A calibration protocol for soil-crop models aimed at reducing prediction error and inter-model variability

Daniel Wallach[1], Samuel Buis[2], Diana-Maria Seserman[3], Taru Palosuo[4], Peter Thorburn[5], Henrike Mielenz[6], Eric Justes[7], Kurt-Christian Kersebaum[3,8], Benjamin Dumont[9] Marie Launay[10], Sabine Julia Seidel[*11]

[1] Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany. dwallach@uni-bonn.de. ORCID 0000-0003-3500-8179
[2]INRAE, UMR 1114 EMMAH, Avignon, France. samuel.buis@inrae.fr
[3]Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany. diana.seserman@gmail.com, ckersebaum@zalf.de
[4]Natural Resources Institute Finland (Luke), Helsinki, Finland. taru.palosuo@luke.fi
[5]CSIRO Agriculture and Food, Brisbane, Queensland, Australia. Peter.Thorburn@csiro.au
[6]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Crop and Soil Science, Braunschweig, Germany. henrike.mielenz@julius-kuehn.de
[7]CIRAD, Persyst Department, Montpellier, France. eric.justes@cirad.fr
[8]Global Change Research Institute CAS, Brno, Czech Republic. ckersebaum@zalf.de
[9]Plant Sciences & TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium. benjamin.dumont@ulg.ac.be
[10]INRAE, US 1116 AgroClim, Avignon, France. marie.launay@inrae.fr
[11]Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany. sabine.seidel@uni-bonn.de

Abstract

Process-based soil-crop models are widely used in agronomic research. They are the tools of choice for evaluating climate change impact on crop production and for testing adaptation and mitigation strategies. Multi-model simulation studies show a wide diversity of results among models given the same inputs, implying that simulation results are very uncertain. A large part of the variability results from uncertainty in parameter values due to different approaches to model calibration. This study proposes an innovative calibration protocol that should improve model predictions compared to common practices and that should reduce variability between modeling groups. The two major innovations concern the treatment of multiple output variables and the choice of parameters to estimate, both of which are here based on standard statistical procedure adapted to the specificities of soil-crop models. The protocol performed well in an artificial-data test with relatively many measured variables and estimated parameters. The protocol is formulated so as to be applicable to a wide range of models and data sets. If widely adopted, it could substantially reduce model error and inter-model variability.

Introduction

Process-based models that describe crop growth and development, soil water and nitrogen dynamics and their interactions (henceforward "soil-crop" models) are an essential research tool for agronomy. They are the tool of choice for evaluating climate change impact on crop production and for testing adaptation and mitigation strategies (Asseng et al., 2019; Ramirez-Villegas et al., 2017; Webber et al., 2014). They are also used as aids in yield forecasting (van der Velde and Nisini, 2019), crop breeding programs (Ramirez Villegas et al., 2020) and for informing crop management decisions (Keating et al., 2003).

A fairly recent practice, largely driven by the Agricultural Modeling Intercomparion and Improvement Project (AgMIP, (Rosenzweig et al., 2013)) is to organize multi-model ensemble studies, where multiple modeling groups use the same inputs and simulate the same output variables. It has been observed that these studies systematically exhibit a large amount of variability between modeling groups (for example (Bruni et al., 2022; Webber et al., 2017) , though this can be mitigated by using the multi-model mean or median (Martre et al., 2015; Wallach et al., 2018). In studies of the impact of global climate change on crop production using multiple soil-crop and multiple climate models, the contribution to total variability of variability among soil-crop models has been found to be even greater than the contribution due to variability in climate projections (Asseng et al., 2013; Li et al., 2015; Wang et al., 2020).

For given inputs, the variability in soil-crop model simulations arises from variability in the model equations ("model structure") and in the values of the parameters. Comparing those two sources of uncertainty, it has generally been found that both are important. Uncertainty in model structure is often found to make the larger contribution to overall variability (Tao et al., 2018; Xiong et al., 2020; Zhang et al., 2017). However, a recent study pointed out that those previous studies assume that there is a fixed set of parameters to estimate, while in fact there is also considerable uncertainty in the choice of parameters to estimate. Taking into account the uncertainty in choice of parameters, it was found that parameter uncertainty in most cases contributed more, often much more, than model structure uncertainty to overall variability in simulations (Wallach et al., 2023a).

The necessity of improving model structure in order to reduce soil-crop model prediction error and inter-model variability has been recognized (Maiorano et al., 2017). Improving parameterization on the other hand, and in particular improving methodology of crop model calibration, has only recently been seen as a major pathway to improve predictions and reduce variability of soil-crop model simulations. In a series of multi-model studies, the AgMIP calibration group (https://agmip.org/crop-model-calibration-3/), considered the simplified situation where only phenology data were used for calibration and found that there was very substantial variability in calibration practices between modeling groups, even between modeling groups using the same model structure (Wallach et al., 2021a, 2021b, 2021c). The "role of users" on soil-crop model simulations has also been found elsewhere (Albanito et al., 2022; Confalonieri et al., 2016). The AgMIP group proposed a protocol for calibration based on phenology data, aimed at improving and homogenizing practices, and found that the new protocol substantially

reduced variability and prediction error compared to the case where each modeling team implemented its usual calibration approach (Wallach et al., 2023b). This shows that it is possible to achieve the twin goals of reduced error and reduced variability of soil-crop model simulations through improved calibration practices.

The present study reports the next stage of the AgMIP calibration activity, which considers the general soil-crop model calibration problem, where one uses all available data and not just phenology data. The two objectives of this study are firstly to propose improved calibration practices for soil-crop models in order to reduce prediction error and secondly to formulate the recommendations in a protocol in such a way that they are applicable to essentially all soil-crop models and data sets, in order to reduce inter-modeling group variability.

Improved calibration practices are necessary in particular with respect to two problems. The first is how to take into account multiple variables. Data available for calibration may include multiple variables, such as days to several development stages, biomass, light interception and soil water at various dates, end of season measurements of yield, grain number and grain protein content and others. Furthermore, the variety of available data is increasing with new sensors, additional remote sensing possibilities and improved data transmission capabilities (Pasquel et al., 2022). It is important to take all observed variables into account, even for example if the focus of the study is only on yield, in order to simulate as realistically as possible the dynamics of the soil-plant system, since this should improve predictions for new environments (Angulo et al., 2013a; Pasley et al., 2023). More realistic simulations of all processes is also important if the model is used as a tool for understanding the contributions of different processes to an overall result

The major difficulty here is that as more variables are considered, the number of parameters to estimate increases, leading to numerical problems. The most common solution to this problem is to split the problem into parts by fitting the model to only one or only a few variables at a time, to avoid estimating a large number of parameters at the same time (Angulo et al., 2013b; Jha et al., 2021; Pasley et al., 2023). A variant of this approach is to also separately fit environments with and without water and nitrogen stresses (Guillaume et al., 2011; Kersebaum et al., 2011). It has been emphasized that it is important to choose the order so that the "most independent" variables are treated first (Pasley et al., 2023). The difficulty with this solution is that crop models describe an interacting system, and in general it is not possible to order the processes in such a way that earlier processes affect later processes but later processes have no effect on earlier processes; i.e. there are feedbacks in the system. As a result, when parameters of later processes are fit to data, this may degrade the fit to variables used in earlier calibration steps (Guillaume et al., 2011). There have not been any proposed procedures that allow one to fit all data simultaneously while nonetheless simplifying the numerical problem of finding the best parameter values.

