How well do crop models predict phenology, with emphasis on the effect of calibration?

Wallach[1], Daniel; Palosuo[2], Taru; Thorburn[3], Peter; Seidel[4], Sabine J.; Gourdain[5], Emmanuelle; Asseng[6], Senthold; Basso[7], Bruno; Buis[8], Samuel; Crout[9], Neil, Dibari[10], Camilla; Dumont[11], Benjamin; Ferrise[10], Roberto; Gaiser[4], Thomas; Garcia[7], Cécile; Gayler[12], Sebastian; Ghahramani[13], Afshin; Hochman[3], Zvi; Hoek[14], Steven; Horan[3], Heidi; Hoogenboom[6,15], Gerrit; Huang[16], Mingxia; Jabloun[9], Mohamed; Jing[17], Qi; Justes[18], Eric; Kersebaum[19], Kurt Christian; Klosterhalfen[20], Anne; Launay[21], Marie; Luo[22], Qunying; Maestrini[7], Bernardo; Moriondo[23], Marco; Nariman Zadeh[24], Hasti; Olesen[25], Jørgen Eivind; Poyda[26], Arne; Priesack[27], Eckart; Pullens[255], Johannes Wilhelmus Maria; Qian[17], Budong; Schütze[28], Niels; Shelia[6,15], Vakhtang; Souissi[29,30], Amir; Specka[19], Xenia; Srivastava[4], Amit Kumar; Stella[19], Tommaso;

[1] INRA, UMR AGIR, Castanet Tolosan, France. ORCID 0000-0003-3500-8179
[2] Natural Resources Institute Finland (Luke), Helsinki, Finland
[3] CSIRO Agriculture and Food, Brisbane, Queensland, Australia
[4] Institute of Crop Science and Resource Conservation, University of Bonn, Germany
[5] ARVALIS - Institut du végétal Paris, France
[6] Agricultural and Biological Engineering Department, University of Florida, Gainesville, Florida
[7] Department of Earth and Environmental Sciences, Michigan State University, East Lansing, Michigan
[8] INRA, UMR 1114 EMMAH, Avignon, France
[9] School of Biosciences, University of Nottingham, Loughborough, UK
[10] Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Italy
[11] Department Terra & AgroBioChem, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium
[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] Environmental Sciences Group, Wageningen University & Research, Wageningen, The Netherlands
[15] Institute for Sustainable Food Systems, University of Florida, Gainesville, Florida
[16] College of Resources and Environmental Sciences, China Agricultural University, Beijing, China
[17] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada
[18] CIRAD, UMR SYSTEM, Montpellier, France
[19] Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany
[20] Institute for Bio- and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
[21] INRA, US 1116 AgroClim, Avignon, France
[22] Hillridge Technology Pty Ltd, Sydney, Australia
[23] CNR-IBIMET, Firenze, Italy
[24] Aalto University School of Science, Espoo, Finland
[25] Department of Agroecology, Aarhus University, Tjele, Denmark
[26] Grass and Forage Science / Organic Agriculture, Institute of Crop Science and Plant Breeding, Kiel University, Kiel, Germany
[27] Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany
[28] Institute of Hydrology and Meteorology, Chair of Hydrology, Technische Universität Dresden, Dresden, Germany

Streck[12], Thilo; Trombi[10], Giacomo; Wallor[19], Evelyn; Wang[16], Jing; Weber[12], Tobias, K.D.; Weihermüller[20], Lutz; de Wit[14], Allard; Wöhling[28,31], Thomas; Xiao[32,6], Liujun; Zhao[6], Chuang; Zhu[322], Yan

[29] National Institute of Agronomic Research of Tunisia (INRAT), Agronomy Laboratory, University of Carthage, Tunis, Tunisia
[30] National Agronomy Institute of Tunisia (INAT), University of Carthage, Tunis, Tunisi
[31] Lincoln Agritech Ltd., Hamilton, New Zealand
[32] 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

ABSTRACT

Plant phenology, which describes the timing of plant development, is a major aspect of plant response to environment and for crops, a major determinant of yield. Many studies have focused on comparing model equations for describing how phenology responds to climate but the effect of crop model calibration, also important for determining model performance, has received much less attention. The objectives here were to obtain a rigorous evaluation of prediction capability of wheat phenology models, to analyze the role of calibration and to document the various calibration approaches. The 27 participants in this multi-model study were provided experimental data for calibration and asked to submit predictions for sites and years not represented in those data. Participants were instructed to use and document their "usual" calibration approach. Overall, the models provided quite good predictions of phenology (median of mean absolute error of 6.1 days) and did much better than simply using the average of observed values as predictor. The results suggest that calibration can compensate to some extent for different model formulations, specifically for differences in simulated time to emergence and differences in the choice of input variables. Conversely, different calibration approaches were associated with major differences in prediction error between the same models used by different groups. Given the large diversity of calibration approaches and the importance of calibration, there is a clear need for guidelines and tools to aid with calibration. Arguably the most important and difficult choice for calibration is the choice of parameters to estimate. Several recommendations for calibration practices are proposed. Model applications, including model studies of climate change impact, should focus more on the data used for calibration and on the calibration methods employed.

Introduction

Crop models are widely used to simulate the effect of weather, soil and management on crops (Rauff & Bello, 2015; van Ittersum et al., 2003). Here we focus specifically on the use of crop models to simulate crop phenology i.e. the cycle of biological events in plants. Matching the phenology of crop varieties to the climate in which they grow is a critical crop production strategy (Hunt et al., 2019; Rezaei, Siebert, & Ewert, 2015; Rezaei, Siebert, Hüging, & Ewert, 2018). Thus, understanding and improving our ability to simulate phenology with crop models is important step in using models for improving crop management, for designing better adapted genotypes and for preparing for and adapting to global change. Process-based models similar to those for crops can be used for natural vegetation, so crop models can serve as examples for studies of phenology in ecosystems (Piao et al., 2019).

Crop model evaluation is an essential aspect of modeling, assessing whether model performance is acceptable for the intended use of the model. For studies of phenology two major questions are a) how accurate are current models for the prediction of crop development stages? and b) what determines model accuracy and what does that imply about how accuracy can be improved? We use here prediction in the sense of determining outputs (dates of development stages) from known inputs (weather, soil, management). The problem of predicting future events, with unknown weather, is not considered.

There have been numerous evaluation studies of crop model simulations, including but not restricted to phenology, both of individual models and of multi-model ensembles. The typical procedure is to first calibrate each model using a part of the available field data and then to evaluate it using the remaining data.

Most crop model evaluation studies focusing on crop phenology have had relatively little data for calibration or evaluation. (Andarzian, Hoogenboom, Bannayan, Shirali, & Andarzian, 2015) for example, used data from one location covering five growing seasons and two or three sowing dates per year. Out of these data, one year was used for calibration and the other two years of data to evaluate the model. (Yuan, Peng, & Li, 2017) used one year of data for calibration and the second year of data from the same location for evaluation of the rice crop model ORYZA. Hussain, Khaliq, Ahmad, & Akhtar (2018) tested four models using data from two locations with two years of data, 11 crop planting dates, and three varieties. Paucity of data means that model parameters are estimated with relatively large uncertainty and model evaluation is quite uncertain.

Another common feature of crop model evaluation is that the data are often such that model error for the evaluation data cannot be assumed to be independent of model error for the calibration data. That holds for the examples listed above since the evaluation and calibration data come from the same sites. In such cases, the evaluation does not give an unbiased estimate of how well the model will predict for other sites not included in the calibration data. Since usually the model is meant for use over a range of sites, this clearly reduces the usefulness of the evaluation information.

