Calibration of crop phenology models: Going beyond recommendations

Running title: Model calibration: Beyond recommendations

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[9], Cecile Garcia[6], Sebastian Gayler[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], Marie Launay[19], Elisabet Lewan[20], Fasil Mequanint[10], 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[9], Giacomo Trombi[8], Tobias K.D. Weber[10], Lutz Weihermüller[25], Thomas Wöhling[22,26], Matthew Harrison[27], Ke Liu[27], Sabine J. Seidel [9*]

[1]INRAE, UMR AGIR, Castanet Tolosan, France. ORCID 0000-0003-3500-8179

[2]Natural Resources Institute Finland (Luke), Helsinki, Finland

[3]CSIRO Agriculture and Food, Brisbane, Queensland, Australia

[4]Institute for Crop and Soil Science, Federal Research Centre for cultivated Plants, Julius Kühn-Institut (JKI), Braunschweig, Germany

[5]INRAE, UMR 1114 EMMAH, Avignon, France

[6]ARVALIS - Institut du végétal 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, Italy

[9]Institute of Crop Science and Resource Conservation, University of Bonn, Germany

[10]Institute of Soil Science and Land Evaluation, Biogeophysics, University of Hohenheim, Stuttgart, Germany

[11]Aalto University School of Science, Espoo, Finland;

[12]Agricultural and Biological Engineering Department, University of Florida, Gainesville, Florida, USA

[13]Food Systems Institute, University of Florida, Gainesville, Florida, USA

[14]Royal Institute of Technology (KTH), Stockholm, Sweden

[15]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada

[16]CIRAD, UMR SYSTEM, 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]Tasmanian Institute of Agriculture, University of Tasmania Launceston, Australia

[*]Correspondence to sabine.seidel@uni-bonn.de

Abstract

A major effect of environment on crops is through crop phenology, and therefore, the capacity to predict phenology as a function of soil, weather, and management is important. Mechanistic crop models are a major tool for such predictions. It has been shown that there is a large variability between predictions by different modeling groups for the same inputs, and therefore, a need for shared improvement of crop models. Two pathways to improvement are through improved understanding of the mechanisms of the modeled system, and through improved model parameterization. This article focuses on improving crop model parameters through improved calibration, specifically for prediction of crop phenology. A detailed calibration protocol is proposed, which covers all the steps in the calibration work-flow, 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. For those aspects where knowledge of the model and target environments is required, the protocol gives detailed guidelines rather than strict instructions. The protocol includes documentation tables, to make the calibration process more transparent. The protocol was applied by 19 modeling groups to three data sets for wheat phenology. All groups first calibrated their model using their "usual" calibration approach. Evaluation was based on data from sites and years not represented in the training data. Compared to usual calibration, calibration following the new protocol significantly reduced the error in predictions for the evaluation data, and reduced the variability between modeling groups by 22%.

Key words

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 is a dominant aspect of plant ecology (Cleland et al. 2007) and has been shown to be affected by warming (Piao et al. 2019; Menzel et al. 2020; Stuble et al. 2021) and phenology of cultivated crops. 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, in particular as a function of climate.

A number of mechanistic crop models have been developed, which include simulation of phenology, and 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. 2014) and others. 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 (we will refer to such models as crop phenology models, though they are usually embedded within more general crop models), 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 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 is essentially universally 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 models. The first is through improvement of model structure through improved understanding of the underlying processes, and the second is through improvement of the model parameters. For a fixed data set, improvement of model parameters implies improvement of model calibration, and that is the topic here.

Crop models, like system models in general, are basically regression models, in that they predict outcomes of the system based on input variables. While calibration of regression models (usually referred to as parameter estimation in a statistical context) is a major topic in statistics, the application of statistical methods to system models is not straightforward. Among the difficulties are the fact 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) and 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, not only crop phenology models but also full crop models, hydrological models, ecology models and models in other fields. 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 and model outputs for crop phenology models furnished with identical data, even between modeling groups using the same model structure (Confalonieri et al. 2016; Wallach et al. 2021a, b).

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 emphasized the need to consider the full flow of the calibration exercise, including the choice of default parameter values, the choice of observed data to include in the objective function, the form of the objective function, the choice of parameters to estimate, and the choice of numerical algorithm for calculating the best parameter values, and proposed recommendations for each of those activities (Wallach et al., 2021c).

Recommendations for improved calibration are useful, but have important limitations. Not only do they generally concern only part of the calibration activity, but in addition they are generally not tested for a wide range of situations, to verify that they have general applicability, and they generally do not have any mechanism for ensuring that they are correctly followed. To date, recommendations have not resulted in a convergence of calibration practices for crop phenology models.

The purpose of this study is to propose, implement, and evaluate a calibration protocol for crop phenology models that does not have the above-mentioned limitations. Compared to usual recommendations, the protocol suggested here, shown schematically in Figure 1, is more detailed, covers the full range of decisions involved in calibration, and includes documentation templates that help in correctly implementing the protocol and make the calibration strategy more transparent. To test the applicability of the protocol to a range of model structures, multiple modeling groups implemented the protocol, using multiple model structures. To test the applicability to differently structured data sets, each modeling group applied the protocol to three different wheat phenology data sets, with two quite different structures. The most important, and exacting, test of the proposed calibration protocol is its ability to improve predictions of phenology for out-of-sample environments, compared to usual calibration practice. All of the modeling groups that applied the protocol in this study first used their usual method of calibration on the same data sets as those used here, in most cases as part of previous studies(Wallach et al. 2021a, b). Thus, we were able to compare the predictive accuracy of usual calibration with calibration using the protocol proposed here. It is also important to reduce the uncertainty in phenology predictions, as measured by the variability between different modeling groups. Therefore, we also compared the variability between groups for the case where each group uses their usual calibration approach or the protocol.

