in silico Plants Vol. 3, No. 1, pp. 1–12
doi:10.1093/insilicoplants/diab004
Advance Access publication 23 March 2021
Special Issue: Linking Crop/Plant Models and Genetics
Research Article

in silico Plants
OXFORD

Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models

Ryan F. McCormick[1], Sandra K. Truong[1], Jose Rotundo[1], Adam P. Gaspar[2], Don Kyle[3], Fred van Eeuwijk[4] and Carlos D. Messina[1,*]

[1]Predictive Agriculture, Corteva Agriscience, Johnston, IA, USA
[2]Integrated Field Sciences, Corteva Agriscience, Jainesville, WI, USA
[3]Plant Breeding, Corteva Agriscience, Princeton, IL, USA
[4]Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands
*Corresponding author's e-mail address: charlie.messina@corteva.com

Guest Editor: Xinyou Yin
Editor-in-Chief: Stephen P. Long

Citation: McCormick RF, Truong SK, Rotundo J, Gaspar AP, Kyle D, van Eeuwijk F, Messina CD. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. *In Silico Plants* **2021:** diab004; doi: 10.1093/insilicoplants/diab004

ABSTRACT

The timing of crop development has significant impacts on management decisions and subsequent yield formation. A large intercontinental dataset recording the timing of soybean developmental stages was used to establish ensembling approaches that leverage both knowledge-based, human-defined models of soybean phenology and data-driven, machine-learned models to achieve accurate and interpretable predictions. We demonstrate that the knowledge-based models can improve machine learning by generating expert-engineered features. The collection of knowledge-based and data-driven models was combined via super learning to both improve prediction and identify the most performant models. Stacking the predictions of the component models resulted in a mean absolute error of 4.41 and 5.27 days to flowering (R1) and physiological maturity (R7), providing an improvement relative to the benchmark knowledge-based model error of 6.94 and 15.53 days, respectively, in cross-validation. The hybrid intercontinental model applies to a much wider range of management and temperature conditions than previous mechanistic models, enabling improved decision support as alternative cropping systems arise, farm sizes increase and changes in the global climate continue to accelerate.

KEYWORDS: crop model; ensemble; machine learning; phenology; soybean; super learner

1. INTRODUCTION

The timing of the changes in the life stages of a crop, also referred to as the crop's phenology, represents primary determinants of the suitability of a crop for a growing region, its yield and major management decisions such as planting date and agronomic treatments (Shaykewich 1995). Allelic diversity in genetic pathways regulating crop phenology introduces significant variation in how, both within and between species, plants integrate environmental information and determine developmental stages. Models built to predict phenology for agronomic decision support are often relevant to only a limited set of prediction tasks such as a limited range of relative maturities (RM), planting dates, locations and/or photoperiod sensitivities, due to the inherent difficulty of building and training a broadly applicable and accurate model (Shaykewich and Bullock 2018; dos Santos *et al.* 2019). However, as new sustainability practices and climate change motivate adoption of alternative management practices, such as planting in new target environments, seed increases in winter nurseries or later planting of a shorter maturity crop in alternative cropping systems, more comprehensive approaches to phenology prediction are necessary.

Soybean is the world's fourth largest crop as measured by area harvested and an important source of protein and oil (FAOSTAT 2016). As in most plants, the soybean life cycle is a well-regulated process that integrates environmental cues and internal states to determine the onset of phenological events. The timings of these events are the primary drivers of plant performance and reproductive success (Andrés and Coupland 2012). Due to the economic and food security consequences that the timing of these events have, the understanding and prediction of soybean phenology have been a focus of research for decades (Brown 1960; Hesketh *et al.* 1973; Wang *et al.* 1987; Cao *et al.* 2017; Shaykewich and Bullock 2018). The present study builds off of this work by leveraging existing knowledge-based models of how the plant integrates key environmental determinants influencing phenology (Grimm *et al.* 1993; Jones *et al.* 2003; Setiyono *et al.* 2007; Salmerón and Purcell 2016). Ultimately, this research was sought to generate modelling strategies for predicting soybean phenology across disparate geographies and training procedures to generate an intercontinentally useful model.

Advances in computing and data science have driven massive increases in data availability and a concomitant increase in models describing biological systems, and models describing these systems vary in their purpose, accuracy, correctness, granularity and interpretability. Prior to the increase in compute power, models of plant phenotypic outcomes given an environment typically existed either as parametric statistical models with explicit G×E interactions or as knowledge-based models composed of functions explicitly defining plant processes (Sinclair 1986; Prusinkiewicz 2004; van Eeuwijk *et al.* 2016). More recent efforts have focused on integrating G×E by coupling whole genome prediction with dynamical crop growth models, thereby generating phenotypic outcomes via non-linear functions of marker effects and environmental inputs (Technow *et al.* 2015; Cooper *et al.* 2016; Onogi *et al.* 2016; Messina *et al.* 2018). Purely data-driven approaches trained via machine learning have also garnered considerable attention as data generation continues to become more routine and higher throughput for plant systems (Liakos *et al.* 2018; Taghavi Namin *et al.* 2018; Shakoor *et al.* 2019). Parametric statistical models and process-based models can provide inferential and predictive ability in relatively data-poor environments and bring interpretability and understanding of the system being modelled, whereas in data-rich environments, data-driven machine-learned models often yield high predictive accuracy at the expense of interpretability. This tradeoff has motivated interest in hybrid modelling approaches in an effort to leverage the strengths of different modelling approaches and to mitigate their respective weaknesses (Fan *et al.* 2015; Hamilton *et al.* 2017; Karpatne *et al.* 2017*a*, *b*; Roberts *et al.* 2017; Oyetunde *et al.* 2018; Pathak *et al.* 2018).

While existing knowledge-based models can perform well under specific conditions or with laborious calibration, they are not sufficiently generalizable to support agronomic and breeding decisions across a variety of environmental conditions and continuous RM (as opposed to discrete maturity groups). One possible approach to a general phenology model is the calibration or fitting of existing discrete-time dynamical systems models such as CROPGRO's phenology module (hereafter referred to as CROPGRO for brevity) or SOYDEV (Boote *et al.* 1998; Jones *et al.* 2003; Setiyono *et al.* 2007; Salmerón and Purcell 2016). However, re-fitting these models to new data often requires specialized experiments whose labour requirements restrict throughput to relatively few assays (Grimm *et al.* 1993). Like many biological models, these models are overparameterized, and they are unidentifiable given only the observations of calendar dates to developmental stages that are typically measured in applied field settings. This problem of fitting unidentifiable models with non-convex cost landscapes and limited data is not unique to the life sciences; while strategies exist to re-parameterize the complex model to assist model fitting, crop growth modelers have generally employed various optimization approaches to regularize and fit the complex crop growth model (Wallach *et al.* 2001; He *et al.* 2010; Archontoulis *et al.* 2014; Sexton *et al.* 2016; Lamsal *et al.* 2017; Messina *et al.* 2018). The best optimization strategy will vary with the size and composition of the training dataset, sensitivity of the output to the selected set of input parameters, choice of priors and regularization schemes and model runtime. Inspired by the success of machine learning approaches to identify optima (even if not the global optimum) that perform well in out-of-sample prediction tasks for models with large numbers of parameters (e.g. neural networks for image processing), this work develops a highly parallelized, gradient-free, multimodal optimization strategy to explore high-dimensional parameter space and find satisfactory fits to available data (Whitley 1994; Spall 1998; Kennedy 2010). This approach is used to recalibrate existing knowledge-based models given a large dataset of field observations, and the recalibrated models are shown to have improved prediction accuracy.

