# Multi model evaluation of phenology prediction for wheat in Australia

1. Wallach1, Daniel; Palosuo2, Taru; Thorburn3, Peter; Hochman3, Zvi; Andrianasolo4, Fety; Asseng5,
2. Senthold; Basso6, Bruno; Buis7, Samuel; Crout8, Neil; Dumont9, Benjamin; Ferrise10, Roberto; Gaiser11, Thomas;
3. Gayler12, Sebastian; Hiremath13, Santosh; Hoek14, Steven; Horan3, Heidi; Hoogenboom5,15, Gerrit; Huang16,
4. Mingxia; Jabloun8, Mohamed; Jansson17, Per-Erik; Jing18, Qi; Justes19, Eric; Kersebaum20,21, Kurt Christian;
5. Launay22, Marie; Lewan23, Elisabet; Luo24, Qunying; Maestrini14, Bernardo; Moriondo25, Marco; Padovan10,
6. Gloria; Olesen26, Jørgen Eivind; Poyda27, Arne; Priesack28, Eckart; Pullens26, Johannes Wilhelmus Maria; Qian18,
7. Budong; Schütze29, Niels; Shelia5,15, Vakhtang; Souissi30,31, Amir; Specka20, Xenia; Srivastava11, Amit Kumar;
8. Stella20, Tommaso; Streck12, Thilo; Trombi10, Giacomo; Wallor20, Evelyn; Wang16, Jing; Weber12, Tobias, K.D.;
9. Weihermüller32, Lutz; de Wit14, Allard; Wöhling29,33, Thomas; Xiao5,34, Liujun; Zhao5, Chuang; Zhu34, Yan;
10. Seidel, Sabine J.11

12

13

1. 1INRA, UMR AGIR, Castanet Tolosan, France. ORCID 0000-0003-3500-8179
2. 2Natural Resources Institute Finland (Luke), Helsinki, Finland
3. 3CSIRO Agriculture and Food, Brisbane, Queensland, Australia
4. 4ARVALIS - Institut du végétal Paris, France
5. 5Agricultural and Biological Engineering Department, University of Florida, Gainesville, Florida
6. 6Department of Earth and Environmental Sciences, Michigan State University, East Lansing, Michigan
7. 7INRA, UMR 1114 EMMAH, Avignon, France
8. 8School of Biosciences, University of Nottingham, Loughborough, UK
9. 9Plant Sciences & TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege,
10. Gembloux, Belgium
11. 10Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Italy
12. 11Institute of Crop Science and Resource Conservation, University of Bonn, Germany
13. 12Institute of Soil Science and Land Evaluation, Biogeophysics, University of Hohenheim, Stuttgart, Germany
14. 13Aalto University School of Science, Espoo, Finland
15. 14 Wageningen University & Research, Wageningen, The Netherlands
16. 15Institute for Sustainable Food Systems, University of Florida, Gainesville, Florida
17. 16College of Resources and Environmental Sciences, China Agricultural University, Beijing, China
18. 17Royal Institute of Technology (KTH), Stockholm, Sweden
19. 18Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada
20. 19CIRAD, UMR SYSTEM, Montpellier, France
21. 20Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany
22. 21Global Change Research Institute CAS, Brno, Czech Republic
23. 22INRA, US 1116 AgroClim, Avignon, France
24. 23Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
25. 24Hillridge Technology Pty Ltd, Sydney, Australia
26. 25CNR-IBE, Firenze, Italy
27. 26Department of Agroecology, Aarhus University, Tjele, Denmark
28. 27Grass and Forage Science / Organic Agriculture, Institute of Crop Science and Plant Breeding, Kiel University,
29. Kiel, Germany
30. 28Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research Center for
31. Environmental Health, Neuherberg, Germany
32. 29Institute of Hydrology and Meteorology, Chair of Hydrology, Technische Universität Dresden, Dresden,
33. Germany
34. 30National Institute of Agronomic Research of Tunisia (INRAT), Agronomy Laboratory, University of Carthage,
35. Tunis, Tunisia
36. 31National Agronomy Institute of Tunisia (INAT), University of Carthage, Tunis, Tunisia
37. 32Institute of Bio- and Geosciences - IBG-3, Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
38. 33Lincoln Agritech Ltd., Hamilton, New Zealand
39. 34National Engineering and Technology Center for Information Agriculture, Jiangsu Key Laboratory for
40. Information Agriculture, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing
41. Agricultural University, Nanjing, Jiangsu, China

55

56

# Abstract

1. Predicting wheat phenology is important for cultivar selection, for effective crop
2. management and provides a baseline for evaluating the effects of global change. Evaluating
3. how well crop phenology can be predicted is therefore of major interest. Twenty-eight wheat
4. modeling groups participated in this evaluation. Model predictions depend not only on model
5. structure but also on the parameter values. This study is thus an evaluation of modeling groups,
6. which choose the structure and fix or estimate the parameters, rather than an evaluation just of
7. model structures. Our target population was wheat fields in the major wheat growing regions
8. of Australia under current climatic conditions and with current local management practices.
9. The environments used for calibration and for evaluation were both sampled from this same
10. target population. The calibration and evaluation environments had neither sites nor years in
11. common, so this is a rigorous evaluation of the ability of modeling groups to predict phenology
12. for new sites and weather conditions. Mean absolute error (MAE) for the evaluation
13. environments, averaged over predictions of three phenological stages and over modeling
14. groups, was 9 days, with a range from 6 to 20 days. Predictions using the multi-modeling group
15. mean and median had prediction errors nearly as small as the best modeling group. For a given
16. modeling group, MAE for the evaluation environments was significantly correlated with MAE
17. for the calibration environments, which suggests that it would be of interest to test ensemble
18. predictors that weight individual modeling groups based on performance for the calibration
19. data. About two thirds of the modeling groups performed better than a simple but relevant
20. benchmark, which predicts phenology by assuming a constant temperature sum for each
21. development stage. The added complexity of crop models beyond just the effect of temperature
22. was thus justified in most cases. Finally, there was substantial variability between modeling
23. groups using the same model structure, which implies that model improvement could be
24. achieved not only by improving model structure, but also by improving parameter values, and
25. in particular by improving calibration techniques.
26. Keywords: evaluation, phenology, wheat, Australia, structure uncertainty, parameter
27. uncertainty 85

# 1. Introduction

* 1. Crop phenology describes the cycle of biological events during plant growth. These
  2. events include, for example, seedling emergence, leaf appearance, flowering, and maturity.
  3. Timing of growing seasons and their critical phases as well as estimates of them are increasingly
  4. important in changing climate (Olesen et al., 2012, Dalhaus et al., 2018). Matching the
  5. phenology of crop varieties to the climate in which they grow is critical for viable crop
  6. production strategies (Rezaei et al., 2018, Hunt et al., 2019). Furthermore, accurate simulation
  7. of phenology is essential for models which simulate plant growth and yield (Archontoulis et

94 al., 2014; Boote et al., 2010, 2008).

1. In this study we focus on wheat phenology in Australia. Australia was the world’s ninth
2. largest producer of wheat in 2018 and the sixth largest exporter (Workman, 2020). Crop model
3. predictions of phenology have been used in various studies related to wheat production in
4. Australia. In a study by Luo et al. (2018), the APSIM model was used to simulate changes in
5. phenology, water use efficiency, and yield to be expected from global climate change. The

100

APSIM model was used to evaluate changes in wheat phenology in Australia as a result of

101

warming temperatures in recent decades (Sadras and Monzon, 2006). That model was also used

102

to determine the flowering date at each location associated with highest average yield (Flohr et

103

al., 2017).

