European Journal of Agronomy 168 (2025) 127659

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Evaluating the AgMIP calibration protocol for crop models; case study and new diagnostic tests

Daniel Wallach [a,1], Kwang Soo Kim [b], Shinwoo Hyun [b], Samuel Buis [c], Peter Thorburn [d], Henrike Mielenz [e], Sabine Julia Seidel [a,f,*], Phillip D. Alderman [g], Benjamin Dumont [h], Mohammad Hassan Fallah [i], Gerrit Hoogenboom [j,k], Eric Justes [l], Kurt-Christian Kersebaum [m,n,o], Marie Launay [p], Luisa Leolini [q], Muhammad Zeeshan Mehmood [g], Marco Moriondo [r], Qi Jing [s], Budong Qian [s], Schulz Susanne [m], Diana-Maria Seserman [m], Vakhtang Shelia [j,k], Lutz Weihermüller [t], Taru Palosuo [u]

[a] Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany
[b] Department of Agriculture, Forestry, and Bioresources, Seoul National University, South Korea
[c] INRAE, UMR 1114 EMMAH, Avignon, France
[d] CSIRO Agriculture and Food, Brisbane, Queensland, Australia
[e] Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Crop and Soil Science, Braunschweig, Germany
[f] Institute of Organic Farming, Department of Agricultural Sciences, University of Natural Resources and Life Sciences, Vienna 1180, Austria
[g] Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK, USA
[h] Plant Sciences & TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium
[i] Department of Agrotechnology, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran
[j] Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL, USA
[k] Global Food Systems Institute, University of Florida, Gainesville, FL, USA
[l] CIRAD, Persyst Department, Montpellier, France
[m] Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany
[n] Global Change Research Institute CAS, Brno, Czech Republic
[o] Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Dep.of Crop Sciences, Georg-August-University of Göttingen, Göttingen, Germany
[p] INRAE, US 1116 AgroClim, Avignon, France
[q] Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Piazzale delle Cascine 18, Florence 50144, Italy
[r] CNR-IBE, Firenze, Italy
[s] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada
[t] Institute of Bio, and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
[u] Natural Resources Institute Finland (Luke), Helsinki, Finland

ARTICLE INFO

Keywords:
Crop model
Weighted least squares
Forward regression
Parameter selection
AICc
Feedback
Diagnostic tools
Squared bias

ABSTRACT

Crop simulation models are important tools in agronomy. Typically, they need to be calibrated before being used for new environments or cultivars. However, there is a large variability in calibration approaches, which contributes to uncertainty in simulated values, so it is important to develop improved calibration procedures that are widely applicable. The AgMIP calibration group recently proposed a comprehensive, generic calibration protocol that is directly based on standard statistical parameter estimation in regression models. Weighted least squares (WLS) is used to handle multiple response variables and forward regression using the corrected Akaike Information Criterion (AICc) is used to select the parameters to be calibrated. The protocol includes two adaptations, which are specific to each model and data set. First, initial approximations to the WLS parameters are obtained by fitting variables one group at a time. Secondly, "major" parameters are identified that are intended to reduce bias, analogously to the constant in linear regression. In this study, new diagnostic tools to be included in the

* Corresponding author at: Institute of Organic Farming, Department of Agricultural Sciences, University of Natural Resources and Life Sciences, 1180 Vienna, Austria.
E-mail addresses: dwallach@uni-bonn.de (D. Wallach), sabine.seidel@uni-bonn.de (S.J. Seidel).
[1] ORCID 0000–0003-3500–8179

https://doi.org/10.1016/j.eja.2025.127659
Received 21 December 2024; Received in revised form 18 April 2025; Accepted 19 April 2025
Available online 5 May 2025

protocol are proposed and tested in a case study. The diagnostics test whether the protocol does indeed lead to good initial approximations to the WLS parameters, and whether the protocol does indeed substantially reduce bias. These diagnostics provide in-depth understanding of the calibration process, reveal problems and help suggest solutions. The diagnostics should increase confidence in the results of the protocol. Having a reliable, generic calibration approach, like the augmented AgMIP protocol, is essential to using crop models more effectively.

1. Introduction

Crop simulation models are widely used in agronomic research to screen management options, to estimate the impact of climate change, to better understand the complex interactions in the soil-plant-atmosphere system, and to address a wide array of agricultural challenges and environmental perturbations (Asseng et al., 2014; Boote, 2019; Boote et al., 2010). The models are built on theoretical understanding of the interacting processes within the studied system, but their use requires numerical parameter values that require model calibration. Typically, crop model simulation studies begin with a data set representative of some "target" population and the model is fitted to those data by calibration. Then the calibrated model is used for further scenario analyses and risk assessment. The calibration step is necessary because model parameters are not constants of nature but rather dependent on the approximations in the model in interaction with a specific set of environments (Fath and Jorgensen, 2011; Janssen and Heuberger, 1995; Wallach, 2011). This also means that the calibration (or more precisely the parameters calibrated) can have a major impact on simulated values, and thus, on the predictive quality of crop models. Therefore, effective calibration is essential for robust predictions (Grassini et al., 2015).

Calibration of crop models is complex and time-consuming (Seidel et al., 2018), with no consensus on the best procedure so far (Ahuja and Ma, 2011; Seidel et al., 2018; Wallach et al., 2021b). Calibration can be also numerically demanding, because of model discontinuities (Liu et al., 2018) and local optima (César Trejo Zúñiga et al., 2014). Several studies have examined specific aspects of calibration, such as the way to conduct sensitivity analysis in order to choose the most sensitive parameters to estimate (Richter et al., 2010; Wang et al., 2023), different ways of searching for the parameter values that minimize the sum of squared errors (Harrison et al., 2019; Jha et al., 2021), or Bayesian parameter estimation (Gao et al., 2021; Van Oijen et al., 2005). There have also been studies for specific crop models, with recommendations as to how to calibrate those models (Adnan et al., 2019; Berton Ferreira et al., 2024; Kersebaum, 2011). All of these studies have concerned a specific crop model and/or specific model parameters rather than the overall calibration activity.

Recently, the AgMIP calibration group (https://agmip.org/crop-model-calibration-3/#) proposed for the first time a calibration protocol that is comprehensive, i.e. it considers all the steps involved in crop model calibration and generic, i.e.it is applicable to essentially any crop model and data set. A simplified version of the protocol was tested for phenology in a multi-model study and found to give, on the average, better predictions and less variability than usual calibration approaches (Wallach et al., 2023). The full protocol was also tested using the STICS model in combination with artificial data, which allowed for exact performance evaluation. The results showed that the protocol was easily applicable and avoided over-fitting, i.e. goodness-of-fit and errors in predicting for new environments were similar (Wallach et al., 2024). Of course the overall results depend on the model and the data set as well as on the calibration approach. To evaluate the protocol specifically, one needs to look in depth at the calibration procedure.

The AgMIP protocol contains two major innovations compared to usual practice. The first concerns the way to handle multiple observed variables, for example days to anthesis, biomass and yield. A common way of dealing with this problem in crop model calibration is to calibrate parameters using only one or only a few types of variable at a time (Hlavinka et al., 2013; Winn et al., 2023; Wolf et al., 1996). One defines an order for considering the variables, and then performs the calibration sequentially according to the predefined order. This has the important advantage that rather than considering all important model parameters at the same time, one only needs to consider the subset of parameters related to the variable being treated, which reduces numerical problems. A difficulty is that, due to the interconnections of processes within the calibrated model, fitting parameters to a variable may affect (degrade) the fit to variables considered earlier in the order. Due to such feedbacks, at the end of the procedure there may be a poor fit to those earlier variables (Guillaume et al., 2011). To mitigate this problem, in some cases one adds a final step in the calibration, where all parameters are allowed to vary. The goal is to find parameter values that give more acceptable overall results (Hlavinka et al., 2013; Winn et al., 2023). However, this is generally done in an ad hoc manner, without an explicit overall objective function. In some cases, with relatively few variables, all variables are treated together, but weights are not generally explicitly based on statistical theory (Guo et al., 2021).

A standard way to handle multiple variables in regression is weighted least squares (WLS). In this approach, the overall objective function is a sum of weighted squared errors, where each variable is weighted by the inverse of the standard deviation of model error. This can be done in two stages. One begins with some approximation to the parameter values, perhaps based on ordinary least squares (OLS). Based on the resulting fit to the model one estimates the standard deviation of model error for each variable, and that provides the weights for the final WLS step (Seber and Wild, 1989). The AgMIP calibration protocol follows the above statistical approach, but uses usual crop model calibration practice to obtain the initial estimate of optimal parameter values. Specifically, according to the protocol, one first treats one variable group at a time to obtain initial estimates of optimal parameter values. Then, those initial estimates are used for simulation, the resulting model error standard deviations are calculated, and are used as weights in the final WLS parameter estimation.