As an alternative to knowledge-based models, a second approach to a general phenology model is the utilization of advances in time-series modelling and model training made by the machine learning community. Specifically, artificial neurons that retain an internal state or memory can modulate their output based on past input; artificial neural networks built from such neurons can learn complex relationships between temporal sequences and output (e.g. speech to text applications). In other words, these networks have the potential to capture the sequence-dependent impact of environmental stimuli, such as a period of cool weather, on plant development. As such, network architectures including recurrent neural networks (RNN), such as the long short term memory (LSTM) networks and their relatives, represent useful tools to model time-series data and predict phenological states (Hochreiter and Schmidhuber 1997). These data-driven models have the potential to learn a mapping between daily inputs over time and the output phenological state on a given day, providing a data-driven analogue to the knowledge-based models mentioned previously. While the use of artificial neural networks to model soybean phenology is not new, the utility of an LSTM network to serve as a daily phenology model has not been explored to our knowledge (Elizondo *et al.* 1994; Zhang *et al.* 2009).

Moreover, it stands to reason that since knowledge-based models represent a formal encapsulation of decades of research into system behaviour, providing the prediction of a knowledge-based model as a feature should reduce the training burden on a machine-learned model and improve its predictive skill (Pathak *et al.* 2018). In this sense, knowledge-based models could be used to generate expert-engineered features from which a data-driven model could learn more effectively, and we explore this possibility in this paper.

The different model training approaches applied herein for the knowledge-based models and for the data-driven models do not have any guarantees regarding convergence on a global best fit, and their learning process is stochastic by nature. Therefore, instead of assuming that there exists a single 'best' (i.e. most generalizable) model structure and parameters, model-selection approaches are eschewed for an ensembling technique that builds reliable meta-models given the predictions from the component models. This meta-model learns from the component model predictions and is often referred to as a super learner. Most any regression procedure can serve as a super learner to stack the model predictions, and modern data science has successfully employed a variety of approaches from simple linear regression to artificial neural networks as super learners (Polley and van der Laan 2010; Naimi and Balzer 2018). In terms of providing uncertainty estimates regarding the prediction, Bayesian model averaging (BMA) and related approaches have emerged as practical ensembling techniques with useful properties, in that they weigh individual models to provide some information on relative importance in the final ensemble prediction (Raftery *et al.* 2005; Yao *et al.* 2018).

In this work, an imbalanced, transcontinental soybean phenology dataset consisting of 13 673 records in 187 unique environments (defined here as unique combinations of planting date, latitude and longitude) was used to demonstrate four concepts: (1) a multimodal optimization strategy that can be used to train a crop growth model with a large dataset, (2) LSTM networks can be trained via machine learning to accurately predict soybean phenology, (3) knowledge-based model features can improve machine learning processes and (4) combining the component model predictions via super learning reliably generates accurate out-of-sample predictions. The combined model applies to a wide variety of environments with a median out-of-sample mean absolute error (MAE) of 4.41 and 5.27 days to R1 and R7, respectively, representing a 34% and 66% improvement over the best pre-existing model tested (6.94 and 15.53 days MAE) across the 10 cross-validation folds. Accuracy for these and other key phenological stages is sufficient to direct management decisions and inform yield formation. We show that this approach yields considerable improvements in predictive power while still retaining some interpretability via the knowledge-based models to guide future experimentation and better understand system behaviour underlying soybean phenology.

2. RESULTS

2.1 Dataset collection, evaluation of existing soybean phenology models and model training overview

Data compilation from different public sources of relevant multi-environment trials and planting date experiments across years and a variety of latitudes was carried out and combined with data from trials internal to Corteva Agriscience™. This resulted in 187 unique environments (defined as unique combinations of latitude, longitude and planting date) with 13 673 records of observations of the calendar day of flowering (R1) and physiological maturity (R7) for a range of RM groups (Fehr and Caviness 1977). Of these 13 673 records, a subset also had observations for emergence and/or beginning seed filling (R5): 7043 records had emergence and 8212 records had R5. 5488 records included all four developmental stages. While additional phenological stages are characterized for soybean, the application for this study prioritized accurate predictions of R1 and R7, followed by R5 and emergence. The sampled environments range from −38.36° to 49.44° in latitude, covering a wide range of latitudes in growing regions of North and South America and representing approximately 83% of annual global soybean production; sampled lines range in RM from −1.5 to 9.2 (i.e. less than maturity group (MG) 0 to greater than MG IX). Studies from which public data were obtained are listed in **Supporting Information—Table S1**.

A variety of soybean phenology models exist which generally serve to convert a time-series of daily temperatures and photoperiod to the calendar day of a developmental event, generally conditioned on an RM group (Shaykewich and Bullock 2018). To approximate the state-of-the-art performance in soybean phenology modelling in the intercontinental dataset, the soybean phenology models described in CROPGRO and SOYDEV were re-implemented and applied without any calibration (Grimm *et al.* 1993; Jones *et al.* 2003; Setiyono *et al.* 2007; Salmerón and Purcell 2016). After generalizing the discrete maturity parameters of the respective models to a continuous range of RM (i.e. make a model parameter a function of continuous RM instead of a discrete maturity group), both models demonstrated acceptable performance in some cases, but failed to generalize well across all environments and maturities (Fig. 1). This is unsurprising, given that the range of photoperiods, temperatures and maturities in the intercontinental dataset is outside of those sampled when the original model developers built their systems of equations and fit their model parameters. When predicting into the entire dataset, CROPGRO displayed useful performance for planting to R1 (MAE = 6.90 days) and exhibited substantial bias towards underprediction of days from planting to R7 (MAE = 15.47 days); SOYDEV displayed substantial error for both planting to R1 and planting to R7 (MAE = 10.59 and 25.14 days, respectively), particularly for maturities outside of those originally parameterized by Setiyono *et al.* (2007). As is generally the case, the models would need to be calibrated prior to use for decision support since predictions with their default parameters would be unsuitable for the range of environments and RM present in the current prediction task.

Given the original dataset of 187 environments, 10 independent, mutually exclusive testing sets were generated on the basis of these environments. Each testing set comprised 10% of the entire dataset, thereby creating 10-folds; the remaining 90% of environments in each fold were randomly divided into a training and a tuning set in a 60:40 split (Fig. 2). The training, tuning, and testing sets were generated on a per-environment basis, in which a single environment could contain multiple records of phenological observations. Set partitioning was done on the basis of environments rather than records because, otherwise, models may have already learned using observations from the target environment; this represents the more trivial use case of predicting outcomes in an already observed environment. The training and tuning sets were used to fit component models and super learners, respectively; final prediction accuracy of the model training pipeline was evaluated on records from the testing set environments.

2.2 Training component models: fitting knowledge-based phenology models

Fitting complex biological models to data remains a challenging problem, and a number of optimization strategies have been

