in silico Plants Vol. 4, No. 2, pp. 1–16
https://doi.org/10.1093/insilicoplants/diac017
Advance Access publication 11 August 2022
Original Research

in silico Plants
OXFORD

Integration of machine learning into process-based modelling to improve simulation of complex crop responses

Ioannis Droutsas[1,2,*], Andrew J. Challinor[1,2], Chetan R. Deva[1] and Enli Wang[3]

[1]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK
[2]Priestley International Centre for Climate, University of Leeds, Leeds LS2 9JT, UK
[3]CSIRO Agriculture and Food, Canberra, ACT 2601, Australia
*Corresponding author's e-mail address: earid@leeds.ac.uk

Handling Editor: Graeme Hammer

Citation: Droutsas I, Challinor AJ, Deva CR, Wang E. 2022. Integration of machine learning into process-based modelling to improve simulation of complex crop responses. *In Silico Plants* **2022**: diac017; doi: 10.1093/insilicoplants/diac017

ABSTRACT

Machine learning (ML) is the most advanced field of predictive modelling and incorporating it into process-based crop modelling is a highly promising avenue for accurate predictions of plant growth, development and yield. Here, we embed ML algorithms into a process-based crop model. ML is used within GLAM-Parti for daily predictions of radiation use efficiency, the rate of change of harvest index and the days to anthesis and maturity. The GLAM-Parti-ML framework exhibited high skill for wheat growth and development in a wide range of temperature, solar radiation and atmospheric humidity conditions, including various levels of heat stress. The model exhibited less than 20 % error in simulating the above-ground biomass, grain yield and the days to anthesis and maturity of three wheat cultivars in six countries (USA, Mexico, Egypt, India, the Sudan and Bangladesh). Moreover, GLAM-Parti reproduced around three-quarters of the observed variance in wheat biomass and yield. Existing process-based crop models rely on empirical stress factors to limit growth potential in simulations of crop response to unfavourable environmental conditions. The incorporation of ML into GLAM-Parti eliminated all stress factors under high-temperature environments and reduced the physiological model parameters down to four. We conclude that the combination of process-based crop modelling with the predictive capacity of ML makes GLAM-Parti a highly promising framework for the next generation of crop models.

KEYWORDS: Crop model; GLAM-Parti; heat stress; machine learning; model development; SEMAC.

1. INTRODUCTION

Climate change brings higher frequency and intensity of drought and heat extremes and there has been particular focus on evaluating the representation of these stresses into the crop models. Studies have conducted model inter-comparisons with the aim to reveal the best-performing methods, algorithms or models to simulate the impact of high temperature and/or limited water availability on plant growth, development and productivity (e.g. Asseng *et al.* 2013; Eitzinger *et al.* 2013; Rosenzweig *et al.* 2013; Bassu *et al.* 2014; Fleisher *et al.* 2017; Maiorano *et al.* 2017; Müller *et al.* 2017). However, the high complexity of weather/plant interactions and the diverse methodologies used to build and calibrate the crop models do not easily allow the selection of 'optimal' methods or models. Asseng *et al.* (2015) and Wallach *et al.* (2018) showed that in their model inter-comparison studies, the way to optimize the model prediction performance was by considering an ensemble instead of relying on any particular crop model.

Major factors affecting the skill of crop models include the choice of equations and the way that they are structured in the prediction algorithm (Challinor *et al.* 2009). Equations that fail to represent the modelled processes and/or inconsistencies in the model structure are common problems that limit performance considerably (Martre *et al.* 2015). This is particularly true in the representation of crop growth and development under environmental stress conditions, where the complexity of the modelled system increases. For instance, process-based

crop models use stress factors to represent abiotic stresses that limit growth and modify phenological development (Ewert *et al.* 2015). The choice of the threshold above which these factors are activated and the way that these factors interact in respect to multi-stress environments are not always clear and modellers may rely on calibration of certain parameters or switch-off optional subroutines to improve the model fit to the observations (e.g. Seidel *et al.* 2018; Wallach *et al.*,2021b). As a result, the internal model consistency may be disrupted and the evaluation of processes that are most limiting for the model skill becomes a difficult task, thus impeding further model development.

Recently, machine learning (ML) methods have been applied for crop yield prediction to overcome the limitations of process-based crop models. Various studies have shown the potential of ML for crop yield prediction (e.g. Crane-Droesch 2018; Cai *et al.* 2019; Leng and Hall 2020; Shahhosseini *et al.* 2020; Newman and Furbank 2021; Lischeid *et al.* 2022), but also their limitations. A main obstacle is that the ML algorithms do not usually incorporate an understanding of the processes that lead to the observed yield level, but instead relate generic growth indices (e.g. monthly/seasonal/annual average temperature, rainfall, solar radiation, soil moisture, etc.) to crop productivity (e.g. Jeong *et al.* 2016; Folberth *et al.* 2019). Thus, the time evolution of the crop response to the environment is not fully represented and this can be limiting for the algorithm's performance and suitability.

The combination of process-based crop modelling with ML can potentially optimize the model performance by taking advantage of the process understanding of the former approach and the predictive power of the latter. Feng *et al.* (2019) incorporated output variables from the APSIM (Agricultural Production Systems sIMulator) crop model into Random Forests (RF) and showed that the hybrid model (crop model + RF) advanced by 33 % in explaining the variance in Australian wheat yield compared to the baseline crop model. Similarly, Shahhosseini *et al.* (2021) applied the APSIM crop model for prediction of US maize productivity and used output variables as inputs into ML. The study showed that the prediction error of maize yield can decrease by up to 20 % in comparison with a baseline ML model with no process-based crop model features in it.

The above modelling studies introduce output variables from process-based crop models as inputs into ML algorithms, which are then used for crop yield prediction. A more integrated approach is the incorporation of ML into a process-based crop model. Here, we embed ML algorithms into the process-based crop model, GLAM-Parti (Fig. 1). We use ML to estimate variables that regularly escape the crop model's predictive capacity with traditional methods. ML is applied for the prediction of radiation use efficiency (RUE) and the rate of change of harvest index (dHI/dt) in daily time step, as well as the days to anthesis and maturity. The aim is to create a new crop modelling/ML framework with high performance in the representation of crop response to a wide range of environments, including stress conditions. The derived crop model does not use stress factors for the simulation of abiotic stresses and the embedded ML algorithms reduce the physiological model parameters down to four. Ultimately, the yield predictions are derived from the process-based crop model, thus incorporating the understanding of the within-season crop/environment interactions.

2. MATERIALS AND METHODS

2.1 Data sets

For the assessment of our framework, we initially used data from the 'Hot Serial Cereal Experiment' for wheat (HSC) (Martre *et al.* 2018).

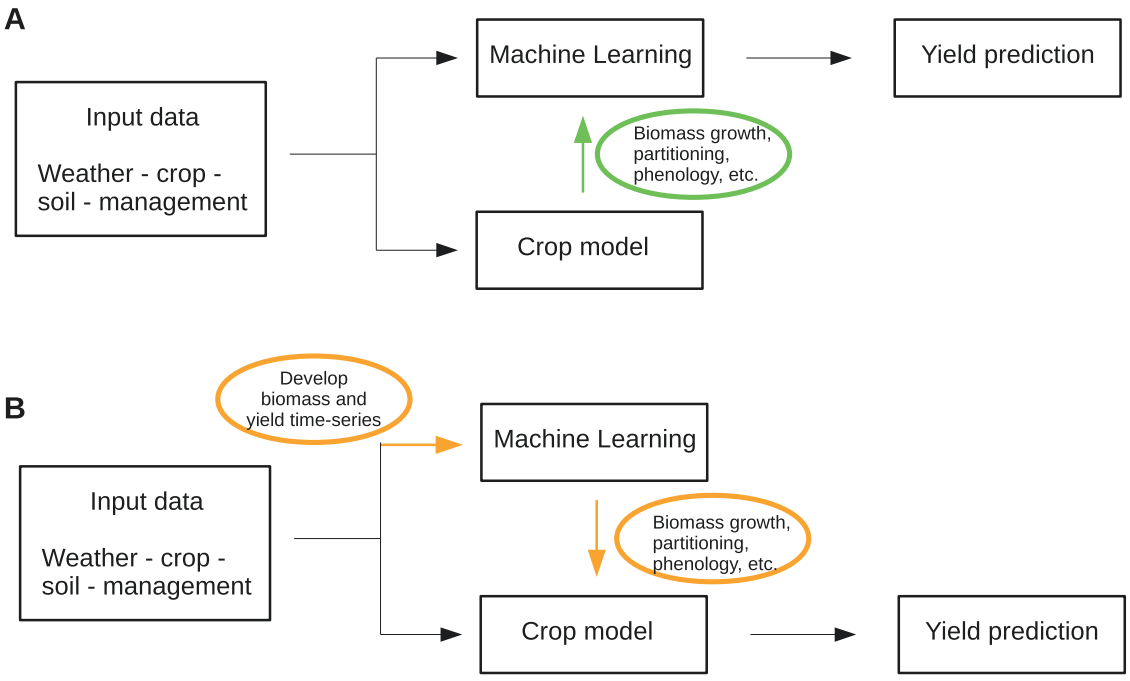

Figure 1. (A) Crop modelling framework developed and applied in Shahhosseini *et al.* (2021); (B) GLAM-Parti-ML framework developed in this study. Ellipses represent transfer of information and respective arrows show direction across which information is transferred.

The data set was chosen since it is designed for crop model development and evaluation studies. HSC contains various experiments with spring wheat (cv. Yecora Rojo) grown in Maricopa, AZ, USA at regular time intervals (every 6 weeks) for a 2-year period (2007–09). The crop experienced a wide range of temperatures (−2 to 42 °C), solar radiation and atmospheric humidity conditions, including exposure to heat stress. As a result, wheat yield varied from as low as 1.07 t ha^{-1} to as high as 8 t ha^{-1} (excluding the experiments where the crop was terminated early due to extreme weather conditions).

Subsequently, we introduced a second data set, the International Heat Stress Genotype Experiment (IHSGE), to facilitate model evaluation on a larger spatial domain. IHSGE contains 28 field experiments with two spring wheat cultivars (cv. Bacanora 88 and Nesser) grown in five low-latitude countries (Martre *et al.* 2017). The experiments were conducted in Mexico (two locations: Ciudad Obregon and Tlaltizapan), Egypt (Aswan), India (Dharwar), the Sudan (Wad Medani) and Bangladesh (Dinajpur). All locations are considered as 'hot' or 'very hot' and the grain yields ranged from 1.91 to 6.3 t ha^{-1} (Martre *et al.* 2017).