The second innovation concerns the choice of which parameters to estimate. Crop models in general have many parameters. In order both to simplify the calculations and to avoid over-fitting, only a small fraction of the parameters is fit to data. Several different approaches for choosing the parameters to estimate are commonly used for crop models (Wallach et al., 2021b). One is to perform a sensitivity analysis to identify the most sensitive parameters and calibrate them first (Zhao et al., 2014). Another is to identify the most important model parameters based on expert knowledge, independently of the available data (He et al., 2017). These might be the parameters identified as being cultivar dependent (Jha et al., 2021). Sensitivity analysis and model expertise can be combined, and might also be combined with testing various parameters in order to identify the parameters that give the greatest improvement to the fit (Wallach et al., 2021b). None of these approaches is specifically designed to avoid over-fitting.

The problem of choosing which parameters to estimate is very similar to choosing the terms and associated parameters to include in a linear regression model, which is a well-studied problem in statistics. There are several standard statistical methods for subset selection in regression models. The order of terms to consider can be based on forward, stepwise, or backward regression. The decision as to whether to include a term can be based on a likelihood ratio test, Bayesian

Information Criterion (BIC), he Akaike Information Criterion (AIC) or the version corrected for small sample size (AICc) (Claeskens and Hjort, 2001; Morozova et al., 2015). Normally, one automatically includes an intercept, i.e. a constant term, which ensures that the model bias is zero (Turgal and Doganay, 2017). The same methods can be applied to nonlinear models like crop models. In nonlinear models, for each parameter considered, one compares the model where that parameter is fitted to the data with the model where that parameter has its default value. If a parameter is not accepted, then it is set to its default value and not to zero. The AgMIP calibration protocol again adopts a standard statistical approach, but adapted to crop models. The choice of parameters to estimate is done separately for each variable group. For each group, a strictly limited number of "major" parameters are identified, which are to be estimated from the data. These parameters should be chosen with the main goal of reducing bias (like the constant term in linear regression). In addition, a list of "candidate" parameters is defined. Those parameters are considered in turn, using the equivalent of forward regression. Each additional candidate is tentatively added to the list of parameters to estimate. If that leads to a decrease in the AICc criterion, the parameter is accepted for estimation. If not, it is fixed at its default value.

Thus, the protocol is directly based on standard statistical approaches, but the implementation also includes choices, which depend on expert knowledge of the model. In particular, the order of fitting the variable groups and the choice of major parameters depend on the specific model and data set. The statistical basis can be considered reliable, but it is important to evaluate the model-specific choices. The purpose of this study is to propose diagnostics for examining the crop model-specific choices made when implementing the AgMIP calibration protocol. These diagnostics provide new insights into the calibration procedure, and should increase confidence in the results. If problems are identified, the diagnostics can indicate alternative choices. The application and interpretation of the diagnostics is illustrated with a case study. Note that this is the first application of the protocol to real (as opposed to artificial) data.

2. Materials and methods

2.1. Data

The experimental data were obtained from multi-year variety trials that were conducted by Arvalis Institut du vegetal, Paris at multiple locations in France. The data here are for a typical winter wheat variety used in France. Eight different variables were measured (Table 1). Date of maturity was not directly observed but rather estimated as 15 calendar days before harvest date (ARVALIS, 2022). The full data set has data from 22 site-years (hereafter environments). A subset of fourteen environments (six different sites, five different years) was used for calibration. The subset with the other eight environments (five different sites, two different years) was used for evaluation. All data are representative of conventionally managed wheat fields in the major wheat growing regions of France, with effective weed, pest and disease control, under current climate (this is the "target population"). The calibration and evaluation subsets had no sites or years in common. Thus, the evaluation of the calibrated model is a rigorous test of how well the model can predict behavior for environments of the target population independent of those used for calibration. Further details about the experiments and the environments can be found in Wallach et al. (2021a).

2.2. The DSSAT-NWheat model

The N-Wheat model, incorporated within the DSSAT Cropping System Model (Hoogenboom et al., 2019), is widely used to simulate the growth and development of specific wheat cultivars under diverse environmental and management conditions (Kassie et al., 2016). This model serves as a scientific and practical tool for the evaluation of agricultural practices and optimization of resource use and the assessment of climate change impacts on crop production (Fallah et al., 2020; Jing et al., 2021). The model requires daily weather data and detailed soil profile data, such as texture, organic matter content, and hydraulic properties as input. Furthermore, crop management practices, including planting dates, irrigation schedules, and fertilizer applications are incorporated to predict key agronomic outcomes, such as biomass and yield, for a given growing season. In particular, the N-Wheat model can estimate the nitrogen content of grains, which influences grain quality. For instance, nitrogen uptake is determined by the interaction between soil nitrogen availability and crop nitrogen demand (see Supplementary Information).

2.3. The calibration protocol

The calibration protocol is described in detail in Wallach et al. ((2024). The protocol involves 8 steps. Steps 1–5 translate the generic recommendations of the protocol into specific variables and parameters of the model and require no calculations. Steps 6–8 concern the calculations. Once steps 1–5 are done, steps 6–8 can be done automatically without further modeler input. The steps of the protocol are described briefly below.

1. Select the default values for all parameters, explain and document the choice. This is important, since in general the large majority of parameters will not be fit to the data, but rather will retain their default values.
2. List the observed variables together with the corresponding simulated variables if any. The recommendation is to use in the calibration all the observed variables that have corresponding simulated variables.
3. Group the variables and order the groups. All phenological variables are in the phenology group. Multiple measurements over time of the same variable (such as biomass in this study) are together in the same group. Other variables are in a group by themselves. The order

Table 1
Overview of observed data with number of data points for each variable that were used for calibration and evaluation.

variable	explanation	number of observations in calibration subset	number of observations in evaluation subset
BBCH30	days after sowing to start of stem elongation (development stage BBCH30)	14	8
BBCH55	days after sowing to start of heading (development stage BBCH55)	14	8
BBCH90	days after sowing to physiological maturity (development stage BBCH90)	14	8
biomass	aboveground biomass dry matter at various dates (g/m^2)	44	31
biomass N	% N in total aboveground biomass at harvest (%)	9	5
grain number	grains/m^2 (number/1000)	13	7
yield	dry grain yield (g/m^2)	13	7
protein	protein in grain (%)	13	6

should be such that fitting later variables in the order will have little or no effect on the fit to earlier variables in the order. Usually, phenology is the first group, since fitting other variables in general has little or no effect on the fit to the phenology data.

4. Select the major parameters for each group. There is a strict maximum for the number of major parameters per group. The phenology group can have as many major parameters as observed development stages. Groups with multiple measurements over time can have two major parameters, which either affect the variable at different periods or affect the average value and the rate of change over time. All other groups have only one variable, and can have at most one major parameter. As far as possible, the major parameters should be parameters that have nearly the same effect in all environments, i.e. that act like additive constants. The main role of the major parameters is to reduce bias. For example, for phenology, the major parameters will usually be degree days to different development stages. In general, a parameter that acts like a constant term in the model, i.e. has a similar effect in all environments, will nearly eliminate bias (see Supplementary equations S5-S8 for the demonstration that a constant term eliminates bias).
5. Select the candidate parameters for each group. The candidate parameters are those parameters that are expected to explain the variability between environments. The candidate parameters should be ordered as far as one knows by importance. There is no strict limit on their number, but increasing the number increases computing time and the risk of choosing unimportant parameters by chance.
6. Choose the parameters to estimate and do the estimation, for each variable group in turn. For each group, first estimate the major parameters using ordinary least squares (OLS), i.e. minimize the sum of squared errors. Here, and throughout the protocol, biomass values are replaced by ln(biomass). The major parameter or parameters for a variable group constitute the initial list of parameters to estimate. Then each candidate parameter is considered in turn. The candidate is added tentatively to the list of parameters to estimate, and all the parameters in the list are estimated using OLS. If the result is a decrease in the corrected Akaike Information Criterion (AICc), the candidate is added definitively to the list of parameters to estimate. If not, it is fixed at its default value and one goes on to the next candidate. *AICc* is calculated as:

$$AICc = n\ln(SS/n) + 2p + \frac{2p(p+1)}{n-p-1} \quad (1)$$

where *SS* is the sum of squared errors for all variables in the group, n is the number of data points and p the number of estimated parameters.

7. Calculate the weighted least squares (WLS) parameter values. The overall objective function is defined as:

$$J = \sum \left(\frac{y_{ij} - \hat{y}_{ij}}{w_i}\right)^2 \quad (2)$$

where y_{ij} is the observed value of observation j of variable group i, $\hat{y}_{ij}$ is the corresponding simulated value, and w_i is the weight for group i. The weight for group i is

$$w_i = \sqrt{SS/(n-p)} \quad (3)$$

8. Evaluate goodness-of-fit and estimate prediction error. Many different metrics and graphs can be used to judge the goodness of the fit. Prediction error can be estimated using data splitting or cross validation. The metrics used here are defined below.

