



HAL
open science

Multi-model evaluation of phenology prediction for wheat in Australia

Daniel Wallach, Taru Palosuo, Peter Thorburn, Zvi Hochman, Fety Andrianasolo, Senthold Asseng, Bruno Basso, Samuel Buis, Neil Crout, Benjamin Dumont, et al.

► To cite this version:

Daniel Wallach, Taru Palosuo, Peter Thorburn, Zvi Hochman, Fety Andrianasolo, et al.. Multi-model evaluation of phenology prediction for wheat in Australia. *Agricultural and Forest Meteorology*, 2021, 298-299, 10.1016/j.agrformet.2020.108289 . hal-03119039

HAL Id: hal-03119039

<https://hal.inrae.fr/hal-03119039>

Submitted on 26 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 Multi-model evaluation of phenology prediction for wheat in Australia

2 Wallach¹, Daniel; Palosuo², Taru; Thorburn³, Peter; Hochman³, Zvi; Andrianasolo⁴, Fety; Asseng⁵,
3 Senthold; Basso⁶, Bruno; Buis⁷, Samuel; Crout⁸, Neil; Dumont⁹, Benjamin; Ferrise¹⁰, Roberto; Gaiser¹¹, Thomas;
4 Gayler¹², Sebastian; Hiremath¹³, Santosh; Hoek¹⁴, Steven; Horan³, Heidi; Hoogenboom^{5,15}, Gerrit; Huang¹⁶,
5 Mingxia; Jabloun⁸, Mohamed; Jansson¹⁷, Per-Erik; Jing¹⁸, Qi; Justes¹⁹, Eric; Kersebaum^{20,21}, Kurt Christian;
6 Launay²², Marie; Lewan²³, Elisabet; Luo²⁴, Qunying; Maestrini¹⁴, Bernardo; Moriondo²⁵, Marco; Padovan¹⁰,
7 Gloria; Olesen²⁶, Jørgen Eivind; Poyda²⁷, Arne; Priesack²⁸, Eckart; Pullens²⁶, Johannes Wilhelmus Maria; Qian¹⁸,
8 Budong; Schütze²⁹, Niels; Shelia^{5,15}, Vakhtang; Souissi^{30,31}, Amir; Specka²⁰, Xenia; Srivastava¹¹, Amit Kumar;
9 Stella²⁰, Tommaso; Streck¹², Thilo; Trombi¹⁰, Giacomo; Wallor²⁰, Evelyn; Wang¹⁶, Jing; Weber¹², Tobias, K.D.;
10 Weihermüller³², Lutz; de Wit¹⁴, Allard; Wöhling^{29,33}, Thomas; Xiao^{5,34}, Liujun; Zhao⁵, Chuang; Zhu³⁴, Yan;
11 Seidel, Sabine J.¹¹

12

13

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

15 ²Natural Resources Institute Finland (Luke), Helsinki, Finland

16 ³CSIRO Agriculture and Food, Brisbane, Queensland, Australia

17 ⁴ARVALIS - Institut du végétal Paris, France

18 ⁵Agricultural and Biological Engineering Department, University of Florida, Gainesville, Florida

19 ⁶Department of Earth and Environmental Sciences, Michigan State University, East Lansing, Michigan

20 ⁷INRAE, UMR 1114 EMMAH, Avignon, France

21 ⁸School of Biosciences, University of Nottingham, Loughborough, UK

22 ⁹Plant Sciences & TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege,
23 Gembloux, Belgium

24 ¹⁰Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Italy

25 ¹¹Institute of Crop Science and Resource Conservation, University of Bonn, Germany

26 ¹²Institute of Soil Science and Land Evaluation, Biogeophysics, University of Hohenheim, Stuttgart, Germany

27 ¹³Aalto University School of Science, Espoo, Finland

28 ¹⁴Wageningen University & Research, Wageningen, The Netherlands

29 ¹⁵Institute for Sustainable Food Systems, University of Florida, Gainesville, Florida

30 ¹⁶College of Resources and Environmental Sciences, China Agricultural University, Beijing, China

31 ¹⁷Royal Institute of Technology (KTH), Stockholm, Sweden

32 ¹⁸Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada

33 ¹⁹CIRAD, UMR SYSTEM, Montpellier, France

34 ²⁰Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany

35 ²¹Global Change Research Institute CAS, Brno, Czech Republic

36 ²²INRAE, US 1116 AgroClim, Avignon, France

37 ²³Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

38 ²⁴Hillridge Technology Pty Ltd, Sydney, Australia

- 39 ²⁵CNR-IBE, Firenze, Italy
- 40 ²⁶Department of Agroecology, Aarhus University, Tjele, Denmark
- 41 ²⁷Grass and Forage Science / Organic Agriculture, Institute of Crop Science and Plant Breeding, Kiel University,
42 Kiel, Germany
- 43 ²⁸Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research Center for
44 Environmental Health, Neuherberg, Germany
- 45 ²⁹Institute of Hydrology and Meteorology, Chair of Hydrology, Technische Universität Dresden, Dresden,
46 Germany
- 47 ³⁰National Institute of Agronomic Research of Tunisia (INRAT), Agronomy Laboratory, University of Carthage,
48 Tunis, Tunisia
- 49 ³¹National Agronomy Institute of Tunisia (INAT), University of Carthage, Tunis, Tunisia
- 50 ³²Institute of Bio- and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
- 51 ³³Lincoln Agritech Ltd., Hamilton, New Zealand
- 52 ³⁴National Engineering and Technology Center for Information Agriculture, Jiangsu Key Laboratory for
53 Information Agriculture, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing
54 Agricultural University, Nanjing, Jiangsu, China

55 Abstract

56 Predicting wheat phenology is important for cultivar selection, for effective crop
57 management and provides a baseline for evaluating the effects of global change. Evaluating
58 how well crop phenology can be predicted is therefore of major interest. Twenty-eight wheat
59 modeling groups participated in this evaluation. Our target population was wheat fields in the
60 major wheat growing regions of Australia under current climatic conditions and with current
61 local management practices. The environments used for calibration and for evaluation were
62 both sampled from this same target population. The calibration and evaluation environments
63 had neither sites nor years in common, so this is a rigorous evaluation of the ability of modeling
64 groups to predict phenology for new sites and weather conditions. Mean absolute error (MAE)
65 for the evaluation environments, averaged over predictions of three phenological stages and
66 over modeling groups, was 9 days, with a range from 6 to 20 days. Predictions using the multi-
67 modeling group mean and median had prediction errors nearly as small as the best modeling
68 group. About two thirds of the modeling groups performed better than a simple but relevant
69 benchmark, which predicts phenology by assuming a constant temperature sum for each
70 development stage. The added complexity of crop models beyond just the effect of temperature
71 was thus justified in most cases. There was substantial variability between modeling groups
72 using the same model structure, which implies that model improvement could be achieved not
73 only by improving model structure, but also by improving parameter values, and in particular
74 by improving calibration techniques.

75 Keywords: evaluation, phenology, wheat, Australia, structure uncertainty, parameter
76 uncertainty

77

78 1. Introduction

79 Crop phenology describes the cycle of biological events during plant growth. These
80 events include, for example, seedling emergence, leaf appearance, flowering, and maturity.
81 Timing of growing seasons and their critical phases as well as estimates of them are increasingly
82 important in changing climate (Olesen et al., 2012, Dalhaus et al., 2018). Matching the
83 phenology of crop varieties to the climate in which they grow is critical for viable crop
84 production strategies (Rezaei et al., 2018, Hunt et al., 2019). Furthermore, accurate simulation
85 of phenology is essential for models which simulate plant growth and yield (Archontoulis et
86 al., 2014; Boote et al., 2010, 2008).

87 In this study we focus on wheat phenology in Australia. Australia was the world's ninth
88 largest producer of wheat in 2018 and the sixth largest exporter (Workman, 2020). Crop model
89 predictions of phenology have been used in various studies related to wheat production in
90 Australia. In a study by Luo et al. (2018), the APSIM model was used to simulate changes in
91 phenology, water use efficiency, and yield to be expected from global climate change. The
92 APSIM model was used to evaluate changes in wheat phenology in Australia as a result of
93 warming temperatures in recent decades (Sadras and Monzon, 2006). That model was also used
94 to determine the flowering date at each location associated with highest average yield (Flohr et
95 al., 2017).

96 Given the interest in using crop models to predict phenology, it is important to evaluate
97 those predictions. How well can wheat phenology be predicted? In trying to answer this
98 question, one must first define exactly what aspect of the models is being evaluated, and then
99 must choose an appropriate methodology for carrying out the evaluation.

100 It is important to distinguish two different types of model evaluation, which might be
101 termed evaluation of extrapolation predictions and evaluation of interpolation predictions. They

102 differ as to whether or not the data provided for calibration are representative of the target
103 population, i.e. of the range of environments of interest. In one type of study, the objective is
104 to evaluate how well models can extrapolate to conditions not represented in the calibration
105 data. For example, in a multi-model ensemble study on the effect of high temperatures on wheat
106 growth (Asseng et al., 2015), detailed crop measurements were provided for one planting date
107 and the models were evaluated using other planting dates, some with additional artificial heating
108 during growth. The evaluation data thus represented a much larger range of temperatures than
109 represented in the calibration data. This was a test of how well the models can extrapolate to
110 more extreme temperatures than those available for calibration. Other studies have evaluated
111 how well crop models can extrapolate to environments with enhanced CO₂, given calibration
112 data for current ambient CO₂ levels (Biernath et al., 2011).

113 In the second type of study, the calibration data are meant to be representative of the
114 target population. This evaluates how well crop models can generalize from the calibration
115 environments to other similar environments. An example is the study by Ceglar et al. (2019),
116 which used data on wheat phenology under current conditions in Europe for calibration and
117 then predicted phenology for other environments from the same target population. This type of
118 evaluation is adapted, for example, to the case where one has data from a network of variety
119 trials and wants to predict for other sites and years from the same target population, as in Bao
120 et al.. (2017) for yield. It is this aspect of crop phenology models, namely their ability to predict
121 when provided with a sample of data from the target population, that is evaluated in the present
122 study.

123 A second aspect of evaluation that must be specified is the modeling group or groups
124 that are being evaluated, where modeling group refers to the combination of crop model and
125 the people responsible for running the simulations. We reserve the term “model” specifically
126 for model structure, i.e. the model equations, while modeling group determines both the model

127 structure and the parameter values, which are chosen or estimated by the group running the
128 model. It is clear that predictions depend not only on the model structure but also on the
129 parameter values, so evaluation really refers to the modeling group. Model evaluation studies
130 may refer to a particular modeling group or to an ensemble of modeling groups. Here, we
131 evaluate an ensemble of 28 different modeling groups. The purpose is not to give information
132 about each specific modeling group, but rather to evaluate how well currently active modeling
133 groups can predict phenology for our target population (e.g. what is the error of the best
134 predicting group), how well can one expect a modeling group chosen at random to predict (e.g.
135 what is the mean or median prediction error), and what is the variability between modeling
136 groups (e.g. what is the spread between the best and worst predictors).

137 It is important to define precisely the evaluation problem (extrapolation or interpolation,
138 single- or multi-group evaluation), but it is also important that the methodology of evaluation
139 be such as to give reliable results. We focus here on the relation of the predictor (model plus
140 parameter values) and evaluation data. It is well-known from statistics that if a predictor is not
141 independent of the evaluation data, then the error for the evaluation data will in general be less
142 than for new environments (Efron, 1986). That is, non-independence in general leads to
143 underestimating prediction errors. The predictor could depend on the evaluation data if, for
144 example, the evaluation data were also used to calibrate the model, or were used to modify the
145 model equations, or were used to tune site characteristics. If the same sites are present in the
146 calibration and evaluation data, then the model has to some extent been tuned to those sites, and
147 so the predictor is not independent of the evaluation data even if the evaluation data have not
148 been used directly to fit the model. Having the same sites in the calibration and evaluation data
149 is often the case for evaluation studies (Andarzian et al., 2015; Asseng et al., 2008; Chauhan et
150 al., 2019; Hussain et al., 2018; Yuan et al., 2017).

