Agronomy for Sustainable Development (2023) 43:46
https://doi.org/10.1007/s13593-023-00900-0

RESEARCH ARTICLE

Proposal and extensive test of a calibration protocol for crop phenology models

Daniel Wallach[1] · Taru Palosuo[2] · Peter Thorburn[3] · Henrike Mielenz[4] · Samuel Buis[5] · Zvi Hochman[3] · Emmanuelle Gourdain[6] · Fety Andrianasolo[6] · Benjamin Dumont[7] · Roberto Ferrise[8] · Thomas Gaiser[1] · Cecile Garcia[6] · Sebastian Gayler[9] · Matthew Harrison[10] · Santosh Hiremath[11] · Heidi Horan[3] · Gerrit Hoogenboom[12,13] · Per-Erik Jansson[14] · Qi Jing[15] · Eric Justes[16] · Kurt-Christian Kersebaum[17,18,27] · Marie Launay[19] · Elisabet Lewan[20] · Ke Liu[10] · Fasil Mequanint[9] · Marco Moriondo[21] · Claas Nendel[17,18,22] · Gloria Padovan[8] · Budong Qian[15] · Niels Schütze[23] · Diana-Maria Seserman[17] · Vakhtang Shelia[12,13] · Amir Souissi[24] · Xenia Specka[17] · Amit Kumar Srivastava[1] · Giacomo Trombi[8] · Tobias K. D. Weber[9,28] · Lutz Weihermüller[25] · Thomas Wöhling[23,26] · Sabine J. Seidel[1]

Accepted: 22 June 2023 / Published online: 13 July 2023

Abstract
A major effect of environment on crops is through crop phenology, and therefore, the capacity to predict phenology for new environments is important. Mechanistic crop models are a major tool for such predictions, but calibration of crop phenology models is difficult and there is no consensus on the best approach. We propose an original, detailed approach for calibration of such models, which we refer to as a calibration protocol. The protocol covers all the steps in the calibration workflow, namely choice of default parameter values, choice of objective function, choice of parameters to estimate from the data, calculation of optimal parameter values, and diagnostics. The major innovation is in the choice of which parameters to estimate from the data, which combines expert knowledge and data-based model selection. First, almost additive parameters are identified and estimated. This should make bias (average difference between observed and simulated values) nearly zero. These are "obligatory" parameters, that will definitely be estimated. Then candidate parameters are identified, which are parameters likely to explain the remaining discrepancies between simulated and observed values. A candidate is only added to the list of parameters to estimate if it leads to a reduction in BIC (Bayesian Information Criterion), which is a model selection criterion. A second original aspect of the protocol is the specification of documentation for each stage of the protocol. The protocol was applied by 19 modeling teams to three data sets for wheat phenology. All teams first calibrated their model using their "usual" calibration approach, so it was possible to compare usual and protocol calibration. Evaluation of prediction error was based on data from sites and years not represented in the training data. Compared to usual calibration, calibration following the new protocol reduced the variability between modeling teams by 22% and reduced prediction error by 11%.

Keywords Crop model · Prediction error · Protocol · Model ensemble · Variability

1 Introduction

Plant phenology is a major aspect of plant response to environment and a major determinant of plant response to climate change. This includes phenology of natural vegetation, which has been shown to be affected by warming (Piao et al. 2019; Menzel et al. 2020; Stuble et al. 2021) as well as phenology of cultivated crops (Rezaei et al. 2018). For the latter, phenology must be taken into account for crop management (Sisheber et al. 2022), choice of cultivar or cultivar characteristics adapted to a particular region (Zhang et al. 2022), and for evaluating the impact of climate change on crop production (Rezaei et al. 2018). It is thus important to be able to predict phenology as a function of environment, and in particular as a function of climate.

A number of mechanistic crop models have been developed, which include simulation of phenology. Such models are regularly used to evaluate management options (McNunn et al. 2019) or the effect of climate change on crops, including wheat (Asseng et al. 2013), rice, (Li et al. 2015), maize (Bassu et al.

Extended author information available on the last page of the article

2014) and soybean (Fodor et al. 2017). Such models are particularly important for taking into account an increasing diversity of combinations of weather events (Webber et al. 2020).

Mechanistic models in general, and models used to simulate crop phenology in particular, are based on our understanding of the processes and their inter-linkages that drive the evolution of the system. This conceptual understanding usually builds on detailed experiments that study specific aspects of the system (e.g., Brisson et al. 2003 for the crop model STICS). The set of model equations, which is the mathematical expression of our understanding, is referred to as "model structure" (Tao et al. 2018).

In addition to model structure, simulation requires values for all the model parameters. In essentially all uses of crop models, the model is first calibrated using observed data that is related to the target population for which predictions are required, for example, observations for the specific variety of interest and/or for the particular set of growing environments of interest. Calibration normally only concerns a fairly small subset of the parameters in a crop model, but is essentially always necessary because mechanistic models are only approximations, without universally valid parameter values (Fath and Jorgensen 2011; Wallach 2011).

There are therefore two main tracks to improvement of crop phenology model predictions. The first is through improvement of model structure through improved understanding of the underlying processes, and the second is through improvement of model calibration, and that is the topic here. The specific context of interest is the use of a phenology model to predict crop phenology for new environments, given a sample of data from similar environments. The calibration problem is how best to use the sample of data in order to minimize prediction error. We do not consider the question of the most appropriate data sets for calibration. Rather, we assume that there is preexisting data, for example, from variety trials, and one is limited to those data. Note that while better equations and better calibration are two separate paths to improving model predictions, they are complementary; if one wants to compare how well different equations predict phenology, one must first calibrate them.

Calibration of crop models is usually patterned on statistical methods used to estimate parameters in regression. However, the application of statistical methods to crop models is not straightforward. Major difficulties include the fact i. that system models often have multiple output variables that can be compared to observed results (e.g., dates of heading and dates of flowering for crop phenology models), ii. that errors for different variables in the same environment are often correlated, and iii. that there are usually many parameters, often more than the number of data points available. While the details differ, these problems apply to essentially all system models. No doubt, as a result, there are no widely accepted standard methods for calibration of system models. It has been found, for example, that there is a wide diversity of calibration approaches between modeling teams furnished with identical data, even between modeling teams using the same model structure (Confalonieri et al. 2016; Wallach et al. 2021a, b). By modeling team we mean a group of people working together on or with a crop model.

Because of the importance of calibration and the lack of standard approaches for calibration, there have been many studies published that make recommendations as to how to calibrate crop models or system models in other fields. One type of study is model-specific and identifies the most important parameters to estimate for a particular model (Ahuja and Ma 2011). Other studies have focused on the methodology of identifying the most important parameters through sensitivity analysis (Khorashadi Zadeh et al. 2022), on the choice between frequentist and Bayesian paradigms (Gao et al. 2021), on the form of the objective function, or on the numerical algorithm for searching for the best parameter values (Rafiei et al. 2022). A recent study has shown that different modeling teams make different choices for all the steps of the calibration procedure (Wallach et al. 2021c). That study showed that the modeling community is far from having a consensus on how to calibrate phenology models and that progress is required for multiple aspects of the calibration procedure.

The purpose of this study was to define and test an original, detailed, comprehensive procedure for crop phenology model calibration that could be applied to a wide range of models. We refer to this new procedure as a "protocol," to emphasize that it contains detailed instructions for calibration. It builds on the recommendations in Wallach et al. (2021c) but goes beyond those more general recommendations, most importantly in proposing an original approach for choosing the parameters to estimate, which is arguably the most important calibration decision. A second major innovation of the protocol is the definition of documentation tables for each step of the calibration procedure, which can be used both for communication within a modeling team and to inform users of the calibrated model. We tested the protocol in a large multimodel ensemble study. Each modeling team first calibrated their model using their "usual" calibration procedure, and then using the protocol proposed here. This is thus a comparison, for each team, of usual versus protocol calibration. The results showed that the protocol reduced the variability between modeling teams compared to usual calibration approaches and, most importantly, that it significantly reduced prediction error compared to usual calibration approaches.