A third feature often found in crop model evaluation is that the range of situations from which the calibration data are drawn (the “calibration population”) is often different than the range of conditions from which the evaluation data are drawn (the “evaluation population”). For example, Hussain et al. (2018) used data from an experiment that included a range of crop stresses to calibrate their model. They used data from the least stressed treatment in the calibration process and evaluated the resultant model on the remaining planting dates at the same location. The evaluation data thus represented a different range of conditions than the calibration data. In a multi-model ensemble study of the effect of high temperatures on wheat growth (Asseng et al., 2015) detailed crop measurements were provided for one planting date and the models were evaluated using other planting dates, some with additional artificial heating. Again, the evaluation data represented a much larger range of temperatures than represented in the calibration data. While the capacity of crop models to extrapolate to conditions quite different than those of the calibration data is obviously of interest, it is a rather different type of evaluation than the case where the calibration and evaluation populations are similar.

Thus, evaluation of crop phenology models to date has mainly concerned situations that would tend to make prediction difficult, because of small amounts of data for calibration and differences between the calibration and target populations. Furthermore, the quality of the evaluation is often questionable, because of the relatively small amounts of data and the non-independence of the errors for the calibration and evaluation situations. There is thus a need for more rigorous assessments of simulation capability of crop phenology models. The first objective of this paper is, therefore, to provide a rigorous evaluation of how well crop models predict wheat phenology, in the situation where there is substantial data for calibration and where the calibration and evaluation data can be assumed to come from the same underlying population. To ensure the rigor of the evaluation, we create a situation where the model errors for the calibration and simulation data can be assumed independent.

The emphasis in model evaluation studies is often on the role of model structure, i.e. model equations (Maiorano et al., 2017; Svystun, Bhalerao, & Jönsson, 2019; Wang et al., 2017). There has been relatively little work on the diversity and importance of calibration approaches in crop modeling. Clearly however the simulated values also depend on the parameter values estimated by calibration and therefore on the calibration approach.

Confalonieri et al., (2016) found that the model user, responsible for calibration, had a very large effect on predictive quality. The second objective of this study then was to investigate the role of calibration in determining prediction quality.

In a wide-ranging survey, (Seidel, Palosuo, Thorburn, & Wallach, 2018) found that there is a wide diversity of calibration strategies used for crop models, but for that survey each response was for a different prediction problem. This did not address the problem of the diversity of calibration approaches by different groups given the same data and the same prediction objectives. The third objective of this study was therefore to obtain detailed information about the diversity of calibration strategies adopted by different modeling groups for the same prediction problem. This is useful as a step toward developing guidelines for calibration of phenology models, in that it helps define the range of possible approaches. This is of practical interest not only for stand-alone phenology models, but also for crop models more generally, since crop models are often calibrated first just for phenology, and then separately for biomass increase and partitioning and soil processes.

Materials and Methods

Experimental data

The data were provided by ARVALIS – Institut du vegetal, a French agricultural technical institute. They run multi-year multi-purpose trials at multiple locations across France, which include variety trials. The data here are from the two winter wheat check varieties, *Apache* which is a common variety grown throughout France and Central Europe and *Bermude*, mainly grown in Northern and Central France. The trials have three repetitions and follow standard agricultural practices, with N fertilization calculated to be non-limiting. Thus, both the calibration and evaluation data are drawn from sites in France where winter wheat is grown, subject to standard management.

The observed data used in model calibration and evaluation are the dates of two development stages, namely beginning of stem elongation (growth stage 30 on the BBCH and Zadoks scales (Zadoks, Chzang, & Konzak, 1974) and middle of heading (growth stage 55 on the BBCH and Zadoks scales). These 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 nitrogen fertilizer applications.

The data were divided into two categories (table 1). One part, the calibration data (six sites, five years for a total of 14 environments i.e. site-year combinations), was provided to participants for calibration. A second part, the evaluation data (five sites, two years for a total of eight environments), was not given to participants. The division of the data was such that the calibration and evaluation data had no sites or years in common. To achieve this some data (denoted "other" data) were used neither for calibration nor evaluation (but were used in the calculation where overall variability in simulated values was evaluated).

Table 1.

Environments (site-year combinations) that provided the data. C= calibration data. E = evaluation data. O = data not used for calibration or evaluation (only used for evaluating variability between models). Blank cells indicate no data.

Site (longitude,latitude)	Harvest year						
	2010	2011	2012	2013	2014	2015	2016
FORESTE (3.20,49.82)			E	E	OO*	O	
MERY	C	C		O	C	C	

4.02,48.33)							
ROUVRES 5.09,47.28)			E	E	O	O	
CESSEVILLE[1] (0.90,49.15)		C					
IVILLE[1] (0.90,49.15)			E				
VILLETTES[1] (0.90,49.15)				E			
EPREVILLE[1] (0.90,49.15)					C		
CRESTOT[1] (0.90,49.15)						C	
OUZOUER (1.52,47.90)		O	E	E	O	O	
BIGNAN (-2.73,47.88)	C	C	O	O	C	C	
BOIGNEVILLE (2.38,48.33)			O	O	C	C	C

*There were two sowing dates at FORESTE in 2014. [1] These are separate sites that are geographically close to one another and share a single weather station.

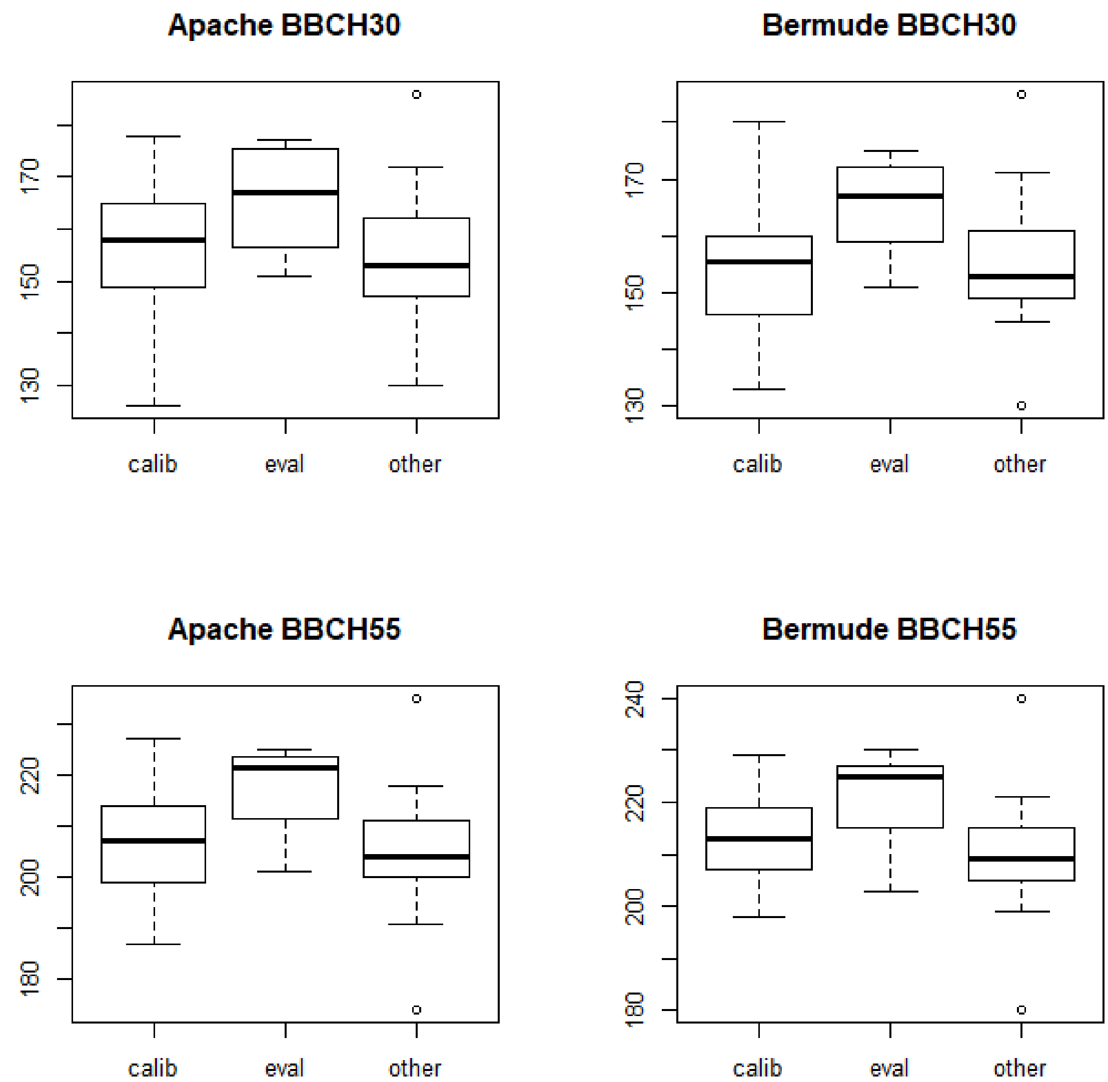
The background and input information provided to the modelers for all environments included information about the sites (location, soil texture, field capacity, wilting point), management (sowing dates, sowing density, irrigation and fertilization dates and amounts), and daily weather data (precipitation, minimum and maximum air temperature, global radiation and potential evapotranspiration). Initial soil water and N content were not measured in these experiments, but best estimates were provided by the experimental scientist. If any models required other input data, modeling groups were asked to derive those values in whatever way that seemed appropriate.