2.4. Metrics for goodness-of-fit and for prediction error

Common measures of goodness-of-fit are mean squared error (MSE), root mean squared error (RMSE), relative root mean squared error (RRMSE) and Nash-Sutcliffe Efficiency (NSE), defined as

$$MSE = (1/n)\sum (y_i - \hat{y}_i)^2 \quad (4)$$

$$RMSE = \sqrt{MSE} \quad (5)$$

$$RRMSE = \frac{RMSE}{\bar{y}} \quad (6)$$

$$NSE = 1 - MSE/MSE_{\bar{y}} \quad (7)$$

where n is the number of observed values, y_i is the i^{th} observed value, $\hat{y}_i$ is the corresponding simulated value, $\bar{y}$ is the average over the observed values and $MSE_{\bar{y}}$ is the MSE value when the predictor is the average of the observed values. It is MSE applied to the calibration data that is minimized in the OLS calculations. The above metrics can also be applied to the evaluation data, to evaluate how well the model predicts for new environments. When the values of the metrics are presented below, we will specify whether they refer to goodness-of-fit, i.e. to the calibration data, or to prediction error, i.e. to the evaluation data. We will also specify what data have been used for calibration of the model.

We also introduce one additional metric, called skill, specifically for evaluating prediction error. Skill measures, often used in climate modeling, compare the prediction error of a model with the error of a simple, "naive" predictor (Hargreaves, 2010). Here, the naive predictor is the mean of the observations in the calibration data. The equation for the skill measure is

$$skill = 1 - MSE/MSE_{\bar{y}calibration} \quad (8)$$

MSE is the mean squared error of the model applied to the evaluation data. $MSE_{\bar{y}calibration}$ is mean squared error for predicting the evaluation data when the predictor is the average of the observed calibration data. The skill metric resembles NSE, but is fundamentally different. NSE applied to the evaluation data compares the model to the average of the evaluation data, which of course one does not know before actually doing the measurements for the evaluation environments. The skill measure on the other hand compares the model with the average of the calibration data, which is known. Thus, the skill measure compares the model with a simple predictor, which could be used in practice. A negative skill value indicates that one would obtain better predictions using the average of the calibration data than by using the model.

MSE can be decomposed into three terms, namely squared bias ($bias^2$), squared differences in standard deviations (SDSD) and lack of correlation weighted by the standard deviations (LCS) (Kobayashi and Salam, 2000).

$$MSE = bias^2 + SDSD + LCS \quad (9)$$

$$\text{bias}^2 = \left[\left(1/n\right)\sum y_i - (1/n)\sum \hat{y}_i\right]^2 \quad (10)$$

$$SDSD = (\sigma_y - \sigma_s)^2(n-1)/n \quad (11)$$

$$LCS = 2\sigma_y\sigma_s(1-r)(n-1)/n \quad (12)$$

where σ_y and σ_s are respectively the sample standard deviations of the observed and simulated values and r is the correlation coefficient of observed and simulated values.

2.5. New diagnostics

After fitting parameters to one variable group at a time, the

estimated parameters are supposed to be a good first approximation to the WLS parameters. The first two diagnostics concern this assumption. A first diagnostic here concerns the extent of feedbacks between variable groups. If feedbacks are important, then some variable groups will be poorly fit after treating all the variable groups, so the estimated parameters will likely not be a good approximation to the overall best (WLS) parameters. The proposed diagnostic here is to plot the simulated values of each variable after each variable group has been used for calibration (see Fig. 2 for an example). If there are no feedbacks, then the simulated values for a variable will not change after the corresponding variable group has been used for calibration. If feedbacks are found to be important, one conclusion might be that the chosen order of treating the variable groups should be changed.

Suppose one has a good first approximation to the WLS parameters. Then in searching for the best WLS parameter values, it should suffice to explore the parameter space in the vicinity of the first approximation. This is important, since often, there will be a large number of parameters to estimate at the WLS stage, so searching the full parameter space would be numerically difficult. According to the protocol, the search for the WLS parameter values has multiple starting points, including a starting point at the first approximation to the WLS parameters. The second diagnostic is to examine the results for different starting points. If the first approximation is good, that starting point should lead to the smallest value of the criterion J (Eq. (2)). If there are other starting points that give a much smaller value of J than the first approximation, then the first priority should be to improve the first approximation (based on the other diagnostics). An alternative would be to improve the search of the parameter space, to obtain more confidence in the WLS parameters.

The second pair of diagnostics is related to the choice of "major" parameters, which might better be called "bias-reducing" parameters. The assumption in the protocol is that, for many variables, squared bias makes a large contribution to MSE and that one can reduce that contribution substantially by fitting the major parameter to the data. The first diagnostic here is to examine squared bias as a fraction of MSE for the default parameter values. It is expected that for most variables, the squared bias contribution will be large. If the squared bias contribution is very small for some variable, there is no point in identifying a major parameter for that variable.

The second diagnostic tool here is to examine squared bias for each variable before and after the major parameter associated with that variable has been fit to the data. One expects a substantial reduction in squared bias. If there is a negligible reduction in squared bias, one must reconsider the choice of major parameter.

2.6. Implementation of the protocol

The modeling group working with DSSAT-NWheat (who are co-authors of this study) followed the above protocol and first filled out the protocol documentation tables for protocol steps 1–5. All calculations were then performed automatically using R scripts based on the use of the R packages CroptimizR (Buis et al., 2023), version 0.7.0, and CroPlotR (Vezy et al., 2023), version 0.10.0, as described in Wallach et al. (2024). All the necessary R scripts, functions, and data for applying the protocol to the datasets used here are freely available and documented on GitHub at https://github.com/sbuis/AgMIP-Calibration-Phase-IV. A generic implementation of the protocol, applicable to any dataset, will be fully integrated into the next version of CroptimizR.

The use of CroptimizR required writing an R wrapper function for DSSAT-NWheat. The wrapper handles the communication between CroptimizR and the crop model, enables the transfer of parameter values to the model and retrieves the required simulated outputs. The optimization algorithm used by CroptimizR is the Nelder-Mead simplex, which is a powerful, robust algorithm for parameter estimation in nonlinear models (Kumar, 2023).

3. Results

3.1. Application of protocol to the DSSAT-NWheat

Parameter values for the winter wheat variety Gamenya within the N-Wheat model were used as the defaults (step 1). This variety seems to be close to the variety used in the experiments in terms of maturity group.

All of the observed variables except ears/m^2 have simulated equivalents in the DSSAT-NWheat model (Table 2) and were used for calibration (step 2). The variables were grouped as prescribed by the protocol, and the order of the groups was chosen to respect the protocol prescription that fitting later variable groups in the order should have a minor or no effect on the fit to earlier groups (step 3).

The major parameters for each variable group are shown in Table 3 (protocol step 4). These are parameters expected to have a similar effect in all environments, based on expert opinion. The candidate parameters for each group, again based on expert opinion, are shown in Table 4 (protocol step 5). That table also shows which candidate parameters reduce AICc and are therefore chosen for estimation (protocol step 6). Of the ten candidate parameters, four were chosen for estimation. The default values of the parameters and the values after protocol steps 6 and 7 are shown in Table 5.

Only a selection of the metrics from step 8 of the protocol are shown. Plots of simulated versus observed values for each variable group are shown in Fig. 1. The RRMSE values for the calibration data, for each simulated variable and after each step of the protocol, are shown in Table 6. For the default parameter values, RRMSE ranges from 9 % to 77 % depending on the simulated variable. At the end of the protocol (after step 7) RRMSE is very small for the phenology variables and for ln (biomass) (maximum 4.3 %) but larger for final biomass, biomass N, grain number, yield and protein content (14–25 %). The RRMSE values for the evaluation data similarly have small values for phenology and ln (biomass), and larger values for the other variables. The similar levels of error for the calibration and evaluation data show that there is no evidence of over-fitting. The skill values obtained show high skill (near 1) for phenology and ln(biomass) and quite high skill for grain number but negative skill for the other variables (Fig. 2, Tables 7 and 8)

3.2. Diagnostics

The first diagnostic tool is a set of graphs to visualize the extent of feedbacks. Fig. 2 shows the simulated values for each variable for each calibration environment using the default parameter values and then after fitting each variable group. Only final biomass is shown for the biomass group. If there are no feedbacks, the simulated values of variables in a variable group will not change after that variable group was used for fitting. This is largely the case. For example, consider the three phenology variables BBCH30, BBCH55 and BBCH90. The simulated values change appreciably in going from step "def" (simulation using the

Table 2
Observed variables and the corresponding simulated variables (protocol step 2). The variables are combined into groups. The order in which the groups are used for parameter estimation is indicated in the last column (protocol step 3).

Observed variable	Name of the simulated variable	Calibration group	order for calibration
BBCH30	Date_BBCH30	phenology	1
BBCH55	Date_BBCH55	phenology	1
BBCH90	Date_BBCH90	phenology	1
biomass	Biomass (all dates)	ln(biomass)	2
biomass N	N_in_biomassHarvest	biomass N	3
grain number	Grain_Number	grain Number	4
yield	Grain_Yield	yield	5
protein	ProteinContentGrain	protein	6
ears number	none	-	-

Table 3
Major parameters for each variable group (protocol step 4). Default values and minimum and maximum values for each parameter are shown.