2 Materials and methods

2.1 Data sets

Three data sets for wheat phenology were used here, where a data set is data from fields representative of some

specific target population. The target population for the first data set was wheat fields in the major wheat-growing regions of France sown with winter wheat variety Apache and using usual management. The data were from cultivar trials. They were separated into a calibration subset with data from 14 environments (6 different sites, sowing years 2010, 2011, 2014-2016 but not every year was represented for every site) and an evaluation subset with data from 8 environments (5 different sites, sowing years 2012 and 2013). The target population for the second data set was identical to that for the first, but with winter wheat variety Bermude in place of Apache. For both data sets, the calibration and evaluation subsets had neither site nor year in common, so the evaluation is a rigorous test of how well a modeling team can simulate phenology for out-of-sample environments. The observed data were days from sowing to the beginning of stem elongation (BBCH30 on the BBCH scale, Meier 1997) and to the middle of heading (BBCH 55).

The target population for the third data set was wheat fields in the major wheat-growing regions of Australia, with usual management and sown with spring wheat variety Janz. The data were from a multi-location multi-year multiple sowing date trial in Australia (Lawes et al. 2016; Wallach et al. 2021b). The calibration subset had data from four sites in 2010 and 2011, with three sowing dates per site (overall 24 environments). The evaluation subset had data from six sites in 2012, with three sowing dates per site (overall, 18 environments). Once again, the calibration and evaluation subsets had neither site nor year in common. In the original trials, the BBCH development stage was observed once weekly in each environment. Based on those data, a graph of BBCH stage versus day was produced, and interpolation was used to obtain the day for each integer BBCH stage from the earliest to the latest recorded in each environment. Those dates were provided to the modeling teams.

2.2 Modeling teams and model structures

Nineteen modeling teams, using 16 different model structures (Table 1), participated in this study, which was carried out within the Agricultural Model Intercomparison and Improvement Project (AgMIP; www.agmip.org). The modeling teams are identified only by a code ("M1," "M2," etc.) without indicating which model structure they used, since it would be misleading to give the impression that the results are determined solely by model structure. The participating teams represent an "ensemble of opportunity," that is, an open call for participants was put out, and all teams that volunteered were accepted. An indication of the resulting variability in phenology prediction is given by the variability in the choice of parameters to calibrate using "usual" calibration (see Supplementary Table S4 in Wallach et al 2021c). Most teams estimated some parameters that represent degree days to various stages, though the number of degree day parameters varied between teams. A few models have parameters that represent development rates rather than degree days to each stage. Only a few teams estimated parameters related to time from sowing to emergence. About half of the teams estimated one or more parameters related to vernalization and to photoperiod sensitivity. A few teams estimated parameters related to the temperature response function, for example, minimum temperature or optimum temperature for development, or related to tillering or leaf

Table 1 List of model structures used by participating modeling teams.

Model structure	Version(s)	References
AgroC	May 2018	(Herbst et al. 2008; Klosterhalfen et al. 2017)
APSIM	7.8, 7.9, 7.10	(Keating et al. 2003; Holzworth et al. 2014)
AquaCrop	4.0	(Vanuytrecht et al. 2014)
CERES-Wheat	DSSAT V4.7.	(Hoogenboom et al. 2019a, 2019b; Jones et al. 2003)
CoupModel	Version 5.4.4	(Jansson 2012; Senapati et al. 2016; Coucheney et al. 2018)
CROPSIM-Wheat	DSSAT V4.7	(Hoogenboom et al. 2019a, 2019b; Jones et al. 2003)
Cropsyst	3.04.08	(Stockle et al. 2001)
HERMES	4.27	(Kersebaum 2007, 2011)
LINTUL	LINTUL5	(Wolf 2012)
MONICA	2.02	(Nendel et al. 2011; Specka et al. 2015, 2019)
PANORAMIX	R version	(Gate 1995; Chatelin et al. 2005)
SPASS	Expert-N 5.0	(Wang 1997)
CERES	Expert-N 5.0	(Jones et al. 2003)
SSM-Wheat		(Soltani et al. 2013)
STICS	8_5_0	(Brisson et al. 2009; Coucheney et al. 2015)
WOFOST	7.1.7	(Boogaard et al. 2013)

appearance rate. Finally, a few teams estimated parameters related to the effect of stress on the development rate.

2.3 Goodness-of-fit and evaluation of predictions

Goodness-of-fit refers to how well a calibrated model fits the data used for calibration. Prediction accuracy refers to how well a calibrated model simulates for environments different than those in the calibration data set. Since the evaluation environments here are for sites and years not represented in the calibration data, the test of simulated values against the evaluation data truly reflects how well a model can predict for new environments.

For both goodness-of-fit and out-of-sample prediction, our basic evaluation metric is the sum of squared errors (SSE) and the related quantities mean squared error (MSE) and root mean squared error (RMSE), where

$$\begin{aligned} SSE &= \sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2 \\ MSE &= SSE/n \\ RMSE &= \sqrt{MSE} \end{aligned} \quad (1)$$

The sum for SSE is over days to BBCH30 and BBCH55 for the French data sets, days to BBCH30, BBCH65, and BBCH90 for the Australian data sets and over environments. Here, y_{ij} is the observed value of variable i for environment j, $\hat{y}_{ij}$ is the corresponding simulated value and n is the number of terms in the sum. We also look at the decomposition of MSE as the sum of three terms, namely squared bias ($bias^2$), a term related to the difference in standard deviations of the observed and simulated values (SDSD), and a term related to the correlation of observed and simulated values (LCS) (Kobayashi 2004).

In addition, we compare the simulated results in this study with two simple benchmark models. The first (the "naive" model) is simply the average number of days to each stage in the calibration data of each data set. This is used as the prediction model for all environments of that data set. The often-used Nash Sutcliffe modeling efficiency is one minus the ratio of MSE of a model to MSE of the naive model. The naive model ignores all variability between environments, so it is a very low bar as a benchmark. We therefore also use a more sophisticated benchmark, the "onlyT" model, as in Wallach et al. (2021a). This benchmark model assumes that the sum of degree days above a threshold of 0°C from sowing to each stage is fixed for spring wheat. For winter wheat, a simple vernalization model is used to determine the start of development. Vernalization is 0 if daily mean air temperature is below −4°C, increases linearly to 1 at 3°C, remains at 1 to 10°C, decreases linearly to 0 at 17°C and is 0 above 17 °C. When the sum of daily vernalization reaches 50, vernalization is complete (van Bussel et al. 2015; Wallach et al. 2021a). Then the fixed number of degree days applied after vernalization is completed. Both benchmark models are quite easily parameterized based on calibration data, and then easily applied to new environments.

2.4 Simulation exercise

The participants received input data (daily weather at the field, soil characteristics, management details, and, where possible, initial conditions) for all environments of every data set. Also, the observed data from the calibration environments were provided to all participants. The participants were asked to use those data to calibrate their models using the calibration protocol described in detail below and then to simulate and report days after sowing to stages BBCH10 (days to emergence), BBCH30, and BBCH55 for the French calibration and evaluation environments, and to stages BBCH10, BBCH30, BBCH65, and BBCH90 for the Australian calibration and evaluation environments. Days to emergence was included to have an example of a variable for which there were no calibration data. The BBCH stages 30 and 55 requested for the French environments represent stages that are used for fertilizer decisions in France. The BBCH stages 30, 65, and 90 requested for the Australian environments represent major transitions that are explicitly simulated by many models.

