



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/131303/>

Version: Accepted Version

Article:

Montesino-San Martin, M, Wallach, D, Olesen, JE et al. (2018) Data requirements for crop modelling-Appling the learning curve approach to the simulation of winter wheat flowering time under climate change. *European Journal of Agronomy*, 95. pp. 33-44. ISSN: 1161-0301

<https://doi.org/10.1016/j.eja.2018.02.003>

(c) 2018, Elsevier Ltd. This manuscript version is made available under the CC BY-NC-ND 4.0 license <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 **Data requirements for crop modelling – applying the learning curve**
2 **approach to the simulation of winter wheat flowering time under climate**
3 **change**

4 Montesino-San Martin, M.^a, Wallach, D.^b, Olesen, J.E.^c, Challinor, A.J.^d, Hoffman, M.P.^e,
5 Koehler, A.K.^d, Rötter, R.P.^{e,f}, Porter, J.R.^{a,g}

6 ^a Department of Plant and Environmental Science, University of Copenhagen, Højbakkegård Allé 30,
7 2630 Taastrup, Denmark

8 ^b National Institute for Agricultural Research (INRA), UMR AGIR, Toulouse, France

9 ^c Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

10 ^d Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds,
11 Leeds LS2 9JT, UK

12 ^e University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS),
13 Grisebachstraße 6, 37077 Göttingen, Germany

14 ^f University of Göttingen, Centre for Biodiversity and Sustainable Land Use, Büsgenweg 1, 37077
15 Göttingen, Germany

16 ^g System Montpellier SupAgro, INRA, CIHEAM-IAMM, CIRAD, Univ Montpellier, 34060
17 Montpellier, France

18

19 **Highlights**

- 20
- Learning curves are useful to diagnose data-model interactions.
- 21
- Phenology model predictions improve asymptotically with size of the calibration
- 22 dataset.
- 23
- More than 7-9 observations of anthesis did not improve model performance of
- 24 phenology models for 2050's (RCP8.5)
- 25
- More abundant but less accurate measurements can lead to similar prediction
- 26 performance.

27 **Data requirements for crop modelling – applying the learning curve**
28 **approach to the simulation of winter wheat flowering time under climate**
29 **change**

30 Montesino-San Martin, M.^a, Wallach, D.^b, Olesen, J.E.^c, Challinor, A.J.^d, Hoffmann, M.P.^e,
31 Koehler, A.K.^d, Rötter, R.P.^{e,f}, Porter, J.R.^{a,g}

32 ^a Department of Plant and Environmental Science, University of Copenhagen, Højbakkegård Allé 30,
33 2630 Taastrup, Denmark

34 ^b National Institute for Agricultural Research (INRA), UMR AGIR, Toulouse, France

35 ^c Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

36 ^d Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds,
37 Leeds LS2 9JT, UK

38 ^e University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS),
39 Grisebachstraße 6, 37077 Göttingen, Germany

40 ^f University of Göttingen, Centre for Biodiversity and Sustainable Land Use, Büsgenweg 1, 37077
41 Göttingen, Germany

42 ^g System Montpellier SupAgro, INRA, CIHEAM-IAMM, CIRAD, Univ Montpellier, 34060
43 Montpellier, France

44 **Abstract**

45 A prerequisite for application of crop models is a careful parameterization based on
46 observational data. However, there are limited studies investigating the link between quality
47 and quantity of observed data and its suitability for model parameterization. Here, we explore
48 the interactions between number of measurements, noise and model predictive skills to
49 simulate the impact of 2050's climate change (RCP8.5) on winter wheat flowering time. The
50 learning curve of two winter wheat phenology models is analysed under different assumptions
51 about the size of the calibration dataset, the measurement error and the accuracy of the model
52 structure. Our assessment confirms that prediction skills improve asymptotically with the size
53 of the calibration dataset, as with statistical models. Results suggest that less precise but larger
54 training datasets can improve the predictive abilities of models. However, the non-linear
55 relationship between number of measurements, measurement error, and prediction skills limit
56 the compensation between data quality and quantity. We find that the model performance does
57 not improve significantly with a theoretical minimum size of 7-9 observations when the model
58 structure is approximate. While simulation of crop phenology is critical to crop model
59 simulation, more studies are needed to explore data needs for assessing entire crop models.

60 **Key words:** Learning curve, Anthesis, *Triticum aestivum*, Dataset, Climate Change

61 **Data requirements for crop modelling – applying the learning curve**
62 **approach to the simulation of winter wheat flowering time under climate**
63 **change**

64 Montesino-San Martin, M.^a, Wallach, D.^b, Olesen, J.E.^c, Challinor, A.J.^d, Hoffmann, M.P.^e,
65 Koehler, A.K.^d, Rötter, R.P.^{e,f}, Porter, J.R.^{a,g}

66 ^a Department of Plant and Environmental Science, University of Copenhagen, Højbakkegård Allé 30,
67 2630 Taastrup, Denmark

68 ^b National Institute for Agricultural Research (INRA), UMR AGIR, Toulouse, France

69 ^c Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

70 ^d Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds,
71 Leeds LS2 9JT, UK

72 ^e University of Göttingen, Centre for Biodiversity and Sustainable Land Use, Büsgenweg 1, 37077
73 Göttingen, Germany

74 ^f University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS),
75 Grisebachstraße 6, 37077 Göttingen, Germany

76 ^g System Montpellier SupAgro, INRA, CIHEAM-IAMM, CIRAD, Univ Montpellier, 34060
77 Montpellier, France

78

79 **1. Introduction**

80 Models are increasingly used in impact assessments of climate change on crop production and
81 food security (Ruane et al., 2017). Models intended for these applications require suitable
82 datasets to minimize the error in the projections (Wallach, 2011). The crop modelling
83 community has repeatedly addressed and improved the definition of suitable datasets (Nix,
84 1983; Boote et al., 1999; Hunt et al., 2001; White et al., 2013). The latest efforts have been
85 made in the context of AgMIP (Rosenzweig et al, 2013) and MACSUR (Rötter et al., 2013)
86 projects. Boote et al., (2016) developed a generic qualitative method that ranks datasets based
87 on the presence or absence of input and state variables. Kersebaum et al., (2015) designed a
88 numerical classification approach where rules based on expert opinion provide scores for
89 several desirable features. The total quality score of a dataset is the summation of scores from
90 each feature. Further contributions to the definition of suitable datasets go through replacing
91 expert opinion by empirically based rules. Hence, further research is needed assessing the
92 impacts of dataset features on simulations and model performance. Confalonieri et al., (2016)
93 worked in this direction by introducing a method for assessing changes in model performance
94 depending on measurement errors. He et al., (2017) quantified the repercussions of the number
95 of seasons and state variables on their effectiveness to calibrate a crop model. The results of
96 these studies are key to elucidate the interactions between data and crop model but their
97 comparison with the rules in Kersebaum et al., (2015) is not straightforward. In order to favour

98 this comparison, features of datasets should be changed and assessed in a progressive and
99 comprehensive manner.

100 The number of observations and the measurement error (as a proxy for number of replicates)
101 are two essential features of datasets in the scoring system by Kersebaum et al., (2015). This is
102 due to their critical role in estimating model parameters and their uncertainty (Wallach et al.,
103 2011; Confalonieri et al., 2016) and the relevance of parameter uncertainty in impact
104 assessments of climate change (Wallach et al., 2011; Wallach et al., 2017). Large and accurate
105 datasets could reduce parameter uncertainty but the crop modelling community has suffered
106 from chronic data scarcity exacerbated by ensemble modelling (Rötter et al., 2011; Jones et al.,
107 2017). The maturation of new information technologies, namely mobile technology and remote
108 sensing, and the implementation of new initiatives, such as crowdsourcing, could help solving
109 this situation (Janssen et al., 2017) at the cost of accuracy. An assessment of suitable datasets
110 for crop modelling in terms of number of observations and measurement error may bring light
111 to the potential benefits of these technologies to improve crop impact projection performance.