104

Given the interest in using crop models to predict phenology, it is important to evaluate

105

those predictions. How well can wheat phenology be predicted? In trying to answer this

106

question, one must first define exactly what aspect of the models is being evaluated, and then

107

must choose an appropriate methodology for carrying out the evaluation.

108

It is important to distinguish two different types of model evaluation, which might be

109

termed evaluation of extrapolation predictions and evaluation of interpolation predictions. They

110

differ as to whether or not the data provided for calibration are representative of the target

111

population, i.e. of the range of environments of interest. In one type of study, the objective is

112

to evaluate how well models can extrapolate to conditions not represented in the calibration

113

data. For example, in a multi-model ensemble study on the effect of high temperatures on wheat

114

growth (Asseng et al., 2015), detailed crop measurements were provided for one planting date

115

and the models were evaluated using other planting dates, some with additional artificial heating

116

during growth. The evaluation data thus represented a much larger range of temperatures than

117

represented in the calibration data. This was a test of how well the models can extrapolate to

118

more extreme temperatures than those available for calibration. Other studies have evaluated

119

how well crop models can extrapolate to environments with enhanced CO2, given calibration

120

data for current ambient CO2 levels (Biernath et al., 2011).

121

In the second type of study, the calibration data are meant to be representative of the

122

target population. This evaluates how well crop models can generalize from the calibration

123

environments to other similar environments. An example is the study by Ceglar et al. (2019),

124

which used data on wheat phenology under current conditions in Europe for calibration and

125

then predicted phenology for other environments from the same target population. This type of

126

evaluation is adapted, for example, to the case where one has data from a network of variety

127

trials and wants to predict for other sites and years from the same target population, as in Bao

128

et al.. (2017) for yield. It is this aspect of crop phenology models, namely their ability to predict

129

when provided with a sample of data from the target population, that is evaluated in the present

130

study.

131

A second aspect of evaluation that must be specified is the modeling group or groups

132

that are being evaluated. We reserve the term “model” specifically for model structure, i.e. the

133

model equations, while modeling group determines both the model structure and the parameter

134

values, which are chosen or estimated by the group running the model. It is clear that predictions

135

depend not only on the model structure but also on the parameter values, so evaluation really

136

refers to the modeling group. Model evaluation studies may refer to a particular modeling group

137

or to an ensemble of modeling groups. Here, we evaluate an ensemble of 28 different modeling

138

groups. The purpose is not to give information about each specific modeling group, but rather

139

to evaluate how well currently active modeling groups can predict phenology for our target

140

population (e.g. what is the error of the best predicting group), how well can one expect a

141

modeling group chosen at random to predict (e.g. what is the mean or median prediction error),

142

and what is the variability between modeling groups (e.g. what is the spread between the best

143

and worst predictors).

144

It is important to define precisely the evaluation problem (extrapolation or interpolation,

145

single- or multi-group evaluation), but it is also important that the methodology of evaluation

146

be such as to give reliable results. We focus here on the relation of the predictor (model plus

147

parameter values) and evaluation data. It is well-known from statistics that if a predictor is not

148

independent of the evaluation data, then the error for the evaluation data will in general be less

149

than for new environments (Efron, 1986). That is, non-independence in general leads to

150

underestimating prediction errors. The predictor could depend on the evaluation data if, for

151

example, the evaluation data were also used to calibrate the model, or were used to modify the

152

model equations, or were used to tune site characteristics. If the same sites are present in the

153

calibration and evaluation data, then the model has to some extent been tuned to those sites, and

154

so the predictor is not independent of the evaluation data even if the evaluation data have not

155

been used directly to fit the model. Having the same sites in the calibration and evaluation data

156

is often the case for evaluation studies (Andarzian et al., 2015; Asseng et al., 2008; Chauhan et

157

al., 2019; Hussain et al., 2018; Yuan et al., 2017).

158

There do not seem to have been any multi-modeling group evaluation studies of

159

prediction of wheat phenology in Australia, where the calibration data are sampled from the

160

target population (i.e. evaluation of interpolation predictions). The purpose of this study is to

161

present such an evaluation, using a rigorous approach where the parameterized model is

162

independent of the evaluation data.

163

# 2. Materials and Methods

164

## 2.1 Experimental data

165

The data are a subset from a multi-cultivar, multi-location, and multi-sowing date trial

166

for wheat in Australia, described in(Lawes et al. (2016). The environments reflect the diversity

167

in the wheat-growing regions of Australia (Fig. 1). Only the data for cultivar Janz, classified as

168

a fast-moderate maturing cultivar, were used here. The data are from 10 sites, located

169

throughout the grain growing region each with one to three sowing years and three planting

170

dates in each year (overall 66 environments, i.e. site-sowing date combinations, [Table 1](#_bookmark0)). The

171

sowing dates at each site correspond to early, conventional, and late sowing. Plant density was

172

100-120 plants/m², and sowing depth was 20-35 mm. Nutrients were managed to be non-

173

174

limiting. There were 1-3 repetitions for each environment (average of 2.1 repetitions).

175



176

### Figure 1

177

### Location of calibration (red circles) and evaluation (blue triangles) sites across the

178

### Australian cropping zones (shaded area; Source: Teluguntla et al., 2018).

179

Plots were visited regularly (about every two weeks) starting soon after emergence of

180

the early sowing and ending after crop maturity, and the Zadoks growth stage (Zadoks et al.,

181

1974), on a scale from 1-100, was determined. Overall, there were 709 combinations of

182

environment and measurement date, with an average of 10.7 stage notations per environment.

183

The stages to be predicted here are stage Z30 (Zadoks stage 30, pseudostem, i.e. youngest leaf

184

sheath erection), stage Z65 (Zadoks stage 65, anthesis half-way, i.e. anthers occurring half way

185

to tip and base of ear), and stage Z90 (Zadoks stage 90, grain hard, difficult to divide). These

186

stages are often used for management decisions or to characterize phenology.

187

In preparing the data for the simulation study, a linear interpolation was performed

188

between each pair of stages, to give the date for every integer Zadoks stage from the first to the

189

last observed stage. If observed Zadoks stage for one date was larger than the observed Zadoks

190

stage for the next measurement date, both stages were replaced by the average of the two Zadoks

191

stages before interpolation. The interpolated values were provided in order to avoid different

192

modeling groups using different methods for interpolating the data, which would have added

193

additional uncertainty unrelated to the model performance.

194

The average standard deviation of observed Zadoks stages based on the replicates was

195

0.93 days. The standard deviation of interpolated days after sowing to Z30, Z65, and Z90 was

196

calculated using a bootstrap. For a day with *r* replicates, a sample of size *r* was obtained by

197

drawing values at random with replacement, independently for each measurement date. Then

198

the Zadoks values were interpolated as for the original data. This was done 1000 times, giving

199

standard deviations of 1.8 days for observed days to Z30, 0.9 days for observed days to Z65,

200

and 0.5 days for observed days to Z90, respectively.

201

Part of the data was provided to the modeling groups for calibration , and part was never

202

revealed to participants and used for evaluation . The calibration data originated from four sites,

203

two years, and three planting dates, so overall 24 environments. The evaluation data were from

204

six sites, one year, and three planting dates for a total of 18 environments (Table 1). The data

205

were divided in such a way that the calibration and evaluation data had neither sites nor years

206

in common.

