1	1	1	-1	1	-1	1	-1	1
RIJC334	-1	-1	-1	-1	-1	-1	1	-1	1	-1	1	-1
RIJC335	1	1	1	1	-1	-1	-1	-1	1	-1	1	1
RIJC337	1	-1	-1	-1	-1	1	-1	-1	1	1	-1	1
RIJC339	-1	-1	-1	-1	-1	-1	1	-1	1	-1	1	1
RIJC340	-1	1	1	1	1	1	1	1	-1	1	1	1
RIJC341	1	1	1	1	1	-1	1	1	1	1	-1	-1
RIJC342	-1	1	-1	-1	-1	1	1	-1	1	-1	-1	-1
RIJC343	1	-1	-1	-1	1	1	1	1	1	1	1	1
RIJC344	-1	-1	-1	-1	1	-1	1	-1	-1	1	1	1
RIJC346	1	-1	-1	-1	-1	-1	-1	1	1	-1	1	1
RIJC347	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	1
RIJC348	1	1	1	1	-1	-1	1	-1	-1	-1	-1	-1
RIJC349	-1	-1	-1	-1	1	-1	1	1	-1	-1	1	1
RIJC350	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	1
RIJC351	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1
RIJC352	-1	-1	1	1	1	1	-1	1	-1	1	1	-1
RIJC353	-1	1	1	1	1	1	-1	1	-1	1	-1	-1
RIJC354	1	1	1	1	1	1	1	1	-1	-1	-1	-1
RIJC355	1	-1	1	1	1	1	-1	1	1	-1	-1	-1
RIJC356	-1	-1	-1	-1	1	1	-1	-1	1	-1	1	1
RIJC357	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
RIJC358	-1	-1	1	1	-1	-1	1	1	-1	-1	-1	-1
RIJC360	1	-1	1	-1	-1	-1	1	-1	1	-1	-1	-1
RIJC361	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
RIJC362	-1	-1	-1	-1	1	1	1	-1	1	1	1	-1
RIJC363	1	1	1	1	1	1	1	-1	-1	1	1	1
RIJC364	1	1	1	1	-1	1	-1	1	-1	1	1	1
RIJC366	1	1	-1	-1	1	1	1	-1	1	-1	-1	-1
RIJC367	1	1	1	1	1	1	-1	-1	1	1	1	1
RIJC368	1	1	1	1	-1	-1	-1	1	-1	1	-1	-1
RIJC369	1	1	1	1	-1	-1	-1	1	1	-1	1	1
RIJC370	-1	-1	1	1	1	-1	-1	-1	1	-1	-1	-1
RIJC371	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1	-1
RIJC372	1	1	1	1	-1	-1	-1	1	-1	1	1	-1

RIJC373	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1
RIJC374	1	1	-1	-1	-1	-1	1	-1	-1	1	-1	-1
RIJC375	-1	1	-1	1	-1	-1	1	-1	1	1	1	1

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

Terms [a]	Estimated	2.5%	97.5%	p-value
Intercept	2.35148×10^{-2}	2.33418×10^{-2}	2.36879×10^{-2}	9.01155×10^{-232}
$Tmax_{s,g}$	5.72311×10^{-4}	5.11997×10^{-4}	6.32626×10^{-4}	4.37596×10^{-62}
$Tmin_{s,g}$	5.29789×10^{-4}	4.87553×10^{-4}	5.72025×10^{-4}	1.32239×10^{-94}
$DayL_{s,g}$	-1.56357×10^{-3}	-1.65041×10^{-3}	-1.47673×10^{-3}	1.53907×10^{-152}
$Srad_{s,g}$	-8.50211×10^{-5}	-1.20053×10^{-4}	-4.99896×10^{-5}	2.42422×10^{-6}
$QTL1_{s,g}$	9.41278×10^{-4}	7.67693×10^{-4}	1.11486×10^{-3}	1.14674×10^{-20}
$QTL2_{s,g}$	1.24887×10^{-3}	1.05220×10^{-3}	1.44554×10^{-3}	6.18612×10^{-26}
$QTL3_{s,g}$	-6.08364×10^{-4}	-8.29898×10^{-4}	-3.86831×10^{-4}	2.26795×10^{-7}
$QTL4_{s,g}$	2.36803×10^{-4}	4.20996×10^{-5}	4.31506×10^{-4}	1.81926×10^{-2}
$QTL5_{s,g}$	5.67194×10^{-6}	-1.86555×10^{-4}	1.97899×10^{-4}	9.53950×10^{-1}
$QTL6_{s,g}$	5.27617×10^{-4}	3.36592×10^{-4}	7.18642×10^{-4}	2.07379×10^{-7}
$QTL7_{s,g}$	-4.11459×10^{-4}	-5.56487×10^{-4}	-2.66430×10^{-4}	9.99240×10^{-8}
$QTL8_{s,g}$	-2.11983×10^{-4}	-3.60161×10^{-4}	-6.38048×10^{-5}	5.61825×10^{-3}
