Parameter variability makes a major contribution to uncertainty of crop phenology simulations

Wallach[1], Daniel; Palosuo[2], Taru; Mielenz[3], Henrike; Buis[4], Samuel; Thorburn[5], Peter; Asseng[6], Senthold; Dumont[7], Benjamin; Ferrise[8], Roberto; Gayler[9], Sebastian; Ghahramani[10], Afshin; Harrison[11], Matthew Tom; Hochman[5], Zvi; Hoogenboom[12], Gerrit; Huang[13], Mingxia; Jing[14], Qi; Justes[15], Eric; Kersebaum[16,17,18], Kurt Christian; Launay[19], Marie; Lewan[20], Elisabet; Liu[11], Ke; Luo[21], Qunying; Mequanint[9], Fasil; Nendel[16,17,22], Claas; Padovan[8], Gloria; Olesen[23], Jørgen Eivind; Pullens[23], Johannes Wilhelmus Maria; Qian[14], Budong; Seserman[16], Diana-Maria; Shelia[12], Vakhtang; Souissi[24], Amir; Specka[16], Xenia; Wang[13], Jing; Weber[9,25], Tobias, K.D.; Weihermüller[26], Lutz, Seidel, Sabine J.*[1]

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

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

[3]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Crop and Soil Science , Braunschweig, Germany

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

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

[6]Technical University of Munich, Department of Life Science Engineering, Digital Agriculture, 85354 Freising, Germany

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

[10]Centre for Sustainable Agricultural Systems, Institute for Life Sciences and the Environment, University of Southern Queensland, Toowoomba, Queensland, Australia

[11]Tasmanian Institute of Agriculture, University of Tasmania, Launceston, Tasmania, Australia, 7248

[12]Agricultural and Biological Engineering Department, University of Florida, Gainesville, Florida, USA & Global Food Systems Institute, University of Florida, Gainesville, Florida, USA

[13]College of Resources and Environmental Sciences, China Agricultural University, Beijing, China

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

[15]CIRAD, Univ Montpellier, Persyst Department, Montpellier, France

[16]Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany

[17]Global Change Research Institute CAS, Brno, Czech Republic

[18]Institute for Tropical Plant Production and Agricultural Systems Modelling, Georg-August University, Göttingen, Germany

[19]INRAE, US 1116 AgroClim, Avignon, France

[20]Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

[21]Hillridge Technology Pty Ltd, Sydney, Australia

[22]Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany

[22]Department of Agroecology, Aarhus University, Tjele, Denmark

[24]Swift Current Research and Development Centre, Agriculture and Agri-Food Canada, Swift Current, Saskatchewan, Canada

[25] Soil Science Section, Faculty of Organic Agricultural Sciences, University of Kassel, Germany

[26]Institute of Bio- and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany

[*]Corresponding author. Email: taru.palosuo@luke.fi

Author contribution:

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: expertise on models and simulation results, writing, review & editing.

Abstract

Crop phenology has a major influence on crop yield and is a major aspect of crop response to global warming. Process-based models of phenology are often used to predict the effect of weather on the development rate of crops through their growth phases, but such models are associated with large uncertainties, as demonstrated by the large variability between simulation results of different models. The purpose of this study is to estimate the relative importance of model structure uncertainty (due to uncertainty in the model equations) and model parameter uncertainty. Previous studies have assumed that the choice of model parameters to be fitted to data is fixed, and have evaluated the effect of uncertainty in those parameters. This underestimates parameter variance, as it ignores uncertainty in many aspects of the calibration approach, in particular the choice of parameters to estimate, as well as uncertainty in the parameters not fitted by calibration. Here, we propose and apply two approaches for estimating parameter variance that take into account uncertainty in all aspects of calibration, and, for one approach, uncertainty in parameters not fitted by calibration. Both approaches are based on previously reported large multi-model studies. The first approach uses components of variance analysis, which is possible because these studies included different modeling groups using the same model structure. The second approach is based on a study where multiple modeling groups each applied two different calibration procedures. The variance calculated from the two calibration approaches gives an estimate of parameter variance. Both methods give an overall result of 69% of total variance coming from parameter variance, for those simulated variables for which there are observed data. We conclude

that in order to reduce uncertainty in crop phenology predictions, improving and sharing calibration procedures is as or more important than improving model structure.

1. Introduction

Crop phenology has a major effect on crop yield and is a major aspect of crop response to global warming (Cleland et al., 2007; Liu et al., 2022; Parmesan, 2006). Process-based models of phenology, usually embedded in crop or agroecosystem models, are often used to predict the effect of weather on phenology (Bindi et al., 2015; Muleke et al., 2022). It has become common to study crop performance in general and crop phenology, in particular, using multi-model ensembles, where multiple models are provided with identical inputs and the range of simulated outputs is examined. These studies have shown that the variability in simulated values is substantial (Asseng et al., 2013; Durand et al., 2018; Sándor et al., 2020; Wallach et al., 2021a, 2021b). In several studies where multiple crop models were combined with multiple global circulation models (GCMs), it was been found that the uncertainty contribution from crop models is as large or larger than that from the GCMs (Asseng et al., 2013; Müller et al., 2021; Wang et al., 2020). A reduction in uncertainty from crop models is necessary in order to increase confidence in the results of these models, especially for the simulation of climate change impact assessment and options for adaptation (Maiorano et al., 2017; Rötter et al., 2011).

For given inputs (i.e. weather, crop management, soil characteristics and initial conditions), the uncertainty in crop model simulations has two origins, the uncertainty in crop model structure, which encompasses both the processes included in the model and the specific forms of the equations used to model those processes, and the uncertainty in the parameter values. Only relatively few studies have targeted the relative importance of both structure and parameter uncertainty, and those studies mostly indicate that the contribution of parameter uncertainty is rather small compared to the contribution

from structure uncertainty (Tao et al., 2018; Wallach et al., 2017; Xiong et al., 2020; Zhang et al., 2017). Here, we argue that those studies do not take into account the full extent of parameter uncertainty, and we explain why it is difficult to do so. Therefore, we present here a new analysis, based on multi-model ensemble simulation studies of crop phenology, that show that parameter uncertainty is the major contributor to overall simulation uncertainty.

Model structure uncertainty refers to the range of plausible model structures, where a "plausible" structure would be deemed acceptable by researchers knowledgeable in the field of crop modeling. Often an explicit or implicit requirement for participating in multi-model studies is that the model structure has been published in the scientific literature, and has thus undergone peer review. This can be considered to satisfy the plausibility condition. We also assume that in a multi-model ensemble study, the represented model structures are a random sample from the population of plausible structures, even though this assumption may be hard to fully justify. The problem is that crop multi-model ensembles, like those in other fields, are generally "ensembles of opportunity" (Knutti, 2010; Wallach et al., 2016). That is, whichever modeling groups wish to participate are accepted. In fact, this openness is a principle of the Agricultural Modeling Intercomparison and Improvement Project (AgMIP, Rosenzweig et al., 2013). As a result, the ensemble may well have the same model structure represented more than once, or there may be complex relationships among models, where some models are derived from others or share similarities in structure (Fronzek et al., 2018; Holzworth et al., 2014; van Ittersum et al., 2003) so that the structures are not independent. There have been attempts to account for model interrelatedness in climate modeling (Sanderson et al., 2015) , but this has not been done for crop models so far.