151 There do not seem to have been any evaluation studies of prediction of wheat phenology
152 in Australia based on results from multiple modeling groups, where the calibration data are
153 sampled from the target population (i.e. evaluation of interpolation predictions). The purpose
154 of this study is to present such an evaluation, using a rigorous approach where the parameterized
155 model is independent of the evaluation data.

156 **2. Materials and Methods**

157 **2.1 Experimental data**

158 The data are a subset from a multi-cultivar, multi-location, and multi-sowing date trial
159 for wheat in Australia, described in Lawes et al. (2016). The environments reflect the diversity
160 in the wheat-growing regions of Australia (Fig. 1). Only the data for cultivar Janz, classified as
161 a fast-moderate maturing cultivar, were used here. The data are from 10 sites, located
162 throughout the grain growing region each with one to three sowing years and three planting
163 dates in each year (overall 66 environments, i.e. site-sowing date combinations, Table 1). The
164 sowing dates at each site correspond to early, conventional, and late sowing. Plant density was
165 100-120 plants/m², and sowing depth was 20-35 mm. Nutrients were managed to be non-
166 limiting. There were 1-3 repetitions for each environment (average of 2.1 repetitions).

167



168

169

Figure 1

170 **Location of calibration (red circles) and evaluation (blue triangles) sites across the**
171 **Australian cropping zones (shaded area; Source: Teluguntla et al., 2018).**

172 Plots were visited regularly (about every two weeks) starting soon after emergence of
173 the early sowing and ending after crop maturity, and the Zadoks growth stage (Zadoks et al.,
174 1974), on a scale from 1-100, was determined. Overall, there were 709 combinations of
175 environment and measurement date, with an average of 10.7 stage notations per environment.
176 The stages to be predicted here are stage Z30 (Zadoks stage 30, pseudostem, i.e. youngest leaf
177 sheath erection), stage Z65 (Zadoks stage 65, anthesis half-way, i.e. anthers occurring half way
178 to tip and base of ear), and stage Z90 (Zadoks stage 90, grain hard, difficult to divide). These
179 stages are often used for management decisions or to characterize phenology.

180 In preparing the data for the simulation study, a linear interpolation was performed
181 between each pair of stages, to give the date for every integer Zadoks stage from the first to the
182 last observed stage. At 10 of the 709 measurement dates, observed Zadoks stage decreased
183 slightly (by an average of 3 on the Zadoks scale) compared to the previous date, due to sampling
184 variability. In that case both observed Zadoks stages were replaced by the average for the two

185 dates, before interpolation. The interpolated values were provided in order to avoid different
186 modeling groups using different methods for interpolating the data, which would have added
187 additional uncertainty unrelated to the model performance.

188 The average standard deviation of observed Zadoks stages based on the replicates was
189 0.93 days. The standard deviation of interpolated days after sowing to Z30, Z65, and Z90 was
190 calculated using a bootstrap. For a day with r replicates, a sample of size r was obtained by
191 drawing values at random with replacement, independently for each measurement date. Then
192 the Zadoks values were interpolated as for the original data. This was done 1000 times, giving
193 standard deviations of 1.8 days for observed days to Z30, 0.9 days for observed days to Z65,
194 and 0.5 days for observed days to Z90, respectively.

195 Part of the data was provided to the modeling groups for calibration , and part was never
196 revealed to participants and used for evaluation . The calibration data originated from four sites,
197 two years, and three planting dates, so overall 24 environments. The evaluation data were from
198 six sites, one year, and three planting dates for a total of 18 environments (Table 1). Dates of
199 Z30, Z65 and Z90 were observed at respectively 16, 18 and 5 of these 18 environments. The
200 data were divided in such a way that the calibration and evaluation data had neither sites nor
201 years in common.

202 **Table 1**

203 **Sites and sowing dates for calibration (underlined) and evaluation (bold). Note that**
204 **the calibration and evaluation data have neither sites nor years in common.**

site\ year	2010	2011	2012
Bungunya			2012-05-10
(Queensland)			2012-05-22

			2012-06-23
Corrigin (West Australia)			2012-05-02 2012-05-21 2012-06-21
Eradu (West Australia)	<u>2010-05-14</u> <u>2010-05-27</u> <u>2010-06-22</u>	<u>2011-04-29</u> <u>2011-05-24</u> <u>2011-06-23</u>	
LakeBolac (Victoria)	<u>2010-05-03</u> <u>2010-05-19</u> <u>2010-07-08</u>	<u>2011-05-09</u> <u>2011-06-03</u> <u>2011-06-16</u>	
Minnipa (South Australia)	<u>2010-04-30</u> <u>2010-05-31</u> <u>2010-06-24</u>	<u>2011-05-13</u> <u>2011-05-27</u> <u>2011-06-24</u>	
Nangwee (Queensland)			2012-05-17 2012-05-31 2012-06-23
Spring Ridge (New South Wales)	<u>2010-05-10</u> <u>2010-06-11</u> <u>2010-07-01</u>	<u>2011-05-09</u> <u>2011-06-06</u> <u>2011-06-23</u>	
Temora (New South Wales)			2012-05-05 2012-05-23 2012-06-25
Turretfield (South Australia)			2012-05-30 2012-06-15 2012-07-05

Walpeup (Victoria)			2012-04-27 2012-06-04 2012-07-18
-----------------------	--	--	---

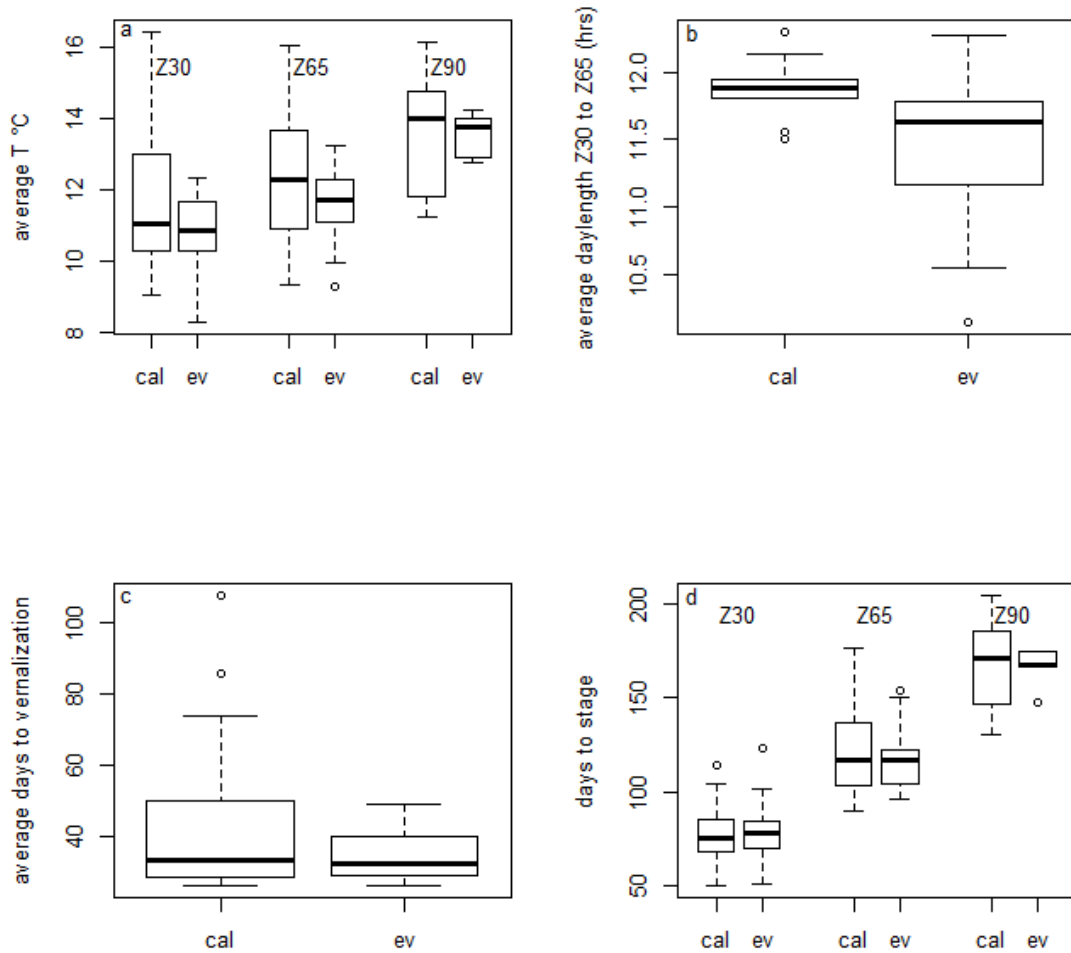
205

206

207

208 To characterize the environments, we calculated for each environment the average
209 temperature from sowing to Z30, Z65, and Z90, the average photoperiod from Z30 to Z65 using
210 the daylength function in the R package *insol* (Corripio, 2019.; R Core Team, 2017) and days
211 to full vernalization using the model in van Bussel et al. (2015) with a required duration of
212 exposure to vernalizing temperatures (V_{sat}) of 25 days, estimated from the figure in their paper.
213 Figure 2 shows the range of average temperature, day length, and days to vernalization for the
214 calibration and evaluation environments as well as the range of observed calendar days to Z30,
215 Z65, and Z90. The range of values for the evaluation data is always within the range of the
216 calibration data, with the single exception of photoperiod. While the median and maximum day
217 lengths were very similar for the two sets of environments, the shortest day length was 11.5
218 hours among calibration environments, while among the evaluation environments the shortest
219 day length was 10.1 hours.

220



221

222

Figure 2

223 **Boxplots of a) average temperatures from sowing to Zadoks stages Z30, Z65, and**

224 **Z90 b) average day length between observed days of Zadoks stages Z30 and Z65 c)**

225 **average days from sowing to complete vernalization d) average days from sowing to**

226 **Zadoks stages Z30, Z65, and Z90. Results are shown separately for the calibration (ca)**

227 **and evaluation (ev) environments. Boxes indicate the lower and upper quartiles. The solid**

228 **line within the box is the median. Whiskers indicate the most extreme data point which is**

229 **no more than 1.5 times the interquartile range from the box, and the outlier dots are those**

230 **observations that are beyond that range.**

231 2.2 Modeling groups

232 Twenty-eight different modeling groups participated in this study, where modeling
233 group refers to the group of people conducting the modeling exercise. Each modeling group is
234 associated with some specific model structure (some specific named model) and also with some
235 specific parameter values. The model structures involved are presented in Supplementary Table
236 S1. Models were considered to have the same structure even if the version number was different,
237 because version differences are expected to be negligible for phenology. Three of the model
238 structures were used by more than one group. Since different groups using the same structure
239 obtained different results, identifying the contributions by the name of the model would be
240 misleading. Furthermore, the performance of specific groups was not of major interest here.
241 Therefore the modeling groups were anonymized, and only identified by a number. There is no
242 model M5 because that group dropped out in the course of the study. The model structures used
243 by more than one group are noted S1 (three groups), S2 (three groups) and S3 (two groups).

