bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The chaos in calibrating crop models

Wallach[1*], Daniel; Palosuo[2], Taru; Thorburn[3], Peter; Hochman[3], Zvi; Gourdain[4], Emmanuelle; Andrianasolo[4], Fety; Asseng[5], Senthold; Basso[6], Bruno; Buis[7], Samuel; Crout[8], Neil; Dibari[9], Camilla; Dumont[10], Benjamin; Ferrise[9], Roberto; Gaiser[11], Thomas; Garcia[6], Cecile; Gayler[12], Sebastian; Ghahramani[13], Afshin; Hiremath[14], Santosh; Hoek[15], Steven; Horan[3], Heidi; Hoogenboom[5,16], Gerrit; Huang[17], Mingxia; Jabloun[8], Mohamed; Jansson[18], Per-Erik; Jing[19], Qi; Justes[20], Eric; Kersebaum[21,22], Kurt Christian; Klosterhalfen[23], Anne; Launay[24], Marie; Lewan[25], Elisabet; Luo[26], Qunying; Maestrini[15], Bernardo; Mielenz[27], Henrike; Moriondo[28], Marco; Nariman Zadeh[14], Hasti; Padovan[9], Gloria; Olesen[29], Jørgen Eivind; Poyda[30], Arne; Priesack[31], Eckart; Pullens[29], Johannes Wilhelmus Maria; Qian[19], Budong; Schütze[32], Niels; Shelia[5,16], Vakhtang; Souissi[33,34], Amir; Specka[21], Xenia; Srivastava[11], Amit Kumar; Stella[21], Tommaso; Streck[12], Thilo; Trombi[9], Giacomo; Wallor[21], Evelyn; Wang[17], Jing; Weber[12], Tobias, K.D.; Weihermüller[23], Lutz; de Wit[15], Allard; Wöhling[32,35], Thomas; Xiao[5,36], Liujun; Zhao[5], Chuang; Zhu[36], Yan; Seidel, Sabine J.[11]

[1]INRAE, UMR AGIR, Castanet Tolosan, France. ORCID 0000-0003-3500-8179
[2]Natural Resources Institute Finland (Luke), Helsinki, Finland
[3]CSIRO Agriculture and Food, Brisbane, Queensland, Australia
[4]ARVALIS - Institut du végétal Paris, France
[5]Agricultural and Biological Engineering Department, University of Florida, Gainesville, Florida
[6]Department of Earth and Environmental Sciences, Michigan State University, East Lansing, Michigan
[7]INRAE, UMR 1114 EMMAH, Avignon, France
[8]School of Biosciences, University of Nottingham, Loughborough, UK
[9]Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Italy
[10]Plant Sciences & TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium
[11]Institute of Crop Science and Resource Conservation, University of Bonn, Germany
[12]Institute of Soil Science and Land Evaluation, Biogeophysics, University of Hohenheim, Stuttgart, Germany
[13]Centre for Sustainable Agricultural Systems, Institute for Life Sciences and the Environment, University of Southern Queensland, Toowoomba, Queensland, Australia
[14]Aalto University School of Science, Espoo, Finland
[15] Wageningen University & Research, Wageningen, The Netherlands
[16]Institute for Sustainable Food Systems, University of Florida, Gainesville, Florida
[17]College of Resources and Environmental Sciences, China Agricultural University, Beijing, China
[18]Royal Institute of Technology (KTH), Stockholm, Sweden
[19]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada
[20]CIRAD, UMR SYSTEM, Montpellier, France

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21]Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany
[22]Global Change Research Institute CAS, Brno, Czech Republic
[23]Institute of Bio- and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
[24]INRAE, US 1116 AgroClim, Avignon, France
[25]Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[26]Hillridge Technology Pty Ltd, Sydney, Australia
[27]Institute for Crop and Soil Science, Federal Research Centre for cultivated Plants, Julius Kühn-Institut (JKI), Braunschweig, Germany
[28]CNR-IBE, Firenze, Italy
[29]Department of Agroecology, Aarhus University, Tjele, Denmark
[30]Grass and Forage Science / Organic Agriculture, Institute of Crop Science and Plant Breeding, Kiel University, Kiel, Germany
[31]Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany
[32]Institute of Hydrology and Meteorology, Chair of Hydrology, Technische Universität Dresden, Dresden, Germany
[33]National Institute of Agronomic Research of Tunisia (INRAT), Agronomy Laboratory, University of Carthage, Tunis, Tunisia
[34]National Agronomy Institute of Tunisia (INAT), University of Carthage, Tunis, Tunisia
[35]Lincoln Agritech Ltd., Hamilton, New Zealand
[36]National Engineering and Technology Center for Information Agriculture, Jiangsu Key Laboratory for Information Agriculture, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing, Jiangsu, China
[*]Corresponding author. Email: Daniel.Wallach@inrae.fr

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Highlights

- We documented calibration procedures in two multi- model studies
- Groups differ in criteria for best parameters, parameters to estimate and software
- There are important differences even between groups using the same model structure

Abstract

Calibration, the estimation of model parameters based on fitting the model to experimental data, is among the first steps in essentially every application of crop models and process models in other fields and has an important impact on simulated values. The goal of this study is to develop a comprehensive list of the decisions involved in calibration and to identify the range of choices made in practice, as groundwork for developing guidelines for crop model calibration starting with phenology. Three groups of decisions are identified; the criterion for choosing the parameter values, the choice of parameters to estimate and numerical aspects of parameter estimation. It is found that in practice there is a large diversity of choices for every decision, even among modeling groups using the same model structure. These findings are relevant to process models in other fields.

Keywords: calibration; modelling practice; parameter estimation, crop model, phenology

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

Crop models are a set of mathematical equations that describe the interactions between the atmosphere, the crop, soil conditions, and human management on growth, development and yield of crop plants. They are widely used to study, understand and optimize crop production in current and future environments e.g.(Ewert et al., 2015; Keating and Thorburn, 2018; Tsuji et al., 1998). Here we are particularly interested in the simulation of crop phenology i.e. the cycle of biological events in plants. The simulation of crop phenology is an essential part of crop models. Matching the phenology of crop varieties to the climate in which they grow is a critical crop production strategy (Hunt et al., 2019; Rezaei et al., 2018, 2015).