112 The learning curve approach evaluates in a progressive manner the impact of the size and
113 measurement error of the calibration dataset on model performance. Learning curves are graphs
114 displaying the evolution of simulation errors with the size of the training dataset (Perlich et al.,
115 2003; Perlich, 2011). Errors usually evolve asymptotically with the size of the training dataset,
116 increasing for the training dataset and decreasing for the testing dataset. The shape of the curves
117 can reveal, for instance, when the model is considered to have a sufficiently large calibration
118 dataset. The size is considered large enough when greater observations produce small changes
119 in the simulation skills. However, defining when the changes are small enough depends on the
120 model application. The learning curve approach has been used in the past with statistical
121 models in the field of machine learning (e.g. Perlich, 2011 or Figueroa et al., 2012). To our
122 knowledge, the method has not been applied yet for the assessment of dataset features in crop
123 modelling.

124 Drawing the learning curves requires calibrating and evaluating the model repeatedly, changing
125 the size of the calibration dataset. This makes the process computationally demanding and data
126 intensive. Phenology combines its relevance for yield (Craufurd and Wheeler, 2009) with its
127 simple mathematical formulation and fast execution (e.g. Ceglar et al., 2011). Within the
128 phenology phases, flowering is particularly critical; it is a very sensitive phase to temperature
129 extremes (Ugarte et al., 2007) and it defines the balance between source-sink organs. Therefore,

130 the simulation of flowering time represents a practical starting point to introduce the learning
131 curve approach into crop modelling. Phenology modelling offers several working solutions
132 with different mathematical formulations (Ceglar et al., 2011; Alderman and Stanfill, 2017).
133 Learning curves are likely influenced by model structures, since prediction skills of different
134 modelling hypotheses vary due to specific error compensations forged during calibration
135 (Wallach et al., 2011). Hence, robust conclusions about data-model interactions with the
136 learning curves require the assessment of multiple structures.

137 Our study aims to analyse the influence of datasets on model simulation performance. More
138 specifically, we seek to elucidate the impact of number and measurement error of crop state
139 variables on the prediction skills of a phenology model intended for climate change
140 applications. We apply the learning curve approach which allows the progressive assessment
141 of properties of datasets and brings the opportunity to compare the evolution of model
142 performance with the scoring rules specified in the data classification system. Additionally, we
143 inspect possible compensations between size and measurement error thanks to their joint
144 analysis.

145 **2. Methods**

146 The generation of learning curves is a two-step process repeated multiple times. The first step
147 is the calibration and evaluation of the models against the training (or calibration) dataset. The
148 second step is the evaluation of the predictive skills of the model against the testing (or
149 evaluation) dataset. The training dataset varies in number of observations (quantity of
150 observations) and levels of measurement error (quality of observations). Long series of records
151 (greater than 10 seasons) of flowering dates required to construct the learning curves are scarce.
152 Hence, data is replaced by the simulations of a “*perfect model*” with structure and parameter
153 values considered to be true. The simulations from such perfect models are masked with
154 different levels of noise. This perfect model approach gives us full control over the number of
155 seasons and errors introduced in the datasets. In addition, it allows the evaluation of the
156 simulation model predictive skills against the perfect model under climate change.

157 Two phenology models for simulating anthesis dates of winter wheat under climate change are
158 considered; the Broken-Sticks (BS) and Continuous Curvilinear (CC) (Wang and Engel, 1998)
159 models. The BS is a wide-spread practical model to simulate phenology whereas the CC model
160 is considered a more realistic version from a biological perspective (Streck et al., 2008).
161 Consequently, we assume that the CC model is the “*perfect model*” and the BS and the CC

162 models are used as simulation models. Thus, two situations concerning model structures are
 163 assessed; (S1) the structure of the simulation model is an exact representation of reality (the
 164 simulation model and the “*perfect model*” are the same, both represented by the CC model),
 165 and (S2) the structure of the simulation model approximates the reality (the BS and the CC
 166 model correspond to the simulation model and the “*perfect model*” respectively). The results
 167 are used to analyse the shape of the learning curves and understand the relationships between
 168 measurements, errors and model structures.

169 2.1. Phenology models

170 The fundamental difference between the BS and the CC model is the smoother reaction of crop
 171 development to changes in temperature and photoperiod with the latter model (Fig. 1b,c). In
 172 addition, our CC model uses the vernalization response proposed by Streck et al. (2003). Here,
 173 vernalization follows a sigmoidal curve instead of the linear response in the BS model (Fig.
 174 1a). Water or nitrogen limitations are not included, assuming models are applied under optimal
 175 conditions.

176 (Fig. 1)

177 2.1.1. Vernalization response

178 The vernalization response (f_{v-BS}) in the BS model is represented from zero to one for un-
 179 vernalized and fully vernalized wheat, respectively. The parameters in this model (Eq. 1) are
 180 the base vernalization (V_{base}) and the vernalization saturation (V_{sat}). Base vernalization is the
 181 minimum vernalization required to start the accumulation of vernal degree days (VDD).
 182 Vernalization saturation is the total accumulation of VDD at which the crop is considered fully
 183 vernalised.

$$184 \quad f_{v-BS} = \min \left[1, \max \left[0, \frac{(VDD - V_{base})}{(V_{sat} - V_{base})} \right] \right] \quad (\text{Eq. 1})$$

185 In our version of the CC model, the vernalization response (f_{v-CC}) follows the description in
 186 Streck et al. (2003) (Eq. 2). Vernalization is accumulated based on a s-shaped curve. The
 187 parameter of this model is the inflection for vernalization ($V_{0.5}$), that defines the VDD
 188 accumulated when the crop is half-way vernalized.

$$189 \quad f_{v-CC} = \frac{(VDD)^5}{(V_{0.5})^5 + (VDD)^5} \quad (\text{Eq. 2})$$

190 The BS and CC models are analogous when; (1) the V_{sat} in the BS model has twice the value
 191 of $V_{0.5}$ in the CC model and V_{base} in the BS model is considered zero. The accumulation of
 192 vernal degree days (VDD) is computed by summing daily rates of vernalization. The daily rates
 193 are calculated using the Eq. 6-8 for the BS model and Eq. 9-11 for the CC model (see section
 194 2.1.3). In these equations, the cardinal temperatures, i.e. T_{base} , T_{opt} and T_{max} , equal -4, 6.5,
 195 and 17°C, for the BS model (Weir et al., 1984).

196 **2.1.2. Photoperiod response:**

197 In the BS model, the photoperiod response (f_{p-BS}) ranges from 0 to 1 when the daylight hours
 198 (dh) are higher than the minimum threshold and lower than the maximum threshold (Eq. 3).
 199 These minimum and maximum thresholds are named base photoperiod (P_{base}) and optimum
 200 photoperiod (P_{opt}), respectively.

$$201 \quad f_{p-BS} = \min \left[1, \max \left[0, \frac{(dh - P_{base})}{(P_{opt} - P_{base})} \right] \right] \quad (\text{Eq. 3})$$

202 In the CC model, the response (f_{p-CC}) also varies between 0 and 1 (Eq. 4), but its shape is
 203 negatively exponential (Fig. 1-B). The model parameters are the base photoperiod (P_{base}) and
 204 the sensitivity to changes in photoperiod (ω). Changes of P_{base} in the BS model involve
 205 modifications in the sensitivity to photoperiod. In the CC model, the sensitivity (ω) is
 206 independent from P_{base} . To resemble the reaction in both models, an empirical relationship
 207 was established between ω and P_{base} and P_{opt} in the CC model (Eq. 5).

$$208 \quad f_p = 1 - e^{[-\omega(dh - P_{base})]} \quad (\text{Eq. 4})$$

$$209 \quad \omega = 1.49 - 2.96 \cdot 10^{-2} P_{base} - 1.14 \cdot 10^{-1} P_{opt} + 2.82 \cdot 10^{-3} P_{base}^2 + 2.41 \cdot 10^{-3} P_{opt}^2$$

210 (Eq. 5)

211 With Eq. 5, the BS and CC model are defined by P_{base} and P_{opt} .

212 **2.1.3. Temperature response:**

213 The response of the crop development (f_{t-BS}) to the daily air temperature (T_a) in the BS model
 214 is considered proportional when air temperatures are between the base (T_{base}) and optimum
 215 (T_{opt}) cardinal temperatures (Eq. 6). If the temperature is above the optimum, but below its
 216 critical temperature (T_{max}), the rate of development reacts inversely proportional to the

217 difference between the air temperature and its optimum (Eq. 7). If the air temperature is below
 218 its base temperature or above its critical temperature, the daily rate of development is zero (Eq.
 219 8).