The recommendations here propose doing the calibration calculations in two steps. In the first step, variables are treated individually. In the second step, the model is fit to all observed variables simultaneously, using weights based on the fit in the first phase. This is similar to the standard statistical approach of first doing ordinary least squares (OLS) regression, and then

using the OLS results to obtain weights for doing weighted least squares (WLS) (Seber and Wild, 1989). This approach takes advantage of a particularity of soil-crop models, which is that while there are feedbacks, these are often quite limited. As a result, the first step is expected to give good staring points for searching for the best parameter values in the second step, thus greatly simplifying the numerical problem.

The second problem is how to choose the parameters to be estimated from the data. In general, crop models have many more parameters than can be estimated from available data, so it is necessary to decide which parameters to estimate. One approach identifies a priori (independently of the available data) the major parameters that should be estimated for a particular model structure (Ahuja and Ma, 2011). Other studies have focused on ways of doing sensitivity analysis for crop models, in order to identify the most important parameters that should be estimated (Ceglar et al., 2011; Li and Ren, 2019). In some cases, selection involves an ad hoc combination of a priori choice, sensitivity analysis and test of various parameters to see how much they can improve the fit to the data. None of these approaches specifically addresses the problem of over-parameterization, or the risk of highly correlated and therefore highly uncertain parameter estimators. The AgMIP study on calibration using only phenology data proposed an original approach to choice of parameters to estimate, based on a standard model selection criterion (Wallach et al., 2023b). An analogous approach is used here, but extended to multiple measured variables. An alternative would be a Bayesian approach to parameter estimation, but this is rarely applied to soil-crop models (though see Dumont 2014 for a Bayesian approach to calibration of the STICS soil-crop model).

Previous calibration studies have very largely concerned calibration of a particular model (for example Ahuja and Ma, 2011). In order to reduce inter-model variability, which is the second objective of this study, it is necessary that the proposed calibration protocol be applicable to essentially any soil-crop model and to any set of measured data. To achieve this genericity, the protocol provides recommendations for procedures, and explains how to combine them with model expertise in order to apply those recommendations to each specific model. While much of the variability in simulations arises from the two problems discussed above, modeling teams may also differ as to other aspects of calibration (Wallach et al., 2021c). To further reduce variability therefore, the proposed protocol covers all the steps involved in soil-crop model calibration,

The proposed protocol was tested using the STICS soil-crop model (Beaudoin et al., 2023) with artificial data for winter wheat. The “measured” data included days to three development stages, biomass at several dates, nitrogen content of final biomass, grain yield and grain number. Altogether 23 parameters were considered. Both the number of different variables and the number of parameters considered are large compared to most studies. The protocol performed well including for out-of-sample predictions.

Results

Description of the calibration protocol

The proposed protocol for soil-crop model calibration, described in detail below, is composed of eight steps. The first five steps (the model expertise part of the protocol) require detailed knowledge of the model and the data. No calculations are performed here. The result of

these steps is a series of tables that contain all the model-specific information needed for the calculations. The last three steps (the calculation steps) describe the calculations to be done. The protocol includes instructions for each step, and the documentation to be produced in each step. The documentation is an integral part of the protocol, insuring transparency and reproducibility of the calibration procedure.

Model expertise steps

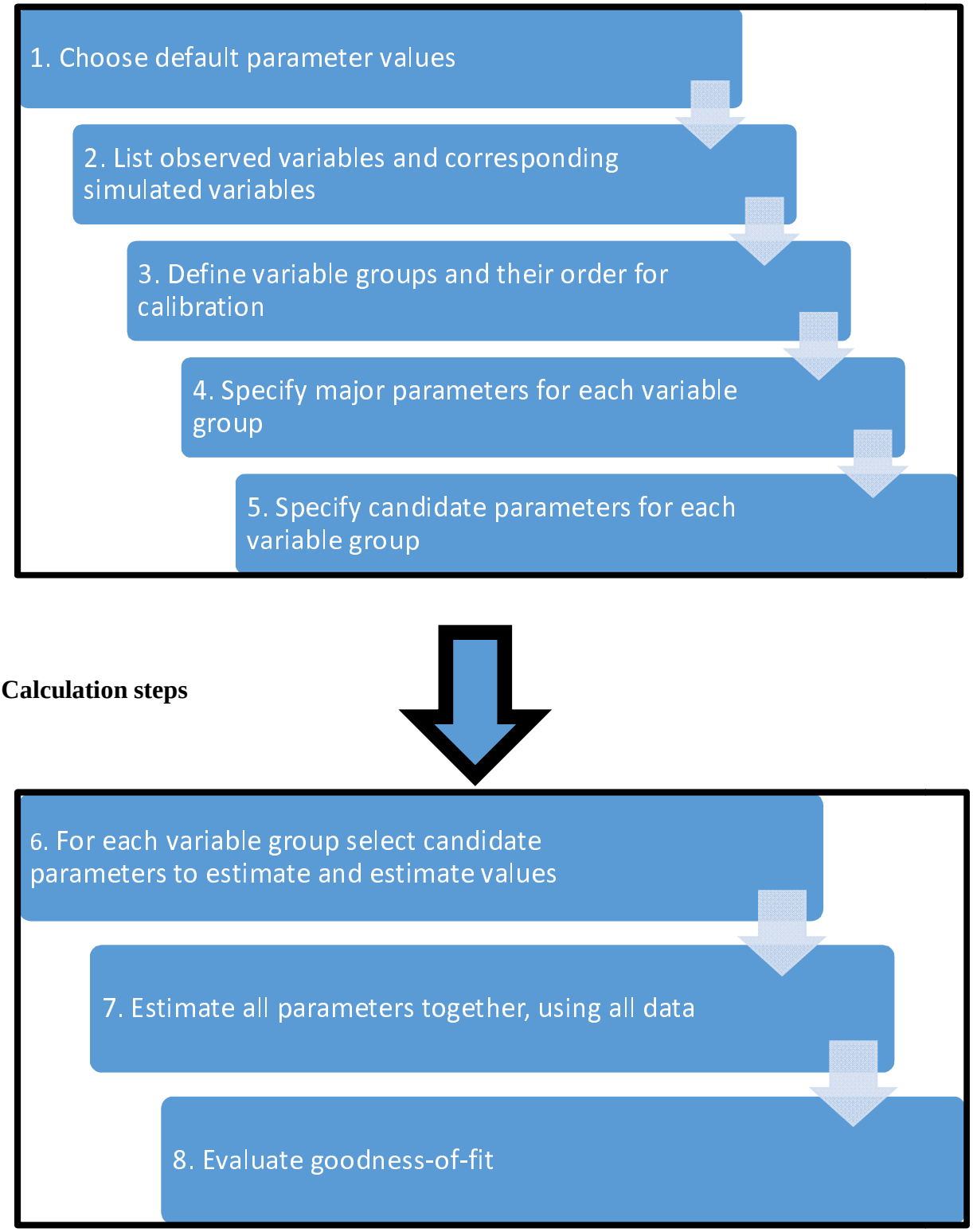

Figure 1

Schema of calibration protocol. The first five steps involve codifying model expertise. Given that information, the calculation steps 6 -8 require no farther model- specific inputs.

Step 1. Explain choice of default parameter values and describe the calibration environments.