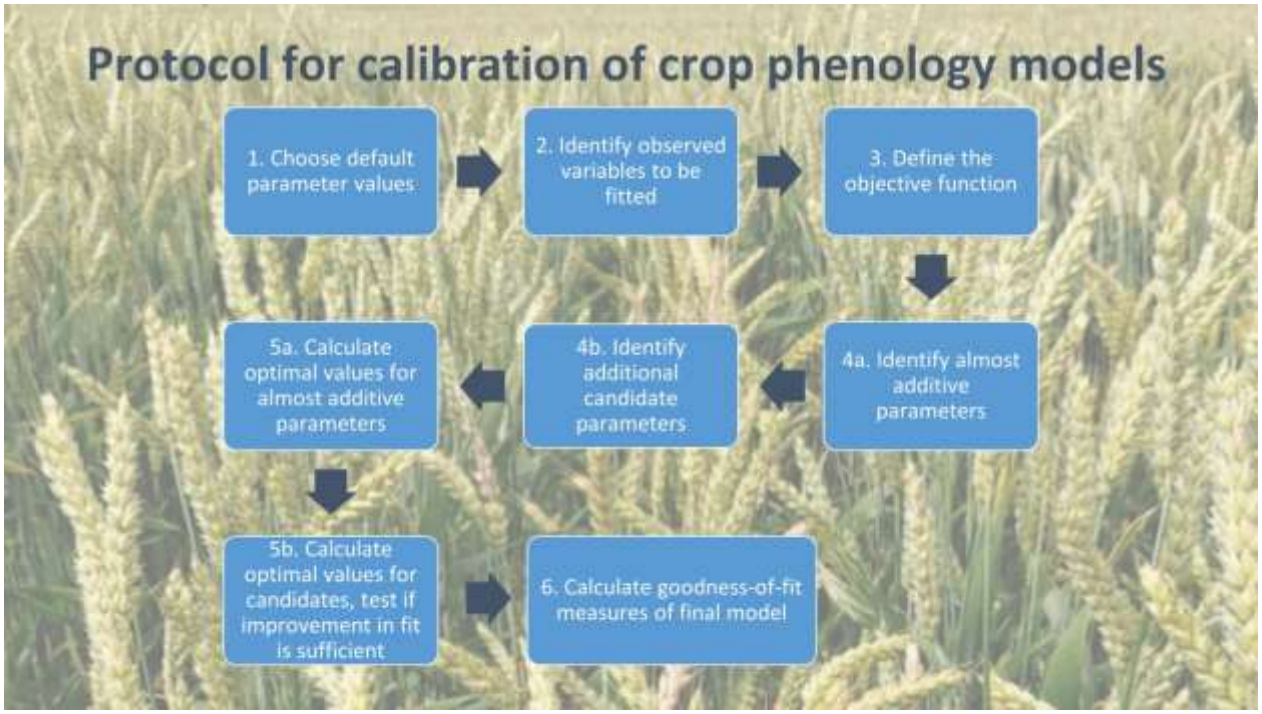
Protocol for calibration of crop phenology models
1. Choose default parameter values
2. Identify observed variables to be fitted
3. Define the objective function
4a. Identify almost additive parameters
4b. Identify additional candidate parameters
5a. Calculate optimal values for almost additive parameters
5b. Calculate optimal values for candidates, test if improvement in fit is sufficient
6. Calculate goodness-of-fit measures of final model

Figure 1: Schematic diagram of steps in the proposed protocol for calibration of crop phenology models.

2. Materials and Methods

Data sets

The phenology stages targeted in this study, using the BBCH scale (Meier 1997)., are BBCH10 (emergence), BBCH30 (beginning of stem elongation), BBCH55 (middle of heading), BBCH65 (anthesis half way) and BBCH90 (fully ripe grain). Simulations were performed for three data sets. The two French data sets had a similar structure, but concerned two different winter wheat varieties. The third data set was from Australia, for a spring wheat variety. In each case, the data could be considered to come from a well-defined target population: conventionally managed wheat fields in the major wheat growing regions of France for the first two data sets, and of Australia for the third data set. The Australian data set included multiple planting dates from within the range of reasonable dates. Each data set was split into two subsets, one for calibration and one for evaluation. The two subsets had neither site nor year in common, so the evaluation is a rigorous test of how well a modeling group can simulate phenology for out-of-sample environments.