All teams calibrated their model with the same data as here using their usual calibration approach, either in previous studies (Wallach et al. 2021a, b) or specifically for this study. It is the results of the usual calibration method that are compared here to the results of using the proposed protocol. At no time were the evaluation data shown to participants, neither in previous studies nor in the present study.

The protocol does not impose a specific software solution. However, several participants used trial and error in their usual approach and requested help in finding and implementing an automated search algorithm, since that is required for the protocol. To answer this need, the CroptimizR R package (Buis et al. 2021) was modified to do the protocol calculations, and many of the participants used this software.

In addition to the individual models, we report on two ensemble models, created by taking the mean (the e-mean model) or the median (the e-median model) of the simulated values. These ensemble models were calculated both for the usual and protocol calibration results.

2.5 AICc and BIC

The protocol prescribes a model selection criterion to decide which parameters to estimate. The corrected Akaike Information Criterion (AICc) and the Bayesian Information

Criterion (BIC) are two different criteria that are often used for model selection (Chakrabarti and Ghosh 2011). Both are based on model error, with a penalization term that increases with the number of estimated parameters. Assuming that model errors are normally distributed, the criteria are:

$$AIC_c = n\ln(MSE) + 2p + \frac{2p(p+1)}{n-p-1}$$
$$BIC = n\ln(MSE) + p\ln(n) \quad (2)$$

where n is the number of data points and p is the number of calibrated parameters. These criteria are only used for comparing models calibrated using the same data.

There have been comparisons between these criteria, but there does not seem to be one that systematically performs better than the other, for choosing the model that predicts best (Kuha 2004). In applying the protocol here, participants were asked to perform the calculations twice, once using the AICc criterion and once using the BIC criterion to choose the parameters to estimate. In almost all cases, the two criteria led to exactly the same choice of parameters. In the few cases where the criteria led to different choices, the final models had very similar RMSE for the evaluation data, with a very slight advantage to BIC (Supplementary tables S24-S25). Therefore, all results shown here are based on the BIC criterion.

3 Results and discussion

3.1 Description of protocol

The protocol is based on the recommendations in Wallach et al. (2021c), and follows the same list of steps (Fig. 1), but has important additions, in particular for the choice of parameters to estimate and the documentation to be produced.

Step 1. Describe environments, choose default parameter values

It is important to describe the environments represented in the data, and of the target population, in particular temperatures and day lengths. Information for the data sets used here can be found in Wallach et al. (2021a, 2021b).

Since most parameters retain their default values, the choice of default values for those parameters that affect phenology is important. For phenology, one would want to have reasonable approximations to the cycle length for the cultivar in question, to photoperiod dependence, and to vernalization requirements. This information and more are usually available from the cultivar developer. The documentation for step 1 (see example in Table 2) specifies the characteristics of the cultivar being modeled and those of the cultivar used to provide default parameter values.

Step 2. Identify correspondence between observed and simulated variables

In the simplest case, there is a simulated variable that corresponds directly to each observed development stage. The documentation for step 2 is a table with one row for each observed variable, showing the corresponding simulated variable if any (see example in Table 3).

Step 3. Define the objective function

The objective function of the protocol is squared error summed over development stages and environments, which is the objective function of ordinary least squares (OLS) regression and is often used in crop model calibration. A major choice of the protocol is to include in the objective function all the observed development stages that have a simulated equivalent, including stages that are not of primary interest. A first reason is that often the same calibrated model will be used for several different objectives, so measured variables that are not of central interest in the current study may be important in future studies. Furthermore, using

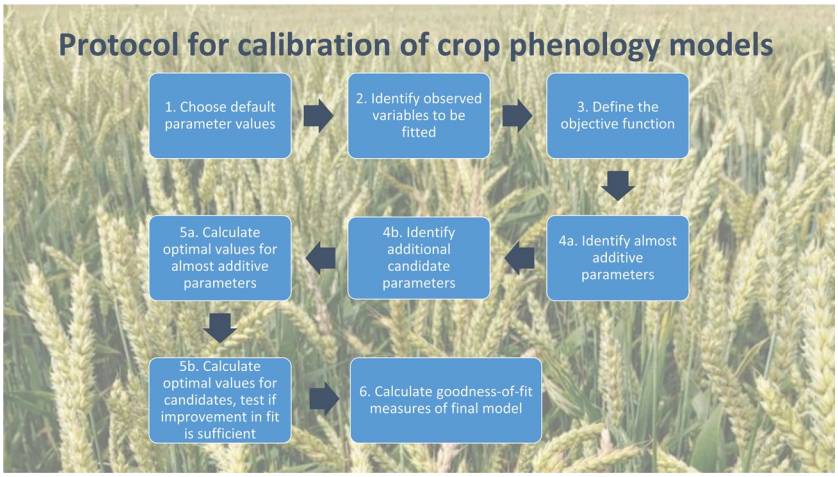

Fig. 1 Schematic diagram of steps in proposed calibration protocol.

Table 2 Example of protocol documentation for Step 1, "Choose default parameter values". The first row shows cultivar characteristics of the observed cultivar. The second row shows characteristics of the cultivar that provides the default parameter values. This example is for the French data set, variety Apache, modeling team M21.

Cultivar	Characteristics
Cultivar of observations: Apache	A soft winter wheat. Stem elongation—semi-early. Heading—early. Vernalization requires 40 days where full vernalization occurs if daily average temperature is between 3 and 10°C. There is no vernalization below −4°C or above 17°C. Otherwise, there is a proportional reduction in vernalization effectiveness.
Cultivar used to provide default parameter values: Soissons	Soissons seems to be close to Apache in terms of vernalization requirements and earliness.

more variables makes the model a better representation of multiple aspects of the system dynamics, which is likely to improve all simulations. The choice in the protocol is to use OLS and to avoid estimating additional statistical parameters representing variance and covariance of errors. However, one should check residual errors to evaluate the extent of heteroscedasticity or correlation of errors. Since the objective function is the sum of squared errors over the variables from step 2, no new decisions are required here and no additional documentation is required.

Step 4. Choose which parameters to estimate

This is arguably the most difficult and the most important decision of the calibration approach. Here we propose a novel approach which combines expert knowledge with a statistical model selection criterion. This approach distinguishes two categories of parameters to estimate: the nearly additive, obligatory parameters (those that will definitely be estimated) and the candidate parameters (those that will be tested, and only changed from the default value if the improvement in the fit to the calibration data is sufficiently large).

Step 4a. Identify the obligatory parameters

The obligatory parameters are parameters that are nearly additive, i.e., such that changing the parameter has a similar effect for all environments for some variable in the objective function. Usually, a parameter that represents degree days to a measured stage is a good choice as an obligatory parameter for time to that stage. Estimating a truly additive parameter, which adds the same constant amount to days to a stage for all environments, will exactly eliminate bias for that stage. That is, the mean of simulated values will exactly equal the mean of observed values (see Supplementary Eqs. 1-2). Estimating an almost additive parameter will nearly eliminate bias. Once bias is nearly eliminated, one may already have a fairly reasonable fit to the data. Each almost additive parameter must affect a different variable or combination of variables. There cannot be more almost additive parameters than the number of variables in the objective function. Otherwise, the parameters would be very poorly estimated, or non-estimable. The protocol does allow fewer almost additive parameters than observed variables. In that case, bias is only nearly eliminated on average over several variables, and not for each variable. The choice of obligatory parameters is up to the modeling team, based on knowledge of the model. However, the protocol gives fairly detailed recommendations, namely that they should be nearly additive, that degree days to stages are usually a good choice, and that the number of obligatory parameters cannot exceed the number of different variables. For each obligatory parameter, one must provide the default value and what one considers a reasonable range for that parameter (for an example of choice of bounds see Tao et al. 2018). The documentation for step 4a is a table with one row for each obligatory parameter (see example in Table 4).