Only a small fraction of crop model parameters will usually be estimated from the data. The majority of the parameters will remain at their default values, so it is important to choose the default values with care. In particular, one should obtain as much information as possible about the cultivar characteristics (maturity class, photoperiod sensitivity, etc.) and choose default parameter values accordingly. The documentation required here (not shown) contains the cultivar characteristics and the rationale for the choice of default parameter values.

Step 2. List observed variables and corresponding simulated variables, if any.

The purpose of this step is to identify the correspondence between observed and simulated variables. The documentation required for this step is a table with one row for each measured variable, showing also the corresponding simulated variable (Table 1 shows an example).

Step 3. Define groups of variables and order them.

The grouping of observed variables is fixed by the protocol All days to development stages are grouped together in a phenology group. All measurements of a given variable at different times (e.g. biomass) will also be in the same group. Other variables (including all final values such as final yield, grain number, grain protein content etc.) will each constitute a separate group. The order of the groups is the order in which they will be used for fitting the model. The order of the groups is important. It should be such that the simulated values of the variables in a group have little or no effect on the simulated values of variables in previous groups. Phenology will usually be the first group, since the days to different development stages are usually not affected or only slightly affected by other variables. The required documentation here, which is combined with the documentation for step 2, shows the group and order for each observed variable (see Table 1).

Step 4. Identify the major parameter or parameters for each group of variables

The purpose of this step is to identify the major parameters that affect each variable group. There is a strict limit on the number of major parameters for each group, to avoid over-parameterization. If there is only one variable in the group, there can only be one major parameter. For variables with at least two measurements in some environments (e.g. biomass with in-season measurements) there can be at most two major parameters (for example, one that determines rate of increase during vegetative growth and a second that determines rate of increase during reproductive growth).. For phenology, there can be as many major parameters as observed development stages with simulated equivalents. However, each major parameter must affect the time to a different stage.

The major parameters for a group of variables should have an effect on the simulated values in all environments. If a parameter is nearly additive, i.e. has nearly the same effect in all environments, then the estimation of the parameter will make the model bias nearly zero for the associated variable, which is desirable. Thus, the first choice of major parameter for a variable is a parameter that is nearly additive. Thermal degree days to a development stage is usually a

nearly additive parameter for days to that stage, since increasing the required number of degree days will, in general, increase the days to the stage by a similar amount for all environments. Parameters that describe the effect of stresses, which only affect the simulated values if the stresses are present, will not be major parameters. The required documentation here is a table which shows the major parameters for each variable group (see example in Table 2).

Step 5. Identify candidate parameters for each group of variables.

The candidate parameters are those parameters that are likely to explain a substantial part of the variability between environments and/or management strategies that remains after the major parameters are estimated. Each of these parameters will be tested (in the next step), and will only be included in the final list of parameters to estimate if estimation leads to a sufficient improvement in fit to the data.

The candidate parameters should be ordered from supposedly most to supposedly least important. The number of candidate parameters is not limited, but it is recommended to keep the number fairly small. The required documentation here is a table with the candidate parameters for each variable group (see example in Table 3).

Step 6. Selection of parameters to estimate for each variable group and first estimation of their values

In this step, each group of variables is treated separately, in the order chosen in step 3. A list of parameters to estimate for each group is initialized with the major parameters. The major parameters for the group are estimated using ordinary least squares (OLS), and the corrected Akaike Information Criterion (AICc Brewer et al., 2016; Chakrabarti and Ghosh, 2011) is calculated as

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

where SS is the sum of squared errors for all variables in the group, n is the number of data points and p the number of estimated parameters. This assumes that all model errors for the group are independent and identically normally distributed.

Once the major parameters have been estimated, each candidate parameter in turn is added tentatively to the list of parameters to be estimated. If estimating all the parameters on the list reduces AICc below the previous smallest value, the candidate is kept on the list of parameters to be estimated. Otherwise, the candidate is removed from the list of parameters to be estimated, and returns to its default value (see flow diagram in Figure 2). AICc is a standard model selection criterion, designed to choose the best predicting model even if none of the proposed models is the true model (Aho et al., 2014).

Biomass should be replaced by the natural logarithm of biomass before the calculation. The reason is that biomass values may go over a wide range of values during the growth period, with an associated increase in the standard deviation of model error. The log transformation will make the standard deviations approximately constant for all dates. Similarly, a log transformation should be used for any other variables expected to vary over a wide range over time.

In the example here, the Nelder-Mead simplex algorithm was used to find optimal parameter values, both in step 6 and step 7 (Nelder and Mead, 1965). However, other optimization algorithms could also be used. When testing candidate parameters, one of the starting points for optimization should be the previous best parameter values, since those will often be close to the new best values.

A first required documentation table here shows the result of adding each new candidate parameter, for each variable group (see example in **Error! Reference source not found.**). The optimum parameter values and the fit to the measured data after this step are combined with the documentation of step 7 (Table 5, Table 6).