Parameter uncertainty for a given model structure and data set refers to the range of plausible parameter vectors for that structure and data. Here "plausible" means that research groups working with that structure would consider those parameters, or the way in which the parameters are obtained,

127 i.e. model calibration, reasonable. In almost all multi-model ensemble studies, models are calibrated
128 using some data set (which is the same for all models). Thus, the uncertainty in parameterization for a
129 given model structure is due to variability in the calibration method, which includes the choice of how
130 many and which parameters to estimate as well as the numerical procedure for estimating parameters,
131 and in the choice of default parameter values (i.e. the values assigned to the parameters whose values
132 are not estimated by calibration).

133 To study the consequences of uncertainty in model structure and/or in model parameters, one
134 needs a sample drawn from the range of plausible structures and/or from the range of plausible
135 parameters for one or more structures. As discussed, the models represented in multi-model ensemble
136 studies are generally accepted as a sample from plausible structures, even though this is not expected to
137 be exactly true. How to sample from plausible parameter vectors on the other hand is not clear.

138 Wallach et al. (2023) identified several steps in crop model calibration, including the choice
139 of default parameter values, the choice of observed data to use in the objective function for frequentist
140 calibration, the choice of parameters to be estimated from the data and finally the choice of algorithm
141 and software for doing the calibration. They showed that different modeling groups made different
142 choices, and that even different modeling groups using the same model structure made different
143 choices. We suggest that sampling from plausible parameter vectors thus involves sampling from the
144 plausible choices at each step of the calibration procedure, including the choice of default parameter
145 values.

146 There has been a range of previous studies that have sampled model parameters Zhang et al.
147 (2017) used five rice phenology models and, for a chosen set of parameters, generated 50 randomly
148 perturbed sets of values to obtain a sample of parameter values. Wallach et al. (2017) estimated
149 parameters for two rice phenology models using weighted least squares, then generated samples of

150 parameters using the estimated variance-covariance matrix of the estimated parameters. Tao et al.
151 (2018) using seven barley models identified a reasonable range for each parameter and generated a
152 super ensemble, where each model was associated with multiple parameter values drawn from the
153 range of that parameter. Xiong et al. (2020) compared three different crop models and four
154 parametrization strategies that differed in the amount of data available for calibration. Other studies
155 have specifically considered the way in which the model parameters are estimated. Kawakita et al.
156 (2019), in a study of time to heading in wheat, assessed three models and three different optimization
157 algorithms for calibration, leading to three different parameter vectors for each model. Jha et al. (2022)
158 compared three calibration approaches for CERES Maize, namely grid search, GLUE, and a genetic
159 algorithm. Gao et al. (2020) compared the predictions and uncertainty estimates for the CERES Rice
160 model using a weighted least squares calibration method, the GLUE method, and a Bayesian approach.
161 Harrison et al. (2019) compared search techniques for calibrating the APSIM model. In these studies, the
162 choice of parameters to be treated as uncertain, as well as the choices for most of the calibration steps
163 listed above, are taken as given. Also, no uncertainty in the default parameter values is taken into
164 account. Thus those studies ignore several causes of uncertainty in parameters that exist in multi-model
165 ensemble studies (Albanito et al., 2022; Wallach et al., 2021c) and could be important, and thus, those
166 studies are likely to underestimate the importance of parameter uncertainty.

167 A straightforward approach to estimating parameter and model structure uncertainty would be
168 to specifically create a multi-model ensemble in which each model structure is used by more than one
169 modeling group, each of which independently estimates the model parameters. In practice, this is
170 difficult to achieve. Confalonieri et al. (2016) organized a study with five crop models and 20 users,
171 leading to different model parameterizations. However, in that case, the groups responsible for
172 calibration were students and not model experts, so the parameter vectors might not be considered
173 "plausible" by model experts.

The goal of this study is to obtain estimates of parameter uncertainty, and of the relative importance of parameter and structure uncertainty, which for the first time fully take into account the uncertainty as to the choice of parameters to estimate, the method of estimation and the default values of the parameters which are not estimated. A first sub-objective is methodological. We show how such estimates can be obtained. A second sub-objective is to obtain parameter uncertainty estimates for several particular cases, as an illustration of how parameter ad structure uncertainty compare. The specific application is the estimation of wheat phenology in multiple environments, so the calculated uncertainties only refer to these situations.

Two different approaches are employed; both involve a new method of analyzing results from previously reported multi-model studies. The first is based on two previously reported multi-model ensemble studies which were not specifically designed to study parameter uncertainty, but where it turned out that some of the model structures were used by more than one modeling group. Those groups independently chose default parameter values and the calibration procedure. The differences between modeling groups using the same structure give information about parameter uncertainty. The second approach is based on a previously reported study where multiple modeling groups each used two different calibration procedures. The differences between the two procedures are used to estimate parameter uncertainty.

2. Materials and Methods

2.1 Data

This study revisits three simulation experiments on wheat phenology that have been previously reported (Wallach et al., 2023, 2021a, 2021b). In the previous studies, the emphasis was on prediction accuracy, whereas in the present study the simulation results are reanalyzed with the objective of

196 estimating the contribution of parameter uncertainty to the overall uncertainty of the phenology
197 prediction.

198 The simulations concern three data sets (**Table 1**). Two of the data sets are from winter wheat
199 variety trials in France. These data sets have identical structures (same sites and sowing dates) but
200 concern two different varieties, "Apache" and "Bermude". Both varieties are of intermediate precocity,
201 but Bermude has a longer cycle. The observed data from each environment consist of days of stem
202 elongation (BBCH30 on the BBCH scale, Meier, 1997) and the middle of heading (BBCH55). The third data
203 set is from a multi-site sowing date experiment on spring wheat variety "Janz" in Australia. The original
204 data were observations of growth stage at weekly intervals. The data provided to modeling groups were
205 the dates of each BBCH stage, from the first to the last observed stage, derived from interpolation of the
206 original data. All data sets were split into two subsets, one for calibration and the other for evaluation.
207 The calibration and evaluation data sets had neither site nor year in common. Thus evaluation was a true
208 test of how well models could predict out-of-sample conditions.

Data set	Number of calibration environments	Number of evaluation environments	Observed development stages*	Simulated development stages used for analysis*
France Variety Apache	14	8	BBCH30, BBCH55	BBCH10, BBCH30, BBCH55
France Variety Bermude	14	8	BBCH30, BBCH55	BBCH10, BBCH30, BBCH55
Australia Variety Janz	24	18	BBCH stage at weekly intervals	BBCH10, BBCH30, BBCH65, BBCH90

*The development stages, using the BBCH scale (Meier, 1997) are emergence (BBCH10), start of stem elongation (BBCH30), middle of heading (BBCH55), full flowering (BBCH65), and maturity (BBCH90).

