

Title: Evaluating differences among crop models in simulating soybean in-season growth

Author names and affiliations: Kritika Kothari^{1,2}, Rafael Battisti³, Kenneth J. Boote⁴, Sotirios V. Archontoulis⁵, Adriana Confalone⁶, Julie Constantin⁷, Santiago V. Cuadra⁸, Philippe Debaeke⁷, Babacar Faye⁹, Brian Grant¹⁰, Gerrit Hoogenboom^{4,11}, Qi Jing¹⁰, Michael van der Laan^{12,13}, Fernando Antônio Macena da Silva⁸, Fabio R. Marin¹⁴, Alireza Nehbandani¹⁵, Claas Nendel^{16, 17,18}, Larry C. Purcell¹⁹, Budong Qian¹⁰, Alex C. Ruane²⁰, Céline Schoving⁷, Evandro H.F.M. Silva¹⁴, Ward Smith¹⁰, Afshin Soltani¹⁵, Amit Srivastava²¹, Nilson A. Vieira Jr.¹⁴, Montserrat Salmerón^{1*}

¹University of Kentucky, Department of Plant and Soil Sciences, Lexington, KY, United States;

²Indian Institute of Technology Roorkee, Department of Water Resources Development and Management, Uttarakhand, India;

³Universidade Federal de Goiás, Escola de Agronomia, Goiânia, Brazil;

⁴University of Florida, Agricultural and Biological Engineering Department, Gainesville, FL, United States;

⁵Iowa State University, Department of Agronomy, Ames, IA, United States;

⁶Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Agronomía, NACT-NAACCE, Azul, Buenos Aires, Argentina;

⁷Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), Université Fédérale de Toulouse, UMR AGIR, Castanet-Tolosan, France;

⁸Empresa Brasileira de Pesquisa Agropecuária, São Paulo, Brazil;

⁹Université du Sine Saloum EL-HÂDJ IBRAHIMA NIASS, UFR 2SE, Kaolack, Sénégal

- 22 ¹⁰Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Canada;
- 23 ¹¹University of Florida, Global Food Systems Institute, Gainesville, FL, United States;
- 24 ¹²University of Pretoria, Department of Plant and Soil Sciences, Pretoria, South Africa;
- 25 ¹³Agricultural Research Council (ARC) – Natural Resources and Engineering, Pretoria, South
26 Africa;
- 27 ¹⁴University of São Paulo, College of Agriculture “Luiz de Queiroz” – Esalq, Biosystems
28 Engineering Department, São Paulo, Brazil;
- 29 ¹⁵Gorgan University of Agricultural Sciences and Natural Resources, Agronomy Group, Gorgan,
30 Iran;
- 31 ¹⁶Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany;
- 32 ¹⁷University of Potsdam, Institute of Biochemistry and Biology, Potsdam, Germany;
- 33 ¹⁸Czech Academy of Sciences, Global Change Research Institute, Brno, Czech Republic;
- 34 ¹⁹University of Arkansas, Department of Crop, Soil, and Environmental Sciences, Fayetteville,
35 Arkansas;
- 36 ²⁰NASA Goddard Institute for Space Studies, New York, NY, United States;
- 37 ²¹University of Bonn, Institute of Crop Science and Resource Conservation, Bonn, Germany;
- 38 ***Corresponding author** (msalmeron@uky.edu)

ABSTRACT

Crop models are useful tools for simulating agricultural systems that require continued model development and testing to increase their robustness and improve how they describe our current understanding of processes. Coordinated and “blind” evaluations of multiple models using same protocols and experimental datasets provides unique opportunities to further improve models and enhance their reliability. For soybean [*Glycine max* (L.) Merr.], there has been limited coordinated multi-model evaluations for the simulation of in-season plant growth dynamics. We evaluated ten dynamic soybean crop models for their simulation of in-season plant growth using data from five experiments conducted in Argentina, Brazil, France, and USA. We evaluated models after a *Blind* (using only phenology data) and a *Full* calibration (with in-season and end-of-season variables). Calibration reduced model uncertainty by reducing standard bias for the simulation of in season variables (biomass, leaf, pod, and stem weights, and leaf area index, LAI). However, we found that most models had difficulty in reproducing leaf growth dynamics, with normalized root mean squared error (nRMSE) of 56% for leaf weight and 43% for LAI (across locations and models after *Full* calibration). Models with different levels of complexity and experience were capable of simulating final seed yield at maturity with reasonable accuracy (nRMSE of 8 to 31% after *Full* calibration). However, the nRMSE for pod weight (of 17 to 64% after *Full* calibration) was two-fold larger than that of seed yield. Moreover, the models differed in how they simulated timing from sowing to beginning seed growth (47 to 93 days) and effective seed filling period (18 to 54 days), owing to model structural differences in defining the reproductive developmental stages. Overall, we identified the following processes that can benefit from further model improvement: leaf expansion and senescence, reproductive phenology, and partitioning to reproductive growth. Simulation of pod wall tissue and individual seed cohorts is

another aspect that many models currently lack. Model improvement can benefit from high-temporal resolution experimental datasets that concurrently account for phenology, plant growth, and partitioning. Further, we recommend collecting reproductive phenology in the field consistent with actual dry matter allocation to organs in the models and collecting multiple observations of seed and pod weight to aid model improvement for simulation of seed growth and yield formation.

Keywords. Agricultural Model Intercomparison and Improvement Project (AgMIP), multi-model evaluation, pod growth, leaf area index, reproductive partitioning

Abbreviations. R1, Beginning flowering stage; R3, Beginning pod stage; R5, Beginning seed stage; R7, Physiological maturity stage; RMSE, Root mean squared error; nRMSE, Normalized root mean squared error; MSE, Mean squared error; LC, Lack of correlation; SB, standard bias; NU, Non-unity of slope

1. INTRODUCTION

With an annual production of 301 million tons worldwide (FAO, 2020), soybean is the most widely grown legume and a major component of the global food supply. Ensuring high yield stability and the sustainability of soybean production under current and future climate conditions is essential to meet the growing food, feed, and nutrition demand. Process-based crop physiological models are powerful tools to simulate crop growth under climate change and to study potential management and genotype adaptations (Rosenzweig et al., 2014; White et al., 2011). However, models require calibration and evaluation with experimental data prior to their application. While the research focusing on crop model applications has increased in the past decades, the efforts for model evaluation and improvement with detailed field experimental data are lagging (Silva and Giller, 2021). Therefore, new studies that evaluate model simulations with

84 experimental data that include in-season crop growth are paramount to reduce model uncertainty
85 before their application.

86 An essential component of model improvement is the calibration and evaluation of models
87 with detailed experimental data that include in-season crop samplings. Model calibration with only
88 end-of-season measured data may result in model parameters that give unrealistic simulation of
89 in-season growth variables (Martre et al., 2015). Uncertainty in simulating in-season growth
90 dynamics could cause incorrect simulated response to climatic factors occurring during critical
91 growth phases. A recent soybean multi-model evaluation revealed high uncertainty across models
92 in their response to temperature and atmospheric CO₂ concentration, as well as unexpected effect
93 of model calibration on the temperature response of yield depending on the model and location
94 (Kothari et al., 2022). A literature review of the ten models participating in the study by Kothari
95 et al. (2022) showed that most of those models were relatively “new” or inexperienced in
96 simulating soybean growth (**Supplementary Fig. S1**). Seven out of ten models had less than five
97 publications that mentioned biomass and leaf area, and only one of the models had published
98 articles that mentioned in-season variables such as pod weight and seed weight (**Supplementary**
99 **Fig. S1**). Therefore, modeling efforts focused on model calibration and evaluation for simulation
100 of in-season growth data are still limited for most models capable of simulating soybean growth.
101 A comprehensive evaluation of soybean models with detailed measurements of in-season growth
102 is therefore essential for assessing the current state of soybean models and for identifying
103 opportunities of model improvement.

104 Multi-model comparison studies can be important contributors to crop model
105 improvement. A unique advantage of coordinated multi-model studies is that all models are
106 evaluated against the same set of independent experimental data. Another advantage is that model

calibration and model evaluation against experimental data is performed by independent teams, allowing to quantify model uncertainty under real “blind” conditions. Coordinated multi-model comparisons under the umbrella of the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013), have been useful in the past to find avenues for model improvement. In a multi-model study of 30 wheat crop models across seven contrasting environments, Asseng et al. (2015) found that simulated yield tended to deviate more from measurements at relatively warmer temperatures, and identified an opportunity to improve wheat model simulations under climate change with improved heat stress responses for plant growth and yield formation (Wang et al., 2017). Hasegawa et al. (2017) evaluated 16 rice models using data from free-air CO₂ enrichment experiments and controlled chambers, and found that yield simulations under elevated CO₂ could be improved by driving leaf area development in models by rice morphology rather than directly by photosynthesis. Lastly, a recent multi-model study for simulation of wheat phenology revealed that better prediction accuracy not only results from better model structure but also depends on the choice of parameters to estimate and the calibration approach (Wallach et al., 2021). This approach of coordinated multi-faceted studies across different sites and multiple models is still limited for the analysis of in-season growth dynamics.

For soybean, there have been few prior multi-model evaluations with experimental data (Battisti et al., 2017b; Kothari et al., 2022). Battisti et al. (2017b) evaluated five soybean models for simulation of yield against field data from southern Brazil, with limited focus on the analysis of model performance for simulation of in-season biomass and leaf area index (LAI). The multi-model study by Kothari et al. (2022) included ten models of variable structure, complexity, and prior experience simulating soybean growth. The study by Kothari et al. (2022) revealed similar performance across most models for simulation of yield after calibration with measured data, but

high uncertainty under blind conditions, or when simulated for hypothetical yield responses to changes in temperature and [CO₂]. A comprehensive evaluation of soybean models for simulation of growth against detailed time-series data from multiple sites can provide new opportunities for improving the simulation of in-season growth dynamics and reproductive partitioning, and reduce model uncertainty for climate-change applications.

The goal of this study was to conduct a multi-model evaluation of soybean crop models with respect to their simulation of in-season growth. For that we used detailed experimental data from five sites (Azul, Argentina; Auzeville, France; Brasilia, Brazil; Fayetteville, Arkansas, and Ames, Iowa, USA). We utilized ten soybean models that had been previously evaluated for the simulation of final grain yield across the same sites (Kothari et al., 2022). The specific objectives were to 1) quantify model performance and uncertainty for simulation of in-season growth and partitioning under two levels of calibration; and 2) identify approaches and mechanisms that can lead to improvement of model performance for simulation of in-season growth.

2. MATERIALS AND METHODS

2.1. Field Experiments and Model Inputs

The experimental data for the simulation exercise were obtained from five locations: Brasilia, Brazil (BRZL; 15°35'33"S, 47°44'12"W); Azul, Argentina (ARGN; 36°27'S, 59°30'W); Auzeville, France (FRNC; 43°31'38" N, 1°30'22" E); and Fayetteville, AR (AKNS; 36°05'52.8"N, 94°10'44.4"W) and Ames, IA (IOWA; 42°01'20.37"N, 93°46'36.05"W) in the United States. These sites were selected to include major soybean growing areas and to provide variability in environmental conditions to evaluate models. Across the five sites, BRZL, IOWA, and ARGN

were the wettest, with a mean precipitation of 640, 576, and 517 mm, respectively, during the crop growing season (from sowing to harvest). FRNC and AKNS were the driest (253 and 265 mm, respectively), although AKNS was irrigated. The five sites in our study also provided a wide range of average growing season air temperatures: ARGN (19.9 °C), FRNC (21.2 °C), IOWA (21.6 °C), BRZL (22.7 °C), and AKNS (25.0 °C).

A complete description of the experimental data from each site is provided in Kothari et al. (2022). In brief, field experiments were conducted in a randomized block design with three to four replications and were carried out over several years, under different planting dates or irrigation management, depending on the location. Experiments in ARGN and FRNC included irrigated and rainfed treatments. Other sites had all treatments managed either under irrigation (AKNS) or under rainfed conditions (BRZL and IOWA). The same soybean cultivar was used across years and treatments within each location. Across locations, the maturity groups (MG) ranged from MG 1 (FRNC) to MG 6 (BRZL). Plot size ranged from 10 to 778 m² depending on the site. Sowing density ranged from 29 to 50 seeds m⁻². Plots consisted of four to six rows planted in 30- to 76-cm row spacing.

Dates of onset of developmental stages were monitored at all sites according to Fehr and Caviness (1977). Stages recorded included the dates for emergence (VE), beginning flower (R1), beginning pod (R3), beginning seed (R5), beginning (physiological) maturity (R7), and harvest maturity (R8). Destructive in-season samplings were collected 3 to 25 times during the growing season depending on the location and year, by sampling 0.5 to 1 m² of crop land area to quantify aboveground biomass. A subsample of 3 to 6 plants was separated into leaves, stems, and pods; then the leaves were used to measure leaf area. The in-season data available for calibration in this study consisted of total aboveground biomass, LAI, and weights of individual plant components

(leaf, stem, and pod). Soybean yield was quantified by harvesting 3.7 to 250 m² by plot combine and is expressed here on a dry weight basis.

Volumetric soil water content data were obtained from individual soil samplings or continuous readings from soil sensors depending on the location, and used to initialize model simulations and calibrate soil parameters to match observed soil moisture. The soil at the IOWA site has a shallow water table that fluctuates from 0.3 to 2.0 m below the soil surface, which can bring as much as 100–250 mm of water into the soil profile per year (Ebrahimi-Mollabashi et al., 2019; Ordóñez et al., 2018). This particular field had no subsurface tile drainage. The soils located in the other sites included in our study do not commonly have a shallow water table. Additional details about experimental sites, information related to management, soil, and weather data are provided in Kothari et al. (2022).