207

### Table 1

208

### Sites and sowing dates for calibration (underlined) and evaluation (bold). Note that

209

### the calibration and evaluation data have neither sites nor years in common.

|  |  |  |  |
| --- | --- | --- | --- |
| site\ year | 2010 | 2011 | 2012 |
| Bungunya (Queensland) |  |  | **2012-05-10**  **2012-05-22**  **2012-06-23** |
| Corrigin  (West Australia) |  |  | **2012-05-02**  **2012-05-21** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | **2012-06-21** |
| Eradu | 2010-05-14 | 2011-04-29 |  |
| (West Australia) | 2010-05-27 | 2011-05-24 |
| 2010-06-22 | 2011-06-23 |
| LakeBolac | 2010-05-03 | 2011-05-09 |  |
| (Victoria) | 2010-05-19 | 2011-06-03 |
| 2010-07-08 | 2011-06-16 |
| Minnipa | 2010-04-30 | 2011-05-13 |  |
| (South Australia) | 2010-05-31 | 2011-05-27 |
| 2010-06-24 | 2011-06-24 |
| Nangwee |  |  | **2012-05-17** |
| (Queensland) | **2012-05-31** |
| **2012-06-23** |
| Spring Ridge | 2010-05-10 | 2011-05-09 |  |
| (New South Wales) | 2010-06-11 | 2011-06-06 |
| 2010-07-01 | 2011-06-23 |
| Temora |  |  | **2012-05-05** |
| (New South Wales) | **2012-05-23** |
| **2012-06-25** |
| Turretfield |  |  | **2012-05-30** |
| (South Australia) | **2012-06-15** |
| **2012-07-05** |
| Walpeup |  |  | **2012-04-27** |
| (Victoria) | **2012-06-04** |
| **2012-07-18** |

210

211

212

213

To characterize the environments, we calculated for each environment the average

214

temperature from sowing to Z30, Z65, and Z90, the average photoperiod from Z30 to Z65 using

215

the daylength function in the R package insol (Corripio, 2019.; R Core Team, 2017) and days

216

to full vernalization using the model in van Bussel et al. (2015) with the value *Vsat* = 25 days,

217

estimated from the figure in their paper. [Figure 2](#_bookmark1) shows the range of average temperature, day

218

length, and days to vernalization for the calibration and evaluation environments as well as the

219

range of observed calendar days to Z30, Z65, and Z90. The range of values for the evaluation

220

data is always within the range of the calibration data, with the single exception of photoperiod.

221

While the median and maximum day lengths were very similar for the two sets of environments,

222

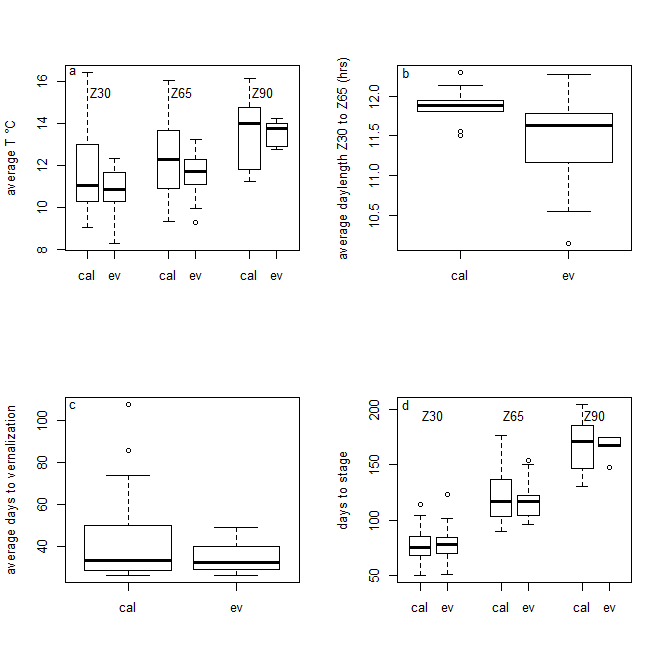
the shortest day length was 11.5 hours among calibration environments, while among the

223

224

evaluation environments the shortest day length was 10.1 hours.

225



226

### Figure 2

227

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

228

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

229

### average days from sowing to complete vernalization d) average days from sowing to

230

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

231

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

232

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

233

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

234

### observations that are beyond that range.

235

## Modeling groups

236

Twenty-eight different modeling groups participated in this study. “Modeling group”

237

here refers to the association of some specific model structure (some specific named model)

238

with some specific parameter values. Three modeling groups used the same underlying model

239

structure, referred to as structure S1, three groups used a second model structure (noted S2),

240

and two groups used a third model structure (noted S3). All other groups used unique model

241

structures. The model structures involved are presented in Supplementary Table S1. Models

242

were considered to have the same structure even if the version number was different, because

243

version differences are expected to be negligible for phenology. Since different groups using

244

the same structure obtained different results, identifying the contributions by the name of the

245

model would be misleading. Furthermore, the performance of specific groups was not of major

246

interest here. Therefore the modeling groups were anonymized, and only identified by a

247

number. There is no model M5 because that group dropped out in the course of the study.

248

The multi-model ensemble here was an “ensemble of opportunity” meaning that any

249

modeling group that asked to join was accepted. The activity was announced on the list server

250

of the Agricultural Modeling Inter-comparison and Improvement Project (AgMIP) and on the

251

list servers of several models. In addition to the original models, we defined two ensemble

252

models. The model e-mean has predictions equal to the mean of the simulated values. The

253

model e-median has predictions equal to the median of the simulated values.

254

## Simulation experiment

255

Each participating modeling group was provided with weather, soil, and management

256

data for all environments, as well as all available observed and interpolated values for days to

257

each Zadoks stage for the calibration data. Participants were requested to return simulated

258

values for number of days from sowing to emergence (even though days to emergence was

259

never observed) and values for number of days from sowing to stages Z30, Z65, and Z90 for

260

all environments, including both the calibration environments and the evaluation environments.

261

## 2.4Evaluation

262

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

263

*m*, MAE is

264

*n*

*MAEm* (1 / *n*)*yi* *y*ˆ*i*,*m*

*i*1

(1)

265

where

*yi* is the observed value for environment *i* and

*y*ˆ*i*,*m* is the value simulated by modeling

266

group *m* for that environment. The sum is over either calibration environments, to evaluate

267

goodness-of-fit, or over evaluation environments, to estimate prediction error. This is

268

preferred over mean squared error (MSE) or root mean squared error (RMSE), because unlike

269

MSE, MAE does not give extra weight to large errors (Willmott and Matsuura, 2005). To test

270

whether MAE is the same for prediction of days to different stages, we used the R function

271

pairwise.t.test, with method=”holm” to correct for multiple comparisons. We also calculated

272

MSE, RMSE, and NRMSE (normalized root mean squared error) for comparison with other

273

studies.

*MSE*

*n*

(1 / *n*)

*y* *y*ˆ 2

274

*m*

*RMSEm* 



*i*1

*MSEm*

*i i*,*m*

(2)

*NRMSEm* *RMSEm* / *y*

275

where *y* is the average of the observed values.

276

We considered two skill measures. A skill measure compares prediction error of the

277

modeling group to be evaluated with the error of a simple model used for comparison. We

278

define two simple models, and therefore two skill measures. Both use MSE, rather than MAE,

279

as the measure of model error, in keeping with usual practice. The first simple model, noted

280

“naive”, predicts that days to each stage will be equal to the average number of days to that

281

stage in the calibration data. The predictions of the naïve model here are 77.1, 123.1, and 166.5

282

days from sowing to stages Z30, Z65, and Z90, respectively. The first skill measure, modeling

283

efficiency (EF), is defined as

284

*EFm* 1*MSEm* / *MSEnaive*

(3)

285

The naive model ignores all variability and predicts that days to any stage will be the same

286

regardless of the environment. A model with EF ≤ 0 is a model that does no better than the

287

naive model, and so would be considered a very poor predictor. A perfect model, with no error,

288

has modeling efficiency of 1. Often modeling efficiency is based on the fit of a calibrated model

289

to the data used for calibration (McCuen et al., 2006). Here, in contrast, the naïve model is

290

based on calibration data and used to predict for independent data.