Table 1. Data sets

The three data sets used in the multi-model ensemble studies analyzed here.

2.2 Participating modeling groups and model structures

The call for participants in the first study was published using the mailing lists of several models and of AgMIP. All modeling groups that volunteered (a modeling group is one or more individuals that work together to run a model) were accepted. All groups used model structures that have been published. A total of 30 modeling groups using 23 different structures participated in one or more of the simulation experiments. For the purposes of this study, each model with a unique name was considered as a different model structure. A list of structures represented here, together with references that describe the model equations, is given in Supplementary Table 1. In all cases, the phenology models were embedded in general crop models, so the model structure here refers to the overall crop model. Most but not all of the groups participated in all the studies.

No specific attempt was made to encourage or discourage participation by different groups that used the same structure. In the event, there were two or three (depending on the simulation study) model structures used by more than one group. In particular, in the first two studies three structures were used by 2-4 different modeling groups (Table 2).

In all studies, the different modeling groups were identified by a code (M1, M2 etc.) without identifying the group or the model structure used. The same codes are used here. In addition, we assign here codes to the different model structures (S1, S2, etc.). The same model structure can be associated with more than one modeling group.

Simulation study	Data set (country, cultivar)	Calibration procedure	Number of modeling groups	Repeated structures*
1	France, Apache	Usual	27	S2 (4), S4 (3), S18 (2)
1	France, Bermude	Usual	27	S2 (4), S4 (3), S18 (2)
2	Australia, Janz	Usual	28	S2 (3), S4 (3), S18 (2)
3	France, Apache	Protocol	19	S2 (3), S4 (2)
3	France, Bermude	Protocol	19	S2 (3), S4 (2)
3	Australia, Janz	Protocol	19	S2 (3), S4 (2)

*The model structures used by more than one modeling group, and in parentheses, the number of groups that used that structure.

Table 2. Simulation studies

The three different simulation studies analyzed here.

2.3 Simulation studies

In the first simulation study (Wallach et al., 2021a), based on the French data sets, each participating modeling group implemented their usual calibration method to calibrate their model using the calibration data subsets and then simulated days from sowing to phenological stages BBCH10, BBCH30, and BBCH55 for both varieties and for both the calibration and evaluation environments (Table 2). In the second study (Wallach et al., 2021b), each participating modeling group again used their usual calibration method and then simulated days from sowing to phenological stages BBCH10, BBCH30,

245 BBCH65, and BBCH90 for the Australian environments. Neither the French nor the Australian data sets
246 had observations of BBCH10, which was included in the variables to simulate in order to have an
247 example of a simulated stage without calibration data.

248 Detailed information about the variability in calibration procedures among the different
249 modeling groups is presented in Wallach et al. (2021c). The detailed descriptions of the calibration
250 procedures used by each group in the first two studies were used to identify the range of decisions for
251 each calibration step. Based on an analysis of the different choices, and following discussions with the
252 participating groups, a new detailed calibration procedure (hereafter the calibration "protocol") was
253 proposed and applied to all three data sets by all participating groups. Any group that participated in the
254 third study but that had not participated in the first two studies first did the calibration using their usual
255 method, before applying the protocol.

256 The proposed calibration protocol is described in detail by Wallach et al. (2023). The following is
257 a brief summary of that protocol. The protocol covers all steps in phenology model calibration and
258 includes templates for documenting the different steps. A major innovation, compared to the usual
259 calibration procedures, is in the choice of parameters to estimate. In the protocol, this is done in two
260 stages. First, parameters that play an almost additive role, i.e. that change the simulated values by a
261 similar amount in all environments, are identified. At most, there can be as many almost additive
262 parameters as the number of different variables used for calibration (for example, two almost additive
263 parameters if BBCH30 and BBCH55 are used for calibration). In most cases, the almost additive
264 parameters are degree days to development stages at or near the observed stages. Those parameters
265 are estimated based on the calibration data. This should remove bias in the simulations, or nearly do so.
266 Then a series of candidate parameters, thought likely to explain part of the remaining unexplained
267 variability between environments, is identified, and those parameters are tested one by one. If the
268 estimation of a candidate parameter, together with all previously chosen parameters, sufficiently

improves the fit to the data, as measured by the Bayesian Information Criterion (BIC) (Chakrabarti and Ghosh, 2011), it is included in the final list of estimated parameters. If not, it is kept at its default value. A comparison between usual and protocol calibration for the participants in the third study showed major differences between the two. For example, the number of estimated parameters differed between the two procedures for the great majority of participants (Wallach et al., 2021c).

2.4 Estimation of structure and parameter variance

Variance components approach

For a given environment, with given inputs, the variability in the simulated values comes from the variability in structures and parameters. We assume that for each simulation experiment, the model structures are a random sample from plausible structures. It is further assumed that the parameter vector used by each modeling group in simulation studies 1 and 2 is a random sample of plausible parameter vectors for the corresponding model structure. Presumably, all the modeling groups were well-versed in using their model, which fulfills the plausibility requirement for the parameters. The random sample assumption is satisfied, since there was no attempt to choose groups that used a particular approach to parameter estimation. That is, for a given model structure, all plausible parameter estimation approaches had an equal chance of being represented in the multi-model studies. Finally, we assume independence between the parameter estimation choices of different modeling groups.

The statistical model for a simulated value is then

$$y_{sp} = \mu + \alpha_s + \beta_{sp}$$

where y_{sp} is the simulated value for model structure *s* and the parameter vector *p*, μ is the overall mean ,α_s is the effect of model structure and β_{sp} is the effect of parameter vector nested within the model

structure. There is no residual term because these are simulated values with no measurement error. We treat α_s and β_{sp} as random effects, since we are interested in the variability in the population not in the specific model structures or parameter vectors in the sample. The variances are noted $\mathrm{var}(\alpha_s) = \sigma^2_{structure}$ for all *s* and $\mathrm{var}(\beta_{sp}) = \sigma^2_{parameters}$ for all *sp*. The random terms are assumed independent so that the total variance for one environment and one simulated variable is the sum of model structure variance and parameter variance (Scheffé, 1959):

$$\sigma^2_{total} = \sigma^2_{structure} + \sigma^2_{parameters} \tag{2}$$

To estimate parameter and model structure variance simultaneously , we use the lmer function of the lme4 package (Bates et al., 2015) in R (R Core Team, 2017), which is designed to analyze linear mixed effect models (SI). Note that while the estimation of parameter variance depends specifically on the variability between modeling groups using the same model structure, the estimation of structure variance uses the variability between all modeling groups.