Group	Parameter	Short explanation	Default value (min, max)
phenology	P1	Thermal time from seedling emergence to the end of the juvenile phase	400 (300,600)
	PHINT	Phyllochron interval	110 (70,120)
	P5	Thermal time (base 0 oC) from beginning of grain fill to maturity	700 (200,800)
ln(biomass)	RUE1	Pre-anthesis radiation use efficiency, g plant dry matter/MJ PAR	3.8 (3,5)
	RUE2	Post-anthesis radiation use efficiency, g plant dry matter/MJ PAR	3.8 (3,5)
biomass N	MXNUP	max N uptake per day	0.6 (0.4,1)
grain number	GRNO	Coefficient of kernel number per stem weight at the beginning of grain filling	22 (20,30)
yield	MXFIL	Potential kernel growth rate	1.9 (1,3)
protein	INGNC	% protein, initial grain N conc	0.03 (0.01,0.04)

default parameter values) to step "phe" (simulation with parameters estimated using the phenology data). In subsequent steps, where parameters are estimated using the biomass data, the biomass N data, the grain number data, the yield data and the protein data, simulated phenology does not change at all. That is, estimating parameters related to those later steps has no effect on the simulated phenology. The only evidence of feedbacks is that the simulated values of final biomass and of biomass N do change somewhat after parameters are fitted to yield. The table of RRMSE values (Table 6) tells much the same story. The RRMSE values in general do not change after the group of a variable is used for calibration, with minor exceptions in the case of final biomass and biomass N. Overall, it seems that the order of fitting the variable groups is acceptable.

The second diagnostic concerns the search for the WLS parameter values. In searching for the best WLS parameters here, the simplex algorithm was started at the parameter values after step 6, and at 19 additional starting values chosen by Latin hypercube sampling covering the full parameter space. The smallest value of the weighted objective function was found when starting from the results of step 6 (detailed results not shown). The conclusion is that there is no evidence that the parameter values obtained at the end of protocol step 6 are a poor first approximation to the WLS parameters.

The third diagnostic is the fraction of MSE for the calibration data due to squared bias for the default parameter values and for the parameter values after protocol step 7. Table 9 shows that for the default parameter values, squared bias represents 0.11–0.94 of MSE. The fraction is over one half for seven of the nine variables. The fraction of MSE due to squared bias after step 7 is below 0.20 for all variables except final biomass, which is not a specific target of calibration (it is included with ln(biomass)). Squared bias is clearly initially an important part of overall error, and that contribution is substantially reduced by the protocol.

The final diagnostic tool is examination of the effectiveness of the major parameters in reducing squared bias. Table 10 shows the values of squared bias after each variable group has been used for calibration. For the groups where candidate parameters are accepted (phenology, ln (biomass), yield), the table also shows squared bias after estimating just the major parameters for that group. For each response variable, the difference between squared bias after estimating the major parameters for the corresponding group and squared bias just before that, measures how effective the major parameters are in reducing squared bias. The major parameters reduce squared bias to less than 4 % of its previous

Table 5
Values of parameters at various stages of the protocol. For each parameter, the table shows the default value, the value after fitting each variable group separately in protocol step 6, and the final WLS value (after protocol step 7).

Group	Parameter	Default value	Value after step 6	Value after step 7
phenology	P1	400	463.03	462.45
phenology	PHINT	110	118.13	118.08
phenology	P5	700	676.76	677.65
phenology	VSEN	1.6	4.98	4.98
ln(biomass)	RUE1	3.8	4.35	4.35
ln(biomass)	RUE2	3.8	3.02	3.04
ln(biomass)	PLGP1	1400	2000	1994.31
ln(biomass)	SLAP1	280	400	399.98
biomass N	MXNUP	0.6	0.4	0.4
grain number	GRNO	22	26.76	26.87
yield	MXFIL	1.9	1.71	1.71
yield	STEMN	0	0.088	0.15
protein	INGNC	0.03	0.033	0.034

Table 4
The candidate parameters for each variable group (protocol step 5). Default values and minimum and maximum values for each parameter are shown. A candidate parameter is chosen for estimation during the optimization process if it leads to a reduction in the AICc criterion (as indicated by "Y" in the last column).

Group	Parameter	Default value (min,max)	Short explanation	Reduces AICc
phenology	VSEN	1.6 (1,5)	sensitivity to vernalisation	Y
	PPSEN	2 (1,5)	sensitivity to photoperiod	N
ln(biomass)	STMMX	1.5 (1,5)	Potential final dry weight of a single tiller (excluding grain)	N
	P2AF	0.6 (0.4,0.8)	threshold AD in a layer becoming effective on root growth	N
	ADLAI	1 (0.5,1)	threshold aeration deficit (AF2) affecting LAI (set to 1.0 for no stress run)	N
	PLGP1	1400 (1000,2000)	for calculating plag: potential leaf growth. plag= plag_p1 *cumph(istage)* *plag_p2	Y
	PLGP2	0.6 (0.4,0.8)	for calculating plag: potential leaf growth. plag= plag_p1 *cumph(istage)* *plag_p2	N
	SLAP1	280 (250,400)	ratio of leaf area to mass at emergence (cm2/g)	Y
	SLAP2	270 (200,300)	ratio of leaf area to mass at end of leaf growth (cm2/g)	N
yield	STEMN	0 (0,1)	0 =original C to grain translocation, > 0–1.0 sets % of C of stem to be transloc. to grain	Y

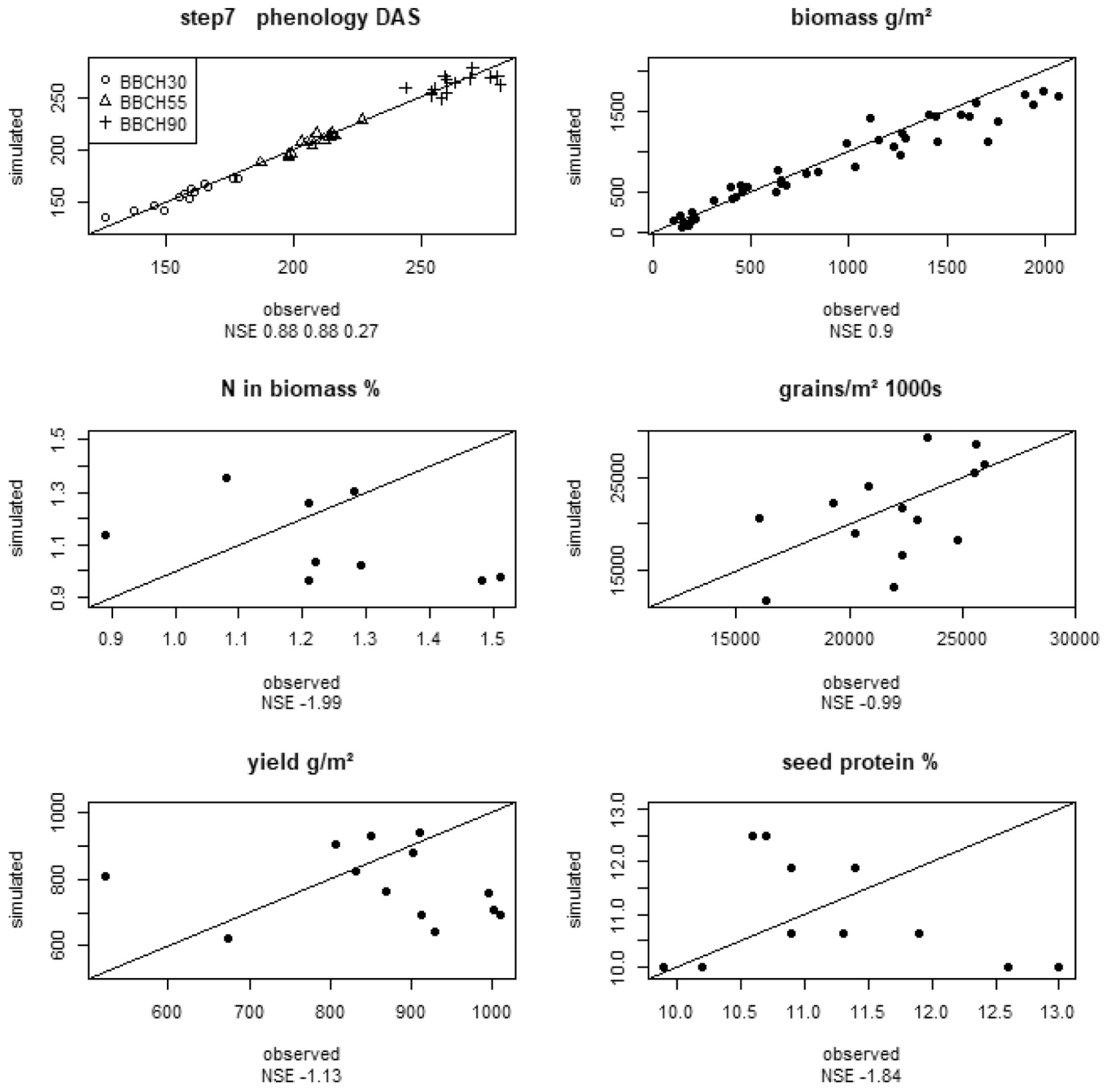

Fig. 1. Goodness-of-fit for each variable group. The observed values are the calibration data. The simulated values are the corresponding results of the model after the full calibration procedure (after protocol step 7). It is biomass and not ln(biomass) that is shown. In the graph of phenology, DAS is days after sowing. For each variable, the Nash Sutcliffe Efficiency (NSE) value is shown. For phenology, the three NSE values refer to stages BBCH30, BBCH55 and BBCH90 in that order.