The range of observed days from sowing to development stages BBCH30 and BBCH55 for the two varieties for each category of data (calibration, evaluation, other hidden data) is shown in figure 1. The spread from minimum to maximum in the evaluation data is between 24 and 27 days depending on stage and variety. The spread is larger for the calibration data, and in fact, the calibration data cover the range of the evaluation data and the range of other hidden data. Thus, the models are not being used to extrapolate outside the observed values of the calibration data.

Figure 1

Boxplots of calibration, evaluation and other data for development stages BBCH30 and BBCH50 and varieties Apache (left) and Bermude (right). The y-axis shows days from sowing to the indicated development stage. 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 go beyond that range.



Crop models

Twenty-seven modeling groups participated in this study, noted M1-M27. Information about the underlying model structures is given in Supplementary table S1. The four groups M2, M3, M4, and M5 all used the same model structure (i.e. models with the same name), denoted as model structure S1. The four models M7, M12, M13, and M25 also shared a common model structure, denoted as S2. As will be seen, different groups using the same model structure had different results. This could be due to differences in model version, but the differences are not in the basic phenology equations, and therefore, should have no or only a negligible effect on the simulated development stages. The differences are assumed to mainly be due to differences in the values of the parameters, either those not fit by calibration or those estimated by calibration. Since groups using the same model structure obtained different results, we refer to the 27 contributions as different models. In the presentation of the results the models are anonymized and are identified simply as M1 to M27. It would be misleading to use the names of the model structures, since the results depend on both model structure and on the values of the parameters.

Two of the models (M9, M18) only simulated days to development stage BBCH55 and not to stage BBCH30. Results for these two models are systematically included with the results for the other models, but averages over development stages for these two models only refer to BBCH55. This is not repeated explicitly every time an average over development stages is discussed.

In addition to the individual model results, we show the results for the model ensemble mean ("e-mean") and the model ensemble median ("e-median"). We also define a very simple predictor, denoted "naive", which was calculated as the average of the observations in the calibration data for prediction. The naive model thus predicts that all days from sowing to stage BBCH30 (BBCH55) will correspond to the average of days from sowing to BBCH30 (BBCH55) in the calibration data, separately for each variety. The naive model predictions for days from sowing to BBCH30 and BBCH55 are respectively 155.9 days and 206.9 days for variety Apache, and 156.1 days and 213.1 days for variety Bermude.

Calibration and simulation experiment

The participants were provided with observed phenology data (dates of BBCH30 and BBCH55) only for the calibration environments. The participants were asked to calibrate their model using those data, and then to use the calibrated model to simulate phenology for all environments (i.e. calibration, evaluation and hidden data environments). No guidelines for calibration were provided. Participants were instructed to calibrate their model in their "usual way" and fill out a questionnaire explaining what they did (Supplementary table S2).

Evaluation

A common metric of error is mean squared error (MSE). We calculated MSE for each model, each development stage (BBCH30 and BBCH55) and for each variety, as well as averaged over stages and varieties. This was done separately for the calibration and evaluation data. For example, MSE for model m, for predicting BBCH30, variety Apache, based on the evaluation data, is:

$$MSE_{eval,m}^{BBCH30,Apache} = (1/8)\sum_{i \in eval}\left(y_i^{BBCH30,Apache} - \hat{y}_{i,m}^{BBCH30,Apache}\right)^2 \quad (1)$$

where the sum is over the eight environments used for evaluation and $y_i^{BBCH30,Apache}$ and $\hat{y}_{i,m}^{BBCH30,Apache}$ are respectively the observed value and value simulated by model m for evaluation environment *i*, development stage BBCH30 and variety Apache. For $MSE_{eval,m}^{all}$, the average is over the eight evaluation environments, both stages and both varieties, so overall 32 predictions.

Mean squared error can be shown to be the sum of three positive terms, namely squared bias, the difference in variance between the observed and simulated values and a term related to the correlation between observed and simulated values (Kobayashi & Salam, 2000). We specifically examined the bias, defined as the average over observed values minus the average over simulated values.

The mean absolute error (MAE) is of interest as a more direct measure of error, that does not give extra weight to large errors as MSE does. For example, MAE for model m for predicting BBCH30, variety Apache, based on the evaluation data, is:

$$MAE_{eval,m}^{BBCH30,Apache} = (1/8)\sum_{i \in eval} \left| y_i^{BBCH30,Apache} - \hat{y}_{i,m}^{BBCH30,Apache} \right|$$

We also look at modeling efficiency (EF) defined for model m as

$$EF_m = 1 - MSE_m / MSE_{naive}$$

where MSE_m is MSE for model m and MSE_{naive} is MSE for the naive model defined above. EF is a skill measure, which compares the predictive capability of a model to that of the naive model. Since the naive model makes the same prediction for all environments, it does not account for any of the variability between environments. A model with $EF \le 0$ is a model that does no better than the naive model, and so would be considered to be a very poor predictor. A perfect model, with no error, has modeling efficiency of 1.

Results

Goodness-of-fit and prediction error

Summary statistics for MSE averaged over both varieties and over both development stages, for the calibration and evaluation data, are shown in table 2. Summary MSE values for the calibration data for each development stage and variety separately are shown in Supplementary table S7, and results for each individual model are given in Supplementary figure S1.

Table 2

Summary statistics of MSE (days²) averaged over both varieties and over both development stages.

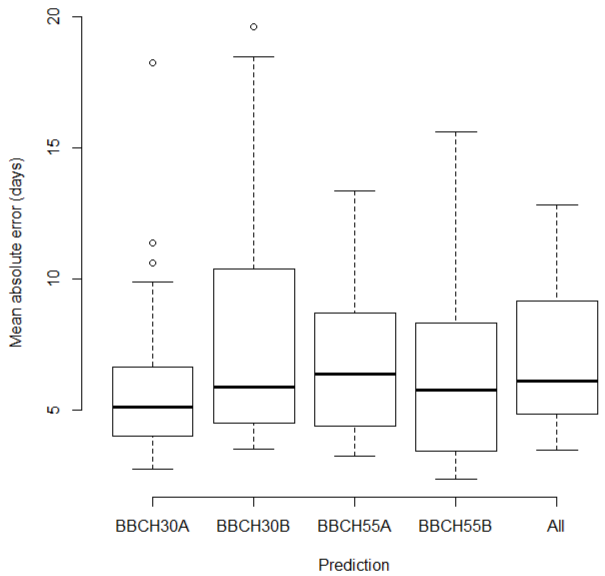
MSE (days²)	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Calibration data	15	28	47	77	63	426
Evaluation data	20	35	62	79	111	235