220 *if* $T_{base} < T_a < T_{opt}$ *then* $f_{t-BS} = (T_a - T_{base})$ (Eq. 6)

221 *if* $T_{opt} < T_a < T_{max}$ *then* $f_{t-BS} = (T_{opt} - T_{base})(T_{max} - T_a)/(T_{max} - T_{opt})$ (Eq. 7)

222 *if* $T_{base} > T_a$ *or* $T_a > T_{opt}$ *then* $f_{t-BS} = 0$ (Eq. 8)

223 In the CC model, the response of the crop development (f_{t-CC}) to the daily air temperature
 224 oscillates between 0 and 1. The daily rate of development is described by a curve (Eq. 9)
 225 between a minimum and maximum temperatures (T_{base} and T_{max} , respectively). The term α
 226 allows to peak the daily rate of development at T_{opt} (Eq. 10). The daily rate of development is
 227 zero if the air temperature does not reach T_{base} or exceeds T_{max} (Eq. 11).

228 *if* $T_{base} < T_a < T_{max}$ *then* $f_{t-CC} = \frac{2(T_a - T_{base})^\alpha (T_{opt} - T_{base})^\alpha - (T_a - T_{base})^{2\alpha}}{(T_{opt} - T_{base})^{2\alpha}}$ (Eq. 9)

229 $\alpha = \frac{\ln 2}{\ln \left[\frac{(T_{max} - T_{base})}{(T_{opt} - T_{base})} \right]}$ (Eq. 10)

230 *if* $T_{base} > T_a$ *or* $T_a > T_{max}$ *then* $f_{t-CC} = 0$ (Eq. 11)

231 T_{base} , T_{opt} and T_{max} are 0, 24 and 35°C in both models (Wang and Engel, 1998).

232 2.1.4. Development phase duration

233 A development stage is reached when the accumulation of the daily rates equals a threshold
 234 (TT) in the BS model. Eq. 12 shows the accumulation of daily rates between emergence and
 235 terminal spikelet. The value of the threshold (TT_{EMTS}) is estimated from field observations
 236 during calibration and is expressed in degree days (°Cd).

237 $TT_{EMTS} = \sum_{i=1}^d f_{t-BC} \cdot f_{v-BC} \cdot f_{p-BC}$ (Eq. 12)

238 In the CC model, a development stage is reached when the accumulation of daily rates (TTN)
 239 equals 1 (e.g., Eq. 13). This is achieved by using a scaling parameter (r_{max}) that represents the
 240 maximum daily development rate. The maximum development rate has an exponential form

241 based on a parameter k (Eq. 14). Eq. 13 is an example of the computation between emergence
242 and terminal spikelet.

$$243 \quad TTN_{EMTS} = r_{max,EMTS} \sum_{i=1}^d f_{t-CC} \cdot f_{v-C} \cdot f_{p-CC} \quad (\text{Eq. 13})$$

$$244 \quad r_{max} = e^{-k} \quad (\text{Eq.14})$$

245 In both models, the period from sowing to anthesis was divided into three phases; (1) from
246 sowing to emergence, (2) from emergence to terminal spikelet and (3) from terminal spikelet
247 to anthesis. The first phase is responsive to temperature, the second to temperature,
248 vernalization and photoperiod and the last one to temperature and photoperiod. We assume that
249 the duration, i.e. TTN_{SWEM} , between sowing and emergence is a constant. We also considered
250 that 45% of the duration between emergence and anthesis corresponds to the development from
251 emergence to terminal spikelet (TTN_{EMTS}), and 65% corresponds to the development from
252 terminal spikelet to anthesis (TTN_{TSAN}).

253 **2.1.5. Phenology model parameters**

254 Key parameters in the BS model reflecting genotypic differences in flowering time are
255 vernalization saturation, base photoperiod and thermal time (V_{sat} , P_{base} and TT , respectively)
256 (Bogard et al., 2014). Therefore, we selected these parameters for calibration. We picked
257 analogous parameters to calibrate the CC model; half-way vernalized, base photoperiod and
258 maximum daily rate of development ($V_{0.5}$, P_{base} and k , respectively).

259 **2.2. Perfect models and artificial flowering date records**

260 A “*perfect model*” will be used in subsequent steps in substitution of the lacking long series of
261 records of flowering dates. The “*perfect model*” has a structure and parameter values
262 considered to be true. Parameter values for this “*perfect model*” were derived from calibration
263 using actual data. These data were collected and used in simulations of the Agricultural Model
264 Inter-comparison Project (Asseng et al., 2015). The information available covered the average
265 flowering date during 1980-2010 (\bar{y}^{actual}), the average sowing date, daily maximum and
266 minimum temperatures for the same period, latitude and longitude and qualitative descriptions
267 of the sensitivities to vernalization and photoperiod of the varieties being grown. A subset of 8
268 locations (Table 1) was selected among the 60-major wheat producing regions worldwide
269 available. The locations are Netherlands, Argentina, USA, China (with continental and oceanic
270 climates), Russia, Turkey and Canada, showing a wide diversity of environmental conditions.

271 The “*perfect model*” was calibrated independently for each location using Ordinary Least
 272 Squares (OLS). The calibration concerned the parameters related to vernalization, photoperiod
 273 and thermal responses (see section 2.1.5). The OLS method searched iteratively for those
 274 parameter values (θ) that minimize the squared distance between the actual flowering date
 275 (\bar{y}^{actual}) and the simulation ($f(\theta, x_i)$) for every season (i) between 1980 and 2010 (Eq. 15).
 276 The calibration was carried out in R (version 3.3.1) using the *optim* function (R Core Team,
 277 2016).

$$278 \theta^{True} \in \underset{\theta}{\operatorname{argmin}}\{\sum_{i=1}^{30}[\bar{y}^{actual} - f(\theta, x_i)]^2\} \quad (\text{Eq. 15})$$

279 Then, we used the calibrated “*perfect model*” to generate two artificial datasets: (1) A training
 280 dataset consisting of annual dates of anthesis ($y_{i-train}^{True}$) for all seasons between 1980 and 2010
 281 using observed weather data from the AgCFSR dataset
 282 (<http://data.giss.nasa.gov/impacts/agmipcf/>) and (2) a testing dataset (y_{i-test}^{True}) consisting of
 283 annual dates of anthesis over 30 years of bias-corrected weather data. The weather data was
 284 sampled from the predicted 2050’s climate under the RCP8.5 by the GDFL-CM3 Global
 285 Climate Model (Asseng et al., 2015). We assume that there is no adaptation to climate change,
 286 hence sowing dates and cultivars were fixed for both time periods in each location.

287 (Table 1)

288 To mimic the sampling error that exists in field measurements (Kersebaum et al., 2015), we
 289 added noise (ε_i) to the flowering time datasets created with the “*perfect model*” (Eq. 16 and 20
 290 in Fig. 2). Noise values were sampled from normal distributions with mean at zero and
 291 variance σ_ε^2 . We assume hereinafter that the resulting values ($y_{i-train}^{Measure}$ or $y_{i-test}^{Measure}$) represent
 292 the long series ($i = \{1, \dots, 30\}$) of records of anthesis dates under baseline and future climate.
 293 The artificial datasets generated for the simulation experiment are listed in Table 2.

294 (Table 2)

295 **2.3. Steps to generate the learning curves**

296 The models were recalibrated (Fig. 2) using OLS (Eq. 17) and n randomly sampled seasons
 297 from the training dataset (Eq. 16). The resulting model ($f^{Sim}(\hat{\theta}, x_i)$) was used to simulate the n
 298 seasons of the calibration dataset (baseline) and the 30 seasons of the testing dataset (i.e. 2050’s
 299 anthesis dates under RCP8.5). The assessment of the performance of $f^{Sim}(\hat{\theta}, x_i)$ was based on

300 its Mean Square Error (MSE) (Eq. 18) and the Mean Square Error of Prediction (*MSEP*) (Eq.
301 20).