The French data sets contained observations of stages BBCH30 and BBCH55. For each variety there were 14 calibration environments and 8 evaluation environments, where an environment is a combinations of site and sowing date. More details about the French data set can be found in Wallach et al (2021a). The Australian data set contained once weekly notations of BBCH stage in each environment. The Australian data were interpolated to give the date of each integer BBCH stage from the earliest that was observed to the latest, and these interpolated data were provided to the modeling groups. The interpolated data were provided to avoid each modeling group doing their own interpolation, which would add unwanted variability to the exercise. The Australian data set had 24 calibration environments and 18 evaluation environments. For more details see Wallach et al (2021b).

Modeling groups

Nineteen modeling groups, 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 groups 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 codes here refer to the same modeling groups as in Wallach et al. (2021a, 2021b).

Table 1: List of model structures used by participating groups

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	(P.-E. 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; Kersebaum, 2011)

Evaluation

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

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

The sum is over variables and 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 (bias²), 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 vales (LCS) (Kobayashi 2004).

In addition, we define 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, introduced 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, and then the fixed number of degree days applies after vernalization is completed (van Bussel et al. 2015; Wallach et al. 2021a). Both benchmark models are quite easily parameterized based on calibration data, and then easily applied to new environments.

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, and then to simulate and report days after sowing to stages BBCH10, 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 (BBCH10) 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.

Seventeen of the 19 participating modeling groups participated in previous phases of this project, and in that framework had already calibrated their model using their usual calibration approach (Wallach et al. 2021a, b). The two remaining groups also calibrated their models using their usual approach before beginning to use the protocol proposed here. It is the results of the usual calibration method that are compared here to the results of using the proposed protocol. If any simulated results were missing from the usual calibration results, the corresponding results from the protocol calibration results were deleted, and vice versa. Thus, the results for the usual and protocol calibrations are comparable. At no time were the evaluation data shown to participants, neither in previous studies nor in the present study. To keep these data confidential, and thus, potentially useable in future studies, no graphs are presented here showing the evaluation data.

In the present study, participants were given the detailed protocol for model calibration and asked to implement it for the three data sets. The protocol was the same as described below, but the textual description was somewhat different. The team leaders used the protocol documentation generated by the modeling teams to identify problems and interact with the participants.

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.

AICc and BIC

The protocol uses a model selection criterion to decide which parameters to estimate. The corrected Akaike Information Crition (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

$$AICc = n\ln(MSE) + 2p + \frac{2p(p+1)}{n-p-1} \tag{2}$$

$$BIC = n\ln(MSE) + p\ln(n)$$

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, so any term that only involves n, the number of data points, has been dropped because it is the same for all models.

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 22-23). All results shown here are based on the BIC criterion.

The protocol

The proposed protocol follows the recommendations from Wallach et al. (2021c), but with additional details to extend, standardize, and document the implementation of those recommendations. The protocol covers all the steps involved in calibration, with detailed instructions for each step.

Step 1. Choose default parameter values

The first step in the protocol is to define the default values for all parameters. This step is rarely if ever mentioned in discussions of system model calibration. It is however very important since most parameters retain their default values. Furthermore, the protocol calculations work best if the default values of the parameters to be estimated are reasonably close to the new best values. 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. The documentation for step 1 (see example in Table 2) specifies the cultivar which is used to provide default parameter values, and why that cultivar was chosen.

Table 2: Example of protocol documentation for step 1, showing cultivar characteristics of the target population cultivar and of the cultivar that provides the default parameter values. This example is for the French data set and modeling group M21.

Variety	Characteristics
Variety of target population: 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°C and 10°C. There is no vernalization below -4°C or above 17°C. Otherwise there is a proportional reduction in vernalization effectiveness.
Default variety : Soissons	Soissons seems to be close to Apache in terms of vernalization requirements and earliness.

Step 2. Identify measured variables to be fit

Here one lists each of the observed variables and the corresponding simulated variables if any. In the simplest case, there is a simulated variable that corresponds directly to each observed variable. The documentation for step 2 is a table with one row for each observed variable (Table 3).

Often this step is straightforward. In some cases, however, there may be stages for which simulations are required, and for which there are observations, but which are not specifically identified in the model. In the study here, for example, several models do not specifically simulate the stage BBCH30, which was observed in the French data set and which was to be simulated. Most models do, however, simulate an internal growth stage variable. One can then treat the internal growth stage that corresponds to BBCH30 as a new parameter, to be estimated. This is the recommendation of the protocol. This approach makes it possible for a much wider range of models to use all the data for calibration, than if only observed variables specifically simulated by the model were used.

Table 3*:* Example of documentation of protocol step 2 with one row for each measured variable, showing the corresponding simulated variable. This example is for the French data set and modeling group 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 3. Define the objective function

The protocol uses as objective function to be minimized SSE, where the sum over variables is over the observed variables from step 2 that have a simulated equivalent. No choices are required here. Once step 2 has been finalized, the objective function follows.

In discussing system models in hydrology, Hernandez-Suarez et al. (2021) distinguish two categories of objective functions, one based on traditional performance metrics and the second adapted to the specific aspects of the system that are of primary interest. This choice applies generally to system models, and here we opt for the first possibility. The objective function of the protocol is simply the total sum of squared errors, which is the objective function of ordinary least squares (OLS) regression and is often used in crop model calibration. Another major choice of the protocol is to include in the objective function all the observed variables that have a simulated equivalent, including variables that are not of primary interest. For example, in the study here, the Australian data set has observations of many stages. The protocol says to use all possible data, and in fact, most modeling groups did indeed use observations of other stages for calibration, in addition to data for the BBCH stages 30, 65, and 90, which are used for evaluation. A first reason for using all data 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 more variables makes the model a better representation of multiple aspects of the system dynamics, which is likely to improve all simulations.