value for the three phenology variables and for grain number. We conclude that for those variables there are parameters that substantially reduce bias and those parameters were correctly identified. The squared bias contribution is reduced to about 20 % of its previous value for the ln (biomass) group (about 30 % for final biomass) after estimating the major parameters for biomass. Here, the bias reduction is again quite effective. On the other hand, the protein major parameter only reduces squared bias for protein to about 70 % of its previous value. The biomass N major parameter actually increases squared bias for biomass N. This suggests that one should make a different choice of major parameter for biomass N, or perhaps choose not to have a major parameter. The squared bias for yield is hardly affected by fitting the major parameter for yield, but is then very substantially reduced by fitting the candidate parameter. It seems that the major and candidate parameters should be exchanged.

4. Discussion

4.1. Overall results

A very important property of the protocol is that it greatly simplifies the overall calibration activity, which is often very time-consuming (Seidel et al., 2018) and requires model expertise throughout. A major simplifying feature is that the protocol clearly separates the model expertise steps from the calculation steps. Once the tables that summarize the model expertise are finalized, the calculations can all be

Table 6
RRMSE values for the calibration data, for each simulated variable at each calibration stage. The table shows results using default parameter values, parameter values after using each variable group separately for calibration, and after using all variable groups together for estimating all parameters, i.e. after protocol step 7. The value in bold in each column is the RRMSE value after fitting the associated variable group. Final biomass i.e. biomass at maturity is included in the ln(biomass) group, but results are also shown specifically for this variable.

Calibration step	BBCH30	BBCH55	BBCH90	ln(biomass)	Final biomass	Biomass N	Grain number	Yield	Protein
default	0.35	0.155	0.088	0.182	0.77	0.18	0.74	0.7	0.2
phenology	**0.03**	**0.016**	**0.034**	0.06	0.18	0.26	0.3	0.4	0.28
ln(biomass)	0.03	0.016	0.034	**0.043**	0.15	0.3	0.27	0.42	0.31
biomass N	0.03	0.016	0.034	0.042	**0.15**	**0.3**	0.27	0.41	0.31
grain number	0.03	0.016	0.034	0.042	0.15	0.3	**0.2**	0.41	0.36
yield	0.03	0.016	0.034	0.043	0.20	0.24	0.2	**0.22**	0.14
protein (end of step 6)	0.03	0.016	0.034	0.043	0.20	0.24	0.2	0.22	**0.14**
Calibration step7	0.03	0.016	0.034	0.043	0.19	0.25	0.2	0.22	0.14

automated, as was done here using scripts based on the CroptimizR software (Buis et al., 2023).

The protocol is intended to be applicable to essentially all models and data sets. It was previously applied to the STICS model with an artificial data set (Wallach et al., 2024). Here, it was applied to DSSAT-NWheat, another widely used crop model, using real data. The application to a new model structure was straightforward. In a previous study a simplified version of the protocol, which uses only phenology data for calibration, was used by 16 different model structures in a multi-model exercise (Wallach et al., 2023). This further illustrates the genericity with respect to model structure. The genericity can also be seen from the formulation of the protocol, which is clearly not model specific. In order to apply the protocol, the only requirements are that the model simulate variables which correspond to at least a subset of the measured variables, and that the model have parameters which affect those simulated variables. This makes the protocol applicable to essentially all crop models.

The structures of the data sets (types of variables, number of environments) were very similar between the previous artificial data study and the present study, so this is not a test of genericity with respect to data. However, the data set here has a fairly large diversity of measured variables, which illustrates the versatility of the protocol. In the multi-model study using only phenology data (Wallach et al., 2023), the protocol was applied to a data set with very different phenology measurements than here, illustrating further that the protocol is not specifically adapted to a particular data set. Again, the formulation of the protocol clearly allows one to take into account a very wide range of types of data. The fact that the protocol allows one to easily use all the observed data with simulated equivalents is important, since it is expected that fitting as many observed variables as possible gives the most realistic overall description of the dynamics of the crop system (Pasley et al., 2023).

The goodness-of-fit and prediction errors after calibration here can be compared to the results in the artificial data study using the STICS model (Wallach et al. (2024). In both studies, calibration largely improves goodness-of-fit for phenology and ln(biomass), which are very well simulated (RRMSE less than 5 % error in both studies). For the calibration data, yield has the highest RRMSE value (16 %) or second highest value (22 %) for the artificial data and the data set here, respectively. It remains to be seen whether similar behavior occurs for other models and data sets. This would not be very surprising, since many more equations are involved in yield simulation than in simulating phenology. In both studies, the errors in predicting for the evaluation environments were similar to the errors for the calibration environments, i.e. there was no evidence of over-fitting (Aho et al., 2014). This is as expected, since the AICc criterion is designed to avoid over-fitting and provide good predictions.

The skill values here for phenology, biomass and grain number are positive, indicating that the model predicts better for new environments than using the average of observed calibration values. The skill measures are negative for biomass N, yield and protein, meaning that for these three variables, the model is a worse predictor than the average of the observed values. There is clearly a problem in predicting N uptake, which may be one of the causes of the poor yield predictions. This should be explored further. Note also that final biomass is poorly predicted, which could also impact yield prediction.

4.2. Diagnostics

The new diagnostics provide an in-depth evaluation of the model -specific choices made in implementing the protocol. The first two diagnostics concern the calculation of a first approximation to the WLS parameter values. An important underlying assumption of the protocol is that estimating parameters for one variable group at a time will provide a good first approximation. It is common practice to calibrate crop models using one or a few variables at a time (Dua et al., 2018; Pasley et al., 2023). However, there is little general guidance as to the way to choose the order of the variables, beyond the suggestion that the order of treating variables should be such that the “most independent” variables are treated first (Pasley et al., 2023). Here, we propose a more concrete definition of the objective, namely to minimize feedbacks, and propose a graphical method for visualizing to what extent the chosen order satisfies that objective. This diagnostic is not specific to the AgMIP calibration protocol. It would be useful for any calibration approach that fits parameters to data in stages. In the case study here, feedbacks are small. Also, the parameter values after step 6 are the best starting point for the search for the WLS parameters. Overall, the diagnostics indicate that the choices related to the first approximation to the WLS parameters are acceptable. If the graphical diagnostic reveals large feedbacks, one should first reconsider the order of the variable groups, based on knowledge of the model structure, and redo the protocol if a different order of variable groups seems more logical. If no better order seems reasonable, but if nonetheless the first approximation to the WLS parameters gives the best fit in the WLS step, one might reasonably accept the results of the protocol. Otherwise, a change in the definitions of the variable groups might be necessary, but that is beyond the subject here.

The second pair of diagnostics concern the choice of major parameters and the effect on bias reduction. Bias reduction, which requires only a single parameter per variable, is the “low-hanging fruit” of calibration. This justifies starting the calibration of each variable group with “major” parameters chosen to reduce bias. The diagnostics evaluate to what extent bias is important, and to what extent it is reduced by estimating the major parameters. The first diagnostic here is to examine the initial fraction of MSE due to square bias, to better understand the importance of bias. If squared bias is negligible, there should be no major parameter, only candidate parameters. The second diagnostic is the observed reduction in squared bias resulting from estimation of the major parameters. If a chosen major parameter does not substantially