302 We repeated the calibration-evaluation process multiple times (Fig. 2), changing the number
303 of measurements (n) and noise levels (σ_ε^2) in the training dataset. The number of measurements
304 ranged from 5 up to 30 seasons, in steps of 2. The lower limit in the number of seasons was set
305 just above the minimum number required to calibrate 3 parameters from a mathematical point
306 of view. We also increased the noise in training set from 0 to 0.25, 1, 2.25 and 4 days². We
307 consider that the upper limit in the level of noise is a rare situation when observations are taken
308 by well-trained experimentalists. A $\sigma_\varepsilon^2 = 4$ represents a 4.6% chance to have a measurement
309 error greater than 4 days. The result of the calibrations and evaluation may vary depending on
310 the seasons and errors sampled in every combination of n and σ_ε^2 . Hence, every situation was
311 repeated 60 times to ensure that the results are independent from the sampling.

312 We consider two model structures, so we had two different situations regarding the choice of
313 the true (f^{True}) and the simulation (f^{Sim}) model. The aim was to explore how the structure
314 affected the learning curves. In the first situation (S1), we assume that the simulation model
315 represents perfectly the mechanisms of the true system (i.e., $f^{Sim} = f^{True} = CC$). The second
316 situation (S2) assumes that the model is just an approximation ($f^{Sim} \neq f^{True}$, being $f^{Sim} =$
317 *BS* and $f^{True} = CC$).

318 (Fig. 2)

319 **2.4. Model performance, number of measurements, noise and data requirements**

320 In statistics, it is known that the *MSEP* reacts to the size of the training dataset (n) following
321 Eq. 21 for linear regressions models (Wallach et al., 2013). The magnitude of *MSEP* depends
322 on model errors (σ_ε^2) and the number of parameters being calibrated (p). The theory is valid
323 when (1) the linear regressions represent suitably the system and (2) the training and testing
324 datasets belong to the same population.

$$325 \quad MSEP = \sigma_\varepsilon^2 \left(\frac{p}{n} + 1 \right) \quad (\text{Eq. 21})$$

326 Phenology models in climate impact assessments contradict both premises; (1) they are far
327 from linear and (2) the baseline (training datasets) and future climate flowering dates (testing
328 dataset) represent different populations. Instead of Eq. 21, the relationship will be expressed
329 according to the power law (Eq. 22). In Eq. 22, a and b represent the learning rate and learning

330 limit, respectively. The learning rate (a) represent the portions of the $MSEP$ that is reducible
 331 with larger training datasets (n). Conversely, the learning limit (b) constitutes the unreducible
 332 part of $MSEP$. Eq. 22 is a more general form of Eq. 21 since a and b can adopt the values $a =$
 333 $p\sigma_\varepsilon^2$ and $b = \sigma_\varepsilon^2$.

$$334 \quad f_{MSEP}(n) = \frac{a}{n} + b \quad (\text{Eq. 22})$$

335 Based on Eq. 22, we explore the model data requirements by estimating the smallest calibration
 336 dataset that does not trigger significant improvement in the prediction errors under future
 337 climate, i.e. the lower value of n that makes $\Delta MSEP = f_{MSEP}(n) - f_{MSEP}(n + 1)$ crossing a
 338 threshold. We will consider that $\Delta MSEP$ is trivial when the error is reduced less than 1 day in
 339 one of the 30 seasons under climate change ($t = 1^2/30 \approx 0.03$). The use of $\Delta MSEP$ to determine
 340 the data requirements focuses on the role of the size of the dataset rather than any other factor
 341 affecting the $MSEP$.

342 **3. Results**

343 **3.1. “Perfect model” calibration, training and testing datasets**

344 The calibration of the “*perfect model*” yielded good representation of the observed average
 345 flowering date under baseline climate (Table 1 and Fig. 3). The 30-year means of the annual
 346 flowering date simulated by the Continuous Curvilinear (CC) model were nearly equal the
 347 actual averages (Table 1). The simulations carried out with the “*perfect model*” under climate
 348 change conditions (Fig. 3) led to earlier flowering dates. Flowering dates with the CC model
 349 occurred between 6-17 days earlier than in the baseline. Russia was the only location where
 350 the model predicted a later flowering (3 days).

351 (Fig. 3)

352 **3.2. Size of the training dataset, measurement error and model performance – S1:** 353 **model structures are correct ($f^{True} = f^{Sim}$)**

354 Several calibrations and evaluations of the CC model were carried out following the algorithm
 355 described above. The calibration dataset was changed with respect to the number of seasons
 356 (n) and levels of noise (σ_ε^2) and the model performance was tested in terms of mean squared
 357 errors (MSE and $MSEP$). The squared errors of the CC models can be seen in Figs. 4-5. In
 358 general, Fig. 4 shows an increase of MSE and a decrease of $MSEP$ with greater sizes of the

359 calibration dataset (n). The MSE and $MSEP$ tend to the variance of noise, i.e. 0.25, 1, 2.25, and
360 4 days², without reaching it for the range of n explored. It should be noted that the graphs differ
361 in the range of squared errors displayed on the y-axis for visualization purposes. Results show
362 that prediction performance ($MSEP$) worsens proportionally with the level of measurement
363 error in both calibration and evaluation ($R^2=0.99$) (Fig. 5a).

364 We adjusted Eq. 22 by estimating the learning rate (a) and learning limit (b) that fitted best the
365 median $MSEs$ and $MSEPs$ among locations (solid lines in Fig. 4). The learning rate is negative
366 when the trajectory ascends (MSE) and positive otherwise ($MSEP$). The curves represented
367 well the increase of the MSE with the number of observations. The variability of the MSE
368 explained by the power law varied between 0.95 and 0.99 for the CC model (Fig. 4). Curves
369 represented slightly worse the results of the $MSEPs$: The coefficients of determination dropped
370 from 0.95-0.99 for the $MSEs$ to 0.93-0.97 for the $MSEPs$ of the CC model. Fig. 4 shows how
371 the $MSEPs$ spread out compared to the $MSEs$, as the errors varied considerably between
372 locations.

373 (Fig. 4)

374 (Fig. 5)

375 We further explored the relationship between our results and theory (Eq. 21). Given the
376 proportionality between $MSEPs$ and σ_ε^2 (Fig. 6a), we computed their ratio ($MSEP/\sigma_\varepsilon^2 =$
377 $MSEP'$) to remove the differences among $MSEPs$ caused by noise. According to theory,
378 $MSEP'$ should follow $p/n + 1$. We adjusted Eq. 22 to represent the $MSEP'$. Based on Eq. 21,
379 a should be equal to p and b equal to 1 (in this case, $a = 3$ and $b = 1$). Our results approached
380 reasonably well to theory (Fig. 7a); the model was significant ($p - value = 3.64 \cdot 10^{-6}$) and
381 represented well the variations of $MSEP'$ ($R^2 = 0.86$). Additionally, the estimated model
382 coefficient remained close to the theoretical values with $\hat{a} = 3.92(\pm 0.46)$ and $\hat{b} =$
383 $1.46(\pm 0.04)$.