Calibration is the practice of adjusting model parameters to reduce the error between the model results and the measured data. Essentially every study involving crop models involves some type of calibration prior to model application. Calibration is also ubiquitous for mechanistic models in other fields, because parameter values are not universally valid, as explained by Fath and Jorgensen (2011) in the context of ecological models, and as explained in the context of crop models, based on statistical principles (Wallach, 2011). Model calibration of nonlinear models is a major area of study in statistics (see for example, Seber and Wild, 1989; Sen and Srivastava, 1990), but crop models in common with system models in other fields have several features which make calibration particularly challenging. Firstly, crop models often have a large number of parameters, for example the DAYCENT model has over 1000 parameters (Del Grosso et al., 2011), often many more than the number of observed data. This disequilibrium often exists, though less extreme, if one considers just the parameters that determine phenology and the observations of phenological stages. Even when one estimates a subset of the model parameters, there is often a problem of equifinality, meaning that various different combinations of parameter values can give the same results, and so calibration does not lead to unique parameter values (Beven and Freer, 2001). Also, mechanistic models usually have multiple different outputs which can be compared with observed data (for example, time to different phenological stages and final yield and leaf area at various times for crop models, or times to different phenological stages if one focusses only on the phenology component of crop models). There is then the problem of combining these different types of information into a single criterion for calibration. Furthermore, crop

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

models are often not continuous functions of the parameters. Crop models often have a daily time step, and phenology outputs are thus often integers (days after sowing to some development stage) rather than continuous variables. As a result, many mathematical methods for parameter estimation are not applicable. Finally, software is an additional problem. System models are usually embodied in software code, and it is a major challenge to recode the model, or even just the phenology component of the model, specifically for calibration (Gao et al., 2020). Rather, usually one 'externally' couples the existing model software to calibration software (Buis et al., 2015; He et al., 2010; Hunt et al., 1993), but this can also require substantial effort. As a result, calibration for crop models is often simply done by manual trial and error without using an automated routine.

In response to these and other difficulties, there have been numerous studies published concerning calibration of crop models. One type of study is model specific; it identifies the most important parameters in that model, and explains how they can be estimated from data (Ahuja et al., 2011). Other studies have focused on the implementation of a Bayesian approach or on the comparison of frequentist and Bayesian approaches (Gao et al., 2020; Jansen and Hagenaars, 2004; Sexton et al., 2016), on numerical methods of seeking best parameter values (Bhar et al., 2020), on the choice of parameters to estimate (Angulo et al., 2013), or on the observed data to use for calibration (Guillaume et al., 2011; Hoogenboom et al., 2012).

It is not clear, however, how variable crop model calibration is in practice. What are the decisions that model users need to make concerning calibration, and what is the range of choices that one finds in practice? As a baseline for comparison for new calibration studies, and in view toward improving calibration practices in the crop modeling community, this is an important first step. One must first comprehensively define the problem, and identify current practices. This has not yet been done. In a survey of crop model calibration approaches Seidel et al. (2018) found a very large diversity of approaches. However, in that study the variability in calibration approach was due to differences in data and in objectives of the calibration as well as to differences between modeling groups. A different possible study would be a detailed analysis of the crop modeling literature, focusing on the explanation of the calibration approach. However, this would again mix results from different

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

contexts (models, data, objectives, tools). Furthermore, the description of calibration methods in the literature is often very succinct, so many details are not available in published papers.

We chose, therefore, a different approach. We organized two relatively large multi-model ensemble studies for prediction of wheat phenology, and in each study asked all modeling groups to use their "usual" calibration approach and to describe this approach in detail. This raw material allowed us to examine calibration practices of a large number of crop modeling groups, all given the same data and tasked with the same predictions. All groups ran their full crop model, and so used calibration methods applicable to those models. However, only phenology data was provided for calibration, so this is a simplified example as to the types of data available for calibration. In these multi-model studies, there were several instances where the same model structure (i.e. model equations) were used by multiple modeling groups. This, therefore, also allowed us to examine both inter- and intra-model structure variability in calibration approach. Additionally, the two multi-model studies had quite different types of observed data for calibration; dates of just two phenological stages in one case, dates of multiple stages in the other. This allowed us to insure, at least to some extent, that our results are not specific to a particular dataset.

The purpose of this study is to shed new light on the general problem of calibration of crop models. We do not intend to compare specific methods, or give detailed recommendations. Rather, the goal here is to take a step back, in order to better define the problem. The first major question that we address is, what exactly are the decisions that one must make when calibrating a crop model? Some of these decisions have been well-studied, such as the choice of algorithm and software for calibration. However, there are many other decisions that have been much less studied. In any case, there does not seem to be a comprehensive list of calibration decisions. We specifically consider the case where only phenology data is available for calibration, but for a large part the same decisions arise for crop models with other observed data and for mechanistic models in other fields. The second question considered here is, what is the range of choices that modelers actually make with respect to the decisions involved in calibration? Once again, though the results here are specific to calibration based on phenology data, they are expected to be relevant also to crop and other mechanistic models.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Materials and Methods

This study is based on two multi-modeling group simulation studies. In those two studies the modeling groups were provided with phenology data representative of wheat crops grown under current climate and management for two common varieties in France (first study) or for one common variety in Australia (second study) (Table 1). The groups were given data from a subset of environments for calibration, and then were asked to simulate phenology for other "evaluation" environments from the same target population. The prediction accuracy was determined from the comparison of simulated with observed values for the evaluation environments. At no point did the participants have access to the evaluation data. More details of these studies, including the prediction errors for each modeling group in each study are given in Wallach et al. (2020b, 2020a).