The weather variables of this study are solar radiation (Srad, MJ m^{-2} day^{-1}), minimum, maximum and dew point temperature (T_{min}, T_{max} and T_{dew}, respectively, °C). All variables are provided in daily time step as part of the experimental data sets (Martre *et al.* 2017, 2018). T_{min}, T_{max} and T_{dew} are used for the calculation of vapour pressure deficit (VPD) according to the following formula (Castellvi *et al.* 1996):

$$\mathrm{VPD} = \frac{e^o(T_{\max}) + e^o(T_{\min})}{2} - e^o(T_{\mathrm{dew}}) \tag{1}$$

where $e^o(T)$ is saturation vapour pressure (kPa), calculated as follows:

$$e^o(T) = 0.61078 \cdot e^{17.27T/(T+237.3)} \tag{2}$$

We also computed accumulated solar radiation (Solrac) and thermal time (TT) as follows:

$$\mathrm{Solrac} = \sum_{i=1}^{n} \mathrm{Srad}_i \tag{3}$$

$$\mathrm{TT} = \sum_{i=1}^{n} \frac{T_{\min_i} + T_{\max_i}}{2} \tag{4}$$

where n is number of days after crop emergence.

2.2 Crop model

We slightly modified the GLAM-Parti crop model developed in Droutsas *et al.* (2019) and Droutsas *et al.* (2020) based on the SEMAC (Simultaneous Equation Modelling for Annual Crops) approach. The details of the model are described below:

The above-ground biomass (W) of wheat is separated into mass of leaves (M_L), stems (M_S) and ears, which are further divided into chaffs (i.e. the non-edible part of ear, (M_C)) and grains (M_G), such that:

$$W = M_L + M_S + M_C + M_G \tag{5}$$

M_G relates to W under the harvest index (HI) approach:

$$M_G = \mathrm{HI} \cdot W \tag{6}$$

Since wheat leaves and chaffs are the main contributors to canopy photosynthesis and growth (Merah and Monneveux 2015; Zhang *et al.* 2020), we consider them as photosynthetic organs (M_P) as follows:

$$M_P = M_L + M_C \tag{7}$$

We relate M_S to M_P under the allometric formula:

$$M_S = a \cdot {M_P}^b \tag{8}$$

where a and b are empirical coefficients estimated using linear regression between log-transformed M_S vs. M_P [**see Supporting Information—Fig. S1**].

We re-write Equation (5) using Equations (6–8) as follows:

$$W = \left(\frac{1}{1-\mathrm{HI}}\right)(M_P + a \cdot {M_P}^b) \tag{9}$$

The growth of biomass (dW/dt) is determined in daily time step under the RUE approach:

$$dW/dt = I_o \cdot \mathrm{RUE} \cdot (1 - e^{-k \cdot \mathrm{SLA} \cdot M_P}) \tag{10}$$

where I_o is incident solar radiation (MJ m^{-2} day^{-1} of photosynthetically active radiation (PAR)), RUE is radiation use efficiency (g MJ^{-1} PAR), k is light extinction coefficient and SLA is canopy specific leaf area. The parameter k for wheat is set to 0.7 (Ratjen and Kage 2016; Wang *et al.* 2017) and SLA is 180 cm^2 g^{-1} (Ratjen and Kage 2013; Ratjen *et al.* 2016).

For any given n day after crop emergence, we apply a simple mass balance equation as follows:

$$W_n - dW/dt - W_{n-1} = 0 \tag{11}$$

and incorporate Equations (9) and (10) into Equation (11) to get:

$$\left(\frac{1}{1-\mathrm{HI}}\right) \cdot (M_P + a \cdot {M_P}^b) - I_o \cdot \mathrm{RUE} \cdot (1 - e^{-k \cdot \mathrm{SLA} \cdot M_P}) - W_{n-1} = 0 \tag{12}$$

Equation (12) is a function of M_P, solved in daily time step from crop emergence to maturity. We implement the Newton–Raphson approach to find the root numerically by iteration. The method stops when convergence is achieved and the distance from the optimal solution is less than 10^{-2} t ha^{-1}. The value of M_P is then used to update the variables M_G, M_S and W in Equations (6), (8) and (9), respectively.

2.3 Integration of ML into GLAM-Parti

In process-based crop modelling, the simulated processes can be organized into three broad categories: those that describe crop growth, processes that are involved in phenological development and biomass partitioning-related procedures. If a crop model accurately describes all processes that fall into these three categories across the course of the crop growing season, then skilful biomass productivity and grain yield predictions are expected at end-of-season output. In accordance with

this, our framework incorporates ML into GLAM-Parti to achieve high performance in the prediction of the following:

- RUE: the most representative variable affecting growth in process-based crop models (e.g. Jamieson *et al.* 1998; Jones *et al.* 2003; Boote *et al.* 2013; Brown *et al.* 2019).
- *dHI/dt*: commonly used measure to describe the partitioning of above-ground biomass to the grains (e.g. Challinor *et al.* 2005; Soltani *et al.* 2005; Ramirez-Villegas *et al.* 2017).
- Days to anthesis and maturity: the most significant stages of crop phenological development and central to process-based crop models, since they signify the appearance, growth and maximum weight of grains (e.g. Wang and Engel 1998; Hussain *et al.* 2018; Ceglar *et al.* 2019).

2.3.1 ML algorithms. The ML models used in our framework are RF and Extreme Gradient Boosting (XGBoost). These are well-established, state-of-the-art ML methods which have been previously developed, discussed in detail and used in the literature. Briefly, both RF (Breiman 2001) and XGBoost (Chen and Guestrin 2016) are tree-based algorithms, which create an ensemble of Classification and Regression Trees (CART) and make predictions after considering the output of all individual trees. RF uses bagging, a technique which builds each tree independently based on a random subset of the data set. XGBoost builds trees sequentially in a dependent manner with the use of weak tree learners. The algorithm applies gradient boosting, an iterative process by which each new tree learns and improves upon mistakes of previous trees. Both algorithms can be used for classification as well as regression problems.

2.3.2 Time series of biomass and yield. The incorporation of ML into GLAM-Parti was performed according to the steps shown in Fig. 2. Initially, we used the flexible sigmoid function of Yin *et al.* (2003) to produce daily time series of above-ground biomass (W) and grain yield (Y) in the experiments as follows:

$$W = W_{max}\left(1 + \frac{t_e - t}{t_e - t_m}\right)\left(\frac{t}{t_e}\right)^{\frac{t_e}{t_e - t_m}} \quad (13)$$

where t is time after emergence (in days), W_{max} is above-ground biomass at crop maturity, t_m is time of maximum growth rate and t_e is total number of days from emergence to maturity. W_{max} and t_e are reported in the experiments, whilst t_m is the day of maximum growth rate, which is defined as:

$$t_m = f \cdot t_e \quad (14)$$

where f is a multiplication factor within the (0, 1) range (since $0 < t_m < t_e$). For the estimation of f, we appended values with iteration (0.05 time step) within the acceptable limits and developed time series of biomass with Equation (13). For each crop treatment, we selected the value of f that minimized root mean squared error (RMSE) between observed and simulated biomass. A graphical example of the f optimization is given in **Supporting Information—Fig. S2**.

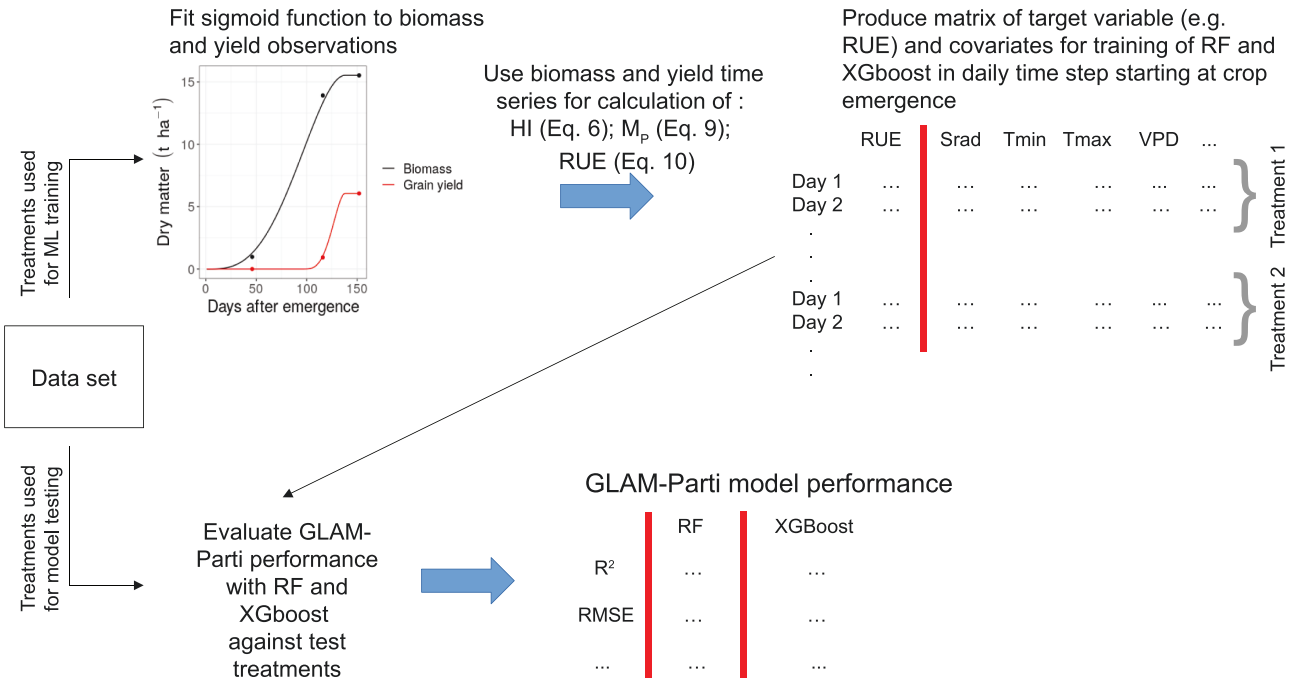

Figure 2. Methodology for integration of ML into GLAM-Parti. The data set is split into training and testing treatments. Crop data from the training treatments are used for fitting time series of biomass and yield, which then derive the target variables RUE and *dHI/dt* for training of RF and XGBoost. The test treatments are used in the evaluation of GLAM-Parti with RF and XGBoost, respectively.

The same sigmoid function was applied for the determination of the time series of yield (Equation (13)). W_{max} was replaced by Y_{max} (grain yield at crop maturity), t is now time after anthesis (in days) and t_e is total number of days from anthesis to maturity. For each crop treatment, the optimal value of f (Equation (14)) was selected to minimize RMSE between observed and simulated yield.

2.3.3 Feature selection and derivation of target variables. The weather inputs are the following variables: T_{min}, T_{max}, VPD, TT, Srad and Solrac. Since all treatments of this study were fully irrigated and fertilized, we did not consider features of soil water and nutrient content or precipitation in ML. All details of the feature selection and the derivation of the target variables are described below and summarized in Table 1.

2.3.3.1 RUE. The fitted time series of above-ground biomass and grain yield (Equation (13)) allowed the calculation of HI and M_p with Equations (6) and (9), respectively. Subsequently, Equation (10) derived the response variable, RUE, in daily time step from crop emergence until maturity was reached.