2.2. Participating soybean models

We used ten process-based eco-physiological crop models that have the capability to simulate soybean growth and development (Table 1). The models selected for this study were the result of the expressed interest of each modeling team along with their expertise with a particular model framework. The models differed in their prior experience simulating soybean growth before this study (**Supplementary Fig. S1**). The DSSAT-CROPGRO-Soybean model (Boote et al., 1998, Jones et al., 2003; Hoogenboom et al., 2019) originated in the 1980s and has been evaluated against over 50 sets of time-series of soybean growth and yield data. In addition to the classic DSSAT-CROPGRO model (hereafter DSSAT), a new energy balance version (DSSAT-CROPGRO-EBL, hereafter DSSAT-EBL) has been developed for the CROPGRO-Soybean model (Cuadra et al., 2020). In both versions, CROPGRO-Soybean is the basic module for the simulation of crop growth and development. However, the DSSAT-EBL model simulates an hourly energy balance and

stomatal conductance, and controls transpiration and photosynthesis response to leaf-to-air vapor pressure difference. In addition, the DSSAT-EBL model simulates canopy temperature (Cuadra et al., 2020), which drives soybean plant processes such as leaf photosynthesis and rate of development. The soybean model within APSIM (Robertson and Carberry, 1998) is later in its origin and has been widely evaluated against Midwestern USA soybean data. The SSM model was designed to represent soybean plant-soil processes in a simplified manner (Sinclair, 1986) and was previously evaluated for simulation of plant C and N dynamics for a limited number of sites in the USA. Recently, the SSM model has been used to simulate soybean in Iran (Nehbandani et al., 2021; Nehbandani et al., 2020). The STICS model was developed around 1996 (Brisson et al., 2003), but the STICS soybean model has limited prior evaluation, mainly with data from France. The LINTUL (Wolf, 2012) and MONICA (Nendel et al., 2011) models were not designed specifically for soybean, but to represent generic crop growth processes. In recent years, the MONICA model has been evaluated with soybean experimental data from Brazil that included in-season and end-of-season data (Battisti et al., 2017a; Battisti et al., 2017b; Battisti et al., 2018) and in Europe with end-of-season data (Nendel et al., 2022). The DNDC model was initially developed to simulate N₂O emissions from agricultural soils (Li et al., 1992), with plant growth processes integrated later into the model (Zhang and Niu, 2016). The plant growth routines in DNDC are focused mainly on simulating yield, biomass and crop N uptake rather than accurately modelling plant fractions (leaf, stem, and pod weight) or LAI. As a result, the DNDC-vCAN model used in this study is still relatively inexperienced for simulating soybean in-season growth. The AQUACROP (Steduto et al., 2009) and SWB (Annandale et al., 2000) models were designed to mimic generic crop growth, with a main emphasis on the simulation of crop evapotranspiration for irrigation scheduling.

Models varied in the level of complexity with respect to simulating the soybean developmental stages (Table 1, **Fig. 1**), types of vegetative and reproductive tissues and approaches for reproductive partitioning (Table 2), and in environmental factors influencing leaf development, growth, and reproductive partitioning (**Supplementary Table 1**). A detailed description of modeling approaches for development, photosynthesis, leaf area development, and vegetative and reproductive tissues is provided in **Supplementary Information SI 1**. In brief, a major difference across models was in their complexity to simulate reproductive development, and the number of developmental stages they simulated (**Fig. 1**, Table 1). Timing of beginning seed and physiological maturity were simulated by all models, while for some models (e.g., DSSAT, APSIM, MONICA) the start of anthesis, and the onset of pod growth were also simulated. Model structure varied among the ten models. For instance, LINTUL and MONICA originated as cereal models where grain growth commences at anthesis. As a result of differences in simulated reproductive stages, the models also varied in the types of reproductive tissues they simulated (Table 1). Most models simulate reproductive growth as a single component (seeds), or as a single component with a seed-to-shell ratio as in the MONICA model. Only two models (APSIM, DSSAT) simulate growth of pod shells (Table 1). Models differed in the level of parameterization required to define cultivar-specific coefficients for different genotypes (vs. species-defined), which varied from as low as seven (for AQUACROP) to as high as 34 (LINTUL) coefficients depending on the model (Table 1). Another major difference across models was in their approach for partitioning to reproductive tissues and yield determination (Table 2). Reproductive partitioning in most models is based on an age-driven harvest index (HI) approach, with the exception of the DSSAT-CROPGRO models, where yield is simulated by adding individual pod and seed cohorts over time, so reproductive partitioning is an indirect result of simulation of addition and individual growth of reproductive

sinks. The number of seeds is only simulated in the STICS and DSSAT-CROPGRO models and seed number set is based on assimilate-carrying supply relatively early in the seed-growth phase.

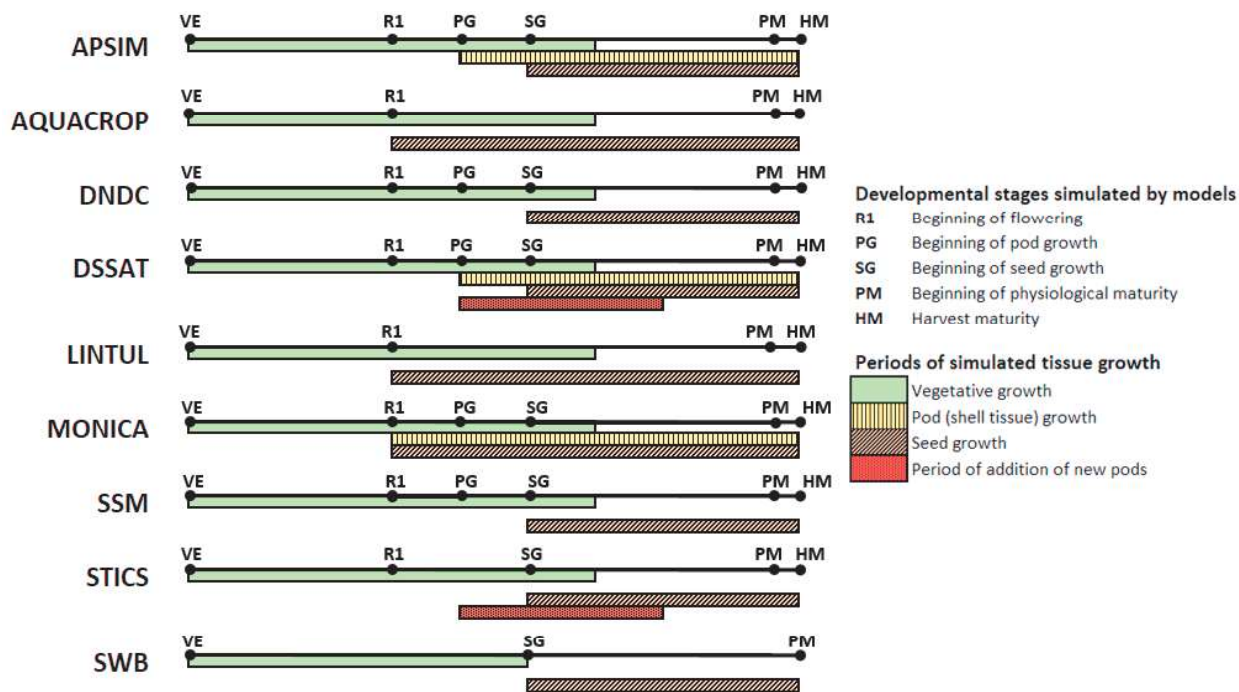


Figure 1. Schematic of soybean reproductive stages and periods of different reproductive tissue growth simulated by crop models participating in this study.

Table 1. Description of soybean crop models used in this study in terms of parameters, reproductive growth stages, and factors affecting rate of development.

Crop Model	# cultivar parameters	Factors driving rate of development towards reproductive stages§					Approach to account for T x P on rate of development†
		Beginning flower	Beginning pod	Beginning seed	Physiological maturity	Harvest maturity	
APSIM	10	T/P/W	T/P/W	T/P/W	T/P/W	Fixed	T
AQUA	7	n/s	n/s	T/W	T/W	Fixed	n/s
DNDC	14	T	T	T	T	Fixed	n/s
DSSAT	18	T/P/W	T/P/W	T/P/W	T/P/W	12 d	M
DSSAT-EBL	18	T/P/W	T/P/W	T/P/W	T/P/W	12 d	M
LINTUL	34	T/P/W	n/s	n/s	T/P	Fixed	M
MONICA	32	T/P/W	T/P/W	T/P/W	T/P/W	T/P/W	M
SSM	12	T/P/W	T/P/W	T/P/W	T/P/W	T/P/W	M
STICS	14	T/P/W	n/s	T/P/W	T/W	T/W	M
SWB	27	n/s	n/s	T	T	n/s	n/s

§ Factor affecting: **T** = Temperature; **P** = Photoperiod; **W** = Water stress; **Fixed** = Fixed number of days;
† Approaches to compute rate of development: **T**: Thermal time target method, **M**: Multiplicative method;
n/s = Developmental stage or process not simulated

Table 2. Description of approaches used by crop models for simulation of partitioning to reproductive growth, for yield component determination, and for leaf area development.

Crop Model	Method for partitioning to reproductive\$	Types of simulated plant tissues¥	Simulation of seed number¶	Factors that drive yield components‡		Method for leaf area development▲	Environmental factors affecting leaf area §		
				Pod addition	Seed growth rate		Duration of new leaf addition	Expansion rate	Senescence
APSIM	S/HI	R/S/L/P/Sd	n/s	n/s	Pn/N/W/T/P	T	T/P/W	T/W	T/W
AQUACROP	HI	Sd	n/s	n/s	n/s	n/s	n/s	T/W	n/s
DNDC	HI	R/S/L/Sd	n/s	n/s	Pn/N/W/T/P	C	n/s	T/W/N	T
DSSAT	S/C	R/S/L/N/P/Sd	S/C	Pn/N/W/T/P	Pn/N/W/T/P	C/T	T/P/W	T/W/Pn	T/W/N/L
DSSAT-EBL	S/C	R/S/L/N/P/Sd	S/C	Pn/N/W/T/P	Pn/N/W/T/P	C/T	T/P/W	T/W/Pn	T/W/N/L
LINTUL	Prt	R/S/L/Sd	n/s	n/s	n/s	C	n/s	T/P/W	T/P
MONICA	Prt	R/S/L/P/Sd	n/s	n/s	n/s	C	T/P/W/N	n/s	T/P/W/N
SSM	HI	S/L/Sd	n/s	n/s	Pn/N/W/T	T	T/P/W	T/W/N	T/W/N
STICS	HI/Prt	R/S/L/Sd	C	n/s	Pn/N/W/T	T	T/W/N	T/W/N	T/W/N
SWB	Prt	R/S/L/Sd	n/s	n/s	Pn/T/W/N	C/T	No	T/W/N	T/W

\$ **HI** = age-driven harvest index; **S** = Sink growth driven; **Prt** = partitioning during reproductive stages
¥ **R** = Roots; **S** = Stem; **L** = Leaf; **N** = Nodules; **P** = Pod; **Sd** = Seed
¶ **S** = New sinks (pods and seeds) cohorts added over a window of time; **C**: seed number computed based on assimilate supply and minimum requirement per seed over a window of time; **n/s**: model does not simulate growth of individual pod cohorts or seed number.
‡ **Pn** = Photosynthesis or growth; **N** = Nitrogen stress; **W** = Water stress; **T** = Temperature effect; **P** = Photoperiod effect; **n/s** = Yield component not simulated by model
▲ **C** = Carbon-based method (leaf area is a function of the assimilate available for leaf growth and is calculated as the daily leaf dry matter growth times the specific leaf area); **T** = Temperature-based method (leaf area is not restricted by assimilate availability but is dependent on the temperature); **C/T** = Hybrid method; **n/s** = not simulated.
§ **T** = Temperature; **P** = Photoperiod; **W** = Water stress; **N** = Nitrogen stress; **L** = PAR intensity; **Pn** = Photosynthesis or growth; **n/s** = Process not simulated.

2.3. Model Calibration

For model parameterization, the modelers were given inputs of daily weather data, crop management, soybean cultivar and maturity group, measured soil properties, and initial soil moisture and inorganic N at planting. The field experimental data were prepared by a central team of researchers who coordinated the processing and analysis of model inputs and outputs. Thus, modelers did not have prior access to the data, until they were provided with measured data for calibration. This provided the opportunity to evaluate models under true “blind” conditions, a principle designed by AgMIP protocols. One exception to this was the IOWA dataset for the APSIM model, since that modeling team provided the experimental data from this location. For model calibration, the measured data were provided to modelers in two steps: 1) *Blind* calibration with phenology data only, and 2) *Full* calibration with detailed in-season crop growth and final yield data, as well as soil moisture. The step-wise calibration methodology is consistent with the prior multi-model intercomparison studies for other crops (Asseng et al., 2015; Bassu et al., 2014).

During *Blind* calibration, modelers optimized phenology coefficients (per cultivar) to best simulate measured dates of developmental stages across all years and treatment levels within a location. In the protocols, modelers were asked to prioritize the accuracy for the simulation of date of R1 and R7 stages, given that reproductive developmental stages of R3 and R5 defined by Fehr and Caviness (1977) refer to the four uppermost nodes in the main stem and may not be equivalent to the timing of start of reproductive partitioning to pods or seeds, respectively. For non-phenology related coefficients, crop modelers accepted and used common or previously calibrated cultivar coefficients for soybean in their model (sometimes referred as “default” coefficients) to the best of their knowledge during *Blind* calibration.

During the *Full* calibration, the calibration protocols recommended the following step-wise approach. First, calibration of soil parameters to match the measured soil moisture data. Second, calibration of cultivar coefficients associated with leaf area development, and crop growth and vegetative partitioning to minimize error in simulation of in-season crop aboveground biomass, LAI, leaf and stem weight. Third, cultivar coefficients associated with partitioning to reproductive organs to minimize error in simulation of in-season total pod weight, final grain yield, and HI. Finally, calibration of cultivar coefficients associated with simulation of unit seed weight and seed yield where applicable, depending on the model. Modelers were recommended to repeat this step-wise approach iteratively if needed. Crop modelers were allowed to modify previously defined phenology-related crop coefficients to better match the measured onset of pod and seed mass growth if needed during *Full* calibration. All models were calibrated manually to optimize simulations within a location, which involved adjusting cultivar and species coefficients depending on the model to match sequential crop growth observations. The complete list of crop coefficients modified during calibration and their parameterization is provided in **Supplementary Table 2**.

2.4. Analysis of in-season growth and development data

For the analysis of data related to timing of developmental stages, we found that timing of reproductive developmental stages simulated by models and the onset of simulated start of pod and seed growth differed depending on the model (see schematic **Fig. 1**). Thus, in addition to timing of developmental stages reported by models, simulated from different photothermal time approaches (Table 1, *Photothermal*), the beginning of pod and seed growth were calculated based on the first day when accumulation of pod and seed growth occurred in simulated daily growth outputs (*Growth*). Using developmental stages based on growth time-series outputs rather than the

phenology dates directly generated by the models, provided a better interpretation for the analysis of in-season reproductive growth in our study.