The French dataset included observed values for days from sowing to two phenological stages, namely beginning of stem elongation (growth stage 30 on the BBCH and Zadoks scales (Lancashire et al., 1991; Zadoks et al., 1974)) and middle of heading (growth stage 55 on the BBCH and Zadoks scales). These two stages are of practical importance because they can easily be determined visually and are closely related to the recommended dates for the second and third N fertilizer applications. Observed data were available for two varieties, namely Apache and Bermude. In all cases, the modeling groups used the same calibration approach for both varieties; therefore, in this study we only report a single calibration approach for each modeling group. The Australian dataset resulted from measurements of Zadoks growth stage (Zadoks et al., 1974), about every two weeks in every plot. The data were interpolated, to give days from sowing to every integer Zadoks stage from the first to the last observed stage, and these were the data provided to each modeling group. The required simulations, for the evaluation environments, were days from sowing to stage Z30 (Zadoks stage 30, pseudostem, i.e. youngest leaf sheath erection), to stage Z65 (Zadoks stage 65, anthesis half-way, i.e. anthers occurring half way to tip and base of ear), and to stage Z90 (Zadoks stage 90, grain hard, difficult to divide). These stages are often used for management decisions or to characterize phenology.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Calibration data		**Evaluation data**	
Dataset	**Number of environments**	**Observed or interpolated phenological stages**	**Number of environments**	**Observed or interpolated phenological stages**
French (repeated for varieties Apache and Bermude)	14	BBCH30, BBCH55	8	BBCH30, BBCH55
Australian (variety Janz)	24	Each integral Zadoks stage from first to last observed Zadoks stage	18	Z30, Z65, Z90

Table 1.

Description of datasets. The French dataset was repeated for two varieties, Apache and Bermude. Data are days from sowing to indicated stages on BBCH or Zadoks scale.

In both simulation exercises, each participating modeling group was asked to calibrate the model in their "usual" way, using the calibration data provided. Each group was also asked to complete a questionnaire, detailing how the calibration was conducted (Supplementary Table S2). The results presented in this study are essentially a synthesis and analysis of those questionnaires.

We use here the term "model structure" to designate a specific set of equations. The model structures used by the participants are listed in Supplementary Table S1. We speak of

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

modeling group to designate the group of researchers that implemented the model in a specific case. The modeling group was responsible for fixing the values of those parameters not determined by calibration, and for determining all aspects of the calibration procedure. There were 27 participating modeling groups in the study using the French dataset, and 28 in the study using the Australian dataset with 25 groups participating in both studies. The modeling groups are simply identified as M1-M29 and the same identifier is used for the same modeling group working with the French and Australian data. The name of the model structure used by each group is not given, since this might give the erroneous impression that the calibration approach and simulation results are specific to that model structure, while in fact the they depend on both the model structure and the modeling group. Three of the model structures, noted S1, S2 and S3, were used by multiple groups. Structure S1 was used by four (French dataset) or three (Australian dataset) groups, structure 2 by three groups and structure S3 by two groups. This provides information about the differences between modeling groups calibrating the same model structure.

For details as to how each model structure simulates phenology, see the references for each structure (Supplementary Table S1). Here we give only a short overview. The basic output of phenology simulation is days from sowing to various phenological stages or between various phenological stages. The number and identity of simulated stages varies with the model, and may include both observable stages (for example anthesis) and stages that are model constructs (for example, start of linear phase of grain filling). The most important inputs that determine spring wheat phenology are daily temperature and photoperiod (Aslam et al., 2017), while for winter wheat it is also important to include the process of vernalization, i.e. the effect of low winter temperatures on development (Li et al., 2013). Most model structures take into account all three factors, though not all factors affect the development rate at all stages of development. Some structures take into account only temperature, or temperature and either photoperiod or vernalization. Most model structures take temperature into account by calculating thermal time, calculated most simply as the sum of the daily temperature above some threshold temperature (a parameter). In other models the daily contribution to degree days may have a plateau above some optimal temperature (a parameter), or decline above the optimum temperature at some rate (a parameter) or be some more complex function of temperature (Kumudini et al., 2014; Wang et al., 2017). The parameters of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

temperature response curve may differ at different stages in the development cycle. Wheat is a long-day plant, which flowers earlier in longer days. Often phenology response to photoperiod is modeled using two parameters, a threshold photoperiod below which development rate increases with increasing photoperiod and a sensitivity coefficient, which describes the rate of increase, though other functions, with other parameterizations, are also used. Vernalization is described as a period of low temperatures, which must be experienced before the plant can flower. Vernalization parameters can include the upper limit for temperature to count as vernalizing, and the required number of vernalizing days. To determine the day of occurrence of a given stage, most models have an internal counter of physiological time, which often is based on degree days modulated by photoperiod. When this counter attains some predetermined value, specific to the stage in question, which is a model parameter, that stage is deemed to have been attained. Some models also relate development to the rate of leaf appearance (called the phyllochron, a parameter) or rate of tillering. Finally, several models also account for cold or drought stress in the simulation of the development rate. Note, that if the development rate depends on drought stress, then it is sensitive to all the parameters in the model that determine soil water content and the soil-plant water dynamics.

3 Results

The major results of this study are firstly a comprehensive list of the decisions required for crop model calibration given phenology data, and secondly information about the choices made by a fairly large number of modeling groups. It is convenient to divide the required decisions into three groups; decisions related to the criterion that defines the best parameter values, decisions related to the choice of parameters to be estimated and decisions related to the. numerical calculation of the best parameter values. The more detailed decisions within each group of decisions are shown in **Error! Reference source not found.**, which also shows the choices made by the participating modeling groups and the number of groups that made each choice. Details of the choices made by each individual modeling group are shown in**Error! Reference source not found.** Supplementary Tables S3, S4, and S5 for the criterion of best parameter values, for the choice of parameters to estimate and for the choices of algorithm and software, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Category	**Decision**	**Choices made by modeling groups** [1]
Criterion for best parameter values	Variables to include in criterion	Only variables to be predicted (27,12) Additional variables (0,16)
	Units for measuring error	Days (27,26) Development stage (0,2)
	Frequentist or Bayesian criterion	Frequentist (25,24) Bayesian (2,4)
	Objective function (for frequentist)	Sum of squared errors (18, 14) Sum of absolute errors (2,2) Other (1,4) No single objective (4,4)
	Simultaneous or sequential calibration (for frequentist)	Simultaneous (23, 20) Sequential (2,4)
	Likelihood (for Bayesian)	Normal likelihood (1,2) Student's t distribution (0,1) Concentrated likelihood (1,1)
	Priors (for Bayesian)	Uniform (1,3) Normal distribution (0,1) Mixture of uniform and normal (1,0)
Parameters to estimate	Which parameters	Parameters related to thermal time between phenological stages (23,26) Parameters related to vernalization (16 ,15) Parameters related to photoperiod