384 (Fig. 6)

385 A larger n and higher σ_ε^2 had positive and negative impacts, respectively, on the prediction
386 performance (Fig. 4-5a). To investigate the compensations between n and σ_ε^2 we rearranged
387 Eq. 21-22 to calculate the n required to reach a specific $MSEP$ ($n = \hat{a}/(MSEP/\sigma_\varepsilon^2) - \hat{b}$).
388 Combined sequences of $MSEP$ and σ_ε^2 were fed into the equation to build the response surfaces

389 seen in Fig. 7a. The graph shows the n (z-axis) depending on the $MSEP$ (x-axis) and the σ_ε^2 (y-
390 axis). The non-equidistant contour lines in Fig. 8a depict the non-linearities between $MSEP$
391 and n captured in Eq. 21 and 22. The straightness of the contour lines reflects the linear
392 relationship between $MSEP$ and σ_ε^2 represented in Eq. 21. We inspected whether larger but less
393 precise datasets could lead lower $MSEPs$ than smaller but more precise datasets. The dashed
394 black line in Fig. 7a shows one case where the $MSEP$ is reduced from 5 day² to 4 day² (in steps
395 of 0.25 day²) by using training datasets with size n equal to 4, 6, 9, 13 and 30 and noise levels
396 equal to 2.22, 2.25, 2.37, 2.41 and 2.51 days², respectively. Eqs. 22-23 and Fig. 7a confirm that
397 it is possible in theory to compensate the lack of precision in the measurements with more
398 seasons observed. However, the equations and the results in Fig. 7a highlight two major
399 limitations for this type of compensations; (1) the noise imposes a minimum limit of the $MSEP$
400 ($\lim_{n \rightarrow \infty} MSEP = \sigma_\varepsilon^2$) and (2) n changes very quickly with $MSEP$ and σ_ε^2 ($n = a/(MSEP - b)$),
401 becoming rapidly very large and practically unfeasible.

402 (Fig. 7)

403 Data required to reach the threshold $\Delta MSEP < 0.03$ was calculated using Eqs. 21-22. The
404 improvements in model performance were not significant when the size of training dataset
405 reached the number of observations appearing in Table 3 (column Situation S1). For instance,
406 models showed no meaningful improvement in prediction skills with training datasets larger
407 than 11(± 1) measurements when noise was $\sigma_\varepsilon^2 = 1$. The data required increased with growing
408 levels of noise.

409 (Table 3)

410 Every square dot in Fig. 4 represents the squared error ($MSE/MSEP$) of a particular location.
411 The dispersion of the $MSEP$ values reveals that the variation between locations is large. To
412 explore the reasons behind these differences, Eq. 22 was adjusted independently for the results
413 of each location. We inspected whether the variance of the training population (flowering dates
414 1980-2010) might be behind the differences in the location-specific learning rates (a) and limits
415 (b) of the $MSEPs$. Fig. 8 displays the a and b obtained from the $MSEPs$ for each location and
416 noise level on the x-axis. On the y-axis, the graph shows the a' and b' obtained from a
417 regression based on noise (σ_ε^2) and the variance of the training dataset (σ_T^2). We found that the
418 variance of the training dataset and the variance of noise in the measurements explained most
419 of the variability in the learning rates (Fig. 8a). The regression of a' based on σ_ε^2 and σ_T^2 shows

420 a good fit between the actual and the estimated learning rates ($R^2=0.85$). The variance of
421 training dataset and its product with the variance of noise ($\sigma_T^2 \cdot \sigma_\epsilon^2$) were highly significant ($p <$
422 0.01) to explain the variations in learning rates. The variability in b' (Fig. 8b) was only
423 significantly explained ($p < 0.01$) by the noise ($R^2=0.98$).

424 (Fig. 8)

425 **3.3. Size of the training dataset, measurement error and model performance – S2:** 426 **model structures are approximations ($f^{True} \neq f^{Sim}$)**

427 The entire process was repeated, but this time the true model and the simulation model were
428 different. In Fig. 9, the CC model represents the true mechanism ($f^{True} = CC$), and the BS
429 model is used as an approximation ($f^{Sim} = BS$). Curves with the shape of Eq. 22 were adjusted
430 to the results of the MSE and $MSEP$ (Fig. 9). $MSEs$ and $MSEPs$ evolved asymptotically with
431 the size of the training dataset as in S1. Eq. 22 represented well the variations of the $MSEs$
432 (grey dots in Fig. 9); R^2 ranged between 0.96 and 0.99 for the BS model simulations (black
433 lines in Fig. 9) and dropped to 54-90% for the $MSEPs$ with the BS model (red lines in Fig. 9).
434 The results show that the prediction error increased linearly with the noise ($R^2=0.99$) (Fig. 5b).
435 The values of $MSEs$ and $MSEPs$ were well represented by a linear regression with an intercept
436 (k) greater than zero. This intercept shows the average cost of an approximated model structure,
437 which was 1.10 and 3.68 days² for the MSE and $MSEP$, respectively. The influence of model
438 structure is also illustrated by a wider spread of $MSEPs$ among locations in S2 than in S1 (red
439 dots in Fig. 9). Structural model errors worsened prediction performance to a greater or lesser
440 extent depending on the location. For instance, the $MSEPs$ were high and roughly decreased
441 with the size of training dataset (n) when applying the BS model in Turkey (outliers in Fig. 9).
442 The flat evolution of the error represents the need of structural model improvements.

443 (Fig. 9)

444 The impact of structural error on $MSEP$ was removed by subtracting the location-specific
445 minimum prediction error obtained with zero noise training datasets (k_{loc}). As in S1, the
446 differences among $MSEPs$ caused by noise were eliminated by dividing $MSEP$ by σ_ϵ^2
447 ($MSEP' = (MSEP - k_{loc})/\sigma_\epsilon^2$). We adjusted Eq. 22 to $MSEP'$ by calibrating a and b (Fig.
448 6b). The model was significant ($p - value = 7.54 \cdot 10^{-6}$) and explained a high portion of the
449 variability in $MSEP'$ ($R^2 = 0.84$). The estimated values of the coefficients \hat{a} and \hat{b} were
450 4.46(± 0.56) and 1.25(± 0.05), so \hat{a} was slightly greater than the value in S1 and \hat{b} was similar

451 to S1 and its theoretical value. Therefore, the model structure hampered the parameter
452 estimation, since \hat{a} is the portion of $MSEP$ attributed to parameter estimation error.

453 We estimated the n (contour lines in Fig. 7b) based on a given $MSEPs$ and σ_ε^2 . The specific
454 version of Eq. 21-22 to S2 was rearranged ($n = \hat{a}/((MSEP - k_{loc})/\sigma_\varepsilon^2) - \hat{b}$). Compared to
455 S1, contour lines in S2 are offset to the lower right corner of the graph. This indicates that the
456 number of observations needed to reach a prediction performance in S2 is larger than in S1.
457 The contours lines are more horizontal than in S1, representing a lower response of n to the
458 noise in the training dataset. Results suggest (black dots in Fig. 7b) that the training datasets of
459 n equal to 5, 7, 12 and 32 can reduce the prediction error from 5 days² to 4.25 days² (in steps
460 of 0.25 day²) with increasing noises (1.06, 1.07, 1.09 and 1.10 days²).

461 Data requirements were estimated by finding the smallest n that surpassed the threshold with
462 the learning rates and limits specific to each location. The models stopped significantly
463 improving model predictions at the n 's specified in Table 3 under the column for Situation S2.
464 There is an increase in data requirements when the model structure changed from perfect to
465 approximate (Table 3).

466 As in S1, Eq. 22 was fitted independently to the results from each location, extracting the values
467 of a and b . To understand the differences between locations, we explored the relationship
468 between the learning rate and limits with the training population variance (σ_T^2) and level of
469 noise (σ_ε^2). Fig. 10 is similar to Fig. 8, but with the results from S2. The results showed a worse
470 approximation between actual and estimated learning rates (a vs. a') ($R^2 = 0.69$) and learning
471 limits (b vs. b') ($R^2 = 0.60$) than in S1 (Fig. 10). The terms σ_T^2 and ($\sigma_T^2 \cdot \sigma_\varepsilon^2$) were highly
472 significant ($p < 0.01$) for explaining the variations of the learning rates among locations. The
473 variation of the learning limit among locations was significantly explained by the terms σ_ε^2 and
474 σ_T^2 . Fig. 10b shows that σ_ε^2 and σ_T^2 alone did not represent well the learning limits in locations
475 such as Turkey (green squares). The shift of the points towards the right while remaining
476 parallel to the 1:1 line indicates existence of an additional locations-specific constant term
477 explaining the learning limit.

478 (Fig. 10)

479 **4. Discussion**

480 As in other disciplines (e.g., Figueroa et al., 2012), the learning curves have proved to be useful
481 for assessing crop phenology models in terms of elucidating the relationship between datasets
482 and prediction performance and defining the suitable size of the calibration datasets given a
483 prediction error target.