To evaluate and summarize model performance for simulation of reproductive in season-growth, we quantified pod growth rate ($\text{kg ha}^{-1} \text{d}^{-1}$) and the rate of increase in seed HI ($\text{g g}^{-1} \text{d}^{-1}$) (Van Roekel et al., 2015) from both simulated and measured in season growth data. The pod growth rate was computed by fitting a linear model to the pod weight time-series data between >10% and <90% of the maximum pod weight. We also computed the duration of pod growth (d), seed growth (d), the days from flowering to beginning of pod growth (d), days from flowering to the beginning of seed growth (d), and the effective seed filling period (d). For the DSSAT, DSSAT-EBL, and STICS models, the only models in the study simulating seed number, we also computed the simulated individual seed growth rate (ISGR, $\text{mg seed}^{-1} \text{d}^{-1}$) by dividing simulated total seed growth rate by seed number at maturity. For measured data, given that only in-season pod weight was available, we computed the rate of increase in pod HI for measured data (instead of the rate of increase in seed HI), which we referred to as the rate of increase in reproductive partitioning. The complete list of variables used for model evaluation and details on how these physiological indices were computed from direct model outputs and measured data is described in Supplementary materials (**Supplementary Table 3, SI 3**).

2.5. Model evaluation

We evaluated model uncertainty in simulation of developmental stages, in-season growth, and physiological growth indices with different statistical performance indicators computed by model and site (eqns. 1–4). We also employed an error decomposition analysis to understand sources of model error (eqns. 5-8) (Gauch et al., 2003).

$$RMSE_{m,s} = \sqrt{\frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i} - Y_{s,i})^2}{N_s}} \quad \dots (1)$$

$$nRMSE_{m,s} = \sqrt{\frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i} - Y_{s,i})^2}{N_s}} \times \frac{100}{\bar{Y}_s} \quad \dots (2)$$

$$d_{m,s} = 1 - \frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i} - Y_{s,i})^2}{\sum_{i=1}^{N_s} (|\hat{Y}_{m,s,i} - \bar{Y}_s| + |Y_{s,i} - \bar{Y}_s|)^2} \quad \dots (3)$$

$$Bias_{m,s} = \frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i})}{N_s} - \frac{\sum_{i=1}^{N_s} (Y_{s,i})}{N_s} \quad \dots (4)$$

$$MSE_{m,s} = \frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i} - Y_{s,i})^2}{N_s} \quad \dots (5)$$

$$SB_{m,s} = \left(\frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i})}{N_s} - \frac{\sum_{i=1}^{N_s} (Y_{s,i})}{N_s} \right)^2 \quad \dots (6)$$

$$LC_{m,s} = (1 - r^2) \frac{\sum_{i=1}^{N_s} (Y_{s,i} - \bar{Y})^2}{N_s} \quad \dots (7)$$

$$NU_{m,s} = (1 - b)^2 \frac{\sum_{i=1}^{N_s} (\hat{Y}_{m,s,i} - \bar{\hat{Y}}_{m,s})^2}{N_s} \quad \dots (8)$$

where $RMSE_{m,s}$, $nRMSE_{m,s}$, $d_{m,s}$, and $MSE_{m,s}$ are the root mean square error, normalized RMSE, index of agreement (d), and mean square error (MSE) for the model m at site s . The MSE is further divided into standard bias (SB), lack of correlation (LC), and non-unity of slope (NU). The $Y_{s,i}$ is the i^{th} measured value at site s , and $\hat{Y}_{m,s,i}$ is the corresponding simulated value by model m ; $\bar{Y}_s$ is the average of measured values at site s ; and N_s is the number of measured values at site s . The

$\overline{\hat{Y}_m}$ is the average of simulated values by model m . The d ranges between 0 to 1 with 1 representing a perfect agreement between the two series.

3. RESULTS

3.1. Soybean developmental stages

Measured phenology data were provided to modelers during the *Blind* calibration for calibrating phenology coefficients. However, in some models there was a need to re-adjust phenology coefficients during *Full* calibration (Table 3) to improve the simulation of in-season pod growth and final yield. As a result, the RMSE and bias for simulation of phenology worsened in some models after *Full* calibration (Table 3), that prioritized improving simulation of in season growth and yield. We compared the accuracy in simulating phenology between the two calibration steps. Most models had minor changes in simulation of phenological stages from *Blind* to *Full* calibration (< 3 days difference in the nRMSE for simulation of reproductive stages between *Blind* and *Full* calibration, Table 3). Of interest, we found that the timing of beginning pod and seed growth reported by models (*Photothermal*) differed some from the timing for these stages computed based on the first day when simulated pod and seed growth occurred (*Growth*, Table 3).

We compared models for the bias in simulating phenological stages. A group of models (AQUACROP, LINTUL, and MONICA) had a higher bias compared to other models in simulating onset of pod and seed growth (*Growth*, Table 3). These three models simulated the start of seed growth more than 30 days earlier on average compared to the observed (*Growth*, Table 3), and this is because their model structure is designed to start allocation of assimilates to seed growth at the time of flowering. This led to an artificially long duration of the time from beginning seed growth to physiological maturity for these models that was inconsistent with the observed data (**Fig. 2**).

Of interest, the MONICA model showed adequate simulations for the Beginning Pod and/or Seed stages based on the *Photothermal* approach (Table 3). These results indicate that the MONICA model would benefit from delaying the start of pod and seed growth to the simulated time of Beginning Pod (*Photothermal*) while maintaining its current model structure for reproductive partitioning.

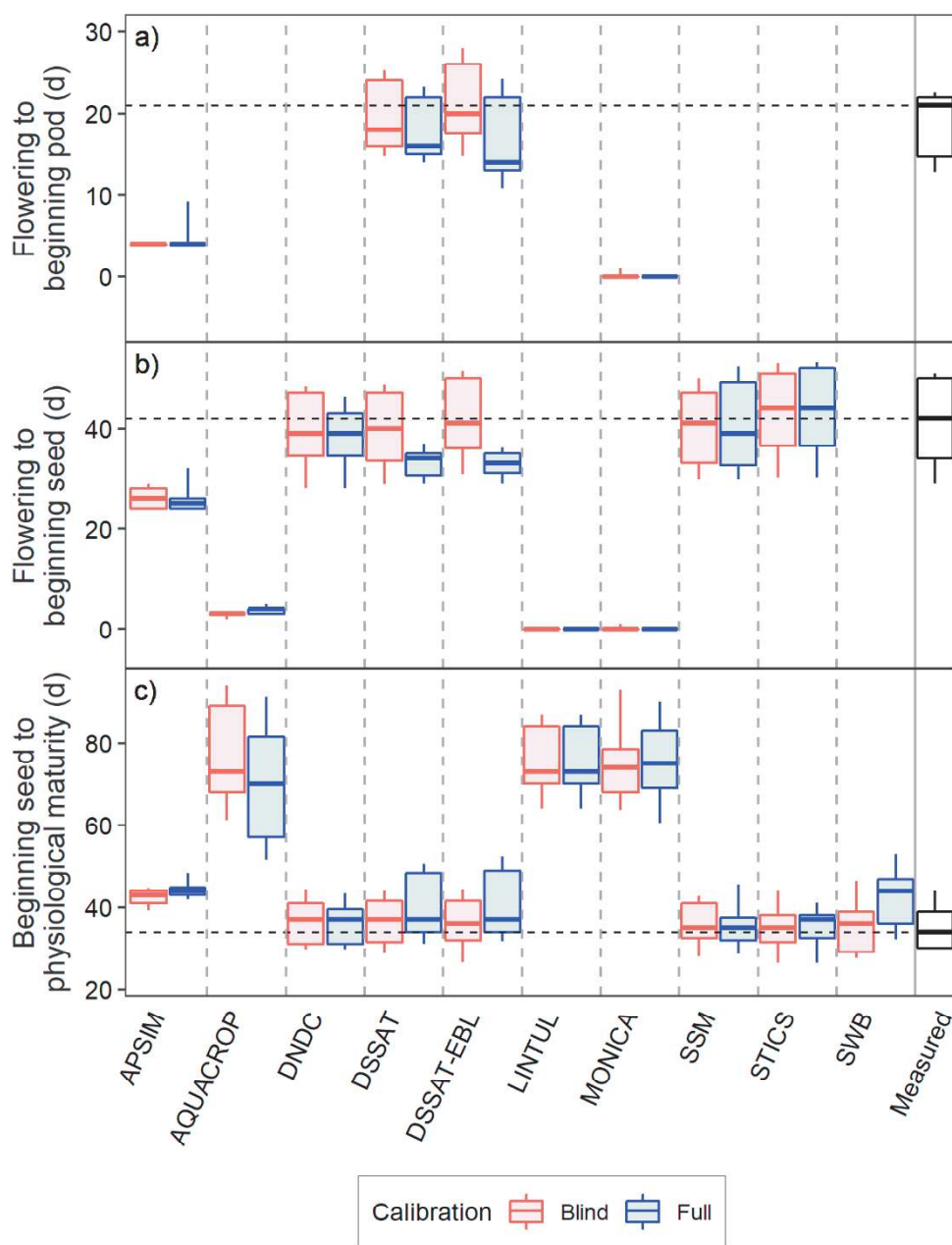


Figure 2. Measured and simulated duration from flowering to beginning pod (a), from flowering to beginning seed (b), and from beginning seed to physiological maturity (c), by model and calibration phase. Boxplots show variability across 19 treatments and all locations. The whiskers show 10th and 90th percentiles, while the bounds are 25th and 75th percentiles, with median shown as the line within the box. For the measured data, durations were calculated using measured stages based on Fehr and Caviness (1977). For the simulated data, durations were calculated using the first day when accumulation of pod and seed growth occurred from simulated daily growth outputs.

Two models simulated pod growth and the lag between flowering, start of pod growth, and the start of seed growth (APSIM and DSSAT models, see **Fig. 1**). For DSSAT and DSSAT-EBL, the bias in beginning seed increased from *Blind* to *Full* calibration (Table 3) to improve calibration to pod growth. Overall, the timing and duration of reproductive growth phases in the DSSAT and DSSAT-EBL models showed good agreement with observed data (**Fig. 2**). In the APSIM model, the model performance for most crop growth stages improved after *Full* calibration (Table 3), but there was still a consistent bias in the simulated timing and duration of reproductive growth compared to the measured data (**Fig. 2**). Another group of models simulated a lag phase in the start of seed growth after flowering, but without simulation of pod tissue (DNDC, SSM, and STICS), and these models showed a very consistent simulation of phenology from *Blind* to *Full* calibration (**Fig. 2**). In addition, two of these models (DNDC, SSM) also simulated the beginning pod developmental stage (*Photothermal*, Table 3). The SWB model did not simulate beginning of flowering, but only the start of seed growth (**Fig. 1**). In SWB model, the bias in simulating physiological maturity increased between the calibration steps (from 0.4 to 5.9 days).

Table 3. Model performance for simulating dates of reproductive phenological stages, expressed as root mean squared error (RMSE) and bias (Measured – Simulated). Measured phenological stages were based on Fehr and Caviness (1977). Simulated time of Beginning Pod and Seed was based on stages reported by models, derived from different photothermal time approaches (*Photothermal*), and based on the first day when accumulation of pod and seed growth occurred in simulated daily growth outputs (*Growth*).

	Crop Model	Anthesis		Beginning Pod <i>Photothermal (Growth)</i>		Beginning Seed <i>Photothermal (Growth)</i>		Physiological maturity	
		Blind	Full	Blind	Full	Blind	Full	Blind	Full
RMSE (days)	APSIM	3.5	4.3	(13.9)	(7.9)	11.5 (11.4)	4.8 (5.1)	6.0	6.9
	AQUACROP	2.6	7.1			(32.4)	(32.4)	9.0	14.7
	DNDC	1.4	1.4	2.8	2.8	(4.2)	(4.7)	5.7	5.5
	DSSAT	2.3	2.5	2.0 (2.2)	4.1 (3.4)	3.3 (3.0)	7.5 (5.9)	3.0	3.3
	DSSAT-EBL	3.4	4.4	3.5 (3.7)	3.6 (3.9)	3.1 (3.7)	3.4 (6.4)	4.3	4.4
	LINTUL	1.9	1.9			(37.8)	(37.8)	7.3	7.3
	MONICA	1.8	1.8	4.4 (17.1)	3.8 (17.2)	11.9 (36.6)	8.8 (36.7)	4.5	4.3
	SSM	2.0	2.0	2.3	2.3	3.2 (3.2)	3.6 (3.6)	3.9	4.8
	STICS	2.6	2.4			3.2 (5.2)	3.2 (5.4)	5.3	4.7
	SWB					4.9 (4.9)	4.9 (4.8)	5.5	11.7
	Average RMSE	2.4	3.0	3.0 (9.2)	3.3 (8.1)	5.9 (14.2)	5.2 (14.3)	5.5	6.8
Bias (days)	APSIM	-1.5	1.3	(-13.6)	(-6.6)	-11.1 (-11.0)	-3.9 (-3.7)	-0.8	3.9
	AQUACROP	1.9	1.5			(-32.1)	(-31.7)	5.6	-2.7
	DNDC	0.1	0.1	-0.1	-0.1	(-0.3)	(-0.9)	0.1	-0.8
	DSSAT	-0.4	0.0	-0.7 (0.7)	-2.2 (-0.2)	-1.6 (0.0)	-6.6 (-4.3)	0.3	-0.9
	DSSAT-EBL	-1.7	3.0	0.0 (1.1)	-0.1 (1.0)	-0.6 (0.6)	-0.9 (-1.6)	2.0	2.2
	LINTUL	-0.8	-0.8			(-37.6)	(-37.6)	-0.3	-0.3
	MONICA	0.2	0.2	-3.4 (-16.8)	-1.7 (-16.9)	-11.0 (-36.3)	-6.3 (-36.4)	-0.1	0.1
	SSM	0.3	0.7	1.2	1.3	1.1 (1.1)	1.5 (1.5)	1.1	1.3
	STICS	-0.3	0.0			-0.3 (2.9)	-0.1 (3.2)	1.2	1.7
	SWB					0.3 (0.3)	1.1 (1.0)	0.4	5.9
	Average bias	-0.2	0.6	-0.6 (-7.2)	-0.5 (-5.7)	-3.3 (-11.2)	-2.2 (-11.0)	0.9	1.0

3.2. Model evaluation for simulation of leaf area development and crop growth

An example of time-series data for simulated and measured aboveground biomass, leaf, stem, pod, and seed weights for two of the nineteen treatments in this study is shown in **Fig. 3**. All ten models simulated biomass, nine models simulated LAI, eight models simulated leaf and stem weights, and only four models simulated pod weight (**Fig. 3**). The limited availability of in-season measured seed weight data for soybean is noteworthy, given that separating seeds from developing pods is a very time-consuming process. Thus, modelers relied on measured in season pod weight data and final seed yield for calibration of reproductive growth.