$QTL9_{s,g}$	-4.42610×10^{-4}	-5.83805×10^{-4}	-3.01415×10^{-4}	5.28426×10^{-9}
$QTL10_{s,g}$	-2.50138×10^{-4}	-3.95428×10^{-4}	-1.04847×10^{-4}	9.14723×10^{-4}
$QTL11_{s,g}$	3.43389×10^{-4}	1.22771×10^{-4}	5.64006×10^{-4}	2.63477×10^{-3}
$QTL12_{s,g}$	-1.53677×10^{-4}	-3.72048×10^{-4}	6.46939×10^{-5}	1.69555×10^{-1}
$QTL1_{s,g} \times QTL2_{s,g}$	3.01579×10^{-4}	1.27364×10^{-4}	4.75794×10^{-4}	8.55786×10^{-4}
$DayL_{s,g} \times QTL3_{s,g}$	-7.66441×10^{-4}	-8.44228×10^{-4}	-6.88654×10^{-4}	4.92482×10^{-66}
$DayL_{s,g} \times QTL7_{s,g}$	-1.62459×10^{-4}	-2.21472×10^{-4}	-1.03446×10^{-4}	9.57607×10^{-8}
$DayL_{s,g} \times QTL12_{s,g}$	-1.45956×10^{-4}	-2.11416×10^{-4}	-8.04965×10^{-5}	1.44756×10^{-5}
$Tmin_{s,g} \times QTL2_{s,g}$	-2.40896×10^{-5}	-5.95337×10^{-5}	1.13545×10^{-5}	1.83307×10^{-1}
$Tmin_{s,g} \times QTL3_{s,g}$	-8.59042×10^{-5}	-1.28746×10^{-4}	-4.30624×10^{-5}	9.42280×10^{-5}
$Tmax_{s,g} \times QTL5_{s,g}$	8.54093×10^{-5}	3.87125×10^{-5}	1.32106×10^{-4}	3.62986×10^{-4}
$Srad_{s,g} \times QTL12_{s,g}$	-3.06273×10^{-5}	-6.04701×10^{-5}	-7.84445×10^{-7}	4.46841×10^{-2}
Fixed effects variance	1.80182×10^{-5}			
Random effects variance	6.34775×10^{-7}			
Residual variance	1.30567×10^{-6}			

[a]Predictor variable are: Intercept represents the overall value of daily development rate, maximum and minimum temperature (Tmax and Tmin, °C), day length (DayL, hours) and solar radiation (Srad, MJ m^{-2} d^{-1}). The rate of progress to flowering rate is (1/duration from plating to flowering in days) for the g^{th} genotype for the s^{th} site; QTL1 to QTL12 allele effects are represented as +1 for Calima alleles and -1 for Jamapa alleles.

Figure S1. Parameter estimation process for predicting first flowering across all sites using the Dynamic Mixed Linear Module (DMLM)

```
## cleaning memory and loading R libraries
rm(list=ls())
library("readr")
library(stats)
library(lmerTest)
library(tidyverse)

### Setup your directory location
setwd("C:\path\to\my\directory")

##Environmental data
we <- read.table("C:\path\to\environment\data",header=T,sep=',')

# Example of table structure to read the environmental data.
# (data.frame(SITE = character(),
#             Srad = numeric(),
#             Tmax = numeric(),
#             Tmin = numeric(),
#             DayL = numeric()))

##Recombinant inbred lines (see Table S1)
r1 <- read_csv("C:\path\to\QTL\data")


## Part 1 - Parameter estimation

## First of all, provide the observed days to first flowering for each RIL
## and store it as a new column (r1$R1). Then extract the mean of
## environmental data from sowing ## to first flower for each RIL and store
## it as a new column for r1$Sradm, r1$Tminm, r1$Tmaxm and r1$DayLm.