Temperature, solar radiation and VPD are the most significant weather determinants of wheat growth for irrigated, well-fertilized conditions (Zhao *et al.* 2017; Ferrante and Mariani 2018; Rashid *et al.* 2018). Consequently, the weather drivers of RUE are Srad, T_{min}, T_{max} and VPD. In addition, the rate of crop growth varies with the phenological stage (e.g. the rate of increase in biomass declines with the onset and progression of plant senescence) and the amount of photosynthetic organs (proxy for leaf and chaff area) defines the level of crop growth. Thus, we also added the ratio of photosynthetic organ mass to total above-ground biomass ($MP_{ratio} = M_p/W$) for the estimation of RUE in daily time step.

2.3.3.2 Days to anthesis and maturity. For the prediction of the number of days to anthesis and maturity, we derive the target variable iphen which consists of three classes:

0—crop is in pre-anthesis stage (starting at crop emergence)
1—crop is between anthesis and maturity
2—crop has been harvested

The main weather drivers of plant phenological development are temperature, photoperiod and solar radiation, which have been extensively used in crop modelling for predicting the days to anthesis and maturity (e.g. Challinor *et al.* 2004; Craufurd and Wheeler 2009; Ottman *et al.* 2013; Wang *et al.* 2017; Baumont *et al.* 2019). Here, photoperiod and vernalization were disregarded, since the wheat cultivars of this study are not sensitive to them (Asseng *et al.* 2015; Martre *et al.* 2017). Daily temperature is considered in terms of T_{min}, T_{max} and TT and solar radiation is considered in terms of Srad and Solrac. Foulkes *et al.* (2011) showed that phenology relates to partitioning and there is a lower threshold of stem:above-ground biomass (MS_{ratio}) before the wheat crop reaches anthesis (around 0.48 in modern cultivars). Consequently, we also included the variable MS_{ratio} into the list of predictors for iphen.

The consideration of all predictors only for the period from crop emergence to maturity would create a highly imbalanced data set, since class 2 of iphen would be represented by only one data point (i.e. the day that the crop reaches physiological maturity). In order to ensure enough observations of class 2, we use equal number of anthesis-to-harvest (class 1) and post-harvest (class 2) data points in the training of ML. For instance, treatment 14C reached maturity 44 days after anthesis; thus, the corresponding inputs start at crop emergence until 44 days post-harvest.

2.3.3.3 Rate of change of harvest index. The time course of HI exhibits three distinct phases: an initial lag phase directly after anthesis where low grain dry matter accumulation occurs, which progresses into a linear phase—where most of the grain growth happens—followed by a maturation phase, where the growth rate of HI is steadily falling to zero at crop maturity (Moot *et al.* 1996). We used ML for the prediction of the rate of change of harvest index (dHI/dt) in daily time step. Since the crop was fully irrigated and fertilized, we only used temperature (T_{min} and T_{max}) and solar radiation (Srad) as weather inputs for dHI/dt. Temperature was selected since it modifies the duration of grain filling, as well as the rate of dry matter accumulation into the grains (Dias and Lidon 2009). Solar radiation affects the rate of increase in grain weight, with low light levels decreasing yield under wet conditions (Shimoda and Sugikawa 2020). Finally, since the time series of HI exhibits the three distinct phases described above, we also used the lag of HI (HI_{n-1}) to increase the predictive capacity of ML.

2.4 GLAM-Parti model runs

The application of the ML models, the GLAM-Parti code and runs and the production of results and figures (package 'ggplot2', Wickham 2016) were implemented in R programming language (R Core Team 2022). The package 'tidymodels' (Kuhn and Wickham 2020) was used to fit the ML models into the GLAM-Parti code. RF and XGBoost were run with the packages 'ranger' (Wright and Ziegler 2017) and 'xgboost' (Chen and Guestrin 2016), respectively. Both ML models were applied for the prediction of the three target variables: RUE (regression), dHI/dt (regression) and iphen (classification). Each of the three algorithms was optimized by tuning the model hyperparameters. In RF, we optimized the hyperparameters 'trees' (total number

Table 1. Target variables and respective inputs of the three ML algorithms embedded into GLAM-Parti. RF and XGBoost were applied for prediction of RUE, the rate of change of harvest index (dHI/dt) and the phenological stage (iphen).

Target variable	Type	Features	ML models
iphen	Categorical	Srad, Solrac, T_{min}, T_{max}, TT, MS_{ratio}	RF, XGBoost
RUE	Numerical	Srad, T_{min}, T_{max}, VPD, MP_{ratio}	RF, XGBoost
dHI/dt	Numerical	Srad, T_{min}, T_{max}, HI_{n-1}	RF, XGBoost

of decision trees in ensemble), 'mtry' (number of input variables randomly selected for splitting at each node) and 'min_n' (minimum number of data points required to split a node further). In XGBoost we optimized the hyperparameters 'trees' (same as above), 'mtry' (same as above) and tree_depth (maximum depth of tree). For the hyperparameter tuning, we used Bayesian search with 10-fold cross-validation in 10 iteration setting (implemented in R package 'tidymodels').

Since RF is one of the most popular ML algorithms and repeatedly used in crop yield prediction (e.g. Shahhosseini *et al.* 2019; Vogel *et al.* 2019; Prasad *et al.* 2021), it was considered as the baseline ML model for the initial GLAM-Parti runs. In Section 3.1, we selected six treatments of the HSC data set for the training of RF (50 % of data) and the remaining six treatments were used for the testing of GLAM-Parti. The selection of treatments was done randomly with the 'set.seed' function of R. In Equation (8), the parameters *a*, *b* were computed using linear regression between log-transformed M_S vs. M_P only for the treatments selected in the training sample (for reference, the same linear regression but for all treatments is given in **Supporting Information—Fig. S1**). The skill of GLAM-Parti was evaluated against the observed end-of-season above-ground biomass and grain yield, as well as the days to anthesis and physiological maturity in the test treatments.

In Section 3.2, we evaluate GLAM-Parti against a wide range of fractions of the HSC data used for ML training, ranging from low (three treatments; around 25 % of data) to high number of observations (nine treatments; 75 % of data). The selection of the treatments exhibits considerable effect on the performance of the model. For instance, when only three treatments are used for the training of ML—if these are the three lowest crop yielding experiments—the model is not expected to show good skill against the unseen data. In order to reduce the impact of treatment selection, we evaluated GLAM-Parti with 10 different training samples in each subset (3–9 training treatments). In each training sample, the parameters *a*, *b* of Equation (8) were computed using linear regression of log-transformed M_S vs. M_P. RF and XGBoost were trained on the three target variables (iphen, RUE and *dHI/dt*), with Bayesian search optimizing the model hyperparameters. Next, we run GLAM-Parti with the optimized ML algorithms and compare the model output to the observations only for the treatments that were not included in the training sample. We test the model performance for above-ground biomass and grain yield, as well as the days to anthesis and maturity. **Supporting Information—Figure S3** is a graphical illustration of the above steps for the generation of the training samples, the optimization of ML and the evaluation of GLAM-Parti.

In Section 3.3, we assess the contribution of ML on the performance of GLAM-Parti. The model is compared to its predecessor, GLAM, a well-established—non-ML—process-based crop model with subroutines designed to capture the impact of high temperature on crop yield (Challinor *et al.* 2004, 2005). GLAM has been previously calibrated and run for the 12 HSC experiments (Maiorano *et al.* 2017). For GLAM-Parti, we use the run of Section 3.1, where RF was trained on 50 % of the experiments. For the model comparison, it should be noted that GLAM has been calibrated with 100 % of the HSC data (all 12 experiments), whilst GLAM-Parti has only seen 50 % of the data (six experiments) during ML training.

In Section 3.4, we evaluate GLAM-Parti on a larger spatial domain. We introduce IHSGE, which is a global network of spring wheat field experiments (Martre *et al.* 2017). IHSGE is combined with six randomly selected HSC treatments to ensure a balanced sample between all locations. The derived data set is comprised of six experiments in Maricopa, AZ, USA with a single cultivar (Yecora Rojo), 14 experiments in Mexico with two locations and two cultivars (Bacanora 88 and Nesser) and a total of 14 experiments in Egypt, India, the Sudan and Bangladesh with single locations and two cultivars (Bacanora 88 and Nesser). More details about the field experiments are given in **Supporting Information—Table S1**.

We randomly selected 50 % of the experiments for the training of RF and XGBoost (three out of six HSC and 14 out of 28 IHSGE treatments) [**see Supporting Information—Table S1**]. Six IHSGE experiments (the late-sown treatments of Obregon, Mexico (1991 and 1993) and the experiments in Aswan, Egypt) were excluded from the training sample, since no within-season biomass values were reported for the computation of the parameter t_m in Equation (13). In addition, the IHSGE experiments did not report within-season values of grain mass; thus, the parameter t_m for biomass was also used in the flexible sigmoid function describing the time series of grain mass in Equation (13).

Since the new data set contains three different cultivars, we introduced the input 'Cultivar' in the feature space of all ML algorithms (Table 1), which consists of three classes as follows:

1—cv. Yecora Rojo
2—cv. Bacanora 88
3—cv. Nesser

For the ML hyperparameter tuning (target variables: iphen, RUE, *dHI/dt*) we used Bayesian search as described above. The parameters *a*, *b* (Equation (8)) were computed using linear regression between the log-transformed M_S vs. M_P for the treatments of the HSC training sample.

2.5 Evaluation metrics

The following metrics are used for the evaluation of the GLAM-Parti model performance:

- Mean bias error (MBE)

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{15}$$

- Root mean squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{16}$$

- normalized root mean squared error (nRMSE)

$$\mathrm{nRMSE}\ (\%) = 100 \cdot \frac{\mathrm{RMSE}}{\bar{O}} \tag{17}$$

- Coefficient of determination (R^2)

$$R^2 = \left[\frac{\sum_{i=1}^{n} (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2 \sum_{i=1}^{n} (P_i - \bar{P})^2}} \right]^2 \quad (18)$$

where P_i and O_i are predictions and observations, respectively, $\bar{P}$ and $\bar{O}$ are mean values, and n equals the number of observations.

MBE computes the mean of the residuals and ranges between $(-\infty, +\infty)$. Negative MBE values show that the model tends to under-predict, while positive values reveal over-prediction compared to the observations. Thus, MBE shows the direction of the bias, with values close to zero signifying low model under-/over-prediction. RMSE measures the magnitude of the model error by squaring the residuals and penalizing large deviations between predictions and observations. It ranges between $[0, +\infty)$ and values close to zero reveal good model fit to the observations. nRMSE divides RMSE by the mean of the observations to normalize the metric across variables with different scales. It ranges between $[0, +\infty)$ and nRMSE < 10 %, 10 % < nRMSE < 20 %, 20 % < nRMSE < 30 % are considered as 'excellent', 'good' and 'fair', respectively, in crop modelling (Jamieson *et al.* 1991). R^2 measures the proportion of the variation in the observed variable that is captured by the model. It ranges between $[0, 1]$, with values closer to 1 indicating that high percentage of the measured variance is reproduced within the simulations. Detailed discussion of the evaluation metrics is given in Bennett *et al.* (2013).