484 We explored the interaction between the number of measurements in the calibration dataset
485 and the prediction skills of two phenology models. The results show a nonlinear relationship
486 between prediction error and the size of the calibration dataset. The system developed by
487 Kersebaum et al. (2015) scores the quality of modelling datasets in a linear fashion with the
488 number of seasons observed. The existing statistical theory and our results suggest that a
489 nonlinear power-law scoring system would be more representative. According to the effect of
490 noise on model squared error, we observed that prediction performance improves
491 proportionally with reductions in measurement error. The relationships between size, noise of
492 datasets and model skills (Eq. 21-22) indicate that it could be possible to improve the
493 predictions skills using less precise but more abundant datasets ($n = a/(MSEP/\sigma_\varepsilon^2) - b$).
494 Therefore, satellite images, for instance, could help observing ground-based phenology
495 (Sakamoto et al., 2005) to improve climate change impact assessments. Their spatial and
496 temporal coverage (large n) may compensate the errors arising from calibration and
497 atmospheric disturbances (high σ_ε^2) (Studer et al., 2007). However, compensations between
498 noise and size of datasets might be limited by the non-linear growth in size needed to
499 compensate for measurement error. Further assessments investigating these synergies are
500 needed.

501 We estimated that 5-7 observations of flowering dates were enough to conduct impact
502 assessments under 2050's climate change conditions. These results correspond to 0.25 day²
503 measurement error and perfect model structures. However, model structures are known to be
504 imperfect representations of the agricultural systems (Rötter et al., 2011). Therefore, S2 is more
505 realistic representation of the situation in crop modelling. In our experiment, structural
506 approximations (S2) translated into an increase of prediction error. The error increase was
507 specific to each model and location. Structural errors also interfered with parameter estimation,
508 increasing the data requirements. Therefore, moving from S1 to S2 caused an increase of data
509 requirements to 7-9 with 0.25 day² of measurement error. The number of field measurements
510 (years) usually available to compare observations and simulation ranges from 5 to 10 before
511 the cultivar becomes obsolete. This number of measurements is around the recommended
512 minimum number estimated in our analysis. However, noise in field observations is likely

513 larger than 0.25 days^2 . To get more measurements in the same time period, multi-
514 environmental trials or experiments with multiple sowing dates have to be conducted, which
515 goes in line with recommendations by He et al. (2017). Strictly, neither structures can be
516 considered correct, nor are parameter values true. For these reasons, the results obtained with
517 this kind of assessment are merely theoretical and advisory. These recommendations can vary
518 among locations: the data required depends on the learning rate and results show that it varies
519 with the inter-annual flowering variability of the training population (Fig. 8 and 10). Therefore,
520 the suitable size of the dataset could be larger in places where there is greater variability among
521 seasons.

522 The estimates of data requirements made in this assessment concern phenology models used
523 on their own for climate change impact assessments for 2050's under the RCP8.5 scenario.
524 Results cannot be extended to phenology models embedded in crop models, even when
525 phenology parameters are independently calibrated as the initial process of model calibration
526 (e.g., Angulo et al., 2013). Generally, the number of parameters being calibrated is greater than
527 3 (p in Eq. 21) since more than one phase of the development is involved (e.g. flowering and
528 maturity). A greater number of parameters may raise the learning rate (a in Eq. 22), therefore
529 increasing the n (number of observations) needed to surpass the threshold. Additionally, the
530 information available to calibrate the models involves observations of multiple phases,
531 meaning more information to calibrate the model. These aspects may change the shape of the
532 learning curves and the suitable number of measurements required for calibration. Another
533 factor influencing the learning rate is the inter-annual variability of the flowering time at the
534 time being projected (σ_T^2). This variability of the flowering time may change over time in some
535 locations, for instance due to more variable temperatures in the future (Craufurd and Wheeler,
536 2009). Therefore, data requirements would vary depending on the time horizon being projected.
537 Future work needs to include more phases and locations and time horizons in the learning curve
538 approach and the upscaling of the learning curves to whole crop models.

539 **5. Conclusions**

540 To our knowledge, there is no study to date giving statistical evidence about the effects of the
541 size and measurement error of the datasets on crop modelling for climate impact assessment.
542 Here we applied the learning curve approach to crop modelling, using phenology models
543 varying the dataset features in a progressive manner. Learning curves might be promising tools

544 to explore the balance between the size of the dataset, measurement error and model
545 performance to provide practical guidance.

546 Prediction skill reacted non-linearly to the size of the training dataset according to power-law.
547 Approximate phenology models required at least 7-9 observations to reach negligible
548 improvements with larger datasets to predict the flowering time for the 2050's under the
549 RCP8.5 scenario. The analysis based on learning curves also suggested that improvements in
550 predictions can be achieved with less precise but more abundant datasets. Based on the theory,
551 these compensations follow $n = a / ((MSEP / \sigma_{\epsilon}^2) - b)$. Therefore, new satellite-based
552 monitoring techniques could potentially improve simulations despite their errors. The extent of
553 improvement will depend on the noise and number of seasons used as a training set and more
554 studies are needed.

555 The estimates made in this study concern the phenology models used independently for impact
556 studies of flowering in 2050's under RCP8.5. We encourage further efforts to adapt the learning
557 curve approach to complete crop models and explore the requirements for projecting different
558 time horizons.

559 **Acknowledgements**

560 The present study was carried out in the context of CropM within the FACCE-MACSUR
561 knowledge hub. The contributions of MMSM, JEO and JRP were funded by Innovation Fund
562 Denmark. RPR and MPH received financial support from the 'Limpopo Living Landscapes'
563 project (SPACES program; 794 grant number 01LL1304A) funded by the German Federal
564 Ministry of Education and Research (<http://www.bmbf.de/en/>), and from the IMPAC3 project,
565 funded by the German Federal Ministry of Education and Research, grant number 031A351A.
566 AKK and AJC are also supported by the CGIAR Research Program on Climate Change,
567 Agriculture and Food Security (CCAFS), which is carried out with support from CGIAR

568 **References:**

- 569 Alderman, P. D., & Stanfill, B. (2017). Quantifying model-structure-and parameter-driven
570 uncertainties in spring wheat phenology prediction with Bayesian analysis. *European Journal*
571 *of Agronomy*, 88, 1-9.
- 572 Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S., & Ewert, F. (2013). Implication of
573 crop model calibration strategies for assessing regional impacts of climate change in Europe.
574 *Agricultural and Forest Meteorology*, 170, 32-46.
- 575 Asseng, S., et al. (2015). Rising temperatures reduce global wheat production. *Nature Climate*
576 *Change*, 5(2), 143-147.
- 577 Bogard, M., et al. (2014). Predictions of heading date in bread wheat (*Triticum aestivum* L.)
578 using QTL-based parameters of an ecophysiological model. *Journal of experimental botany*,
579 eru328.
- 580 Boote, K.J. 1999. Data requirements for model evaluation and techniques for sampling crop
581 growth and development. In: G. Hoogenboom, P.W. Wilkens, and G.Y. Tsuji, editors, *DSSAT*
582 *Version 3. A decision support system for agrotechnology transfer*. Vol. 4. University of Hawaii,
583 Honolulu. p. 201–216.
- 584 Boote, K. J., et al., (2016). Sentinel site data for crop model improvement—definition and
585 characterization. *Improving Modeling Tools to Assess Climate Change Effects on Crop*
586 *Response*, (advagricsystem7), 125-158.
- 587 Ceglar, A., Črepinšek, Z., Kajfež-Bogataj, L., & Pogačar, T. (2011). The simulation of
588 phenological development in dynamic crop model: the Bayesian comparison of different
589 methods. *Agricultural and Forest Meteorology*, 151(1), 101-115.
- 590 Confalonieri, R., Bregaglio, S., & Acutis, M. (2016). Quantifying uncertainty in crop model
591 predictions due to the uncertainty in the observations used for calibration. *Ecological*
592 *Modelling*, 328, 72-77.
- 593 Craufurd, P. Q., & Wheeler, T. R. (2009). Climate change and the flowering time of annual
594 crops. *Journal of Experimental Botany*, 60(9), 2529-2539.