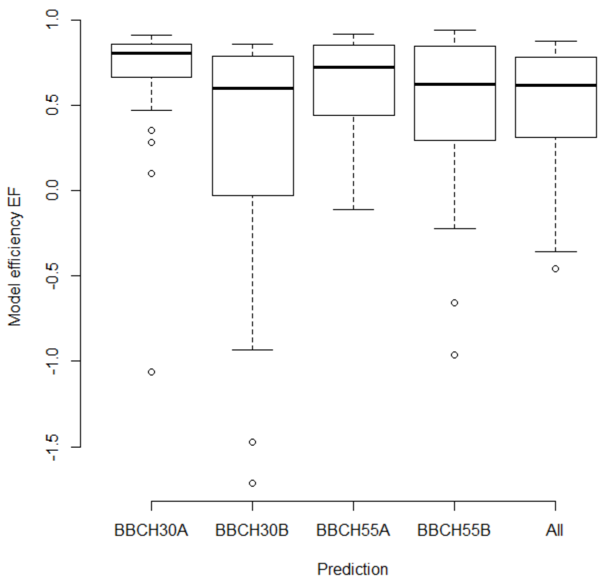
Figure 3 and Supplementary tables S4-S6 show results for each development stage and variety and averaged over development stages and varieties for the evaluation data. Results for each model are given in Supplementary table S3. The median of MAE for the evaluation data is 6.1 days. The median of overall efficiency is 0.62, signifying that half of the models have MSE values for the evaluation data that are at most 38% as large as that of the naive predictor. Only two models have negative values of EF, indicating that one would do better to predict using the average of the calibration data. For the four individual predictions (two development stages, two varieties), the median of MAE ranges from 5.1 to 6.4 and the median of EF ranges from 0.6 to 0.8. The ensemble models e-median and e-mean, though not the best predictors, are among the best, with e-median being rated second best and e-mean fourth best.

The range of results among individual models is appreciable. The mean absolute errors for the evaluation data averaged over all predictions (MAE_{eval}^{all}) go from 3.5 to 13 days. The MSE_{eval}^{all} values vary by over a factor of 10, from a minimum of 20 days² to a maximum of 235 days².

Figure 3

Box and whisker diagrams of absolute errors for evaluation data for each prediction and on average (top panel) and modeling efficiency for each prediction and on average (bottom panel). BBCH30A and BBCH30B refer respectively to prediction of days to BBCH30 for variety Apache and variety Bermude. BBCH55A and BBCH55B refer respectively to prediction of days to BBCH55 for variety Apache and variety Bermude. The variability comes from differences between models.





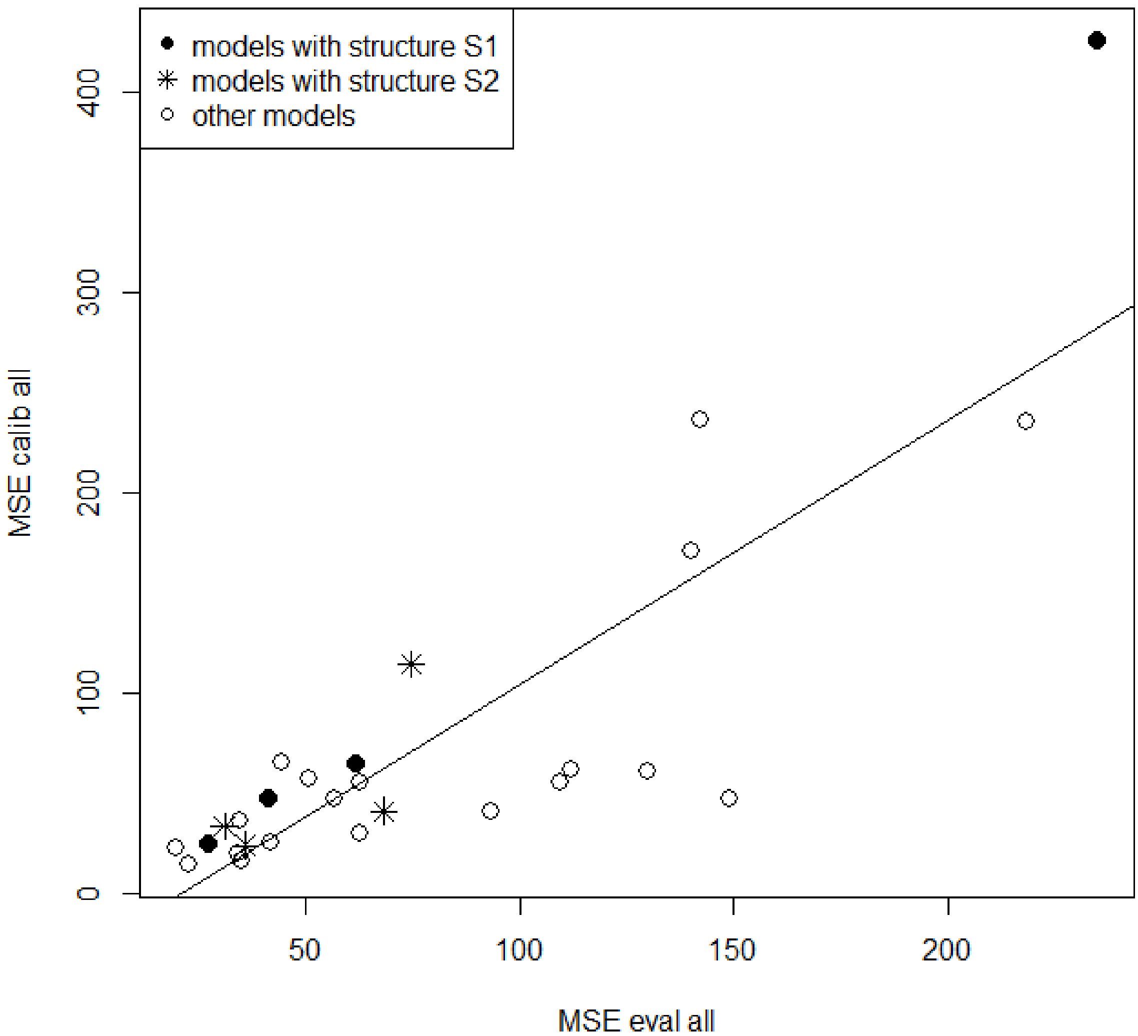
Role of calibration

The relationship between overall MSE for the evaluation data and overall MSE for the calibration data is quite close (adjusted R^2=0.70, figure 4). That is, much of the variability between models in MSE for the evaluation data can be explained by the variability in MSE values for the calibration data, which emphasizes the importance of obtaining a good fit to the calibration data, which in turn depends to a large extent on the method of calibration.

The four models that have model structure S1 and the four models that have model structure S2 are identified in figure 4. Models with the same structure have different MSE values; the differences are particularly large for S1. The models with structure S1 are ranked 3rd, 9th, 14th and 27th best for overall evaluation MSE among the 27 individual models. The models with structure S2 are ranked 4th, 8th, 17th and 18th best.