This "components of variance" approach to estimating parameter variance is not used for simulation study 3, since there were too few model structures used by more than one modeling group. An illustration of the data used in this approach is given in **Error! Reference source not found.**. As that figure shows, three of the model structures (S2, S4, and S18) were used by more than one modeling group. It is the variability between the different groups using the same model structure that allows us to estimate parameter variance. The results for all groups contribute to the estimation of structure variance. Note that the simplified approach to estimating parameter uncertainty using these same data in Wallach et al. (2021a, 2021b) led to under-estimation of the contribution of parameter uncertainty.

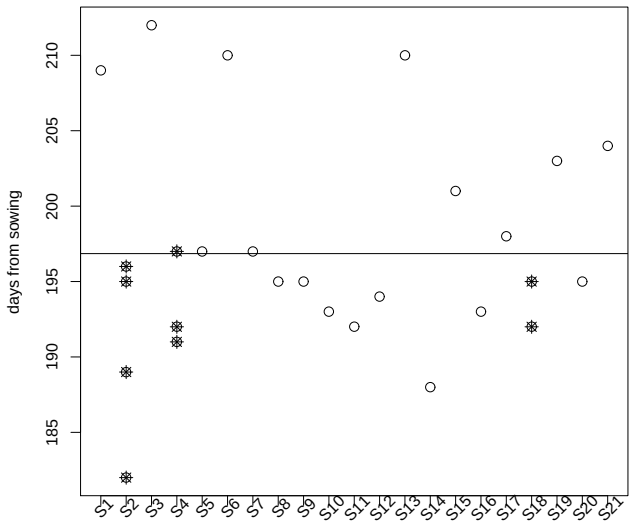

Figure 1

Example of data used for estimating parameter and structure uncertainty in the "variance components" approach. Each circle or star is a simulated value of days from sowing to mid-heading (stage BBCH55) for variety Apache in one environment from simulation study 1. All modeling groups used the same inputs. The x-axis shows the model structure. In most cases, each structure was used by a single modeling group (open circles), but structures S2, S4, and S18 were used respectively by 4, 3, and 2 groups (data shown as stars). It is the variability between simulations from different groups using those model structures that is the basis for estimating parameter uncertainty. The horizontal line is the average of all simulated values.

Comparison of usual and protocol calibration

The second estimation of parameter uncertainty is based on all three simulation studies (see Table 2). In studies 1 and 2 every group used their usual calibration approach and in study 3 used the

newly proposed protocol calibration approach to obtain parameters for each data set. The estimate of parameter uncertainty here is based on the difference between these two calibration methods. The two simulation results for every modeling group for one environment are shown in Figure 2

Each group's usual calibration approach can be considered to be a random draw from plausible calibration approaches for their model structure. The protocol calibration approach, which was developed in collaboration with all participating modeling groups, is also by construction a plausible calibration approach. It is essentially independent of the usual approach since the contribution of any single group to the protocol was limited. Furthermore, the protocol did not mimic closely any single usual calibration approach. Therefore, for each modeling group we treat the usual and the protocol calibration as two random draws from the population of plausible calibration approaches for the model structure used by that group. For each environment, we calculate the variance of the two calibration approaches, noted var_i for modeling group *i* (eq. 1), and then average over modeling groups to obtain an estimated parameter variance for that environment (eq. 2).

$$\mathrm{var}_i = \left(y^{usual} - \overline{y}\right)^2 + \left(y^{protocol} - \overline{y}\right)^2 \quad (1)$$

$$\hat{\sigma}^2_{parameters} = 1/R \sum_{i=1}^{R} \mathrm{var}_i \quad (2)$$

where $\overline{y} = (y^{usual} + y^{protocol)} / 2$ and R is the number of modeling groups.

We do not have a rigorous method of estimating structure variance from these data. The difficulty is that the assumptions underlying the random effects model described above are not satisfied; the parameter vectors obtained by protocol calibration are not independent between modeling groups, since all use the same protocol. Therefore, or each environment we use the estimate of structure

variance from the variance components approach above.

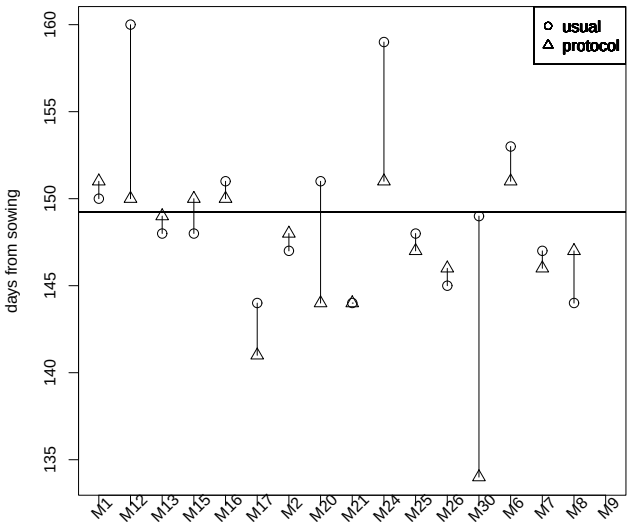

Figure 2

Example of data used for usual compared to protocol approach to estimating parameter uncertainty Simulated days from sowing to stem elongation (stage BBCH30) for variety Bermude in one environment from simulation study 3, which compared usual and protocol calibration for each modeling group. Each code (M1, etc.) refers to a specific modeling group (x-axis). For each modeling group, there are two simulated values, one using parameters based on usual calibration and the other using parameters based on protocol calibration. The differences between the two values are used to estimate parameter variance. The horizontal line shows the mean of simulations using usual calibration.

3. Results

First, all calculations were done for the calibration and evaluation data subsets separately. However, we found that there were no systematic differences between the two (Fig. S1 and Fig. S2). All results here are therefore based on the combined calibration and evaluation data.

Figure 3 shows parameter variance as a fraction of total variance for each environment of the two French data sets, using variance components or usual compared to protocol calibration. Figure 4 shows the same information for the Australian data set. **Table 3** and

Table 4 show parameter variance, and parameter variance as a fraction of total variance, averaged over all environments for the French and Australian data sets, respectively.

Table 4

Parameter and structure variance averaged over environments of the Australian data set. The variance components estimate is based on ensemble studies where some model structures were used by multiple modeling groups. The usual compared to protocol estimate is based on comparing usual and protocol calibration results for each modeling group.

shows results averaged over all environments of both the French and Australian data sets. Separate values are given for BBCH10 (for which there was no observed data) and for all other stages.

Two aspects of the results in Table 3 and

Table **4** are worth emphasizing. First, the estimated contribution of parameter variance to total variance is always much less for BBCH10 than for the other simulated development stages. This is the case for both approaches, but the difference is particularly important for the approach based on a comparison of usual and protocol calibration.

Secondly, for development stages other than BBCH10, the estimated contribution of parameter uncertainty is large. It ranges from 24 to 93%, and is larger than 50% for 11 out of the 14 combinations of data set, development stage and approach for estimating the contribution of parameter variance.