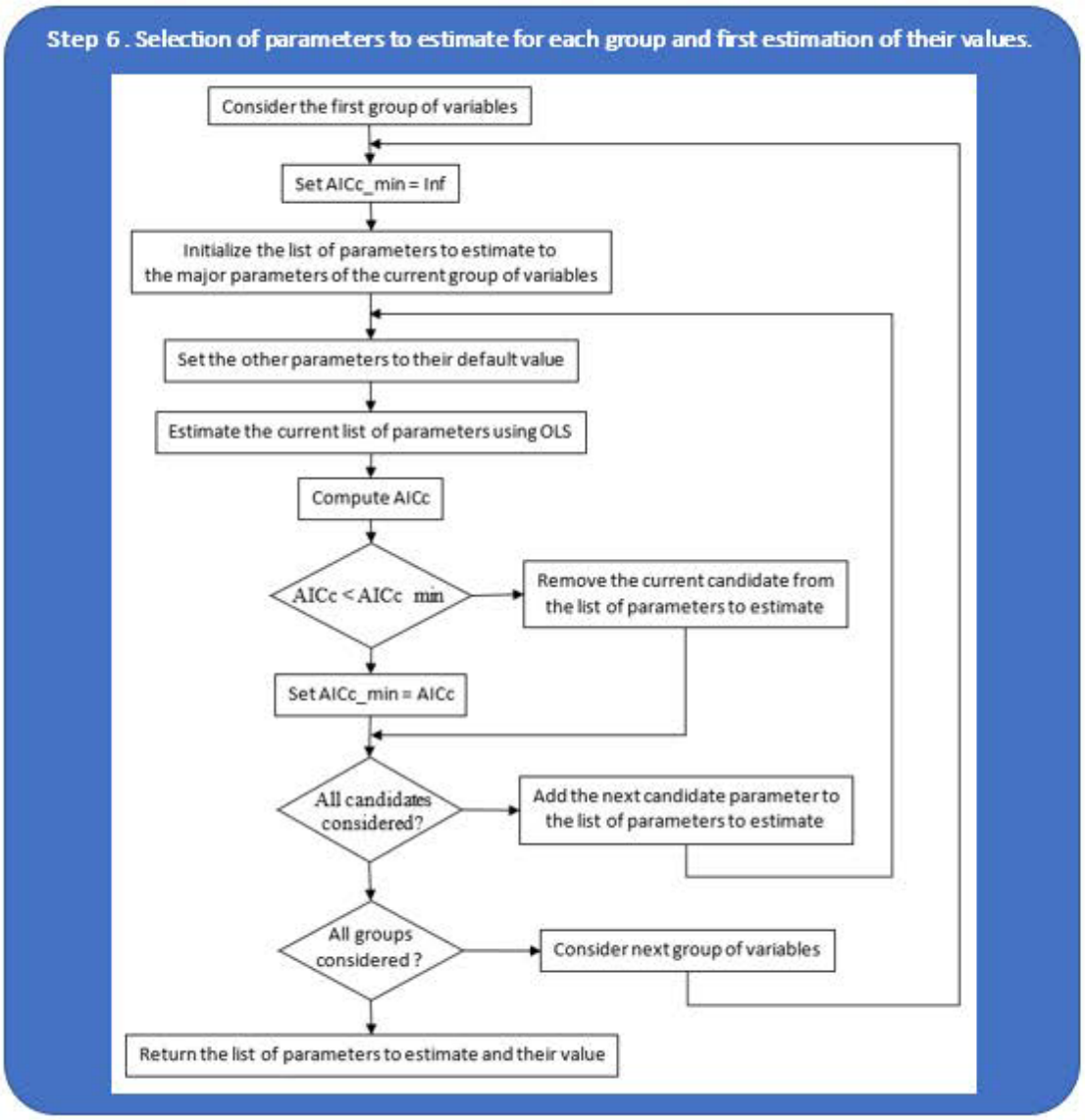

Figure 2

Flow diagram for the selection of parameters to estimate, and first estimation of their values, for one variable group, for step 6 of the protocol

Step 7. Re- estimation of all selected parameters using all variables simultaneously

In this step, all the selected parameters from step 6 are estimated together, using all the data, using weighted least squares (WLS). Each group of variables is weighted by the inverse of the standard deviation of model error (s), estimated from the results of step 6 as

$$s = SS / (n - p) \tag{2}$$

where SS is the sum of squared errors for all variables in the group, n is the number of data points and p the number of estimated parameters. The required documentation table here shows the estimated parameter values after steps 6 and 7 (see example in Table 5).

Step 8. Evaluation of goodness-of-fit

In this step metrics of goodness of fit are calculated for the simulations using the default parameter values, using the parameter values after step 6 and using the parameter values after step 7. The required documentation table here shows the metrics for goodness-of-fit at each stage. An example is shown in Table 6 and Supplementary Table S1.. Additional metrics could also be calculated. Graphs of simulated versus observed values for each variable should also be produced (see example in Supplementary Figures S1-S3).

Evaluation of the protocol

The STICS soil-crop model was used both to generate the artificial data and as the model calibrated using the protocol. The use of artificial data makes it possible to add measurement error as desired, rather than working with uncontrolled measurement error. It also makes it possible to compare estimated parameter values with the true values.

The six different variable groups in the artificial data are shown in Table 2. Data from fourteen environments were used for calibration, and data from 8 different environments were used to evaluate out-of-sample prediction error. For testing the protocol, 23 parameters of the STICS model were set to default values different than those used to generate the artificial data.

For all variable groups, the calibration substantially improved the fit to the calibration data (Table 6, Supplementary Table S1 and Figs. S1-S3). The improvement was most pronounced for the phenology group, where relative root mean squared error (RRMSE) decreased from 16%-18%, depending on the simulated development stage, using the default parameters to less than 1% for all stages using the parameters estimated in step 7. RRMSE for yield decreased from 25% with default parameter values to 16% after calibration., which was the largest final RRMSE value.

Calibration also substantially improved the fit to the out-of-sample data (Table 7, Supplementary Table S2). For the phenology group, RRMSE decreased from 14%-16% using the default parameters to less than 1% using the parameters estimated in step 7. RRMSE for yield

decreased from 17% with default parameter values to 9% after calibration. The similar RRMSE values for the calibration and out-of-sample data show that there is not a problem of over-parameterization. For both calibration and out-of-sample data, RRMSE was smaller in almost every case after step 7 than after step 6, but the differences were in all cases small, at most a difference of 1.5% in RRMSE value. It seems that in this example feedbacks, which are ignored in step 6 where each variable group is fit separately, but which are taken into account in step 7; are relatively slight.

All 23 parameters that had default values different than the values used to generate the artificial data were chosen as either major parameters or candidate parameters. However, only 13 of those parameters were finally estimated. The remainder were candidate parameters that did not reduce AICc, and so remained at their default values. Estimated values of the major parameters for the phenology group were much closer to the true values than were the default parameter values (Table 5). For other parameters the final estimated values could be closer or farther from the" true" values than were the default values. This suggests that there is relatively little compensation of errors for the major phenology parameters, because the major parameters have a preponderant effect on phenology, while the results for other variables depend on multiple parameters and so compensation of errors plays a larger role.

Discussion

The two objectives of the protocol are first to define calibration practices that are effective in reducing prediction error and secondly to reduce inter-model variability by formulating the protocol in such a way that it can be implemented by essentially all soil-crop models and for all data sets. The results here are very encouraging with respect to both objectives. The protocol was effective in substantially reducing errors both for the calibration environments and for out-of-sample environments compared to initial error, for all variables, despite the relatively large diversity of measured variables and the relatively large number of parameters that were considered. The protocol was in no way specifically adapted to the STICS model or to the artificial data set used here, so its successful application here suggests that it can be used as a standardized procedure for a large range of soil-crop models and data sets.

It would be of interest to test alternative calibration procedures, using this protocol as a baseline. One alternative is to directly use all the data simultaneously for calibration, rather than first using one variable group at a time, despite the large number of parameters to estimate that that implies. One possibility here would be to use the PEST calibration package (Doherty et al., 2010), which uses regularization techniques to make the parameter estimation feasible even for highly correlated parameter estimators and for ill-conditioned models. There is still however the problem of choosing a limited number of parameters to consider. This might be done using sensitivity analysis (Necpálová et al., 2015).

Another alternative would be to do a Bayesian analysis, where one estimates the distribution of the parameters rather than the best value. In principle, the choice of which parameters to estimate is less crucial here than with a frequentist approach, since neither uninfluential nor highly correlated parameters create particular difficulties. However, even here it is not possible to consider all parameters, so some selection of parameters would still be

necessary. Bayesian methods have so far been applied principally to calibration problems with relatively few variable types and parameters (Dumont et al., 2014; Iizumi et al., 2009; López-Cruz et al., 2017).

An important feature of the protocol here is the separation into two parts. The first part, based on model expertise, is where all the specificities of a particular model structure and data set are taken into account. The calculations in the second part can be completely automated, once the documentation from the first part is available. This could greatly simplify the calibration activity for soil-crop models. In the example here, all the calculations (steps 6-8) were done automatically, with the tables from steps 2-5 as inputs. The next step in the AgMIP calibration project is a multi-model application of the protocol to real data. This is currently underway.

Methods

Structure of the artificial data set

In order to make the artificial data as realistic as possible, the data set is based on a real data set for winter wheat variety Apache grown in France, from variety trials carried out by Arvalis – Institut du vegetal Paris. The same environments (weather, soil characteristics, management) and the same measured variables as in that data set are used for the artificial data. The only difference is that the measured values are replaced by values simulated using the STICS model. The simulated variables are those shown in Table 1.