291

The naïve model is a very low baseline for evaluating a crop model. We therefore

292

introduce a more realistic, but still simple model which takes into account the effect of

293

temperature on phenology. This “onlyT” model predicts that degree days (°D) from sowing to

294

each stage will be equal to the number of degree days from sowing to that stage in the calibration

295

data, where degree days on any calendar day is equal to average temperature that day. The

296

predictions of the onlyT model are that Z30 will occur 893.7 °D after sowing, Z65 will occur

297

1476.0 °D after sowing, and Z90 will occur 2245.7 °D after sowing. The second skill measure,

298

noted skillT, is then

299

*skillTm* 1*MSEm* / *MSEonlyT*

(4)

300

where

*MSEonlyT*

is MSE for the onlyT model. As for any skill measure, a perfect model has

301

skillT = 1 and a model that does no better than the onlyT model has skillT ≤ 0

302

## 2.5 Within- and between-model structure variability

303

Three of the model structures are used by more than one modeling group. This makes it

304

possible to estimate separately the variance in simulated values due to structure and the variance

305

due to modeling group nested within structure. We treat the simulated values as a sample from

306

the distribution of plausible model structures and plausible parameter values. According to the

307

law of total variance (Casella and Berger, 1990), the total variance of simulated values can be

308

decomposed into two parts as

309

var( *y*ˆ) var *E* *y*ˆ | *S* *E* var *y*ˆ | *S*  (5)

310

where *y*ˆ

are the simulated values, *S* is model structure, E is the expectation, var is the variance,

311

and the notation |S means that the expectation (in the first term on the right hand side) or the

312

variance (in the second term on the right hand side) is taken separately for each value of model

313

structure. We estimated the first term by first calculating the average simulated value for each

314

structure (if a structure is represented by a single modeling group, this is just the value simulated

315

by that group), and then calculating the variance of those average values. This is the between-

316

structure variability. To estimate the second term, we first calculated the variance between

317

simulated values for each of the three structures with multiple groups. Then we calculated the

318

average of those variances. This is the within-structure variability (i.e. variability due to

319

parameters).

320

# 3.Results

321

## Prediction error and skill

322

MAE values for the evaluation data are shown in [Figure 3](#_bookmark3) and summarized in [Table 2.](#_bookmark2)

323

Results for individual modeling groups are given in Supplementary Table S2. Median MAE

324

values (and ranges) were 12 days (8-25 days) for days to Z30, 10 days (5-24 days) for days to

325

Z65, and 7 days (1-22 days) for days to Z90. The difference between MAE for prediction of

326

days to Z30 and MAE for prediction of days to Z65 was significant (*p*=0.041) as was the

327

difference between MAE for prediction of days to Z30 and MAE for prediction of days to Z90

328

(*p*=0.011). On the other hand, the difference between MAE for prediction of days to Z65 and

329

to Z90 had a p value of 0.11. The median (and range) of MAE averaged over the three stages

330

was 9 days (6-20 days). The ensemble predictors e-mean and e-median both had averaged MAE

331

values of 7 days. They were both only marginally worse than the best two individual modeling

332

groups, and e-median was marginally better than e-mean. For comparison with other studies,

333

we also report other criteria of error in [Table 2.](#_bookmark2)

334

### Table 2

335

### Summary of prediction errors for the evaluation and calibration environments,

336

### in each case averaged over predictions of days to stages Z30, Z65, and Z90 except for

337

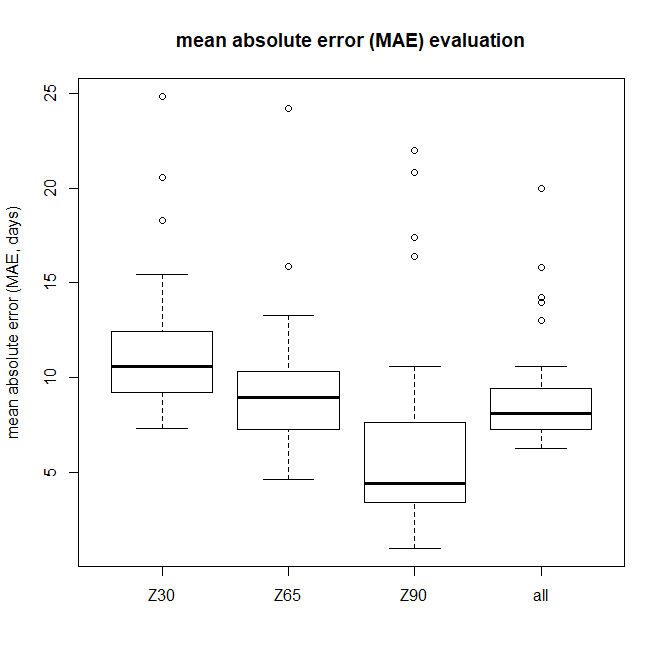
### NRMSE, where the values refer to predictions of number of days to stage Z65. The

338

### 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 |

339



340

### Figure 3

341

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

342

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

343

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

344

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

345

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

346

347

348

### beyond that range.

349

Boxplots of EF and skillT for the evaluation data are shown in [Figure 4.](#_bookmark4) The median

350

EF value of the individual modeling groups, averaged over stages, was 0.51, and 86 % of the

351

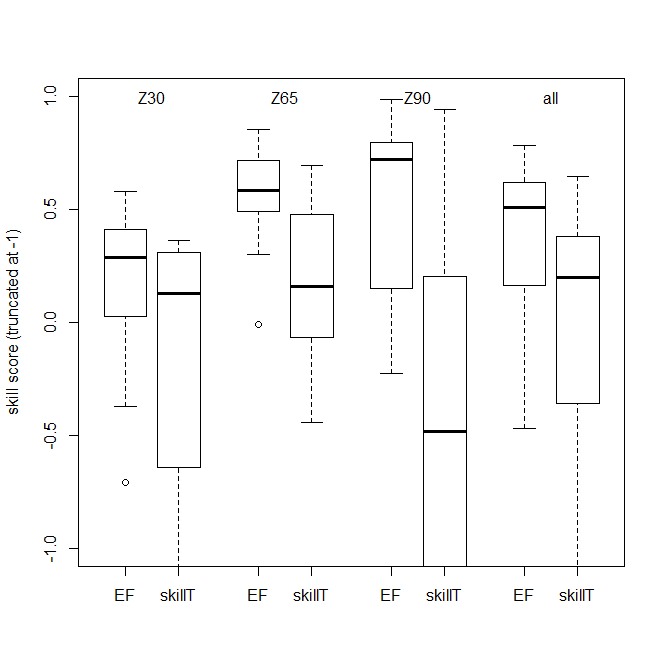
modeling groups had EF > 0. The median skillT value of the individual modeling groups,

352

353

averaged over stages, was 0.20, and 68% of the modeling groups had skillT > 0.