3.2.1. Vegetative growth and partitioning

After *Blind* calibration, the nRMSE for simulation of in-season aboveground biomass ranged from 22% to 109% depending on the model, with the nRMSE below 30% for a group of models (APSIM, DSSAT, DSSAT-EBL, LINTUL, and SSM) (**Fig. 4a**). After *Full* calibration, the nRMSE for simulation of in-season biomass ranged from 16% to 48% and six models simulated aboveground biomass with nRMSE < 30% (**Fig. 4a**). Overall, the nRMSE for simulation of in-season leaf and stem weight in the *Blind* calibration was higher (67% and 64% averaged across models, respectively), compared to the nRMSE for simulation of aboveground biomass (45%).

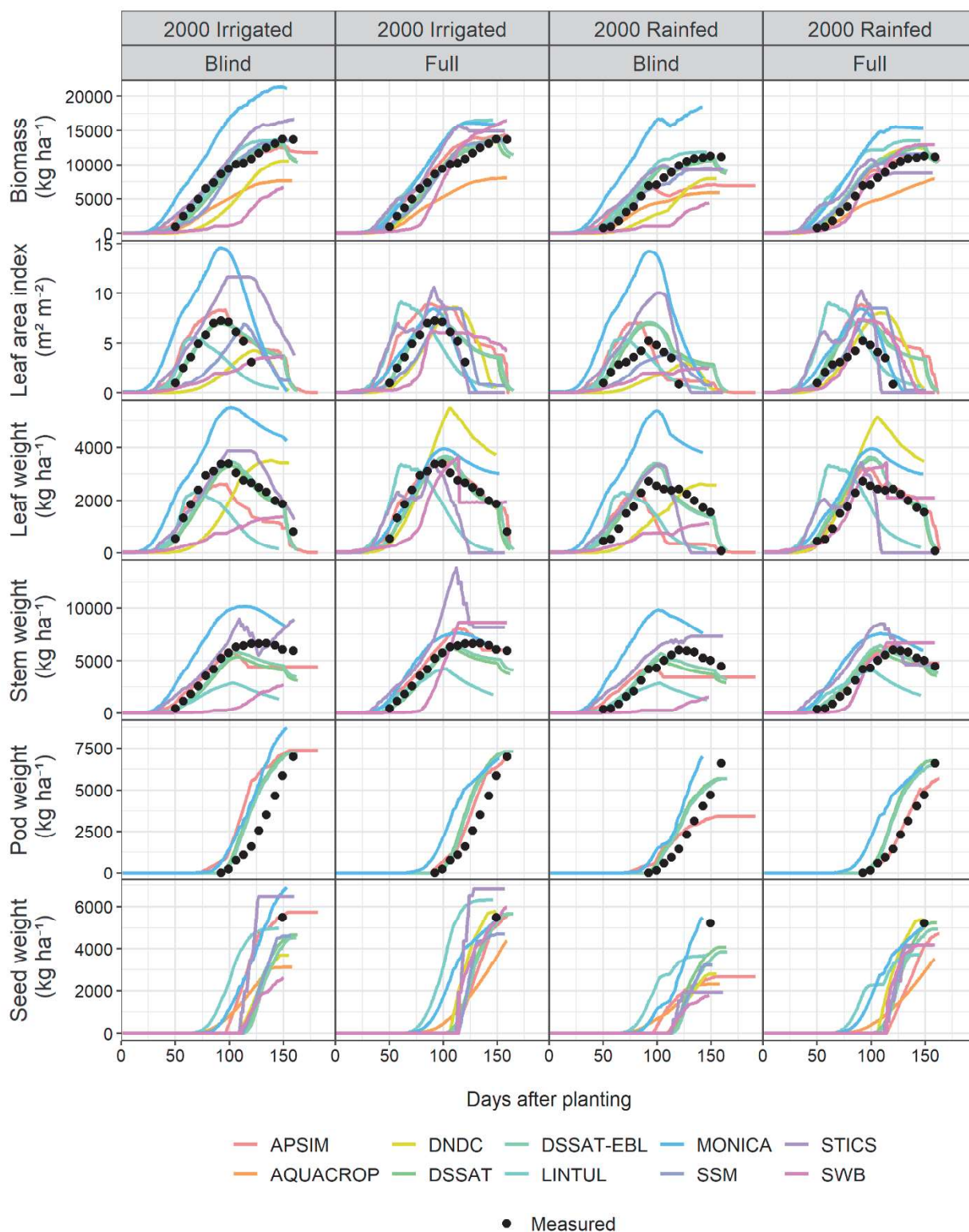


Figure 3. Measured (symbols) and simulated (lines) biomass, leaf area index, leaf, stem, pod, and seed weight for 10 crop models, after *Blind* and after *Full* calibration. Data from irrigated and rainfed treatments in 2000 at the ARGN site. All models simulated biomass and seed weight,

22 nine models simulated LAI, eight models simulated stem and leaf weights, while only four
 23 models simulated pod weight.

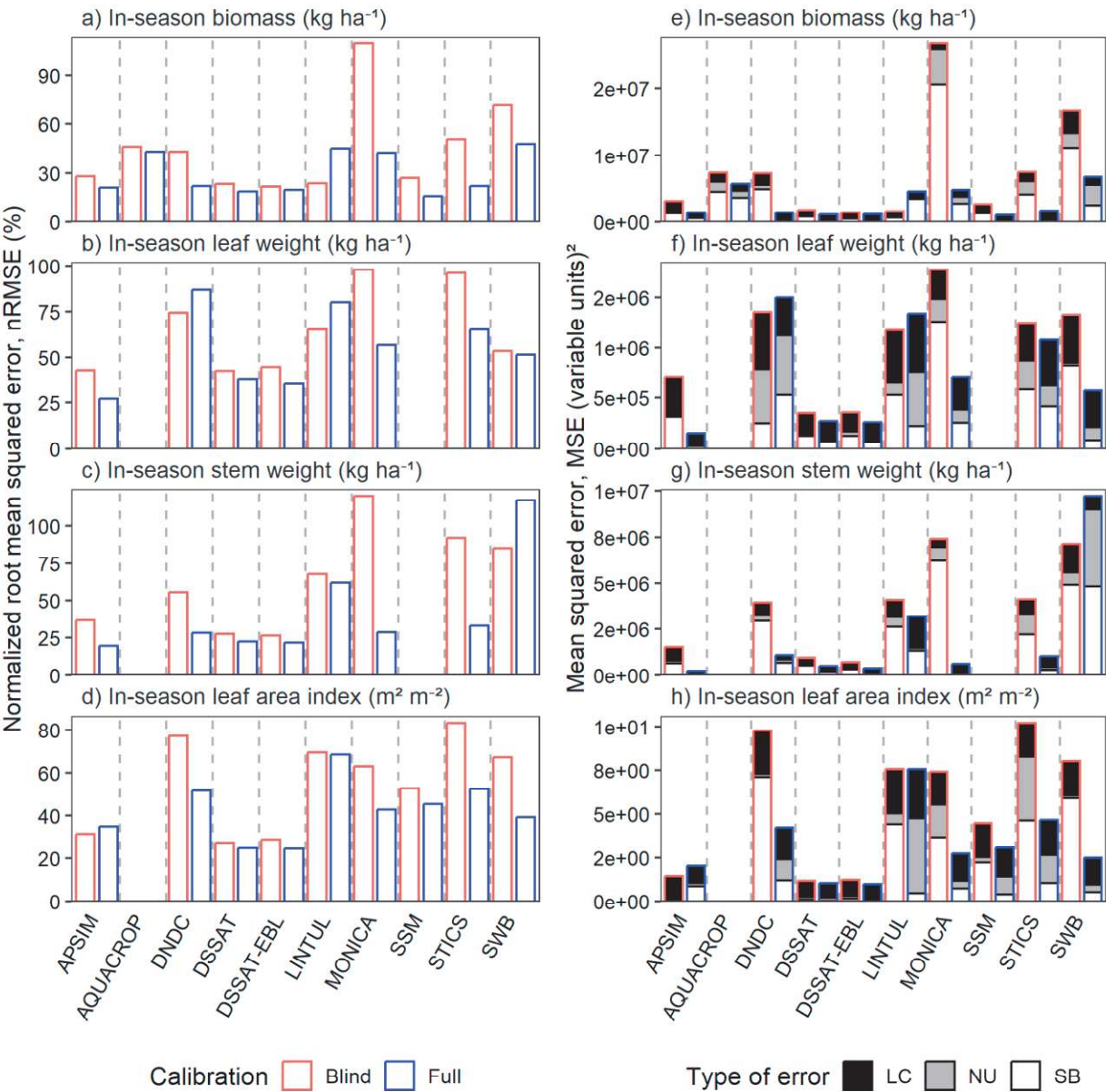


Figure 4. Normalized root mean square error (nRMSE) and mean square error (MSE) for the simulation of in-season biomass (a and d), leaf weight (b and e), and stem weight (c and f) by model after *Blind* and *Full* calibration. Statistics calculated across all treatments and locations. The MSE is decomposed into lack of correlation (LC), non-unity of slope (NU), and standard bias (SB).

The MSE decomposition for simulation of in-season total biomass and vegetative growth showed that model performance improved after *Full* calibration mainly due to a reduction in standard bias (SB), compared to *Blind* calibration (**Fig. 4d, e, and f**). In the *Blind* calibration, SB was the largest contributor to the total error, accounting for 53%, 40%, and 62% of the MSE for simulation of in-season biomass, leaf weight, and stem weight, respectively (**Fig. 4d, e, and f**). After *Full* calibration, lack of correlation (LC) was the largest contributor to MSE, accounting for 50%, 55%, and 49% of the total MSE for simulation of in-season biomass, leaf weight, and stem weight, respectively. While calibration was effective in reducing SB, substantial error due to LC and NU remained for simulation of in-season leaf weight (**Fig. 4e**). The larger source of error due to LC (in better performing models) and due to NU and LC (in models with poorer leaf growth performance) indicates that improvement is needed in simulating leaf growth dynamics.

We explored how models differ in partitioning biomass into leaves by plotting the simulated and measured leaf to vegetative weight (leaf plus stem) ratio over time (**Fig. 5**). The observed data showed a linear decrease in the leaf weight ratio over time, whereas the simulations show a different shape in several models. The leaf weight ratio was close to 0.68–0.72 at the earliest sampling in the measured data (15 to 25 days after planting) and declined linearly over time (**Fig. 5**). In two of the models (DNDC and STICS), the leaf weight ratio started at values of 0.5,

48 indicating equal partitioning between leaves and stems in these two models during early vegetative
49 growth. Towards the end of the season, near harvest maturity, the leaf weight ratio was reduced to
50 almost zero in most models including DSSAT, DSSAT-EBL, APSIM, LINTUL, and STICS,
51 which is expected and was also seen from the measured data. However, in some models the leaf
52 weight ratio at the time of harvest was still high after *Full* calibration, with values of 0.50 and 0.39
53 on average in the DNDC and MONICA models, respectively (**Fig. 5**). A noteworthy finding from
54 this analysis was that for most models, calibration had a minor effect on the shape of the leaf
55 weight ratio curves over time.

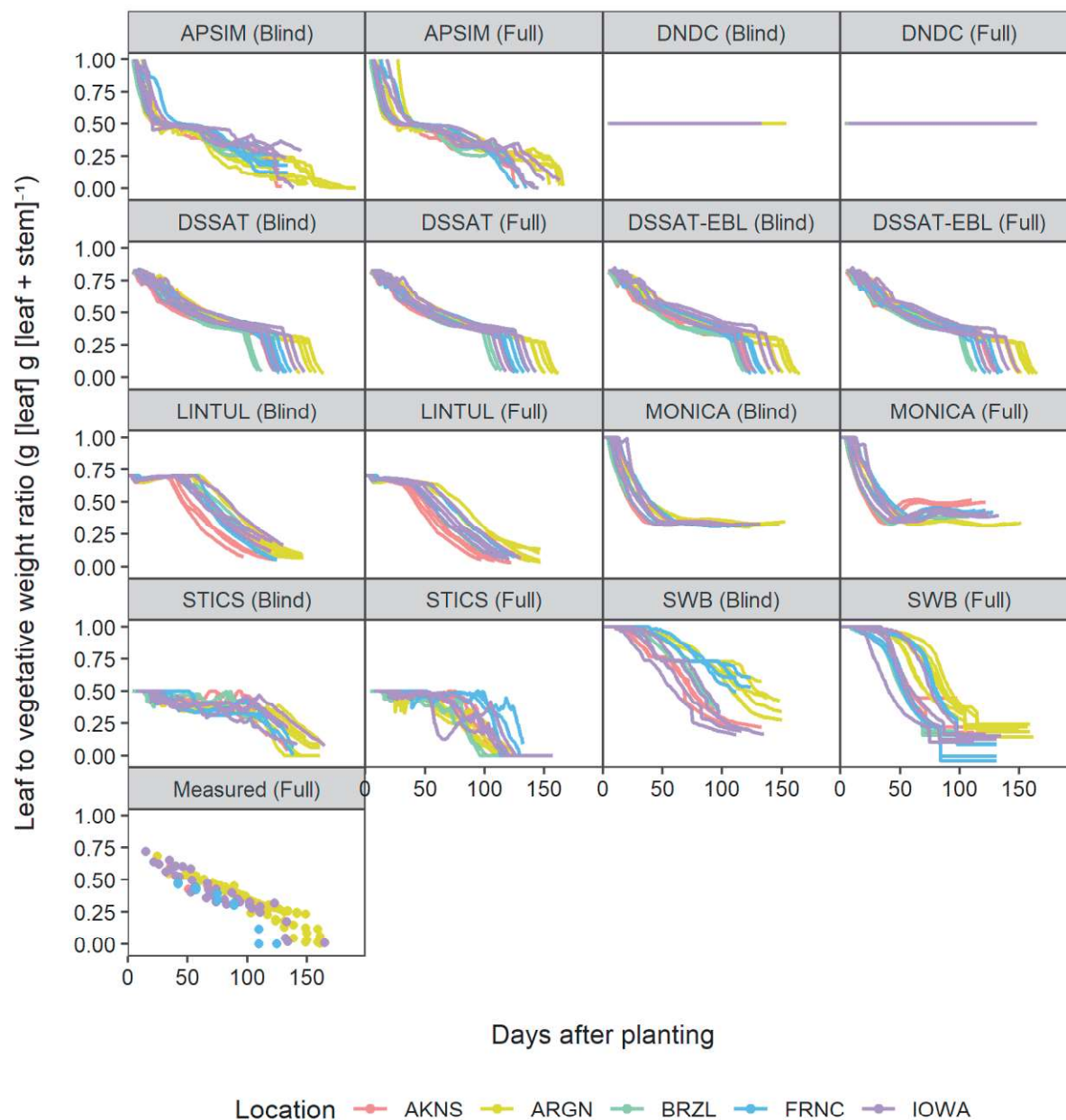


Figure 5. Leaf to vegetative weight (leaf + stem) ratio simulated by eight models and measured in field. Individual lines represent the nineteen treatments and the different colors distinguish the five locations.