OLS has attractive optimal properties if the standard assumptions are satisfied (Seber and Wild 1989). It has been argued that crop models probably do not fully satisfy these assumptions, not even the first assumption that the model is correctly specified, i.e. that there are parameter values such that the expectation of model error is zero (Wallach 2011). However, even in this case, OLS has desirable properties (White 1981).

The assumptions underlying the optimal properties of OLS also include homoscedasticity (equal variances of error for all variables) and independence between errors for all variables. Once again, these assumptions are probably not fully satisfied for crop phenology models. If problems of heteroscedasticity (unequal error variances) and non-independence are ignored, the calibration does not make optimal use of the data. One could correct the problem of heteroscedasticity by using weighted least squares, and the problem of non-independence by using generalized least squares (Seber and Wild 1989), but in both cases that would require estimating additional parameters, related to the standard deviation or to the variance-covariance matrix of model error. The assumption here is that the problems of heteroscedasticity and non-independence are often not too important for crop phenology models, since the variables are all times to various development stages and thus, may have similar error variances, and different environments often have similar amounts of data, so taking non-independence into account would not greatly change the relative importance of different environments. The choice in the protocol is, therefore, to use OLS and to avoid estimating additional parameters related to variance and covariance of errors.

An alternative approach would be to do a Bayesian calculation, where one calculates the posterior parameter distribution rather than a single best parameter vector. However, this is more difficult computationally, and furthermore requires specification of the distribution of model error, including variances and covariances. Also, while in principle one can estimate any number or parameters using a Bayesian approach, in practice there is always a choice of which parameters to include in the calculation, as with the approach here.

Step 4. Choose which parameters to estimate

Usual calibration approaches include various methods of choosing which parameters to estimate, including sensitivity analysis, expert opinion, and testing different options to find the best fit to the data (Wallach et al. 2021c). The protocol here combines expert opinion and a data-based criterion of choice.

The protocol 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 nearly additive, 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. Usually, a parameter that represents degree days to a measured stage is a good choice as an obligatory parameter for time to that stage. If the calculation of a variable in the model includes a constant (i.e. there is an exactly additive parameter), then estimating that constant using OLS ensures that model bias will be exactly zero (i.e. the average of observed and simulated values will be equal), An approximately additive parameter will almost eliminate bias. Reducing bias reduces MSE for the calibration data, since squared bias is one of the three terms that add up to MSE (Kobayashi and Salam 2000). Once bias is nearly eliminated, one may already have a fairly reasonable fit to the data.

Ideally, the number of almost additive parameters will be identical to the number of variables in the objective function. The number of almost additive parameters cannot be greater than the number of variables in the objective function, and each must be nearly additive for a different variable or combination of variables. 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 documentation for protocol step 4a has one row for each ,early additive parameter (Table 4), which gives the default value and upper and lower bounds. Our definition of upper and lower bounds is that the modeler would be very surprised if the true best value for this target population were outside these bounds. This is admittedly a rather vague definition, but it is meant to translate the fact that while we do not know the true best value of each parameter, we do have some knowledge about a reasonable range.

Table 4*:* Example of documentation for protocol step 4a, showing obligatory parameters for the French data set, variety Apache for modeling group M21.

Name of 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 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 parameters that seem likely, when estimated, to explain a substantial part of the unexplained variability between environments, and to order them by amount of variability likely to be explained. Each parameter is associated with a particular process, so the choice here is of the process and then of the specific parameter in the equations describing that process. For example, in Table 5, the first and last candidate parameters are both associated with crop vernalization, but affect different aspects of that process.

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. There should be only a limited number of candidate parameters, with those thought most important first, in order to reduce the risk of selection bias where unimportant parameters are chosen because of random error (Lukacs et al. 2010). The choice of almost additive and candidate parameters is the calibration step which requires the most detailed knowledge of the model.

Table *5:* Example of documentation for protocol step 4b, showing candidate parameters for the French data set, variety Apache for modeling group M21.

Candidate parameter (include units and a brief explanation)	Default value (Lower and upper bounds)
jvc (days): number of vernalizing days	38 (25 – 60)
sensrsec (no unit): index of root sensitivity to drought (1=insensitive)	0.5 (0 – 1)
belong (degree.day-1): parameter of the	0.012 (0.005 – 0.03)

curve of coleoptile elongation	
JVCmini (days): minimum vernalizing days required	7.0 (2 – 15)
stressdev (no unit): maximum phasic delay allowed due to stresses	0 (0 – 1)

Step 5. Calculation of the optimal parameter values

The protocol prescribes using an algorithm based on the Nelder-Mead simplex (Nelder and Mead 1965), to minimize SSE. This is a robust, derivative-free method, which is appropriate for crop models which may have multiple discontinuities.