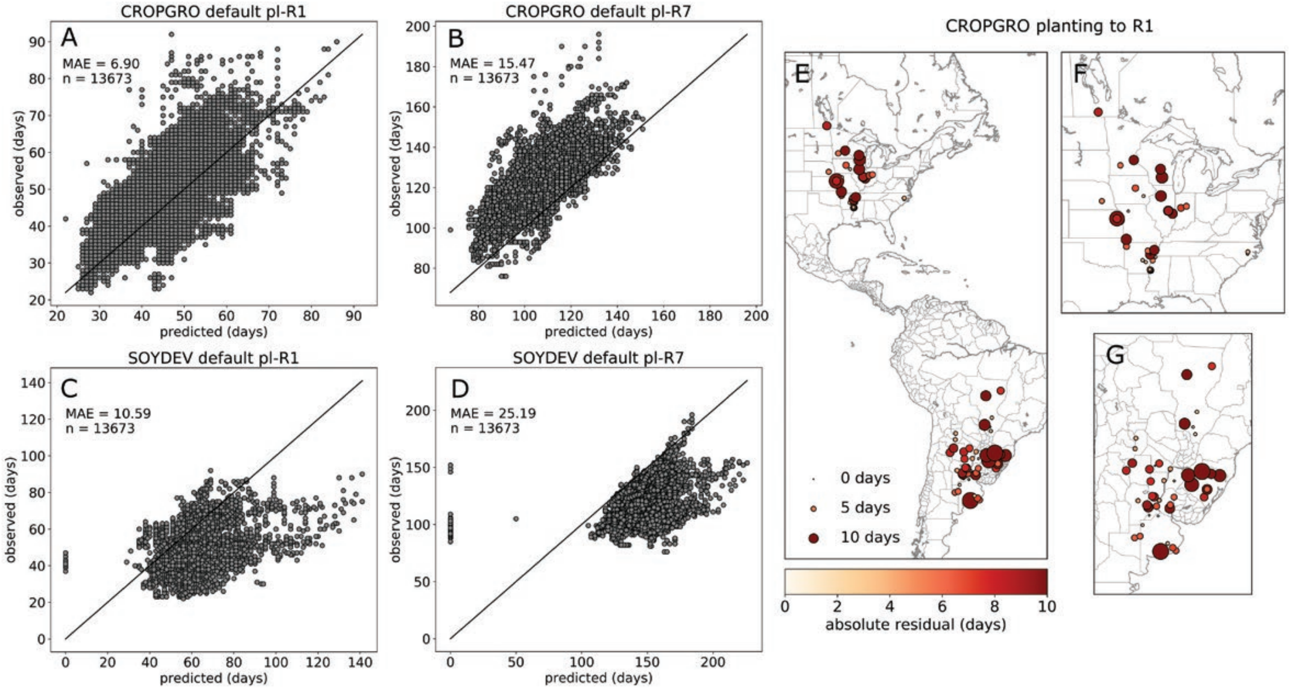

Figure 1. Performance of state-of-the-art phenology models from CROPGRO and SOYDEV. Black line in scatter plots shows the one-to-one line for observed versus predicted CROPGRO planting to R1 and planting to R7 (A, B) and likewise for SOYDEV (C, D). Predictions of zero by SOYDEV occurred due to RM less than 0 failing to ever reach any stage, resulting in a prediction assigned as zero. For CROPGRO, default parameters generalized well for planting to R1 and R7 (A), though showed consistent underprediction for planting to R7 (B). Mean absolute residual by site using CROPGRO for planting to R1 (E–G). Residuals were not uniformly distributed with respect to factors like RM, reflecting high model error when predicting cases outside of those the model was originally built for (data not shown).

employed to obtain parameter estimates for crop growth models or their phenology models (Wallach *et al.* 2001; He *et al.* 2010; Archontoulis *et al.* 2014; Sexton *et al.* 2016; Lamsal *et al.* 2017; Messina *et al.* 2018). The approach used here assumes that there may be more than one equivalently good solution given the data (sometimes referred to as equifinality) and employs a highly parallelized, multimodal optimization strategy (Wong 2015). In this manner, large populations of particles, each representing a parameter vector with a corresponding goodness of fit to the dataset, are used to explore the parameter space. Because the observations may fail to constrain the fit to a unique solution given the model structure, distinct particles may have similar goodness of fits. While this optimization strategy provides the opportunity to examine families of parameters that may cluster differently in parameter space but have equivalent fits, a single particle with the best fit (i.e. lowest cost) was chosen to represent the best model parameters given the training set and used for subsequent analyses of a given fold. Thus, while the fit model generalizes well for prediction tasks, there are no guarantees that the variation explained by the best fitting model is driven by the biologically correct parameters, nor even the globally optimum parameters, and the fit parameters should either not be used for inference or used with caution. Most every parameter in the model, including parameters such as optimal temperatures and photoperiod responses for different developmental phases, were jointly estimated, for a total of 36 free parameters in CROPGRO [**see Supporting Information**]. Of note, poor performance of both the default SOYDEV model and the re-fit SOYDEV model relative to CROPGRO led to subsequently dropping SOYDEV from further analyses; the lower performance of SOYDEV was potentially a consequence of the functions SOYDEV uses to relate RM to parameter values, where the shapes of these functions were unable to represent behaviour outside of the RM (roughly RMs 1–4) fit by Setiyono *et al.* (2007).

Re-fitting the CROPGRO model resulted in improved out-of-sample performance across the 10-folds when predicting into the tuning set, particularly for developmental stages past R1 where bias towards underprediction had been observed; the median across-fold MAEs for the optimized model were 6.25 and 6.26 days to R1 and R7, respectively, and 6.88 and 15.95 days to R1 and R7 for the default model parameters (Fig. 3). For each phase, the differences in the folds' paired MAEs were statistically significant via a Wilcoxon signed-rank test; EM: d = 1.24 and p = 5e−3*; R1: d = 0.65 and P = 0.01*; R5: d = 8.08 and p = 5e−3*; R7: d = 9.29 and p = 5e−3*, where d is the mean paired difference of MAEs of the default model and the re-fit model. While the re-fit model has improved out-of-sample prediction error, it is notable that the prediction skill for R1 of the default model is performant, and that the prediction skill for

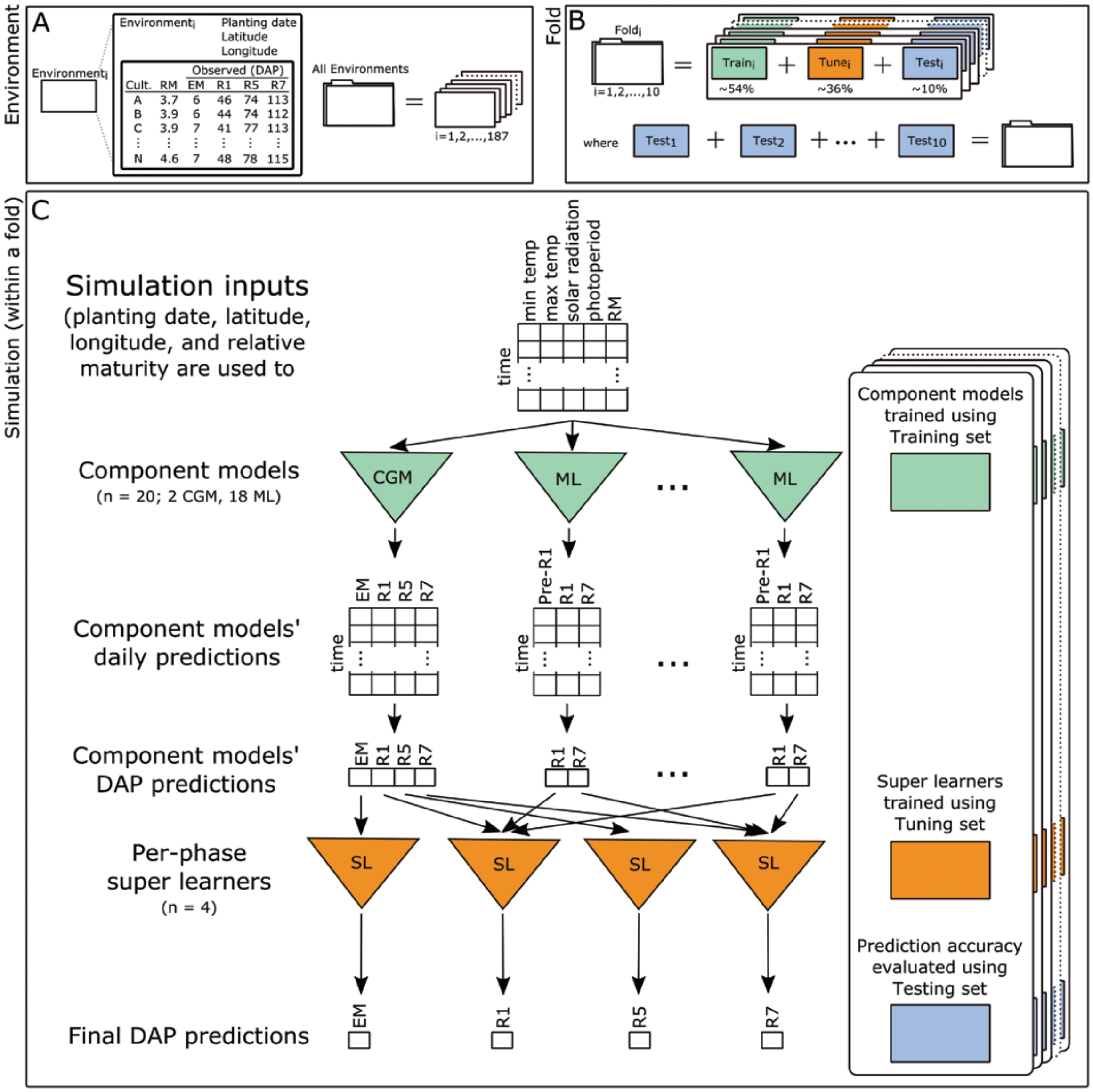
Environment
A
Environment$_i$
Environment$_i$ Planting date Latitude Longitude
Observed (DAP)
Cult. RM EM R1 R5 R7
A 3.7 6 46 74 113
B 3.9 6 44 74 112
C 3.9 7 41 77 113
N 4.6 7 48 78 115
All Environments
=
i=1,2,...,187
Fold
B
Fold$_i$
i=1,2,...,10
=
Train$_i$ + Tune$_i$ + Test$_i$
~54% ~36% ~10%
where Test$_1$ + Test$_2$ + ... + Test$_{10}$ =
Simulation (within a fold)
C
Simulation inputs (planting date, latitude, longitude, and relative maturity are used to
min temp
max temp
solar radiation
photoperiod
RM
time
Component models (n = 20; 2 CGM, 18 ML)
CGM
ML
ML
Component models' daily predictions
EM R1 R5 R7
Pre-R1 R1 R7
time
Component models' DAP predictions
Per-phase super learners (n = 4)
SL
Final DAP predictions
EM
R1
R5
R7
Component models trained using Training set
Super learners trained using Tuning set
Prediction accuracy evaluated using Testing set