In summary, calibration reduced the overall error for most models in the simulation of total aboveground biomass, stem, and leaf weight with a few exceptions (**Fig. 4**). However, calibration did not modify greatly the general shape of the simulated leaf to vegetative weight ratio over time

(Fig. 5), which is an indirect effect of mechanisms hard-coded in models. Across all models, APSIM, DSSAT, and DSSAT-EBL had the lowest nRMSE and MSE for simulation of in-season biomass, leaf and stem weight (Fig. 4), and also followed most closely the observed shape of the leaf to vegetative weight ratio over time (Fig. 5).

3.2.2. Leaf area development

When evaluating performance of individual models for the simulation of LAI, we found that DSSAT, DSSAT-EBL, and APSIM had the lowest nRMSE ($< 31\%$) in *Blind* calibration (Fig. 4d). After *Full* calibration, MONICA, SWB, and SSM performed relatively better for simulation of in-season LAI (nRMSE = 39–45%) than DNDC, STICS, and LINTUL (52–69%) (Fig. 4d). The analysis of the MSE decomposition indicated that models with the lowest MSE for the simulation of in-season LAI under *Blind* conditions (APSIM and DSSAT models) had negligible SB ($< 12\%$ of total MSE), and most of the error ($> 76\%$) came from LC (Fig. 4h). The remaining models had a larger contribution of MSE from SB (45–73%) for simulation of in-season LAI in the *Blind* condition. After *Full* calibration, models with high MSE for simulation of LAI reduced MSE associated to SB (DNDC, MONICA, STICS, SWB), but the error associated to LC remained similar, and the one associated to NU increased for some models (DNDC, SSM, SWB). The LINTUL model showed no improvement in the simulation of LAI after *Full* calibration, but a change in the source of error, with a decrease in SB, and an increase in NU compared to *Blind* conditions.

Overall, calibration reduced MSE for simulation of LAI for most models, mainly due to a reduction in bias. Error due to bias can be typically adjusted by changing cultivar parameters affecting dry matter accumulation, duration of leaf area development, and simulation of soil water balance and drought stress that might limit leaf area development. However, the lack of correlation

and non-unity of the slope for the simulation of LAI (**Fig. 4h**) suggest the absence of robust leaf senescence mechanisms in many of the models.

3.2.3. Reproductive growth and partitioning to seed

For a group of four models (APSIM, DSSAT, DSSAT-EBL, and STICS), reproductive growth started with growth of pods. For other models, it started with the growth of seeds. Thus, models that did not simulate pod growth could only rely on measured data for final seed yield and HI for model evaluation (**Fig. 6**). Final seed HI varied from 0.15 to 0.67 in the models and from 0.37 to 0.64 in the measured data, with a general trend of under-prediction by many models (AQUACROP, DSSAT, DSSAT-EBL, DNDC, LINTUL, MONICA, SSM, STICS, and SWB) under *Blind* condition (**Fig. 6c**). Likewise, most of the models under-predicted final seed yield under *Blind* condition. When comparing the range of seed yield for measured and simulated data under *Blind* and *Full* calibration, it is clear that differences among models and the deviation of models from measured values reduced substantially following *Full* calibration (**Fig. 6**).

When evaluating the performance of individual models, results showed that pod weight was simulated with the lowest nRMSE by the DSSAT and DSSAT-EBL (nRSME <31%), followed by the APSIM model (37%) under *Blind* conditions (**Fig. 7a**). After *Full* calibration, the nRMSE for simulation of pod weight decreased in APSIM (nRMSE<27%) due to a reduction in the nRMSE of pod growth rate (**Fig. 7b**). For DSSAT and DSSAT-EBL, the nRMSE for simulation of pod weight decreased during *Full* calibration (**Fig. 7a**), mostly due to advancing the onset of pod and seed growth (**Fig. 2a-b**), and less associated to a reduction in the nRMSE for the simulation of rate of pod growth, or rate of increase in pod harvest index (**Fig. 7b-c**). In the MONICA model, the simulation of in-season pod weight showed a high nRMSE under the *Blind* calibration, mainly due to an over-prediction of pod growth rate (**Fig. 7b**) and total aboveground biomass. After *Full*

110 calibration, the nRMSE for pod weight in the MONICA model was reduced but remained high
111 because pod and seed growth began immediately after anthesis due to the model structure for
112 initiating reproductive growth (**Fig. 2a-b**).

113

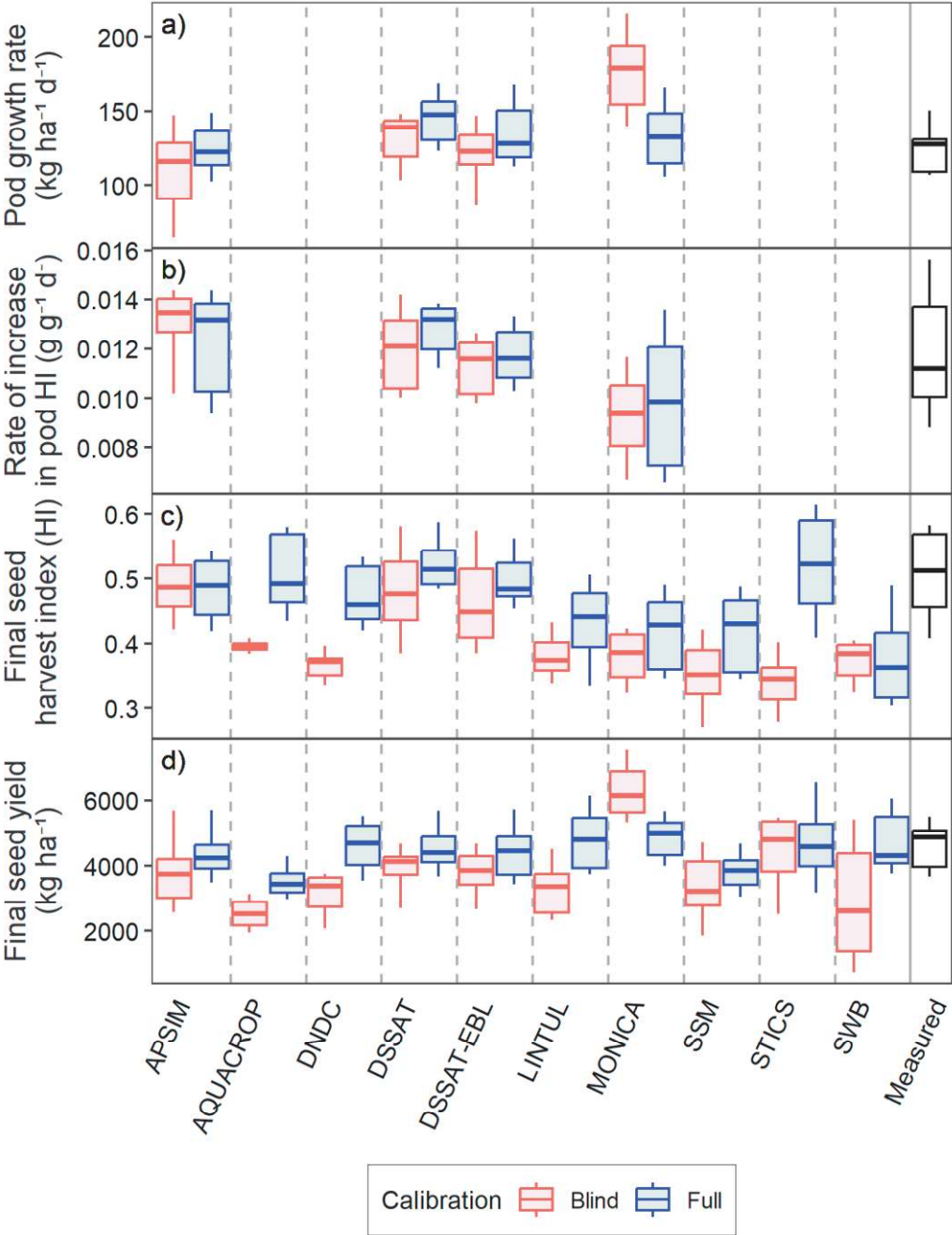


Figure 6. Measured and simulated rate of pod growth (a), rate of increase in pod harvest index (b), final seed harvest index (c), and final seed yield (d) by model and calibration. Boxplot show variability across 19 treatments and all locations. The whiskers show 10th and 90th percentiles, while the bounds are 25th and 75th percentiles, with median shown as the line within the box.

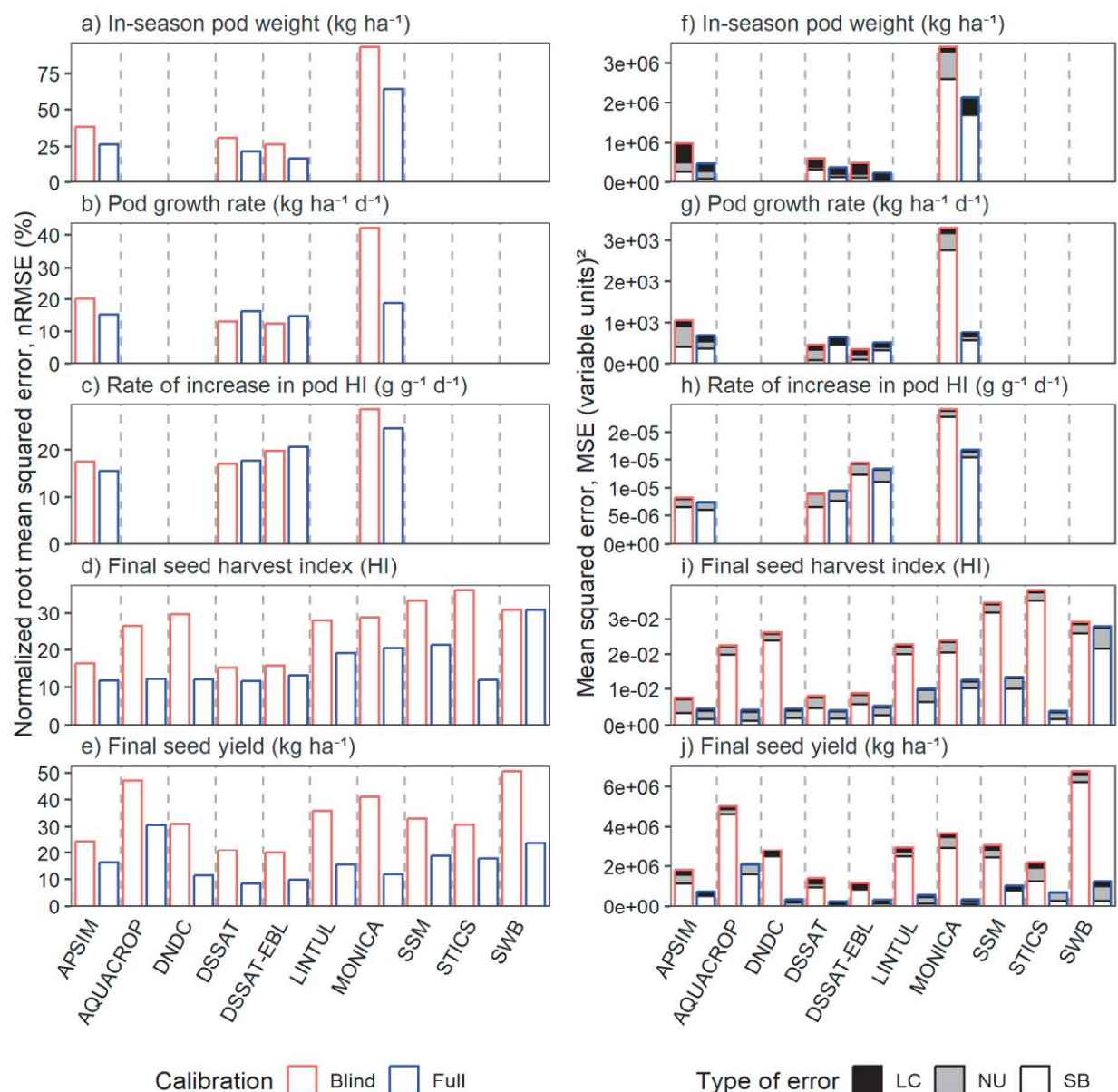


Figure 7. Normalized root mean square error (nRMSE) and mean square error (MSE) for the simulation of in-season pod weight (a and f), pod growth rate (b and g), rate of increase in pod harvest index (c and h), final seed harvest index (d and i), and final seed yield (e and j) by model after *Blind* and *Full* calibration. Statistics calculated across all treatments and locations. The MSE is further decomposed into lack of correlation (LC), non-unity of slope (NU), and standard bias (SB).

For the simulation of final seed yield, the nRMSE ranged from 20% (DSSAT-EBL) to 51% (SWB) after *Blind* calibration, and from 8% (DSSAT) to 31% (AQUACROP) after *Full* calibration (Fig. 7e). DSSAT and DSSAT-EBL followed by APSIM generally had the lowest nRMSE (<24%) in the *Blind* calibration for simulation of final seed yield and HI (Fig. 7d-e). Similar to the other variables analyzed, the analysis of the MSE decomposition revealed that high nRMSE for simulation of yield and HI was largely associated to a large bias (SB) (Fig. 7i-j).

We further analyzed the reproductive growth over time using the time-series of simulated seed HI (Fig. 8). We did not have observed seed growth data to compute seed HI, but the pod HI is shown in Fig. 8 as reference, with the premise that the rate of pod HI increase is likely similar in pattern to the rate of seed HI. The shapes of seed HI curves were different across models, and treatments (Fig. 8). Models differed in the onset of growth allocation to seed, and in turn, the rate of increase in HI also varied to match the final measured yield. Calibration also modified the shape of the simulated seed HI over time for some models, due to adjustments of parameters to improve simulation of the final seed yield without regards to the in-season seed growth dynamics. The range of the rate of increase in seed HI among models was 0.003–0.030 HI d⁻¹ after *Blind* and 0.002–0.064 HI d⁻¹ after *Full* calibration (Fig. 9a). In contrast, the rate of increase in pod HI from the measured data in our study ranged from 0.008 to 0.0192 pod HI d⁻¹ across treatments. Based on our literature review, the rate of increase in seed HI from other experimental sites ranged from 0.0100 to 0.0210 HI d⁻¹ (Fig. 9b). The rate of increase in seed HI was higher for the STICS model (median of 0.0329 HI d⁻¹) compared to the other models and the reported values in the literature (Fig. 9a-b). In contrast, the AQUACROP MONICA, and SWB models had relatively lower rates of increase in seed HI after Full calibration compared to values reported in the literature (Fig. 9a-b).