The optimization is done in several steps, first for the obligatory parameters and then for each candidate parameter. At each step, one estimates the previously chosen parameters plus the next candidate parameter. If this leads to a smaller value of the BIC criterion than the previous best model, the candidate parameter is added to the list of parameters to estimate. If not, the candidate parameter returns to its default value and is not added to the list of parameters to estimate. The model with the lowest value of BIC is the final model. Table 6 shows the documentation of protocol step 5, which has one row for each minimization step.

The results of the simplex are sensitive to starting values (Wang and Shoup 2011), and there is no guarantee that the algorithm will converge to the global optimum. Therefore, the protocol calls for multiple starting points. For the initial step, with more than one parameter to estimate, the protocol recommends using at least 20 starting values, chosen within the lower to upper bound for each parameter. Latin hypercube sampling could be used to distribute the starting values within the space of possible values. When a new candidate parameter is estimated, it is expected that the optimal values of the other estimated parameters will often be similar to their previous optimal values. Therefore, the protocol recommends starting from the parameter values that gave the previous lowest BIC value, and five different starting values for the new parameter. In this way the calculations take advantage of knowledge of the system to focus the search for best parameters on a limited portion of the parameter space, thus reducing calculation time.

Table *6:* Example of documentation for protocol step 5, showing the calculations for the French data set, variety Apache, modeling group M21. The first line shows the optimization results for the obligatory parameters, then each subsequent line corresponds to a candidate parameter. In this example, the first candidate parameter (jvc) is accepted, and all the subsequent candidate parameters increase BIC, and are therefore, rejected. The model finally chosen (minimum BIC) has three estimated parameters.

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

Step 6 Diagnostics

a. The protocol evaluation graph shows simulated versus observed values, with a different symbol for each variable (Figure 1).

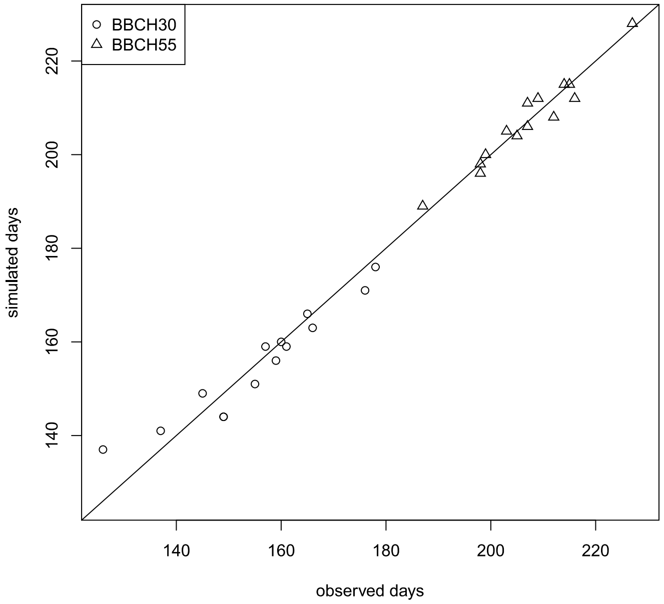

Figure 2*:* Graph for protocol documentation of step 6a, showing simulated and observed values of BBCH30 and BBCH55. The results are for the French data set, variety Apache, modeling group M21.

b. Calculate MSE and its components for each variable (7)

Table *7:* Example of documentation of protocol step 6b, showing the calculations for the French data set, variety Apache, modeling group M21. Mean squared error (MSE) is a sum of bias², SDSD, and LCS.

	MSE (days²)	bias² (days²)	SDSD (days²)	LCS (days²)
BBCH30	19.64	0.25	5.93	13.46
BBCH55	5.29	0.02	0.03	5.24

c. Compare results to benchmark models (Table 8).

Table 8*:* Example of documentation for protocol step 6c, showing results for benchmark models for the French data set, variety Apache, modeling group M21. The benchmark “naive” assumes a constant number of days to each stage. The benchmark “onlyT” assumes a constant temperature sum to each stage.

	naive	onlyT	M21
variable	RMSE (days)	RMSE (days)	RMSE (days)
BBCH30	12.5	8.4	3.1
BBCH55	8.3	9.5	3.8

3. Results and discussion

Comparison of usual and protocol calibration for individual models

Number of parameters and simulated values

There were substantial differences between the consequences of doing usual or protocol calibration. For the large majority of modeling groups, the number of estimated parameters in the final model was different between protocol and usual calibration (Supplementary Figure 1). Table 9 shows that on average over modeling groups, protocol calibration led to fewer parameters (French data sets) or the same number of parameters (Australian data set) to be estimated in comparison to usual calibration. The modeling groups considered a larger number of parameters in the protocol calibration than in the usual calibration, but then rejected most of the candidates.

Table 9*:* Number of estimated parameters in the final model after protocol or usual calibration, averaged over modeling groups.

Data set	Protocol calibration				Usual calibration
	Additive parameters	Candidate parameters	Accepted candidate parameters	Total parameters	Total parameters
France Apache	1.9	3.6	0.7	2.6	3.5
France Bermude	1.9	3.6	0.9	2.8	3.6
Australia	4.0	4.5	1.7	5.7	5.7

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. The mean absolute difference between simulated values using usual or protocol calibration is shown in Table 10. Results by modeling group are shown in Supplementary Figure 2.

Table 10*:* Mean over modeling groups and environments of the absolute difference in simulated values between usual calibration and protocol calibration for the calibration data.