244 Details about the way phenology is modeled by each model structure can be found in
245 the references for each model (Supplementary Table S1). Here we give only a brief overview.
246 The principal factors that affect winter wheat developmental rate are temperature, day length
247 and degree of vernalization (Johnen et al., 2012). Most, but not all, model structures take into
248 account all three factors. The simplest approach to modeling the effect of temperature is to
249 assume that development rate increases linearly with daily average temperature above some
250 base temperature (a parameter). In other models the rate may be constant above some optimal
251 temperature (a parameter), development rate may decline above the optimum temperature at
252 some rate (a parameter), or development rate may be some more complex function of
253 temperature (Kumudini et al., 2014; Wang et al., 2017). The parameters of the temperature
254 response curve may differ depending on development stage. The effect of photoperiod on
255 development rate is often modeled as a multiplier that is a piecewise linear function of

256 photoperiod. The function increases with some slope (a parameter) up to a threshold
257 photoperiod (a parameter), and then is 1 for photoperiods longer than the threshold.
258 Vernalization, which must be accomplished before the plant can flower, requires a period of
259 cold temperatures. Vernalization parameters can include the upper limit for temperature to
260 count as vernalizing, and the required number of vernalizing days. Some models also relate
261 development to the rate of leaf appearance (called the phyllochron, a parameter) or rate of
262 tillering. Finally, several models also take into account the effect of cold or drought stress on
263 development rate. If drought stress is taken into account, then development rate is related to all
264 the processes that determine soil moisture and plant water uptake.

265 The multi-model ensemble here was an “ensemble of opportunity” meaning that any
266 modeling group that asked to join was accepted. The activity was announced on the list server
267 of the Agricultural Modeling Inter-comparison and Improvement Project (AgMIP) and on the
268 list servers of several models. In addition to the original models, we defined two ensemble
269 models. The model e-mean has predictions equal to the mean of the simulated values. The
270 model e-median has predictions equal to the median of the simulated values.

271 2.3 Simulation experiment

272 Each participating modeling group was provided with weather, soil, and management
273 data for all environments, as well as all available observed and interpolated values for days to
274 each Zadoks stage for the calibration data. Participants were requested to return simulated
275 values for number of days from sowing to emergence (even though days to emergence was
276 never observed) and values for number of days from sowing to stages Z30, Z65, and Z90 for
277 all environments, including both the calibration environments and the evaluation environments.

278 2.4 Evaluation

279 As our basic metric of model error, we use the mean absolute error (MAE). For a model
280 m , MAE is

$$281 \quad MAE_m = (1/n) \sum_{i=1}^n |y_i - \hat{y}_{i,m}| \quad (1)$$

282 where y_i is the observed value for environment i and $\hat{y}_{i,m}$ is the value simulated by modeling
283 group m for that environment. The sum is over either calibration environments, to evaluate
284 goodness-of-fit, or over evaluation environments, to estimate prediction error. This is
285 preferred over mean squared error (MSE) or root mean squared error (RMSE), because unlike
286 MSE, MAE does not give extra weight to large errors (Willmott and Matsuura, 2005). To test
287 whether MAE is the same for prediction of days to different stages, we used the R function
288 `pairwise.t.test`, with `method="holm"` to correct for multiple comparisons. We also calculated
289 MSE, RMSE, and NRMSE (normalized root mean squared error) for comparison with other
290 studies.

$$291 \quad \begin{aligned} MSE_m &= (1/n) \sum_{i=1}^n (y_i - \hat{y}_{i,m})^2 \\ RMSE_m &= \sqrt{MSE_m} \\ NRMSE_m &= RMSE_m / \bar{y} \end{aligned} \quad (2)$$

292 where \bar{y} is the average of the observed values.

293 We considered two skill measures. A skill measure compares prediction error of the
294 modeling group to be evaluated with the error of a simple model used for comparison. We
295 define two simple models, and therefore two skill measures. Both use MSE, rather than MAE,
296 as the measure of model error, in keeping with usual practice. The first simple model, noted
297 “naive”, predicts that days to each stage will be equal to the average number of days to that

298 stage in the calibration data. The predictions of the naïve model here are 77.1, 123.1, and 166.5
299 days from sowing to stages Z30, Z65, and Z90, respectively. The first skill measure, modeling
300 efficiency (EF), is defined as

$$301 \quad EF_m = 1 - MSE_m / MSE_{naive} \quad (3)$$

302 The naïve model ignores all variability and predicts that days to any stage will be the same
303 regardless of the environment. A model with $EF \leq 0$ is a model that does no better than the
304 naïve model, and so would be considered a very poor predictor. A perfect model, with no error,
305 has modeling efficiency of 1. Often modeling efficiency is based on the fit of a calibrated model
306 to the data used for calibration (McCuen et al., 2006). Here, in contrast, the naïve model is
307 based on calibration data and used to predict for independent data.

308 The naïve model is a very low baseline for evaluating a crop model. We therefore
309 introduce a more realistic, but still simple model which takes into account the effect of
310 temperature on phenology. This “onlyT” model predicts that degree days ($^{\circ}\text{D}$) from sowing to
311 each stage will be equal to the number of degree days from sowing to that stage in the calibration
312 data, where degree days on any calendar day is equal to average temperature that day. The
313 predictions of the onlyT model are that Z30 will occur 893.7 $^{\circ}\text{D}$ after sowing, Z65 will occur
314 1476.0 $^{\circ}\text{D}$ after sowing, and Z90 will occur 2245.7 $^{\circ}\text{D}$ after sowing. The second skill measure,
315 noted skillT, is then

$$316 \quad skillT_m = 1 - MSE_m / MSE_{onlyT} \quad (4)$$

317 where MSE_{onlyT} is MSE for the onlyT model. As for any skill measure, a perfect model has
318 $skillT = 1$ and a model that does no better than the onlyT model has $skillT \leq 0$

319 2.5 Sources of variability

320 A major interest of ensemble studies is that they provide information on the variability
321 in simulation results between different modeling groups. This variability can arise from
322 differences in model structure between different modeling groups or differences in parameter
323 values for groups that use the same model structure. In this study, three of the model structures
324 are used by more than one modeling group. This makes it possible to estimate separately the
325 variance in simulated values due to structure and the variance due to modeling group nested
326 within structure (i.e. due to differences in parameter values). We treat the simulated values as a
327 sample from the distribution of plausible model structures and plausible parameter values.
328 According to the law of total variance (Casella and Berger, 1990), the total variance of
329 simulated values can be decomposed into two parts as

$$330 \quad \text{var}(\hat{y}) = \text{var}[E(\hat{y} | S)] + E[\text{var}(\hat{y} | S)] \quad (5)$$

331 where \hat{y} are the simulated values, S is model structure, E is the expectation, var is the variance,
332 and the notation $|S$ means that the expectation (in the first term on the right hand side) or the
333 variance (in the second term on the right hand side) is taken separately for each value of model
334 structure. We estimated the first term by first calculating the average simulated value for each
335 structure (if a structure is represented by a single modeling group, this is just the value simulated
336 by that group), and then calculating the variance of those average values. This is the between-
337 structure variability. To estimate the second term, we first calculated the variance between
338 simulated values for each of the three structures with multiple groups. Then we calculated the
339 average of those variances. This is the within-structure variability (i.e. variability due to
340 parameters).

341 3.Results

342 3.1 Prediction error and skill

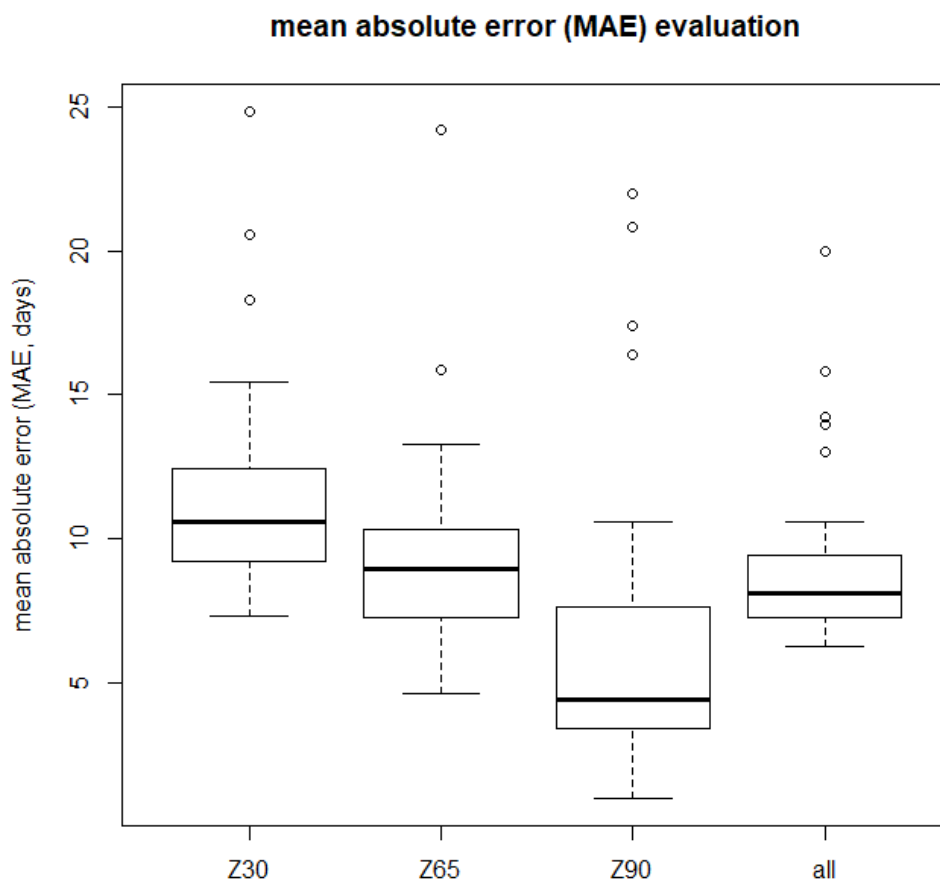
343 MAE values for the evaluation data are shown in Figure 3 and summarized in Table 2.
344 Results for individual modeling groups are given in Supplementary Table S2. Median MAE
345 values (and ranges) were 12 days (8-25 days) for days to Z30, 10 days (5-24 days) for days to
346 Z65, and 7 days (1-22 days) for days to Z90. The median (and range) of MAE averaged over
347 the three stages was 9 days (6-20 days). The ensemble predictors e-mean and e-median both
348 had averaged MAE values of 7 days. They were both only marginally worse than the best two
349 individual modeling groups, and e-median was marginally better than e-mean. For comparison
350 with other studies, we also report other criteria of error in Table 2.