Figure 2. Graphical overview of data partitioning, training procedures and processing pipeline. (A) Each environment comprises a unique latitude, longitude and planting date associated with one or more records; each record corresponds to phenological observations or traits of days after planting (DAP) to reach a developmental stage for a soybean cultivar with a known RM. (B) 10-folds are generated by building 10 mutually exclusive testing sets of environments, where each testing set comprises 10% of the entire set of environments; the remaining 90% of the environments for each fold is randomly assigned to training and tuning sets on a 60:40 basis. (C) Overview of the processing pipeline for a single record. The record's environment and its RM define the matrix of daily inputs that component models receive. Component models, trained using environments from the training set, generate daily outputs that are converted to a prediction of calendar days after planting (DAP) that a phenological phase is reached. Predictions made by component models are used as input to per-phase super learners that were trained using the tuning set. The output of the super learners represents the final predictions; prediction accuracy was evaluated using the testing set. Note that this does not depict the input of the crop growth models' daily predictions to the machine-learned models, a process that is described below.

R5 and R7 with the default model could be improved by a simple bias correction of the predictions; for example, when the mean residual of the default CROPGRO R7 predictions for the entire dataset was added to the predicted R7 values, the resulting MAE was 6.98 days [**see Supporting Information—Fig. S1**]. These observations suggest that the default model structure and parameters, while relatively

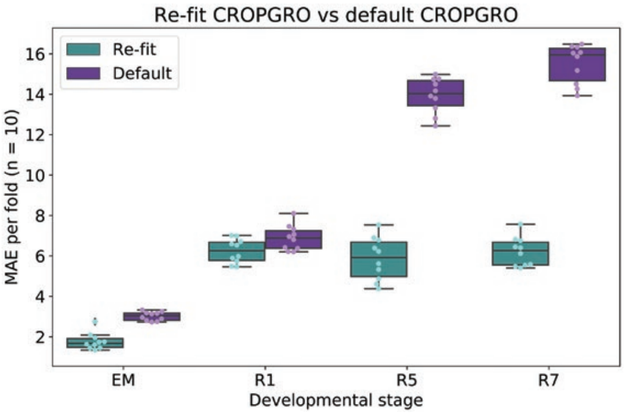

Figure 3. Comparison of out-of-sample accuracy for the re-fit CROPGRO with the default CROPGRO across 10-folds. In all cases, refitting the model using the training set improved prediction accuracy in the out-of-sample tuning set.

simple, are capable of capturing large components of observed variation in the duration of developmental phases (Figs 1 and 3).

2.3 Training component models: machine-learned models of soybean phenology using knowledge-based features

Recurrent neural network architectures such as LSTM networks have demonstrated good performance in learning temporal or sequence-dependent structure in data, and they could potentially serve as a data-driven analogue to the knowledge-based soybean phenology models. Thus, a training strategy was devised to train LSTM networks to predict daily soybean developmental stages and to identify relevant hyperparameters for network architecture. LSTM networks were trained using inputs of minimum daily temperature, maximum daily temperature, solar radiation, night length as the complement of photoperiod and RM, where RM was constant with respect to time (Fig. 2). Thus, for each day in the sequence, the daily output of the LSTM network was a set of probabilities defining which of the discrete phenological phases the plant was likely in that day.

Hyperparameter tuning of neural networks is an outstanding challenge in machine learning, often accomplished by grid searches or evolutionary algorithms. A cursory grid search over the number of nodes and layers was performed using a training set (and evaluating accuracy in a held-out subset of the training set). A set of nine different architectures (i.e. number of nodes and layers), including shallow and deep architectures, were retained. Once the architectures were defined, two separate models of each of these architectures with different random initializations were trained, generating a pool of 18 trained models of various architecture complexities within a fold. Notably, the relative performance for a given architecture varied by fold, and most architectures received non-zero weights in at least one-fold [**see Supporting Information**—Table **S2**]. This observation is consistent with the premise of super learning that, rather than the 'best' model being represented by a single model structure, utilizing a diversity of component models can lead to improved outcomes (Page 2018).

Since knowledge-based models represent a formal encapsulation of decades of research into the processes governing the system and are capable of explaining considerable variation in outcomes, their utility in machine learning was also examined. It stands to reason that providing this expert encapsulation of knowledge to a data-driven model as a feature should improve model training and generalization, so whether or not providing predictions from knowledge-based models as a feature to the data-driven model would improve performance was tested. To test this, the LSTM network architectures were trained using the aforementioned inputs only, or with the daily predictions from the default and re-fit CROPGRO model as additional features which themselves are simply functions of the same inputs.

Inclusion of the predictions made by the CROPGRO models as features to the neural network model improved the prediction accuracy of the collection of machine-learned models (Fig. 4). The CROPGRO model outputs are simply functions of the same input data the neural network already received and thus represent an expert-engineered feature. An interpretation of this result is that encapsulating prior knowledge as an engineered feature (i.e. the knowledge-based model's prediction) improves the data-driven training process to fit a more generalizable model (Fig. 4). That is, despite the theoretical ability of neural networks to faithfully represent arbitrary mapping functions, an imperfect training process identifies a poor local optimum. Expert features in the form of the knowledge-based model's prediction assist the training process to better find improved, more generalizable models. The median MAEs for models trained without CROPGRO features were 4.71 and 5.8 days for planting to R1 and planting to R7, respectively, whereas the median MAEs for models trained with CROPGRO features were 4.58 and 5.48 days for planting to R1 and planting to R7, respectively. For both phases, the differences in the folds' paired MAEs were statistically significant via a Wilcoxon signed-rank test; R1: $d = 0.13$ and $p = 1\text{e}{-3}^*$; R7: $d = 0.45$ and $p = 4\text{e}{-17}^*$, where d is the mean paired difference of MAEs of models trained without CROPGRO features and with CROPGRO features. Due to the improvement in accuracy, subsequent analyses were performed using the data-driven models that included the knowledge-based models' predictions as inputs.

Unlike the knowledge-based models, the LSTM network models were trained only to predict R1 and R7. This was performed in order to maximize the amount of data the networks could be trained on, as there were considerably fewer observations containing all stages, and the handling of missing data in the context of training a recurrent neural network is not well established. As such, the neural networks were not trained to predict emergence or R5; the knowledge-based models generate predictions for those stages in subsequent analyses.