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		response (13 ,13) Parameters related to temperature response curve (6,8) Parameters related to phyllochron or tillering (6 ,4) Parameters related to water stress or cold stress (2,2) Parameters related to emergence (2,3)
	Number of parameters	French dataset 1-9 (median=3.0) Australian dataset 2-10 (median=4.5)
	Basis for choosing parameters to estimate	Expert knowledge (18,20) Expert knowledge and data based (4,4) Sensitivity analysis (5,4)
Numerical methods	Algorithm	Gradient-free search algorithm (6,6) Gradient based (5,4) Trial and error (12,12) MCMC (4,6)
	software	See text and Supplementary Table S5

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. The numbers in parentheses are the numbers of groups that made the choice for the French dataset (out of 27 groups) and for the Australian dataset (out of 28 groups), respectively

Table 2.

Decisions required for crop model calibration, and the choices made by the various modeling groups

3.1 Criterion for best parameters

A first calibration decision in this category is which variables to use in the criterion that defines the best parameter values, and in particular whether to use only those variables for which predictions are sought, or additional observed variables. The French dataset only had observations for two phenological stages (BBCH30 and BBCH55, see Table 1), and simulations were required for these two stages. Thus, it was natural that for almost all groups, the criterion of best parameters included observations of both those variables. Two groups (M9 and M18) used model structures that did not simulate days to stage BBCH30, so these groups only used a subset of the observed variables (i.e. the observations of days to stage BBCH55) in the criterion defining best parameters. The Australian dataset on the other hand had many observed variables for which simulation was not required, so here the choice of variables to include in the criterion was not so straightforward. For the Australian dataset about 40% of the groups used only the variables to be simulated (days to Z30, Z65, and Z90 or a subset if the model structure didn't simulate all those variables,) while about 60% of the modeling groups included other observed variables in the criterion of best parameters. Different groups that used the same model structure did not necessarily all make the same choice here. Consider structure S1, used by groups M2, M3, and M4, and the Australian dataset. All three groups used minimum sum of squared errors as the criterion defining the best parameter values, However, group M2 included additional variables in addition to the variables to be simulated in their sum of squared errors, group M3 used only squared errors

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for simulating Z65 and Z90, and only group M4 used exactly the same variables to be simulated in their criterion, i.e. Z30, Z65, and Z90.

A second decision concerns the definition of error. Almost all groups expressed error in terms of days to reach a specified stage (Table S3). However, it is also possible to express error in terms of phenological stage. In the simplest case, suppose that a model structure calculates Zadoks stage each day (i.e. the internal counter each day is directly or is translated into a value for Zadoks stage). Suppose that for a particular environment it is observed that stage Z30 is attained on day 45, but the simulated day is 40. The error in days is 5 days. Suppose that the simulated stage on day 45 is Z33.4. Then the error in terms of development stage is 30-33.4=-3.4. For the French dataset all groups calculated error in days, but for the Australian dataset three groups expressed error in terms of development stage rather than days.

A third decision is whether to use a frequentist or Bayesian perspective. If a frequentist perspective is chosen, one must then define the mathematical form of the objective function. If a Bayesian perspective is chosen, one must define the form of the likelihood and the prior distributions for the parameters. The large majority of groups adopted a frequentist approach, where the estimated parameter values are those values that minimize some measure of error between the simulated and observed values (Table S3). Most of the frequentist groups sought to minimize the sum of squared errors, where the sum is over calibration environments and over all variables included in the criterion. This is the ordinary least squares (OLS) criterion. One (French data) or four (Australian data) groups used a different measure of distance between observed and simulated values, namely the sum of root mean squared errors for the different variables, or a weighted sum of squared errors. Two groups chose to minimize the sum of absolute errors. This is the least absolute value criterion (LAV). Four groups for the French dataset, and the same four groups for the Australian dataset, did not define an explicit objective function to be minimized, but rather sought parameter values to give a “best fit” to the data, where “best fit” was determined visually or by some subjective combination of mean squared error, R^2, or other fit metrics. Groups using the same model structure did not make the same decisions. For example, among the three groups that used model structure S2, for the Australian dataset, one used the OLS criterion and two had no explicit objective function.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Another decision for the frequentist perspective is whether to fit all the observed variables in a single calculation step or to use multiple steps, adjusting parameters to different variables in each step. Almost all groups estimated all parameters simultaneously (Table S3). However, two (French data) or four (Australian data) groups estimated parameters in more than one step, fitting for example three parameters to the BBCH30 data in the French dataset, and then fixing those parameters at their estimated values and fitting another parameter to the BBCH55 data. Groups using the same model structure did not make the same decisions. For example, among the two groups that used model structure S3, for the Australian dataset, group M23 estimated all parameters simultaneously while group M24 estimated parameters in two steps.

Of the total of six Bayesian calibrations (two for the French dataset, four for the Australian dataset), three assumed a normal distribution of errors and one a Student's *t* distribution. One group worked with the concentrated likelihood, which replaces the model variance for each variable by its maximum likelihood value. For the Bayesian groups, parameters were assumed to have either uniform of truncated normal prior distributions.

No group took correlations of errors for different variables in the same environment into account. That is, all groups treated all the errors as though they were independent. Only two groups (M19, M21, see Table S3) took into account the possibility of different variances for different variables, M21 by using the method of concentrated likelihood, (Seber and Wild, 1989) and M19 by dividing the standard deviation of the error of each variable by the number of observations of that variable.

3.2 Choice of parameters to estimate

Each model structure is parameterized differently, so it is not possible to directly compare names of parameters between model structures. It is, however, possible to identify the role of estimated parameters in the model and base the comparison between groups on that. Details related to the choice of parameters by each group are given in Supplementary Table S4.