French data set (Apache and Bermude)	stage	BBCH10	BBCH30	BBCH55	
	Mean absolute	0.8	4.5	3.1	

	difference (days)				
Australian data set	stage	BBCH10	BBCH30	BBCH65	BBCH90
	Mean absolute difference (days)	1.8	7.9	7.9	8.5

Goodness of fit and evaluation

Figure 3 and 4 and Supplementary Tables 2-7 show RMSE using usual and protocol calibration for the French and Australian data sets respectively. If a point is below the diagonal, RMSE is smaller for protocol calibration than for usual calibration.

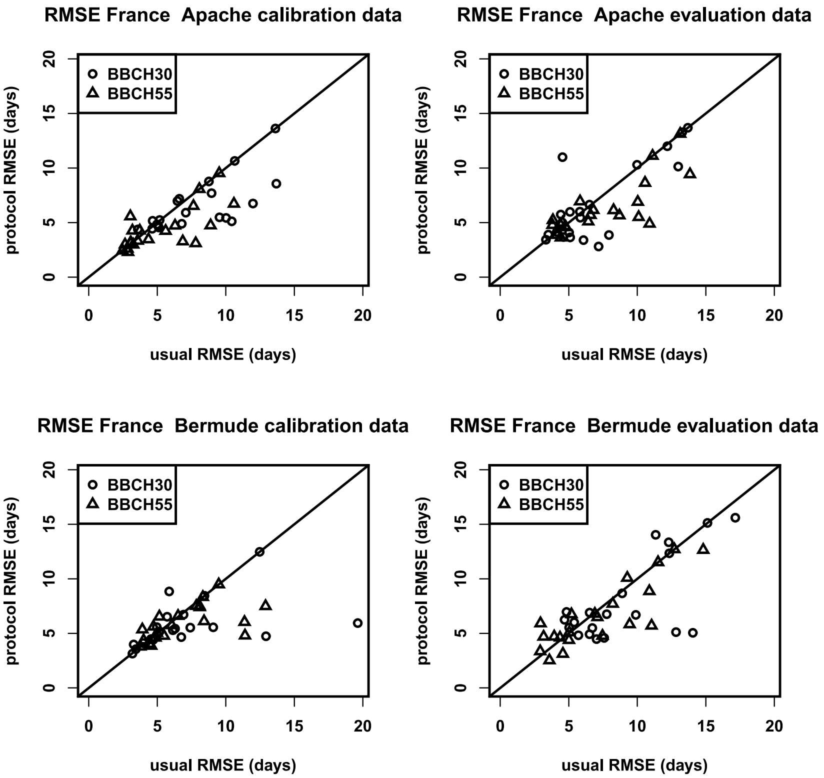

Figure 3: RMSE for usual and protocol calibration, for each French datasets, for calibration and evaluation data.

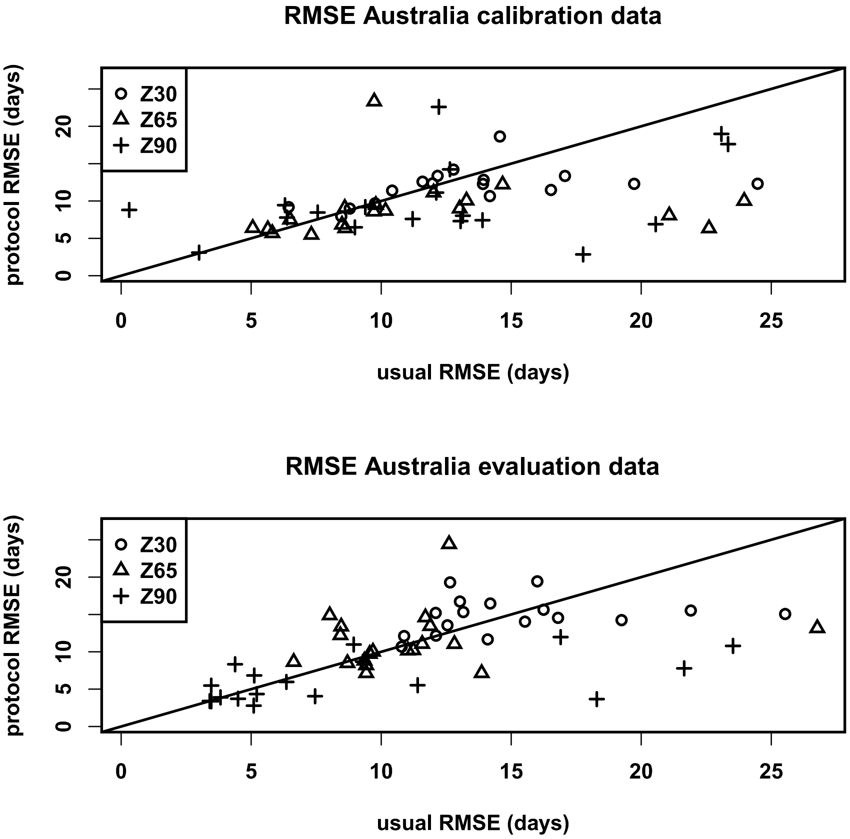

Figure 4: RMSE for usual and protocol calibration, for each Australian dataset, for calibration and evaluation data.