351 **Table 2**

352 **Summary of prediction errors for the evaluation and calibration environments,**
353 **in each case averaged over predictions of days to stages Z30, Z65, and Z90 except for**
354 **NRMSE, where the values refer to predictions of number of days to stage Z65. The**
355 **median, minimum, and maximum error over modeling groups are shown.**

		median	minimum	maximum
Evaluation data	MAE (days)	9	6	20
	RMSE (days)	12	9	25
	NRMSE	0.094	0.056	0.227
	EF	0.51	-1.51	0.70
	skillT	0.2	-3.34	0.49
Calibration data	MAE (days)	8	6	19
	RMSE (days)	11	6	24

	NRMSE	0.068	0.041	0.197
--	-------	-------	-------	-------



356

357

Figure 3

358 **Boxplot of mean absolute error (days) for each development stage and averaged**

359 **over stages, for the evaluation data. The variability is between different modeling groups.**

360 **Boxes indicate the lower and upper quartiles. The solid line within the box is the median.**

361 **Whiskers indicate the most extreme data point which is no more than 1.5 times the**

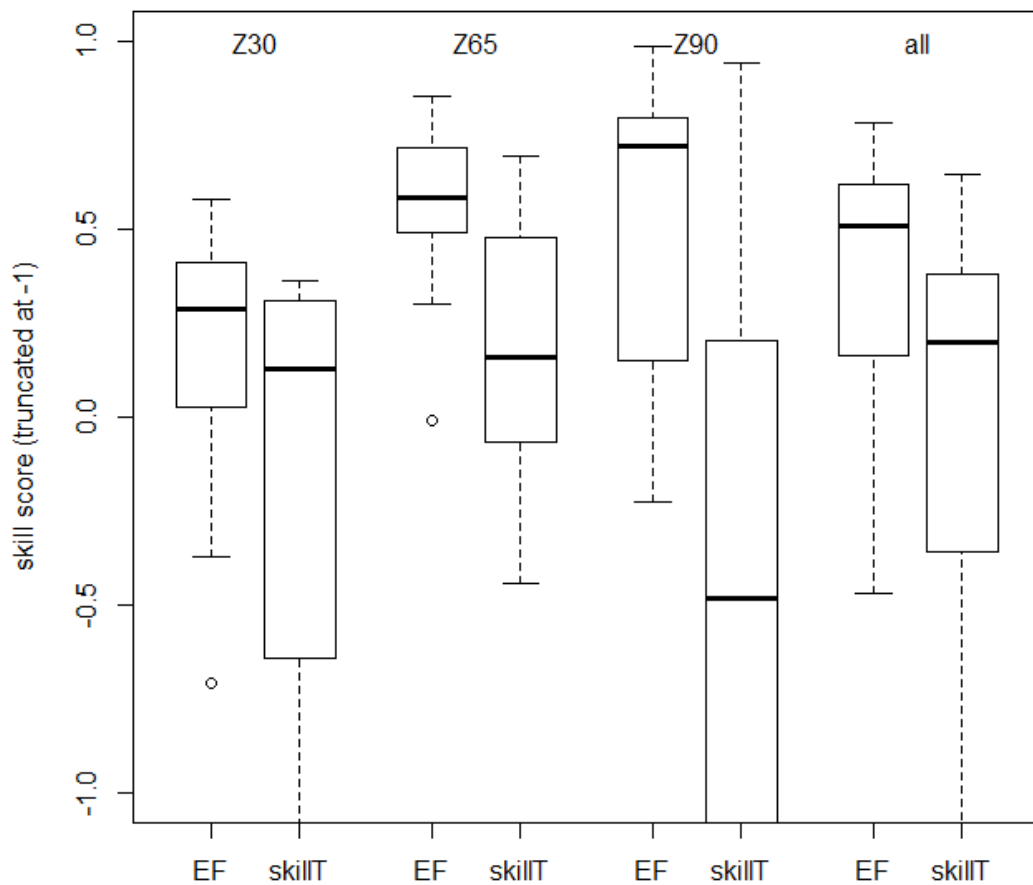
362 **interquartile range from the box, and the outlier dots are those observations that are**

363 **beyond that range.**

364

365

366 Boxplots of EF and skillT for the evaluation data are shown in Figure 4. The median
367 EF value of the individual modeling groups, averaged over stages, was 0.51, and 86 % of the
368 modeling groups had $EF > 0$. The median skillT value of the individual modeling groups,
369 averaged over stages, was 0.20, and 68% of the modeling groups had $skillT > 0$.



370

371

Figure 4

372

373

374

375

Boxplots of skill scores for prediction of days to Zadoks stages Z30, Z65, and Z90, and averaged over stages (all) for the evaluation data. Skill score is 1 for a modeling group that predicts perfectly, and is less than or equal to 0 for a modeling group that does no better than using average days to each stage in the calibration data (EF skill score) or than

376 **using the average number of degree days to each stage in the calibration data (skillT skill**
377 **score). Boxes indicate the lower and upper quartiles. The solid line within the box is the**
378 **median. Whiskers indicate the most extreme data point which is no more than 1.5 times**
379 **the interquartile range from the box, and the outlier dots are those observations that are**
380 **beyond that range. For readability the y axis is cut off at -1.**

381

382 Overall MAE for the evaluation data and the calibration data for the same modeling
383 group were correlated. The calibration value explains 46 % of the variability in the evaluation
384 data ($R^2 = 0.46$).

385

386 3.2 Sources of variability

387 There was substantial variability between modeling groups for each individual
388 prediction, including between modeling groups that share the same model structure
389 (Supplementary Figure S1). Averaged over the evaluation environments and over all three
390 stages Z30, Z65, and Z90, the estimated within-structure standard deviation was 4.3 days and
391 the estimated between-structure standard deviation was 11.9 days, so the within-structure
392 standard deviation was 36 % as large as the between-structure standard deviation.

393

394 4. Discussion

395 4.1 Comparison of calibration and evaluation environments

396 The calibration and evaluation environments were drawn from the same target
397 population, namely wheat crops in the major wheat growing regions in Australia, with current

398 climate and local management practices. We compared the calibration and evaluation
399 environments for the main characteristics that are likely to affect phenology, namely
400 temperature, day length, and accumulation of vernalizing temperatures. Temperatures and
401 vernalizing durations of the evaluation environments were within the ranges of the calibration
402 environments, but the evaluation data had a larger range of day lengths than the calibration data.
403 This is the result of sampling variability, and may have led to larger prediction errors than if
404 the calibration data had a range of day lengths comparable to that of the evaluation data.
405 However, the range of days to each phenology stage for the evaluation data was always within
406 the range for the calibration data. We conclude that this study represents a case where the
407 calibration and evaluation data represent a similar range of conditions (with the caveat just
408 mentioned concerning photoperiod). This type of situation is of particular importance, for
409 example, where one wants to calibrate a crop model using current conditions and subsequently
410 test possible sowing dates within a limited range, or to compare phenology of multiple potential
411 cultivars at specific sites within the calibration domain.

412 4.2 Prediction error

413 The evaluation here was based on data which had neither sites nor years in common
414 with the calibration data. This was thus a rigorous estimate of how well crop modeling groups
415 can predict wheat phenology for unseen sites and weather, when provided with calibration data
416 sampled from the target population. The median MAE among models averaged over phenology
417 stages was 9 days, which was substantially larger than the standard deviation of observed stages,
418 which was in the range 1-2 days. The best modeling group had an average MAE of 7 days,
419 which was still substantially larger than the standard deviation of observed stages. MAE values
420 were significantly larger for prediction of days to Z30 than for prediction of days to later Zadoks
421 stages. This may be due to the large variability between groups in predicting time to emergence,

422 which is discussed in more detail below. Time to emergence is a major part of the time to Z30,
423 but a smaller fraction of time to Z65 or Z90.

424 Chauhan et al. (2019) reported a value of NRMSE of 0.062 for prediction of time to
425 flowering of wheat in Australia, for a version of APSIM taking the effect of water stress on
426 phenology into account. In that study, the model was adjusted to some extent to the data used
427 for evaluation, so the reported error probably underestimates the error for new environments.
428 That reported value was in any case within the range of NRMSE values found for different
429 modeling groups here, for both the evaluation data (NRMSE here from 0.056 to 0.227) and the
430 calibration data (NRMSE here from 0.041 to 0.197). Asseng et al. (2008), using the APSIM
431 model, found RMSE of 4 days for wheat phenology predictions (mostly predictions of days to
432 anthesis) for 44 different environments in Western Australia, a level of error which was smaller
433 than the minimum RMSE of 9 days found here for the evaluation data, and even smaller than
434 the minimum RMSE of 6 days found here for the calibration data. In that study, the phenology
435 model was again adjusted to some extent to the data (S. Asseng, 2020, pers. comm.), which
436 could explain the smaller errors.

437 The above comparisons suggest that prediction errors are very roughly similar between
438 studies, but that there are differences depending on the details of the prediction problem and
439 the way prediction error is evaluated. It is clearly useful to build up a knowledge base
440 concerning phenology prediction error, as a baseline for comparison for future studies or even
441 as a default value if evaluation is not done. Contributions to the knowledge base will be all the
442 more useful, to the extent that the details of the prediction problem are clearly specified
443 (including whether it is of type interpolation or extrapolation and including a characterization
444 of the target population) and to the extent that the evaluation has a rigorous separation between
445 the predictor and the evaluation data. The present study should therefore be a valuable
446 contribution to such a knowledge base.

447 It is of interest to compare the results here with those from a study structured like the
448 present study (calibration and evaluation environments with similar characteristics, evaluation
449 data not used for model development or tuning) but where the evaluation concerned prediction
450 of two phenological stages of wheat in France, namely BBCH30 (equivalent to Z30) and
451 BBCH55 (equivalent to Z55) (Wallach et al., 2019). To a large extent, the same modeling
452 groups participated in both studies. Specifically, the French study included 27 different
453 modeling groups, 26 of which participated in the present study. A comparison between the two
454 studies gives an indication of variability in prediction error for the same modeling groups but
455 for different target populations (Australian wheat in one case, French wheat in the other) and
456 for somewhat different calibration data and predicted stages.

457 MAE averaged over the evaluation environments and over predicted stages ranged from
458 3 to 13 days (median 6 days) for the French data compared to 6 to 20 days (median 9 days) for
459 the Australian data. The target population (wheat fields in Australia versus wheat fields in
460 France) thus had a substantial effect on prediction errors. A detailed analysis of the underlying
461 reasons for the larger errors in Australia is beyond the scope of this study. However, one
462 possible contributing cause is the simulation of time to emergence. The average simulated time
463 to emergence for all French environments was 10 days after sowing, and the mean standard
464 deviation between modeling groups was 4 days. The corresponding values for the Australian
465 environments were a mean emergence time of 15 days after sowing, and a mean standard
466 deviation between modeling groups of 18 days. This very large standard deviation for the
467 Australian environments, pointing at major differences between modeling groups, may be due
468 to dry conditions in some environments and the uncertainty regarding initial soil conditions,
469 leading some models to simulate very long times to emergence (up to 107 days, Supplementary
470 Figure S1). This suggests that for Australian environments, it would be valuable to have
471 observations of time to emergence for calibration. It seems that for many modeling groups, it

472 would be worthwhile to revisit the predictions of time to emergence under conditions like those
473 of the Australian environments, taking advantage of specific modeling studies of time to
474 emergence for wheat (Lindstrom et al., 1976; Wang et al., 2009).

475 An important question in modeling is whether the same modeling groups perform best
476 for all target populations, or whether different groups are best for different target populations.
477 There is quite a bit of scatter in the graph of MAE for the Australian versus French environments
478 (Supplementary Fig. S2), but the rank correlation between the two (Kendall's tau) is 0.31, which
479 is statistically significant ($p=0.013$). This suggests that there are modeling groups which
480 perform better than others over a wide range of environments. Once again, it is prudent to repeat
481 that this applies to the case where calibration is based on environments that are sampled from
482 the target distribution. Prediction errors for extrapolation to conditions very different than those
483 of the calibration data might behave very differently.

484 4.3 Skill measures

485 While prediction error is of course of interest, skill scores may be even more useful, as
486 they indicate how models compare to alternative methods of prediction. Note that the EF skill
487 score used here is somewhat different than the usual definition. Here, the naïve model is based
488 solely on the calibration data, so this is in fact a feasible predictor. The more usual definition
489 of the naïve model is the mean of all the data, including the data used for evaluation. Overall,
490 all except four modeling groups had smaller MSE (were better predictors) than the naïve model.

491 The EF criterion is a rather low baseline for evaluating the usefulness of crop models
492 for predicting phenology. Our second skill measure compares model MSE and MSE of the
493 onlyT model, which assumes a constant number of degree days from sowing to each Zadoks
494 stage, and estimates that number based on the calibration data. This should be a better predictor
495 than the naïve model if photoperiod and vernalization effects are limited, and so is a more

496 stringent test of usefulness of process models. We found that the onlyT model was indeed a
497 better predictor than the naïve model. Nonetheless, 19 of the modeling groups performed better
498 than the onlyT model. It seems that in most cases here, the added complexity in crop models
499 beyond a simple sum of degree days is warranted. More generally, we suggest that
500 systematically calculating a skill measure like skillT would give valuable information about the
501 usefulness of more complex models.