2.4 Combining component model predictions with super learning

Given a pool of models, model selection techniques are often used to choose a single 'best' model for prediction and inference. Alternatively, since the best model for inference may not be the best model for prediction, model ensembling approaches such as meta-regression or super learning have gained popularity for prediction tasks; these approaches take advantage of model diversity to improve predictions (Breiman 1996; Naimi and Balzer 2018). The simplest ensembling

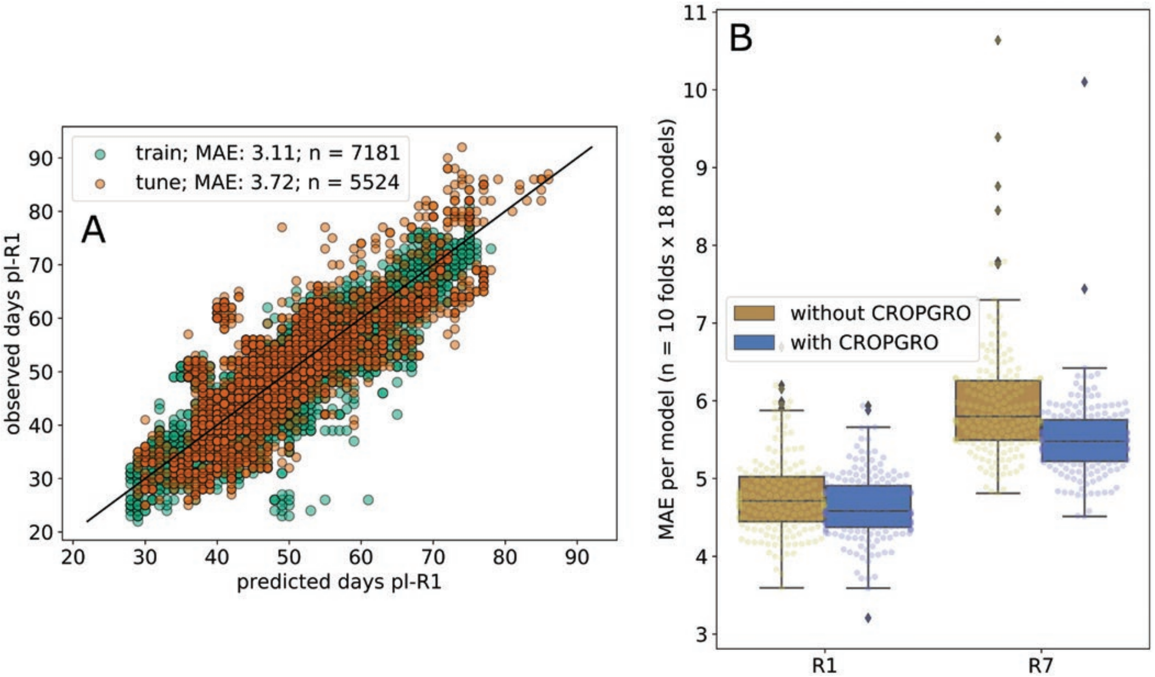

Figure 4. Data-driven approaches can learn temporal dynamics to predict phenology, and they are improved by inclusion of predictions from knowledge-based models as an expert-engineered feature. (A) Example of a well-performing model for planting to R1 of one initialization of one LSTM network architecture from 1-fold (where the out-of-sample tuning set MAE corresponds to one 'R1 with CROPGRO' point in panel B). (B) Comparison of out-of-sample performance in the tuning set for the collection of machine-learned models trained using the base environmental features alone (labelled 'without CROPGRO') or with both the default CROPGRO and optimized CROPGRO predictions as additional features (labelled 'with CROPGRO'). The median performance of the entire collection of models is improved by inclusion of the knowledge-based model predictions during training even though the knowledge-based model is simply a transform of the same environmental inputs already provided to the data-driven model.

approach is an averaging of the predictions of every model in the pool, such that each individual model's prediction is equally weighted in the final ensemble prediction. For applications in which the relative utility of individual models is of interest, BMA has emerged as a useful approach capable of weighing models, thereby providing some indication of which models are performing best while also leveraging their combined information (Hoeting *et al.* 1999; Raftery *et al.* 2005).

The pool of component models consisted of 18 data-driven and 2 knowledge-based models, all of which operated on daily time steps (Fig. 2). For each model and for each developmental stage that the model predicted, the number of calendar days between planting and the developmental stage (DAP) was obtained. All models predicted R1 and R7, whereas only the knowledge-based models predicted emergence and R5. In a conceptually similar approach to Raftery *et al.* (2005), the records from the tuning set were used to regress the observations on the predictions to obtain a bias-corrected Bayesian linear regression model, generating a pool of Bayesian linear regression models that were then ensembled and weighted via stacking of predictive distributions described by Yao *et al.* (2018).

For predictions of R1 and R7, the super learner had access to the full pool of 20 models, whereas emergence and R5 only contained the default CROPGRO and re-fit CROPGRO models in their pool (Fig. 2). Median out-of-sample MAEs for predictions into the testing set across the 10-folds

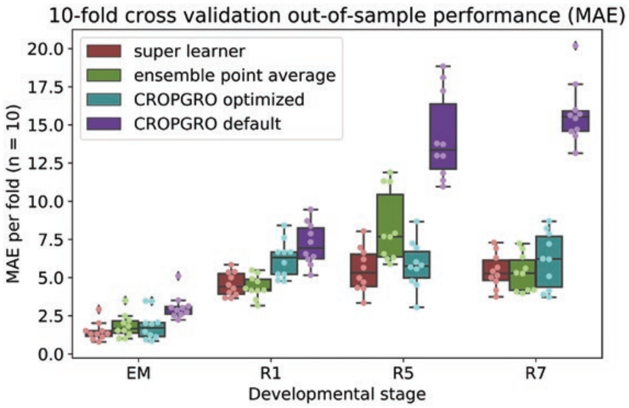

Figure 5. Comparison of out-of-sample performance across folds for the super learners, the ensemble point average (i.e. all models weighted equally), CROPGRO with optimized parameters and CROPGRO with default parameters. Ensemble models for emergence (EM) and R5 only contain the optimized and default CROPGRO, whereas ensemble models for R1 and R7 additionally contain the 18 neural network models. In this case, out-of-sample performance is evaluated on only the testing set of each fold.

were compared for (i) the super learner, (ii) the simple ensemble point average of each model's predictions (i.e. all models weighted equally), (iii) CROPGRO with optimized parameters, and (iv) CROPGRO with default parameters (Fig. 5). For each stage, the median across-fold MAEs of the final predictions from the super learners were 1.32, 4.41, 5.31 and 5.27 days for planting to emergence, to R1, to R5 and to R7, respectively.

As expected, all approaches that had been trained using a subset of the dataset had improved prediction accuracy for the out-of-sample testing set relative to CROPGRO with default parameters. In all cases, the super learners performed at parity or better than the ensemble point averages (i.e. all models' point prediction weighed equally). EM and R5 but not R1 and R7 showed statistically significant differences between the folds' paired MAEs via a two-sided Wilcoxon signed-rank test; EM: d = 0.36 and P = 0.02*; R1: d = –0.17 and P = 0.14; R5: d = 2.76 and p = 7e–3*; R7: d = –0.13 and P = 0.58, where d is the mean paired difference of MAEs of the ensemble average and the super learner across folds. Even when performing at parity, the super learners had the added benefit of identifying which models could be reasonably excluded from the ensemble. Because many of the models in the pool are given no weight by the super learner in a given fold, this procedure has the dramatic benefit of retaining the predictive accuracy of the whole ensemble but reducing the number of models that need to be maintained and their deployment footprint for production settings. Similarly, in model deployment and production settings, this also facilitates the examination of potential tradeoffs between accurate models with larger memory footprints and runtimes versus those with smaller footprints and runtimes (e.g. for deployment on a mobile device). Also of note, the significant difference for EM and R5 emphasizes the value of the weighted ensemble when poor performing component models are a large proportion of the collection; simply assigning equal weights to average the model predictions can lead to suboptimal predictions. Ultimately, the super learners resulted in a median out-of-sample MAE of 4.41 and 5.27 days to R1 and R7, respectively, providing sufficient accuracy for decision support in a number of applications.