3. RESULTS

3.1 GLAM-Parti performance with 50 % of HSC data for training of RF

Figure 3 illustrates the performance of GLAM-Parti when six HSC treatments are used for the training of ML and the remaining six treatments for model evaluation. GLAM-Parti successfully reproduced the observed patterns of above-ground biomass and grain yield, as well as the days to anthesis and maturity in the training treatments (Fig. 3; left of red line in A–D). The model also exhibited high skill in the simulations against the test treatments. GLAM-Parti reproduced 98 % of the variance in end-of-season biomass and grain yield (Fig. 3A and B; right of red line) and the respective RMSE (nRMSE) was 2.15 (17.06 %) and 1.06 (19.87 %) t ha^{-1}. Nevertheless, the model exhibited a negative bias in the prediction of both variables and MBE for above-ground biomass and grain yield was −1.76 and −1.02 t ha^{-1}, respectively (more details about model bias in Figs 4 and 5).

With regards to phenology, GLAM-Parti exhibited good skill in the simulation of the observed days to anthesis and maturity in the test treatments (Fig. 3C and D; right of red line) and the variance explained by the model was 99 and 98 %, respectively. Neverthelesss, the model underestimated the onset of both phenological stages, thus predicting anthesis and maturity earlier than observed (5.5 days earlier for anthesis and 5 days earlier for maturity). RMSE (nRMSE) was 6.49 days (10.5 %) for anthesis and 6.38 days (6.82 %) for maturity. Overall, the above results show good model skill, revealing the power of our methodology, which benefits from the combination of process understanding in crop modelling with the predictive capacity of ML.

3.2 GLAM-Parti performance with different fractions of HSC data for training of ML

RF and XGBoost were trained against various fractions of the HSC data (i.e. different numbers of treatments) and the skill of GLAM-Parti was compared to the end-of-season measurements of above-ground biomass and grain yield, as well as the days to anthesis and maturity in the test treatments (Fig. 4). We present the results of the model performance based on the median values of the evaluation metrics, unless otherwise stated. **Supporting Information—Tables S2–S5** contain all information discussed below and illustrated in the boxplots of Fig. 4. In addition to the end-of-season analysis, limited mid-season above-ground biomass and grain mass measurements were available for model testing. The results of the model performance against the mid-season biomass and grain mass observations are illustrated in **Supporting Information—Fig. S4** and discussed in Supporting Information.

With regards to the prediction of the end-of-season above-ground biomass, the use of 3 out of the 12 treatments for training of ML (i.e. 25 % of data) equipped GLAM-Parti with the ability to reproduce 84/82 % (RF/XGBoost) of the observed variance in the test treatments (Fig. 4). R^2 increased to 91/89 % (RF/XGBoost) at six and 95/96 % (RF/XGBoost) at nine training treatments. The application of RF resulted in biomass error of 2.83 t ha^{-1}/23.03 % (RMSE/nRMSE) at three training treatments, which decreased to 1.76 t ha^{-1}/14.57 % at six and 1.55 t ha^{-1}/12.72 % at nine training treatments. For XGBoost, the equivalent error was 2.84 t ha^{-1}/24.43 % at three training treatments, 1.78 t ha^{-1}/15.45 % at six and 1.76 t ha^{-1}/14.32 % at nine training treatments. In addition, both ML models exhibited low bias against all training settings, with XGBoost resulting in the highest underestimation at three training treatments (−0.64 t ha^{-1}) and the largest overestimation at nine training treatments (0.64 t ha^{-1}).

With regards to grain yield, GLAM-Parti reproduced 79/73 % (RF/XGBoost) of the observed variance, when three treatments were used for the training of ML. R^2 increased with the addition of more data and the crop model reproduced 84/83 % (RF/XGBoost) of the observed variance at six training treatments and 94/92 % (RF/XGBoost) at nine training treatments. Similarly, RMSE started at 1.44/1.75 t ha^{-1} (RF/XGBoost) at three training treatments, decreased to 1.16/0.98 t ha^{-1} (RF/XGBoost) at six training treatments and a minimum of 0.96/0.89 t ha^{-1} (RF/XGBoost) at nine training treatments. This translated to a relative error (nRMSE) of 28.31/36.49 % (RF/XGBoost) at three training treatments, which reduced to 22.31/18.3 % (RF/XGBoost) at six and 18.31/17.19 % (RF/XGBoost) at nine training treatments. Moreover, GLAM-Parti exhibited a relatively small bias in yield prediction. XGBoost resulted in the highest underestimation at three training treatments (−0.33 t ha^{-1}), as well as the largest overestimation at nine training treatments (0.63 t ha^{-1}).

With regards to phenological development, GLAM-Parti reproduced at least 97 % of the observed variance in the days to anthesis and maturity with both RF and XGBoost, under all fractions of treatments used for ML training. The model error for anthesis (expressed as RMSE) was 4.45/3.55 days (RF/XGBoost) at three training treatments and decreased to 3.84/3.08 days (RF/XGBoost) at nine training treatments. Similarly, RMSE for maturity started at 7.63/7.78 days (RF/XGBoost) with three training treatments and minimized at

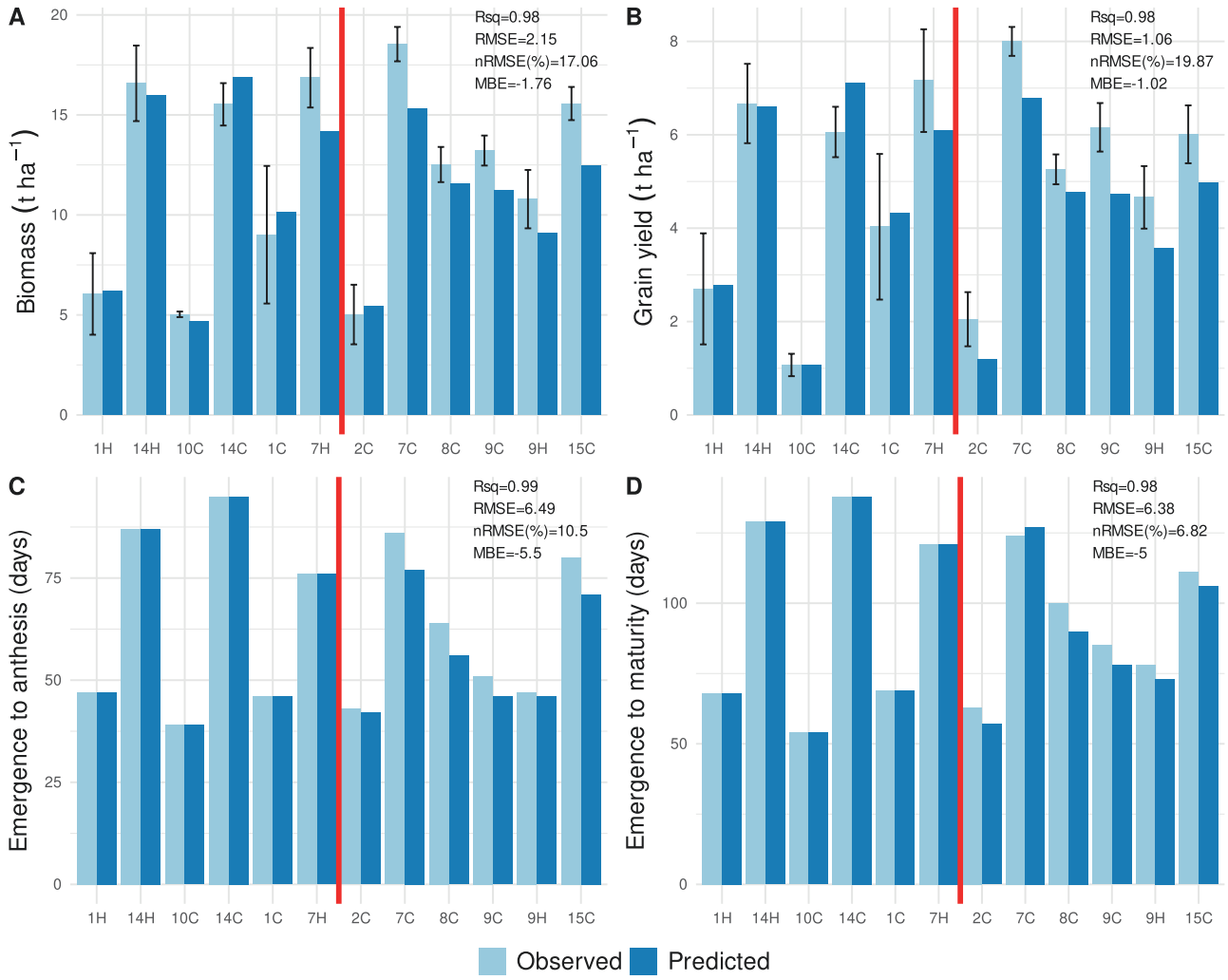

Figure 3. Barplots of observed and predicted (GLAM-Parti) end-of-season above-ground biomass (A), grain yield (B), days from emergence to anthesis (C) and days from emergence to maturity (D) for the wheat treatments of the 'Hot Serial Cereal Experiment' (HSC). Vertical lines separate the treatments used for training of RF (left of vertical line) and the treatments used for model testing (right of vertical line). Top right corner shows the evaluation metrics (coefficient of determination (R2), root mean squared error (RMSE), normalized root mean squared error (nRMSE) and mean bias error (MBE)), which are computed only for the test treatments (right of vertical line). Error bars in (A) and (B) show standard errors in measurements of above-ground biomass and grain yield, respectively.

5.36/4.88 days (RF/XGBoost) with nine training treatments. In terms of percentage error, nRMSE remained lower than 10 % for anthesis and maturity, using both RF and XGBoost in the total range of training treatments (3–9 treatments; 25–75 % of training data). Also, both ML models exhibited a negative bias in the prediction of anthesis against low training data (3–5 treatments). As a result, GLAM-Parti underestimated the days to anthesis by 1.31/1.21 days (RF/XGBoost; median values), with a maximum underestimation of 5/6 days (RF/XGBoost). For six or more training treatments, the model exhibited a change to positive bias in most runs. RF and XGBoost overestimated the days to anthesis by 0.83/0.5 days (RF/XGBoost; median values), with a maximum MBE of 5.2/4.83 days (RF/XGBoost). Finally, GLAM-Parti underestimated the days to maturity with both RF and XGBoost against almost all settings (with the exception of nine training treatments). For RF, the highest underestimation was seen at four training treatments (−2.62 days), whilst XGBoost resulted in the largest negative bias at three training treatments (−2.72 days).