502 **4.4 Model averaging**

503 As found in many studies, e-median and e-mean had prediction errors comparable to
504 the best modeling groups. This confirmed previous evidence and theoretical considerations
505 showing that the use of e-mean or e-median is often a good strategy (Bassu et al., 2014; Palosuo
506 et al., 2011; Rötter et al., 2012; Wallach et al., 2018). The e-mean model is based on a simple
507 average over simulated values, so the results from every modeling group are weighted equally.
508 An open question in using model ensembles is whether it would be better to give more weight
509 to models that have smaller prediction errors for the calibration data (Christensen et al., 2010),
510 for example using Bayesian Model Averaging (Wöhling et al., 2015). The results here show
511 that phenology predictive performance for the calibration environments is significantly
512 correlated with predictive performance for new environments. This was also found to be the
513 case for a study evaluating phenology prediction by modeling groups based on phenology in
514 French environments (Wallach et al., 2019) and suggests that in these cases, it may be
515 worthwhile to use performance-weighted model ensembles. This may be due to the fact that in
516 these studies, the calibration and evaluation environments were similar to one another. In cases
517 where one is extrapolating to conditions quite different than those represented by the calibration
518 environments, performance weighting may be less useful. This once again emphasizes that it is
519 important to define for each evaluation study whether it is an evaluation of type “interpolation”
520 or “extrapolation”.

521 4.5 Sources of variability

522 A major outcome of model ensemble studies is the variability in simulated values
523 between modeling groups, which is an indication of the uncertainty of model-based predictions
524 (Asseng et al., 2013). Beyond a measure of the variability, it is of interest to understand the
525 origins of the variability. One important aspect here is how differences in the model equations
526 between model structures affect the simulated values. This however is difficult to untangle,
527 given the multiple differences between structures. It seems that specific studies, for example
528 modifying one specific aspect of multiple models, are needed to understand the various sources
529 of structure uncertainty (Maiorano et al., 2016). The present study does not allow us to relate
530 specific differences in model structure to differences in simulated results. However, it does
531 allow us to separate two contributions to variability, namely the overall variability between
532 model structures and the variability between different parameter values for the same model
533 structure. An important question is the relative importance of the two, to determine priorities
534 for reducing overall uncertainty. Parameter uncertainty can arise from uncertainty in the default
535 values of those parameters that are fixed, from uncertainty in the choice of calibration approach
536 (for example, the form of the objective function or the choice of parameters to estimate) and
537 from the values of the estimated parameters, which are uncertain because there is always a
538 limited amount of data. The within-structure variability here is a measure of the uncertainty due
539 to choice of default values and calibration approach, but not of uncertainty in the values of the
540 calibrated parameters. The within-structure standard deviation here is 4.3 days, compared to a
541 between-structure standard deviation (contribution of structure) of 11.9 days. The study based
542 on French environments found a within-structure standard deviation of 5.6 days and a between-
543 structure standard deviation of 8.0 days (Wallach et al., 2019). Confalonieri et al. (2016) also
544 found that the within-structure effect was in general, but not in all cases, smaller than the
545 between-structure effect on variability.

546 Other studies have on the contrary focused on structural uncertainty versus uncertainty
547 in the calibrated parameters, without taking into account uncertainty in all the default parameter
548 values, nor uncertainty in the calibration approach chosen. Zhang et al. (2017) found that model
549 structure explained about 80 % of the variability in simulated time to heading in rice and about
550 92 % of the variability in simulated time to maturity in rice, the remainder of the variability
551 being due to parameter uncertainty. Wallach et al. (2017) found that model structure uncertainty
552 contributed about twice as much variance as parameter uncertainty to overall simulation
553 variance. It would be of interest to have a fuller treatment of parameter uncertainty, including
554 both different groups using the same model structure and an estimate of the uncertainty in the
555 parameters estimated by each group.

556 5. Conclusions

557 We evaluated how well 28 crop modeling groups simulate wheat phenology in
558 Australia, in the case where both the calibration data and the evaluation data were sampled from
559 fields in the major wheat growing areas in Australia under current climate and local
560 management. It is important to distinguish between interpolation type prediction, as here, and
561 extrapolation type, since they are not evaluating the same properties of modeling groups. It is
562 also important to emphasize that evaluation concerns both model structure and parameter
563 values, and therefore the modeling group and not just the underlying model structure. MAE for
564 the evaluation data here ranged from 6 to 20 days depending on the modeling group, with a
565 median of 9 days. About two thirds of the modeling groups performed better than a simple but
566 relevant benchmark, which predicts phenology assuming a constant temperature sum for each
567 development stage. The added complexity of crop models beyond just the effect of temperature
568 is therefore justified in most cases. As found in many other studies, the multi-modeling group
569 mean and median had prediction errors nearly as small as the best modeling group, suggesting
570 that using these ensemble predictors is a good strategy. Prediction errors for calibration and

571 evaluation environments were found to be significantly correlated, which suggests that for
572 interpolation type studies, it would be of interest to test ensemble predictors that weight
573 individual models based on performance for the calibration data. The variability due to
574 modeling group for a given model structure, which reflects part of parameter uncertainty, was
575 found to be smaller than the variability due to model structure, but was not negligible. This
576 implies that model improvement could be achieved not only by improving model structure but
577 also by improving parameter values.

578

579 **Acknowledgements**

580 This work was in part supported by the Collaborative Research Center 1253 CAMPOS
581 (Project 7: Stochastic Modelling Framework), funded by the German Research Foundation
582 (DFG, Grant Agreement SFB 1253/1 2017), the Academy of Finland through projects AI-
583 CropPro (316172 and 315896) and DivCSA (316215) and Natural Resources Institute Finland
584 (Luke) through a strategic project BoostIA, the BonaRes projects "Soil3" (BOMA 03037514)
585 and "IAS" (031B0513I) of the Federal Ministry of Education and Research (BMBF),
586 Germany, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under
587 Germany's Excellence Strategy - EXC 2070 – 390732324, the project BiomassWeb of the
588 GlobeE programme (Grant number: FKZ031A258B) funded by the Federal Ministry of
589 Education and Research (BMBF, Germany), the EU funded SustEs project
590 (CZ.02.1.01/0.0/0.0/16_019/0000797), the INRA ACCAF meta-programme, the German
591 Federal Ministry of Education and Research (BMBF) in the framework of the funding
592 measure "Soil as a Sustainable Resource for the Bioeconomy – BonaRes", project "BonaRes
593 (Module B): BonaRes Centre for Soil Research, subproject B" (grant 031B0511B), the
594 National Key Research and Development Program of China (2017YFD0300205), the

595 National Science Foundation for Distinguished Young Scholars (31725020), the Priority
596 Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the 111
597 project (B16026), and China Scholarship Council, the Agriculture and Agri-Food Canada's
598 Project 1387 under the Canadian Agricultural Partnership, the DFG Research Unit FOR 1695
599 'Agricultural Landscapes under Global Climate Change – Processes and Feedbacks on a
600 Regional Scale, the U.S. Department of Agriculture National Institute of Food and
601 Agriculture (award no. 2015-68007-23133) and USDA/NIFA HATCH grant N. MCL02368,
602 the National Key Research and Development Program of China (2016YFD0300105), The
603 Broadacre Agriculture Initiative, a research partnership between University of Southern
604 Queensland and the Queensland Department of Agriculture and Fisheries, the JPI FACCE
605 MACSUR2 project, funded by the Italian Ministry for Agricultural, Food, and Forestry
606 Policies (D.M. 24064/7303/15 of 26/Nov/2015). The field work was jointly funded by CSIRO
607 and the Grains Research and Development Corporation (GRDC) under the "Adding Value to
608 GRDC's National Variety Trial Network" project (CSA00027). The order in which the donors
609 are listed is arbitrary

610

611 References

- 612 Andarzian, Bahram, Hoogenboom, G., Bannayan, M., Shirali, M., Andarzian, Behnam, 2015.
613 Determining optimum sowing date of wheat using CSM-CERES-Wheat model. *J. Saudi*
614 *Soc. Agric. Sci.* 14, 189–199. <https://doi.org/10.1016/J.JSSAS.2014.04.004>
- 615 Archontoulis, S. V., Miguez, F.E., Moore, K.J., 2014. A methodology and an optimization
616 tool to calibrate phenology of short-day species included in the APSIM PLANT model:
617 Application to soybean. *Environ. Model. Softw.* 62, 465–477.
618 <https://doi.org/10.1016/j.envsoft.2014.04.009>
- 619 Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J.,
620 Thorburn, P.J., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P.,
621 Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A.J., Doltra, J., Gayler,
622 S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurrealde,
623 R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O’Leary, G., Olesen,
624 J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak,
625 I., Steduto, P., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M.,
626 Waha, K., Wallach, D., White, J.W., Williams, J.R., Wolf, J., 2013. Uncertainty in
627 simulating wheat yields under climate change. *Nat. Clim. Chang.* 3, 827–832.
628 <https://doi.org/10.1038/nclimate1916>
- 629 Asseng, S., Keating, B.A., Fillery, I.R.P., Gregory, P.J., Bowden, J.W., Turner, N.C., Palta,
630 J.A., Abrecht, D.G., 2008. Performance of the APSIM-wheat model in Western
631 Australia. *F. Crop. Res.* 57, 163–179.
- 632 Bao, Y., Hoogenboom, G., McClendon, R., Vellidis, G., 2017. A comparison of the
633 performance of the CSM-CERES-Maize and EPIC models using maize variety trial data.
634 *Agric. Syst.* 150, 109–119. <https://doi.org/10.1016/J.AGSY.2016.10.006>

- 635 Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C.,
636 Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S.,
637 Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S.,
638 Izaurrealde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.-H., Kumar,
639 N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F.,
640 Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize
641 crop models vary in their responses to climate change factors? *Glob. Chang. Biol.* 20,
642 2301–20. <https://doi.org/10.1111/gcb.12520>
- 643 Biernath, C., Gayler, S., Bittner, S., Klein, C., Högy, P., Fangmeier, A., Priesack, E., 2011.
644 Evaluating the ability of four crop models to predict different environmental impacts on
645 spring wheat grown in open-top chambers. *Eur. J. Agron.* 35, 71–82.
646 <https://doi.org/10.1016/j.eja.2011.04.001>
- 647 Boote, K.J., Jones, J.W., Hoogenboom, G., 2008. Crop simulation models as tools for agro-
648 advisories for weather and disease effects on production. *J. Agrometeorol.* 10, 9–17.
- 649 Boote, K.J., Jones, J.W., Hoogenboom, G., White, J.W., 2010. The Role of Crop Systems
650 Simulation in Agriculture and Environment. *Int. J. Agric. Environ. Inf. Syst.* 1, 41–54.
- 651 Casella, G., Berger, R.L., 1990. *Statistical Inference*. Wadsworth and Brooks, Pacific Grove,
652 CA.
- 653 Ceglar, A., van der Wijngaart, R., de Wit, A., Lecerf, R., Boogaard, H., Seguni, L., van den
654 Berg, M., Toreti, A., Zampieri, M., Fumagalli, D., Baruth, B., 2019. Improving
655 WOFOST model to simulate winter wheat phenology in Europe: Evaluation and effects
656 on yield. *Agric. Syst.* 168, 168–180. <https://doi.org/10.1016/J.AGSY.2018.05.002>
- 657 Chauhan, Y.S., Ryan, M., Chandra, S., Sadras, V.O., 2019. Accounting for soil moisture
658 improves prediction of flowering time in chickpea and wheat. *Sci. Rep.* 9, 7510.