3. DISCUSSION

Fundamentally, all phenotypes are a consequence of genotype by environment interactions by nature of the genome's integration of environmental signals. Historically, breeding progress has been driven by efforts that model genetic effects as largely independent of the environment, and this approach is particularly successful when the target population of environments displays relatively little environmental variation. However, as data generation and computational capacity continue to advance, extensions of these traditional approaches have been made that formalize the knowledge that overt phenotypes are non-linear functions of genetic states and the environment over time (Messina *et al.* 2018). This work further builds on these premises, seeking to maximize the ability to transfer information learned about how given genetics interact with environments in one geography (e.g. late maturities in Brazil) to another (e.g. the USA).

Data acquisition in the life sciences will continue to present challenging modelling tasks for the foreseeable future, where $p \gg n$ due to the complexity of the system and relative inability to experimentally sample it (where p is the number of model parameters and n is the number of observations). Some measurements can be readily obtained at low costs (e.g. marker genotypes), whereas others are more expensive and laborious (e.g. multi-environment trials, longitudinal data), generating both data-rich and data-poor settings from which actionable conclusions need to be drawn. Moreover, even in data-rich environments in which machine-learned models can perform well, there is a drive to bring improved interpretability to these models or to constrain them with prior knowledge (Marcus 2018; Hazard *et al.* 2019). The current work hybridizes the modelling strategies by training data-driven models using knowledge-based predictions as features, as well as ensembling the knowledge-based and data-driven predictions by training a meta-model; considerable work remains to create truly hybrid systems that are easily used, capable of providing system understanding and inference and identify novel components of the system (Fan *et al.* 2015; Hamilton *et al.* 2015, 2017; Karpatne *et al.* 2017*b*). These interpretable and adaptive models will enable both understanding and prescriptive decision support and may continue to develop in the form of probabilistic programming and model-based machine learning (Bishop 2013; Salvatier *et al.* 2016). Moreover, as super learners become more sophisticated, it stands to reason that they will be able to learn to use the signal processing provided by individual component models as abstractions to build hierarchical representations of the world, not unlike the human mind (Iten *et al.* 2018; Marcus 2018).

Given that the neural network models outperform the optimized CROPGRO, it suggests that our current understanding of the system may need to be adapted to incorporate aspects of the system that the data-driven approach captures, which the knowledge-based model does not. That said, that the performance improvement of the super learner over the default model is not more drastic is a testament to the body of knowledge developed regarding soybean phenology over the past decades. This body of knowledge helped to define the relevant features used as input to the data-driven model, namely, temperature and photoperiod. Additionally, it is notable that the addition of CROPGRO predictions as features for the data-driven approach enabled the training procedure to generally identify better optima than the environmental features alone across an entire collection of tested network architectures. This implies that the knowledge-based model provided an informative integration of the environmental features that the training procedure for the LSTM was otherwise unable to identify.

This work builds on a long history of research into the genetic basis of and modelling of soybean phenology and represents a practical milestone in integrating some of this information. However, the current work assumes that the genetic variation in soybean phenology can be compressed into a single scalar: the RM. While the model achieves acceptable accuracy for the target applications, additional work should examine the utility of additional genetic information, such as markers for the soybean E-loci or markers for stem growth habit to distinguish between determinate and indeterminate lines (Messina *et al.* 2006; Tian *et al.* 2010). Future work will also benefit from advances in remote-sensing efforts to accurately measure phenological events (Zeng *et al.* 2016), as well as advances in optimization that enable efficient searches over parameter spaces, hyperparameter spaces and optimization strategies for data-driven models (Li and Malik 2016).

The end application of the model is to support decisions across a wide variety of geographies by enabling growers and agronomists a forecast to plan crop management events, including planting and harvest dates, pesticide applications and irrigation events, as well as to examine how maturities planted outside of their typical zone will perform. As the world's geopolitics and climate continue to change, accurate prediction of outcomes for agricultural systems outside of the norm will be critical for rapid adaptation.

4. METHODS

4.1 Soybean phenology and weather database

The soybean phenology database used in this study was derived from public and private sources of data, with phenology stages based on Fehr and Caviness (1977). Phenology data for North America were derived predominantly from Corteva Agriscience's internal resources. Data from Argentina were obtained from publicly available soybean tests coordinated by the National Institute of Agricultural Technology (INTA). Data on soybean phenology from Brazil were derived from available technical reports. A list of the public sources of observations included in the analysis is available in **Supporting Information—Table S1**. A proprietary Corteva Agriscience™ weather repository and interpolation system was used to acquire daily historical weather observations from public and private weather stations for North and South America at the experimental coordinates.

4.2 Data harmonization and partitioning

The data were partitioned on the basis of unique environments, defined as unique latitude, longitude and planting date (defined by month, day and year) combinations to ensure that trained models were not trained using records that shared the same environment with out-of-sample records; here, a record is defined as group of phenological observations (i.e. traits) for an RM in an environment. It could be argued that different planting dates within a site-year have sufficiently correlated weather to cause information leakage between in-sample and out-of-sample records and lead to over-optimistic out-of-sample accuracy metrics, but we consider the photoperiods to be sufficiently different to merit consideration as unique environments. Some data sources recorded observations of full maturity (R8) instead of physiological maturity (R7); while factors like relative humidity, temperature, wind speed, and rainfall can introduce variation in the duration between R7 and R8 across different environments, the duration can be practicably treated as a constant number of calendar days for this application (Gaspar *et al.* 2017; Martinez-Feria *et al.* 2017). Thus, observations of R8 were harmonized to R7 prior to any analyses by subtracting a constant 9 calendar days, and the harmonized observations were considered equivalent to R7 observations. The dataset consisted of 187 unique environments with a total of 13 673 records. All 13 673 records contained planting date, flowering (R1) date and physiological maturity (R7) date. A subset of 7043 records also contained emergence date, 8212 recorded beginning seed filling (R5) date and 5488 records included all four developmental stages.

The 187 environments were partitioned into training, tuning and testing sets in 10-folds (Fig. 2). Each fold withheld a random selection of 10% of the environments as the testing set, where the environments in each fold's testing set were mutually exclusive of all other folds. The testing set of the first 9-folds contained 18 environments, and the final 10th-fold contained 25. The remaining environments in a fold were then randomly split on a 60:40 basis into a training and tuning set, with 101 environments in the training set and 68 in the tuning set (fold 10 contained 61 in the tuning set).

Individual component models in the model pool or library (i.e. the collection of models whose predictions were provided to the super learners) were trained using records from the training set. The super learners were trained using the component models' predictions and records from the tuning set. The performance of the trained super learners was evaluated using the testing set (Fig. 2).

4.3 Knowledge-based phenology model implementation and optimization

The phenology models for soybean from CROPGRO and SOYDEV were re-implemented with OpenCL and interfaced with PyOpenCL (PyOpenCL v2020.3.1 and pocl v1.5) to enable highly parallelized execution of the models (Jones *et al.* 2003; Setiyono *et al.* 2007; Stone *et al.* 2010; Klöckner *et al.* 2012). A swarm of particles, each representing a parameter vector, was used to explore the parameter space, where each particle moved down an approximated gradient using simultaneous perturbation stochastic approximation (SPSA) in parallel to find a satisfactory minimum of the cost function (Spall 1998). In brief, SPSA takes a random subset of the parameters and generates two new parameter vectors, one resulting from a small positive adjustment to the subset of parameters and one resulting from a small negative adjustment. The cost of both is evaluated, and the parameters are updated in the direction that minimizes cost. For this application, the cost was defined as the negative log-likelihood of the parameter vector given the data, where missing observations (e.g. missing R5 date) did not contribute to cost. In this manner, all observations for a record (e.g. planting to emergence, R1, R5 and R7) could be used simultaneously to fit the dynamical model.