The full data set has data from 22 environments, which were divided into two groups. The data from fourteen environments (six different sites, five different years) were used for calibration (the “calibration” data). The data from the eight other environments (five different sites, two different years) were used for testing (referred to as the “evaluation” or “out of sample” data). None of the sites or years present in the calibration data were also present in the evaluation data. Thus, the simulation errors for the evaluation data measure how well the calibrated model simulates for environments different than those used for calibration, but drawn from the same population (conventionally managed wheat fields in the major wheat growing regions of France, under current climate, sown with variety Apache), for the case where the data-generating mechanism is the same as for the calibration data. Further details about the environments can be found in (Wallach et al., 2021a).

Generation of artificial data

The STICS soil-crop model, using parameter values previously estimated for the French winter wheat variety Soissons, was used to generate the artificial data for both the calibration and the evaluation environments. Those are henceforward referred to as the “true” parameter values. Then random noise was added to each generated value for the calibration environments. The amount of noise was chosen independently for each measurement by drawing from a Gaussian distribution with mean 0 and standard deviation equal to 2 days for the phenology data and equal to 10% of the generated value for the other variables, truncated at ± 3 standard deviations. Noise was not added to the evaluation data, since we are interested in prediction of the true values.

Default parameter values for the calibration

The parameter values used as starting values for the calibration (the “default” parameter values) of all the parameters shown in Tables 1 and 2 are different than their true values (Supplementary Table S3). The default values were chosen as the true values plus 60% of the distance to the assumed upper limit of the parameter or minus 60% of the distance to the assumed lower limit. The choice of whether to move the starting value toward the upper or lower limit was made at random independently for each parameter.

Calculations

For the test of the protocol, the R packages CroptimizR (Vezy et al., 2023) and CroPlotR (Vezy et al., 2021) were used. All the calculations for steps 6 -8 were done automatically, using the tables prepared in steps 1-5 as input. A wrapper function for STICS available in the R package SticsOnR (Lecharpentier et al., 2023) was used to handle the communication between CroptimizR and the crop model (sending parameter values to the model and recovering simulated values). CroptimizR could be used with any crop model, but a different model wrapper would be required in each case.

The algorithm used in the example for searching the parameter space was the Nelder-Mead simplex algorithm (Nelder and Mead, 1965), as implemented in the R package nloptr R (Ypma and Johnson, 2022). This is a robust algorithm which is well-adapted to crop models since it does not use derivatives and does not require that the model be a continuous function of the parameters (Wang and Shoup, 2011). However, performance may become less efficient in high-dimension (Han and Neumann, 2006),The final result can depend on the starting values and the algorithm is not guaranteed to converge to a global minimum, so the implementation here used multiple starting points for the algorithm. For variable groups with more than one major parameter, the algorithm used 20 starting points within the upper and lower bounds of each major parameter, chosen by Latin Hypercube sampling. For variable groups with a single major parameter, five different starting points were used. For each candidate parameter, all previously chosen parameters had initial values equal to the optimal values previously found, and five different starting values were used for the new candidate. For step 7, one starting point was the best parameter values found when treating each variable group separately. In addition, 19 other starting points were generated at random using Latin Hypercube Sampling. In all cases, using the previous best values as starting point led to the lowest sum of squared errors.

Evaluation of the protocol

The following evaluation metrics were calculated

$$
\begin{aligned}
&SS = \sum (y_i - \hat{y}_i)^2 \\
&MSE = (1/n)SS \\
&RMSE = \sqrt{MSE} \\
&RRMSE = RMSE / \bar{y} \\
&bias = \sum (y_i - \hat{y}_i) \\
&NSE = 1 - SS / SS_{\bar{y}}
\end{aligned}
\tag{3}
$$

where MSE = mean squared error, RMSE = root mean squared error, RRMSE = relative root mean squared error, bias = model bias and NSE = Nash Sutcliffe efficiency. The sum in SS is over all measurements of the variable in question, y_i is the ith measurement and $\hat{y}_i$ is the corresponding simulated value. In RRMSE, $\bar{y}$ is the average of the measured values, and in NSE, $SS_{\bar{y}}$ is the sum of squared errors for the model that uses $\bar{y}$ to predict for all environments. The above criteria are calculated with respect to the measured values, which includes measurement error, for the calibration data, and with respect to the true values, without measurement error, for the evaluation data.

References

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

Ahuja, L.R., Ma, L. (Eds.), 2011. Methods of Introducing System Models into Agricultural Research, Advances in Agricultural Systems Modeling. American Society of Agronomy and Soil Science Society of America, Madison, WI, USA. https://doi.org/10.2134/advagricsystmodel2

Albanito, F., McBey, D., Harrison, M., Smith, P., Ehrhardt, F., Bhatia, A., Bellocchi, G., Brilli, L., Carozzi, M., Christie, K., Doltra, J., Dorich, C., Doro, L., Grace, P., Grant, B., Léonard, J., Liebig, M., Ludemann, C., Martin, R., Meier, E., Meyer, R., De Antoni Migliorati, M., Myrgiotis, V., Recous, S., Sándor, R., Snow, V., Soussana, J.-F., Smith, W.N., Fitton, N., 2022. How Modelers Model: the Overlooked Social and Human Dimensions in Model Intercomparison Studies. Environ. Sci. Technol. 56, 13485–13498. https://doi.org/10.1021/acs.est.2c02023

Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S., Ewert, F., 2013a. Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe. Agric. For. Meteorol. 170, 32–46. https://doi.org/10.1016/j.agrformet.2012.11.017

Angulo, C., Rötter, R., Trnka, M., Pirttioja, N., Gaiser, T., Hlavinka, P., Ewert, F., 2013b. Characteristic 'fingerprints' of crop model responses to weather input data at different spatial resolutions. Eur. J. Agron. 49, 104–114. https://doi.org/10.1016/J.EJA.2013.04.003

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P.J., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J.W., Williams, J.R., Wolf, J., 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., Rötter, R.P., O'Leary, G.J., Fitzgerald, G.J., Girousse, C., Motzo, R., Giunta, F., Babar, M.A., Reynolds, M.P., Kheir, A.M.S., Thorburn, P.J., Waha, K., Ruane, A.C., Aggarwal, P.K., Ahmed, M., Balkovič, J., Basso, B., Biernath, C., Bindi, M., Cammarano, D., Challinor, A.J., De Sanctis, G., Dumont, B., Eyshi Rezaei, E., Fereres, E., Ferrise, R., Garcia□Vila, M., Gayler, S., Gao, Y., Horan, H., Hoogenboom, G., Izaurralde, R.C., Jabloun, M., Jones, C.D., Kassie, B.T., Kersebaum, K.-C., Klein, C., Koehler, A., Liu, B., Minoli, S., Montesino San Martin, M., Müller, C., Naresh Kumar, S., Nendel, C., Olesen, J.E., Palosuo, T., Porter, J.R., Priesack, E., Ripoche, D., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Van der Velde, M., Wallach, D., Wang, E., Webber, H., Wolf, J., Xiao, L., Zhang, Z., Zhao, Z., Zhu, Y., Ewert, F., 2019. Climate change impact and adaptation for wheat protein. Glob. Chang. Biol. 25, 155–173. https://doi.org/10.1111/gcb.14481

Beaudoin, N., Lecharpentier, P., Ripoche-Wachter, D., Strullu, L., Mary, B., J., L., Launay, M., Justes, E., 2023. STICS Soil crop model. Conceptual Framework, Equations and Uses. Editions Quae. https://doi.org/https://doi.org/10.35690/978-2-7592-3679-4