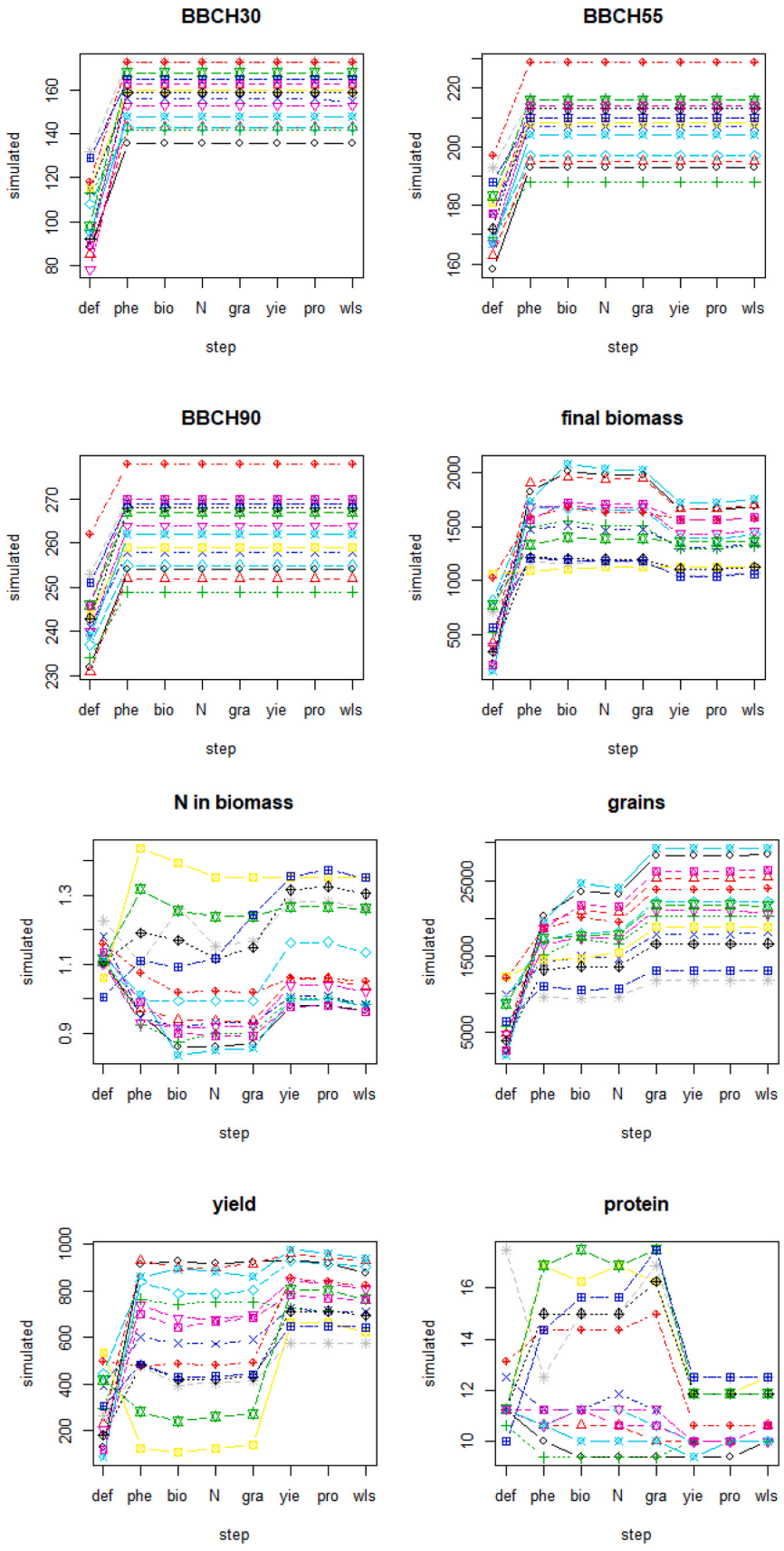

Fig. 2. Simulated values for each calibration environment at each stage of the protocol. Each plot refers to a single variable. Each line in a plot shows the simulated results for a single calibration environment, at each calibration step. The x-axis notations indicate simulated values that use the default parameter values ("def"), parameter values after fitting the phenology group ("phe"), the ln(biomass) group ("bio"), biomass N ("N"), grain number ("gra"), yield ("yie"), protein ("pro") and after the WLS step, step7 ("wls").

Table 7
RRMSE values for the evaluation data. RRMSE values using the default parameter values and parameter values after steps 6 and 7 of the protocol are shown. Final biomass (i.e. biomass at maturity) is included in the ln(biomass) group, but results are also shown specifically for this variable.

Calibration Step	BBCH30	BBCH55	BBCH90	ln(biomass)	Final biomass	Biomass N	Grain number	Yield	Protein
default	0.301	0.14	0.075	0.15	0.64	0.29	0.651	0.607	0.084
step 6	0.039	0.02	0.028	0.043	0.27	0.3	0.096	0.094	0.207
step7	0.039	0.02	0.028	0.042	0.26	0.31	0.095	0.109	0.193

Table 8
Skill values. The skill values measure how well the model predicts for the evaluation environments. The skill values using the default parameter values and parameter values after steps 6 and 7 of the protocol are shown. A positive value means that the predictions of the model have smaller MSE than predictions using the average of the observed values. Final biomass (i.e. biomass at maturity) is included in the ln(biomass) group, but results are also shown specifically for this variable.

Calibration step	BBCH30	BBCH55	BBCH90	ln(biomass)	Final biomass	Biomass N	Grain number	Yield	Protein
default	−12.25	−4.45	−2.23	−0.32	−32.89	−0.42	−16.23	−57.18	0.59
step6	0.78	0.89	0.55	0.89	−5.10	−0.51	0.63	−0.4	−1.51
step7	0.78	0.89	0.55	0.9	−5.10	−0.6	0.63	−0.87	−1.19

Table 9
Fraction of MSE for the calibration data contributed by squared bias. Values refer to simulations of each variable based on default parameter values, parameter values at the end of step 6 and parameter values after the WLS step, step 7.

Calibration step	BBCH30	BBCH55	BBCH90	ln(biomass)	Final biomass	Biomass N	Grain number	Yield	Protein
default	0.9188	0.9405	0.7766	0.588	0.86	0.38	0.885	0.871	0.111
step 6	0.00024	0.0043	0.0052	0.079	0.79	0.14	0.034	0.102	0.022
step7	0.00095	0.0043	0.0052	0.074	0.74	0.17	0.031	0.175	0.003

Table 10
Values of bias^2 for the calibration data. Values refer to simulations of each variable using the default parameter values and after each stage of the calibration. For the variable groups phenology, ln(biomass) and yield, both major and candidate parameters were estimated. For these groups, bias^2 after estimating just the major parameters are shown, and then the results after estimating both major and candidate parameters. The value in bold in each column is the value of bias^2 after fitting the major parameters of the corresponding **variable group.**

Calibration step	BBCH30	BBCH55	BBCH90	ln(biomass)	final biomass	biomass N	grain number	yield	protein
default	2.70E+ 03	969.878	414.41	0.8172	1611016	0.018	233.92	321166	0.5397
phenology major	**7.0e+ 01**	**13.796**	**0.13**	0.0220	68744	0.024	36.17	40228	0.6126
phenology	5.10E−03	0.046	0.41	0.0334	41169	0.023	34.77	53967	2.4085
biomass major	5.1e−03	0.046	0.41	**0.0067**	**12492**	0.052	14.42	72150	4.3296
biomass	5.10E−03	0.046	0.41	0.0014	12564	0.045	20.34	67656	3.5663
biomass N	5.10E−03	0.046	0.41	0.0029	16006	**0.050**	21.16	66025	3.5663
grain number	5.10E−03	0.046	0.41	0.0029	16082	0.048	**0.65**	60998	4.739
yield major	5.1e−03	0.046	0.41	0.0029	15651	0.048	0.66	**58542**	3.9387
yield	5.10E−03	0.046	0.41	0.0062	98584	0.013	0.65	2571	0.0756
protein	5.10E−03	0.046	0.41	0.0062	99332	0.012	0.65	3647	**0.0515**
(end of step 6)									
step7	2.00E−02	0.046	0.41	0.0057	84980	0.016	0.59	6462	0.0068

reduce bias, that is an inappropriate choice of major parameter. The choice should be reconsidered, based on understanding of the model. (One should not simply try multiple choices of major parameter, without an underlying rationale). If the conclusion is that there is a better choice of major parameter, it should be used. If there is no better choice of major parameter, then there should be no major parameter, only candidates. In the case study, the major parameter for biomass N does not reduce bias, but no better choice was identified. The protocol should be rerun with the parameter MXNUP as a candidate, and no major parameter. The diagnostics show that the major parameter for yield does not reduce bias, but the candidate parameter does (Table 10). Here, one should switch major and candidate parameters.

The best choice of order of treating variable groups and the choice of bias-reducing parameters are model-specific choices. This implies that it would be very helpful to have recommendations for each specific model.

4.3. Documentation

The formulation of the protocol includes several documentation tables, which describe the results of the model expertise steps and the calculation steps. This is intended to facilitate collaboration, reproducibility of the calibration process and transparency when reporting results. We suggest that reporting on the diagnostics proposed here should also be a systematic part of the documentation. That would be useful for the modeling group doing the calibration, but also as justification of the choices made in applying the protocol.

4.4. Limitations

This is a case study, with a particular crop model and data set. It is important to test the calibration protocol much more widely, with other models, crops, and data sets.

In this case study, it was possible to arrange the order of variable

groups so that feedbacks are small. This may not always be the case. For example, some models might have LAI dependent on biomass and biomass dependent on LAI. If there are observations of both, there would be substantial feedbacks regardless of the order of variables. More experience in applying the protocol is necessary in order to determine the extent of such problems, and to what the consequences are for calculation of the WLS parameters.