## Mean of environmental data from sowing to first flower across all
## genotypes, sites, years.
#sradMean <- mean(r1$Sradm)
#dayMean  <- mean(r1$DayLm)
#tminMean <- mean(r1$Tminm)
#tmaxMean <- mean(r1$Tmaxm)


##Centering continuous variable
r1$Srad_c <-r1$Sradm - sradMean
r1$Day_c  <-r1$DayLm - dayMean
r1$Tmin_c <-r1$Tminm - tminMean
r1$Tmax_c <-r1$Tmaxm - tmaxMean

### 1/days to first flowering
r1$R1rate <- 1/r1$R1

##Dynamic mixed-effects linear model for first flowering rate
modelr1rate = lmer((R1rate)~
                     Tmax_c+
                     Tmin_c+
                     Day_c+
                     Srad_c+
                     QTL1+QTL2+QTL3+QTL4+QTL5+QTL6+
                     QTL7+QTL8+QTL9+QTL10+QTL11+QTL12+
                     QTL1*QTL2+
```

```
                      Day_c*QTL3+
                      Day_c*QTL7+
                      Day_c*QTL12+
                      Tmin_c*QTL2+
                      Tmin_c*QTL3+
                      Tmax_c*QTL5+
                      Srad_c*QTL12+
                      (1|RIL),data=r1)
summary(modelr1rate)

## Part 2 - Estimating daily flowering rate for each RIL at each SITE
predictionlist <- list()
predictionlist[[1]] =
rbind(predictionlist, c("SITE","RIL",'Observed','Simulated'))
rowcount = 1
stepper <-1

## Change the site based on your data
si <-  c("ND","FL","PR","PA","PO")
for (s in si){

  r2<-subset(r1,r1$SITE==s)  # subset QTL data file
  we1 <- subset(we,SITE==s)  # subset weather data file

  for (i in 1:nrow(r2)) {
    SITE  <- c(as.character(r2$SITE[i]))
    rsite <- SITE
    RIL   <- c(as.character(r2$RIL[i]))
    Observed <- c(r2$R1[i])
    QTL1  <- c(r2$QTL1[i])
    QTL2  <- c(r2$QTL2[i])
    QTL3  <- c(r2$QTL3[i])
    QTL4  <- c(r2$QTL4[i])
    QTL5  <- c(r2$QTL5[i])
    QTL6  <- c(r2$QTL6[i])
    QTL7  <- c(r2$QTL7[i])
    QTL8  <- c(r2$QTL8[i])
    QTL9  <- c(r2$QTL9[i])
    QTL10 <- c(r2$QTL10[i])
    QTL11 <- c(r2$QTL11[i])
    QTL12 <- c(r2$QTL12[i])

    CounterR1 = 0
    DayCount = 0


    for (j in 1:nrow(we1)){
      DayCount <- we1$DAP[j]

      # Adjusting weather variables
      Srad_c <- c(we1$Srad[j])-sradMean
      Day_c <- c(we1$DayL[j])-dayMean
      Tmax_c <- c(we1$Tmax[j])-tmaxMean
      Tmin_c <- c(we1$Tmin[j])-tminMean
      r1day <- as.data.frame(SITE,RIL,QTL1,QTL2,QTL3,QTL4,QTL5,
                             QTL6,QTL7,QTL8,QTL9,QTL10,QTL11,QTL12,
                             Srad_c,Tmin_c,Tmax_c,Day_c)

      ## Estimating daily flowering rate gain based on "modelr1rate" model
      flowerNow <- predict(modelr1rate,re.form=NA,newdata=r1day)
```

```
      ## counting cumulative flowering rate
      CounterR1 = CounterR1 + flowerNow[[1]]
      dailyRateR1 = flowerNow[[1]]
      rowcount = rowcount + 1

      ## exiting loop when rate reaches 1.00
      if (CounterR1[[1]] >= 1.00){
        Simulated <- DayCount
        predictionlist[[rowcount]] <- c(SITE,RIL,Observed,Simulated)
        break}
    }

    stepper = stepper +1
    print(c(SITE,stepper))
  }
}

# converting list to matrix
predictionRate <- as.matrix(do.call("rbind", predictionlist))
colnames(predictionRate)<-predictionRate[1,]  # fixing header row names
predictionRate<-predictionRate[-1,] # removing old header row

##View daily rate prediction for QTL allele combination
View(predictionRate)

#END
```

Figure S2. Computer code for the Dynamic Piecewise Linear Module (DPLM) coupled with the CSM-CROPGRO-Drybean model

```
!=====================================================================
!  Dynamic piecewise linear module, Program,
!=====================================================================
!---------------------------------------------------------------------

      SUBROUTINE DPLM(CONTROL, ISWITCH,                    &  !Control
          WEATHER, YRPLT,                                  &  !Input
          NR1G, SumFRD)                                       !Output

!---------------------------------------------------------------------
      USE ModuleDefs

      IMPLICIT none
      SAVE
!---------------------------------------------------------------------
      CHARACTER*6   GENID
      CHARACTER*30 FILEIO

      LOGICAL FRSTFL

      INTEGER RUN, DYNAMIC, DAS, YRDOY, YR, DOY, YRPLT
      INTEGER DAP, TIMDIF, FDOY, NR1G

      REAL DAYL, SRAD, TMAX, TMIN
      REAL SumFRD, FR
      REAL FRMAX, DLm, Sradm, Tmaxm, Tminm
      REAL, DIMENSION(70) :: QTL

      TYPE (ControlType) CONTROL
      TYPE (SwitchType) ISWITCH
      TYPE (WeatherType) WEATHER

      DYNAMIC = CONTROL % DYNAMIC