Brewer, M.J., Butler, A., Cooksley, S.L., 2016. The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. Methods Ecol. Evol. 7, 679–692. https://doi.org/10.1111/2041-210X.12541

Bruni, E., Chenu, C., Abramoff, R.Z., Baldoni, G., Barkusky, D., Clivot, H., Huang, Y., Kätterer, T., Pikuła, D., Spiegel, H., Virto, I., Guenet, B., 2022. Multi□modelling predictions show high uncertainty of required carbon input changes to reach a 4‰ target. Eur. J. Soil Sci. 73, e13330. https://doi.org/10.1111/ejss.13330

Ceglar, A., Črepinšek, Z., Kajfež-Bogataj, L., Pogačar, T., 2011. The simulation of phenological development in dynamic crop model: The Bayesian comparison of different methods. Agric. For. Meteorol. 151, 101–115. https://doi.org/10.1016/j.agrformet.2010.09.007

Chakrabarti, A., Ghosh, J.K., 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

Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., Pagani, V., Cappelli, G., Vertemara, A., Alberti, L., Alberti, P., Atanassiu, S., Bonaiti, M., Cappelletti, G., Ceruti, M., Confalonieri, A., Corgatelli, G., Corti, P., Dell'Oro, M., Ghidoni, A.,

Lamarta, A., Maghini, A., Mambretti, M., Manchia, A., Massoni, G., Mutti, P., Pariani, S., Pasini, D., Pesenti, A., Pizzamiglio, G., Ravasio, A., Rea, A., Santorsola, D., Serafini, G., Slavazza, M., Acutis, M., 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

Doherty, J.E., Hunt, R.J., Tonkin, M.J., 2010. Approaches to highly parameterized inversion: A guide to using PEST for model-parameter and predictive-uncertainty analysis: U.S. Geological Survey Scientific Investigations Report 2010–5211.

Dumont, B., Leemans, V., Mansouri, M., Bodson, B., Destain, J.-P., Destain, M.-F., 2014. Parameter identification of the STICS crop model, using an accelerated formal MCMC approach. Environ. Model. Softw. 52, 121–135. https://doi.org/http://dx.doi.org/10.1016/j.envsoft.2013.10.022

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

Han, L., Neumann, M., 2006. Effect of dimensionality on the Nelder–Mead simplex method. Optim. Methods Softw. 21, 1–16. https://doi.org/10.1080/10556780512331318290

Iizumi, T., Yokozawa, M., Nishimori, M., 2009. Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach. Agric. For. Meteorol. 149, 333–348.

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

Keating, B.., Carberry, P.., Hammer, G.., Probert, M.., Robertson, M.., Holzworth, D., Huth, N.., Hargreaves, J.N.., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.., Silburn, M., Wang, E., Brown, S., Bristow, K.., Asseng, S., Chapman, S., McCown, R.., Freebairn, D.., Smith, C.., 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, K.C., Ahuja, L.R., Ma, L., 2011. Special Features of the HERMES Model and Additional Procedures for Parameterization, Calibration, Validation, and Applications. https://doi.org/10.2134/advagricsystmodel2.c2

Lecharpentier, P., Vezy, R., Buis, S., Ginet, M., 2023. STICSonR: Manage STICS Simulations Running the Executable or JavaStics [WWW Document]. URL https://doi.org/10.5281/zenodo.8142442 (2023)

Li, P., Ren, L., 2019. Evaluating the effects of limited irrigation on crop water productivity and reducing deep groundwater exploitation in the North China Plain using an agro-hydrological model: I. Parameter sensitivity analysis, calibration and model validation. J. Hydrol. 574, 497–516. https://doi.org/10.1016/J.JHYDROL.2019.04.053

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S.,

Confalonieri, R., Fumoto, T., Gaydon, D., Marcaida, M., Nakagawa, H., Oriol, P., Ruane, A.C., Ruget, F., Singh, B.-, Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z., Bouman, B., 2015. Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob. Chang. Biol. 21, 1328–41. https://doi.org/10.1111/gcb.12758

López-Cruz, I.L., Ruiz-García, A., Fitz-Rodríguez, E., Salazar-Moreno, R., Rojano-Aguilar, A., 2017. A comparison of Bayesian and classical methods for parameter estimation in greenhouse crop models. Acta Hortic. 241–248. https://doi.org/10.17660/ActaHortic.2017.1182.29

Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R.P., Ruane, A.C., Semenov, M.A., Wallach, D., Wang, E., Alderman, P.D., Kassie, B.T., Biernath, C., Basso, B., Cammarano, D., Challinor, A.J., Doltra, J., Dumont, B., Rezaei, E.E., Gayler, S., Kersebaum, K.C., Kimball, B.A., Koehler, A.-K., Liu, B., O'Leary, G.J., Olesen, J.E., Ottman, M.J., Priesack, E., Reynolds, M., Stratonovitch, P., Streck, T., Thorburn, P.J., Waha, K., Wall, G.W., White, J.W., Zhao, Z., Zhu, Y., 2017. Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. F. Crop. Res. 202. https://doi.org/10.1016/j.fcr.2016.05.001

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rötter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Glob. Chang. Biol. 21, 911–25. https://doi.org/10.1111/gcb.12768

Necpálová, M., Anex, R.P., Fienen, M.N., Del Grosso, S.J., Castellano, M.J., Sawyer, J.E., Iqbal, J., Pantoja, J.L., Barker, D.W., 2015. Understanding the DayCent model: Calibration, sensitivity, and identifiability through inverse modeling. Environ. Model. Softw. 66, 110–130. https://doi.org/10.1016/J.ENVSOFT.2014.12.011

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

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

Pasquel, D., Roux, S., Richetti, J., Cammarano, D., Tisseyre, B., Taylor, J.A., 2022. A review of methods to evaluate crop model performance at multiple and changing spatial scales. Precis. Agric. 23, 1489–1513. https://doi.org/10.1007/s11119-022-09885-4

Ramirez-Villegas, J., Koehler, A.-K., Challinor, A.J., 2017. Assessing uncertainty and complexity in regional-scale crop model simulations. Eur. J. Agron. 88, 84–95. https://doi.org/10.1016/J.EJA.2015.11.021

Ramirez□Villegas, J., Molero Milan, A., Alexandrov, N., Asseng, S., Challinor, A.J., Crossa, J., Eeuwijk, F., Ghanem, M.E., Grenier, C., Heinemann, A.B., Wang, J., Juliana, P., Kehel, Z., Kholova, J., Koo, J., Pequeno, D., Quiroz, R., Rebolledo, M.C., Sukumaran, S., Vadez, V., White, J.W., Reynolds, M., 2020. CGIAR modeling approaches for resource□constrained scenarios: I. Accelerating crop breeding for a changing climate. Crop Sci. 60, 547–567. https://doi.org/10.1002/csc2.20048

Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agric. For. Meteorol. 170. https://doi.org/10.1016/j.agrformet.2012.09.011

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

Tao, F., Rötter, R.P., Palosuo, T., Gregorio Hernández Díaz□Ambrona, C., Mínguez, M.I., Semenov, M.A., Kersebaum, K.C., Nendel, C., Specka, X., Hoffmann, H., Ewert, F., Dambreville, A., Martre, P., Rodríguez, L., Ruiz□Ramos, M., Gaiser, T., Höhn, J.G., Salo, T., Ferrise, R., Bindi, M., Cammarano, D., Schulman, A.H., 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 der Velde, M., Nisini, L., 2019. Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015. Agric. Syst. 168, 203–212. https://doi.org/10.1016/J.AGSY.2018.06.009

Vezy, R., Buis, S., Lecharpentier, P., Giner, M., 2023. CoPlotR: A Package to Analyse Crop Model Simulations Outputs with Plots and Statistic [WWW Document]. URL https://doi.org/10.5281/zenodo.4442330 (2023)

Vezy, R., Buis, S., Lecharpentier, P., Giner, M., 2021. CroPlotR: A Package to Analyse Crop Model Simulations Outputs: Plots and Statistics.