Table 3 Example of documentation for protocol Step 2, "Identify correspondence between observed and simulated variables". The table has one row for each measured variable, showing the corresponding simulated variable if any. This example is for the French data sets and modeling team M21.

Measured variable	Corresponding simulated value
Days to BBCH stage 30	Days to end juvenile stage
Days to BBCH stage 55	Days at maximum LAI

Step 4b. Identify candidate parameters

The role of the candidate parameters is to reduce the variability between environments that remains after estimation of the obligatory parameters. It is the role of the modeler to identify the candidate parameters and to order them by amount of variability likely to be explained. In the calculation step (step 5), each candidate parameter is tested, and only those that lead to a reduction in the BIC criterion are retained for estimation. Otherwise, the parameter is kept at its default value. The documentation for step 4b is a table with one row for each candidate parameter (see example in Table 5).

Table 4 Example of documentation for protocol Step 4a, "Identify the obligatory parameters." These are parameters that are almost additive, i.e., that have nearly the same effect for all environments. There is one row for each obligatory parameter. The number of obligatory parameters cannot exceed the number of observed variables which have a simulated equivalent, and each obligatory parameter must be nearly additive for a different variable or combination of variables. This example is for the French data set, variety Apache for modeling team M21.

Obligatory parameter	Explanation	Default value (lower–upper limits)
stlevamf	Degree days sowing to end juvenile stage	233 (150–400)
stamflax	Degree days sowing to maximum LAI	354 (150–500)

Step 5. Calculation of the optimal parameter values

The protocol prescribes the use of a simplex algorithm for searching for the optimal parameter values. The Nelder-Mead simplex method (Nelder and Mead 1965) is a robust, derivative-free method, which is appropriate for crop models which may have multiple discontinuities. The results of the simplex are sensitive to starting values (Press et al. 2007), so the protocol calls for multiple starting points. In the first calculation step, the obligatory parameters are estimated and the BIC value is calculated. This is the initial list of parameters to estimate. Then each candidate parameter is tested in turn. If estimating the new candidate together with the previous list of parameters to estimate leads to a reduction in BIC, the candidate is added to the list of parameters to estimate. If not, the candidate returns to its default value and will not be estimated. The documentation for step 5 is a table with one row for each step in the calculation (see example in Table 6).

The first calculation step, searching for the optimal values for the obligatory parameters, only involves a relatively small number of parameters, and furthermore, these parameters are chosen to be nearly additive. It is expected then that this step should not present serious numerical difficulties. Subsequently, only one new candidate is added at a time. The previously chosen parameters are also estimated, but it is expected that their previous best values should be good starting values, so that once again there should not be serious numerical difficulties. Overall, the protocol takes advantage of the particular structure of crop phenology models, in particular the possibility of identifying the most important, almost additive parameters, in order to separate the search for optimal parameter values into relatively easy steps.

Step 6. Examine goodness-of-fit

Many diagnostics are possible and useful. We emphasize particularly a graph of simulated versus observed values, calculated MSE and its decomposition for each variable (see example in Table 7) and comparison with the two benchmark models (see example in Table 8). It is expected that squared bias should be small after calibration, and this is the case in the example of Table 7.

3.2 Comparison of protocol and usual calibration

None of the modeling teams, in their usual procedure, used the same procedure as the protocol for choosing parameters to estimate. In most cases, for usual calibration, the choice of parameters was based solely on expert opinion. A few teams used expert opinion but tested a few alternative choices to see which gave the best fit to the data. Finally, some teams based the choice on sensitivity analysis (Wallach et al. 2021c). Using usual or protocol calibration led to important differences in the calibrated models. For example, the number of estimated parameters in the final model was different between protocol and usual calibration (Supplementary Figure S1, Table S1).

The differences between simulated values after usual and protocol calibration were small for BBCH10, for which there were no calibration data. For the other stages, the simulated values differed appreciably (Supplementary Figure S2, Table S2).

Table 5 Example of documentation for protocol Step 4b, "Identify candidate parameters." These are parameters that seem likely to explain a substantial part of the variability between environments that remains after fitting the obligatory parameters. There is one row for each candidate, which should be in the order of presumed importance. This example is for the French data set, variety Apache for modeling team M21.

Candidate parameter	Brief explanation (units)	Default value (lower–upper bounds)
jvc	Number of vernalizing days (days)	38 (25–60)
sensrsec	Index of root sensitivity to drought (1=insensitive) (unitless)	0.5 (0–1)
belong	Parameter of the curve of coleoptile elongation (1/°days)	0.012 (0.005–0.03)
JVCmini	Minimum vernalizing days required (days)	7.0 (2–15)
stressdev	Maximum development delay allowed due to stresses (unitless)	0 (0–1)

Table 6 Example of documentation for Step 5 "Calculation of the optimal parameter values". The first line shows the optimization results for the obligatory parameters, and the resulting sum of squared errors and BIC criterion. Each subsequent line corresponds to a candidate parameter. If estimating the candidate together with the previously selected parameters leads to a decrease in BIC compared to the smallest value so far, the candidate is added to the list of parameters to estimate. If not, the candidate returns to its default value and is not considered further. In this example, the first candidate parameter (jvc) is accepted. All the subsequent candidate parameters increase BIC and are therefore rejected. The model finally chosen (minimum BIC) has three estimated parameters. This example is for the French data set, variety Apache, modeling team M21.

Estimated parameters	Initial parameter values	Final values	Sum of squared errors	BIC
stlevamf, stamflax	Multiple	227, 360	405	81.47
stlevamf, stamflax, jvc	227, 360, multiple	212, 367, 55.91	349	80.64
stlevamf, stamflax, jvc, sensrsec	212, 367, 55.91, multiple	209, 367, 58.40, 0.057	322	81.71
stlevamf, stamflax, jvc, belong	212, 367, 55.91, multiple	212, 367, 55.91, 0.012	349	83.97
stlevamf, stamflax, jvc, jvcmini	212, 367, 55.91, multiple	197, 362, 55.28, 20.88	319	81.45
stlevamf, stamflax, jvc, stressdev	212, 367, 55.91, multiple	212, 367, 55.91, 0.00	349	83.97

3.3 Goodness of fit and evaluation of prediction error for usual and protocol calibration

Nineteen modeling teams participated, and were all able to implement the protocol, showing that the protocol, though detailed, is nonetheless sufficiently flexible to be applicable to a wide range of models. The protocol was tested in comparison with the "usual" calibration procedure for each modeling team, which was possible because each modeling team had previously calibrated their model using the same data as here. To our knowledge, this is the first example of such a stringent test for a new calibration procedure. It provides a realistic test of whether the proposed protocol really improves calibration.

Figures 2 and 3 show RMSE using usual and protocol calibration for the French and Australian data sets respectively. Table 9 shows RMSE values for each data set, averaged over modeling teams, for usual and protocol calibration for the calibration and evaluation data (results by modeling team are in Supplementary Tables S3-S8). The protocol reduces RMSE by 10–22% compared to the usual calibration method. The p-values for a one-sided paired t-test of the hypothesis that RMSE is larger for usual calibration than for protocol calibration are also shown. On average over stages other than BBCH10, all three data sets have significantly larger RMSE values with usual calibration than with protocol calibration for the calibration data ($p<0.05$). For the evaluation data, the reduction in RMSE is highly significant ($p<0.01$) for the two French data sets, but less significant ($p=0.15$) for the Australian data set. Table 9 also shows the proportion of modeling teams where RMSE is larger for the usual calibration than for the protocol calibration. Looking at the averages over stages and then averaging over data sets, 75% of models have lower RMSE for protocol calibration than for usual calibration for the calibration data (Supplementary Tables S3, S5, S7). For the evaluation data, 60% of models have lower RMSE for protocol calibration than for usual calibration (Supplementary Tables S4, S6, S8). Presumably, a major reason that protocol calibration reduces RMSE compared to usual calibration is that the protocol uses an improved method of choosing the parameters to estimate, which combines expert knowledge and a statistical model selection criterion. Using a model selection criterion has the advantage that it avoids overfitting and in general will avoid estimation of parameters whose estimators are highly correlated.