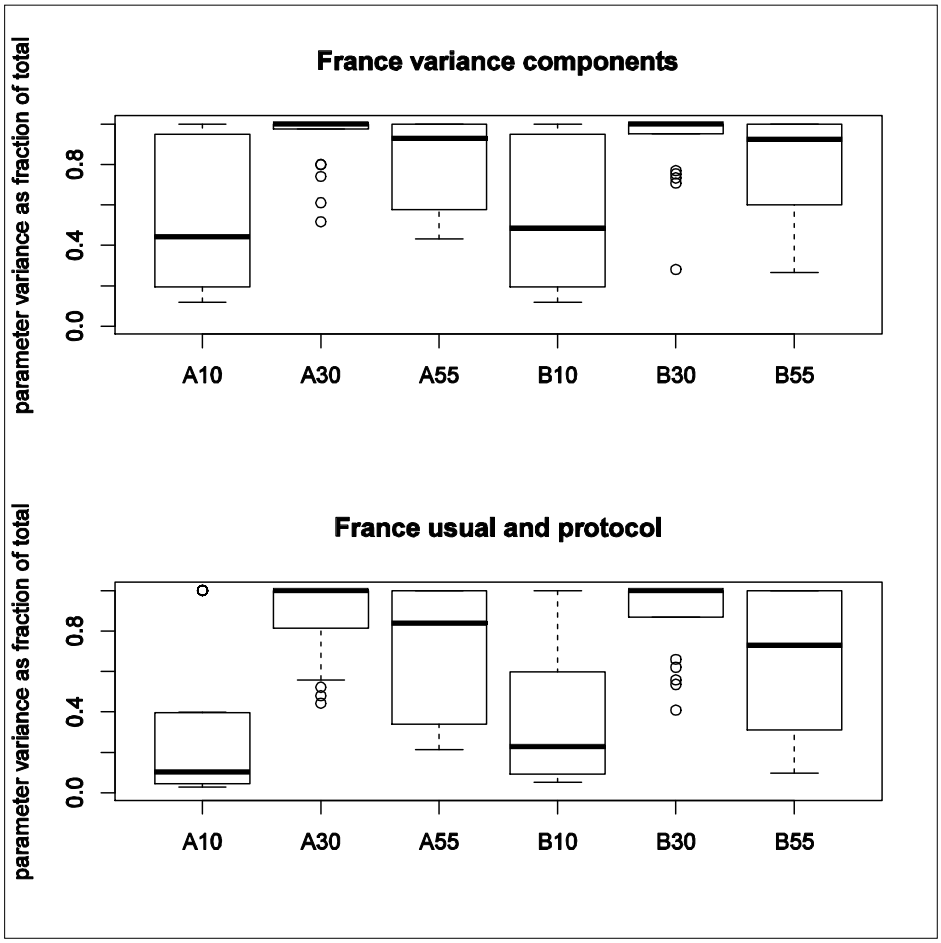

Figure 3

Fraction of total variance due to parameter variance, for the French data sets, based on two different approaches: based on variance components (top) or based on the comparison of usual and protocol calibration (bottom). A10, A30, and A55 (B10, B30, B55) refer respectively to simulations of days from sowing to emergence, to start of stem elongation and to middle of heading for the variety Apache (Bermude). Boxes indicate the lower and upper quartiles. The solid line within the box is the median. Whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the outlier dots are those observations that are beyond that range.

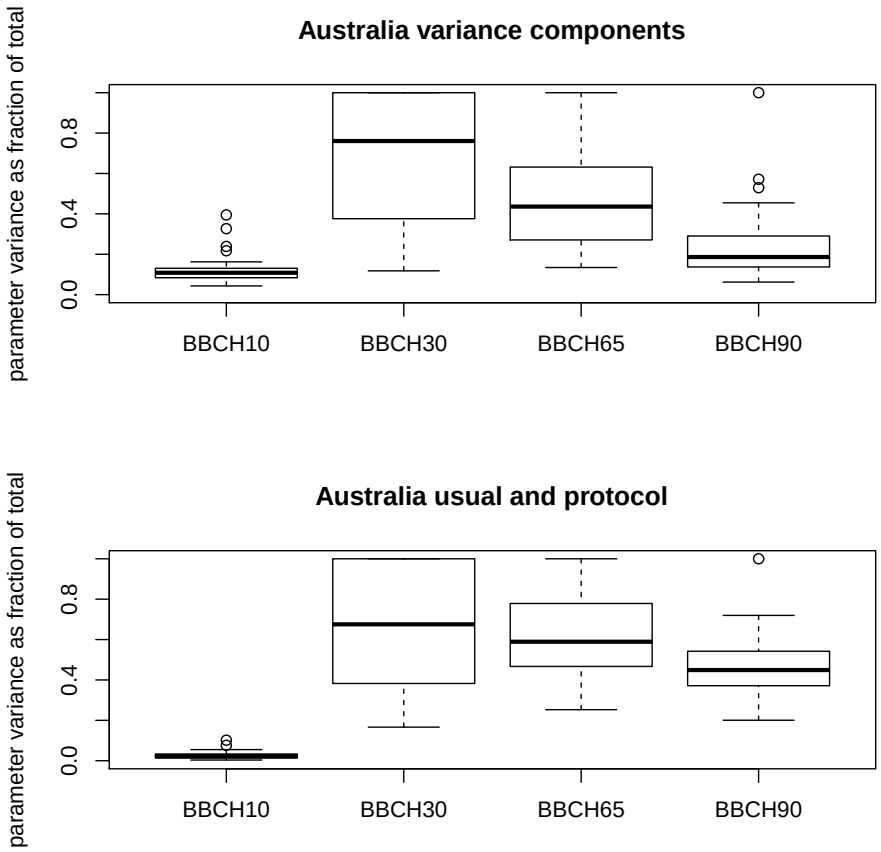

Figure 4

: Fraction of total variance due to parameter variance, for the Australia data set, based on two different approaches: based on variance components (top) or based on comparison of usual and protocol calibration (bottom). Simulations are of days from sowing to emergence, start pf stem elongation, full flowering and maturity. Boxes indicate the lower and upper quartiles. The solid line within the box is the median. Whiskers indicate the most extreme data point which is no more than 1.5 times the interquartile range from the box, and the outlier dots are those observations that are beyond that range.

407

		Apache			Bermude		
Method of estimating parameter variance as fraction of total	Development stage	BBCH10	BBCH30	BBCH55	BBCH10	BBCH30	BBCH55
Variance components	Average number of modeling groups	23	25	27	23	25	27
	Number of model structures	17	19	21	17	19	21
	Parameter variance	11.74	52.94	26.77	12.05	71.25	36.41
	Structure variance	5.05	4.29	9.65	4.99	10.86	12.73
	Parameter variance as fraction of total	0.55	0.93	0.81	0.56	0.92	0.8
Usual compared to protocol calibration	Average number of modeling groups	14	16	17	14	16	16.9
	Parameter variance	0.75	16.62	10.35	2.57	28.52	10.01

	Parameter variance as fraction of total	0.28	0.88	0.71	0.39	0.89	0.65

Table 3

Parameter and structure variance averaged over environments of the French data sets. The variance components estimate is based on ensemble studies where some model structures were used by multiple modeling groups. The usual compared to protocol estimate is based on comparing usual and protocol calibration results for each modeling group. Parameter variance, structure variance and the ratio of parameter variance to total are averages over the values for each environment. The latter is somewhat different than the ratio of average parameter variance to average total variance.