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P.J., Ittersum, M., Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G., Dumont, B., Eyshi Rezaei, E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C., Jones, C.D., Kassie, B.T., Kersebaum, C.C., Klein, C., Koehler, A.-K., Maiorano, A., Minoli, S., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G.J., Palosuo, T., Priesack, E., Ripoche, D., Rötter, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Wolf, J., Zhang, Z., 2018. Multimodel ensembles improve predictions of crop–environment–management interactions. Glob. Chang. Biol. 0. https://doi.org/10.1111/gcb.14411

Wallach, D., Palosuo, T., Mielenz, H., Buis, S., Thorburn, P., Asseng, S., Dumont, B., Ferrise, R., Gayler, S., Ghahramani, A., Harrison, M.T., Hochman, Z., Hoogenboom, G., Huang, M., Jing, Q., Justes, E., Kersebaum, K.C., Launay, M., Lewan, E., Liu, K., Luo, Q., Mequanint, F., Nendel, C., Padovan, G., Olesen, J.E., Pullens, J.W.M., Qian, B., Seserman, D.-M., Shelia, V., Souissi, A., Specka, X., Wang, J., Weber, T.K.D., Weihermüller, L., Seidel, S.J., 2023a. Uncertainty in crop phenology simulations is driven primarily by parameter variability. bioRxiv 2023.02.03.526931. https://doi.org/10.1101/2023.02.03.526931

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

Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., Dumont, B., Ferrise, R., Gaiser, T., Gayler, S., Hiremath, S., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M., Jansson, P.-E., Jing, Q., Justes, E., Kersebaum, K.C., Launay, M., Lewan, E., Luo, Q., Maestrini, B., Moriondo, M., Olesen, J.E., Padovan, G., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Kumar Srivastava, A., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S.J., 2021b. 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., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hiremath, S., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M., Jansson, P.-E., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Lewan, E., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Padovan, G., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S.J., 2021c. The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environ. Model. Softw. 145, 105206. https://doi.org/10.1016/J.ENVSOFT.2021.105206

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

Wang, B., Feng, P., Liu, D.L., O'Leary, G.J., Macadam, I., Waters, C., Asseng, S., Cowie, A., Jiang, T., Xiao, D., Ruan, H., He, J., Yu, Q., 2020. Sources of uncertainty for wheat yield projections under future climate are site-specific. Nat. Food 1, 720–728. https://doi.org/10.1038/s43016-020-00181-w

Wang, P.C., Shoup, T.E., 2011. Parameter sensitivity study of the Nelder–Mead Simplex Method. Adv. Eng. Softw. 42, 529–533. https://doi.org/10.1016/J.ADVENGSOFT.2011.04.004

Webber, H., Asseng, S., Kimball, B., White, J., Ottman, M., Wall, G.W., De Sanctis, G., Doltra, J., Grant, R., Kassie, B., Maiorano, A., Olesen, J.E., Ripoche, D., Rezaei, E.E., Semenov, M.A., Stratonovitch, P., 2017. Canopy temperature for simulation of heat stress in irrigated wheat in a semi-arid environment: A multi-model comparison. F. Crop. Res. 202, 21–35. https://doi.org/10.1016/J.FCR.2015.10.009

Webber, H., Gaiser, T., Ewert, F., 2014. What role can crop models play in supporting climate change adaptation decisions to enhance food security in Sub-Saharan Africa? Agric. Syst. 127, 161–177. https://doi.org/10.1016/J.AGSY.2013.12.006

Xiong, W., Asseng, S., Hoogenboom, G., Hernandez-Ochoa, I., Robertson, R., Sonder, K., Pequeno, D., Reynolds, M., Gerard, B., 2020. Different uncertainty distribution between high and low latitudes in modelling warming impacts on wheat. Nat. Food 1, 63–69. https://doi.org/10.1038/s43016-019-0004-2

Ypma, J., Johnson, S.G., 2022. Introduction to nloptr: an R interface to NLopt [WWW Document]. URL https://astamm.github.io/nloptr/index.html (2022)

Zhang, S., Tao, F., Zhang, Z., 2017. Uncertainty from model structure is larger than that from model parameters in simulating rice phenology in China. Eur. J. Agron. 87, 30–39. https://doi.org/10.1016/j.eja.2017.04.004

Acknowledgements

This study was carried out in the framework of the Agricultural Model Intercomparison and Improvement Project (AgMIP). KCK was supported by AdAgriF - Advanced methods of greenhouse gases emission reduction and sequestration in agriculture and forest landscape for climate change mitigation (CZ.02.01.01/00/22_008/0004635). Partial funding was provided by the BonaRes project Soil3 (BOMA 03037514, 031B0515C) of the Federal Ministry of Education and Research (BMBF), Germany, by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC 2070 – 390732324 and by the European Union (EU horizon project IntercropValuES, grant agreement No 101081973).

Author contributions

DW was responsible for conceptualization, project administration and writing of the first draft. SB adapted the CroptimizR software to automatically apply the protocol, and ran the software for the STICS model. DMS tested the software using the model HERMES. SB, SS, TP, PT and HM participated in the definition of the protocol and project organization. EJ, CK, BD and ML provided model expertise for the STICS or HERMES models. All authors contributed critical evaluation of the protocol and the manuscript. All authors read and approved the final manuscript.

Data availability

All the results of this study can be reproduced based on input data and artificial measured data, using the STICS model. The model is publicly available. The input and artificial measured data are available from the authors on reasonable request.

Competing Interests

None

Figure Legends

Figure 1. Schema of calibration protocol. The first five steps involve codifying model expertise. Given that information, the calculation steps 6 -8 require no farther model-specific inputs.

Figure 2. Flow diagram for the selection of parameters to estimate, and first estimation of their values, for one variable group. This is done in step 6 of the protocol

Tables

Table 1

Measured and simulated variables, as an example of documentation for steps 2 and 3 of the protocol. There is one row for each measured variable, which also shows the corresponding simulated variable, if any, and the units of the simulated variable. The variables are grouped as explained in the protocol. The order of the groups is that in which the variable groups will be used for calibration. This example is for the STICS model applied to the artificial data for calibration used here. The development stages are stem elongation (BBCH30), heading (BBCH55) and maturity (BBCH90). Biomass refers to aboveground biomass.