- 659 <https://doi.org/10.1038/s41598-019-43848-6>
- 660 Christensen, J., Kjellström, E., Giorgi, F., Lenderink, G., Rummukainen, M., 2010. Weight
661 assignment in regional climate models. *Clim. Res.* 44, 179–194.
662 <https://doi.org/10.3354/cr00916>
- 663 Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., Pagani, V.,
664 Cappelli, G., Vertemara, A., Alberti, L., Alberti, P., Atanassiu, S., Bonaiti, M.,
665 Cappelletti, G., Ceruti, M., Confalonieri, A., Corgatelli, G., Corti, P., Dell’Oro, M.,
666 Ghidoni, A., Lamarta, A., Maghini, A., Mambretti, M., Manchia, A., Massoni, G., Mutti,
667 P., Pariani, S., Pasini, D., Pesenti, A., Pizzamiglio, G., Ravasio, A., Rea, A., Santorsola,
668 D., Serafini, G., Slavazza, M., Acutis, M., 2016. Uncertainty in crop model predictions:
669 What is the role of users? *Environ. Model. Softw.* 81, 165–173.
670 <https://doi.org/10.1016/j.envsoft.2016.04.009>
- 671 Corripio, J.G., n.d. *insol: Solar Radiation*. R package version 1.2. 2019.
- 672 Efron, B., 1986. How Biased is the Apparent Error Rate of a Prediction Rule? *J. Am. Stat.*
673 *Assoc.* 81, 461–470. <https://doi.org/10.1080/01621459.1986.10478291>
- 674 Flohr, B.M., Hunt, J.R., Kirkegaard, J.A., Evans, J.R., 2017. Water and temperature stress
675 define the optimal flowering period for wheat in south-eastern Australia. *F. Crop. Res.* v.
676 209, 108–119. <https://doi.org/10.1016/j.fcr.2017.04.012>
- 677 Hussain, J., Khaliq, T., Ahmad, A., Akhtar, J., 2018. Performance of four crop model for
678 simulations of wheat phenology, leaf growth, biomass and yield across planting dates.
679 *PLoS One* 13, e0197546. <https://doi.org/10.1371/journal.pone.0197546>
- 680 Johnen, T., Boettcher, U., Kage, H., 2012. A variable thermal time of the double ridge to flag
681 leaf emergence phase improves the predictive quality of a CERES-Wheat type

- 682 phenology model. *Comput. Electron. Agric.* 89, 62–69.
683 <https://doi.org/10.1016/J.COMPAG.2012.08.002>
- 684 Kumudini, S., Andrade, F.H., Boote, K.J., Brown, G.A., Dzotsi, K.A., Edmeades, G.O.,
685 Gocken, T., Goodwin, M., Halter, A.L., Hammer, G.L., Hatfield, J.L., Jones, J.W.,
686 Kemanian, A.R., Kim, S.-H., Kiniry, J., Lizaso, J.I., Nendel, C., Nielsen, R.L., Parent,
687 B., Stöckle, C.O., Tardieu, F., Thomison, P.R., Timlin, D.J., Vyn, T.J., Wallach, D.,
688 Yang, H.S., Tollenaar, M., 2014. Predicting maize phenology: Intercomparison of
689 functions for developmental response to temperature. *Agron. J.* 106, 2087–2097.
690 <https://doi.org/10.2134/agronj14.0200>
- 691 Lawes, R.A., Huth, N.D., Hochman, Z., 2016. Commercially available wheat cultivars are
692 broadly adapted to location and time of sowing in Australia’s grain zone. *Eur. J. Agron.*
693 77, 38–46. <https://doi.org/10.1016/J.EJA.2016.03.009>
- 694 Lindstrom, M.J., Papendick, R.I., Koehler, F.E., 1976. A Model to Predict Winter Wheat
695 Emergence as Affected by Soil Temperature, Water Potential, and Depth of Planting1.
696 *Agron. J.* 68, 137–141. <https://doi.org/10.2134/agronj1976.00021962006800010038x>
- 697 Luo, Q., O’Leary, G., Cleverly, J., Eamus, D., 2018. Effectiveness of time of sowing and
698 cultivar choice for managing climate change: wheat crop phenology and water use
699 efficiency. *Int. J. Biometeorol.* 62, 1049–1061. <https://doi.org/10.1007/s00484-018->
700 1508-4
- 701 Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R.P., Ruane, A.C.,
702 Semenov, M.A., Wallach, D., Wang, E., Alderman, P.D., Kassie, B.T., Biernath, C.,
703 Basso, B., Cammarano, D., Challinor, A.J., Doltra, J., Dumont, B., Rezaei, E.E., Gayler,
704 S., Kersebaum, K.C., Kimball, B.A., Koehler, A.-K., Liu, B., O’Leary, G.J., Olesen, J.E.,
705 Ottman, M.J., Priesack, E., Reynolds, M., Stratonovitch, P., Streck, T., Thorburn, P.J.,

- 706 Waha, K., Wall, G.W., White, J.W., Zhao, Z., Zhu, Y., 2016. Crop model improvement
707 reduces the uncertainty of the response to temperature of multi-model ensembles. *F.*
708 *Crop. Res.* <https://doi.org/10.1016/j.fcr.2016.05.001>
- 709 McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency
710 Index. *J. Hydrol. Eng.* 11, 597–602. [https://doi.org/10.1061/\(ASCE\)1084-](https://doi.org/10.1061/(ASCE)1084-)
711 [0699\(2006\)11:6\(597\)](https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597))
- 712 Palosuo, T., Kersebaum, K.C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J.E., Patil,
713 R.H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F.,
714 Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B., Rötter, R., 2011. Simulation of winter
715 wheat yield and its variability in different climates of Europe: A comparison of eight
716 crop growth models. *Eur. J. Agron.* 35, 103–114.
717 <https://doi.org/10.1016/j.eja.2011.05.001>
- 718 R Core Team, 2017. A language and Environment for Statistical Computing.
- 719 Rötter, R.P., Palosuo, T., Kersebaum, K.C., Angulo, C., Bindi, M., Ewert, F., Ferrise, R.,
720 Hlavinka, P., Moriondo, M., Nendel, C., Olesen, J.E., Patil, R.H., Ruget, F., Takáč, J.,
721 Trnka, M., 2012. Simulation of spring barley yield in different climatic zones of
722 Northern and Central Europe: A comparison of nine crop models. *F. Crop. Res.* 133, 23–
723 36. <https://doi.org/10.1016/j.fcr.2012.03.016>
- 724 Sadras, V.O., Monzon, J.P., 2006. Modelled wheat phenology captures rising temperature
725 trends: Shortened time to flowering and maturity in Australia and Argentina. *F. Crop.*
726 *Res.* 99, 136–146. <https://doi.org/10.1016/J.FCR.2006.04.003>
- 727 Teluguntla, P., Thenkabail, P.S., Oliphant, A., Xiong, J., Gumma, M.K., Congalton, R.G.,
728 Yadav, K., Huete, A., 2018. A 30-m landsat-derived cropland extent product of Australia
729 and China using random forest machine learning algorithm on Google Earth Engine

- 730 cloud computing platform. *ISPRS J. Photogramm. Remote Sens.* 144, 325–340.
731 <https://doi.org/10.1016/J.ISPRSJPRS.2018.07.017>
- 732 van Bussel, L.G.J., Stehfest, E., Siebert, S., Müller, C., Ewert, F., 2015. Simulation of the
733 phenological development of wheat and maize at the global scale. *Glob. Ecol. Biogeogr.*
734 24, 1018–1029. <https://doi.org/10.1111/geb.12351>
- 735 Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P.J., van Ittersum, M.,
736 Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J.,
737 De Sanctis, G., Dumont, B., Eyshi Rezaei, E., Fereres, E., Fitzgerald, G.J., Gao, Y.,
738 Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C.,
739 Jones, C.D., Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A.-K., Maiorano, A.,
740 Minoli, S., Müller, C., Naresh Kumar, S., Nendel, C., O’Leary, G.J., Palosuo, T.,
741 Priesack, E., Ripoche, D., Rötter, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P.,
742 Streck, T., Supit, I., Tao, F., Wolf, J., Zhang, Z., 2018. Multimodel ensembles improve
743 predictions of crop-environment-management interactions. *Glob. Chang. Biol.* 24, 5072–
744 5083. <https://doi.org/10.1111/gcb.14411>
- 745 Wallach, D., Nissanka, S.P., Karunaratne, A.S., Weerakoon, W.M.W., Thorburn, P.J., Boote,
746 K.J., Jones, J.W., 2017. Accounting for both parameter and model structure uncertainty
747 in crop model predictions of phenology: A case study on rice. *Eur. J. Agron.* 88.
748 <https://doi.org/10.1016/j.eja.2016.05.013>
- 749 Wallach, D., Palosuo, T., Thorburn, P., Seidel, S.J., Gourdain, E., Asseng, S., Basso, B., Buis,
750 S., Crout, N.M.J., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S.,
751 Ghahramani, A., Hochman, Z., Hoek, S., Horan, H., Hoogenboom, G., Huang, M.,
752 Jabloun, M., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Luo,
753 Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Olesen, J.E., Poyda,

- 754 A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka,
755 X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber,
756 T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., 2019.
757 How well do crop models predict phenology, given calibration data from the target
758 population? bioRxiv 708578. <https://doi.org/10.1101/708578>
- 759 Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R.P., Kimball, B.A., Ottman,
760 M.J., Wall, G.W., White, J.W., Reynolds, M.P., Alderman, P.D., Aggarwal, P.K.,
761 Anothai, J., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G.,
762 Doltra, J., Fereres, E., Garcia-Vila, M., Gayler, S., Hoogenboom, G., Hunt, L.A.,
763 Izaurralde, R.C., Jabloun, M., Jones, C.D., Kersebaum, K.C., Koehler, A.-K., Liu, L.,
764 Müller, C., Naresh Kumar, S., Nendel, C., O’Leary, G., Olesen, J.E., Palosuo, T.,
765 Priesack, E., Eyshi Rezaei, E., Ripoche, D., Ruane, A.C., Semenov, M.A., Shcherbak, I.,
766 Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Thorburn, P., Waha, K.,
767 Wallach, D., Wang, Z., Wolf, J., Zhu, Y., Asseng, S., 2017. The uncertainty of crop yield
768 projections is reduced by improved temperature response functions. *Nat. Plants* 3, 1–13.
769 <https://doi.org/10.1038/nplants.2017.102>
- 770 Wang, H., Cutforth, H., McCaig, T., McLeod, G., Brandt, K., Lemke, R., Goddard, T.,
771 Sprout, C., 2009. Predicting the time to 50% seedling emergence in wheat using a Beta
772 model. *NJAS - Wageningen J. Life Sci.* 57, 65–71.
773 <https://doi.org/https://doi.org/10.1016/j.njas.2009.07.003>
- 774 Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the
775 root mean square error (RMSE) in assessing average model performance. *Clim. Res.* 30,
776 79–82.
- 777 Wöhling, T., Schöniger, A., Gayler, S., Nowak, W., 2015. Bayesian model averaging to

- 778 explore the worth of data for soil-plant model selection and prediction. *Water Resour.*
779 *Res.* 51, 2825–2846. <https://doi.org/10.1002/2014WR016292>
- 780 Workman, D., 2020. Worldstopexports [WWW Document]. URL
781 <http://www.worldstopexports.com/wheat-exports-country/> (accessed 3.10.20).
- 782 Yuan, S., Peng, S., Li, T., 2017. Evaluation and application of the ORYZA rice model under
783 different crop managements with high-yielding rice cultivars in central China. *F. Crop.*
784 *Res.* 212, 115–125. <https://doi.org/10.1016/J.FCR.2017.07.010>
- 785 Zadoks, J.C., Chzang, T.T., Konzak, C.F., 1974. A decimal code for the growth stages of
786 cereals. *Weed Res.* 14, 415–421. <https://doi.org/10.1111/j.1365-3180.1974.tb01084.x>
- 787 Zhang, S., Tao, F., Zhang, Z., 2017. Uncertainty from model structure is larger than that from
788 model parameters in simulating rice phenology in China. *Eur. J. Agron.* 87, 30–39.
789 <https://doi.org/10.1016/j.eja.2017.04.004>
- 790