3.3 Comparison between GLAM and GLAM-Parti

Table 2 compares GLAM and GLAM-Parti in the HSC experiments. Moreover, **Supporting Information—Fig. S5** presents the outputs of both models against the observations. GLAM was calibrated with 12 experiments, whilst GLAM-Parti used only six experiments for ML training. Nevertheless, GLAM-Parti exhibited lower RMSE for biomass (1.96/1.79 t ha^{-1}), grain yield (1.40/0.87 t ha^{-1}), the days to anthesis (6.34/4.59 days) and maturity (6.45/4.51 days) (GLAM/

Figure 4. Boxplots of GLAM-Parti model performance for end-of-season above-ground biomass, grain yield, the days to anthesis and maturity using RF and XGBoost. Both ML models were trained on 3 to 9 out of 12 treatments of the 'Hot Serial Cereal Experiment' for wheat (HSC) and GLAM-Parti was evaluated against the unseen data. Evaluation metrics are coefficient of determination (R^2), root mean squared error (RMSE), normalized root mean squared error (nRMSE) and mean bias error (MBE).

GLAM-Parti) in the simulations. In addition, R^2 of GLAM-Parti was higher for grain yield (0.83/0.88), the days to anthesis (0.92/0.96) and maturity (0.96/0.98), while GLAM has higher R^2 for biomass (0.93/0.9) (GLAM/GLAM-Parti). On the other hand, GLAM-Parti showed larger MBE in the prediction of biomass (0.12/−0.96), grain yield (−0.42/0.49), the days to anthesis (0.17/−2.75) and maturity (−0.5/−2.5) (GLAM/GLAM-Parti). According to the above results, GLAM-Parti exhibits similar or improved performance compared to GLAM using half of the data for model training.

3.4 GLAM-Parti performance with global data set

Figure 5 illustrates the performance of GLAM-Parti in simulating the above-ground biomass, grain yield and the days to anthesis and maturity of three spring wheat cultivars in six countries [**see Supporting Information—Table S1**]. Since RF and XGBoost resulted in similar model skill, we only present the results of GLAM-Parti using RF. **Supporting Information—Figures S6 and S7** show barplots of model performance in each experiment of the data set with RF and XGBoost, respectively.

The model exhibited good skill in the prediction of the end-of-season above-ground biomass and reproduced 73 % of the variation (R^2) across locations and cultivars. RMSE (nRMSE) for biomass was 1.61 t ha^{-1} (14.93 %) and no significant bias was observed (MBE = −0.19 t ha^{-1}). Similarly, R^2 for grain yield was 0.76 and RMSE (nRMSE) was 0.68 (16.02 %). No systematic error was observed in the simulation of grain yield (MBE = 0.06 t ha^{-1}). With regards to the crop phenological development, GLAM-Parti was more skilful in predicting the days to maturity (R^2 = 0.79) compared to anthesis (R^2 = 0.66). RMSE (nRMSE) was 8.95 days (13.15 %) for anthesis and 10.26 days (9.89 %) for maturity, respectively. Moreover, there was a negative bias in the

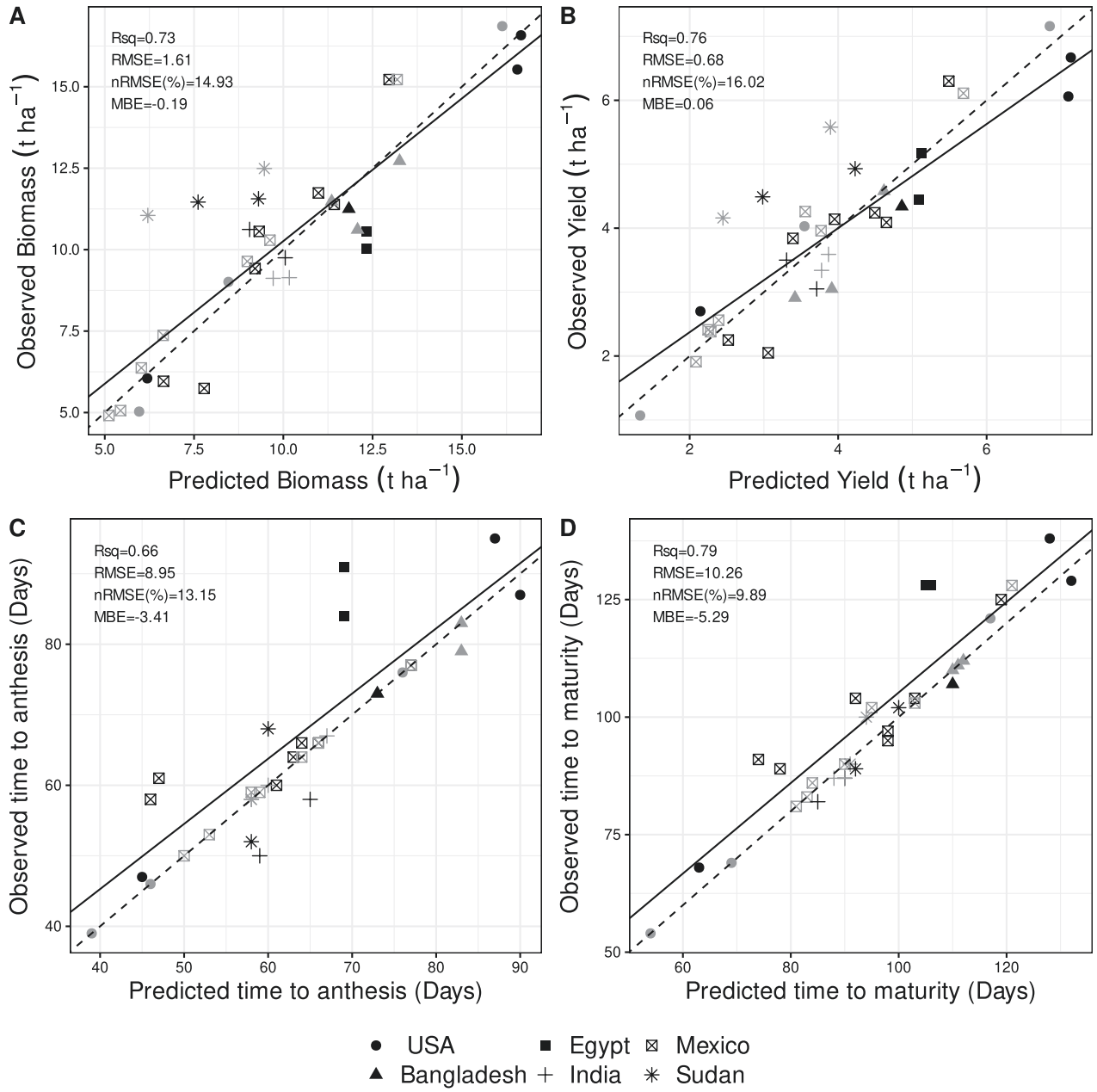

Figure 5. Comparison between observed and GLAM-Parti simulated end-of-season above-ground biomass (A), grain yield (B), days from emergence to anthesis (C) and days from emergence to maturity (D). The field experiments used to derive the plots are reported in Supporting Information—Table S1. In each panel, grey points are experiments used for ML training and black points are experiments used for GLAM-Parti model testing. Evaluation metrics are coefficient of determination (R^2), root mean squared error (RMSE), normalized root mean squared error (nRMSE) and mean bias error (MBE). The linear regression (solid line) and all evaluation metrics are computed only for the experiments used in model testing (black points). The dashed line is the 1:1 line.

prediction of both phenological stages. GLAM-Parti estimated the days to anthesis and maturity earlier than observed with a systematic error of −3.41 and −5.29 days, respectively. The above results reveal that the model is better in simulating wheat biomass and yield, while improvement in the prediction of crop phenology is needed to decrease the systematic error in the progression of the phenological stages.

4. DISCUSSION

Three ML algorithms were embedded into the process-based crop model GLAM-Parti to estimate variables that play a pivotal role in the determination of crop biomass and grain yield. The first target variable, RUE, drives crop growth in many crop models—including GLAM-Parti—under a well-established light interception approach (Equation

Table 2. Performance of GLAM and GLAM-Parti in the 'Hot Serial Cereal Experiment' for wheat (HSC). Evaluation metrics are coefficient of determination (R^2), root mean squared error (RMSE), normalized root mean squared error (nRMSE) and mean bias error (MBE).

	R^2		RMSE		nRMSE (%)		MBE	
	GLAM	GLAM-Parti	GLAM	GLAM-Parti	GLAM	GLAM-Parti	GLAM	GLAM-Parti
Biomass	0.93	0.9	1.96	1.79	16	15	0.12	−0.96
Yield	0.83	0.88	1.40	0.87	28	17	−0.42	−0.49
Anthesis	0.92	0.96	6.34	4.59	10	7	0.17	−2.75
Maturity	0.96	0.98	6.45	4.51	7	5	−0.5	−2.5

(10)). In the HSC experiment, the wheat crop was planted in regular time intervals during a 2-year period, resulting in a wide range of temperature, solar radiation and VPD conditions. Both RF and XGBoost exhibited high efficiency in learning the patterns between inputs and RUE during the course of the growing season. This resulted in good model performance in the prediction of crop biomass with both ML models. GLAM-Parti reproduced at least 82 % of the observed variance in wheat biomass (Fig. 4; median values) and the model error (expressed as nRMSE) was less than 20 % (Fig. 4; median values), when four treatments (i.e. 33 % of data) were used for the training of ML. Moreover, the evaluation of GLAM-Parti in the extended data set (HSC + IHSGE) revealed good performance for the three wheat cultivars in six countries. The model error for above-ground biomass and grain yield was less than 20 % and GLAM-Parti reproduced around three-quarters of the observed variance in both variables (Fig. 5).

Stress factors are the most common approach to limit growth potential in process-based crop models (e.g. Ewert and Porter 2000; Asseng *et al.* 2011; Jin *et al.* 2016; Brown *et al.* 2018). These factors do not usually describe plant processes explicitly, but instead consist of a convenient way to modify the crop response to unfavourable environmental conditions. In this study, the incorporation of ML into GLAM-Parti eliminates the use of stress factors in high-temperature environments. This is a novel and significant improvement in crop modelling, since crops in the field are usually impacted by multiple environmental stresses (e.g. heat stress combined with drought/high VPD/limited soil nitrogen/air pollution, etc.) (Mittler 2006), thus requiring the simultaneous use of many stress factors. In such cases, the high complexity of crop/environment interactions may not allow a clear way of combining stress factors in a crop model, since the overall plant response cannot be derived directly from the sum of plant responses to individual stresses (Barnabás *et al.* 2008; Parent *et al.* 2017; Rashid *et al.* 2018). We instead apply ML to reveal the signals between weather conditions (temperature, solar radiation and VPD—including multiple extremes) and crop growth and development. In other words, we do not attempt to prescribe the plant response to the environment through mechanistic, stress-factor-based equations, but instead learn those interactions from data using state-of-the-art ML methods. Hence, we take advantage of the predictive power of ML to deal with complexity that is otherwise extremely difficult to handle.