A given parameter vector (i.e. a particle) would be updated according to SPSA using randomized minibatches over the records; once all particles finished all minibatches, the worst (i.e. highest cost or worst fit) particles were repopulated with parameters that were recombinations of the best particles' parameters. The space searched for each parameter was bounded by a defined range [**see Supporting Information**]. The search was initialized by gridding a swarm of 7200 particles across the parameter space, and performing parameter updates on minibatches of size 128. The bounds of the range a given parameter could occupy were manually assigned, and per-parameter learning rates were automatically assigned using the magnitude of the range to be searched for the parameter. Thirty rounds of evolution were performed, meaning that each of the 7200 particles completed all minibatches and had the potential to be recombined 30 times. If a training set had 7000 records, this would lead to 7200 * (7000 * 2) * 30 simulations, or 3.024 billion simulations explored during training. While this multimodal optimization strategy allows the exploration of equifinality and similarly behaving families of parameters, the parameter vector with the minimum cost found at the end of any round of evolution was retained and used as the set of optimized parameters for subsequent analyses. This optimization was performed on AWS c5.18×large instances, requiring around 30 h of wall clock time per fold on the 72 vCPUs (first-generation Intel Xeon Platinum 8000).

All Wilcoxon signed-rank tests to compare performance were performed using the stats module of Scipy v1.2.1 (Virtanen *et al.* 2020).

4.4 Machine-learned models and super learners

All machine-learned models were varying architectures of LSTM networks trained using Keras (v2.2.4) and Tensorflow (v1.13.1) (Chollet *et al.* 2015; Abadi *et al.* 2016). Models were trained using daily inputs of minimum temperature, maximum temperature, night length, solar radiation, RM group (which was constant with respect to time) and an indicator variable to indicate whether planting had occurred or not to predict the developmental stage reached that day (as an integer). Model architectures ranged from shallow architectures with one hidden layer of 64 nodes, to deeper architectures with two or three hidden layers and up to 2024 nodes using a tanh activation function [**see Supporting Information—Table S2**]; the output layer used a softmax activation function. Categorical cross entropy was used as the loss function, models were trained using Adam as the optimization strategy for up to 50 epochs with early stopping and the best performing model on the tuning set out of the epochs was retained for each architecture, generating a population of 18 models for each fold. This model fitting was performed on AWS p2.xlarge instances, requiring up to roughly 2 h of wall clock time for the largest network architectures using an NVIDIA Tesla K80 GPU; all 18 models in 1-fold could be trained on one GPU in less than 24 h.

BMA as implemented in PyMC3 (v3.6 with Theano 1.0.4) was used to integrate the predictions of the component models (Raftery *et al.* 2005; Salvatier *et al.* 2016; Yao *et al.* 2018). First, for each component model, a bias-corrected linear model for each developmental stage was fit using the number of calendar days from planting to the developmental stage derived from the model's predictions as input and the observed calendar days as output using the tuning set to estimate the posterior (Raftery *et al.* 2005). For emergence and R5, this involved only the default and optimized CROPGRO models; for R1 and R7 this involved those two models and the additional 18 trained neural network architectures. Stacking of predictive distributions given the population of these bias-corrected models was performed for each developmental stage to find an optimal weighing of component models for each stage (Yao *et al.* 2018). That is, different stages have different subsets of models contributing to the final predictions. Final performance was evaluated on the testing set (Fig. 2).

SUPPORTING INFORMATION

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

Table S1. Public data sources from which soybean phenological data were obtained.

Table S2. Weights assigned by the super learner for each model averaged across all folds for the planting to R1 super learner and the planting to R7 super learner. Machine-learned model architectures are labelled by the <number of hidden layers>x<number of nodes>_round-<number model initialization>. For example, 3x128_round-0 indicates the first of two initializations of a model with three fully connected hidden layers each of 128 nodes. 2x256-128_round-1 indicates the second of two initializations of a model with 2 fully connected hidden layers, the first with 256 nodes and the second with 128 (these labels do not include the implicit input and output layers). Most models received a non-zero weighting in at least one fold. Stochasticity in the training process and variation in the folds led to different weight assignments in each fold, and most models within a given single fold received a small or zero weighting.

Figure S1. Bias-adjusted predictions from default CROPGRO for days from planting to R7. When adding the mean residual calculated from the entire dataset (15.22 days) to the default CROPGRO prediction, the MAE is reduced to 6.98 days. While this bias correction represents a fit to the data and thus the error does not represent out-of-sample prediction, it demonstrates that the default CROPGRO is capturing relevant GxE behaviour of the system; this is a relevant aspect for its suitability as a machine-learning feature. Black line represents the 1:1 line between observed and predicted.

Supplemental Methods Figure 1. Thermal (A) and photoperiod (B) functions used to determine how many PTD would be accumulated on a given day.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Marie Dolleman, Scott Birkett, Kirk Hansen, Riley McDowell and Duke Takle for their technical contributions in application hosting and productionization.

AUTHOR CONTRIBUTION

All authors contributed to conceiving the research, interpreting the results, and writing and editing of the manuscript. JR, APG, and DK performed data collection and curation. RFM and SKT performed the analyses.

SOURCES OF FUNDING

No public funding sources were involved in the preparation of this manuscript.

CONFLICT OF INTEREST

R. M., S. T., J. R., A. G., D. K. and C. M. were employees of Corteva Agriscience™ when this manuscript was written. F. v. E. was engaged in a scientific collaboration with Corteva Agriscience™ when this manuscript was written.

LITERATURE CITED

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, et al. 2016. TensorFlow: a system for large-scale machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association, pp. 265–83.

Andrés F, Coupland G. 2012. The genetic basis of flowering responses to seasonal cues. *Nature Reviews Genetics* **13**:627–639.

Archontoulis SV, Miguez FE, Moore KJ. 2014. A methodology and an optimization tool to calibrate phenology of short-day species included in the APSIM PLANT model: application to soybean. *Environmental Modelling and Software* **62**:465–477.

Bishop CM. 2013. Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**:20120222.

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

Breiman L. 1996. Stacked regressions. *Machine Learning* **24**:49–64.

Brown DM. 1960. Soybean ecology. I. Development-temperature relationships from controlled environment studies. *Agronomy Journal* **52**:493–496.

Cao D, Takeshima R, Zhao C, Liu B, Jun A, Kong F. 2017. Molecular mechanisms of flowering under long days and stem growth habit in soybean. *Journal of Experimental Botany* **68**:1873–1884.

Chollet F *et al.* 2015. *Keras*. https://keras.io (14 January 2020).

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

dos Santos C, Salmerón M, Purcell LC. 2019. Soybean phenology prediction tool for the US midsouth. *Agricultural and Environmental Letters* **4**:190036.

Elizondo DA, McClendon RW, Hoogenboom G. 1994. Neural network models for predicting flowering and physiological maturity of soybean. *Transactions of the ASAE* **37**:981–988.

Fan X-R, Kang M-Z, Heuvelink E, de Reffye P, Hu B-G. 2015. A knowledge-and-data-driven modeling approach for simulating plant growth: A case study on tomato growth. *Ecological Modelling* **312**:363–373.

FAOSTAT. 2016. Food and Agriculture Organization of the United Nations (FAO). FAOSTAT Database. http://faostat.fao.org/site/291/default.aspx

Fehr W, Caviness C. 1977. Stages of soybean development. *Special Report*.

Gaspar AP, Laboski CAM, Naeve SL, Conley SP. 2017. Dry matter and nitrogen uptake, partitioning, and removal across a wide range of soybean seed yield levels. *Crop Science* **57**:2170–2182.

Grimm SS, Jones JW, Boote KJ, Hesketh JD. 1993. Parameter estimation for predicting flowering date of soybean cultivars. *Crop Science* **33**:137–144.