Table S1

Model structures used in this study

Model structure	Version(s)	References
AgroC	May2018	Herbst M., Hellebrand H.J. , Bauer J., Huisman J.A., Šimůnek J., Weihermüller L., Graf A., Vanderborght J., Vereecken H. (2008). Multiyear heterotrophic soil respiration: Evaluation of a coupled CO ₂ transport and carbon turnover model. <i>Ecological Modelling</i> . 214: 271-283. Klosterhalfen, A., Herbst M., Weihermüller L., Graf A., Schmidt M., Stadler A., Schneider K., Subke J.-A., Huisman J.A., Vereecken H. (2017). Multi-site calibration and validation of a net ecosystem carbon exchange model for croplands. <i>Ecological Modelling</i> . 363: 137-156.
APSIM	7.8, 7.9, 7.10	Keating B.A., P.S. Carberry, G.L. Hammer, M.E. Probert, M.J. Robertson, D. Holzworth, N.I. Huth, J.N.G. Hargreaves, H. Meinke, Z. Hochman, G. McLean, K. Verburg, V. Snow, J.P. Dimes, M. Silburn, E. Wang, S. Brown, K.L. Bristow, S. Asseng, S. Chapman, R.L. McCown, D.M. Freebairn and C.J.Smith. (2003). An overview of APSIM, a model designed for farming systems simulation. <i>European Journal of Agronomy</i> 18: 267-288. Holzworth D.P., Huth N.I., DeVoil P.G. et al. (2014) APSIM - Evolution towards a new generation of agricultural systems simulation. <i>Environmental Modelling & Software</i> , 62, 327-350
AquaCrop	4.0	Vanuytrecht E., Raes D., Steduto P., Hsiao T.C., Fereres E., Heng L.K., Garcia Vila M., Mejias Moreno, P. (2014). AquaCrop: FAO'S crop water productivity and yield response model. <i>Environmental Modelling & Software</i> , 62: 351-360

CERES-Wheat	DSSATV4.7,V 4.7., Expert-N 3.0	<p>Hoogenboom, G., C.H. Porter, K.J. Boote, V. Shelia, P.W. Wilkens, U. Singh, J.W. White, S. Asseng, J.I. Lizaso, L.P. Moreno, W. Pavan, R. Ogoshi, L.A. Hunt, G.Y. Tsuji, and J.W. Jones. 2019. The DSSAT crop modeling ecosystem. In: p.173-216 [K.J. Boote, editor] <i>Advances in Crop Modeling for a Sustainable Agriculture</i>. Burleigh Dodds Science Publishing, Cambridge, United Kingdom (http://dx.doi.org/10.19103/AS.2019.0061.10).</p> <p>Hoogenboom, G., C.H. Porter, V. Shelia, K.J. Boote, U. Singh, J.W. White, L.A. Hunt, R. Ogoshi, J.I. Lizaso, J. Koo, S. Asseng, A. Singels, L.P. Moreno, and J.W. Jones. 2019. Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.7 (www.DSSAT.net). DSSAT Foundation, Gainesville, Florida, USA.</p>
CoupModel	Version 5.4.4	<p>Coucheney E, Eckersten H, Hoffmann H, Jansson PE, Gaiser T, Ewert F, Lewan E. 2018. Key functional soil types explain data aggregation effects on simulated yield, soil carbon, drainage and nitrogen leaching at a regional scale. <i>Geoderma</i>, 318: 167-181. DOI: 10.1016/j.geoderma.2017.11.025.</p> <p>Jansson, P-E. (2012). CoupModel: model use, calibration, and validation. <i>Transactions of the ASABE</i>, 55 (4):1337-1344. (American Society of Agricultural and Biological Engineers).</p> <p>Senapati, N., Jansson, P-E., Smith, P., Chabbi, A. (2016). Modelling heat, water and carbon fluxes in mown grassland under multi-objective and multi-criteria constraints. <i>Environmental modelling & software</i>, 80: 201-224.</p>
CROPSIM-Wheat	DSSAT V4.7	<p>Hoogenboom G., Porter C. H., Shelia V., Boote K. J., Singh U., White J. W., Hunt L. A., Ogoshi R., Lizaso J. I., Koo J., Asseng S., Singels A., L.P. Moreno, Jones J. W. (2017). Decision</p>

		Support System For Agrotechnology Transfer (DSSAT). Version 4.7. DSSAT Foundation, Gainesville, Florida, USA.
Cropsyst	3.04.08	Stöckle C. O., Donatelli M., Nelson R. (2003). CropSyst, a cropping systems simulation model. <i>European Journal of Agronomy</i> , 18(3-4), 289-307.
DAISY	5.59	Hansen S., P. Abrahamsen C. T. Petersen, Styczen M.. (2012). Daisy: Model Use, Calibration, and Validation. <i>Transactions of the ASABE</i> , 55, 1317–1335.
Nwheat	DSSAT	Kassie B.T., Asseng S., Porter C.H. and Royce F.S. (2016). Performance of DSSAT-Nwheat across a wide range of current and future growing conditions. <i>European Journal of Agronomy</i> , 81, 27-36.
GECROS	Expert-N 3.0	Yin X., van Laar H. H. (2005). Crop systems dynamics. An ecophysiological simulation model for genotype-by-environment interactions. Wageningen Academic Publishers, 155 pp., Wageningen, The Netherlands.
HERMES	4.27	Kersebaum K.C. (2007). Modelling nitrogen dynamics in soil-crop systems with HERMES. <i>Nutrient Cycling in Agroecosystems</i> , 77, 39-52. Kersebaum K.C. (2011). Special features of the HERMES model and additional procedures for parameterization, calibration, validation, and applications In: L.R. Ahuja and L. Ma (ed.): <i>Advances in Agricultural Systems Modeling Series 2</i> . 65-94. ASA, CSSA, SSSA, Madison, USA.
LINTUL	LINTUL5	Wolf J. (2012). User guide for LINTUL5: Simple generic model for simulation of crop growth under potential, water limited and nitrogen, phosphorus and potassium limited conditions. Wageningen UR.

MONICA	2.02	<p>Nendel C., Berg M., Kersebaum K.C., Mirschel W., Specka X., Wegehenkel M., Wenkel K.O., Wieland R. (2011). The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. <i>Ecological Modelling</i> 222(9), 1614 - 1625.</p> <p>Specka X., Nendel C., Wieland R. (2015). Analysing the parameter sensitivity of the agro-ecosystem model MONICA for different crops. <i>European Journal of Agronomy</i>, 71, 73-87.</p> <p>Specka X., Nendel C., Wieland R. (2019). Temporal Sensitivity Analysis of the MONICA Model: Application of Two Global Approaches to Analyze the Dynamics of Parameter Sensitivity. <i>Agriculture</i> 9(2), 37.</p>
OpenCrop		<p>OpenCrop: An Open Source Crop Model – Model Description. Crout NMJ, Karanaratne, A & Jabloun, M (2018). School of Biosciences, University of Nottingham, UK</p>
PANORAMIX	R version	<p>Gate, P., 1995. <i>Écophysiologie du blé</i>. Lavoisier-Technique et documentation.</p> <p>Chatelin, M.H., Aubry, C., Poussin, J.C., Meynard, J.M., Massé, J., Verjux, N., Gate, P., Le Bris, X., 2005. DéciBlé, a software package for wheat crop management simulation. <i>Agric. Syst.</i> 83, 77–99. https://doi.org/10.1016/J.AGSY.2004.03.003</p>
Salus		<p>Basso B, Ritchie JT, Grace PR, Sartori L (2006) Simulation of tillage systems impact on soil biophysical properties using the SALUS model. <i>Italian Journal of Agronomy</i>,1, 677-688.</p> <p>Basso B. and J.T. Ritchie. 2015. Simulating Crop Growth and Biogeochemical Fluxes in Response to Land Management using the SALUS Model. In S. K. Hamilton, J. E. Doll, and G. P. Robertson, editors. <i>The ecology of agricultural landscapes: long-term research on the path to sustainability</i>. Oxford University Press, New York, NY USA</p>
SPASS	Expert-N 3.0	<p>Wang, E. (1997). <i>Development of a Generic Process-Oriented Model for Simulation of Crop Growth</i>. München, Herbert Utz Verlag Wissenschaft. 195 pp.</p>

SSM-Wheat		Soltani A., Maddah V., Sinclair T. (2013). SSM-Wheat: a simulation model for wheat development, growth and yield. <i>International Journal of Plant Production</i> , 7, 711-740.
STICS	8_5_0	Brisson N., Launay M., Mary B., Beaudoin N. (2009). Conceptual basis, formalisations and parametrization of the STICS crop model. <i>Quae</i> , 304pp Coucheney E., Buis S., Launay M. Constantin J., Mary B., Garcia de Cortazar-Atauri I., Ripoche D., Beaudoin N., Ruget F., Andrianorisoa S., Le Bas C., Justes E., Léonard J. (2015). Accuracy, robustness and behavior of the STICS 8.2.2 soil-crop model for plant, water and nitrogen outputs: evaluation over a wide range of agro-environmental conditions in France. <i>Environmental Modelling & Software</i> , 64, 177-190
SUCROS	Expert-N 3.0	van Laar, H.H. , J. Goudriaan, und H. van Keulen, 1992: Simulation of crop growth for potential and water-limited production situations (as applied to spring wheat).: Simulation Report CABO-TT no. 27. Wageningen: Centre for Agrobiological Research and Department of Theoretical Production Ecology, Wageningen Agricultural University; Vanclooster, M., Viaene P., Diels J., Christiaens K., 1994: WAVE a mathematical model for simulating water and agrochemicals in the soil and vadose environment. Reference and user's manual (release 2.0). Leuven: Institute for Land and Water Management, Katholieke Universiteit Leuven.
PCWOFOST	5.3.3	Ceglar A. , van der Wijngaart R. , de Wit A., Lecerf R., Boogaard H., Seguini L. , van den Berg M., Toreti A., Zampieri M., Fumagalli D., Baruth B. (2019). Improving WOFOST model to simulate winter wheat phenology in Europe: Evaluation and effects on yield. <i>Agricultural Systems</i> . 168, 168-180.
WCCWOFOST	7.1.7	Boogaard, H.L., Van Diepen, C.A., Rötter, R.P., Cabrera, J.M.C.A., Van Laar, H.H., 1998. User's guide for the WOFOST 7.1 crop growth simulation model and WOFOST control center 1.5. Technical Document 52. Winand Staring Centre, Wageningen, the Netherlands, 144 pp.