      FILEIO  = CONTROL % FILEIO
      RUN     = CONTROL % RUN
      DAS     = CONTROL % DAS
      YRDOY   = CONTROL % YRDOY


!***********************************************************************
!***********************************************************************
!     Run Initialization - Called once per simulation
!***********************************************************************
      IF (DYNAMIC .EQ. RUNINIT) THEN
!-----------------------------------------------------------------------
!       Read Genetic input data
!-----------------------------------------------------------------------
        CALL IPGENE(FILEIO, TF, GENID)

!***********************************************************************
!***********************************************************************
!     Seasonal initialization - run once per season
!***********************************************************************
      ELSEIF (DYNAMIC .EQ. SEASINIT) THEN
!-----------------------------------------------------------------------
!     Set sowing/start day of year for flowering model to start
!     Initialize progress toward flowering, SumFRD, & Day of First Flower
!-----------------------------------------------------------------------
```

```
        SumFRD = 0.0
        DAYL    = 0.0
        SRAD    = 0.0
        TMAX    = 0.0
        TMIN    = 0.0
        FDOY    = 0
        NR1G    = 0
        FRSTFL  = .FALSE.
        GENID   = ''
!-----------------------------------------------------------------------
!     Averages across 5 environments in datasets used to estimate model
!     mean values of environmental variables.
!-----------------------------------------------------------------------
        DLm    = 12.722559638877
        Sradm  = 18.2718980213904
        Tmaxm  = 27.4529030160428
        Tminm  = 16.1181873475936

!-----------------------------------------------------------------------
!     Limit maximum rate for a genotype based on QTLs, (FRMAX)
!-----------------------------------------------------------------------
        FRMAX  = 0.02798564527544710                &
               + 0.00107126855594358 * QTL(1)       &
               + 0.00124937987119220 * QTL(2)       &
               - 0.00035350501343249 * QTL(3)       &
               + 0.00039945509894467 * QTL(4)       &
               + 0.00007085168560867 * QTL(5)       &
               + 0.00056306276150221 * QTL(6)       &
               - 0.00042854934463711 * QTL(7)       &
               - 0.00023509947596009 * QTL(8)       &
               - 0.00060905231969296 * QTL(9)       &
               - 0.00029702065347147 * QTL(10)      &
               + 0.00063538481240068 * QTL(11)      &
               - 0.00023169840751149 * QTL(12)

!***********************************************************************
!***********************************************************************
!     Daily Rate calculations
!***********************************************************************
      ELSE IF (DYNAMIC .EQ. RATE) THEN
!-----------------------------------------------------------------------
!       Read Weather data
!-----------------------------------------------------------------------
        CALL YR_DOY(YRPLT,YR,DOY)
        DAP = MAX (0, TIMDIF(YRPLT, YRDOY))

        SRAD = WEATHER%SRAD
        TMAX = WEATHER%TMAX
        TMIN = WEATHER%TMIN
        DAYL = WEATHER%DAYL
!-----------------------------------------------------------------------
!       The dynamic gene-based mixed effects linear model, Bean
!-----------------------------------------------------------------------
         FR = 0.02351482064754700                                  &
            + 0.00057231114859053 * (TMAX - Tmaxm)                 &
            + 0.00052978909146124 * (TMIN - Tminm)                 &
            - 0.00156356767055738 * (DAYL - DLm)                   &
            - 0.00008502111815422 * (SRAD - Sradm)                 &
            + 0.00094127790142489 * QTL(1)                         &
            + 0.00124887042689091 * QTL(2)                         &
            - 0.00060836426965787 * QTL(3)                         &
```

```
             + 0.00023680275595848 * QTL(4)                          &
             + 0.00000567194207821 * QTL(5)                          &
             + 0.00052761690606345 * QTL(6)                          &
             - 0.00041145860322971 * QTL(7)                          &
             - 0.00021198302954652 * QTL(8)                          &