595 Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size
596 required for classification performance. *BMC medical informatics and decision making*, 12(1),
597 8.

598 He, D., Wang, E., Wang, J., & Robertson, M. J. (2017). Data requirement for effective
599 calibration of process-based crop models. *Agricultural and Forest Meteorology*, 234, 136-148.

600 Hunt, L. A., White, J. W., & Hoogenboom, G. (2001). Agronomic data: advances in
601 documentation and protocols for exchange and use. *Agricultural Systems*, 70(2), 477-492.

602 Janssen, S. J., Porter, C. H., Moore, A. D., Athanasiadis, I. N., Foster, I., Jones, J. W., & Antle,
603 J. M. (2017). Towards a new generation of agricultural system data, models and knowledge
604 products: Information and communication technology. *Agricultural systems*, 155, 200-212.

605 Jones, J. W., et al. (2017). Toward a new generation of agricultural system data, models, and
606 knowledge products: State of agricultural systems science. *Agricultural Systems*.

607 Kersebaum, K. C., et al. (2015). Analysis and classification of data sets for calibration and
608 validation of agro-ecosystem models. *Environmental Modelling & Software*, 72, 402-417.

609 Nix, H. A. (1983). Minimum data sets for agrotechnology transfer. In *Proceedings of the*
610 *International Symposium on Minimum Data Sets for Agrotechnology Transfer* (pp. 181-188).

611 Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: A
612 learning-curve analysis. *Journal of Machine Learning Research*, 4(Jun), 211-255.

613 Perlich, C. (2011). Learning curves in machine learning. In *Encyclopedia of Machine Learning*
614 (pp. 577-580). Springer US.

615 R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation
616 for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.

617 Rosenzweig, C., et al. (2013). The agricultural model intercomparison and improvement
618 project (AgMIP): protocols and pilot studies. *Agricultural and Forest Meteorology*, 170, 166-
619 182.

620 Rötter, R.P., et al., (2013). Challenges for agro-ecosystem modelling in climate change risk
621 assessment for major European crops and farming systems. In: *Impacts World 2013*

622 Conference Proceedings. Potsdam Institute for Climate Impact Research, Potsdam, pp. 555-
623 564.

624 Ruane, A. C., et al. (2017). An AgMIP framework for improved agricultural representation in
625 IAMs. *Environmental Research Letters*.

626 Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., & Ohno, H. (2005).
627 A crop phenology detection method using time-series MODIS data. *Remote sensing of*
628 *environment*, 96(3), 366-374.

629 Streck, N. A., Weiss, A., Xue, Q., & Baenziger, P. S. (2003). Improving predictions of
630 developmental stages in winter wheat: a modified Wang and Engel model. *Agricultural and*
631 *Forest Meteorology*, 115(3), 139-150.

632 Streck, N. A., Weiss, A., & Baenziger, P. S. (2003). A generalized vernalization response
633 function for winter wheat. *Agronomy journal*, 95(1), 155-159.

634 Streck, N. A., Lago, I., Gabriel, L. F., & Samboranza, F. K. (2008). Simulating maize
635 phenology as a function of air temperature with a linear and a nonlinear model. *Pesquisa*
636 *Agropecuária Brasileira*, 43(4), 449-455.

637 Studer, S., Stöckli, R., Appenzeller, C., & Vidale, P. L. (2007). A comparative study of satellite
638 and ground-based phenology. *International Journal of Biometeorology*, 51(5), 405-414.

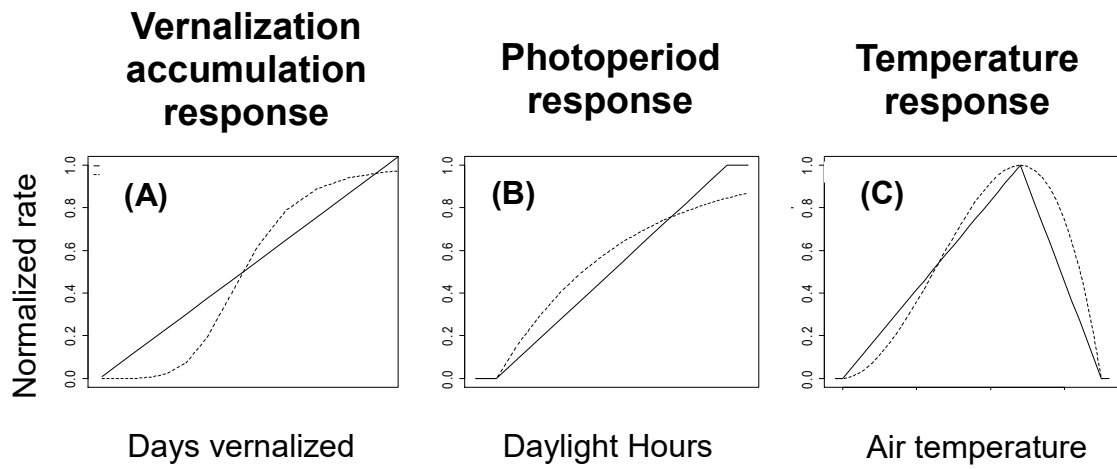
639 Ugarte, C., Calderini, D. F., & Slafer, G. A. (2007). Grain weight and grain number
640 responsiveness to pre-anthesis temperature in wheat, barley and triticale. *Field Crops Research*,
641 100(2), 240-248.

642 Wallach, D. (2011). Crop model calibration: a statistical perspective. *Agronomy Journal*,
643 103(4), 1144-1151.

644 Wallach, D., Makowski, D., Jones, J. W., & Brun, F. (2013). Working with dynamic crop
645 models: methods, tools and examples for agriculture and environment. Academic Press.

646 Wallach, D., Nissanka, S. P., Karunaratne, A. S., Weerakoon, W. M. W., Thorburn, P. J.,
647 Boote, K. J., & Jones, J. W. (2017). Accounting for both parameter and model structure
648 uncertainty in crop model predictions of phenology: a case study on rice. *European Journal of*
649 *Agronomy*, 88, 53-62.

- 650 Wang, E., & Engel, T. (1998). Simulation of phenological development of wheat crops.
651 *Agricultural systems*, 58(1), 1-24.
- 652 Weir, A. H., Bragg, P. L., Porter, J. R., & Rayner, J. H. (1984). A winter wheat crop simulation
653 model without water or nutrient limitations. *The Journal of Agricultural Science*, 102(2), 371-
654 382.
- 655 White, J. W., et al., (2013). Integrated description of agricultural field experiments and
656 production: The ICASA Version 2.0 data standards. *Computers and electronics in agriculture*,
657 96, 1-12.



659

660 **Fig. 1: Normalized responses of crop development to vernalization (A), photoperiod (B)**
661 **and temperatures (C) simulated by the Broken-Sticks Model (solid line) and the**
662 **Continuous Curvilinear Model (dashed line)**

Steps to obtain the learning curves:

a. Sample n measurement errors ε_i from $N(0, \sigma_\varepsilon^2)$

b. Select n measurements randomly from 1980-2010

c. Build the calibration dataset

$$\begin{aligned} \text{calibration dataset} &= \{y_{1-\text{train}}^{\text{Measure}}, \dots, y_{n-\text{train}}^{\text{Measure}}\} = \\ &= \{y_{1-\text{train}}^{\text{True}} + \varepsilon_1, \dots, y_{n-\text{train}}^{\text{True}} + \varepsilon_n\} = \\ &= \{f^{\text{True}}(\theta^{\text{True}}, x_{1-\text{train}}) + \varepsilon_1, \dots, f^{\text{True}}(\theta^{\text{True}}, x_{n-\text{train}}) + \varepsilon_n\} \end{aligned} \quad (\text{Eq.16})$$

d. Calibrate the model by OLS using *calibration dataset*

$$\hat{\theta} \in \operatorname{argmin} \left\{ \sum_{i=1}^n [y_{i-\text{train}}^{\text{Measure}} - f^{\text{Sim}}(\theta, x_{i-\text{train}})]^2 \right\} \quad (\text{Eq. 17})$$