Figure 4

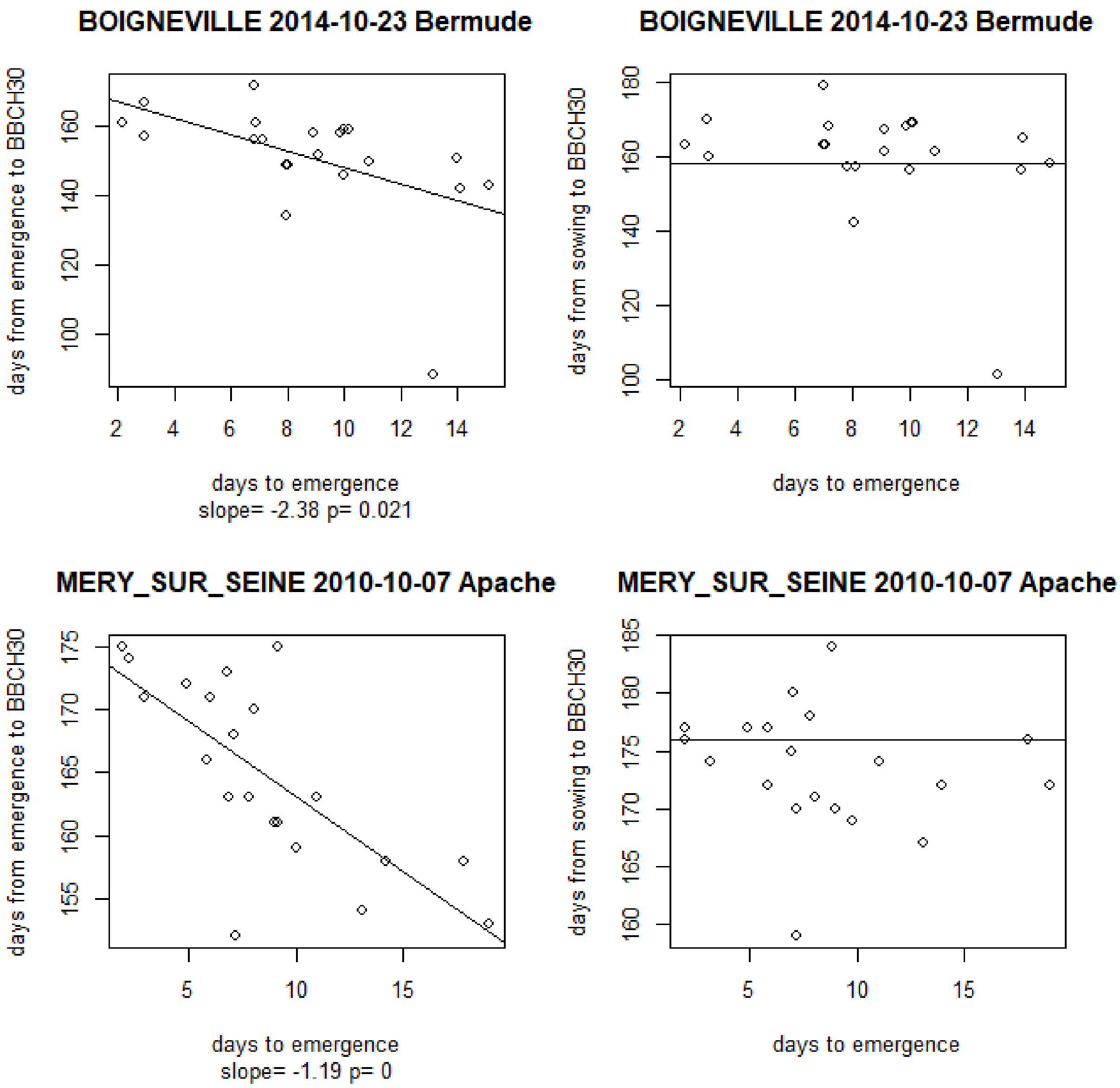
Mean squared error (MSE) for the calibration data, averaged over environments, development stages and varieties (MSE_{calib}^{all} days²), as related to MSE for the evaluation data (MSE_{eval}^{all}, days²). The regression line MSE_{calib}^{all} = -27.6 + 1.32 MSE_{eval}^{all} is shown (R^2=0.70). ● indicates models with structure S1. ⁎ indicates models with structure S2. ○ indicates other models.



Twenty-one models simulated and reported time from sowing to emergence. For these models, we can separate simulated time from sowing to BBCH30 (sow_30) into two contributions, the simulated time from sowing to emergence (sow_em) plus the simulated time from emergence to BBCH30 (em_30), so that sow_30=sow_em+em_30. Figure 5 shows results from two environments, typical of essentially all environments and both varieties, for the relation between em_30 or sow_30 and sow_em. The average slope of the regression em_30=a+b*sow_em over all environments (including calibration, evaluation and other environments) and both varieties is b=-1.04, so that each day increase in simulated days to emergence is on average associated with a 1.04 day decrease in simulated time from emergence to BBCH30. The negative correlation between sow_em and em_30 leads to a between-model variance for sow_30 (average variance 92 days^2) that is smaller than the sum of the variances of sow_em (average variance 20 days^2) and em_30 (average variance 101 days^2). The right panels of figure 5 show that different models could simulate almost exactly the observed value of sow_30 with quite different values of sow_em.

Figure 5

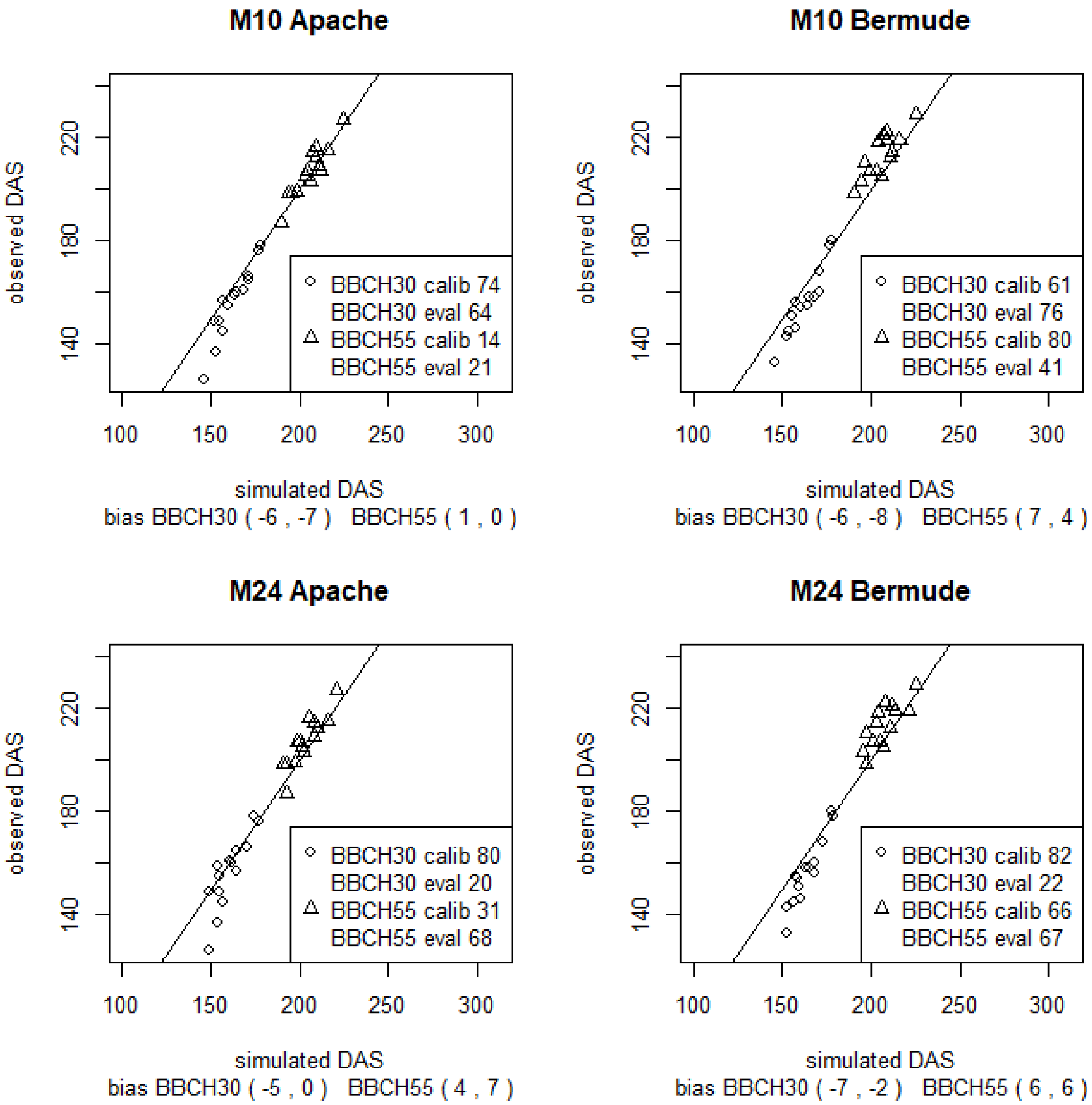
Relation between simulated days from emergence to BBCH30 and simulated days from sowing to emergence as reported by 21 crop models for two environments (left panels). Relation between simulated days from sowing to BBCH30 and simulated days from sowing to emergence for the same environments (right panels). A small amount of noise has been applied to avoid overlap. The slope of the linear regression line and the p-value for testing slope=0 are shown for the left panels. The observed days from sowing to BBCH30 is shown as a horizontal line in the right panels.