             - 0.00044260992328382 * QTL(9)                          &
             - 0.00025013760365754 * QTL(10)                         &
             + 0.00034338877075107 * QTL(11)                         &
             - 0.00015367693845455 * QTL(12)                         &
             + 0.00030157871435221 * QTL(1) * QTL(2)                 &
             - 0.00076644059984278 * (DAYL - DLm) * QTL(3)           &
             - 0.00016245875627569 * (DAYL - DLm) * QTL(7)           &
             - 0.00014595603452997 * (DAYL - DLm) * QTL(12)          &
             - 0.00002408961229346 * (TMIN - Tminm) * QTL(2)         &
             - 0.00008590422032661 * (TMIN - Tminm) * QTL(3)         &
             + 0.00008540931051535 * (TMAX - Tmaxm) * QTL(5)         &
             - 0.00003062728932614 * (SRAD - Sradm) * QTL(12)

     ! Note that one can replace the above mixed effects linear model with
     ! any function that computes each day's rate of progress toward
     ! first flowering

!***********************************************************************
!***********************************************************************
!     Daily integration
!***********************************************************************
      ELSEIF (DYNAMIC .EQ. INTEGR) THEN

!-----------------------------------------------------------------------
!     Compute time integral of development to pass back as cumulative
!       progress toward development each day
!     In the equation for computing SumFRD, the time step is assumed
!       to be 1.0 d for this module (fixed)
!-----------------------------------------------------------------------
!     Limit rate of development to positive values;
!       initial value=0.0. When SumFRD first reaches 1.00,
!       flowering will occur
!-----------------------------------------------------------------------

        IF (FR < 1E-5) THEN
          FR = 0.0
        ENDIF

        IF (FR > FRMAX) THEN
          FR = FRMAX
        ENDIF

        SumFRD = SumFRD + FR*1.0

        IF (SumFRD >= 1.0 .AND. FDOY < 1) THEN   !First flower occurs
          FRSTFL = .TRUE.
          FDOY = DAP + DOY
          NR1G = DAS
        ENDIF

!***********************************************************************
!***********************************************************************
!     END OF DYNAMIC IF CONSTRUCT
!***********************************************************************
      ENDIF
```

```
      RETURN
    END ! DPLM

!-----------------------------------------------------------------------
!     DPLM VARIABLES LIST
!-----------------------------------------------------------------------
! DAP         Number of days after planting (d).
! DAS         Days after start of simulation (d).
! DAYL        Day length on day of simulation (from sunrise to sunset) (hr).
! DLm         Mean day length across all five sites, all genotypes.
! DOY         Current day of simulation (d).
! DYNAMIC     Module control variable.
! FDOY        Number of days after planting when the first flower occurs
(d).
! FRSTFL      Flag to identify that the first flower occurs (true/false).
! GENID       Identifier for reading in the input file '.gen'.
! NR1G        Number of days after start of simulation when the first flower
!             occurs (d).
! FR          Daily rate of progress from planting to first flower
appearance
!             for selected genotype and environmental factors on the current
!             environment & day.
! FRMAX       Maximum rate of progress toward first flower.
! SRAD        Solar radiation (MJ/m2-d).
! Sradm       Mean solar radiation transplanting to first flower across all
!             genotypes, sites, years.
! SumFRD      Current progress toward first flowering of FR.
! QTL(n)      Alleles at QTL(1) : QTL(n) in jth genotype.
! TIMDIF      Integer function which calculates the number of days between
!             two Julian dates (da).
! TMAX        Maximum daily temperature (Celsius).
! Tmaxm       Mean of maximum temperature from transplanting to first flower
!             across all genotypes, sites, years.
! TMIN        Minimum daily temperature (Celsius)
! Tminm       Mean of minimum temperature from transplanting to first flower
!             across all genotypes, sites, years.
! YR          Year portion of date
! YRDOY       Current day of simulation (YYYYDDD)
! YRPLT       Planting date (YYYYDDD)
```

UPF
UNIVERSIDADE
DE PASSO FUNDO

UPF Campus I - BR 285, São José
Passo Fundo - RS - CEP: 99052-900
(54) 3316 7000 - www.upf.br