Table 11 shows RMSE values for each data set, averaged over modeling groups, for usual and protocol calibration for the calibration and evaluation data. 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, p<0.01 for the two French data sets, but $p=0.15$ for the Australian data set. The table also shows the proportion of modeling groups 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, and 66% for the evaluation data.

The evaluation data here are from environments similar to those of the calibration data. This has the important advantage of being a standard situation that does not introduce the additional complication of the degree of dissimilarity between the calibration and evaluation environments. However, it leaves open the question of how useful protocol calibration would be if the models were used to extrapolate to quite different conditions, such as projected future climate. Nonetheless, the improvement here for out-of-sample predictions is certainly encouraging.

Almost all modeling groups did better than the two benchmark models for all stages. Averaged over stages, for the evaluation data, only 2, 1, and 4 modeling groups using usual calibration did worse than the onlyT model for the French Apache, French Bermude, and Australian data sets, respectively. Using protocol calibration, 0, 1, and 1 modeling groups did worse than the onlyT model for the French Apache, French Bermude, and Australian data sets respectively.

Table 11*:* RMSE averaged over modeling groups for each stage and averaged over stages for the calibration and evaluation data, using usual or protocol calibration, for each data set. The p value is for the test that RMSE using usual calibration is larger than RMSE using protocol calibration, and below that is the proportion of models for which RMSE using usual calibration is larger than RMSE using protocol calibration.

		Calibration data			Evaluation data		
		Usual RMSE	protocol RMSE	*p*-value $RMSE_u$>$RMSE_p$	usual RMSE	protocol RMSE	*p*-value $RMSE_u$>$RMSE_p$
France	BBCH3	7.7	6	0.007	6.7	6.2	0.20

Apache	0			10/16			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	12.0	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

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 9-24).

Comparison of usual and protocol calibration for ensemble

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 (not including BBCH10) 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 3, 5, 7, 8).

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). It has, thus, become quite common to base projections of climate change impact on crop production on the ensemble median (e.g. Asseng et al., 2019). The e-mean and e-median results here, compared to the individual modeling groups, 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.

The variability between simulated results for different modeling groups is shown in Table 12. The standard deviation is similar for usual and protocol calibration for BBCH10, 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.

Table 12.*:* Standard deviation of values simulated by modeling groups (days). The average is over stages without BBCH10.

		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

The variability among simulated values is a measure of uncertainty in impact assessments (Asseng et al. 2013). This variability arises from differences in model structure, but also from parameter uncertainty and uncertainty in climate projections (Tao et al. 2018) The variability due to the method of calibration has not been specifically studied. The comparison here between the variability after usual and protocol calibration indicates whether, and how much, variability between modeling groups can be reduced if all groups apply the same calibration procedure. The results show that using a standard calibration approach can be an effective way of substantially reducing variability of crop phenology model simulations.

4. Conclusion

We have shown that it is possible to define a detailed protocol for calibration of crop phenology models that is applicable to a wide range of wheat models and to data sets with different structures. The protocol is designed to provide strict instructions where possible and clear guidelines where input from the modeling team is required. While the application here is to wheat phenology models, the same protocol could undoubtedly be used more generally for phenology models of other crops.

Comparison with usual calibration practices shows that, on average over modeling groups, the protocol leads to a better fit to the calibration data and to a better fit to out-of-sample data. Use of the protocol would be advantageous not just for individual modeling studies, but also for studies based on ensembles of models, including projections of climate impact. In particular, we have shown that if all modeling groups use the protocol, between-model variability can be substantially reduced, thus reducing the uncertainty of projections.

Models of crop phenology are in general relatively simple, compared to models that simulate not only phenology, but also crop growth and soil processes. It seems likely that having a standardized protocol for calibration would be even more important for these more complex models than for simpler models. Defining a protocol for more complex models would introduce new problems compared to the study here, in particular because of the number of output variables. However, the protocol here would be a useful starting point for treating more complex situations.

Declarations

Funding

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): 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-programme and AgroEcoSystem department. The order in which the donors are listed is arbitrary.

Conflicts of interests / Competing interests

The authors have no financial or proprietary interests in any material discussed in this article.

Code availability / Availability of data and material

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

Ethics approval/declarations

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Authors' contributions

Daniel Wallach: Conceptualization, methodology, project administration, writing - original draft, validation. Taru Palosuo, Henrike Mielenz, Peter Thorburn, Samuel Buis, Sabine Seidel: Conceptualization, methodology, project administration, writing , review &editing. All other authors: Simulations, model expertise, writing, review & editing.