Most groups estimated some parameters that concern the physiological time required to attain one or more phenological stages. Fifteen groups for each of the datasets estimated one or more parameters related to vernalization, and 13 groups for both of the datasets estimated one or more parameters related to photoperiod sensitivity. A smaller number of groups estimated

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parameters related to the temperature response function (for example minimal temperature or optimal temperature for development) or to tillering or leaf appearance rate (phyllochron). Two groups for each dataset estimated parameters related to the effect of stress on the development rate, and two (French dataset) or three (Australian dataset) groups estimated parameters related to time to emergence.

The number of estimated parameters ranged from one to nine for the French dataset and from two to ten for the Australian dataset. In most cases, the choice of parameters to estimate was based on expert opinion, but four groups for each dataset combined expert opinion with data-based information (for example, testing various combinations of parameters to see which gives the best fit). Five (French dataset) or four (Australian dataset) groups based the choice of parameters to estimate on sensitivity analysis. We have separated the categories of expert knowledge and data-based choice of parameters to estimate, but it should be noted that in fact, expert knowledge also adapts, at least to some extent, the choice of parameters to estimate to the dataset. Thus, almost every group that based the choice of parameters on expert knowledge estimated a different (and in most cases larger) set of parameters based on the Australian dataset, with more observed variables, than based on the French dataset.

There were important differences even between groups using the same model structure. Consider for example structure S2. The three groups that used this model structure (M7, M12, and M13) estimated respectively four, three, and two parameters for the French dataset and nine, four, and two parameters for the Australian dataset. Two of the groups based the choice on expert opinion, while group M7 made a partially data-driven choice.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.3 Numerical methods

Two basic decisions here are which algorithm to use for estimating the parameters and what software to implement that algorithm. The choices made by each modeling group are shown in Supplementary Table S5, and information about the specific software used is given in Table S6. Among the groups that chose a frequentist approach, slightly over half used trial and error to search for the optimal parameter values. In some of those cases, available software was used as an aid, but the final values were found by simply trying different parameter values. The remaining half was split between groups that used a derivative-free search algorithm, usually an algorithm designed to find a global optimum, and those that used a gradient-based algorithm. Many different software solutions were used, including multi-purpose software packages as well as software written expressly for calibration of that particular model structure. Four groups (French data) or six groups (Australian data) used a Markov-Chain Monte-Carlo (MCMC) algorithm to estimate the posterior distribution, using various software packages. That included one group that used an MCMC algorithm though the objective was to minimize the sum of absolute errors. The groups that used the same model structure, in general, did not use the same algorithm and software. For example, considering the two groups using model structure S3, group M23 used a combination of global and local search algorithms and available software packages, while group M24 used trial and error and no software packages.

4 Discussion

There are multiple decisions to make in crop model calibration, as shown in **Error! Reference source not found.**. We have split these decisions into three main groups: firstly decisions related to the criteria for defining the best model parameters, secondly the choice of which parameters to estimate, and thirdly the algorithm and software used to find the best parameter values. In each category, there are multiple decisions that together define the calibration approach. For all those decisions, there is substantial variability between modeling groups, even though all have the same calibration data and the same predictions to make. There are multiple differences even between groups that use the same model structure. Thus,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a first overall conclusion is that we are far from having a consensus on how to calibrate crop models, even for a given model structure and dataset.

4.1 Criteria for best parameters

A major decision is the observed variables (here observed development stages) to include in the criterion of best fit. From a modeling point of view, using as many variables as possible for fitting the model reduces the risk of "getting the right answer for the wrong reason", i.e. getting a good fit for some variables while other variables, that describe other aspects of system behavior, are poorly simulated (Wang et al., 2011). Fitting the model to more variables will reduce the aspects of the system that could unknowingly be poorly simulated. This would be important for all applications of crop models. For example, process-based crop simulation models are argued to be meaningful tools for understanding crop growth and production in response to climate variability and change (Keating and Thorburn, 2018), particularly as they cover interconnections of different system variables in their structures (Ewert et al. 2015). Calibration using multiple observed variables should improve the representation of these interconnections. From a statistical point of view, more data in general leads to predictors with smaller variance, which argues for using all the available data. However, this assumes that the model is correctly specified, in the statistical sense meaning that model errors have expectation zero for all values of the explanatory variables. It has been argued, that crop models are most likely statistically incorrectly specified, and as a result the best parameters for predicting one variable may be different than the best parameters for predicting a different variable (Wallach, 2011). In that case, using additional variables in the objective function may degrade predictive accuracy for the variables of primary interest. This was found to be the case in the study of Guillaume et al. (2011). If, however, one is willing to assume that statistical misspecification is not too extreme, then it seems worthwhile to include as many of the observed variables as possible in the objective function.

Most groups defined error as the difference between the simulated and observed days to reach a given phenological stage, but in a few cases error was defined as the difference between the simulated development stage and the observed stage, on the day of observation. Defining error in terms of development stage is specific to phenology data, but could still be used for phenology even if there are also other types of observations, such as yield. This option

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

requires that the model include some internal counter whose observed and simulated values at observed phenological stages are known, but this is often the case. It has been argued, that the problem of minimizing errors is much better behaved numerically when errors are in terms of development stage rather than days (Wallach et al., 2018). In practical terms, this might make feasible the use of derivative based search algorithms that do not converge when errors are in terms of days.

For the groups that adopted a frequentist perspective, the large majority framed the problem as an ordinary least squares (OLS) problem. Two modeling groups chose parameters to minimize the sum of absolute errors, which has been argued to have advantages over OLS, in that it is less sensitive to outliers (Willmott and Matsuura, 2005). Four groups did not have an explicit objective function. One obvious disadvantage of this approach is its subjectivity, adding uncertainty in the definition of best-fit to other uncertainties in calibration. A second disadvantage is that one cannot automate the search for the best parameters.

In most cases, a single objective function, combining all errors, was used. In a few cases however, parameters were fit sequentially (first to one variable then to the next etc.). This sequential technique has often been recommended for full crop models (Anothai et al., 2008; Ahuja et al., 2011). This simplifies the mechanics of finding the best parameter values, but it will lead to sub-optimal results with respect to an overall objective function. If the objective is to minimize the total sum of squared errors, for example, the best parameter values are those that minimize exactly that objective function.