Method for estimating parameter variance as fraction of total	Development stage	BBCH10	BBCH30	BBCH65	BBCH90
Variance components	Average number of modeling groups	22.95	23.93	27.64	26.05
	Average number of model structures	17.95	19.93	22.64	21.05
	Parameter variance	40.69	102.88	47.1	30.3

	Structure variance	276.33	45.74	48.4	90.71
	Parameter variance as fraction of total	0.12	0.66	0.49	0.24
Usual compared to protocol calibration	Average Number of modeling groups	13.9	16.9	18.9	17.3
	Parameter variance	8.41	62.54	67.12	80.76
	Parameter variance as fraction of total	0.026	0.65	0.61	0.46

Table 4

Parameter and structure variance averaged over environments of the Australian data set. The variance components estimate is based on ensemble studies where some model structures were used by multiple modeling groups. The usual compared to protocol estimate is based on comparing usual and protocol calibration results for each modeling group.

Method for estimating parameter variance as fraction of total	BBCH10	All other stages
Variance components	0.41	0.69
Usual compared to protocol calibration	0.23	0.69

Table 5

Fraction of total variance due to parameter variance based on two estimation procedures: variance components and comparison of usual and protocol calibration. Averages are over all environments of both the French and Australian data sets, either for the development stage BBCH10 or for all other development stages.

4. Discussion

There are relatively few studies that have estimated separately the contributions of parameter uncertainty and model structure uncertainty to total simulation uncertainty in ensemble studies with crop models or crop phenology models. Those studies focused on parameter variability when the list of parameters to be estimated is fixed. This includes studies where the parameter distribution is based on expert opinion (Tao et al., 2018; Zhang et al., 2017) where parameters differ due to the use of different levels of gridded information (Xiong et al., 2020) or where the variance-covariance matrix of the parameter vector is estimated during least squares parameter estimation (Wallach et al., 2017). These studies have systematically found that structural uncertainty is substantially more important than parameter uncertainty.

Those studies sampled from the distribution of values for some fixed subset of model parameters. However, this does not really take into account the full parameter uncertainty in crop or phenology models, which arises from all aspects of the calibration approach, not just from uncertainty in a fixed subset of parameters (Albanito et al., 2022; Wallach et al., 2021c). In addition, there is uncertainty in the values of the fixed parameters. The two approaches we propose here take into account more realistically the full uncertainty in model parameterization.

Our first approach is based on variability in simulated values between different modeling groups using the same model structure. The underlying hypothesis is that the most realistic way to sample true parameter uncertainty in crop or phenology models is to sample established crop modeling groups using

the same model structure. If they are provided with the same input data and the same observational data for calibration, the parameter vectors they obtain are by definition plausible parameter vectors for that structure and those data, since our definition of plausible is parameter vectors that a group well-versed in modeling could obtain. It was found in the studies that were analyzed here, that different groups took different decisions at all steps in the calibration procedure, including the choice of which parameters to estimate, the form of the objective function, and the algorithm used for searching for the best parameters (Wallach et al., 2021c). This confirms that uncertainty in calibrated parameters arises from uncertainty in all aspects of the calibration procedure. Our approach automatically takes all those differences in the calibration procedure into account, as well as differences in the choice of default parameter values. A limitation of the present study is that only three model structures were used by more than one group, and at most four groups used the same model structure.

Our second estimate of parameter uncertainty is based on a comparison, for multiple modeling groups, between two different calibration procedures, their usual procedure, and a new protocol procedure. In general, there were major differences between the two calibration procedures for each group (Wallach et al., 2023). Both are plausible in the sense that they were developed and used by modeling groups well-versed in the use of their model structure. It seems reasonable then to assume that for each modeling group, the two are a random sample of size two of plausible procedures for the corresponding model structure. One limitation of this second estimate is that the same modeling group implemented both calibration procedures. It seems likely that the default parameter values were mostly unchanged between the two calibration approaches. Since the variability in the fixed parameter values is not sampled, this approach underestimates the contribution of parameter uncertainty to the total uncertainty.

Both methods assume that the model structures in the multi-model studies area random sample of plausible structures. At best, this will only be approximately true. If the sample is under-representative

476 (i.e. has less variability than actually exists among crop models), then structure uncertainty will be
477 underestimated. The same holds for calibration methods. If the actual variability in calibration methods
478 is larger than is sampled in the multi-model ensembles used here, then parameter uncertainty is
479 underestimated.

480 In the estimates obtained here, parameter variance as a fraction of total variance is
481 systematically smaller for emergence (stage BBCH10) than for the other stages. The fraction is
482 particularly small for the estimates based on usual compared to protocol calibration. This is likely due to
483 the fact that there were no data for days to emergence, and so simulated days to emergence were
484 usually determined by the default parameter values. Probably, most modeling groups used the same
485 default values for usual and protocol calibration, leading to no difference in simulated days to
486 emergence between usual and protocol calibration. The contribution of parameter uncertainty to total
487 uncertainty for stages other than BBCH10, as estimated here, is similar for the two approaches to
488 estimation of parameter uncertainty. This suggests that when there are data related to a development
489 stage, differences between groups are mostly due to differences in calibration procedure rather than to
490 differences in default values.

491 The contribution of parameter uncertainty to total uncertainty, averaged over development
492 stages and data sets, is 69%. This implies that in a multi-model ensemble study simulating phenology,
493 over two-thirds of the variability is due to parameter uncertainty and only 31%, or about one-third is due
494 to structure uncertainty, i.e. to the variability among plausible model structures. The estimates obtained
495 here come from three different data sets, with 22 or 42 different environments. The French and
496 Australian data sets not only had quite different environmental conditions and plant material (winter
497 wheat in France and spring wheat in Australia), but also quite different measured variables (two
498 development stages for the French data set, weekly development stage for the Australian data set). At
499 least 19 modeling groups participated in each study used here. There is substantial variability in results

for different environments and different simulated variables, but the range of situations analyzed gives some confidence in the overall conclusion, that parameter uncertainty is a major contribution to overall model uncertainty when simulating phenology.

It is important to obtain information on the contribution of parameter uncertainty in other conditions than those considered here. The most realistic estimate of parameter uncertainty is based on independent modeling groups using the same model structure. One possibility would be to proactively create multi-model ensembles where the same structure is represented by multiple modeling groups. Another approach would be to take advantage of those multi-model ensemble studies where, by chance, the same structure is represented by more than one modeling group. The variance components calculation applied here could be applied quite generally to those studies. In some cases, the results of past multi-model ensemble studies have been documented in data journals (e.g. Martre et al., 2018). Those studies could be revisited in order to estimate parameter and structure uncertainty.