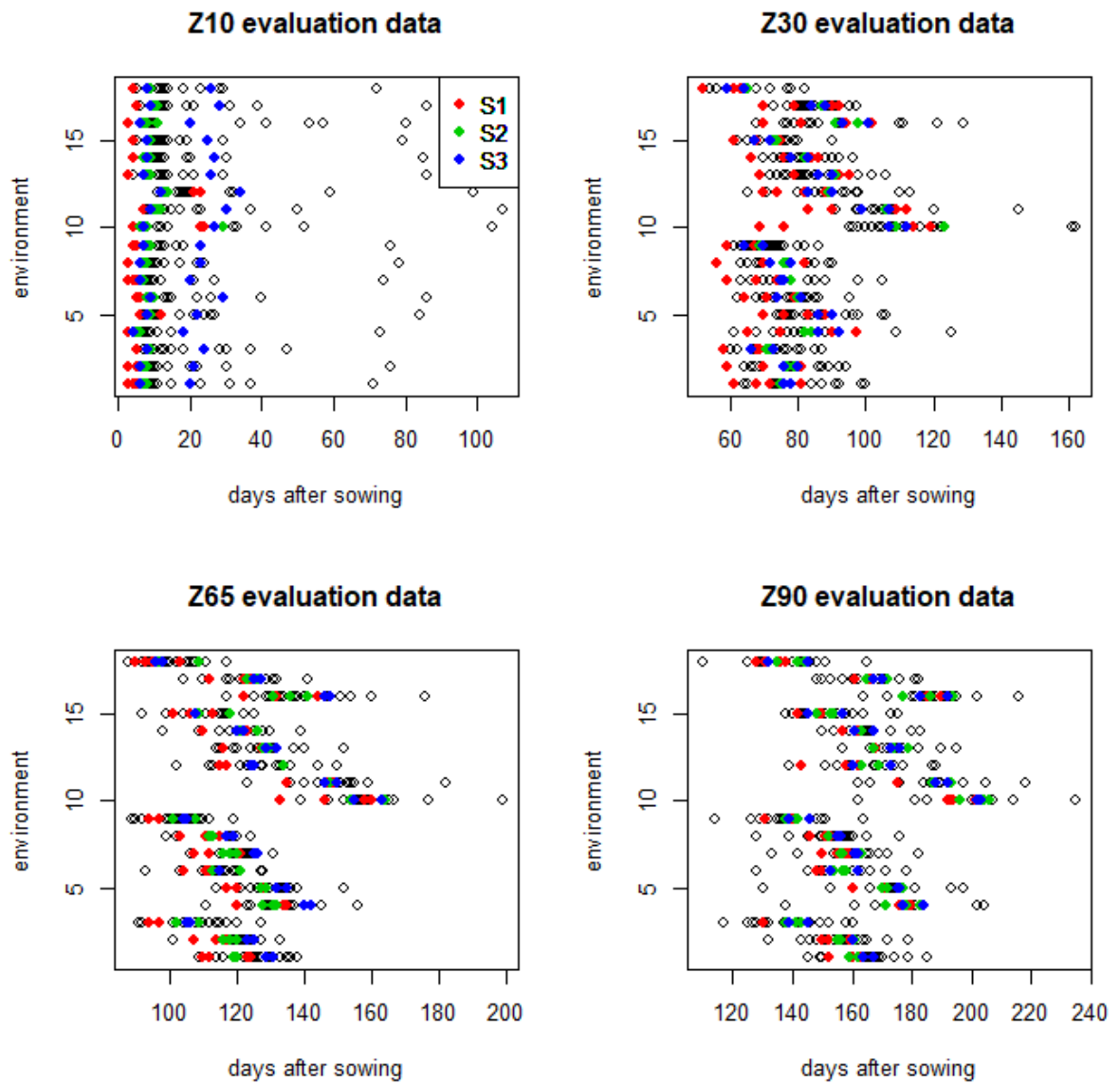
Wheat-Grow	3.1	<p>Zhu Y.; Liu L.; Liu, B. WheatGrow: A simulation model for predicting growth and productivity in wheat. In Proceedings of the Workshop on Modeling Wheat Response to High Temperature, Texcoco, Mexico, 19–21 June 2013.</p> <p>Lv Z., Liu X., Tang L., Liu, L., Cao, W. and Zhu, Y., 2016. Estimation of ecotype-specific cultivar parameters in a wheat phenology model and uncertainty analysis. <i>Agricultural and Forest Meteorology</i>, 221: 219-229.</p>
------------	-----	---

794

795

796

797



798

799

Figure S1

800

801

802

803

804

Predictions of days from sowing to Zadoks stages Z10 (emergence), Z30, Z65 and Z90 by each modeling group for each evaluation environment. Modeling groups that used the same model structure are identified by color (red for structure S1, green for structure S2, blue for structure S3).

805

Table S2

806 Prediction errors for each modeling group and for the models e-mean, e-median, naive and onlyT. The columns are
 807 MAE averaged over the stages Z30, Z65 and Z90 for the evaluation environments (days), MAE for each of the stages Z30, Z65
 808 and Z90 for the evaluation environments (days), root mean squared error (RMSE) averaged over the stages Z30, Z65 and Z90
 809 for the evaluation environments (days), the skill measures EF and skillT averaged over the stages Z30, Z65 and Z90 for the
 810 evaluation environments (unitless) and MAE averaged over the stages Z30, Z65 and Z90 for the calibration environments
 811 (days). The models are ordered by average MAE (value in first column). NA indicates that that modeling group didn't predict
 812 the time to the indicated stage.

	MAE _eval	MAE _Z30	MAE _Z65	MAE _Z90	RMSE _eval	EF_ eval	skillT _eval	MA E_cal
M9	6.3	NA	9.3	3.2	7.2	0.7	0.489	6.2
an eme	6.3	8.8	7.3	2.9	8.1	0.6 4	0.38	6
M24	6.4	9	6.8	3.2	8.6	0.6 2	0.351	8.5
dian eme	6.4	8.6	7.4	3.3	8.3	0.6 3	0.367	5.9
M21	6.6	8.7	6.6	4.4	8.5	0.6 4	0.379	5.9
M4	6.7	9.8	6.4	3.8	8.4	0.6 5	0.39	5.7
M2	6.8	10.4	7.3	2.8	8.8	0.5 7	0.263	6.3
M13	7.2	10.6	7.9	3.2	9	0.5 5	0.231	8.3
M18	7.2	NA	10.8	3.6	8.6	0.6 2	0.336	7.7
M15	7.3	11.7	4.6	5.6	8.5	0.6 4	0.383	8

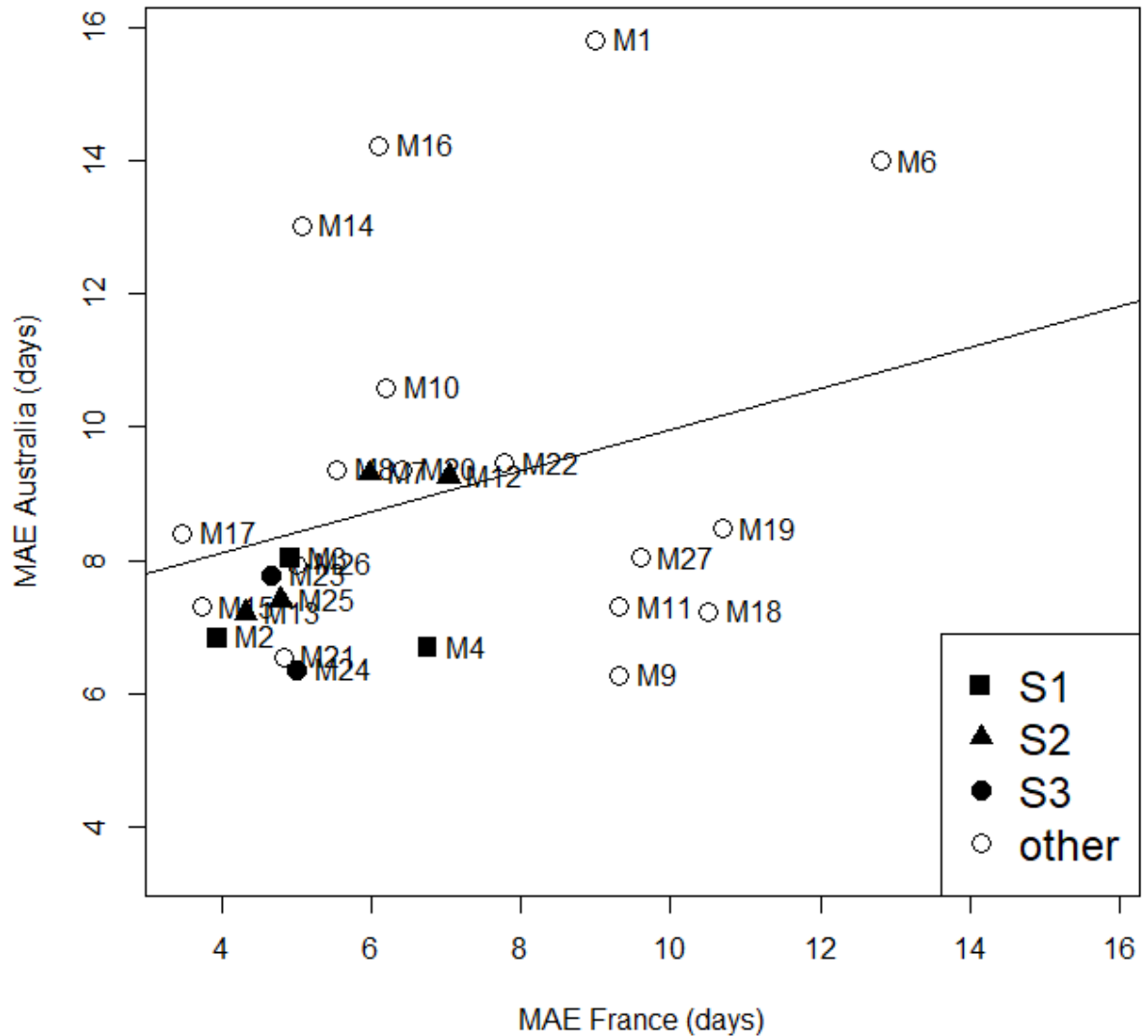
M11	7.3	7.7	7.4	6.8	10	0.5 3	0.18	9
M25	7.4	10.7	6.7	4.8	8.9	0.6	0.307	6.1
M23	7.8	10.6	8.5	4.2	10	0.5	0.14	6.7
M26	7.9	9.3	10.3	4.2	10.3	0.4 7	0.082	7.4
M3	8	NA	10.1	6	9	0.6 1	0.322	7.7
M27	8	9.7	10.9	3.4	10.1	0.4 6	0.066	6.7
M29	8.2	8.2	8.1	NA	11.2	0.4 4	0.029	9.7
only T	8.2	10.7	10.6	3.2	10.5	0.4 2	0	8
M17	8.4	7.3	7.3	10.6	10.4	0.5 2	0.165	7
M19	8.5	11.4	10.2	3.8	10.4	0.4 4	0.032	7.9
M12	9.3	12.4	7.3	8	11.4	0.3 9	- 0.058	8.1
M7	9.3	15.4	9	3.4	11.5	0.2 3	- 0.323	13.3
M20	9.3	NA	11.5	7.2	11.4	0.4	- 0.041	8.3
M8	9.4	12.4	8.8	6.8	11.6	0.3 5	- 0.117	12
M22	9.5	9.1	15.8	3.4	12.1	0.2 1	- 0.362	7.3
M10	10.6	24.8	5.9	1	13.3	- 0.54	- 1.664	12

e	naiv	11.3	12.2	14.3	7.5	14.6	0	-	17.1
								0.727	
	M14	13	15.2	13.3	10.5	16.2	-	-	18.7
							0.18	1.044	
	M6	14	9.9	10	22	17.6	-	-	8.1
							0.47	1.537	
	M16	14.2	11.8	10	20.8	16.6	-	-	13.9
							0.29	1.228	
	M1	15.8	20.6	10.4	16.4	16.8	-	-	12.8
							0.33	1.301	
	M28	20	18.3	24.2	17.4	23.5	-	-	17.4
							1.51	3.343	

813

814

815



816

817

Fig. S2

818

819

820

821

822

823

Relation between mean absolute error (MAE) for the Australian environments and MAE for the French environments, for modeling groups that participated in both studies. Values are averages over predicted development stages. Points are identified by modeling group. Modeling groups that shared the same structure (S1, S2 or S3) are identified by filled squares, triangles or circles, respectively. The regression line is $y=8.23+0.24x$.