354



### Figure 4

355

### Boxplots of skill scores for prediction of days to Zadoks stages Z30, Z65, and Z90,

356

### and averaged over stages (all) for the evaluation data. Skill score is 1 for a modeling group

357

### that predicts perfectly, and is less than or equal to 0 for a modeling group that does no

358

### better than using average days to each stage in the calibration data (EF skill score) or than

359

### using the average number of degree days to each stage in the calibration data (skillT skill

360

### score). Boxes indicate the lower and upper quartiles. The solid line within the box is the

361

### median. Whiskers indicate the most extreme data point which is no more than 1.5 times

362

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

363

364

### beyond that range. For readability the y axis is cut off at –1.

365

The relation between overall MAE for the evaluation data and the calibration data for

366

the same modeling group is shown in [Figure 5.](#_bookmark5) The calibration value explains 46 % of the

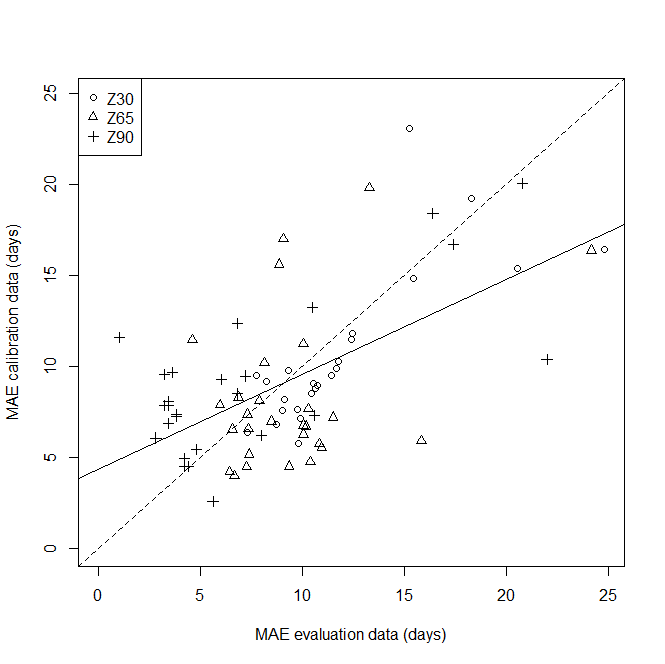
367

variability in the evaluation data (R² = 0.46) and the slope of the linear regression was

368

significantly different than 0 at the 1% level (p<0.01).

369



370

### Figure 5

371

### Relation between MAE for the evaluation data and MAE for the calibration data.

372

### Each point represents one Zadoks stage (Z30, Z65 or Z90) and one modeling group. The

373

### equation of the regression line (solid line) is y=4.3+0.52x and the slope is significantly

374

375

### different than 0 (*p*<0.01). The dashed line is the 1:1 line.

376

## Within- and between-model structure variability

377

There was substantial variability between modeling groups for each individual

378

prediction, including between modeling groups that share the same model structure

379

(Supplementary Figure S1). Averaged over the evaluation environments and over all three

380

stages Z30, Z65, and Z90, the estimated within-structure standard deviation was 4.3 days and

381

the estimated between-structure standard deviation was 11.9 days, so the within-structure

382

383

standard deviation was 36 % as large as the between-structure standard deviation.

384

1. Discussion

385

## Comparison of calibration and evaluation environments

386

The calibration and evaluation environments were drawn from the same target

387

population, namely wheat crops in the major wheat growing regions in Australia, with current

388

climate and local management practices. We compared the calibration and evaluation

389

environments for the main characteristics that are likely to affect phenology, namely

390

temperature, day length, and accumulation of vernalizing temperatures. Temperatures and

391

vernalizing durations of the evaluation environments were within the ranges of the calibration

392

environments, but the evaluation data had a larger range of day lengths than the calibration data.

393

This is the result of sampling variability, and may have led to larger prediction errors than if

394

the calibration data had a range of day lengths comparable to that of the evaluation data.

395

However, the range of days to each phenology stage for the evaluation data was always within

396

the range for the calibration data. We conclude that this study represents a case where the

397

calibration and evaluation data represent a similar range of conditions (with the caveat just

398

mentioned concerning photoperiod). This type of situation is of particular importance, for

399

example, where one wants to calibrate a crop model using current conditions and subsequently

400

test possible sowing dates within a limited range, or to compare phenology of multiple potential

401

cultivars at specific sites within the calibration domain.

402

## Prediction error

403

The evaluation here was based on data which had neither sites nor years in common

404

with the calibration data. This was thus a rigorous estimate of how well crop modeling groups

405

can predict wheat phenology for unseen sites and weather, when provided with calibration data

406

sampled from the target population. The median MAE among models averaged over phenology

407

stages was 9 days, which was substantially larger than the standard deviation of observed stages,

408

which was in the range 1-2 days. The best modeling group had an average MAE of 7 days,

409

which was still substantially larger than the measurement error. MAE values were significantly

410

larger for prediction of days to Z30 than for prediction of days to later Zadoks stages. This may

411

be due to the large variability between groups in predicting time to emergence, which is

412

discussed in more detail below. Time to emergence is a major part of the time to Z30, but a

413

smaller fraction of time to Z65 or Z90.

414

Chauhan et al. (2019) reported a value of NRMSE of 0.062 for prediction of time to

415

flowering of wheat in Australia, for a version of APSIM taking the effect of water stress on

416

phenology into account. In that study, the model was adjusted to some extent to the data used

417

for evaluation, so the reported error probably underestimates the error for new environments.

418

That reported value was in any case within the range of NRMSE values found for different

419

modeling groups here, for both the evaluation data (NRMSE here from 0.056 to 0.227) and the

420

calibration data (NRMSE here from 0.041 to 0.197). Asseng et al. (2008), using the APSIM

421

model, found RMSE of 4 days for wheat phenology predictions (mostly predictions of days to

422

anthesis) for 44 different environments in Western Australia, a level of error which was smaller

423

than the minimum RMSE of 9 days found here for the evaluation data, and even smaller than

424

the minimum RMSE of 6 days found here for the calibration data. In that study, the phenology

425

model was again adjusted to some extent to the data (S. Asseng, 2020, pers. comm.), which

426

could explain the smaller errors.

427

The above comparisons suggest that prediction errors are very roughly similar between

428

studies, but that there are differences depending on the details of the prediction problem and

429

the way prediction error is evaluated. It is clearly useful to build up a knowledge base

430

concerning phenology prediction error, as a baseline for comparison for future studies or even

431

as a default value if evaluation is not done. Contributions to the knowledge base will be all the

432

more useful, to the extent that the details of the prediction problem are clearly specified

433

(including whether it is of type interpolation or extrapolation and including a characterization

434

of the target population) and to the extent that the evaluation has a rigorous separation between

435

the predictor and the evaluation data. The present study should therefore be a valuable

436

contribution to such a knowledge base.

437

It is of interest to compare the results here with those from a study structured like the

438

present study (calibration and evaluation environments with similar characteristics, evaluation

439

data not used for model development or tuning), but where the evaluation concerned prediction

440

of wheat phenology in France (Wallach et al., 2019). To a large extent, the same modeling

441

groups participated in both studies. Specifically, the French study included 27 different

442

modeling groups, 26 of which participated in the present study. A comparison between the two

443

studies for those 26 groups is an indication of the variability of prediction errors between target

444

populations for the same modeling groups.