Bias (average over environments of observed values – average of simulated values) is one aspect of goodness-of-fit. For most models, the bias for the calibration data is quite small. Considering absolute bias for both development stages and both varieties, the median value over models was 2 days (Supplementary table S8). In cases where the bias is relatively large, it is often of opposite sign for BBCH30 and BBCH55, as in the examples of figure 2.

Figure 2

Observed vs. simulated days after sowing (DAS) for calibration data for models M10 and M24. The legend shows MSE (days²) for each stage and for calibration and evaluation data. (The individual evaluation results are not displayed). In the subtitles, bias values (days) for each stage are shown. The first number in parentheses is for the calibration data, the second number is for the evaluation data.



Calibration approach

Each participant was asked to calibrate the model in the "usual" way, using the calibration data provided. The questionnaire about calibration focused on three aspects of calibration; the criterion of error to be minimized, the software used and the choice of parameters to estimate. The choices of the participants are summarized in table 3 and choices for each model are shown in Supplementary table S9.

Table 3

Summary of calibration approaches. Numbers are number of models with indicated choice. The specific models associated with each choice are shown in Supplementary tables S3 and S9. More information about the software is presented in Supplementary table S10.

Number of parameters[1]	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	1.00	2.00	3.00	3.63	4.50	9.00
Which parameters	Thermal time to a single development stage 16					
	Thermal time to two or more development stages 6					
	Related to vernalization 11					
	Related to photoperiod 11					
	Related to effect of temperature (e.g. base temperature) 6					
	Related to phyllochron 6					
	Related to tiller appearance 2					

	Related to time to emergence 3 Parameters unrelated to calibration data[2] 6
Objective function	Sum of squared errors or of root mean squared errors 21 Sum of absolute errors 2 Concentrated likelihood 1 No single explicit objective function 3
Software[3]	Trial and error 10 DIRECT-L (Gablonsky & Kelley, 2001; Johnson, n.d.) 2 Ucode (E. P. Poeter, Hill, Banta, Mehl, & Christensen, 2005; Eileen P. Poeter & Hill, 1999) 3 DE Optim (Mullen, Ardia, Gil, Windover, & Cline, 2011) 3 PEST (Doherty, Hunt, & Tonkin, 2010) 2 SCE (Duan, Gupta, & Sorooshian, 1993; Houska, Kraft, Chamorro-Chavez, & Breuer, 2015) 2 GLUE (Beven & Binley, 2014; J. He, Jones, Graham, & Dukes, 2010) 1 DREAM (J. A. Vrugt et al., 2009; Jasper A. Vrugt, 2016) 2 Wrote code [4] 4

358 **[1] Summary of number of estimated parameters for models M1-M27. [2] These are**
359 **parameters that do not affect simulated days to BBCH30 or BBCH55. [3] Some modeling**
360 **groups used more than one software package. [4]Modeling groups that wrote their own**
361 **software.**

362 **Objective function**

363 Most modeling groups defined the sum of squared errors or the sum of root mean
364 squared errors as the objective function to be minimized, where the sum is over the two

stages. (In all cases, the calibration was done separately for the two varieties). Two groups minimized the sum of absolute errors. Calibration for model M21 was based on maximizing the concentrated likelihood (Seber & Wild, 1989) assuming a normal distribution of errors with possibly different error variances for the two development stages. In this case, the objective function involves a product of errors for the two outputs, rather than a sum. Four of the participants (M12, M16, M18) did not define an explicit objective function to be minimized. In these cases, the parameter values were chosen to obtain a “good fit” to the data by visual inspection. Finally, two of the models (M7, M8) divided the calibration into two steps. In these cases three of the parameters were used to fit the BBCH30 data, and then in another step another parameter was used to fit the BBCH55 data.

Minimizing the sum of squared errors is a standard statistical approach to model calibration, which has highly desirable properties if certain assumptions about model error are satisfied, including equal variance of model error for all data points and non-correlation of model errors. Only model M21 took into account the possibility that the error variances are different for BBCH30 and BBCH55, and none of the modeling groups took into account possible correlations between errors for BBCH30 and BBCH55 in the same field. Based on the errors for all the data and all the models, it was found that there is a highly significant difference in variance between errors for BBCH30 (variance of error 100.7 days^2) and BBCH55 (variance of error 67.3 days^2). Also, the correlation between the error for BBCH30 and the error for BBCH55 in the same field is 0.53 and highly significant. However, if only results for a single model are considered, then for most models the differences in variance and the correlation are not significant.

Two models defined a posterior probability of the parameters equal to the likelihood times the prior probability, as usually assumed in a Bayesian approach. The parameters used for prediction were those that maximized the posterior probability (i.e., the estimated mode of the posterior distribution). In both cases, the likelihood was assumed Gaussian with independent errors, and the prior distribution was assumed uniform between some minimum and maximum value. This approach is equivalent to minimizing the mean squared error, with constraints on the parameter values.

Software

Seven participants simply used trial and error to search for the optimal parameters. The other participants used software specifically adapted to minimizing the objective function, either written specifically for their model or, in most cases, available from other sources (Supplementary table S10).

Choice of parameters to estimate

The choice of parameters to estimate was based on expert judgement in most cases. The participants declared that they chose parameters known to affect phenology in the model, or more specifically parameters expected to have a major effect on time to BBCH30 and BBCH55 and expected to differ between varieties. Five participants did a sensitivity analysis to aid in the choice of parameters to estimate. The number of estimated parameters ranged from 1 to 9. In almost all cases, the number of parameters to estimate was decided a priori. In three cases, the number was the result of testing the fit with different numbers of parameters. In one of those cases the Akaike Information Criteria (AIC, Akaike, 1973) and adjusted R^2 were used to test whether additional parameters should be estimated.

Almost all modeling groups estimated one or more parameters that represent thermal time between development stages (table 6). Some adjustments were necessary for models that did not explicitly calculate time of BBCH30 or BBCH55. In model M2, for example, a new parameter was added to the model, and estimated, representing the fraction of thermal time from double ridge to heading at which BBCH30 occurs. Thirteen groups estimated a parameter related to the effect of photoperiod. Ten groups estimated a parameter related to vernalization. Six groups modified one or more parameters related to the temperature response (for example model M6 estimated *Tbase*, the temperature below which there is no development). Only three models modified parameters related to the time from sowing to emergence, and only one model modified a parameter related to the effect of water stress. Six models included among the parameters to estimate, parameters that have no effect on the variables furnished as calibration data. Such parameters included thermal times for development stages after BBCH55, potential kernel growth rate, kernel number per stem weight and the temperature below which there is 50% death due to cold (Supplementary table S9).

Discussion

Prediction error

The challenge in this study was to predict the time from sowing to beginning of stem elongation and to heading in winter wheat field trials performed across France. This is a problem of practical importance, since these two development stages are important for wheat management (e.g. fertilization). The evaluation concerned years and sites not included in the calibration data, making this one of the most rigorous evaluations to date of how well crop models simulate phenology.

Twenty-seven modeling groups participated in the exercise. Most models predicted times to stem elongation and heading quite well (median MAE of 6 days). Half the models had MSE values of prediction that were 36% or less than MSE of a naive predictor. It must be kept in mind that this study is a rather favorable situation for prediction, with a substantial amount of calibration data and predictions for environments similar to those of the calibration data.

Role of calibration

What is the role of calibration in determining prediction accuracy? We cannot answer this exactly, because differences between models result not only from differences in calibration approach, but also from differences in structure and from differences in the values of parameters not estimated by calibration. However, several aspects of the results indicate that calibration is important.

Consider first the comparison between models with the same structure. There are fairly large differences in MSE between models with the same structure. This could partially be due to different values for the parameters not estimated by calibration. However, since there are major differences in calibration approach, and in general the parameters estimated by calibration are among the most important controlling phenology, it seems reasonable to conclude that the differences between models with the same structure are largely due to differences in calibration.