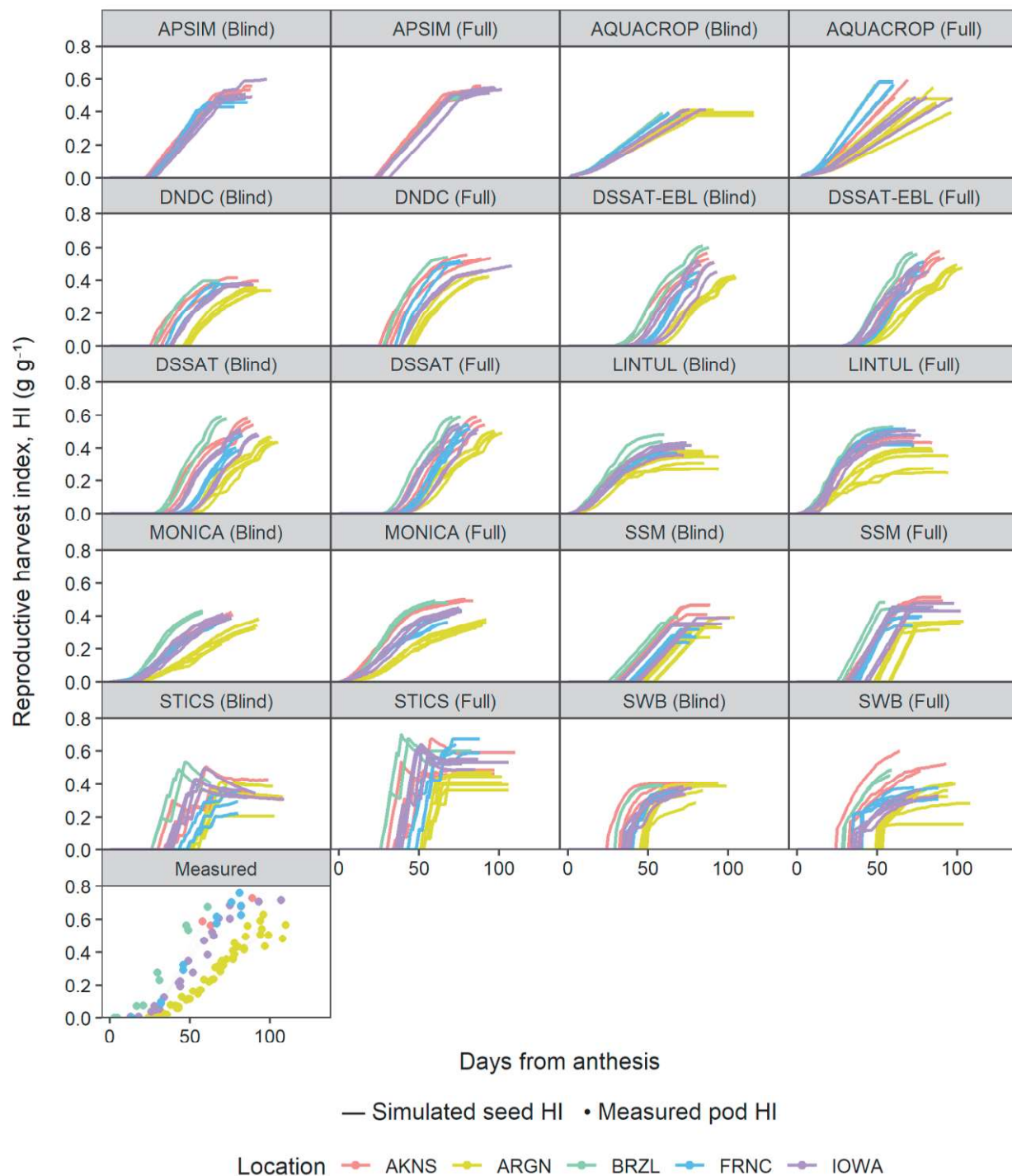


Figure 8. Seed harvest index over time simulated by models and pod harvest index over time measured in the field. Individual lines represent the nineteen treatments and the different colors distinguish the five locations.

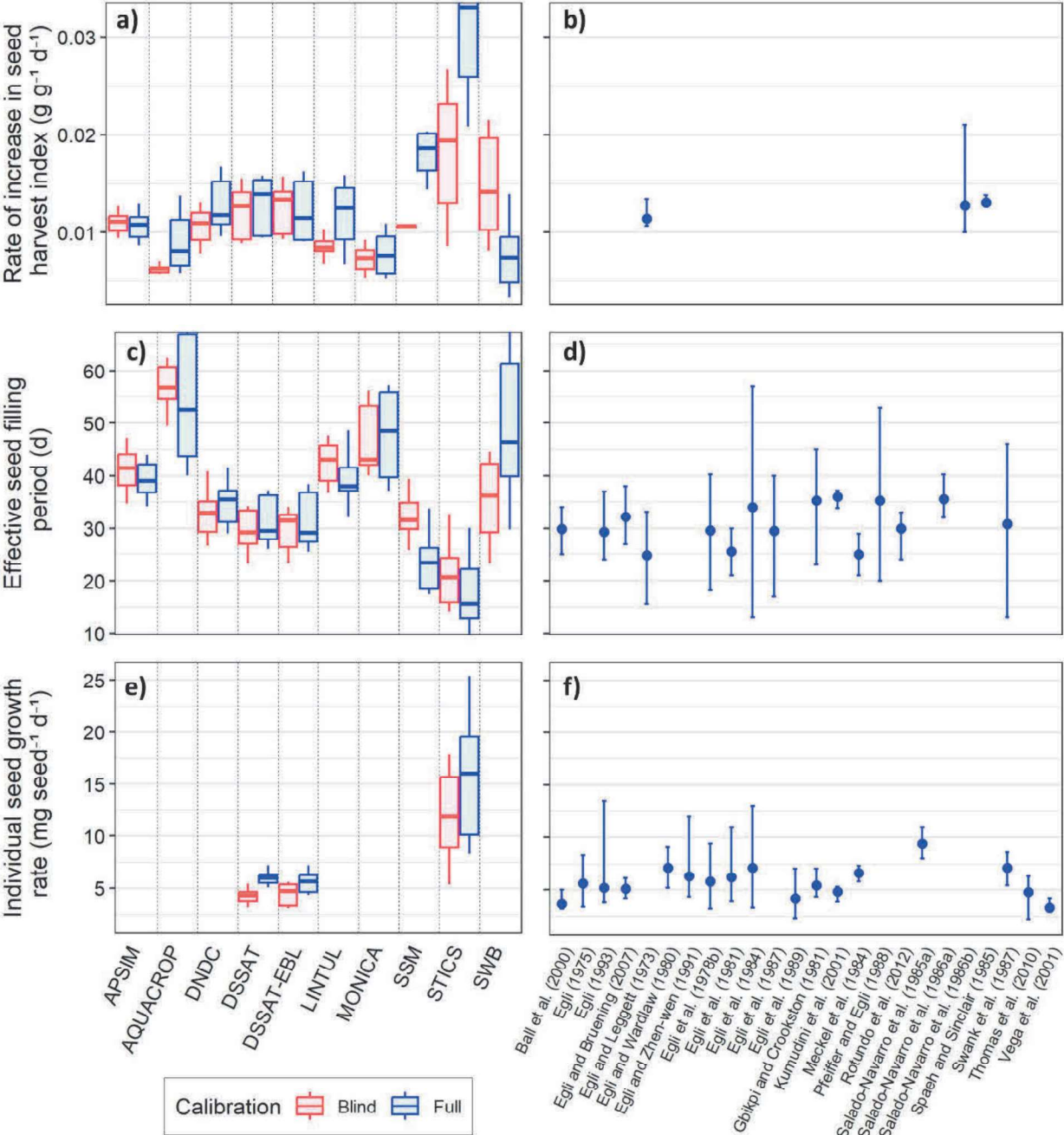


Figure 9. Rate of increase in seed harvest index (a and b), effective seed filling period (c and d), and individual seed growth rate (e and f), calculated from the simulated data by model and compared to independent data from other sites published in the literature. Boxplot show variability across 19 treatments and all locations. The whiskers show 10th and 90th percentiles, while the bounds are 25th and 75th percentiles, with median shown as the line within the box.

We computed the duration of the simulated seed growth as the effective filling period and compared it with literature-reported values (**Fig. 9c-d**). In the literature, the effective seed-filling period ranged by as much as 13–57 days (**Fig. 9c-d**). The effective seed-filling period simulated by most models remained within the range of values typically reported in the literature (**Fig. 9d**). However, we found that models that began accumulating seed weight on the day of anthesis (AQUACROP, MONICA, LINTUL, and SWB) had longer effective seed filling period (median 38–52 days after *Full* calibration) compared to other models (**Fig. 9c**), and in the high range compared to those in the literature (**Fig. 9d**). In contrast, the STICS model had a shorter seed-fill period (16 days) compared to other models (**Fig. 9c**) and literature-reported values (**Fig. 9d**).

Lastly, a group of models (STICS, DSSAT, and DSSAT-EBL) simulated seed number. Given that seed number is a plastic yield component that will change across environments, we calculated the individual seed growth rate to compare with values in the literature (**Fig. 9e-f**). The individual seed growth rate simulated by DSSAT and DSSAT-EBL was within the range of literature reported values, while the rate from STICS model was higher than the reported values.

4. DISCUSSION

4.1. Vegetative growth and partitioning: In-season biomass, leaf, and stem weights

We found that models that performed best under *Blind* calibration for the simulation of vegetative in-season growth (DSSAT, DSSAT-EBL and APSIM) had benefitted from prior evaluations with soybean experimental datasets compared to the other models that had conducted fewer to no evaluations (**Supplementary Fig. S1**). The AQUACROP model ranked third in prior experience (**Supplementary Fig. S1**), but it is a model driven by crop transpiration (Steduto et al.,

2009) with relatively simple crop growth mechanisms that does not differentiate between stem and leaf tissues (**Tables 1-2**). The list of model parameters adjusted to improve the simulation of dry matter production from *Blind* to *Full* calibration (**Supplementary Table 2**) shows that this study provided the opportunity to parameterize some of the less experienced models (MONICA, SWB, and DNDC) to help improve the simulations of in-season vegetative growth during *Full* calibration (**Fig. 4**). In addition, some models (STICS and SWB) modified parameters related to leaf senescence and translocation of assimilates from vegetative to reproductive tissues during *Full* calibration (**Supplementary Table 2**). The measured leaf to vegetative weight ratio showed a linear decline over time, which was mimicked most closely by the DSSAT and DSSAT-EBL models followed by APSIM and LINTUL, even after *Blind* calibration (**Fig. 5**). However, model calibration did not greatly affect the pattern of the leaf to vegetative weight ratio curves over time in most instances, indicating that models continued to partition new vegetative growth similarly after calibration (**Fig. 5**). These results indicate that detailed in-season data for model calibration can partially reduce uncertainty in simulation of in-season growth associated with bias, but to achieve additional model improvement, the refinement of growth and partitioning mechanisms in these soybean models is required.

We identified some opportunities for improving simulation of soybean in-season leaf and stem growth. In the DNDC and STICS, the initial leaf to vegetative weight ratio was too small (value of 0.5) and could be increased to better match measured data (**Fig. 5**). The stem to leaf weight ratio in the DNDC remained constant throughout the growing season and could benefit from adopting a dynamic partitioning approach. In most models, the decline in the leaf to vegetative growth ratio was simulated by partitioning coefficients that are phenology-based, and also indirectly due to leaf senescence. In the DSSAT and DSSAT-EBL, the decline in the leaf to

vegetative mass ratio is mostly driven by new stem growth along with leaf and petiole (stem) senescence as a function of N remobilization (Boote et al., 1998; Jones et al., 2003), but it is also affected by stored carbohydrate and protein accumulation in stems and leaves during early reproductive growth and then subsequent mobilization of carbohydrate and protein to seed. This approach in the DSSAT and DSSAT-EBL models is consistent with experimental data showing that assimilate source for seed growth in soybean is a combination of new photoassimilates, and to a minor extent carbohydrate and protein remobilization from stems and pod shells (Sedigh and Jolliff, 1986). It is possible that under some stress conditions or high yielding environments, models may benefit from accurate simulation of assimilate partitioning into stem and leaf tissue, and accounting for assimilate remobilization from these tissues to final seed yield.

We also observed that the MONICA, SWB, and DNDC models had a high leaf to vegetative weight ratio at harvest since these models did not simulate total abscission of senesced leaf mass. This was partially due to these models originating from cereal crops, where most of the senesced leaf mass remains attached to the stem and is accounted for with total aboveground biomass at harvest. This has implications for using observed total aboveground biomass at harvest or HI data to evaluate these models, and for accurate simulation of organic C and N inputs to the soil and turnover. There is a clear opportunity to improve some models for simulation of end-of-season growth, as well as C and N dynamics by programming senesced leaf mass to become part of the surface soil residue.

Overall, we found that the analysis of the ratios of different tissues in this study was useful to elucidate model differences and identify opportunities to improve models for simulation of in-season growth. The partitioning of growth into different plant components is a topic that has received little attention in prior crop model evaluations, which typically focus on evaluating

models for simulation of total biomass over time or just final yield. Interestingly, our observed data across five contrasting locations showed consistent trends across treatments in the dynamics of ratios for different tissues over time. These results suggest there is an opportunity to develop protocols that simplify sampling operations in experimental trials that collect detailed data for model calibration.

4.2. Leaf area development

The models evaluated in our study also differed in the approaches to simulate LAI, the factors that affect leaf area development, such as temperature, nitrogen, and drought stress (Table 2), and the cardinal temperatures governing leaf appearance rate and expansion (**Supplementary Table 1**). Consequently, models differed in their LAI curves (**Fig. 3**). Battisti et al. (2017b) found differences in the shape of LAI curves among three soybean models and found a strong effect of irrigation on the curves, with faster decline in LAI for rainfed compared to irrigated treatments. This was consistent with results from our study, where under *Blind* calibration, differences in simulated LAI were partially due to uncertainty in simulating water balance and crop water stress. During *Full* calibration, in addition to adjusting soil and crop parameters associated with soil water balance and crop water uptake, modelers adjusted specific leaf area (DSSAT, DSSAT-EBL, LINTUL, MONICA, and SWB), phyllochron (SSM), leaf fraction in biomass (DNDC, LINTUL, and SWB), and relative rate of increase in LAI (LINTUL and STICS) (**Supplementary Table 2**). To improve the simulation of LAI during the leaf senescence phase, the modelers adjusted parameters related to the start of leaf senescence (LINTUL and SWB), relative death rate of leaves (LINTUL, MONICA), and lifespan of an adult leaf (STICS).