Almost all modeling teams did better than the two benchmark models for all stages, both for usual calibration and protocol calibration, with slightly better results for protocol calibration (Supplementary Tables S3-S8). Since the protocol specifically aims to reduce bias, one would expect squared bias to be a smaller fraction of MSE for protocol calibration than for usual calibration, and this is the case, both for the calibration data and the evaluation data (Supplementary Tables S9-S23).

3.4 Between-model variability

The variability between simulated results for different modeling teams is shown in Table 10. The standard deviation is similar for usual and protocol calibration for BBCH10,

Table 7 First example of documentation for protocol Step 6, "Examine goodness-of-fit." In this table, there is one row for each observed variable with a simulated equivalent, showing mean squared error (MSE) and its decomposition into three terms. Of particular interest is the bias^2 contribution, which should be small if there is an almost obligatory parameter corresponding to this variable. This example is for the French data set, variety Apache, modeling team M21.

	MSE (days^2)	Bias^2 (days^2)	SDSD (days^2)	LCS (days^2)
BBCH30	19.64	0.25	5.93	13.46
BBCH55	5.29	0.02	0.03	5.24

Table 8 Second example of documentation for protocol Step 6, "Examine goodness-of-fit". In this table, there is one row for each observed variable with a simulated equivalent, showing root mean squared error (RMSE) for the calibrated model and for two benchmark models. The "naïve" benchmark assumes that all environments have the same number of days to the given development stage, equal to the average of the observed days to that stage. The "onlyT" benchmark assumes a constant number of degree days to the stage in question, with a simple vernalization calculation in the case of winter wheat. This example is for the French data set, variety Apache, modeling team M21.

	Naive	onlyT	M21
Variable	RMSE (days)	RMSE (days)	RMSE (days)
BBCH30	12.5	8.4	3.1
BBCH55	8.3	9.5	3.8

for which there are no data for calibration, but is systematically smaller for protocol calibration for the other stages. Considering the average over stages other than BBCH10 and taking the average over data sets, protocol calibration reduced the standard deviation of simulated values by 31% for the calibration data and by 22% for the evaluation data (see Table 10).

3.5 Comparison of usual and protocol calibration for ensemble of models

The choice of usual or protocol calibration has little effect on the predictive accuracy of the ensemble models e-mean and e-median. Averaged over development stages and over data sets, for the evaluation data, RMSE for e-median is respectively 5.7 and 5.8 days for usual and protocol calibration. The values for RMSE of e-mean are 6.1 and 6.2 days for usual and protocol calibration, respectively (Supplementary Tables S4, S6, S8).

Recently, many crop model studies have been based on ensembles of models (Jägermeyr et al. 2021). Many studies have found that the ensemble mean and median are good predictors, sometimes better than even the best individual model (Martre et al. 2015; Wallach et al. 2018; Farina et al. 2021). It has thus, become quite common to base projections of climate change impact on crop production on the

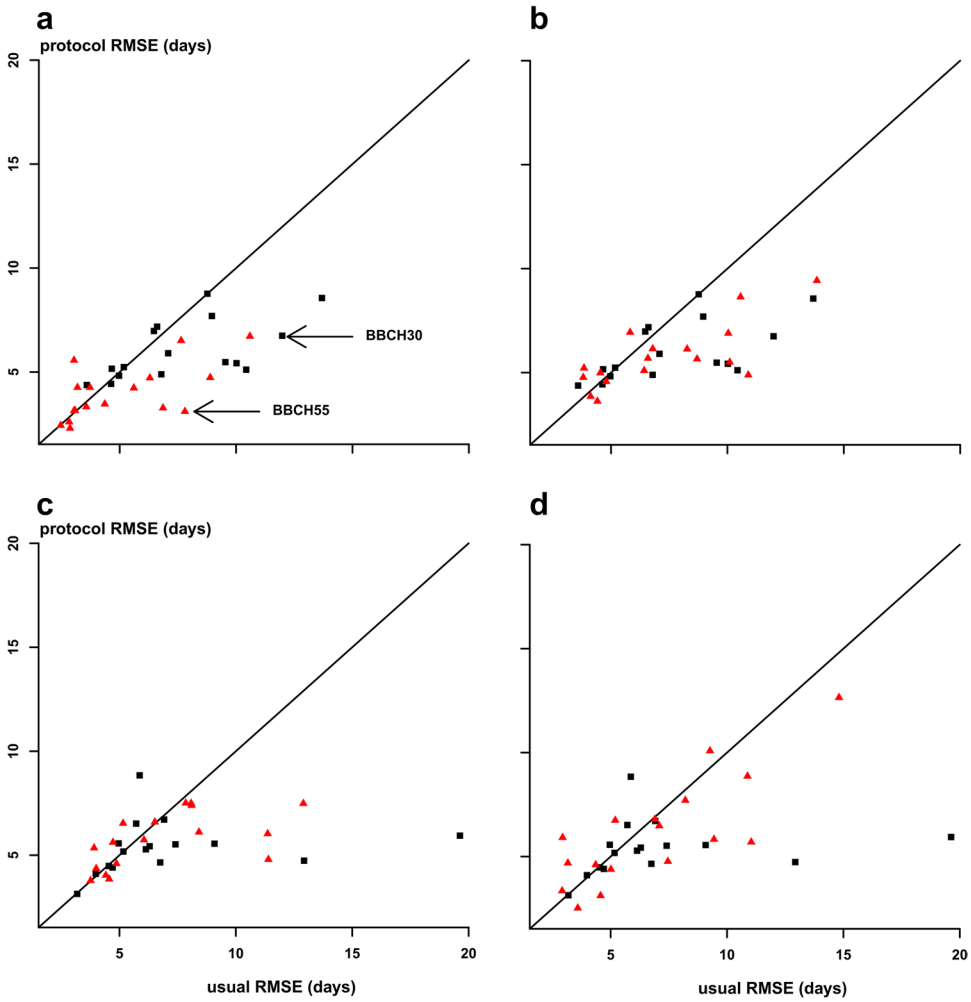

Fig. 2 Root mean squared error (RMSE) following protocol calibration versus RMSE following usual calibration, for each modeling group. Points below the diagonal indicate smaller RMSE for protocol calibration than for usual calibration. **a** French calibration data, variety Apache. **b** French evaluation data, variety Apache. **c** French calibration data, variety Bermude. **d** French evaluation data, variety Bermude.