The major role of parameter uncertainty emphasizes the importance of documenting all parameter values, both fixed and calibrated, for each model in each study, since these may have as much or more effect on simulated values as the model equations. Without that information, the results of a simulation study cannot be reproduced.

Much, though not all, of the parameter uncertainty, comes from the variability in calibration approaches between modeling groups. The obvious path to the reduction of this source of uncertainty is to have a single calibration approach for each model structure. Of course, this is only worthwhile if the common calibration approach leads to a parameterized model that predicts well. There is no consensus on the best calibration approach for crop models. However, there have been recent attempts to define a calibration protocol that could be applied generally to phenology models, and that has been found both to reduce uncertainty and to reduce the mean squared error of prediction (Wallach et al., 2023).

The results here focus on simulating phenology, one component of crop models. Full crop models simulate many other processes and not just development, which may lead to larger structure uncertainty. However, full crop models have many more parameters than phenology models and parameter estimation of an entire crop model is more complex, which may increase parameter uncertainty. It is thus not clear how the results here scale up to more complex models. In any case, however, it is important to take into account the full extent of parameter uncertainty in analyzing structural versus parameter uncertainty for full crop models.

5. Conclusions

The large uncertainties associated with crop models in general and crop phenology models in particular clearly limit the utility and acceptance of these models for impact assessment and other practical applications. Previous studies that have aimed at quantifying the separate contributions of model structure and model parameter uncertainty have concluded that model parameter uncertainty is relatively less important, leading to the implicit conclusion that research should be focused on the processes and equations in crop models. We find here that when one takes into account the full uncertainty in the calibration procedure and also uncertainty in the default parameter values, parameter uncertainty is the major cause of uncertainty in simulated values of crop phenology models. These results are of course specific to the situations examined here, but they strongly suggest that The development and application of improved, shared calibration procedures as well as increased attention to the choice of default parameter values should be a major goal of the crop modeling community, on a par with work on improving model equations.

6. Acknowledgements

This work was in part supported by the Collaborative Research Center 1253 CAMPOS (Project 7: Stochastic Modelling Framework), funded by the German Research Foundation (DFG, Grant Agreement

SFB 1253/1 2017), the Academy of Finland through projects AICropPro (316172) and DivCSA (316215) and Natural Resources Institute Finland (Luke) through a strategic project BoostIA, the BonaRes projects "Soil3" (BOMA 03037514) and "I4S" (031B053I) 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 project BiomassWeb of the GlobeE programme (Grant number: FKZ031A258B) funded by the Federal Ministry of Education and Research (BMBF, Germany), the INRA ACCAF meta-programme, 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): BonaRes Centre for Soil Research, subproject B" (grant 031B0511B), the National Key Research and Development Program of China (2017YFD0300205), the National Science Foundation for Distinguished Young Scholars (31725020), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the 111 project (B16026), and China Scholarship Council, the Agriculture and Agri-Food Canada's Project J-002303 under the Interdepartmental Research Initiative in Agriculture, the DFG Research Unit FOR 1695 'Agricultural Landscapes under Global Climate Change – Processes and Feedbacks on a Regional Scale, the U.S. Department of Agriculture National Institute of Food and Agriculture (award no. 2015-68007-23133) and USDA/NIFA HATCH grant N. MCL02368, the National Key Research and Development Program of China (2016YFD0300105), The Broadacre Agriculture Initiative, a research partnership between University of Southern Queensland and the Queensland Department of Agriculture and Fisheries, the Academy of Finland through project AI-CropPro (315896), 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), the SYSTEMIC funded by JPI HDHL, JPI-OCEANS and FACCE-JPI under ERA-NET (n.696295), the SustEs project funded by the Ministry of Education, Youth and Sports of the Czech Republic (CZ.02.1.01/0.0/0.0/16_019/000797). The time of MTH and KL was supported by the Grains Research

and Development Corporation (Project UOT1906-002RTX). The observed Australian phenology data were jointly funded by CSIRO, the Grains Research and Development Corporation (GRDC) under the "Adding Value to GRDC's National Variety Trial Network" project (CSA00027). The order in which the donors are listed is arbitrary.

7. References

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

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

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01

592 Bindi, M., Palosuo, T., Trnka, M., Semenov, M., 2015. Modelling climate change impacts on crop
593 production for food security. Clim. Res. 65, 3–5. https://doi.org/10.3354/cr01342

594 Chakrabarti, A., Ghosh, J.K., 2011. AIC, BIC and Recent Advances in Model Selection. Philos. Stat. 583–
595 605. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

596 Cleland, E.E., Chuine, I., Menzel, A., Mooney, H.A., Schwartz, M.D., 2007. Shifting plant phenology in
597 response to global change. Trends Ecol. Evol. https://doi.org/10.1016/j.tree.2007.04.003

598 Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., Pagani, V., Cappelli, G.,
599 Vertemara, A., Alberti, L., Alberti, P., Atanassiu, S., Bonaiti, M., Cappelletti, G., Ceruti, M.,
600 Confalonieri, A., Corgatelli, G., Corti, P., Dell'Oro, M., Ghidoni, A., Lamarta, A., Maghini, A.,
601 Mambretti, M., Manchia, A., Massoni, G., Mutti, P., Pariani, S., Pasini, D., Pesenti, A., Pizzamiglio, G.,
602 Ravasio, A., Rea, A., Santorsola, D., Serafini, G., Slavazza, M., Acutis, M., 2016. Uncertainty in crop
603 model predictions: What is the role of users? Environ. Model. Softw. 81, 165–173.
604 https://doi.org/10.1016/j.envsoft.2016.04.009

605 Durand, J.-L., Delusca, K., Boote, K., Lizaso, J., Manderscheid, R., Weigel, H.J., Ruane, A.C., Rosenzweig, C.,
606 Jones, J., Ahuja, L., Anapalli, S., Basso, B., Baron, C., Bertuzzi, P., Biernath, C., Deryng, D., Ewert, F.,
607 Gaiser, T., Gayler, S., Heinlein, F., Kersebaum, K.C., Kim, S.-H., Müller, C., Nendel, C., Olioso, A.,
608 Priesack, E., Villegas, J.R., Ripoche, D., Rötter, R.P., Seidel, S.I., Srivastava, A., Tao, F., Timlin, D.,
609 Twine, T., Wang, E., Webber, H., Zhao, Z., 2018. How accurately do maize crop models simulate the
610 interactions of atmospheric CO2 concentration levels with limited water supply on water use and
611 yield? Eur. J. Agron. 100, 67–75. https://doi.org/10.1016/J.EJA.2017.01.002

612 Fronzek, S., Pirttioja, N., Carter, T.R., Bindi, M., Hoffmann, H., Palosuo, T., Ruiz-Ramos, M., Tao, F., Trnka,
613 M., Acutis, M., Asseng, S., Baranowski, P., Basso, B., Bodin, P., Buis, S., Cammarano, D., Deligios, P.,
614 Destain, M.-F., Dumont, B., Ewert, F., Ferrise, R., François, L., Gaiser, T., Hlavinka, P., Jacquemin, I.,