445

MAE averaged over the evaluation environments and over predcted stages ranged from

446

3 to 13 days (median 6 days) for the French data compared to 6 to 20 days (median 9 days) for

447

the Australian data. The target population (wheat fields in Australia versus wheat fields in

448

France) thus had a substantial effect on prediction errors. A detailed analysis of the underlying

449

reasons for the larger errors in Australia is beyond the scope of this study. However, one

450

possible contributing cause is the simulation of time to emergence. The average simulated time

451

to emergence for all French environments was 10 days after sowing, and the mean standard

452

deviation between modeling groups was 4 days. The corresponding values for the Australian

453

environments were a mean emergence time of 15 days after sowing, and a mean standard

454

deviation between modeling groups of 18 days. This very large standard deviation for the

455

Australian environments, pointing at major differences between modeling groups, may be due

456

to dry conditions in some environments and the uncertainty regarding initial soil conditions,

457

leading some models to simulate very long times to emergence (up to 107 days, Supplementary

458

Figure S1). This suggests that for Australian environments, it would be valuable to have

459

observations of time to emergence for calibration. It seems that for many modeling groups, it

460

would be worthwhile to revisit the predictions of time to emergence under conditions like those

461

of the Australian environments, taking advantage of specific modeling studies of time to

462

emergence for wheat (Lindstrom et al., 1976; Wang et al., 2009).

463

An important question in modeling is whether the same modeling groups perform best

464

for all target populations, or whether different groups are best for different target populations.

465

There is quite a bit of scatter in the graph of MAE for the Australian versus French environments

466

(Supplementary Fig. S2), but the rank correlation between the two (Kendall’s tau) is 0.31, which

467

is statistically significant (*p*=0.013). This suggests that there are modeling groups which

468

perform better than others over a wide range of environments. Once again, it is prudent to repeat

469

that this applies to the case where calibration is based on environments that are sampled from

470

the target distribution. Prediction errors for extrapolation to conditions very different than those

471

of the calibration data might behave very differently.

472

## Skill measures

473

While prediction error is of course of interest, skill scores may be even more useful, as

474

they indicate how models compare to alternative methods of prediction. Note that the EF skill

475

score used here is somewhat different than the usual definition. Here, the naïve model is based

476

solely on the calibration data, so this is in fact a feasible predictor. The more usual definition

477

of the naïve model is the mean of all the data, including the data used for evaluation. Overall,

478

all except four modeling groups had smaller MSE (were better predictors) than the naïve model.

479

The EF criterion is a rather low baseline for evaluating the usefulness of crop models

480

for predicting phenology. Our second skill measure compares model MSE and MSE of the

481

onlyT model, which assumes a constant number of degree days from sowing to each Zadoks

482

stage, and estimates that number based on the calibration data. This should be a better predictor

483

than the naïve model if photoperiod and vernalization effects are limited, and so is a more

484

stringent test of usefulness of process models. We found that the onlyT model was indeed a

485

better predictor than the naïve model. Nonetheless, 19 of the modeling groups performed better

486

than the onlyT model. It seems that in most cases here, the added complexity in crop models

487

beyond a simple sum of degree days is warranted. More generally, we suggest that

488

systematically calculating a skill measure like skillT would give valuable information about the

489

usefulness of more complex models.

490

## Model averaging

491

As found in many studies, e-median and e-mean had prediction errors comparable to

492

the best modeling groups. This confirmed previous evidence and theoretical considerations

493

showing that the use of e-mean or e-median is often a good strategy (Bassu et al., 2014; Palosuo

494

et al., 2011; Rötter et al., 2012; Wallach et al., 2018). The e-mean model is based on a simple

495

average over simulated values, so the results from every modeling group are weighted equally.

496

An open question in using model ensembles is whether it would be better to give more weight

497

to models that have smaller prediction errors for the calibration data (Christensen et al., 2010),

498

for example using Bayesian Model Averaging (Wöhling et al., 2015). The results here show

499

that phenology predictive performance for the calibration environments is significantly

500

correlated with predictive performance for new environments. This was also found to be the

501

case for a study evaluating phenology prediction by modeling groups based on phenology in

502

French environments (Wallach et al., 2019) and suggests that in these cases, it may be

503

worthwhile to use performance-weighted model ensembles. This may be due to the fact that in

504

these studies, the calibration and evaluation environments were similar to one another. In cases

505

where one is extrapolating to conditions quite different than those represented by the calibration

506

environments, performance weighting may be less useful. This once again emphasizes that it is

507

important to define for each evaluation study whether it is an evaluation of type “interpolation”

508

or “extrapolation”.

509

## Structure uncertainty and parameter uncertainty

510

Uncertainty in simulated values can arise from uncertainty in structure, from uncertainty

511

in parameter values and from uncertainty in the values of explanatory variables (Luo and

512

Schuur, 2019; Wallach et al., 2016). Here we focus on structure and parameter uncertainty. An

513

important question is the relative importance of the two, to determine priorities for reducing

514

overall uncertainty. Parameter uncertainty can arise from uncertainty in the default values of

515

those parameters that are fixed, from uncertainty in the choice of calibration approach (for

516

example, the form of the objective function or the choice of parameters to estimate) and from

517

the values of the estimated parameters, which are uncertain because there is always a limited

518

amount of data. The within-structure variability here is a measure of the uncertainty due to

519

choice of default values and calibration approach, but not of uncertainty in the values of the

520

calibrated parameters. The within-structure standard deviation here is 4.3 days, compared to a

521

between-structure standard deviation (contribution of structure) of 11.9 days. The study based

522

on French environments found a within-structure standard deviation of 5.6 days and a between-

523

structure standard deviation of 8.0 days (Wallach et al., 2019). Confalonieri et al. (2016) also

524

found that the within-structure effect was in general, but not in all cases, smaller than the

525

between-structure effect on variability.

526

Other studies have on the contrary focused on structural uncertainty versus uncertainty

527

in the calibrated parameters, without taking into account uncertainty in all the default parameter

528

values, nor uncertainty in the calibration approach chosen. Zhang et al. (2017) found that model

529

structure explained about 80 % of the variability in simulated time to heading in rice and about

530

92 % of the variability in simulated time to maturity in rice, the remainder of the variability

531

being due to parameter uncertainty. Wallach et al. (2017) found that model structure uncertainty

532

contributed about twice as much variance as parameter uncertainty to overall simulation

533

variance. It would be of interest to have a fuller treatment of parameter uncertainty, including

534

both different groups using the same model structure and an estimate of the uncertainty in the

535

parameters estimated by each group.

536

# 5. Conclusions

537

We evaluated how well 28 crop modeling groups simulate wheat phenology in

538

Australia, in the case where both the calibration data and the evaluation data were sampled from

539

fields in the major wheat growing areas in Australia under current climate and local

540

management. It is important to distinguish between interpolation type prediction, as here, and

541

extrapolation type, since they are not evaluating the same properties of modeling groups. It is

542

also important to emphasize that evaluation concerns both model structure and parameter

543

values, and therefore the modeling group and not just the underlying model structure. MAE for

544

the evaluation data here ranged from six to 20 days depending on the modeling group, with a

545

median of 9 days. About two thirds of the modeling groups performed better than a simple but

546

relevant benchmark, which predicts phenology assuming a constant temperature sum for each

547

development stage. The added complexity of crop models beyond just the effect of temperature

548

is therefore justified in most cases. As found in many other studies, the multi-modeling group

549

mean and median had prediction errors nearly as small as the best modeling group, suggesting

550

that using these ensemble predictors is a good strategy. Prediction errors for calibration and

551

evaluation environments were found to be significantly correlated, which suggests that for

552

interpolation type studies, it would be of interest to test ensemble predictors that weight

553

individual models based on performance for the calibration data. The variability due to

554

modeling group for a given model structure, which reflects part of parameter uncertainty, was

555

found to be smaller than the variability due to model structure, but was not negligible. This

556

implies that model improvement could be achieved not only by improving model structure but

557

558

also by improving parameter values.