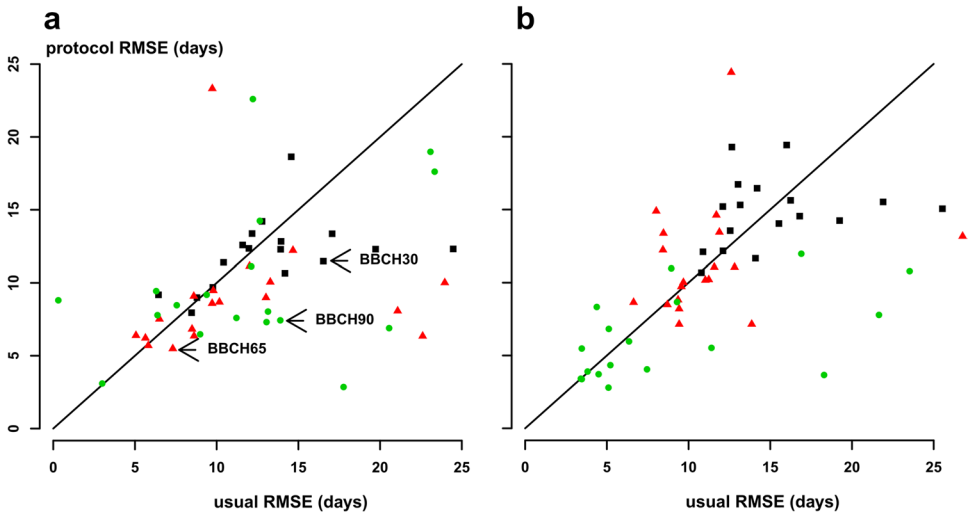

Fig. 3 Root mean squared error (RMSE) following protocol calibration versus RMSE following usual calibration, for each modeling group. Points below the diagonal indicate smaller RMSE for protocol calibration than for usual calibration. **a** Australian calibration data. **b** Australian evaluation data.

ensemble median (e.g., Asseng et al. 2019). The e-mean and e-median results here, compared to the individual modeling teams, are in keeping with previous results. The e-median model is among the better predictors though not the very best, and is somewhat better than e-mean. However, this does not imply that individual model results are unimportant. First, there will continue to be studies based on a single model or on a very small ensemble, and for those studies, it is important to improve individual models. Also, even for larger ensembles, it is important to reduce variability between models because this reduces variability in the ensemble results.

Table 9 Comparison of errors for usual and protocol calibration. The table shows root mean squared error (RMSE) averaged over modeling teams for each stage and for the average over stages, separately for the calibration and evaluation data. For each calibration and evaluation data, the first column is RMSE for simulations using the usual calibration method, the second column is RMSE using protocol calibration, and the third column is the *p*-value of a one-sided paired *t*-test that tests whether RMSE for usual calibration ($RMSE_u$) is larger than for protocol calibration ($RMSE_p$). Below the *p*-value is the fraction of modeling teams for which RMSE for usual calibration is larger than for protocol calibration.

		Calibration data			Evaluation data		
		Usual RMSE	Protocol RMSE	*p*-value	Usual RMSE	Protocol RMSE	*p*-value
France Apache	BBCH30	7.7	6	0.007 10/16	6.7	6.2	0.20 9/16
	BBCH55	5	4	0.02 11/17	7.3	5.8	0.005 13/17
	average	6.4	5	0.004 14/17	7	6	0.006 14/17
France Bermude	BBCH30	7.1	5.4	0.05 11/16	8.8	7.5	0.07 9/16
	BBCH55	6.8	5.7	0.04 11/17	6.9	6.1	0.08 11/17
	average	7	5.6	0.029 13/17	7.8	6.8	0.008 11/17
Australia	BBCH30	13.3	2.0	0.09 9/17	15.1	14.8	0.38 8/17
	BBCH65	11.4	9.0	0.06 14/19	11.1	11.4	0.60 10/19
	BBCH90	11.9	9.9	0.096 11/18	9.0	6.2	0.20 12/18
	average	12.2	10.2	0.049 13/19	11.7	10.7	0.15 10/19

Table 10 Variability of simulated values. This table shows the standard deviations of values simulated by the different modeling teams (days), for each simulated development stage and for the average over stages other than BBCH10. BBCH10 is not considered, since there were no observed values for BBCH10. Separate standard deviations are given for the calibration and evaluation data, and simulation using usual calibration or protocol calibration.

		Calibration data		Evaluation data	
		Usual	Protocol	Usual	Protocol
France Apache	BBCH10	4.2	4.1	4.8	5.2
	BBCH30	6.4	4.3	6.2	5.5
	BBCH55	4.5	3.0	6.3	3.7
	average	5.4	3.6	6.2	4.6
France Bermude	BBCH10	4.3	4.4	4.8	6.6
	BBCH30	6.9	4.3	6.7	6.2
	BBCH55	4.8	3.5	5.8	4.3
	average	5.9	3.9	6.2	5.3
Australia	BBCH10	8.6	7.5	9.6	8.0
	BBCH30	11.3	7.3	10.2	8.1
	BBCH65	10.3	7.4	8.2	7.1
	BBCH90	11.2	9.3	11.1	6.9
	average	10.9	8.0	9.8	7.4

4 Conclusions

We propose an original, detailed protocol for calibration of crop phenology models and showed that it can be applied by a wide range of wheat models and for data sets with different structures. The application here is to wheat phenology models. However, the same protocol could undoubtedly be used for phenology models of other crops.

This protocol could also be the basis for a calibration protocol for general crop models using multiple kinds of data. The procedure proposed here for the choice of parameters to estimate could be applied in the more general case. However, there would be the additional problem of dealing with multiple types of data, for example, phenology, biomass, yield, etc.

The protocol was tested with data sets representing a diversity of conditions. Comparison with usual calibration practices showed that on average over modeling teams, the protocol led to a better fit to the calibration data and to a better fit to out-of-sample data. The error of the mean or median of simulations was nearly identical with usual or protocol calibration, but the protocol substantially reduced between-model variability compared to usual calibration, which reduces the uncertainty of the mean or median. Thus, application of the protocol would be advantageous not just for individual modeling studies, but also for studies based on ensembles of models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13593-023-00900-0.

Authors' contributions DW did the conceptualization, methodology, project administration, writing (original draft), and model validation. TP, HM, PT, SB, SJS worked on conceptualization, methodology, project administration, and contributed to the writing, review & editing. ZH, EG, FA, BD, RF, TG, CG, MH, SH, HH, GH, P-EJ, QJ, EJ, K-CK, ML, EL, KL, FM, MM, CN, GP, BQ, NS, N-MS, VS, AM, XS, AKS, GT, WTKW, LW, TW conducted the model simulations, provided the model expertise, and contributed to writing, review, and editing. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was implemented as a co-operative project under the umbrella of the Agricultural Model Intercomparison and Improvement Project (AgMIP). This work was supported by the Academy of Finland through projects AI-CropPro (316172 and 315896) and DivCSA (316215) and Natural Resources Institute Finland (Luke) through a strategic project EFFI, the German Federal Ministry of Education and Research (BMBF) in the framework of the funding measure "Soil as a Sustainable Resource for the Bioeconomy - BonaRes", project "BonaRes (Module B, Phase 3): BonaRes Centre for Soil Research, subproject B" (grant 031B1064B), the BonaRes project "I4S" (031B0513I) of the Federal Ministry of Education and Research (BMBF), Germany, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 -390732324 EXC (PhenoRob), the Ministry of Education, Youth and Sports of Czech Republic through SustES - Adaption strategies for sustainable ecosystem services and food security under adverse environmental conditions (project no. CZ.02.1.01/0.0/0.0/16_019/000797), the Agriculture and Agri-Food Canada's Project J-002303 "Sustainable crop production in Canada under climate change" under the Interdepartmental Research Initiative in Agriculture, the JPI FACCE MACSUR2 project, funded by the Italian Ministry for Agricultural, Food, and Forestry Policies (D.M. 24064/7303/15 of 6/Nov/2015), and the INRAE CLIMAE meta-program and AgroEcoSystem department. The order in which the donors are listed is arbitrary.

Data availability The datasets generated during and/or analyzed during the current study are not publicly available but are available from the authors on reasonable request.

Code availability The codes (custom code) are available from the authors on reasonable request.

Declarations

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

Competing interests The authors declare no competing interests.