A few groups chose a Bayesian rather than a frequentist perspective. There are fundamental differences between frequentist and Bayesian approaches (Berger and Bayarri, 2004). However, for the practical prediction problem here, there are also important similarities. A major difference is that the Bayesians approach focuses on the posterior distribution, which is a distribution of predicted values, while the frequentist approach focuses on point predictions, i.e. one single predicted value. Here, however, all groups were asked for point predictions, so the groups that used Bayesian approach had to choose a single result from the posterior distribution. In all cases, they chose the parameter values that maximized the posterior distribution, which then plays the same role as the objective function for the frequentist approach. Another important difference is that for the Bayesian approach the prior

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

information about values of the parameters is included in the calculation, while this is not in general included in a frequentist approach. However, the frequentist approach here, in almost all cases, included lower and upper bounds on the parameter values (table S5), which is also prior information. In fact, a Bayesian approach with normal likelihood and uniform priors leads to exactly the same criterion for best fit, namely minimum squared error subject to the constraints on the parameters, as OLS.

In almost all cases the calibration approach was directly based on regression methods in statistics, either frequentist or Bayesian. This seems logical, insofar as these statistical methods have desirable properties. However, these properties in general require that certain assumptions be satisfied. The standard assumptions for the OLS method are that the model errors be independent and identically distributed, with expectation 0 (Seber and Wild, 1989; Sen and Srivastava, 1990). For the Bayesian methods, one must make explicit assumptions about the distribution of errors, including whether all errors have the same distribution and whether errors for different variables are correlated. In the case of crop models, with multiple observed variables in each environment, the assumptions of independent identically distributed errors with expectation 0 are not likely to be satisfied (Wallach et al., 2019). Most obviously, errors for different variables in the same environment (e.g. days to development stages Z30 and Z65) may be correlated, since any particularities of the environment affect all variables for that environment. Also, the variances of errors for different variables may be different, in which case the assumption of identical distributions for all errors is violated. No group took correlations of errors into account. Two groups took into account the possibility of different variances for different variables, M21 by using the method of concentrated likelihood, (Seber and Wild, 1989) M19 by dividing the standard deviation of the error of each variable by the number of observations of that variable. In general, it would seem worthwhile to go a step further in applying statistical methods, beyond employing standard techniques, in order to examine whether the standard assumptions about model error are satisfied. To detect unacceptably large violations of the standard assumptions, it seems worthwhile to examine the model residuals (observed minus simulated values) after calibration, as is standard procedure in regression (see for example NIST/SEMATECH, 2013)). One could examine overall bias of model residuals, which should be zero, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variances of residuals for different variables, which should be similar, and correlations between residuals for different variables in the same environment, which should be small.

Only phenology data were available in the datasets here, and so all errors had the same units (days or phenological stage). If other data had been available, for example final yield or soil moisture content, with different units, it would be meaningless to simply combine errors. In that case, a first step could be to divide all simulated and observed values by an estimated standard deviation of error for that variable, as in weighted least squares (Seber and Wild, 1989). Then all errors would be unitless and could be combined. It would still be important to test residuals after calibration.

4.2 Choice of parameters to estimate

There is some agreement about the categories of parameters to estimate between groups using different model structures; for example, most groups estimated parameters related to physiological time required to achieve different phenological stages, but the detailed choices are quite different. A further indication of the diversity of choices is the range in the number of estimated parameters, i.e., one to nine for the French dataset and two to ten for the Australian dataset.

Of particular interest is the rationale behind the choice of parameters to estimate, and what this implies for the adaptation of the choice of parameters to the dataset. In most cases, the choice of parameters to estimate was based on “expert knowledge” of the model. To some extent, this takes into account the dataset. However, expert knowledge only takes the amount and type of observed data into account approximately. An alternative would be to formally consider the choice of parameters to estimate as a problem of model selection, where the selection is of the subset of parameters to estimate by calibration, while the other parameters retain their default values. For example, one could use the Akaike Information Criterion (AIC; Akaike (1973)), which has been widely used for model choice in ecology (Burnham et al., 2011), to choose the parameters to estimate. The use of a model selection rule would automatically adapt the choice of parameters to estimate to the calibration dataset. However, given the large number of possible parameters to estimate, it would probably be necessary to combine expert opinion, in order to choose a fairly small number of candidate parameters, with a formal model selection criterion.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

All parameters that were not estimated using the calibration data retained their default values, and this, in general, applies to the majority of model parameters. While some parameters will have no effect on the simulated values, many others will have an effect. It is clear then, that the choice of these default values is extremely important, and should reflect whatever information one has about the cultivars and environments of interest. The choice of default values probably merits more attention than it usually receives.

4.3 Algorithm and software

Somewhat over a third of groups used trial and error to search for the best-fit parameters. There are several disadvantages to this approach; it is time-consuming, it is likely to end in a non-optimal solution, especially with several parameters, and it cannot be replicated for example to estimate prediction error using cross-validation.

Among the groups that automated the search for the best-fit parameters or for the posterior distribution, there was no consensus on the software to use even among groups using the same model structure. The problem of choosing a calibration algorithm and software to search for optimal parameter values has received much attention in the field of hydrological modeling (Skahill and Doherty, 2006). Gradient based algorithms are, in general, very efficient, but may converge to a local rather than global optimum (Blasone et al., 2006). Furthermore, most crop models have multiple discontinuities when the outputs are considered as functions of the parameters, which may make gradient based algorithms unusable. Removing these discontinuities may be possible, but at the price of detailed intervention in the model code (Liu et al., 2018). Global search algorithms, such as a grid search or a genetic algorithm, may avoid converging to a local optimum but in general require many more executions of the model. A third possibility is a gradient-free search algorithm such as the simplex method (Nelder and Mead, 1965).