We have not discussed the choice of algorithm for searching for the best parameter values. The protocol focuses rather on simplifying the numerical problem, by fitting variable groups one at a time and by making a good choice of the major parameters. An underlying assumption is that estimating many parameters will be extremely difficult because of model discontinuities (Liu et al., 2018) and local optima (César Trejo Zúñiga et al., 2014). However, it would be of interest to explore the use of global minimizers (César Trejo Zúñiga et al., 2014; Jha et al., 2022) with the protocol.

The protocol is based directly on frequentist methods of parameter estimation in regression. An alternative would be a Bayesian approach, where one calculates a distribution of parameter values rather than a single best value (Berton Ferreira et al., 2024). However, a Bayesian approach would probably also require a choice of parameters to include in the calculation, since crop models typically have dozens if not hundreds of parameters. A Bayesian approach also requires a way of combining different variables in the overall likelihood function. A comprehensive Bayesian approach would need to specify how to handle such problems. It would certainly be of interest to compare such a Bayesian approach with the protocol here.

5. Conclusions

The AgMIP calibration protocol for crop models is directly based on statistical methods. However, the application of those methods requires choices, which are specific to each model and data set. The new diagnostic tools proposed here serve to evaluate those choices. They provide in-depth understanding of the calibration process, reveal problems and help suggest solutions.

Many fundamental questions related to crop models, including the importance of different observed variables for calibration, the relation between amount of data and prediction accuracy, the difference in prediction error between different model structures or the possibility of extrapolation, require a generic calibration approach that can be relied on to use the available data effectively. We suggest that the AgMIP calibration protocol, including the new diagnostics, is a promising candidate for such a calibration approach.

CRediT authorship contribution statement

Leolini Luisa: Writing – review & editing. **Mehmood Muhammad Zeeshan:** Writing – review & editing. **Mielenz Henrike:** Writing – review & editing, Resources, Project administration, Conceptualization. **Moriondo Marco:** Writing – review & editing. **Seidel Sabine J.:** Writing – review & editing, Resources, Project administration, Conceptualization. **Jing Qi:** Writing – review & editing. **Thorburn Peter:** Writing – review & editing, Resources, Project administration, Conceptualization. **Qian Budong:** Writing – review & editing. **Alderman Phillip D:** Writing – review & editing. **Schulz Susanne:** Writing – review & editing. **Dumont Benjamin:** Writing – review & editing. **Seserman Diana-Maria:** Writing – review & editing. **Fallah Mohammad Hassan:** Writing – review & editing. **Shelia Vakhtang:** Writing – review & editing. **Hoogenboom Gerrit:** Writing – review & editing. **Weihermüller Lutz:** Writing – review & editing. **Justes Eric:** Writing – review & editing. **Wallach Daniel:** Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Conceptualization. **Palosuo Taru:** Writing – review & editing, Resources, Project administration, Conceptualization. **Kersebaum Kurt-Christian:** Writing – review & editing. **Kim Kwang Soo:** Writing – review & editing, Software, Resources. **Launay Marie:** Writing – review & editing. **Hyun Shinwoo:** Writing – review & editing, Software, Resources. **Buis Samuel:** Writing – review & editing, Software, Resources, Project administration.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

All authors of this manuscript disclose any potential sources of conflict of interest.

Acknowledgements

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 – 390732324, by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE), grant number 2822ABS010, the European Union (EU horizon project IntercropVALUES, grant agreement No 101081973), the Federal Ministry of Education and Research (BMBF) BonaRes Project I4S (grant Number 031B1069B), the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) under the Excellence Strategy of the Federal and State Governments, and Rural Development Administration (RDA), Republic of Korea under the Cooperative Research Program for Agriculture Science & Technology Development (Project No. RS-2024-00361442), the Ministry of Education, Youth and Sports of the Czech Republic (grant AdAgriF - Advanced methods of greenhouse gases emission reduction and sequestration in agriculture and forest landscape for climate change mitigation (CZ.02.01.01/00/22_008/0004635), the INRAE CLIMAE meta-program and AgroEcoSystem department.

The graphical abstract was created in BioRender. Mielenz, H. (2025) https://BioRender.com/kxzd509

All authors of this manuscript disclose any potential sources of conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eja.2025.127659.

Data availability

The data used for calibration has been published in Wallach et al. (2021) https://doi.org/10.1016/j.eja.2020.126195, the evaluation data is confidential.

References

Adnan, A.A., Diels, J., Jibrin, J.M., Kamara, A.Y., Craufurd, P., Shaibu, A.S., Mohammed, I.B., Tonnang, Z.E.H., 2019. Options for calibrating CERES-maize genotype specific parameters under data-scarce environments. PLoS One 14, e0200118. https://doi.org/10.1371/journal.pone.0200118.

Aho, K., Derryberry, D., Peterson, T., 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95, 631–636. https://doi.org/10.1890/13-1452.1.

Ahuja, L.R., Ma, L., 2011. A synthesis of current parameterization approaches and needs for further improvements, in: methods of introducing system models into agricultural research. Adv. Agric. Syst. Model. 427–440. https://doi.org/10.2134/advagricsystmodel2.c15.

ARVALIS, 2022. Les températures chaudes en mai accélèrent la maturité des céréales d'hiver [WWW Document]. Infos Tech. URL ⟨https://www.arvalis.fr/infos-techniques/les-temperatures-chaudes-en-mai-accelerent-la-maturite-des-cereales-dhiver⟩ (accessed 10.29.24).

Asseng, S., Zhu, Y., Basso, B., Wilson, T., Cammarano, D., 2014. Simulation modeling: applications in cropping systems. Encycl. Agric. Food Syst. 102–112. https://doi.org/10.1016/B978-0-444-52512-3.00233-3.

Berton Ferreira, T., Shelia, V., Porter, C., Moreno Cadena, P., Salmeron Cortasa, M., Sohail Khan, M., Pavan, W., Hoogenboom, G., 2024. Enhancing crop model parameter estimation across computing environments: utilizing the GLUE method and parallel computing for determining genetic coefficients. Comput. Electron. Agric. 227, 109513. https://doi.org/10.1016/J.COMPAG.2024.109513.

Boote, K. (Ed.), 2019. Advances in crop modelling for a sustainable agriculture. Burleigh Dodds Science Publishing. ⟨https://doi.org/10.1201/9780429266591⟩.

Boote, K.J., Jones, J.W., Hoogenboom, G., White, J.W., 2010. The role of crop systems simulation in agriculture and environment. Int. J. Agric. Environ. Inf. Syst. 1, 41–54.

Buis, S., Lecharpentier, P., Vezy, R., Ginet, M., 2023. CroptimizR: A Package to Estimate Parameters of Crop Models [WWW Document]. URL https://doi.org/10.5281/zenodo.4066451 (2023).

César Trejo Zúñiga, E., López Cruz, I.L., García, A.R., 2014. Parameter estimation for crop growth model using evolutionary and bio-inspired algorithms. Appl. Soft Comput. 23, 474–482. https://doi.org/10.1016/J.ASOC.2014.06.023.

Claeskens, G., Hjort, N.L., 2001. Model Selection and Model Averaging. Cambridge University Press. ⟨https://doi.org/10.1017/CBO9780511790485⟩.

Dua, V.K., Minhas, J.S., Rawal, S., Singh, S.P., Singh, S.K., KUMAR, P., PATHANIA, R., Kapoor, T., SHARMA, J., Sharma, S.K., Mankar, P., Kawat, S., Singh, B.P., Chakrabarti, S.K., 2018. Calibration and validation of WOFOST model for seven potato (Solanum tuberosum) cultivars in India. Indian J. Agron. 63, 357–365. https://doi.org/10.59797/IJA.V63I3.5661.

Fallah, M.H., Nezami, A., Khazaie, H.R., Nassiri Mahallati, M., 2020. Evaluation of DSSAT-Nwheat model across a wide range of climate conditions in Iran. J. Agroecol 12, 561–580. https://doi.org/10.22067/jag.v12i4.77250.

Fath, B., Jorgensen, S.E., 2011. Elsevier, Amsterdam.

Gao, Y., Wallach, D., Hasegawa, T., Tang, L., Zhang, R., Asseng, S., Kahveci, T., Liu, L., He, J., Hoogenboom, G., 2021. Evaluation of crop model prediction and uncertainty using Bayesian parameter estimation and Bayesian model averaging. Agric. For. Meteor. 311, 108686. https://doi.org/10.1016/J.AGRFORMET.2021.108686.

Grassini, P., van Bussel, L.G.J., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., de Groot, H., van Ittersum, M.K., Cassman, K.G., 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. F. Crop. Res. 177, 49–63. https://doi.org/10.1016/j.fcr.2015.03.004.

Guillaume, S., Berez, J.-E., Wallach, D., Justes, E., 2011. Methodological comparison of calibration procedures for durum wheat parameters in the STICS model. Eur. J. Agron. 35, 115–126.