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

Asseng S, Ewert F, Rosenzweig C, et al (2013) Uncertainty in simulating wheat yields under climate change. Nat Clim Chang 3:827–832. doi: 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. doi: 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 20:2301–20. doi: 10.1111/gcb.12520

Boogaard HL, Van Diepen CA, Rötter RP, et al (1998) User's guide for the WOFOST 7.1 crop growth simulation model and WOFOST control center 1.5. Technical Document 5. Wageningen, The Netherlands

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

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

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

Chakrabarti A, Ghosh JK (2011) AIC, BIC and Recent Advances in Model Selection. Philos Stat 583–605. doi: 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. doi: 10.1016/J.AGSY.2004.03.003

Cleland EE, Chuine I, Menzel A, et al (2007) Shifting plant phenology in response to global change. Trends Ecol Evol 22:357–365. doi: 10.1016/j.tree.2007.04.003

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. doi: 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. doi: http://dx.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. doi: 10.1016/J.GEODERMA.2017.11.025

Fath B, Jorgensen SE (2011) Fundamentals of ecological modelling: Applications in environmental management and research. 4th edition. Elsevier, Amsterdam

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. doi: 10.1016/J.AGRFORMET.2021.108686

Gate P (1995) Ecophysiologie du blé. Tec & Doc-Lavoisier, Paris

Godwin D, Ritchie J, Singh U, Hunt L A User’s Guide to CERES Wheat -V2. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiKrM7K0vr4AhURmRoKHYYhC5sQFnoECAUQAQ&url=https%3A%2F%2Fpdf.usaid.gov%2Fpdf_docs%2

FPNABU270.pdf&usg=AOvVaw2qvwr3lq1RUwuFyPClQ6aq

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. doi: 10.1016/J.ECOLMODEL.2008.02.007

Hernandez-Suarez JS, Nejadhashemi AP, Deb K (2021) A novel multi-objective model calibration method for ecohydrological applications. Environ Model Softw 144:105161. doi: 10.1016/J.ENVSOFT.2021.105161

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. doi: 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

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

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. doi: 10.1038/s43016-021-00400-y

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. doi: 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. doi: 10.1007/s10705-006-9044-8

Kersebaum KC, Ahuja LR, Ma L (2011) Special Features of the HERMES Model and Additional Procedures for Parameterization, Calibration, Validation, and Applications

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

Kobayashi K, Salam MU (2000) Comparing simulated and measured values using mean squared deviation and its components. Agron J 92:345–352

Kuha J (2004) AIC and BIC: Comparisons of Assumptions and Performance. Sociol Methods Res 33:188–229. doi: 10.1177/0049124103262065

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 21:1328–41. doi: 10.1111/gcb.12758

Lukacs PM, Burnham KP, Anderson DR (2010) Model selection bias and Freedman's paradox. Ann Inst Stat Math 62:117–125. doi: 10.1007/s10463-009-0234-4

Martre P, Wallach D, Asseng S, et al (2015) Multimodel ensembles of wheat growth: many models are better than one. Glob Chang Biol 21:911–25. doi: 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. doi: 10.3389/fsufs.2019.00108

Meier U (1997) Growth stages of mono- and dicotyledonous plants : BBCH-Monograph. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjBo-Ln4_r4AhVM3RoKHc-TA5sQFnoECAcQAQ&url=https%3A%2F%2Fwww.politicheagricole.it%2Fflex%2FAppData%2FWebLive%2FAgrometeo%2FMIEPFY800%2FBBCHengl2001.pdf&usg=AOvVaw1G_eUh_-qnVt91_qbj44d. Accessed 9 Jan 2019

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

Nelder JA, Mead R (1965) A Simplex Method for Function Minimization. Comput J 7:308–313. doi: 10.1093/comjnl/7.4.308

Nendel C, Berg M, Kersebaum KC, et al (2011) The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. Ecol Modell 222:1614–1625. doi: 10.1016/J.ECOLMODEL.2011.02.018

P.-E. Jansson P-E (2012) CoupModel: Model Use, Calibration, and Validation. Trans ASABE 55:1337–1346. doi: 10.13031/2013.42245

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

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. doi: 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. doi: 10.1038/s41598-018-23101-2

Seber GAF, Wild CJ (1989) Nonlinear regression. Wiley , New York

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. doi: 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. doi: 10.1016/J.JAG.2021.102670

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. doi: 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. doi: 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

Stockle CO, Donatelli M, Nelson R (2001) CropSyst, a cropping systems simulation model. Eur J Agron 18:289–307

Stuble KL, Bennion LD, Kuebbing SE (2021) Plant phenological responses to experimental warming – A synthesis. Glob Chang Biol gcb.15685. doi: 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. doi: 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. doi: 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. doi: 10.1016/J.ENVSOFT.2014.08.005

Wallach D (2011) Crop model calibration: A statistical perspective. Agron J 103:1144–1151

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

Wallach D, Palosuo T, Thorburn P, et al (2021a) How well do crop modeling groups predict wheat phenology, given calibration data from the target population? Eur J Agron 124:126195. doi: https://doi.org/10.1016/j.eja.2020.126195

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

Wallach D, Palosuo T, Thorburn P, et al (2021c) The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environ Model Softw 145:105206. doi: 10.1016/J.ENVSOFT.2021.105206

Wallach D, Palosuo T, Thorburn P, et al (2021d) Multi model evaluation of phenology prediction for wheat in Australia. Agric For Meteorol in press: doi: 10.1101/2020.06.06.133504

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

Wang PC, Shoup TE (2011) Parameter sensitivity study of the Nelder–Mead Simplex Method. Adv Eng Softw 42:529–533. doi: 10.1016/J.ADVENGSOFT.2011.04.004

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

White H (1981) Consequences and detection of misspecified nonlinear regression models. J Am Stat Assoc 374:419–433

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

Zhang 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. doi: 10.1016/J.AGRFORMET.2022.108856