There is calibration software that has been developed specifically for some crop models (Buddhaboon et al., 2018; Buis et al., 2015; Hunt et al., 1993), and also some software that is designed to be easily coupled to any model (Doherty et al., 2010). Coupling parameter estimation software to a crop model is not simple and so modeling groups tend to use available software or even no software rather than developing new calibration software themselves. This implies that to improve calibration approaches it is not sufficient to propose

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

guidelines for good calibration practices. For the guidelines to be effective, they must include software solutions that can be used by any model.

5 Conclusions

The results here are based on calibration of crop models, given data on crop phenology, but are quite relevant for other types of observed data and for system models in other fields. Calibration of crop models involves multiple decisions, which can be grouped into choice of criteria for defining the best parameter values, choice of parameters to estimate and choice of algorithm and software. Different modeling groups make quite different decisions, even for modeling groups using the same model structure. It seems that we are far from having a consensus on how to calibrate crop models, even in the simple case with only phenology data.

We found that the choice of objective function is usually based on statistical methods for regression, but without testing whether the usual statistical assumptions are valid. It is suggested that this should be done. Many modeling groups search for the best parameter values by trial and error, which is unadvisable since it is laborious and may not lead to optimal parameter values. For those groups that automate the parameter estimation, algorithm and software are often based on existing software, which highlights the importance of providing calibration software with crop models. Arguably the most difficult decisions concern the choice of parameters to estimate. In most cases, this is based on expert opinion. It is suggested that it would be better to combine expert knowledge of the model and the modeled system with statistical methods of model selection. Guidelines for calibration of crop phenology models would be very helpful, but need to include software solutions to be of practical use. Overall, crop model calibration probably cannot be fully automatic and in particular requires understanding of the system and the model.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 AI-CropPro (316172) and DivCSA (316215) and Natural Resources Institute Finland (Luke) through a strategic project BoostIA, the BonaRes project "Soil3" (BOMA 03037514) 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 1387 under the Canadian Agricultural Partnership, 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 26/Nov/2015). The order in which the donors are listed is arbitrary.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Ahuja, L. (Lajpat), Ma, L., American Society of Agronomy., Crop Science Society of America., Soil Science Society of America., 2011. Methods of introducing system models into agricultural research. American Society of Agronomy.

Ahuja, L.R., Ma, L., (eds.), 2011. Methods of introducing system models into agricultural research. American Society of Agronomy.

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle, in: Petrov, B.N., Csaki, F. (Eds.), In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S., 2013. Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe. Agric. For.

Aslam, M.A., Ahmed, M., Stöckle, C.O., Higgins, S.S., Hassan, F. ul, Hayat, R., 2017. Can Growing Degree Days and Photoperiod Predict Spring Wheat Phenology? Front. Environ. Sci. 5, 57. https://doi.org/10.3389/fenvs.2017.00057

Berger, J.O., Bayarri, M.J., 2004. The Interplay of Bayesian and Frequentist Analysis. Stat. Sci. 19, 58–80. https://doi.org/10.1214/088342304000000116

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Bhar, A., Kumar, R., Qi, Z., Malone, R., 2020. Coordinate descent based agricultural model calibration and optimized input management. Comput. Electron. Agric. 172, 105353. https://doi.org/10.1016/J.COMPAG.2020.105353

Blasone, R.S., Madsen, H., Rosbjerg, D., 2006. Calibration and reliability in groundwater modelling, in: Bierkens, M.F.., Gehrels, J.C.(, Kovarik, K. (Eds.), Calibration and Reliability in Groundwater Modelling: From Uncertainty to Decision Making (Proceedings of

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ModelCARE'2005, The Hague, The Netherlands, June 2005). IAHS, p. 316.

Buddhaboon, C., Jintrawet, A., Hoogenboom, G., 2018. Methodology to estimate rice genetic coefficients for the CSM-CERES-Rice model using GENCALC and GLUE genetic coefficient estimators. J. Agric. Sci. 156, 482–492. https://doi.org/10.1017/S0021859618000527

Buis, S., Wallach, D., Guillaume, S., Varella, H., Lecharpentier, P., Launay, M., Guérif, M., Bergez, J.-E., Justes, E., 2015. The STICS Crop Model and Associated Software for Analysis, Parameterization, and Evaluation. John Wiley & Sons, Ltd, pp. 395–426. https://doi.org/10.2134/advagricsystmodel2.c14

Burnham, K.P., Anderson, D.R., Huyvaert, K.P., 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65, 23–35. https://doi.org/10.1007/s00265-010-1029-6

Del Grosso, S.J., Parton, W.J., Keough, C.A., Reyes-Fox, M., 2011. Special features of the DayCent modeling package and additional procedures for parameterization, calibration, validation, and applications, in: Ahuja, L.R., Ma, L. (Eds.), Methods of Introducing System Models into Agricultural Research. American Society of Agronomy, Madison, pp. 155–176.

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.

Ewert, F., Rötter, R.P., Bindi, M., Webber, H., Trnka, M., Kersebaum, K.C., Olesen, J.E., van Ittersum, M.K., Janssen, S., Rivington, M., Semenov, M.A., Wallach, D., Porter, J.R., Stewart, D., Verhagen, J., Gaiser, T., Palosuo, T., Tao, F., Nendel, C., Roggero, P.P., Bartosˇová, L., Asseng, S., 2015. Crop modelling for integrated assessment of risk to food production from climate change. Environ. Model. Softw. 72. https://doi.org/10.1016/j.envsoft.2014.12.003

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

He, J., Jones, J.W., Graham, W.D., Dukes, M.D., 2010. Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method. Agric. Syst. 103, 256–264. https://doi.org/10.1016/j.agsy.2010.01.006

Hoogenboom, G., Jones, J.W., Traore, P.C.S., Boote, K.J., 2012. Experiments and Data for Model Evaluation and Application BT - Improving Soil Fertility Recommendations in Africa using the Decision Support System for Agrotechnology Transfer (DSSAT), in: Kihara, J., Fatondji, D., Jones, J.W., Hoogenboom, G., Tabo, R., Bationo, A. (Eds.), . Springer Netherlands, Dordrecht, pp. 9–18. https://doi.org/10.1007/978-94-007-2960-5_2

Hunt, J.R., Lilley, J.M., Trevaskis, B., Flohr, B.M., Peake, A., Fletcher, A., Zwart, A.B., Gobbett, D., Kirkegaard, J.A., 2019. Early sowing systems can boost Australian wheat yields despite recent climate change. Nat. Clim. Chang. 9, 244–247. https://doi.org/10.1038/s41558-019-0417-9

Hunt, L.A., Pararajasingham, S., Jones, J.W., Hoogenboom, G., Imamura, D.T., Ogoshi, R.M., 1993. GENCALC: Software to Facilitate the Use of Crop Models for Analyzing Field Experiments. Agron. J. 85, 1090–1094.