Conversely, our results indicate that calibration can result in models with very different structures achieving similar values of MSE. One essential aspect of model structure is the choice of input variables. In fact, MSE can be expressed as a sum of two terms, the first of which depends only on the choice of the model input variables, while the second measures the distance between the model used and the optimal model for those inputs (Wallach, Makowski, Jones, & Brun, 2019). Calibration has a major effect of the second term, and in fact the objective of calibration is to minimize that term. The most important inputs that determine spring wheat phenology are daily temperature and photoperiod (Aslam et al., 2017) and for winter wheat it is also necessary to include the process of vernalization, i.e. the effect of low winter temperatures on development (Li et al., 2013). Five of the best eight predicting models here, with MSE_{eval}^{all} < 40 days², do use all three of those variables (daily temperature, photoperiod, vernalizing temperatures) as inputs. Two of those best eight models however do not use vernalizating temperatures, and one of those best eight does not use photoperiod. Thus there are similarly low values of MSE for prediction even for models so fundamentally different in structure that they use different input variables. It seems likely that this is largely due to the fact that the different models are calibrated using the same data.

Another indication that calibration compensates for differences in structure is the result that there is less variability between models for predicting days from sowing to BBCH30, which is provided as calibration data, than would be expected if the uncertainties in days from sowing to emergence and days from emergence to BBCH30 simply added up. Compensation is usually discussed in the context of single models. For example, equifinality, which is a well-known phenomenon of complex models, means that different combinations of parameter values, and thus different quantitative descriptions of processes, can lead to the same results for outputs because there is compensation between the processes (Beven, 2006; D. He et al., 2017). However, this phenomenon has not been described in the context of multi-model studies. Here, we have an example of compensation for differences between models in the way they partition days from sowing to BBCH30 into days from sowing to emergence plus days from emergence to BBCH30. Models with longer simulated times from sowing to emergence tend to have a shorter simulated time from emergence to development stage BBCH30 and vice versa. In fact, each extra day from sowing to emergence is associated on average with almost exactly one less day from emergence to BBCH30. The result is that models with quite different simulated days from sowing to emergence can have nearly identical times from sowing to BBCH30. This can be expressed in terms of model uncertainty, as quantified by between-model variance. The variance of days from sowing to BBCH30 is less than the sum of variances of days from sowing to emergence and days from emergence to BBCH30. That is, calibration reduces, but does not eliminate, model uncertainty for the variable provided for calibration.

We do not have observed time to emergence, but in any case, the models with different simulated days to emergence can't all be right. This is an example of how models can get the right answer (correct days to BBCH30, thanks to calibration) for the wrong reasons (wrong days to emergence and compensating wrong days from emergence to BBCH30), illustrating the problem pointed out for example by (Challinor, Martre, Asseng, Thornton, & Ewert, 2014). The same compensation of errors between sowing to emergence and emergence to BBCH30 will not be appropriate for all environments. This is one of the main reasons that extrapolation to populations different than the calibration population is dangerous.

Much previous work on improving the predictive capability of crop models has focused on the model equations, for instance the way temperature is taken into account in various processes (Maiorano et al., 2016; Wang et al., 2017). Here we show that models with the same structure can have very different levels of prediction error, if the calibration methods differ, while models with quite different structures can have very similar prediction accuracy, thanks to calibration using the same data. This means that model comparison studies may often be comparing calibration approaches as much or more as they are comparing model equations. This is in line with the conclusions of Confalonieri et al. (2016), who argued that one should not speak of evaluation of a model but rather of the combination of a model and a model user, where a major role of the user is in implementing calibration.

Calibration approach

This study was designed to identify how different groups do calibration, given the same data and prediction objectives. We focused here on three specific aspects of the calibration approach; the choice of objective function, the software used and the choice of parameters to estimate. The results show the diversity of approaches. Since different models differ in multiple ways, the study does not allow us to define best practices for each aspect of calibration. However, it is possible to point out practices which should probably be avoided.

Objective function

Most participants defined an objective function based on what one would use in a statistical approach to non-linear regression, namely a sum of squared errors to be minimized or a likelihood to be maximized. However three models (see Supplementary table S9) did not have an explicit quantitative objective function. Those models all had relatively large values of overall MSE for the evaluation data (MSE_{eval}^{all}) , having 15th, 16th, and 18th largest MSE_{eval}^{all} values out of the 25 models that predicted both BBCH30 and BBCH55. It seems reasonable to suppose that the lack of a quantitative objective function can be a drawback since then one does not have a clear criterion for deciding on the best parameter values.

Among the models that chose to minimize a sum of squared errors or to maximize a likelihood, all but one implicitly or explicitly assumed that all model errors had equal variance and were independent. This will in general not be the case when there are multiple measurements in the same field, as is the case here (measurement of days to BBCH30 and days to BBCH55 in each field). Ignoring unequal variances and correlated errors in non-linear regression leads to inefficient estimators (Seber & Wild, 1989). One should at least test whether heteroscedasticity and non-independence are important.

Software

Several different software solutions were used for calibration by the different models. There does not seem to be any clear connection between the software used and predictive quality. Various different software solutions were used by the best predicting models, but largely the same software solutions were also found among the models with the largest prediction errors.

A problem that may arise concerns the test for convergence to the parameter values that minimize the chosen objective function. Having such a test allows the user to have confidence that the best parameter values have been found. With trial and error, there is no such test, which is a major drawback of this approach. Algorithms to estimate a Bayesian posterior distribution normally test convergence to the posterior distribution, which may not be relevant if one is using just the mode of the distribution. It would be good practice to adopt a software option that includes an appropriate test of convergence.

Choice of parameters to estimate

There was a large diversity of choices of parameters to estimate by calibration, and this had in certain cases an important effect on prediction error. One rather unexpected observation was that several participants included, among the parameters to estimate, parameters that have no effect on the variables furnished as calibration data among the parameters to estimate. The data cannot in those cases give any information about the parameter value. At best, including such parameters among the parameters to estimate is useless, and those parameters will simply have final values exactly equal to their initial values. However, there may also be serious disadvantages to including such parameters. It gives the erroneous impression that one is estimating parameters that cannot in fact be estimated, it increases computation time and it can cause problems for the parameter estimation algorithm. The very poor fit of model M5 to the calibration data seems to be directly related to the fact that for this model, several parameters unrelated to the calibration data were chosen to be fitted. The software used here was PEST (Doherty et al., 2010), with the singular value decomposition option, which allows one to deal with non-estimable parameters, but at the cost of introducing bias in the estimated parameter values. This bias may be at the origin of the poor performance. Obviously, one should not include non-estimable parameters among the parameters to estimate.

The choice of parameters to estimate may be the principal cause of bias in fitting the calibration data for some models. If a model includes an additive constant term, and squared error is minimized, bias will be 0 for the calibration data. Even for more complex models, calibration can bring bias close to 0, as illustrated here by the fact that many of the models had very small biases for the calibration data. Eliminating bias is important, since squared bias is one component of MSE, and therefore the bias necessarily adds on to MSE (Kobayashi & Salam, 2000). If one does not have a parameter with a nearly additive effect for each of the development stages BBCH30 and BBCH55, the elimination of bias for both outputs is not assured. Model M24 estimated only a single parameter. In such a case, at best one can estimate a parameter value that gives the best compromise between errors in BBCH30 and BBCH55. This may lead to a negative bias for one of those outputs and more or less corresponding positive bias for the other. This is exactly the behavior illustrated in figure 2. Model M10 also had fairly large biases. Here three parameters were estimated, but one is unrelated to the observed data and a second concerns time to emergence, which was only allowed to vary in a limited range. Apparently in this case also there was not enough flexibility to eliminate bias for both development stages. Models with large bias for the calibration data tended to have large MSE values for the evaluation data (Supplementary figure S2). This suggests that the parameters to estimate should include one parameter that is nearly additive (i.e. that adds an amount that is nearly the same for all environments) for each observed output, and that is not too limited in the allowed range of values.