Our framework differs from existing ML models for crop yield prediction, as it does not exclusively rely on ML. Purely ML-based crop yield algorithms deal with a large, complex feature space, since the weather data often exist in different temporal resolution from the target variable (i.e. daily weather features vs. end-of-season crop yield). In order to deal with this issue, modellers post-process the weather inputs by averaging them in weekly/monthly/annual time step (e.g. Folberth *et al.* 2019; Shahhosseini *et al.* 2021; Srivastava *et al.* 2022). As a result, information may be lost (e.g. extreme weather values) and the dimensionality increases significantly (e.g. the weather variable 'daily temperature' can become up to 12 variables if averaged monthly). In GLAM-Parti, the weather predictors are in the same temporal resolution (i.e. daily time step) as the three target variables (i.e. RUE, *dHI/dt* and iphen), thus allowing the use of observations without post-processing.

The HSC and IHSGE data sets consist of 40 seasonal yield observations available for ML training. This is a far too small sample for training and evaluation of a pure ML model. However, the consideration of the same experiments in terms of daily observations comprises a significantly larger sample size. In this study, the 40 field experiments make approximately 4000 daily observations (i.e. 40 experiments × average 100 days from crop emergence to maturity per experiment). Hence, in the GLAM-Parti framework, the selection of output variables with daily time step results in an adequately large sample for ML model development. In other words, the relatively small number of field experiments provides an adequate sample of daily crop growth, development and partitioning data. ML is used for the prediction of daily changes in crop-related processes, which are then incorporated into GLAM-Parti for biomass and yield estimations. Thus, the GLAM-Parti framework consists of a possible avenue for the use of ML in data sets with a small number of field experiments (i.e. limited end-of-season yield observations).

With regards to the crop phenological development, the model predicted the days to anthesis and physiological maturity earlier than observed, leading to a negative bias in the simulations (Figs 3 and 5). One possible explanation is the relatively small training sample of this study. Additional field experiments can lead to the development of more skilful ML models. Another reason for the systematic error could be the omission of relevant explanatory variables. The weather inputs used for phenology are radiation (Srad, Solrac) and temperature-based (T_{min}, T_{max}, TT). Most significantly, TT was computed in the simplest form, summing only the mean of T_{min} and T_{max} (Equation (4)). We did not attempt to incorporate a more complex temperature response curve (e.g. Wang and Engel 1998), since we wanted to test the model performance with minimal number of parameters. Nevertheless, the use of a simple sum of degree days for phenology often leads to lower performance than more complex process-based algorithms (Wallach *et al.* 2021a). Wang *et al.* (2017) make skilful predictions with a

curvilinear temperature response function based on a minimum, optimum and maximum cardinal temperature for wheat. In the future, the introduction of a feature of thermal time accumulation with a more complex, wheat-based function could increase the ML skill and may decrease the bias in the simulations for the progression of the crop phenological stages.

GLAM-Parti contains the following four physiological parameters: allometric coefficients *a*, *b* (Equation (8)); *k* and SLA (Equation (10)). Together with the ML model hyperparameters, they consist of the complete set of parameters. In this study, the only calibration process applied was the optimization of the ML hyperparameters for predicting RUE, *dHI/dt* and iphen. Also, it should be noted that ML is fully integrated into GLAM-Parti, meaning that neither of the two pieces (the process-based part of GLAM-Parti or the ML models) can produce grain yield output individually. Hence, there is no baseline crop model without ML. Nevertheless, GLAM-Parti retains its process-based nature, thus providing information other than end-of-season grain yield (the model outputs daily estimations of crop biomass, individual organ mass, etc.). Therefore, the model can be used for applications other than end-of-season crop yield prediction. The model evaluation against within-season crop measurements revealed a significantly better skill in predicting mid-season above-ground biomass compared to grain mass [**see Supporting Information—Fig. S4**]. This points to the need of further model improvement in the representation of the time series of HI (for more details, see discussion in SI). Additionally, more mid-season observations of grain mass would be particularly useful to constrain and evaluate the crop model in the future.

GLAM and GLAM-Parti were compared against the HSC experiments. Prior to the evaluation, GLAM was calibrated with 100 % of the data (all 12 experiments), whilst GLAM-Parti used only 50 % (six experiments) during ML training. Nevertheless, GLAM-Parti exhibited 27.6–37.9 % lower RMSE for biomass, grain yield, the days to anthesis and maturity (Table 2). In addition, R^2 of GLAM-Parti was higher for grain yield and the days to anthesis and maturity, whilst GLAM had higher R^2 for biomass. On the other hand, GLAM-Parti exhibited larger MBE values for all variables tested. Nevertheless, GLAM-Parti was trained on 50 % of the data, while GLAM was calibrated with 100 % of the experiments, thus minimizing the systematic errors in the simulations. Overall, the above comparison reveals the benefit of the ML parameterizations in the GLAM-Parti framework, since the model shows similar or improved performance compared to its predecessor, GLAM, using only half of the data for model training.

Here, GLAM-Parti was tested with wheat treatments of well-irrigated and fertilized conditions. Thus, the variation in crop growth, development and yield was not driven by the availability of soil water and nutrient contents. Model application to a larger and more complex spatial domain (e.g. gridded regional runs) would require additional ML features to reproduce the spatio-temporal wheat yield variation. Even though this is out of the scope of this study, some ideas on how to incorporate additional inputs for model runs in larger domains are provided below:

- Precipitation-related drivers: These can be incorporated into the model by using observed rainfall or cumulative sums at daily time step during the crop growing season.
- Soil properties and water content: Soil moisture levels can be derived from time series of remote sensing data. Additionally, the most important soil properties can be taken into account in developing the ML model.
- Management practices: Levels of irrigation and fertilizer application can be introduced into ML feature space.

The use of three ML algorithms (target variables: iphen, RUE, *dHI/dt*) involves the challenge of selecting the correct models and features in each one, as well as optimizing the model hyperparameters. Most importantly, feature selection is an essential element in building an ML model, since the use of incorrect drivers of plant growth and development can result in relationships that do not correspond to the underlying patterns of the real world. Hence, the remarkable predictive capacity of ML should be harnessed in close contact with the progress in plant science to develop models of increased utility, which provide the ability to explain patterns and interactions of the real world. Finally, an additional challenge of GLAM-Parti is the requirement of at least one in-season measurement of above-ground biomass and grain mass for the construction of the sigmoid curves (see Equation (13) and Fig. 2). Therefore, experiments like HSC and IHSGE provide ideal data sets, essential for achieving progress in crop modelling and we are very grateful to all the scientists and people who worked hard to create these data sets, leading to the model development presented in this study.

5. CONCLUSION

We integrated ML algorithms into the process-based crop model, GLAM-Parti. The new framework exhibited high skill in the prediction of wheat growth, development and yield against a wide range of heat stress experiments. The incorporation of ML into GLAM-Parti eliminates stress factors under high temperature and decreases the physiological model parameters down to four (the full set of parameters includes ML hyperparameters too). Our methodology is highly data-driven, relying on the remarkable capacity of ML in picking up the signals between input (weather and crop) features and target variables. This leads to high model performance in wheat growth and development at daily time step against a wide range of environmental conditions. Here, GLAM-Parti was primarily tested for high temperature; however, the model does not include any heat stress-focused procedures. Given the right data, GLAM-Parti should—in principle—be similarly applied to other crop growing conditions. Finally, open-source data sets such as Martre *et al.* (2017, 2018) are of extremely high importance in taking advantage of ML techniques to develop novel methodologies in crop modelling, such as the one presented in this study.

SUPPORTING INFORMATION

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

Section1 GLAM-Parti model performance for mid-season above-ground biomass and grain mass

Table S1 HSC and IHSGE wheat field treatments used for ML training and GLAM-Parti model Evaluation

Table S2 GLAM-Parti skill for above-ground biomass in the HSC treatments
Table S3 Same as above but for grain yield
Table S4 Same as above but for days from emergence to anthesis
Table S5 Same as above but for days from emergence to maturity
Figure S1 Log-transformed MS vs. MP for the HSC treatments
Figure S2 Time series of wheat biomass with sigmoid function
Figure S3 Methodology for selection of HSC wheat treatments for training of ML and GLAM-Parti model evaluation
Figure S4 GLAM-Parti model performance for predicting mid-season above-ground biomass and grain mass
Figure S5 Observed, GLAM and GLAM-Parti above-ground biomass, grain yield, days from emergence to anthesis and days from emergence to maturity
Figure S6 GLAM-Parti model performance in the IHSGE treatments with RF
Figure S7 GLAM-Parti model performance in the IHSGE treatments with XGBoost

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their constructive comments on the manuscript.

SOURCES OF FUNDING

This research was supported with funding from the European Union's Horizon 2020 program through the projects CONFER Co-production of Climate Services for East Africa (grant 869720) and AfriCultuReS 'Enhancing Food Security in African Agricultural Systems with the Support of Remote Sensing' (grant 774652). C.R.D. was supported by a Biotechnology and Biological Sciences Research Council (BBSRC)-funded project named Bean Breeding for Adaptation to a Changing Climate and Post-Conflict Colombia (BBACO). Grant number BB/S018964/1.

CONFLICT OF INTEREST

None declared.

CONTRIBUTIONS BY THE AUTHORS

Ioannis Droutsas: Conceptualization, Formal Analysis, Software, Methodology, Validation, Data Curation, Writing - Original Draft Preparation, Writing - Review & Editing: Andrew Challinor: Conceptualization, Supervision, Funding Acquisition, Writing - Review & Editing: Chetan Deva: Conceptualization, Methodology, Writing - Review & Editing: Enli Wang: Writing - Review & Editing.

DATA AVAILABILITY

The GLAM-Parti-ML model is open-source and licenced under the GNU General Public License v3.0. For the production of the results and figures of this manuscript, the model can be downloaded from the Gtihub repository: https://github.com/GLAM-Leeds/GLAM-Parti-ML. The data sets 'Hot Serial Cereal Experiment' (HSC) and 'International Heat Stress Genotype Experiment' (IHSGE) were downloaded from Harvard Dataverse with doi:10.7910/DVN/ECSFZG for HSC and doi:10.7910/DVN/1WCFHK for IHSGE.

LITERATURE CITED

Asseng S, Ewert F, Martre P, Rötter RP, Lobell DB, Cammarano D, Kimball BA, Ottman MJ, Wall GW, White JW, Reynolds MP, Alderman PD, Prasad PVV, Aggarwal PK, Anothai J, Basso B, Biernath C, Challinor AJ, De Sanctis G, Doltra J, Fereres E, Garcia-Vila M, Gayler S, Hoogenboom G, Hunt LA, Izaurralde RC, Jabloun M, Jones CD, Kersebaum KC, Koehler A-K, Müller C, Naresh Kumar S, Nendel C, O'Leary G, Olesen JE, Palosuo T, Priesack E, Eyshi Rezaei G, Ruane AC, Semenov MA, Shcherbak I, Stockle C, Stratonovitch P, Streck T, Supit I, Tao F, Thorburn PJ, Waha K, Wang E, Wallach D, Wolf J, Zhao Z, Zhu Y. 2015. Rising temperatures reduce global wheat production. *Nature Climate Change* **5**:143–147.