559

# Acknowledgements

560

This work was in part supported by the Collaborative Research Center 1253 CAMPOS

561

(Project 7: Stochastic Modelling Framework), funded by the German Research Foundation

562

(DFG, Grant Agreement SFB 1253/1 2017), the Academy of Finland through projects AI-

563

CropPro (316172 and 315896) and DivCSA (316215) and Natural Resources Institute Finland

564

(Luke) through a strategic project BoostIA, the BonaRes projects ''Soil3'' (BOMA 03037514)

565

and “I4S” (031B0513I) of the Federal Ministry of Education and Research (BMBF), Germany,

566

the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s

567

Excellence Strategy - EXC 2070 – 390732324, the project BiomassWeb of the GlobeE

568

programme (Grant number: FKZ031A258B) funded by the Federal Ministry of Education and

569

Research (BMBF, Germany), the EU funded SustEs project

570

(CZ.02.1.01/0.0/0.0/16\_019/0000797), the INRA ACCAF meta-programme, the German

571

Federal Ministry of Education and Research (BMBF) in the framework of the funding measure

572

“Soil as a Sustainable Resource for the Bioeconomy – BonaRes”, project “BonaRes (Module

573

B): BonaRes Centre for Soil Research, subproject B” (grant 031B0511B), the National Key

574

Research and Development Program of China (2017YFD0300205), the National Science

575

Foundation for Distinguished Young Scholars (31725020), the Priority Academic Program

576

Development of Jiangsu Higher Education Institutions (PAPD), the 111 project (B16026), and

577

China Scholarship Council, the Agriculture and Agri-Food Canada's Project 1387 under the

578

Canadian Agricultural Partnership, the DFG Research Unit FOR 1695 ‘Agricultural

579

Landscapes under Global Climate Change – Processes and Feedbacks on a Regional Scale, the

580

U.S. Department of Agriculture National Institute of Food and Agriculture (award no. 2015-

581

68007-23133) and USDA/NIFA HATCH grant N. MCL02368, the National Key Research and

582

Development Program of China (2016YFD0300105), The Broadacre Agriculture Initiative, a

583

research partnership between University of Southern Queensland and the Queensland

584

Department of Agriculture and Fisheries, the Academy of Finland through project AI-CropPro

585

(315896), the JPI FACCE MACSUR2 project, funded by the Italian Ministry for Agricultural,

586

Food, and Forestry Policies (D.M. 24064/7303/15 of 26/Nov/2015). The field work was jointly

587

funded by CSIRO and the Grains Research and Development Corporation (GRDC) under the

588

"Adding Value to GRDC’s National Variety Trial Network" project (CSA00027). The order in

589

590

which the donors are listed is arbitrary

591

# References

592

Andarzian, Bahram, Hoogenboom, G., Bannayan, M., Shirali, M., Andarzian, Behnam, 2015.

593

Determining optimum sowing date of wheat using CSM-CERES-Wheat model. J. Saudi

594

Soc. Agric. Sci. 14, 189–199. https://doi.org/10.1016/J.JSSAS.2014.04.004

595

Archontoulis, S. V., Miguez, F.E., Moore, K.J., 2014. A methodology and an optimization

596

tool to calibrate phenology of short-day species included in the APSIM PLANT model:

597

Application to soybean. Environ. Model. Softw. 62, 465–477.

598

https://doi.org/10.1016/j.envsoft.2014.04.009

599

Asseng, S., Keating, B.A., Fillery, I.R.P., Gregory, P.J., Bowden, J.W., Turner, N.C., Palta,

600

J.A., Abrecht, D.G., 2008. Performance of the APSIM-wheat model in Western

601

Australia. F. Crop. Res. 57, 163–179.

602

Bao, Y., Hoogenboom, G., McClendon, R., Vellidis, G., 2017. A comparison of the

603

performance of the CSM-CERES-Maize and EPIC models using maize variety trial data.

604

Agric. Syst. 150, 109–119. https://doi.org/10.1016/J.AGSY.2016.10.006

605

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C.,

606

Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S.,

607

Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S.,

608

Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.-H., Kumar,

609

N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F.,

610

Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize

611

crop models vary in their responses to climate change factors? Glob. Chang. Biol. 20,

612

2301–20. https://doi.org/10.1111/gcb.12520

613

Biernath, C., Gayler, S., Bittner, S., Klein, C., Högy, P., Fangmeier, A., Priesack, E., 2011.

614

Evaluating the ability of four crop models to predict different environmental impacts on

615

spring wheat grown in open-top chambers. Eur. J. Agron. 35, 71–82.

616

https://doi.org/10.1016/j.eja.2011.04.001

617

Boote, K.J., Jones, J.W., Hoogenboom, G., 2008. Crop simulation models as tools for agro-

618

advisories for weather and disease effects on production. J. Agrometeorol. 10, 9–17.

619

Boote, K.J., Jones, J.W., Hoogenboom, G., White, J.W., 2010. The Role of Crop Systems

620

Simulation in Agriculture and Environment. Int. J. Agric. Environ. Inf. Syst. 1, 41–54.

621

Casella, G., Berger, R.L., 1990. Statistical Inference. Wadsworth and Brooks, Pacific Grove,

622

CA.

623

Ceglar, A., van der Wijngaart, R., de Wit, A., Lecerf, R., Boogaard, H., Seguini, L., van den

624

Berg, M., Toreti, A., Zampieri, M., Fumagalli, D., Baruth, B., 2019. Improving

625

WOFOST model to simulate winter wheat phenology in Europe: Evaluation and effects

626

on yield. Agric. Syst. 168, 168–180. https://doi.org/10.1016/J.AGSY.2018.05.002

627

Chauhan, Y.S., Ryan, M., Chandra, S., Sadras, V.O., 2019. Accounting for soil moisture

628

improves prediction of flowering time in chickpea and wheat. Sci. Rep. 9, 7510.

629

https://doi.org/10.1038/s41598-019-43848-6

630

Christensen, J., Kjellström, E., Giorgi, F., Lenderink, G., Rummukainen, M., 2010. Weight

631

assignment in regional climate models. Clim. Res. 44, 179–194.

632

https://doi.org/10.3354/cr00916

633

Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., Pagani, V.,

634

Cappelli, G., Vertemara, A., Alberti, L., Alberti, P., Atanassiu, S., Bonaiti, M.,

635

Cappelletti, G., Ceruti, M., Confalonieri, A., Corgatelli, G., Corti, P., Dell’Oro, M.,

636

Ghidoni, A., Lamarta, A., Maghini, A., Mambretti, M., Manchia, A., Massoni, G., Mutti,

637

P., Pariani, S., Pasini, D., Pesenti, A., Pizzamiglio, G., Ravasio, A., Rea, A., Santorsola,

638

D., Serafini, G., Slavazza, M., Acutis, M., 2016. Uncertainty in crop model predictions:

639

What is the role of users? Environ. Model. Softw. 81, 165–173.

640

https://doi.org/10.1016/j.envsoft.2016.04.009

641

Corripio, J.G., n.d. insol: Solar Radiation. R package version 1.2. 2019.

642

Efron, B., 1986. How Biased is the Apparent Error Rate of a Prediction Rule? J. Am. Stat.

643

Assoc. 81, 461–470. https://doi.org/10.1080/01621459.1986.10478291

644

Flohr, B.M., Hunt, J.R., Kirkegaard, J.A., Evans, J.R., 2017. Water and temperature stress

645

define the optimal flowering period for wheat in south-eastern Australia. F. Crop. Res. v.