Jansen, M.J.W., Hagenaars, T.J., 2004. Calibration in a Bayesian modelling framework, in: van Boekel, M.A.J.S., A., S., A.H.C., van B. (Eds.), Bayesian Statistics and Quality Modelling in the Agro-Food Production Chain . Kluwer Academic , Dordrecht.

Keating, B.A., Thorburn, P.J., 2018. Modelling crops and cropping systems—Evolving purpose, practice and prospects. Eur. J. Agron. 100, 163–176. https://doi.org/10.1016/j.eja.2018.04.007

Kumudini, S., Andrade, F.H., Boote, K.J., Brown, G.A., Dzotsi, K.A., Edmeades, G.O., Gocken, T., Goodwin, M., Halter, A.L., Hammer, G.L., Hatfield, J.L., Jones, J.W., Kemanian, A.R., Kim, S.-H., Kiniry, J., Lizaso, J.I., Nendel, C., Nielsen, R.L., Parent, B., St□ckle, C.O., Tardieu, F., Thomison, P.R., Timlin, D.J., Vyn, T.J., Wallach, D., Yang, H.S., Tollenaar, M., 2014. Predicting maize phenology: Intercomparison of functions for developmental response to temperature. Agron. J. 106, 2087–2097. https://doi.org/10.2134/agronj14.0200

Lancashire, P.D., Bleiholder, H., Boom, T. VAN DEN, LangelüddekeE, P., Stauss, R., Weber, E., Witzenberger, A., 1991. A uniform decimal code for growth stages of crops and weeds. Ann.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appl. Biol. 119, 561–601. https://doi.org/10.1111/j.1744-7348.1991.tb04895.x

Li, G., Yu, M., Fang, T., Cao, S., Carver, B.F., Yan, L., 2013. Vernalization requirement duration in winter wheat is controlled by TaVRN-A1 at the protein level. Plant J. 76, 742–53. https://doi.org/10.1111/tpj.12326

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

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

NIST/SEMATECH, 2013. e-Handbook of Statistical Methods. https://doi.org/https://doi.org/10.18434/M32189

Rezaei, E.E., Siebert, S., Ewert, F., 2015. Intensity of heat stress in winter wheat—phenology compensates for the adverse effect of global warming. Environ. Res. Lett. 10, 024012. https://doi.org/10.1088/1748-9326/10/2/024012

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

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

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

Sen, A., Srivastava, M., 1990. Regression Analysis□: Theory, Methods, and Applications. Springer New York.

Sexton, J., Everingham, Y., Inman-Bamber, G., 2016. A theoretical and real world evaluation of two Bayesian techniques for the calibration of variety parameters in a sugarcane crop model. Environ. Model. Softw. 83, 126–142. https://doi.org/10.1016/j.envsoft.2016.05.014

Skahill, B.E., Doherty, J., 2006. Efficient accommodation of local minima in watershed model calibration. J. Hydrol. 329, 122–139. https://doi.org/10.1016/J.JHYDROL.2006.02.005

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), 1998. Understanding Options for Agricultural Production, Systems Approaches for Sustainable Agricultural Development. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-017-3624-4

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

Wallach, D., Hwang, C., Correll, M.J., Jones, J.W., Boote, K., Hoogenboom, G., Gezan, S., Bhakta, M., Vallejos, C.E., 2018. A dynamic model with QTL covariables for predicting flowering time of common bean (Phaseolus vulgaris) genotypes. Eur. J. Agron. 101, 200–209. https://doi.org/10.1016/J.EJA.2018.10.003

Wallach, D., Makowski, D., Jones, J.W., Brun, F., 2019. Working with Dynamic Crop Models□: Methods, Tools and Examples for Agriculture and Environment., Third. ed. Academic Press, London, U.K.

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., Horan, H., Hoogenboom, G., 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., 2020a. How well do crop modeling groups predict wheat phenology, given calibration data from the target population? bioRxiv 708578. https://doi.org/10.1101/708578

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., 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., 2020b. Multi model evaluation of phenology prediction for wheat in Australia. bioRxiv 2020.06.06.133504. https://doi.org/10.1101/2020.06.06.133504

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.12.294744; this version posted September 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R.P., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., Reynolds, M.P., Alderman, P.D., Aggarwal, P.K., Anothai, J., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G., Doltra, J., Fereres, E., Garcia-Vila, M., Gayler, S., Hoogenboom, G., Hunt, L.A., Izaurralde, R.C., Jabloun, M., Jones, C.D., Kersebaum, K.C., Koehler, A.-K., Liu, L., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Palosuo, T., Priesack, E., Eyshi Rezaei, E., Ripoche, D., Ruane, A.C., Semenov, M.A., Shcherbak, I., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Thorburn, P., Waha, K., Wallach, D., Wang, Z., Wolf, J., Zhu, Y., Asseng, S., 2017. The uncertainty of crop yield projections is reduced by improved temperature response functions. Nat. Plants 3, 1–13. https://doi.org/10.1038/nplants.2017.102

Wang, X., Kemanian, A., Williams, J., 2011. Special features of the EPIC and APEX modeling package and procedures for parameterization, calibration, validation, and applications, in: Ahuja, L.R., Ma, L. (Eds.), Methods of Introducing System Models into Agricultural Research. American Society of Agronomy, Madison, pp. 177–208.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82.

Zadoks, J.C., Chzang, T.T., Konzak, C.F., 1974. A decimal code for the growth stages of cereals. Weed Res. 14, 415–421. https://doi.org/10.1111/j.1365-3180.1974.tb01084.x