Asseng S, Ewert F, Rosenzweig C, Jones JW, Hatfield JL, Ruane AC, Boote KJ, Thorburn PJ, Rötter RP, Cammarano D, Brisson N, Basso B, Martre P, Aggarwal PK, Angulo C, Bertuzzi P, Biernath C, Challinor AJ, Doltra J, Gayler S, Goldberg R, Grant R, Heng L, Hooker J, Hunt LA, Ingwersen J, Izaurralde RC, Kersebaum KC, Müller C, Naresh Kumar S, Nendel C, O'Leary G, Olesen JE, Osborne TM, Palosuo T, Priesack E, Ripoche D, Semenov MA, Shcherbak I, Steduto P, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Travasso M, Waha K, Wallach D, White JW, Williams JR, Wolf J. 2013. Uncertainty in simulating wheat yields under climate change. *Nature Climate Change* **3**:827–832.

Asseng S, Foster IAN, Turner NC. 2011. The impact of temperature variability on wheat yields. *Global Change Biology* **17**:997–1012.

Barnabás B, Jäger K, Fehér A. 2008. The effect of drought and heat stress on reproductive processes in cereals. *Plant, Cell & Environment* **31**:11–38.

Bassu S, Brisson N, Durand J-L, Boote K, Lizaso J, Jones JW, Rosenzweig C, Ruane AC, Adam M, Baron C, Basso B, Biernath C, Boogaard H, Conijn S, Corbeels M, Deryng D, Sanctis GD, Gayler S, Grassini P, Hatfield J, Hoek S, Izaurralde C, Jongschaap R, Kemanian AR, Kersebaum KC, Kim S-H, Kumar NS, Makowski D, Müller C, Nendel C, Priesack E, Pravia MV, Sau F, Shcherbak I, Tao F, Teixeira E, Timlin D, Waha K. 2014. How do various maize crop models vary in their responses to climate change factors? *Global Change Biology* **20**:2301–2320.

Baumont M, Parent B, Manceau L, Brown HE, Driever SM, Muller B, Martre P. 2019. Experimental and modeling evidence of carbon limitation of leaf appearance rate for spring and winter wheat. *Journal of Experimental Botany* **70**:2449–2462.

Bennett ND, Croke BF, Guariso G, Guillaume JH, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LT, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andréassian V. 2013. Characterising performance of environmental models. *Environmental Modelling & Software* **40**:1–20.

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

Breiman L. 2001. Random forests. *Machine Learning* **45**:5–32.

Brown H, Huth N, Holzworth D. 2018. Crop model improvement in APSIM: using wheat as a case study. *European Journal of Agronomy* **100**:141–150.

Brown HE, Huth NI, Holzworth DP, Teixeira EI, Wang E, Zyskowski RF, Zheng B. 2019. A generic approach to modelling, allocation and redistribution of biomass to and from plant organs. *In Silico Plants* **1**:diy004; doi:10.1093/insilicoplants/diy004.

Cai Y, Guan K, Lobell D, Potgieter AB, Wang S, Peng J, Xu T, Asseng S, Zhang Y, You L, Peng B. 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* **274**:144–159.

Castellvi F, Perez PJ, Villar JM, Rosell JI. 1996. Analysis of methods for estimating vapor pressure deficits and relative humidity. *Agricultural and Forest Meteorology* **82**:29–45.

Ceglar A, Van der Wijngaart R, De Wit A, Lecerf R, Boogaard H, Seguini L, Van den Berg M, Toreti A, Zampieri M, Fumagalli D, Baruth B. 2019. Improving WOFOST model to simulate winter wheat phenology in Europe: evaluation and effects on yield. *Agricultural Systems* **168**:168–180.

Challinor AJ, Ewert F, Arnold S, Simelton E, Fraser E. 2009. Crops and climate change: progress, trends, and challenges in simulating impacts and informing adaptation. *Journal of Experimental Botany* **60**:2775–2789.

Challinor AJ, Wheeler TR, Craufurd PQ, Slingo JM. 2005. Simulation of the impact of high temperature stress on annual crop yields. *Agricultural and Forest Meteorology* **135**:180–189.

Challinor AJ, Wheeler TR, Craufurd PQ, Slingo JM, Grimes DIF. 2004. Design and optimisation of a large-area process-based model for annual crops. *Agricultural and Forest Meteorology* **124**:99–120.

Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'16. New York, NY, USA: ACM, pp. 785–794. doi:10.1145/2939672.2939785.

Crane-Droesch A. 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters* **13**:114003.

Craufurd PQ, Wheeler TR. 2009. Climate change and the flowering time of annual crops. *Journal of Experimental Botany* **60**:2529–2539.

Dias AS, Lidon FC. 2009. Evaluation of grain filling rate and duration in bread and durum wheat, under heat stress after anthesis. *Journal of Agronomy and Crop Science* **195**:137–147.

Droutsas I, Challinor AJ, Arnold SR, Mikkelsen TN, Ø Hansen EM. 2020. A new model of ozone stress in wheat including grain yield loss and plant acclimation to the pollutant. *European Journal of Agronomy* **120**:126125.

Droutsas I, Challinor AJ, Swiderski M, Semenov MA. 2019. New modelling technique for improving crop model performance—application to the GLAM model. *Environmental Modelling & Software* **118**:187–200.

Eitzinger J, Thaler S, Schmid E, Strauss F, Ferrise R, Moriondo M, Bindi M, Palosuo T, Rötter R, Kersebaum KC, Olesen JE, Patil RH, Saylan L, Çaldağ B, Çaylak O. 2013. Sensitivities of crop models to extreme weather conditions during flowering period demonstrated for maize and winter wheat in Austria. *The Journal of Agricultural Science* **151**:813–835.

Ewert F, Porter JR. 2000. Ozone effects on wheat in relation to CO_2: modelling short-term and long-term responses of leaf photosynthesis and leaf duration. *Global Change Biology* **6**:735–750.

Ewert F, Rötter RP, Bindi M, Webber H, Trnka M, Kersebaum KC, Olesen JE, van Ittersum MK, Janssen S, Rivington M, Semenov MA, Wallach D, John Roy P, Stewart D, Verhagen J, Gaiser T, Palosuo T, Tao F, Nendel C, Roggero PP, Bartosová L, Asseng S. 2015. Crop modelling for integrated assessment of risk to food production from climate change. *Environmental Modelling & Software* **72**:287–303.

Feng P, Wang B, Liu DL, Waters C, Yu Q. 2019. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. *Agricultural and Forest Meteorology* **275**:100–113.

Ferrante A, Mariani L. 2018. Agronomic management for enhancing plant tolerance to abiotic stresses: high and low values of temperature, light intensity, and relative humidity. *Horticulturae* **4**:21.

Fleisher DH, Condori B, Quiroz R, Alva A, Asseng S, Barreda C, Bindi M, Boote KJ, Ferrise R, Franke AC, Govindakrishnan PM, Harahagazwe D, Hoogenboom G, Naresh Kumar S, Merante P, Nendel C, Olesen JE, Parker PS, Raes D, Raymundo R, Ruane AC, Stockle C, Supit I, Vanuytrecht E, Wolf J, Woli P. 2017. A potato model intercomparison across varying climates and productivity levels. *Global Change Biology* **23**:1258–1281.

Folberth C, Baklanov A, Balkovič J, Skalský R, Khabarov N, Obersteiner M. 2019. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and Forest Meteorology* **264**:1–15.

Foulkes MJ, Slafer GA, Davies WJ, Berry PM, Sylvester-Bradley R, Martre P, Calderini DF, Griffiths S, Reynolds MP. 2011. Raising yield potential of wheat. III. Optimizing partitioning to grain while maintaining lodging resistance. *Journal of Experimental Botany* **62**:469–486.

Hussain J, Khaliq T, Ahmad A, Akhtar J. 2018. Performance of four crop model for simulations of wheat phenology, leaf growth, biomass and yield across planting dates. *PLoS One* **13**:e0197546.

Jamieson PD, Porter JR, Wilson DR. 1991. A test of the computer simulation model arcwheat1 on wheat crops grown in New Zealand. *Field Crops Research* **27**:337–350.

Jamieson PD, Semenov MA, Brooking IR, Francis GS. 1998. Sirius: a mechanistic model of wheat response to environmental variation. *European Journal of Agronomy* **8**:161–179.

Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, Timlin DJ, Shim K-M, Gerber JS, Reddy VR, Kim S-H. 2016. Random forests for global and regional crop yield predictions. *PLoS One* **11**:e0156571.

Jin Z, Zhuang Q, Tan Z, Dukes JS, Zheng B, Melillo JM. 2016. Do maize models capture the impacts of heat and drought stresses on yield? Using algorithm ensembles to identify successful approaches. *Global Change Biology* **22**:3112–3126.

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

Kuhn M, Wickham H. 2020. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org.

Leng G, Hall JW. 2020. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental Research Letters* **15**:044027.

Lischeid G, Webber H, Sommer M, Nendel C, Ewert F. 2022. Machine learning in crop yield modelling: a powerful tool, but no surrogate for science. *Agricultural and Forest Meteorology* **312**:108698.

Maiorano A, Martre P, Asseng S, Ewert F, Müller C, Rötter RP, Ruane AC, Semenov MA, Wallach D, Wang E, Alderman P, Kassie B, Biernath C, Basso B, Cammarano D, Challinor A, Doltra J, Dumont B, Rezaei E, Gayler S, Kersebaum K, Kimball B,

Koehler A, Liu B, O'Leary G, Olesen J, Ottman M, Priesack E, Reynolds M, Stratonovitch P, Streck T, Thorburn P, Waha K, Wall G, White J, Zhao Z, Zhu Y. 2017. Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research* **202**:5–20.

Martre P, Kimball BA, Ottman MJ, Wall GW, White JW, Asseng S, Ewert F, Cammarano D, Maiorano A, Aggarwal PK, Anothai J, Basso B, Biernath C, Challinor AJ, De Sanctis G, Doltra J, Dumont B, Fereres E, Garcia-Vila M, Gayler S, Hoogenboom G, Hunt LA, Izaurralde RC, Jablou M, Jones CD, Kassie BT, Kersebaum KC, Koehler A-K, Muller C, Kumar SN, Liu B, Lobell DB, Nendel C, O'Leary G, Olesen JE, Palosuo T, Priesack E, Rezaei EE, Ripoche D, Rotter RP, Semenov MA, Stockle C, Stratonovitch P, Streck T, Supit I, Tao F, Thorburn P, Waha K, Wang E, Wolf J, Zhao Z, Zhu Z. 2018. The hot serial cereal experiment for modeling wheat response to temperature: field experiments and AgMIP-wheat multi-model simulations. *Open Data Journal for Agricultural Research* **4**:28–34.