Measured variable	Corresponding simulated variable	Units	Number of measurements in calibration data	Variable group	Order for calibration
days from sowing to BBCH30	iamfs	days after sowing	13	phenology	1
days from sowing to BBCH55	ilaxs	days after sowing	13	phenology	1
days from sowing to BBCH90	imats	days after sowing	13	phenology	1
ears	none	NA	NA	NA	NA
biomass at various dates	masec_n	t /ha	44	biomass	2
N in biomass	calculated from QNplanteN_	%	13	plant N	3
grain number	chargefruit	number/m²	13	grain number	4
grain yield	mafruit	t/ha	13	grain yield	5
grain protein	calculated from CNgrain	%	13	grain protein	6

Table 2

Major parameters. Example of documentation for step 4 of the protocol. There is one row for each major parameter of each variable group. The number of major parameters for each group is strictly limited, as explained in the protocol. The upper and/or lower bounds for each parameter can be specified. This example is for the STICS model applied to the artificial calibration data used here.

Group	Major parameter	Default value (bounds)	Short explanation (units)
phenology	stlevamf	324.8 (150,400)	cumulative thermal time from emergence to end of juvenile phase (°C d)
phenology	stamflax	446.8 (150,500)	cumulative thermal time between end of juvenile phase and, end of leaf growth (°C d)
phenology	stdrpmat	820 (500, 900)	cumulative thermal time start of grain filling and maturity (°C d)
biomass	efcroiveg	5.3 (3,6)	maximum radiation use efficiency during vegetative phase (g/MJ)
biomass	efcroirepro	3.5 (3,6)	maximum radiation use efficiency during grain filling (g/MJ)
N in biomass	Vmax2	0.08 (0.002,0.1)	maximum nitrogen uptake rate (µmole/cm/h)
grain_number	cgrain	0.0324 (0.03,0.04)	slope of the relationship between grain number and grain growth rate (grains/(g/d))
grain yield	vitircarbT	0.00031 (0.00005,0.002)	rate of increase of harvest index (1/°C)
grain_protein	vitirazo	0.0064 (0.001, 0.04)	rate of increase of nitrogen harvest index (1/d)

Table 3

Candidate parameters. Example of documentation for step 5 of the protocol. There is one row for each candidate parameter for each variable group. The documentation includes the default value (i.e. best guess) of the parameter. Also, upper and/or lower limits for the parameter can be specified. This example is for the STICS model applied to the artificial calibration data used here.

Group	parameter	Default value (bounds)	Short explanation (units)
phenology	jvc	58.364 (25,60)	number of vernalizing days (d)
phenology	sensrsec	0.8 (0,1)	index of root sensitivity to drougl (1=insensitive) (dimensionless)
phenology	belong	0.0228 (0.005;0.03)	parameter of curve of coleoptil elongation (1/°C)
phenology	jvcmini	11 (2,15)	minimum vernalizing days required (d)
phenology	stressdev	0.6 (0,1)	maximum development delay due t stress (dimensionless)
biomass	dlaimaxbrut	0.003188 (0.000005,0.005)	maximum rate of LAI increase (1/°C)
biomass	durvieF	260 (40,300)	maximum lifespan of an adult lea (dimensionless)
biomass	vlaimax	2.38 (1.5,2.5)	defines shape of LAI curv (dimensionless)
biomass	psisto	12.6 (11,25)	soil water pressure head for stomata closure (bars)
biomass	psiturg	10.6 (1,15)	soil water pressure head at start c decline of cell extension (bars)
N in biomass	croirac	0.348 (0,0.5)	elongation rate of the root apex (cm.degree d-1)
N in biomass	draclong	632 (1,1000)	maximum rate of root length increas (cm/plant/°C)
grain_number	nbjgrain	36 (5,40)	number of days used to compute th number of viable grains (d)
grain yield	pgrainmaxi	0.05528 (0.03,0.065)	maximum grain weight (g)
grain yield	cgrainv0	0.042 (0,0.07)	fraction of maximum grain number fo 0 growth

Table 4

Selection of candidate parameters to estimate. Example of documentation for step 6 of the protocol. Each row shows the list of parameters to be estimated at that stage of the calculations. Candidate parameters that lead to a reduction in AICc compared to the previous smallest value are kept in the list of parameters to estimate. Otherwise, the candidate is removed from the list. This example is for the STICS model applied to the artificial calibration data used here, and for the variable group "biomass". There will be an analogous table for each group of variables.

Group	Parameters to be estimated	AICc	Candidate to be fit to data?
biomass	efcroiveg,efcroirepro	-127.89	yes (automatically)
biomass	efcroiveg,efcroirepro dlaimaxbrut	-166.13	yes
biomass	efcroiveg,efcroirepro dlaimaxbrut, durvieF	-164.97	no
biomass	efcroiveg,efcroirepro dlaimaxbrut, vlaimax	-178.70	yes
biomass	efcroiveg,efcroirepro dlaimaxbrut, vlaimax,psisto	-176.39	no
biomass	efcroiveg,efcroirepro dlaimaxbrut, vlaimax,psiturg	-177.02	no

Table 5

Parameter values. There is a row for each parameter that is estimated. This table is part of the documentation for steps 6 and 7 of the protocol. This example is for the STICS model applied to the artificial calibration data used here. The column of true values has been added here to facilitate evaluation of the protocol. In practical situations, the true values are unknown.

Group	Estimated parameter	True value	Default value	Value after step 6	Value after step 7
phenology	stlevamf	212	324.8	202.14	204.54
phenology	stamflax	367	446.8	356.63	358.95
phenology	stdrpmat	700	820	686.23	686.31
phenology	belong	0.012	0.0228	0.0061	0.0066
phenology	stressdev	0	0.6	0.087	0.093
biomass	efcroiveg	4.25	5.3	4.43	4.55
biomass	efcroirepro	4.25	3.5	3.91	3.43
biomass	dlaimaxbrut	0.00047	0.00318	0.00031	0.00032
biomass	vlaimax	2.2	2.38	2.08	2.09
N in biomass	Vmax2	0.05	0.08	0.014	0.022
grain_number	cgrain	0.036	0.0324	0.037	0.035
grain yield	vitircarbT	0.0007	0.00031	0.00067	0.00068
grain_protein	vitirazo	0.0145	0.0064	0.015	0.014

Table 6

Relative root mean squared error (RRMSE) for the calibration data. The table shows RRMSE for the default parameter values and after parameter estimation in steps 6 and 7, for each variable. This table is part of the documentation for steps 6 and 7 of the protocol. This example is for the STICS model applied to the artificial calibration data used here.

	BBCH30	BBCH55	BBCH90	ln(biomass)	N in biomass	grain number	grain yield	grain protein
Default parameter values	0.156	0.182	0.1588	0.051	0.22	0.31	0.25	0.286
After step 6	0.013	0.013	0.0066	0.021	0.13	0.12	0.17	0.071
After step 7	0.012	0.013	0.0065	0.018	0.12	0.11	0.16	0.076

Table 7

Relative root mean squared error (RRMSE) for simulation of out-of-sample data. The table shows RRMSE for the default parameter values and after parameter estimation in steps 6 and 7, for each variable. This example is for the STICS model applied to the artificial evaluation data used here.

	BBCH30	BBCH55	BBCH90	ln(biomass)	N in biomass	grain number	grain yield	grain protein
Default parameter values	0.1402	0.1563	0.1387	0.047	0.245	0.249	0.166	0.333
After step 6	0.0044	0.0044	0.003	0.015	0.065	0.11	0.084	0.062
After step 7	0.0038	0.0041	0.0023	0.012	0.05	0.095	0.094	0.063