Models most experienced with prior soybean datasets (APSIM, DSSAT, and DSSAT-EBL) ranked first for simulation of in-season LAI under both *Blind* and *Full* calibration (**Fig. 4d**).

Interestingly, these models use contrasting approaches for simulation of LAI (Table 2). In the DSSAT and DSSAT-EBL models, leaf area development uses a carbon and temperature hybrid approach, with carbon-based assimilate production and an allocation to leaf that varies as a function of vegetative stage or thermal time (Boote et al., 1998; Hoogenboom et al., 1992). In the APSIM model, leaf area development is strictly temperature-based using thermal time for node/leaf appearance rate with no influence of assimilation (<https://www.apsim.info/>). Even though the different approaches for simulation of LAI did not affect model performance for experienced models, our previous study found that simulation of LAI across models and in turn final yield differed across the same models when evaluating model responses to varying temperature and atmospheric CO₂ concentration (Kothari et al. 2022). Thus, one could hypothesize that the specific approaches for simulation of LAI have limited impact on the performance of experienced models under baseline conditions, but may be a source of uncertainty for climate-change model applications.

For many models, there was high uncertainty for simulation of in-season LAI during *Blind* calibration that was mostly due to large standard bias (**Fig. 4h**). In models that derive leaf expansion based on carbon allocation entirely (MONICA, LINTUL, SWB, and DNDC), the bias in simulation of in-season LAI could be associated to the error simulating in-season aboveground biomass and leaf weight under *Blind* calibration (**Fig. 4a-b**). After *Full* calibration, a lack of correlation between measured and simulated LAI remained the largest source of error (**Fig. 4h**). The leaf area development can be more challenging to mimic than the total aboveground biomass curves because while the biomass increases over time, the LAI curves have both increasing and receding components. The variation among models for LAI was greater in the receding side of the curve in part because of different model approaches for leaf senescence (see example from two

treatments in **Fig. 3**). Our results are consistent with Jégo et al. (2010), who used the STICS model to simulate LAI in Canada and attributed poor model performance for LAI to the inadequate simulation of leaf senescence. Similar results were observed by Sampaio et al. (2021) using DSSAT for different plant densities under a humid low-latitude environment for two different maturity groups. Under no water limitation and high LAI, modest leaf senescence can occur in soybean as a result of decreased PAR intensity in the lower canopy (Pons & Pearcy, 1994). The effect of PAR intensity on leaf senescence is accounted for in DSSAT and DSSAT-EBL (Table 2). Other factors such as leaf age or temperature, water, and N limitation further influence leaf senescence, and were accounted for in most models (Table 2). After the start of rapid seed growth, nitrogen and carbohydrates from senesced tissues are mobilized to the seed. The seed demand for nitrogen in the SSM model is proposed to trigger an “auto-destruction” of stem and leaf tissues that drives leaf senescence in this model, based on the mechanism hypothesized by Sinclair and de Wit (1976). There are contrasting theories on whether seed N demand drives senescence of vegetative tissues in soybean. Other studies demonstrated that leaf senescence does not occur due to seed N demand, but may be regulated by life-cycle processes in the leaf (Hayati et al., 1995; Seddigh and Jolliff, 1986). Overall, leaf senescence remains a major source of uncertainty for simulation of daily LAI in soybean models.

4.3. Reproductive development

Partitioning of assimilates to reproductive growth in crop models is tightly linked to the simulated timing of different reproductive stages, which indicate the start of assimilate allocation to reproductive tissues. For this reason, observed timing of reproductive developmental stages can be a relatively easy non-destructive measurement with high impact to reduce uncertainty for simulation of crop yield. Previous multi-model studies for cereal crops showed the value of

observed phenology data to provide relatively good model simulations with limited observed crop growth data for calibration (Bassu et al., 2014). However, our inter-comparison of soybean models revealed profound differences among models in their structure for simulation of different reproductive stages (**Fig. 1**, Table 1) that have important consequences that limit model calibration with observed data on developmental stages. Overall, models showed large differences in the simulated time of the start of seed growth after beginning flowering (**Fig. 2**). For example, in ‘2000 Irrigated’ treatment in ARGN after *Full* calibration, seed filling started as early as 54 days after planting in LINTUL and as late as 112 days in SSM model (**Fig. 2**). These large variations in the onset of start of reproductive growth were caused by two main reasons: (i) several models originated from cereal crops, and in some of these models the timing of beginning seed growth is still programmed to occur close to anthesis (**Fig. 1**), and (ii) modelers modified the time to start seed growth and its duration to match measured pod growth and final yield during *Full* calibration, with less regard to the timing of measured reproductive stages. As we will discuss below, the first source of uncertainty for simulation of the onset of reproductive growth in soybean is derived from differences in model structure, whereas the second reason is derived from current uncertainty in experimental data for soybean developmental stages.

We found that reducing uncertainty in soybean models with observed phenology data is currently limited by model structure for simulation of reproductive stages. In some models initially developed for cereal crops (AQUACROP, LINTUL, MONICA, and SWB) the onset of seed growth is programmed to coincide with the onset of anthesis (Battisti et al., 2017a; Nendel et al., 2011). This is true for cereal models that already have the rachis and glume structures supporting the kernel at the moment of anthesis. But this approach does not apply well for grain legumes, where after flowering, pods and seed tissues need time to develop, and rapid seed growth can begin

6 to 20 days after beginning of flowering for any individual flower (Ali et al., 2022). In addition, soybean plants keep adding new flowers as well as new nodes over a several-week period. This means that at a given moment in time, soybean plants can have flowers, early developing pods, and seeds undergoing rapid seed growth co-occurring (**Fig. 1**). Of interest, a few models (DNDC, MONICA, SSM) closely simulated the timing of the Beginning pod stage based on different photothermal approaches (Table 1 and 3), but these models either did not simulate pod growth tissue yet (DNDC, SSM), or it was still programmed to coincide with the onset of anthesis (MONICA).

On average, the simulated date of beginning seed growth in some models (AQUACROP, LINTUL, MONICA) occurred 35 days earlier compared to the observed data (Table 3), and increased the simulated duration of seed fill. Alternatively, the SWB model had a delayed start of seed growth that matched the measured beginning seed stage (**Fig. 2**). To compensate for an early start of seed growth, some models with co-occurring date of flowering and beginning seed growth had a reduced rate of increase in seed HI (AQUACROP, MONICA, and LINTUL) or they ended physiological maturity relatively sooner compared to observed data in some instances (AQUACROP) (**Fig. 9**, Table 3). The lag between flowering and the beginning of seed growth in soybean is, in fact, not negligible. In our study, the period from the observed beginning flowering and beginning seed (Fehr and Caviness, 1977) ranged from 24 (AKNS) to 52 (ARGN) days and averaged 41 days among the 19 treatments. In addition, it is important that the soybean models account for the interactive effect of temperature and photoperiod on the rate of development from flowering to beginning pod and seed growth (Table 1). For example, in a regional study conducted in the Mid-South USA (30.6 to 38.9 °N), the time from beginning flowering to beginning seed ranged from 20 to 61 days across locations and planting dates for MG 4 cultivars (Salmerón and

Purcell, 2016). In soybean, long photoperiods are known to prolong the window from flowering to beginning seed growth, and the duration of the window of pod addition (Nico et al., 2016; Summerfield et al., 1998). Including the effect of photoperiod on the rate of development towards reproductive stages should be a priority for soybean models that currently do not account for this effect (Table 1).

The second factor causing variation in the simulated onset of reproductive growth was the uncertainty in experimental data for soybean developmental stages. The standardized monitoring of soybean pod and seed developmental stages is typically based on the four upper-most nodes of the main stem (Fehr and Caviness, 1977). However, field-observed phenology dates for beginning pod and seed based on Fehr and Caviness (1977) occur later than the “real” start of allocation of growth to pods and seeds. Thus the Fehr and Caviness (1977) stages can be indicative but cannot be directly compared to simulated onset of pod and seed addition in crop models. As a result, even for the experienced models in our study such as APSIM and the DSSAT models, the simulated time of beginning seed, was modified during *Full* calibration to better fit observed pod growth data, compared to simulations after *Blind* calibration based on observed phenology dates by Fehr and Caviness (1977) (Table 3). Other studies had previously raised this issue of the discrepancy between typically observed R3 and R5 dates versus modeled start of pod and seed growth (Battisti et al., 2017b; Salmerón and Purcell, 2016). Based on a regional study in the Mid-South US, the lag time between the R5 beginning seed stage by Fehr and Caviness (1977) and the time when seeds of the same developmental stage occur anywhere on the plant averaged 7 days, but varied as much as 0 to 25 days across latitudes, planting dates, and cultivar MGs (Salmeron and Purcell, 2016).

In summary, we identified a clear opportunity for some of the soybean models to include a lag between anthesis and the start of seed growth for a more realistic simulation of soybean reproductive growth. Simulating this lag-phase and the effect of photoperiod on its duration, will be particularly critical for those models attempting to simulate the number of seeds (Nico et al., 2016). We also suggest recording the occurrence of first developing pods and seeds in the plant in field trials, instead or in addition to, the stages based on Fehr and Caviness (1977), to reduce the current uncertainty in simulation of start of reproductive growth in models when only phenology data is available for calibration.

4.4. Reproductive growth and yield formation

Models also differed in their approaches to allocate new growth to reproductive tissues, and in simulating growth of pod shells (or not) and seed (Tables 1 and 2). The approaches for yield determination in the models in this study included setting the rate of harvest index increase or setting the final seed harvest index (APSIM, AQUACROP, DNDIC, and SSM), partitioning fractions to reproductive organs (MONICA, LINTUL, SWB, and STICS), and simulation of individual pod and seed cohorts (DSSAT and DSSAT-EBL) (Table 1). The nRMSE among models in simulating final seed yield was below 20% in seven out of nine models (**Fig. 7e**). This suggests that if the goal of model application is to estimate final seed yield, and if measured crop data for calibration is available, then models with different complexity and experience can perform similarly. The reasonable accuracy for seed yield despite the lack of pod growth simulation in many models could be attributed to the prior calibration and model development efforts that heavily focused on the final seed yield. However, this may not be true where models are used blindly with no local data for calibration. Furthermore, large differences remain in model-

396 simulated yield in response to CO₂ and temperature for climate change applications, outside of the
397 range of conditions that were used for model calibration (Kothari et al., 2022).

398 While calibration improved the simulation of final seed yield, this was not always associated
399 with improved simulation of in-season pod growth dynamics. After *Full* calibration, the analysis
400 of the rate of increase in seed HI (**Fig. 8**), revealed large differences across models. Reducing
401 uncertainty for simulation of seed growth dynamics in soybean models will require a better match
402 between model structure and common data available for calibration from experimental trials. On
403 one hand, observed time-series pod growth data is usually recorded from frequent destructive crop
404 samplings in soybean trials, whereas seed growth data is seldom available (the seed-separating
405 process is very time-consuming for developing seeds). On the other hand, few models are currently
406 capable of simulating pod tissue growth (APSIM, DSSAT, MONICA), limiting the ability to
407 calibrate these models with the most commonly available in-season reproductive growth data. We
408 complemented our model evaluation for seed growth traits with data from the literature (**Fig. 9**),
409 which showed that reproductive growth dynamics varied across models and deviated from the
410 literature for variables such as duration of the seed-filling phase, single seed growth rate, and rate
411 of increase in HI (**Fig. 9**). Further model development to simulate pod tissue growth as well as
412 collecting experimental in-season growth data for evaluating soybean models is essential.

413 During *Full* calibration, the modelers typically adjusted the seed harvest index or
414 partitioning fraction to improve the simulation of in-season reproductive growth and final yield
415 (**Supplementary Table 2**). In addition, in a group of models (DNDC, SWB, and APSIM),
416 parameters affecting crop evapotranspiration and water demand were adjusted to improve
417 simulation in drought-stressed treatments. In MONICA and LINTUL, partitioning fractions are
418 modifiable by developmental stage, allowing more flexibility to the model user. In DSSAT and

DSSAT-EBL, parameters related to maximum weight per seed, pod and seed fill duration, and threshing percent were adjusted during *Full* calibration. The weight per seed was used to adjust seed number and weight per seed in the DSSAT model but had limited effect on simulation of final yield. In the STICS model, the slope of relationship between seed number and growth rate were modified to simulate seed number, and the rate of increase in harvest index was modified to simulate pod growth and final yield. Unexpectedly, the changes made in STICS for improving final seed yield led to large deviation in the individual seed growth rate, and the rate of increase in seed harvest index when compared with other models and literature-reported values (**Fig. 9**).

Out of the ten models evaluated in this study, only DSSAT and STICS models simulated seed number. Detailed data on seed number over the growing season, and on the duration of the window of flower and pod addition would be helpful to evaluate and better parameterize these models. However, there are currently no standardized approaches to quantify the duration of the window of flowering and pod addition in field trials. Simulation of seed number may be relevant because for soybean and other grain crops, seed number is a yield component of high plasticity that helps adapt the size of the reproductive sink to the resources in a given environment (Borras et al., 2004; Sadras, 2007). Thus, differences in soybean yield across environments are best associated with differences in seed number, relative to the other major yield component, weight per seed (Egli et al., 1998). Yield components in soybean are determined in a sequential and partially overlapping way, with seed number determined before seed weight (see window of pod addition in **Fig. 1**). Seed size in soybean can also have considerable variation in response to changes in assimilate supply during the seed filling phase (Borrás et al., 2004; Chiluwal et al., 2022). After *Full* calibration, most models with a partitioning or HI approach simulated final yield with similar error compared to models that have a sequential simulation of yield components and

growth of individual sinks (DSSAT, DSSAT-EBL and STICS) (**Fig. 7e**). However, considering the sequential determination of yield components in more mechanistic crop models may increase their capacity to accurately account for environmental stressors and simulate soybean yield outside of calibration environments. The logical assumption is that HI and allocation approaches may “fail” by overpredicting yield in extreme stress conditions, unless empirical equations that account for stress factors differently during developmental stages are implemented in the code. For instance, the effect of high temperatures around anthesis is implemented in the DNDC model (**Supplementary Table 1**). Simulation of different yield components and sequential addition of pod cohorts, may provide opportunities for expanding the model applications and improving model performance under high stress conditions and extreme events, for the models that do not currently include this complexity.