Conclusions

Overall, we have shown in a rigorous evaluation of prediction for new environments that most of the 27 crop models tested, given calibration data, provide good predictions of phenology in winter wheat and do much better than predicting with the average of the calibration data. Calibration has a major effect on predictive quality. Calibration can compensate to some extent even for different choices of input variables. It reduces variability between models for outputs used for calibration, but may lead to models getting the right answer for the wrong reason. Poor practices of calibration can seriously degrade predictive capability. Arguably the most difficult aspect of calibration, and yet the least studied, is the choice of parameters to estimate. Unlike the choice of objective function and of software, there is little guidance here from other fields. Furthermore, the problem is specific to each model, since each model has a different set of parameters. Given the large diversity of calibration approaches and the importance of calibration, there is a clear need for guidelines and tools to aid model users with respect to calibration. Model applications, including model studies of climate change impact, should focus more on the data used for calibration and on the calibration methods employed.

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 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.

References

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

Andarzian, B., Hoogenboom, G., Bannayan, M., Shirali, M., & Andarzian, B. (2015). Determining optimum sowing date of wheat using CSM-CERES-Wheat model. *Journal of the Saudi Society of Agricultural Sciences*, *14*(2), 189–199. https://doi.org/10.1016/J.JSSAS.2014.04.004

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? *Frontiers in Environmental Science*, *5*, 57. https://doi.org/10.3389/fenvs.2017.00057

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., … Zhu, Y. (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, *5*(2), 143–147. https://doi.org/10.1038/nclimate2470

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36.

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, *28*(24), 5897–5918. https://doi.org/10.1002/hyp.10082

Challinor, A., Martre, P., Asseng, S., Thornton, P., & Ewert, F. (2014). Making the most of climate impacts ensembles. *Nature Climate Change*, *4*(2), 77–80. https://doi.org/10.1038/nclimate2117

Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., … Acutis, M. (2016). Uncertainty in crop model predictions: What is the role of users? *Environmental Modelling & Software*, *81*, 165–173. https://doi.org/10.1016/j.envsoft.2016.04.009

Doherty, J. E., Hunt, R. J., & Tonkin, M. J. (2010). *Approaches to highly parameterized inversion: A guide to using PEST for model-parameter and predictive-uncertainty analysis: U.S. Geological Survey Scientific Investigations Report 2010–5211*. Retrieved from http://pubs.usgs.gov/sir/2010/5211

Duan, Q. Y., Gupta, V. K., & Sorooshian, S. (1993). Shuffled complex evolution approach for effective and efficient global minimization. *Journal of Optimization Theory and Applications*, *76*(3), 501–521. https://doi.org/10.1007/BF00939380

Gablonsky, J. M., & Kelley, C. T. (2001). A Locally-Biased form of the DIRECT Algorithm. *Journal of Global Optimization*, *21*(1), 27–37. https://doi.org/10.1023/A:1017930332101

He, D., Wang, E., Wang, J., Lilley, J., Luo, Z., Pan, X., … Yang, N. (2017). Uncertainty in canola phenology modelling induced by cultivar parameterization and its impact on simulated yield. *Agricultural and Forest Meteorology*, *232*, 163–175. https://doi.org/10.1016/j.agrformet.2016.08.013

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. *Agricultural Systems*, *103*(5), 256–264. https://doi.org/10.1016/j.agsy.2010.01.006

Houska, T., Kraft, P., Chamorro-Chavez, A., & Breuer, L. (2015). SPOTting Model Parameters Using a Ready-Made Python Package. *PLOS ONE*, *10*(12), e0145180. https://doi.org/10.1371/journal.pone.0145180

Hunt, J. R., Lilley, J. M., Trevaskis, B., Flohr, B. M., Peake, A., Fletcher, A., … Kirkegaard, J. A. (2019). Early sowing systems can boost Australian wheat yields despite recent climate change. *Nature Climate Change*, *9*(3), 244–247. https://doi.org/10.1038/s41558-019-0417-9

Hussain, J., Khaliq, T., Ahmad, A., & Akhtar, J. (2018). Performance of four crop model for simulations of wheat phenology, leaf growth, biomass and yield across planting dates. *PLOS ONE*, *13*(6), e0197546. https://doi.org/10.1371/journal.pone.0197546

Johnson, S. G. (n.d.). The NLopt nonlinear-optimization package.

Kobayashi, K., & Salam, M. U. (2000). Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal*, *92*, 345–352.

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. *The Plant Journal : For Cell and Molecular Biology*, *76*(5), 742–753. https://doi.org/10.1111/tpj.12326

Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., … Zhu, Y. (2017). Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research*, *202*.

https://doi.org/10.1016/j.fcr.2016.05.001

Maiorano, Andrea, Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., … Zhu, Y. (2016). Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research*. https://doi.org/10.1016/j.fcr.2016.05.001

Mullen, K., Ardia, D., Gil, D., Windover, D., & Cline, J. (2011). "DEoptim": An R Package for Global Optimization by Differential Evolution. *Journal of Statistical Software*, *40*, 1–26.

Piao, S., Liu, Q., Chen, A., Janssens, I. A., Fu, Y., Dai, J., … Zhu, X. (2019). Plant phenology and global climate change: current progresses and challenges. *Global Change Biology*, gcb.14619. https://doi.org/10.1111/gcb.14619

Poeter, E. P., Hill, M. C., Banta, E. R., Mehl, S., & Christensen, S. (2005). *UCODE_2005 and Six Other Computer Codes for Universal Sensitivity Analysis, Calibration, and Uncertainty Evaluation: U.S. Geological Survey Techniques and Methods 6-A11*.

Poeter, Eileen P., & Hill, M. C. (1999). UCODE, a computer code for universal inverse modeling. *Computers & Geosciences*, *25*(4), 457–462. https://doi.org/10.1016/S0098-3004(98)00149-6

Rauff, K. O., & Bello, R. (2015). A Review of Crop Growth Simulation Models as Tools for Agricultural Meteorology. *Agricultural Sciences*, *06*(09), 1098–1105. https://doi.org/10.4236/as.2015.69105

Rezaei, E. E., Siebert, S., & Ewert, F. (2015). Intensity of heat stress in winter wheat—phenology compensates for the adverse effect of global warming. *Environmental Research Letters*, *10*(2), 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. *Scientific Reports*, *8*(1), 4891. https://doi.org/10.1038/s41598-018-23101-2

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York: Wiley .

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? *European Journal of Agronomy*, *94*, 25–35. https://doi.org/10.1016/J.EJA.2018.01.006

Svystun, T., Bhalerao, R. P., & Jönsson, A. M. (2019). Modelling Populus autumn phenology: The importance of temperature and photoperiod. *Agricultural and Forest Meteorology*, *271*, 346–354. https://doi.org/10.1016/J.AGRFORMET.2019.03.003

van Ittersum, M. ., Leffelaar, P. ., van Keulen, H., Kropff, M. ., Bastiaans, L., & Goudriaan, J. (2003). On approaches and applications of the Wageningen crop models. *European Journal of Agronomy*, *18*(3–4), 201–234. https://doi.org/10.1016/S1161-0301(02)00106-5

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, *10*(3). https://doi.org/10.1515/IJNSNS.2009.10.3.273

Vrugt, Jasper A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, *75*, 273–316. https://doi.org/10.1016/J.ENVSOFT.2015.08.013

Wallach, D., Makowski, D., Jones, J. W., & Brun, F. (2019). *Working with Dynamic Crop*

Models: Methods, Tools and examples for Agriculture and Environment. London, U.K.: Academic Press.

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R. P., … Asseng, S. (2017). The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nature Plants*, *3*. https://doi.org/10.1038/nplants.2017.102

Yuan, S., Peng, S., & Li, T. (2017). Evaluation and application of the ORYZA rice model under different crop managements with high-yielding rice cultivars in central China. *Field Crops Research*, *212*, 115–125. https://doi.org/10.1016/J.FCR.2017.07.010

Zadoks, J. C., Chzang, T. T., & Konzak, C. F. (1974). A decimal code for the growth stages of cereals. *Weed Research*, *14*(6), 415–421. https://doi.org/10.1111/j.1365-3180.1974.tb01084.x