646

209, 108–119. https://doi.org/10.1016/j.fcr.2017.04.012

647

Hussain, J., Khaliq, T., Ahmad, A., Akhtar, J., 2018. Performance of four crop model for

648

simulations of wheat phenology, leaf growth, biomass and yield across planting dates.

649

PLoS One 13, e0197546. https://doi.org/10.1371/journal.pone.0197546

650

Lawes, R.A., Huth, N.D., Hochman, Z., 2016. Commercially available wheat cultivars are

651

broadly adapted to location and time of sowing in Australia’s grain zone. Eur. J. Agron.

652

77, 38–46. https://doi.org/10.1016/J.EJA.2016.03.009

653

Lindstrom, M.J., Papendick, R.I., Koehler, F.E., 1976. A Model to Predict Winter Wheat

654

Emergence as Affected by Soil Temperature, Water Potential, and Depth of Planting1.

655

Agron. J. 68, 137–141. https://doi.org/10.2134/agronj1976.00021962006800010038x

656

Luo, Q., O’Leary, G., Cleverly, J., Eamus, D., 2018. Effectiveness of time of sowing and

657

cultivar choice for managing climate change: wheat crop phenology and water use

658

efficiency. Int. J. Biometeorol. 62, 1049–1061. https://doi.org/10.1007/s00484-018-

659

1508-4

660

Luo, Y., Schuur, E.A.G., 2019. Model Parameterization to Represent Processes at Unresolved

661

Scales and Changing Properties of evolving Systems. Glob. Chang. Biol. gcb.14939.

662

https://doi.org/10.1111/gcb.14939

663

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency

664

Index. J. Hydrol. Eng. 11, 597–602. https://doi.org/10.1061/(ASCE)1084-

665

0699(2006)11:6(597)

666

Palosuo, T., Kersebaum, K.C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J.E., Patil,

667

R.H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F.,

668

Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B., Rötter, R., 2011. Simulation of winter

669

wheat yield and its variability in different climates of Europe: A comparison of eight

670

crop growth models. Eur. J. Agron. 35, 103–114.

671

https://doi.org/10.1016/j.eja.2011.05.001

672

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

673

Rötter, R.P., Palosuo, T., Kersebaum, K.C., Angulo, C., Bindi, M., Ewert, F., Ferrise, R.,

674

Hlavinka, P., Moriondo, M., Nendel, C., Olesen, J.E., Patil, R.H., Ruget, F., Takáč, J.,

675

Trnka, M., 2012. Simulation of spring barley yield in different climatic zones of

676

Northern and Central Europe: A comparison of nine crop models. F. Crop. Res. 133, 23–

677

36. https://doi.org/10.1016/j.fcr.2012.03.016

678

Sadras, V.O., Monzon, J.P., 2006. Modelled wheat phenology captures rising temperature

679

trends: Shortened time to flowering and maturity in Australia and Argentina. F. Crop.

680

Res. 99, 136–146. https://doi.org/10.1016/J.FCR.2006.04.003

681

Teluguntla, P., Thenkabail, P.S., Oliphant, A., Xiong, J., Gumma, M.K., Congalton, R.G.,

682

Yadav, K., Huete, A., 2018. A 30-m landsat-derived cropland extent product of Australia

683

and China using random forest machine learning algorithm on Google Earth Engine

684

cloud computing platform. ISPRS J. Photogramm. Remote Sens. 144, 325–340.

685

https://doi.org/10.1016/J.ISPRSJPRS.2018.07.017

686

van Bussel, L.G.J., Stehfest, E., Siebert, S., Müller, C., Ewert, F., 2015. Simulation of the

687

phenological development of wheat and maize at the global scale. Glob. Ecol. Biogeogr.

688

24, 1018–1029. https://doi.org/10.1111/geb.12351

689

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P.J., van Ittersum, M.,

690

Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J.,

691

De Sanctis, G., Dumont, B., Eyshi Rezaei, E., Fereres, E., Fitzgerald, G.J., Gao, Y.,

692

Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C.,

693

Jones, C.D., Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A.-K., Maiorano, A.,

694

Minoli, S., Müller, C., Naresh Kumar, S., Nendel, C., O’Leary, G.J., Palosuo, T.,

695

Priesack, E., Ripoche, D., Rötter, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P.,

696

Streck, T., Supit, I., Tao, F., Wolf, J., Zhang, Z., 2018. Multimodel ensembles improve

697

predictions of crop-environment-management interactions. Glob. Chang. Biol. 24, 5072–

698

5083. https://doi.org/10.1111/gcb.14411

699

Wallach, D., Nissanka, S.P., Karunaratne, A.S., Weerakoon, W.M.W., Thorburn, P.J., Boote,

700

K.J., Jones, J.W., 2017. Accounting for both parameter and model structure uncertainty

701

in crop model predictions of phenology: A case study on rice. Eur. J. Agron. 88.

702

https://doi.org/10.1016/j.eja.2016.05.013

703

Wallach, D., Palosuo, T., Thorburn, P., Seidel, S.J., Gourdain, E., Asseng, S., Basso, B., Buis,

704

S., Crout, N.M.J., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S.,

705

Ghahramani, A., Hochman, Z., Hoek, S., Horan, H., Hoogenboom, G., Huang, M.,

706

Jabloun, M., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Luo,

707

Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Olesen, J.E., Poyda,

708

A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka,

709

X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber,

710

T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., 2019.

711

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

712

bioRxiv 708578. https://doi.org/10.1101/708578

713

Wallach, D., Thorburn, P., Asseng, S., Challinor, A.J., Ewert, F., Jones, J.W., Rotter, R.,

714

Ruane, A., 2016. Estimating model prediction error: Should you treat predictions as

715

fixed or random? Environ. Model. Softw. 84, 529–539.

716

https://doi.org/10.1016/j.envsoft.2016.07.010

717

Wang, H., Cutforth, H., McCaig, T., McLeod, G., Brandt, K., Lemke, R., Goddard, T.,

718

Sprout, C., 2009. Predicting the time to 50% seedling emergence in wheat using a Beta

719

model. NJAS - Wageningen J. Life Sci. 57, 65–71.

720

https://doi.org/https://doi.org/10.1016/j.njas.2009.07.003

721

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the

722

root mean square error (RMSE) in assessing average model performance. Clim. Res. 30,

723

79–82.

724

Wöhling, T., Schöniger, A., Gayler, S., Nowak, W., 2015. Bayesian model averaging to

725

explore the worth of data for soil‐plant model selection and prediction. Water Resour.

726

Res. 51, 2825–2846. https://doi.org/10.1002/2014WR016292

727

Workman, D., 2020. Worldstopexports [WWW Document]. URL

728

<http://www.worldstopexports.com/wheat-exports-country/> (accessed 3.10.20).

729

Yuan, S., Peng, S., Li, T., 2017. Evaluation and application of the ORYZA rice model under

730

different crop managements with high-yielding rice cultivars in central China. F. Crop.

731

Res. 212, 115–125. https://doi.org/10.1016/J.FCR.2017.07.010

732

Zadoks, J.C., Chzang, T.T., Konzak, C.F., 1974. A decimal code for the growth stages of

733

cereals. Weed Res. 14, 415–421. https://doi.org/10.1111/j.1365-3180.1974.tb01084.x

734

Zhang, S., Tao, F., Zhang, Z., 2017. Uncertainty from model structure is larger than that from

735

model parameters in simulating rice phenology in China. Eur. J. Agron. 87, 30–39.

736

737

https://doi.org/10.1016/j.eja.2017.04.004