5. CONCLUSIONS

We conducted a multi-model evaluation of ten crop models with detailed multi-site soybean growth data to investigate the current state of eco-physiological models for simulating in-season growth dynamics. In addition to evaluating models for in-season dry matter accumulation, which is typically explored in modeling studies, we analyzed the ratios of dry matter partitioning to vegetative and reproductive tissues, which has received little attention in the past studies. We found that our assessment of current crop model performance for simulation of soybean in-season growth dynamics was helpful to identify opportunities for model improvement, in particular for models with limited prior evaluation with experimental data. In addition, our study highlights bottlenecks in experimental data used for model evaluation, and recommends collection of key crop developmental and growth data in future studies.

Acknowledgements

This work was supported by the National Institute of Food and Agriculture, United States Department of Agriculture [grant number 2019-67019-29470], and the University of Kentucky [hatch funds]

Author contributions

MS, KK, KJB, and RB coordinated the study, analyzed the data, and prepared the first draft of the manuscript. Crop modelers team comprised of RB, KJB, SVA, JC, SVC, PD, BF, BG, GH, QJ, MvdL, FRM, AN, CN, BQ, CS, EHFMS, WS, AS, AS, and NAV, who conducted crop model simulations. AC, LCP, FAMS, and SVA provided field experimental data. ACR prepared the weather data; all authors read and edited the final manuscript.

1. REFERENCES

- Ali, M.F., Brown, P., Thomas, J., Salmeron, M., Kawashima, T. 2022. Effect of assimilate competition during early seed development on the pod and seed growth traits in soybean. *Plant Reproduction*, 35(3): 179-188.
- Annandale, J.G., Campbell, G.S., Olivier, F.C. and Jovanovic, N.Z., 2000. Predicting crop water uptake under full and deficit irrigation: An example using pea (*Pisum sativum* L. cv. Puget). *Irrigation Science*, 19(2): 65-72.
- Asseng, S. et al., 2015. Rising temperatures reduce global wheat production. *Nat. Clim. Change*, 5(2): 143-147.
- Ball, R.A., Purcell, L.C. and Vories, E.D., 2000. Short-season soybean yield compensation in response to population and water regime. *Crop Science*, 40(4): 1070-1078.

485 Bassu, S. et al., 2014. How do various maize crop models vary in their responses to climate
486 change factors? *Global Change Biology*, 20(7): 2301-2320.

487 Battisti, R., Parker, P.S., Sentelhas, P.C. and Nendel, C., 2017a. Gauging the sources of
488 uncertainty in soybean yield simulations using the MONICA model. *Agricultural*
489 *Systems*, 155: 9-18.

490 Battisti, R., Sentelhas, P.C. and Boote, K.J., 2017b. Inter-comparison of performance of soybean
491 crop simulation models and their ensemble in southern Brazil. *Field Crops Research*, 200:
492 28-37.

493 Battisti, R. et al., 2018. Assessment of crop-management strategies to improve soybean
494 resilience to climate change in Southern Brazil. *Crop and Pasture Science*, 69(2): 154-
495 162.

496 Boote, K.J., Jones, J.W., Hoogenboom, G. and Pickering, N.B., 1998. The CROPGRO model for
497 grain legumes. In: G.Y. Tsuji, G. Hoogenboom and P.K. Thornton (Editors),
498 *Understanding Options for Agricultural Production*. Springer Netherlands, Dordrecht, pp.
499 99-128.

500 Borrás, L., Slafer, G.A. and Otegui, M.a.E., 2004. Seed dry weight response to source–sink
501 manipulations in wheat, maize and soybean: a quantitative reappraisal. *Field Crops*
502 *Research*, 86(2): 131-146.

503 Brisson, N. et al., 2003. An overview of the crop model STICS. *European Journal of Agronomy*,
504 18(3): 309–332.

505 Chiluwal, A., Kawashima, T. and Salmeron, M. 2021. Soybean seed weight responds to
506 increases in assimilate supply during late seed-fill phase. *Journal of Crop Improvement*
507 (36): 222-238.

508 Cuadra, S.V., Kimball, B.A., Boote, K.J., Suyker, A.E. and Pickering, N., 2020. Energy balance
509 in the DSSAT-CSM-CROPGRO model. *Agricultural and Forest Meteorology*: 108241.

510 Ebrahimi-Mollabashi, E. et al., 2019. Enhancing APSIM to simulate excessive moisture effects
511 on root growth. *Field Crops Res.*, 236: 58-67.

512 Egli, D., 1975. Rate of accumulation of dry weight in seed of soybeans and its relationship to
513 yield. *Canadian Journal of Plant Science*, 55(1): 215-219.

514 Egli, D.B., 1993. Cultivar maturity and potential yield of soybean. *Field Crops Research*, 32(1):
515 147-158.

516 Egli, D.B. and Bruening, W.P., 2007. Accumulation of nitrogen and dry matter by soybean seeds
517 with genetic differences in protein concentration. *Crop Science*, 47(1): 359-366.

518 Egli, D.B., Fraser, J., Leggett, J.E. and Poneleit, C.G., 1981. Control of seed growth in soya
519 beans [*Glycine max* (L.) Merrill]. *Annals of Botany*, 48(2): 171-176.

520 Egli, D.B., Leggett, J.E. and Wood, J.M., 1978. Influence of soybean seed size and position on
521 the rate and duration of filling. *Agronomy Journal*, 70(1): 127-130.

522 Egli, D.B., Orf, J.H. and Pfeiffer, T.W., 1984. Genotypic variation for duration of seedfill in
523 soybean. *Crop Science*, 24(3): 587-592.

524 Egli, D.B., Ramseur, E.L., Zhen-wen, Y. and Sullivan, C.H., 1989. Source-sink alterations affect
525 the number of cells in soybean cotyledons. *Crop Science*, 29(3): 732-735.

526 Egli, D.B. and Wardlaw, I.F., 1980. Temperature response of seed growth characteristics of
527 soybeans. *Agronomy Journal*, 72(3): 560-564.

528 Egli, D.B. and Zhen-wen, Y., 1991. Crop growth rate and seeds per unit area in soybean. *Crop*
529 *Science*, 31(2): 439-442.

530 Egli, D.B. (1998) Seed Biology and the yield of grain crops. CAB International, Wallingford,
531 UK.

532 FAO, 2020. Food and Agriculture Organization crop statistical data, retrieved from
533 fao.org/faostat on June 24 2020.

534 Fehr, W.R. and Caviness, C.E., 1977. Stages of soybean development. Cooperative Extension
535 Service, Agriculture and Home Economics Experiment Station, Iowa State University,
536 Ames, Iowa.

537 Gauch, H.G., Hwang, J.G. and Fick, G.W., 2003. Model evaluation by comparison of model-
538 based predictions and measured values. *Agronomy Journal*, 95(6):1442-1446.

539 Gbikpi, P.J. and Crookston, R.K., 1981. Effect of flowering date on accumulation of dry matter
540 and protein in soybean seeds. *Crop Science*, 21(5): 652-655.

541 Hasegawa, T. et al., 2017. Causes of variation among rice models in yield response to CO₂
542 examined with Free-Air CO₂ Enrichment and growth chamber experiments. *Scientific*
543 *Reports*, 7(1): 14858.

544 Hayati, R., Egli, D.B. and Crafts-Brandner, S.J., 1995. Carbon and Nitrogen Supply during Seed
545 Filling and Leaf Senescence in Soybean. *Crop Science*, 35(4):
546 [cropsci1995.0011183X003500040024x](https://doi.org/10.2135/cropsci1995.0011183X003500040024x).

547 Hoogenboom, G., Jones, J. and Boote, K., 1992. Modeling growth, development, and yield of
548 grain legumes using SOYGRO, PNUTGRO, and BEANGRO: a review. *Transactions of*
549 *the ASAE*, 35(6): 2043-2056.

550 Jégo, G. et al., 2010. Calibration and performance evaluation of soybean and spring wheat
551 cultivars using the STICS crop model in Eastern Canada. *Field Crops Research*, 117(2):
552 183-196.

553 Jones, J.W. et al., 2003. The DSSAT cropping system model. *European Journal of Agronomy*,
554 18(3): 235-265.

555 Kothari, K. et al., 2022. Are soybean models ready for climate change food impact assessments?
556 *European Journal of Agronomy*, 135 (2022): 126482.

557 Kumudini, S., Hume, D.J. and Chu, G., 2001. Genetic improvement in short season soybeans: I.
558 Dry matter accumulation, partitioning, and leaf area duration. *Crop Science*, 41(2): 391-
559 398.

560 Li, C., Frolking, S. and Frolking, T.A., 1992. A model of nitrous oxide evolution from soil
561 driven by rainfall events: 1. Model structure and sensitivity. *Journal of Geophysical*
562 *Research: Atmospheres*, 97(D9): 9759-9776.

563 Martre, P. et al., 2015. Multimodel ensembles of wheat growth: many models are better than one.
564 *Global Change Biology*, 21(2): 911-925.

565 Meckel, L., Egli, D.B., Phillips, R.E., Radcliffe, D. and Leggett, J.E., 1984. Effect of moisture
566 stress on seed growth in soybeans. *Agronomy Journal*, 76(4): 647-650.

567 Nehbandani, A., Soltani, A., Rahemi-Karizaki, A., Dadrasi, A. and Noubakhsh, F., 2021.
568 Determination of soybean yield gap and potential production in Iran using modeling
569 approach and GIS. *Journal of Integrative Agriculture*, 20(2): 395-407.

570 Nehbandani, A., Soltani, A., Taghdisi Naghab, R., Dadrasi, A. and Alimagham, S.M., 2020.
571 Assessing HC27 Soil Database for Modeling Plant Production. *International Journal of*
572 *Plant Production*, 14(4): 679-687.

573 Nendel, C. et al., 2011. The MONICA model: Testing predictability for crop growth, soil
574 moisture and nitrogen dynamics. *Ecological Modelling*, 222(9): 1614-1625.

575 Nendel, C. et al., 2022. Area expansion outweighs increasing drought risk for soybean in Europe
576 (Under review). Glob. Change Biol.

577 Nico, M., Mantese, A.I., Miralles, D.J. and Kantolic, A.G., 2016. Soybean fruit development and
578 set at the node level under combined photoperiod and radiation conditions. Journal of
579 Experimental Botany, 67(1): 365-377.

580 Ordóñez, R.A. et al., 2018. Maize and soybean root front velocity and maximum depth in Iowa,
581 USA. Field Crops Res., 215: 122-131.

582 Robertson, M. and Carberry, P., 1998. Simulating growth and development of soybean in
583 APSIM.

584 Rosenzweig, C. et al., 2014. Assessing agricultural risks of climate change in the 21st century in
585 a global gridded crop model intercomparison. Proceedings of the National Academy of
586 Sciences, 111(9): 3268.

587 Rosenzweig, C. et al., 2013. The Agricultural Model Intercomparison and Improvement Project
588 (AgMIP): Protocols and pilot studies. Agricultural and Forest Meteorology, 170: 166-
589 182.

590 Sadras, V. 2007. Evolutionary aspects of the trade-off between seed size and number in crops.
591 Field Crops Research 100: 125-138

592 Salado-Navarro, L.R., Sinclair, T.R. and Hinson, K., 1986. Yield and reproductive growth of
593 simulated and field-grown soybean. II. Dry matter allocation and seed growth rates. Crop
594 Science, 26(5): 971-975.

595 Salmerón, M. and Purcell, L.C., 2016. Simplifying the prediction of phenology with the DSSAT-
596 CROPGRO-soybean model based on relative maturity group and determinacy.
597 Agricultural Systems, 148: 178-187.

598 Sampaio, L.S, Battisti, R., Lana, M.A., Boote, K.J. 2021. Assessment of sowing dates and plant
599 densities using CSM-CROPGRO-Soybean for soybean maturity groups in low latitude.
600 Journal of Agricultural Science, 158(10): 819-832.

601 Seddigh, M. and Jolliff, G.D., 1986. Remobilization Patterns of C and N in Soybeans with
602 Different Sink-Source Ratios Induced by Various Night Temperatures 1. Plant
603 Physiology, 81(1): 136-141.

604 Silva, J.V. and Giller, K.E., 2021. Grand challenges for the 21st century: what crop models can
605 and can't (yet) do. The Journal of Agricultural Science: 1-12.

606 Sinclair, T.R., 1986. Water and nitrogen limitations in soybean grain production I. Model
607 development. Field Crops Res., 15(2): 125-141.

608 Sinclair, T.R. and de Wit, C.T., 1976. Analysis of the Carbon and Nitrogen Limitations to
609 Soybean Yield1. Agronomy Journal, 68(2): 319-324.

610 Spaeth, S.C., Sinclair., T.R. 1985. Linear increase in soybean harvest index during seed-filling.
611 Agronomy Journal, 77(2): 207-211.

612 Steduto, P., Hsiao, T.C., Raes, D. and Fereres, E., 2009. AquaCrop—The FAO Crop Model to
613 Simulate Yield Response to Water: I. Concepts and Underlying Principles. Agronomy
614 Journal, 101(3): 426-437.

615 Summerfield, R.J., Asumadu, H., Ellis, R.H. and Qi, A., 1998. Characterization of the
616 Photoperiodic Response of Post-flowering Development in Maturity Isolines of Soyabean
617 [Glycine max (L.) Merrill] ‘Clark’. Annals of Botany, 82(6): 765-771.

618 Swank, J.C., Egli, D.B. and Pfeiffer, T.W., 1987. Seed growth characteristics of soybean
619 genotypes differing in duration of seed fill. Crop Science, 27(1): 85-89.

- Thomas, J., Boote, K., Pan, D. and Allen, L., 2010. Elevated temperature delays onset of reproductive growth and reduces seed growth rate of soybean. *J. AgroCrop Sci*, 1: 19-32.
- Van Roekel, R.J., Purcell, L.C. and Salmerón, M., 2015. Physiological and management factors contributing to soybean potential yield. *Field Crops Res.*, 182: 86-97.
- Vega, C.R.C., Andrade, F.H. and Sadras, V.O., 2001. Reproductive partitioning and seed set efficiency in soybean, sunflower and maize. *Field Crops Research*, 72(3): 163-175.
- Wallach, D. et al., 2021. How well do crop modeling groups predict wheat phenology, given calibration data from the target population? *European Journal of Agronomy*, 124: 126195.
- Wang, E. et al., 2017. The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nature Plants*, 3(8): 17102.
- White, J.W., Hoogenboom, G., Kimball, B.A. and Wall, G.W., 2011. Methodologies for simulating impacts of climate change on crop production. *Field Crops Res.*, 124(3): 357-368.
- Wolf, J., 2012. User guide for LINTUL5: Simple generic model for simulation of crop growth under potential, water limited and nitrogen, phosphorus and potassium limited conditions.
- Zhang, Y. and Niu, H., 2016. The development of the DNDC plant growth sub-model and the application of DNDC in agriculture: A review. *Agric. Ecosyst. Environ.*, 230: 271-282.