References

Ahuja LR, Ma L (eds) (2011) Methods of introducing system models into agricultural research. American Society of Agronomy and Soil Science Society of America, Madison, WI, USA. ISBN 13: 9780891181804

Asseng S, Ewert F, Rosenzweig C et al (2013) Uncertainty in simulating wheat yields under climate change. Nat Clim Chang 3:827–832. https://doi.org/10.1038/nclimate1916

Asseng S, Martre P, Maiorano A et al (2019) Climate change impact and adaptation for wheat protein. Glob Chang Biol 25:155–173. https://doi.org/10.1111/gcb.14481

Bassu S, Brisson N, Durand J-L et al (2014) How do various maize crop models vary in their responses to climate change factors? Glob Chang Biol 20:2301–2320. https://doi.org/10.1111/gcb.12520

Boogaard H, Wolf J, Supit I et al (2013) A regional implementation of WOFOST for calculating yield gaps of autumn-sown wheat across the European Union. F Crop Res 143:130–142. https://doi.org/10.1016/j.fcr.2012.11.005

Brisson N, Gary C, Justes E et al (2003) An overview of the crop model stics. Eur J Agron 18:309–332. https://doi.org/10.1016/S1161-0301(02)00110-7

Brisson N, Beaudoin N, Mary B, Launay; M. (2009) Conceptual basis, formalisations and parameterization of the STICS crop model. Quæ. ISBN 2759202909, 9782759202904

Buis S, Lecharpentier P, Vezy R et al. (2021) SticsRPacks/CroptimizR: v0.4.0. 10.5281/ZENODO.5121194

Chakrabarti A, Ghosh JK (2011) AIC, BIC and recent advances in model selection. Philos Stat 583–605. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

Chatelin MH, Aubry C, Poussin JC et al (2005) DéciBlé, a software package for wheat crop management simulation. Agric Syst 83:77–99. https://doi.org/10.1016/j.agsy.2004.03.003

Confalonieri R, Orlando F, Paleari L et al (2016) Uncertainty in crop model predictions: what is the role of users? Environ Model Softw 81:165–173. https://doi.org/10.1016/j.envsoft.2016.04.009

Coucheney E, Buis S, Launay M et al (2015) Accuracy, robustness and behavior of the STICS soil–crop model for plant, water and nitrogen outputs: evaluation over a wide range of agro-environmental conditions in France. Environ Model Softw 64:177–190. https://doi.org/10.1016/j.envsoft.2014.11.024

Coucheney E, Eckersten H, Hoffmann H et al (2018) Key functional soil types explain data aggregation effects on simulated yield, soil carbon, drainage and nitrogen leaching at a regional scale. Geoderma 318:167–181. https://doi.org/10.1016/j.geoderma.2017.11.025

Farina R, Sándor R, Abdalla M et al (2021) Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils. Glob Chang Biol 27:904–928. https://doi.org/10.1111/gcb.15441

Fath B, Jorgensen SE (2011) Fundamentals of ecological modelling: applications in environmental management and research. 4th edition. Elsevier, Amsterdam. ISBN 10: 0444535675 ISBN 13: 9780444535672

Fodor N, Challinor A, Droutsas I et al (2017) Integrating plant science and crop modeling: assessment of the impact of climate change on soybean and maize production. Plant Cell Physiol 58:1833–1847. https://doi.org/10.1093/pcp/pcx141

Gao Y, Wallach D, Hasegawa T et al (2021) Evaluation of crop model prediction and uncertainty using Bayesian parameter estimation and Bayesian model averaging. Agric For Meteorol 311:108686. https://doi.org/10.1016/j.agrformet.2021.108686

Gate, P (1995) Ecophysiologie du blé: de la plante à la culture. Lavoisier Editeur, Paris, France p 424

Herbst M, Hellebrand HJ, Bauer J et al (2008) Multiyear heterotrophic soil respiration: evaluation of a coupled CO2 transport and carbon turnover model. Ecol Modell 214:271–283. https://doi.org/10.1016/j.ecolmodel.2008.02.007

Holzworth DP, Huth NI, deVoil PG et al (2014) APSIM – evolution towards a new generation of agricultural systems simulation. Environ Model Softw 62:327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

Hoogenboom G, Porter CH, Boote KJ et al. (2019a) The DSSAT crop modeling ecosystem. In: Boote KJ (ed) Advances in crop modeling for a sustainable agriculture. Burleigh Dodds Science , Cambridge, United Kingdom, pp 173–216 https://doi.org/10.1201/9780429266591

Hoogenboom G, Porter CH, Shelia V et al. (2019b) Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.7. In: DSSAT found. Gainesville, Florida, USA. www.DSSAT.net, Accessed 05/07/2023

Jägermeyr J, Müller C, Ruane AC et al (2021) Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. Nat Food 2:873–885. https://doi.org/10.1038/s43016-021-00400-y

Jansson P-E (2012) CoupModel: model use, calibration, and validation. Trans ASABE 55:1337–1346. https://doi.org/10.13031/2013.42245

Jones JW, Hoogenboom G, Porter CH et al (2003) The DSSAT cropping system model. Eur J Agron 18:235–265. https://doi.org/10.1016/S1161-0301(02)00107-7

Keating B, Carberry P, Hammer G et al (2003) An overview of APSIM, a model designed for farming systems simulation. Eur J Agron 18:267–288. https://doi.org/10.1016/S1161-0301(02)00108-9

Kersebaum KC (2007) Modelling nitrogen dynamics in soil–crop systems with HERMES. Nutr Cycl Agroecosystems 77:39–52. https://doi.org/10.1007/s10705-006-9044-8

Kersebaum KC (2011) Special features of the HERMES model and additional procedures for parameterization, calibration, validation, and applications. In: Ahuja LR, Ma L (eds) Methods of introducing system models into agricultural research. American Society of Agronomy, Madison, pp 65–94. ISBN 13: 9780891181804

Khorashadi Zadeh F, Nossent J, Woldegiorgis BT et al (2022) A fast and effective parameterization of water quality models. Environ Model Softw 149:105331. https://doi.org/10.1016/j.envsoft.2022.105331

Klosterhalfen A, Herbst M, Weihermüller L et al (2017) Multi-site calibration and validation of a net ecosystem carbon exchange model for croplands. Ecol Modell 363:137–156. https://doi.org/10.1016/j.ecolmodel.2017.07.028

Kobayashi K (2004) Comments on another way of partitioning mean squared deviation proposed by Gauch et al. (2003). With reply. Agron J 96:1206–1207

Kuha J (2004) AIC and BIC: Comparisons of assumptions and performance. Sociol Methods Res 33:188–229. https://doi.org/10.1177/0049124103262065

Lawes RA, Huth ND, Hochman Z (2016) Commercially available wheat cultivars are broadly adapted to location and time of sowing in Australia's grain zone. Eur J Agron 77:38–46. https://doi.org/10.1016/j.eja.2016.03.009

Li T, Hasegawa T, Yin X et al (2015) Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob Chang Biol 21:1328–1341. https://doi.org/10.1111/gcb.12758

Martre P, Wallach D, Asseng S et al (2015) Multimodel ensembles of wheat growth: many models are better than one. Glob Chang Biol 21:911–925. https://doi.org/10.1111/gcb.12768

McNunn G, Heaton E, Archontoulis S et al (2019) Using a crop modeling framework for precision cost-benefit analysis of variable seeding and nitrogen application rates. Front Sustain Food Syst 3:108. https://doi.org/10.3389/fsufs.2019.00108

Meier, U (Ed.) (1997) BBCH-Monograph. Growth stages of plants. Entwicklungsstadien von Pflanzen. Estadios de las plantas. Stadesdedéveloppement des plantes. Blackwell Wissenschafts-VerlagBerlin, Wien p 622

Menzel A, Yuan Y, Matiu M et al (2020) Climate change fingerprints in recent European plant phenology. Glob Chang Biol 26:2599–2612. https://doi.org/10.1111/gcb.15000

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313. https://doi.org/10.1093/comjnl/7.4.308

Nendel C, Berg M, Kersebaum K, Mirschel W (2011) The MONICA model: testing predictability for crop growth, soil moisture and nitrogen dynamics. Ecol Model 222(9):1614–1625. https://doi.org/10.1016/j.ecolmodel.2011.02.018