Kersebaum, K.C., Kollas, C., Krzyszczak, J., Lorite, I.J., Minet, J., Minguez, M.I., Montesino, M., Moriondo, M., Müller, C., Nendel, C., Öztürk, I., Perego, A., Rodríguez, A., Ruane, A.C., Ruget, F., Sanna, M., Semenov, M.A., Slawinski, C., Stratonovitch, P., Supit, I., Waha, K., Wang, E., Wu, L., Zhao, Z., Rötter, R.P., 2018. Classifying multi-model wheat yield impact response surfaces showing sensitivity to temperature and precipitation change. Agric. Syst. 159, 209–224. https://doi.org/10.1016/J.AGSY.2017.08.004

Gao, Y., Wallach, D., Liu, B., Dingkuhn, M., Boote, K.J., Singh, U., Asseng, S., Kahveci, T., He, J., Zhang, R., Confalonieri, R., Hoogenboom, G., 2020. Comparison of three calibration methods for modeling rice phenology. Agric. For. Meteorol. 280, 107785. https://doi.org/10.1016/J.AGRFORMET.2019.107785

Harrison, M.T., Roggero, P.P., Zavattaro, L., 2019. Simple, efficient and robust techniques for automatic multi-objective function parameterisation: Case studies of local and global optimisation using APSIM. Environ. Model. Softw.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM – Evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

Jha, P.K., Ines, A.V.M., Han, E., Cruz, R., Vara Prasad, P.V., 2022. A comparison of multiple calibration and ensembling methods for estimating genetic coefficients of CERES-Rice to simulate phenology and

yields. F. Crop. Res. 284, 108560. https://doi.org/10.1016/J.FCR.2022.108560

Kawakita, S., Inaba, S., Takahashi, T., Kamada, E., Ishikawa, N., Takahashi, H., Okuno, R., 2019. Evaluation of non-linear wheat development models and optimization methods for their parameter determination. shiJournal Agric. Meteorol. 75, 120–128. https://doi.org/10.2480/agrmet.D-18-00034

Knutti, R., 2010. The end of model democracy? Clim. Change 102, 395–404. https://doi.org/10.1007/s10584-010-9800-2

Liu, Y., Bachofen, C., Wittwer, R., Silva Duarte, G., Sun, Q., Klaus, V.H., Buchmann, N., 2022. Using PhenoCams to track crop phenology and explain the effects of different cropping systems on yield. Agric. Syst. 195, 103306. https://doi.org/10.1016/J.AGSY.2021.103306

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., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., Asseng, S., Ewert, F., Cammarano, D., Maiorano, A., Aggarwal, P.K., Anothai, J., Basso, B., Biernath, C., Challinor, A.J., De Sanctis, G., Doltra, J., Dumont, B., Fereres, E., Garcia-Vila, M., Gayler, S., Hoogenboom, G., Hunt, L.A., Izaurralde, R.C., Jabloun, M., Jones, C.D., Kassie, B.T., Kersebaum, K.C., Koehler, A.-K., Müller, C., Kumar, S.N., Liu, B., Lobell, D.B., Nendel, C., O'Leary, G., Olesen, J.E., Palosuo, T., Priesack, E., Rezaei, E.E., Ripoche, D., Rötter, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit,

I., Tao, F., Thorburn, P., Waha, K., Wang, E., Wolf, J., Zhao, Z., Zhu, Y., 2018. The Hot Serial Cereal Experiment for modeling wheat response to temperature: field experiments and AgMIP-Wheat multi-model simulations. Open Data J. Agric. Res. 4, 28–34. https://doi.org/10.18174/odjar.v4i0.15830

Meier, U., 1997. Growth stages of mono- and dicotyledonous plants: BBCH-Monograph [WWW Document]. URL 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 1.9.19).

Muleke, A., Harrison, M.T., de Voil, P., Hunt, I., Liu, K., Yanotti, M., Eisner, R., 2022. Earlier crop flowering caused by global warming alleviated by irrigation. Environ. Res. Lett. 17, 044032. https://doi.org/10.1088/1748-9326/ac5a66

Müller, C., Franke, J., Jägermeyr, J., Ruane, A.C., Elliott, J., Moyer, E., Heinke, J., Falloon, P.D., Folberth, C., Francois, L., Hank, T., Izaurralde, R.C., Jacquemin, I., Liu, W., Olin, S., Pugh, T.A.M., Williams, K., Zabel, F., 2021. Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. Environ. Res. Lett. 16, 034040. https://doi.org/10.1088/1748-9326/abd8fc

Parmesan, C., 2006. Ecological and Evolutionary Responses to Recent Climate Change. Annu. Rev. Ecol. Evol. Syst. 37, 637–669. https://doi.org/10.1146/annurev.ecolsys.37.091305.110100

R Core Team, 2017. A language and Environment for Statistical Computing.

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

Rötter, R.P., Carter, T.R., Olesen, J.E., Porter, J.R., 2011. Crop–climate models need an overhaul. Nat. Clim. Chang. 1, 175–177. https://doi.org/10.1038/nclimate1152

Sanderson, B.M., Knutti, R., Caldwell, P., Sanderson, B.M., Knutti, R., Caldwell, P., 2015. Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. J. Clim. 28, 5150–5170. https://doi.org/10.1175/JCLI-D-14-00361.1

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.-S., Moore, A., Myrgiotis, V., Pattey, E., Rolinski, S., Sharp, J., Skiba, U., Smith, W., Wu, L., Zhang, Q., Bellocchi, G., 2020. Ensemble modelling of carbon fluxes in grasslands and croplands. F. Crop. Res. 252, 107791. https://doi.org/10.1016/j.fcr.2020.107791

Scheffé, H., 1959. The analysis of variance. John Wiley & Sons, 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 Ittersum, M.., Leffelaar, P.., van Keulen, H., Kropff, M.., Bastiaans, L., Goudriaan, J., 2003. On approaches and applications of the Wageningen crop models. Eur. J. Agron. 18, 201–234. https://doi.org/10.1016/S1161-0301(02)00106-5

Wallach, D., Mearns, L.O., Ruane, A.C., Rötter, R.P., Asseng, S., 2016. Lessons from climate modeling on the design and use of ensembles for crop modeling. Clim. Change 139, 551–564. https://doi.org/10.1007/s10584-016-1803-1

Wallach, D., Nissanka, S.P., Karunaratne, A.S., Weerakoon, W.M.W., Thorburn, P.J., Boote, K.J., Jones, J.W., 2017. Accounting for both parameter and model structure uncertainty in crop model predictions of phenology: A case study on rice. Eur. J. Agron. 88. https://doi.org/10.1016/j.eja.2016.05.013

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., 2023. Proposal and extensive test of a calibration protocol for crop phenology models. Agron. Sustain. Dev. 43, 46. https://doi.org/10.1007/s13593-023-00900-0

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

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

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