Guo, D., Olesen, J.E., Pullens, J.W.M., Guo, C., Ma, X., 2021. Calibrating AquaCrop model using genetic algorithm with multi-objective functions applying different weight factors. Agron. J. 113, 1420–1438. https://doi.org/10.1002/agj2.20588.

Hargreaves, J.C., 2010. Skill and uncertainty in climate models. WIREs Clim. Chang 1, 556–564. https://doi.org/10.1002/wcc.58.

Harrison, M.T., Roggero, P.P., Zavattaro, L., 2019. Simple, efficient and robust techniques for automatic multi-objective function parameterisation: Case studies of local and global optimisation using APSIM. Environ. Model. Softw. 117, 109–133. https://doi.org/10.1016/j.envsoft.2019.03.010.

He, D., Wang, E., Wang, J., Robertson, M.J., 2017. Data requirement for effective calibration of process-based crop models. Agric. For. Meteor. https://doi.org/10.1016/j.agrformet.2016.12.015.

Hlavinka, P., Trnka, M., Kersebaum, K., čermák, P., Pohanková, E., Orság, M., Pokorný, E., Fischer, M., Brtnický, M., žalud, Z., 2013. Modelling of yields and soil nitrogen dynamics for crop rotations by HERMES under different climate and soil conditions in the Czech Republic. J. Agric. Sci. 152, 188–204. https://doi.org/10.1017/S0021859612001001.

Hoogenboom, G., Porter, C.H., Boote, K.J., Shelia, V., Wilkens, P.W., Singh, U., White, J. W., Asseng, S., Lizaso, J.I., Moreno, L.P., Pavan, W., Ogoshi, R., Hunt, L.A., Tsuji, G. Y., Jones, J.W., 2019. The DSSAT crop modeling ecosystem, in: Boote, K. J. (Ed.), Advances in Crop Modeling for a Sustainable Agriculture. Burleigh Dodds Sci. Publ., Camb. U. Kingd. 173–216. https://doi.org/10.19103/AS.2019.0061.10.

Janssen, P.H.M., Heuberger, P.S.C., 1995. Calibration of process-oriented models. Ecol. Model. 83, 55–66.

Jha, P.K., Ines, A.V.M., Singh, M.P., 2021. A multiple and ensembling approach for calibration and evaluation of genetic coefficients of CERES-Maize to simulate maize phenology and yield in Michigan. Environ. Model. Softw. 135, 104901. https://doi.org/10.1016/j.envsoft.2020.104901.

Jha, P.K., Ines, A.V.M., Han, E., Cruz, R., Vara Prasad, P.V., 2022. A comparison of multiple calibration and ensembling methods for estimating genetic coefficients of CERES-Rice to simulate phenology and yields. F. Crop. Res. 284, 108560. https://doi.org/10.1016/J.FCR.2022.108560.

Jing, Q., McConkey, B., Qian, B., Smith, W., Grant, B., Shang, J., Liu, J., Bindraban, P., Luce, M., St, 2021. Assessing water management effects on spring wheat yield in the Canadian Prairies using DSSAT wheat models. Agric. Water Manag. 244, 106591. https://doi.org/10.1016/J.AGWAT.2020.106591.

Kassie, B.T., Asseng, S., Porter, C.H., Royce, F.S., 2016. Performance of DSSAT-Nwheat across a wide range of current and future growing conditions. Eur. J. Agron. 81, 27–36. https://doi.org/10.1016/J.EJA.2016.08.012.

Kersebaum, K.C., 2011. Special Features of the HERMES Model and Additional Procedures for Parameterization, Calibration, Validation, and Applications. In: Ahuja, L.R., Ma, L. (Eds.), Methods of Introducing System Models into Agricultural Research. John Wiley & Sons, Ltd, pp. 65–94. ⟨https://doi.org/10.2134/advagricsystmodel2.c2⟩.

Kobayashi, K., Salam, M.U., 2000. Comparing simulated and measured values using mean squared deviation and its components. Agron. J. 92, 345–352.

Kumar, K., 2023. Exploring Optimization Techniques for Parameter Estimation in Nonlinear System Modeling.

Liu, L., Wallach, D., Li, J., Liu, B., Zhang, L., Tang, L., Zhang, Y., Qiu, X., Cao, W., Zhu, Y., 2018. Uncertainty in wheat phenology simulation induced by cultivar parameterization under climate warming. Eur. J. Agron. 94, 46–53. https://doi.org/10.1016/J.EJA.2017.12.001.

Morozova, O., Levina, O., Uusküla, A., Heimer, R., 2015. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Med. Res. Method. 15, 71. https://doi.org/10.1186/s12874-015-0066-2.

Pasley, H., Brown, H., Holzworth, D., Whish, J., Bell, L., Huth, N., 2023. How to build a crop model. A review. Agron. Sustain. Dev. 43, 2. https://doi.org/10.1007/s13593-022-00854-9.

Richter, G.M., Acutis, M., Trevisiol, P., Latiri, K., Confalonieri, R., 2010. Sensitivity analysis for a complex crop model applied to Durum wheat in the Mediterranean. Eur. J. Agron. 32, 127–136. https://doi.org/10.1016/j.eja.2009.09.002.

Seber, G.A.F., Wild, C.J., 1989. Nonlinear regression. Wiley, New York.

Seidel, S.J., Palosuo, T., Thorburn, P., Wallach, D., 2018. Towards improved calibration of crop models – Where are we now and where should we go? Eur. J. Agron. 94, 25–35. https://doi.org/10.1016/J.EJA.2018.01.006.

Turgal, E., Doganay, B., 2017. Incl. Exclud. a Constant Term. Regres. Anal.

Van Oijen, M., Rougier, J., Smith, R., 2005. Bayesian calibration of process-based forest models: bridging the gap between models and data. Tree Physiol. 25, 915–927.

Vezy, R., Buis, S., Lecharpentier, P., Giner, M., 2023. 2023. CroPlotR: a package to analyse crop model simulations outputs with plots and statistics. WWW Doc. https://doi.org/10.5281/zenodo.4066451.

Wallach, D., 2011. Crop model calibration: a statistical perspective. Agron. J. 103, 1144–1151.

Wallach, D., Palosuo, T., Thorburn, P., Gourdain, E., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hochman, Z., Hoek, S., Hoogenboom, G., Horan, H., Huang, M., Jabloun, M., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S.J., 2021a. How well do crop modeling groups predict wheat phenology, given calibration data from the target population? Eur. J. Agron. 124, 126195. https://doi.org/10.1016/j.eja.2020.126195.

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, P.-E., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Lewan, E., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Padovan, G., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K. D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S.J., 2021b. The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environ. Model. Softw. 145, 105206. https://doi.org/10.1016/J.ENVSOFT.2021.105206.

Wallach, D., Palosuo, T., Thorburn, P., Mielenz, H., Buis, S., Hochman, Z., Gourdain, E., Andrianasolo, F., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Harrison, M., Hiremath, S., Horan, H., Hoogenboom, G., Jansson, P.-E., Jing, Q., Justes, E., Kersebaum, K.-C., Launay, M., Lewan, E., Liu, K., Mequanint, F., Moriondo, M., Nendel, C., Padovan, G., Qian, B., Schütze, N., Seserman, D.-M., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Trombi, G., Weber, T.K.D., Weihermüller, L., Wöhling, T., Seidel, S.J., 2023. Proposal and extensive test of a calibration protocol for crop phenology models. Agron. Sustain. Dev. 43, 46. https://doi.org/10.1007/s13593-023-00900-0.

Wallach, D., Buis, S., Seserman, D.-M., Palosuo, T., Thorburn, P.J., Mielenz, H., Justes, E., Kersebaum, K.-C., Dumont, B., Launay, M., Seidel, S.J., 2024. A calibration protocol for soil-crop models. Environ. Model. Softw. 180, 106147. https://doi.org/10.1016/J.ENVSOFT.2024.106147.

Wang, X., Jeong, J., Park, S., Zhang, X., Gao, J., Silvero, N., 2023. DayCent-CUTE: a global sensitivity, auto-calibration, and uncertainty analysis tool for DayCent. Environ. Model. Softw., 105832 https://doi.org/10.1016/J.ENVSOFT.2023.105832.

Winn, C.A., Archontoulis, S., Edwards, J., 2023. Calibration of a crop growth model in APSIM for 15 publicly available corn hybrids in North America. Crop Sci. 63, 511–534. https://doi.org/10.1002/csc2.20857.

Wolf, J., Evans, L.G., Semenov, M.A., Eckersten, H., Iglesias, A., 1996. Comparison of wheat simulation models under climate change. I. Model calibration and sensitivity analyses. Clim. Res. 7, 253–270.

Zhao, G., Bryan, B.A., Song, X., 2014. Sensitivity and uncertainty analysis of the APSIM-wheat model: Interactions between cultivar, environmental, and management parameters. Ecol. Model. 279, 1–11. https://doi.org/10.1016/j.ecolmodel.2014.02.003.