Piao S, Liu Q, Chen A et al. (2019) Plant phenology and global climate change: current progresses and challenges. Glob Chang Biol gcb.14619. https://doi.org/10.1111/gcb.14619

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical recipes : the art of scientific computing, 3rd ed. Cambridge University Press, Cambridge ISBN 978-0-521-88068-8

Rafiei V, Nejadhashemi AP, Mushtaq S et al (2022) An improved calibration technique to address high dimensionality and non-linearity in integrated groundwater and surface water models. Environ Model Softw 149:105312. https://doi.org/10.1016/j.envsoft.2022.105312

Rezaei EE, Siebert S, Hüging H, Ewert F (2018) Climate change effect on wheat phenology depends on cultivar change. Sci Rep 8:4891. https://doi.org/10.1038/s41598-018-23101-2

Senapati N, Jansson P-E, Smith P, Chabbi A (2016) Modelling heat, water and carbon fluxes in mown grassland under multi-objective and multi-criteria constraints. Environ Model Softw 80:201–224. https://doi.org/10.1016/j.envsoft.2016.02.025

Sisheber B, Marshall M, Mengistu D, Nelson A (2022) Tracking crop phenology in a highly dynamic landscape with knowledge-based Landsat–MODIS data fusion. Int J Appl Earth Obs Geoinf 106:102670. https://doi.org/10.1016/j.envsoft.2016.02.025

Soltani A, Maddah V, Sinclair TR (2013) SSM-Wheat: a simulation model for wheat development, growth and yield. Int J Plant Prod 7:711–740. https://doi.org/10.22069/IJPP.2013.1266

Specka X, Nendel C, Wieland R (2015) Analysing the parameter sensitivity of the agro-ecosystem model MONICA for different crops. Eur J Agron 71:73–87. https://doi.org/10.1016/j.eja.2015.08.004

Specka X, Nendel C, Wieland R (2019) Temporal sensitivity analysis of the MONICA model: application of two global approaches to analyze the dynamics of parameter sensitivity. Agriculture 9:1–29. https://doi.org/10.3390/agriculture9020037

Stockle CO, Donatelli M, Nelson R (2001) CropSyst, a cropping systems simulation model. Eur J Agron 18:289–307. https://doi.org/10.1016/S1161-0301(02)00109-0.

Stuble KL, Bennion LD, Kuebbing SE (2021) Plant phenological responses to experimental warming – a synthesis. Glob Chang Biol gcb.15685. https://doi.org/10.1111/gcb.15685

Tao F, Rötter RP, Palosuo T et al (2018) Contribution of crop model structure, parameters and climate projections to uncertainty in climate change impact assessments. Glob Chang Biol 24:1291–1307. https://doi.org/10.1111/gcb.14019

van Bussel LGJ, Stehfest E, Siebert S et al (2015) Simulation of the phenological development of wheat and maize at the global scale. Glob Ecol Biogeogr 24:1018–1029. https://doi.org/10.1111/geb.12351

Vanuytrecht E, Raes D, Steduto P et al (2014) AquaCrop: FAO's crop water productivity and yield response model. Environ Model Softw 62:351–360. https://doi.org/10.1016/j.envsoft.2014.08.005

Wallach D (2011) Crop model calibration: a statistical perspective. Agron J 103:1144–1151. https://doi.org/10.2134/agronj2010.0432

Wallach D, Martre P, Liu B et al (2018) Multimodel ensembles improve predictions of crop-environment-management interactions. Glob Chang Biol 24:5072–5083. https://doi.org/10.1111/gcb.14411

Wallach D, Palosuo T, Thorburn P et al (2021) 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 et al (2021) Multi-model evaluation of phenology prediction for wheat in Australia. Agric For Meteorol 298–299:108289. https://doi.org/10.1016/j.agrformet.2020.108289

Wallach D, Palosuo T, Thorburn P et al (2021) 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

Wang E (1997) Development of a generic process-oriented model for simulation of crop growth. Utz, Wissenschaft. ISBN 3896752332, 9783896752338

Webber H, Lischeid G, Sommer M et al. (2020) No perfect storm for crop yield failure in Germany. Environ Res Lett 15: https://doi.org/10.1088/1748-9326/aba2a4

Wolf J (2012) User guide for LINTUL5: Simple generic model for simulation of crop growth under potential, water limited and

nitrogen, phosphorus and potassium limited conditions Research report, Wageningen University p 63
Zhang L, Zhang Z, Tao F et al (2022) Adapting to climate change precisely through cultivars renewal for rice production across China: when, where, and what cultivars will be required? Agric For Meteorol 316:108856. https://doi.org/10.1016/j.agrformet.2022.108856

Authors and Affiliations

Daniel Wallach[1] · Taru Palosuo[2] · Peter Thorburn[3] · Henrike Mielenz[4] · Samuel Buis[5] · Zvi Hochman[3] · Emmanuelle Gourdain[6] · Fety Andrianasolo[6] · Benjamin Dumont[7] · Roberto Ferrise[8] · Thomas Gaiser[1] · Cecile Garcia[6] · Sebastian Gayler[9] · Matthew Harrison[10] · Santosh Hiremath[11] · Heidi Horan[3] · Gerrit Hoogenboom[12,13] · Per-Erik Jansson[14] · Qi Jing[15] · Eric Justes[16] · Kurt-Christian Kersebaum[17,18,27] · Marie Launay[19] · Elisabet Lewan[20] · Ke Liu[10] · Fasil Mequanint[9] · Marco Moriondo[21] · Claas Nendel[17,18,22] · Gloria Padovan[8] · Budong Qian[15] · Niels Schütze[23] · Diana-Maria Seserman[17] · Vakhtang Shelia[12,13] · Amir Souissi[24] · Xenia Specka[17] · Amit Kumar Srivastava[1] · Giacomo Trombi[8] · Tobias K. D. Weber[9,28] · Lutz Weihermüller[25] · Thomas Wöhling[23,26] · Sabine J. Seidel[1]

✉ Sabine J. Seidel
sabine.seidel@uni-bonn.de

1 Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany
2 Natural Resources Institute Finland (Luke), Helsinki, Finland
3 CSIRO Agriculture and Food, Brisbane, Queensland, Australia
4 Institute for Crop and Soil Science, Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Braunschweig, Germany
5 INRAE, UMR 1114 EMMAH, Avignon, France
6 ARVALIS - Institut du végétal Paris, Paris, France
7 Plant Sciences & TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium
8 Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Florence, Italy
9 Institute of Soil Science and Land Evaluation, Biogeophysics, University of Hohenheim, Stuttgart, Germany
10 Tasmanian Institute of Agriculture, University of Tasmania, Launceston, Tasmania, Australia
11 Aalto University School of Science, Espoo, Finland
12 Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL, USA
13 Global Food Systems Institute, University of Florida, FL, Gainesville, USA
14 Royal Institute of Technology (KTH), Stockholm, Sweden
15 Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada
16 PERSYST Department, CIRAD, Montpellier, France
17 Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany
18 Global Change Research Institute CAS, Brno, Czech Republic
19 INRAE, US 1116 AgroClim, Avignon, France
20 Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
21 CNR-IBE, Firenze, Italy
22 Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany
23 Institute of Hydrology and Meteorology, Chair of Hydrology, Technische Universität Dresden, Dresden, Germany
24 Swift Current Research and Development Centre, Agriculture and Agri-Food Canada, Swift Current, Saskatchewan, Canada
25 Institute of Bio- and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
26 Lincoln Agritech Ltd., Hamilton, New Zealand
27 Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), University of Göttingen, Göttingen, Germany
28 Present Address: Faculty of Organic Agriculture, Soil Science Section, University of Kassel, Witzenhausen, Germany