Hamilton F, Berry T, Sauer T. 2015. Predicting chaotic time series with a partial model. *Physical Review E* **92**:010902.

Hamilton F, Lloyd AL, Flores KB. 2017. Hybrid modeling and prediction of dynamical systems. *PLoS Computational Biology* **13**:e1005655.

Hazard CJ, Fusting C, Resnick M, Auerbach M, Meehan M, Korobov V. 2019. Natively interpretable machine learning and artificial intelligence: preliminary results and future directions. *arXiv*. **arXiv**:1901.00246v2.

He J, Jones JW, Graham WD, Dukes MD. 2010. Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method. *Agricultural Systems* **103**:256–264.

Hesketh JD, Myhre DL, Willey CR. 1973. Temperature control of time intervals between vegetative and reproductive events in soybeans. *Crop Science* **13**:250–254.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Computation* **9**:1735–1780.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian model averaging: a tutorial. *Statistical Science* **14**:382–401.

Iten R, Metger T, Wilming H, del Rio L, Renner R. 2018. Discovering physical concepts with neural networks. *arXiv*. **arXiv**:1807.10300v2.

Jones JW, Hoogenboom G, Porter CH, Boote KJ, Batchelor WD, Hunt LA, Wilkens PW, et al. 2003. The DSSAT cropping system model. *European Journal of Agronomy, Modelling Cropping Systems: Science, Software and Applications* **18**:235–265.

Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S et al. 2017a. Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* **29**:2318–2331.

Karpatne A, Watkins W, Read J, Kumar V. 2017b. Physics-guided neural networks (PGNN): an application in lake temperature modeling. *arXiv* **arXiv**:1710.11431v2.

Kennedy J. 2010. Particle swarm optimization. In: Sammut C, Webb GI, eds. *Encyclopedia of machine learning*. Boston, MA: Springer US, 760–766.

Klöckner A, Pinto N, Lee Y, Catanzaro B, Ivanov P, Fasih A. 2012. PyCUDA and PyOpenCL: A scripting-based approach to GPU run-time code generation. *Parallel Computing* **38**(3):157–174.

Lamsal A, Welch SM, Jones JW, Boote KJ, Asebedo A, Crain J, Wang X, et al. 2017. Efficient crop model parameter estimation and site characterization using large breeding trial data sets. *Agricultural Systems* **157**:170–184.

Li K, Malik J. 2016. Learning to optimize. *arXiv*. **arXiv**:1606.01885v1.

Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. 2018. Machine learning in agriculture: a review. *Sensors* **18**(8):2674.

Marcus G. 2018. Deep learning: a critical appraisal. *arXiv*. **arXiv**:1801.00631.

Martinez-Feria R, Archontoulis SV, Licht MA. 2017. How fast do soybeans dry down in the field? https://crops.extension.iastate.edu/cropnews/2017/09/how-fast-do-soybeans-dry-down-field (14 January 2020).

Messina CD, Jones JW, Boote KJ, Vallejos CE. 2006. A gene-based model to simulate soybean development and yield responses to environment. *Crop Science* **46**(1):456–466.

Messina CD, Technow F, Tang T, Totir R, Gho C, Cooper M. 2018. Leveraging biological insight and environmental variation to improve phenotypic prediction: integrating crop growth models (CGM) with whole genome prediction (WGP). *European Journal of Agronomy* **100**:151–162.

Naimi AI, Balzer LB. 2018. Stacked generalization: an introduction to super learning. *European Journal of Epidemiology* **33**: 459–464.

Onogi A, Watanabe M, Mochizuki T, Hayashi T, Nakagawa H, Hasegawa T, Iwata H. 2016. Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. *Theoretical and Applied Genetics* **129**: 805–817.

Oyetunde T, Bao FS, Chen J-W, Martin HG, Tang YJ. 2018. Leveraging knowledge engineering and machine learning for microbial biomanufacturing. *Biotechnology Advances* **36**:1308–1315.

Page S. 2018. *The model thinker: What you need to know to make data work for you*. UK: Hachette.

Pathak J, Wikner A, Fussell R, Chandra S, Hunt BR, Girvan M, Ott E. 2018. Hybrid forecasting of chaotic processes: using machine learning in conjunction with a knowledge-based model. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28**: 041101.

Polley E, van der Laan M. 2010. Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Prusinkiewicz P. 2004. Modeling plant growth and development. *Current Opinion in Plant Biology* **7**:79–83.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**:1155–1174.

Roberts MJ, Braun NO, Sinclair TR, Lobell DB, Schlenker W. 2017. Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters* **12**:095010.

Salmerón M, Purcell LC. 2016. Simplifying the prediction of phenology with the DSSAT-CROPGRO-soybean model based on relative maturity group and determinacy. *Agricultural Systems* **148**:178–187.

Salvatier J, Wiecki TV, Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2**:e55.

Setiyono TD, Weiss A, Specht J, Bastidas AM, Cassman KG, Dobermann A. 2007. Understanding and modeling the effect of temperature and daylength on soybean phenology under high-yield conditions. *Field Crops Research* **100**:257–271.

Sexton J, Everingham Y, Inman-Bamber G. 2016. A theoretical and real world evaluation of two Bayesian techniques for the calibration of variety parameters in a sugarcane crop model. *Environmental Modelling and Software* **83**:126–142.

Shakoor N, Northrup D, Murray S, Mockler TC. 2019. Big data driven agriculture: big data analytics in plant breeding, genomics, and the use of remote sensing technologies to advance crop productivity. *The Plant Phenome Journal* **2**:1–8.

Shaykewich CF. 1995. An appraisal of cereal crop phenology modeling. *Canadian Journal of Plant Science* **75**:329–341.

Shaykewich CF, Bullock PR. 2018. Modeling soybean phenology. *Agroclimatology: Linking Agriculture to Climate*, agronomymonogra/agronmonogr60.

Sinclair TR. 1986. Water and nitrogen limitations in soybean grain production I. Model development. *Field Crops Research* **15**:125–141.

Spall JC. 1998. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems* **34**:817–823.

Stone JE, Gohara D, Shi G. 2010. OpenCL: a parallel programming standard for heterogeneous computing systems. *Computing in Science Engineering* **12**:66–73.

Taghavi Namin S, Esmaeilzadeh M, Najafi M, Brown TB, Borevitz JO. 2018. Deep phenotyping: deep learning for temporal phenotype/genotype classification. *Plant Methods* **14**:66.

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

Tian Z, Wang X, Lee R, Li Y, Specht JE, Nelson RL, McClean PE, et al. 2010. Artificial selection for determinate growth habit in soybean. *Proceedings of the National Academy of Sciences* **107**: 8563–8568.

van Eeuwijk FA, Bustos-Korts DV, Malosetti M. 2016. What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Science* **56**: 2119–2140.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**:261–272.

Wallach D, Goffinet B, Bergez J-E, Debaeke P, Leenhardt D, Aubertot J-N. 2001. Parameter estimation for crop models. *Agronomy Journal* **93**:757–766.

Wang J, McBlain BA, Hesketh JD, Woolley JT, Bernard RL. 1987. A data base for predicting soybean phenology. *Biotronics* **16**:25–38.

Whitley D. 1994. A genetic algorithm tutorial. *Statistics and Computing* **4**:65–85.

Wong K-C. 2015. Evolutionary multimodal optimization: a short survey. *arXiv*. **arXiv**:1508.00457v1.

Yao Y, Vehtari A, Simpson D, Gelman A. 2018. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis* **13**:917–1007.

Zeng L, Wardlow BD, Wang R, Shan J, Tadesse T, Hayes MJ, Li D. 2016. A hybrid approach for detecting corn and soybean phenology with time-series MODIS data. *Remote Sensing of Environment* **181**:237–250.

Zhang J-Q, Zhang L-X, Zhang M-H, Watson C. 2009. Prediction of soybean growth and development using artificial neural network and statistical models. *Acta Agronomica Sinica* **35**:341–347.