Martre P, Reynolds MP, Asseng S, Ewert F, Alderman P, Cammarano D, Maiorano A, Ruane AC, Aggarwal PK, Anothai J, Basso B, Biernath C, Challinor AJ, De Sanctis G, Doltra J, Dumont B, Fereres E, Garcia-Vila M, Gayler S, Hoogenboom G, Hunt LA, Izaurralde RC, Jabloun M, Jones CD, Kassie BT, Kersebaum KC, Koehler AK, Müller C, Kumar SN, Liu B, Lobell CB, Nendel C, O'Leary G, Olesen JE, Palosuo T, Priesack E, Eyshi Rezaei E, Ripoche D, Rötter RP, Semenov MA, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Thorburn P, Waha K, Wang E, White JW, Wolf J, Zhao Z, and Zhu Y. 2017. The International Heat Stress Genotype Experiment for modeling wheat response to heat: field experiments and AgMIP-wheat multi-model simulations. *Open Data Journal for Agricultural Research* **3**:23–28.

Martre P, Wallach D, Asseng S, Ewert F, Jones JW, Rötter RP, Boote KJ, Ruane AC, Thorburn PJ, Cammarano D, Hatfield JL, Rosenzweig C, Aggarwal PK, Angulo C, Basso B, Bertuzzi P, Biernath C, Brisson N, Challinor AJ, Doltra J, Gayler S, Goldberg R, Grant RF, Heng L, Hooker J, Hunt LA, Ingwersen J, Izaurralde RC, Kersebaum KC, Müller C, Kumar SN, Nendel C, O'Leary G, Olesen JE, Osborne TM, Palosuo T, Priesack E, Ripoche D, Semenov MA, Shcherbak I, Steduto P, Stöckle CO, Stratonovitch P, Streck T, Supit I, Tao F, Travasso M, Waha K, White JW, Wolf J. 2015. Multimodel ensembles of wheat growth: many models are better than one. *Global Change Biology* **21**:911–925.

Merah O, Monneveux P. 2015. Contribution of different organs to grain filling in durum wheat under Mediterranean conditions. I. Contribution of post-anthesis photosynthesis and remobilization. *Journal of Agronomy and Crop Science* **201**:344–352.

Mittler R. 2006. Abiotic stress, the field environment and stress combination. *Trends in Plant Science* **11**:15–19.

Moot DJ, Jamieson PD, Henderson AL, Ford MA, Porter JR. 1996. Rate of change in harvest index during grain-filling of wheat. *The Journal of Agricultural Science* **126**:387–395.

Müller C, Elliott J, Chryssanthacopoulos J, Arneth A, Balkovic J, Ciais P, Deryng D, Folberth C, Glotter M, Hoek S, Iizumi T, Izaurralde RC, Jones C, Khabarov N, Lawrence P, Liu W, Olin S, Pugh TAM, Ray DK, Reddy A, Rosenzweig C, Ruane AC, Sakurai G, Schmid E, Skalsky R, Song CX, Wang X, De Wit Hong Yang A. 2017. Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. *Geoscientific Model Development* **10**:1403–1422.

Newman SJ, Furbank RT. 2021. Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. *Nature Plants* **7**:1354–1363.

Ottman MJ, Hunt LA, White JW. 2013. Photoperiod and vernalization effect on anthesis date in winter-sown spring wheat regions. *Agronomy Journal* **105**:1017–1025.

Parent B, Bonneau J, Maphosa L, Kovalchuk A, Langridge P, Fleury D. 2017. Quantifying wheat sensitivities to environmental constraints to dissect genotype × environment interactions in the field. *Plant Physiology* **174**:1669–1682.

Prasad NR, Patel NR, Danodia A. 2021. Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research* **29**:195–206.

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramirez-Villegas J, Koehler A-K, Challinor AJ. 2017. Assessing uncertainty and complexity in regional-scale crop model simulations. *European Journal of Agronomy* **88**:84–95.

Rashid MA, Andersen MN, Wollenweber B, Zhang X, Olesen JE. 2018. Acclimation to higher VPD and temperature minimized negative effects on assimilation and grain yield of wheat. *Agricultural and Forest Meteorology* **248**:119–129.

Ratjen AM, Kage H. 2013. Is mutual shading a decisive factor for differences in overall canopy specific leaf area of winter wheat crops? *Field Crops Research* **149**:338–346.

Ratjen AM, Kage H. 2016. Nitrogen-limited light use efficiency in wheat crop simulators: comparing three model approaches. *The Journal of Agricultural Science* **154**:1090–1101.

Ratjen AM, Neukam D, Kage H. 2016. A simple drought-sensitive model for leaf: stem partitioning of wheat. *Journal of Agronomy and Crop Science* **202**:300–308.

Rosenzweig C, W Jones J, L Hatfield J, C Ruane A, J Boote K, Thorburn P, Antle JM, Nelson GC, Porter C, Janssen S, Asseng S, Basso B, Ewert F, Wallach D, Baigorria G, Winter JM. 2013. The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. *Agricultural and Forest Meteorology* **170**:166–182.

Seidel SJ, Palosuo T, Thorburn P, Wallach D. 2018. Towards improved calibration of crop models—where are we now and where should we go? *European Journal of Agronomy* **94**:25–35.

Shahhosseini M, Hu G, Archontoulis S. 2020 Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science* **11**:1120.

Shahhosseini M, Hu G, Huber I, Archontoulis SV. 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports* **11**:1–15.

Shahhosseini M, Martinez-Feria RA, Hu G, Archontoulis SV. 2019. Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters* **14**:124026.

Shimoda S, Sugikawa Y. 2020. Grain-filling response of winter wheat (*Triticum aestivum* L.) to post-anthesis shading in a humid climate. *Journal of Agronomy and Crop Science* **206**:90–100.

Soltani A, Torabi B, Zarei H. 2005. Modeling crop yield using a modified harvest index-based approach: application in chickpea. *Field Crops Research* **91**:273–285.

Srivastava AK, Safaei N, Khaki S, Lopez G, Zeng W, Ewert F, Gaiser T, Rahimi J. 2022. Winter wheat yield prediction using convolutional

neural networks from environmental and phenological data. *Scientific Reports* **12**(1):1–14.

Vogel E, Donat MG, Alexander LV, Meinshausen M, Ray DK, Karoly D, Meinshausen N, Frieler K. 2019. The effects of climate extremes on global agricultural yields. *Environmental Research Letters* **14**:054010.

Wallach D, Martre P, Liu B, Asseng S, Ewert F, Thorburn PJ, van Ittersum M, Aggarwal PK, Ahmed M, Basso B, Biernath CJ, Cammarano D, Challinor AJ, de Sanctis G, Dumont B, Eyshi Rezaei E, Fereres E, Fitzgerald GJ, Gao Y, Garcia-Vila M, Gayler S, Girousse C, Hoogenboom G, Horan H, Izaurralde RC, Jones CD, Kassie BT, Kersebaum KC, Klein C, Koehler A-K, Maiorano A, Minoli S, Müller C, Naresh Kumar S, Nendel C, O'Leary GJ, Palosuo T, Priesack E, Ripoche D, Rötter RP, Semenov MA, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Wolf J, Zhang Z. 2018. Multimodel ensembles improve predictions of crop–environment–management interactions. *Global Change Biology* **24**:5072–5083.

Wallach D, Palosuo T, Thorburn P, Hochman Z, Andrianasolo F, Asseng S, Basso B, Buis S, Crout N, Dumont B, Ferrise R, Gaiser T, Gayler S, Hiremath S, Hoek S, Horan H, Hoogenboom G, Huang M, Jabloun M, Jansson PE, Jing Q, Justes É, Kersebaum KC, Launay M, Lewan E, Luo Q, Maestrini B, Moriondo M, Olesen JE, Padovan G, Poyda A, Priesack E, Pullens JWM, Qian B, Schütze N, Shelia V, Souissi A, Specka X, Kumar Srivastava A, Stella T, Streck T, Trombi G, Wallor E, Wang J, Weber TKD, Weihermüller L, de Wit A, Wöhling T, Xiao L, Zhao C, Zhu Y, Seidel SJ. 2021a. Multi-model evaluation of phenology prediction for wheat in Australia. *Agricultural and Forest Meteorology* **298**:108289.

Wallach D, Palosuo T, Thorburn P, Hochman Z, Gourdain E, Andrianasolo F, Asseng S, Basso B, Buis S, Crout N, Dibari C, Dumont B, Ferrise R, Gaiser T, Garcia C, Gayler S, Ghahramani A, Hiremath S, Hoek S, Horan H, Hoogenboom G, Huang M, Jabloun M, Jansson PE, Jing Q, Justes E, Kersebaum KC, Klosterhalfen A, Launay M, Lewan E, Luo Q, Maestrini B, Mielenz H, Moriondo M, Zadeh HN, Padovan G, Olesen JE, Poyda A, Priesack E, Pullens JWM, Qian B, Schütze N, Shelia V, Souissi A, Specka X, Srivastava AK, Stella T, Streck T, Trombi G, Wallor E, Wang J, Weber TKD, Weihermüller L, Wit A de, Wöhling T, Xiao L, Zhao C, Zhu Y, Seidel SJ. 2021b. The chaos in calibrating crop models: lessons learned from a multi-model calibration exercise. *Environmental Modelling & Software* **145**:105206.

Wang E, Engel T. 1998. Simulation of phenological development of wheat crops. *Agricultural Systems* **58**:1–24.

Wang E, Martre P, Zhao Z, Ewert F, Maiorano A, Rötter RP, Kimball BA, Ottman MJ, Wall GW, White JW, Reynolds MP, Alderman PD, Aggarwal PK, Anothai J, Basso B, Biernath C, Cammarano D, Challinor AJ, De Sanctis G, Doltra J, Dumont B, Fereres E, Garcia-Vila M, Gayler S, Hoogenboom G, Hunt LA, Izaurralde RC, Jabloun M, Jones CD, Kersebaum KC, Koehler AK, Liu L, Müller C, Naresh Kumar S, Nendel C, O'Leary G, Olesen JE, Palosuo T, Priesack E, Eyshi Rezaei E, Ripoche D, Ruane AC, Semenov MA, Shcherbak I, Stöckle CO, Stratonovitch P, Streck T, Supit I, Tao F, Thorburn P, Waha K, Wallach D, Wang Z, Wolf J, Zhu Y, Asseng S. 2017. The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nature Plants* **3**:1–13.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Wright MN, Ziegler A. 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* **77**:1–17.

Yin X, Goudriaan JAN, Lantinga EA, Vos JAN, Spiertz HJ. 2003. A flexible sigmoid function of determinate growth. *Annals of Botany* **91**:361–371.

Zhang M, Gao Y, Zhang Y, Fischer T, Zhao Z, Zhou X, Wang Z, Wang E. 2020. The contribution of spike photosynthesis to wheat yield needs to be considered in process-based crop models. *Field Crops Research* **257**:107931.

Zhao J, Pu F, Li Y, Xu J, Li N, Zhang Y, Guo J, Pan Z. 2017. Assessing the combined effects of climatic factors on spring wheat phenophase and grain yield in Inner Mongolia, China. *PLoS One* **12**:e0185690.