e. Compute *MSE* of the model for those *obs*

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(y_i^{\text{Measure}} - f^{\text{Sim}}(\hat{\theta}, x_i) \right)^2 \quad (\text{Eq.18})$$

f. Build the testing dataset

$$\begin{aligned} \text{testing} &= \{y_{1-\text{test}}^{\text{Measure}}, \dots, y_{30-\text{test}}^{\text{Measure}}\} = \{y_{1-\text{test}}^{\text{True}} + \varepsilon_1, \dots, y_{n-\text{test}}^{\text{True}} + \varepsilon_n\} = \\ &= \{f^{\text{True}}(\theta^{\text{True}}, x_{1-\text{test}}) + \varepsilon_1, \dots, f^{\text{True}}(\theta^{\text{True}}, x_{30-\text{test}}) + \varepsilon_{30}\} \end{aligned} \quad (\text{Eq.19})$$

g. Estimate the *MSEP* of the model under climate change

$$\text{MSEP} = \frac{1}{30} \sum_{i=1}^{30} \left(y_i^{\text{Measure}} - f^{\text{Sim}}(\hat{\theta}, x_i) \right)^2 \quad (\text{Eq.20})$$

h. Repeat b-g 60 times

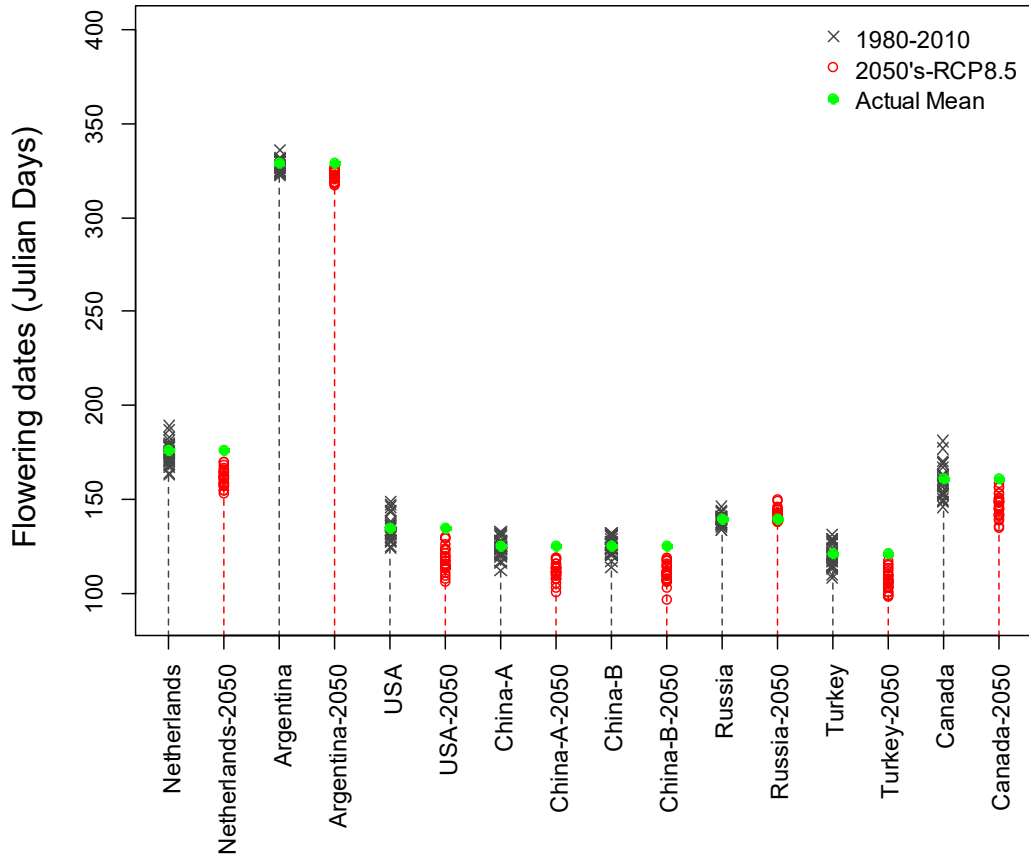
i. Repeat b-g increasing n from 5 to 30 in steps of 2

j. Repeat a-b increasing σ_ε from 0 to 2 in steps of 2.

663

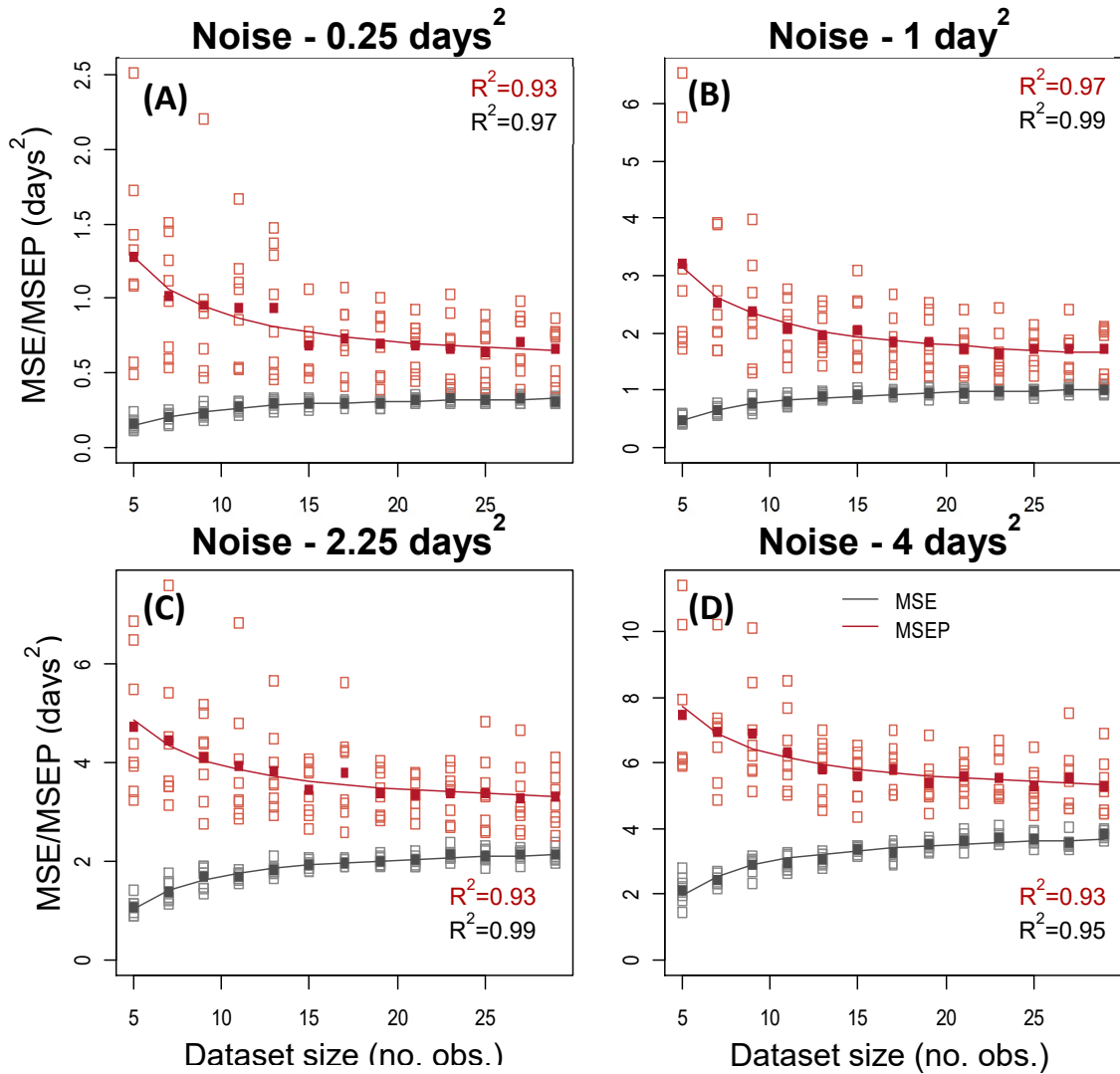
664 **Fig. 2: Outline of the process to obtain the learning curves.**

Generated datasets



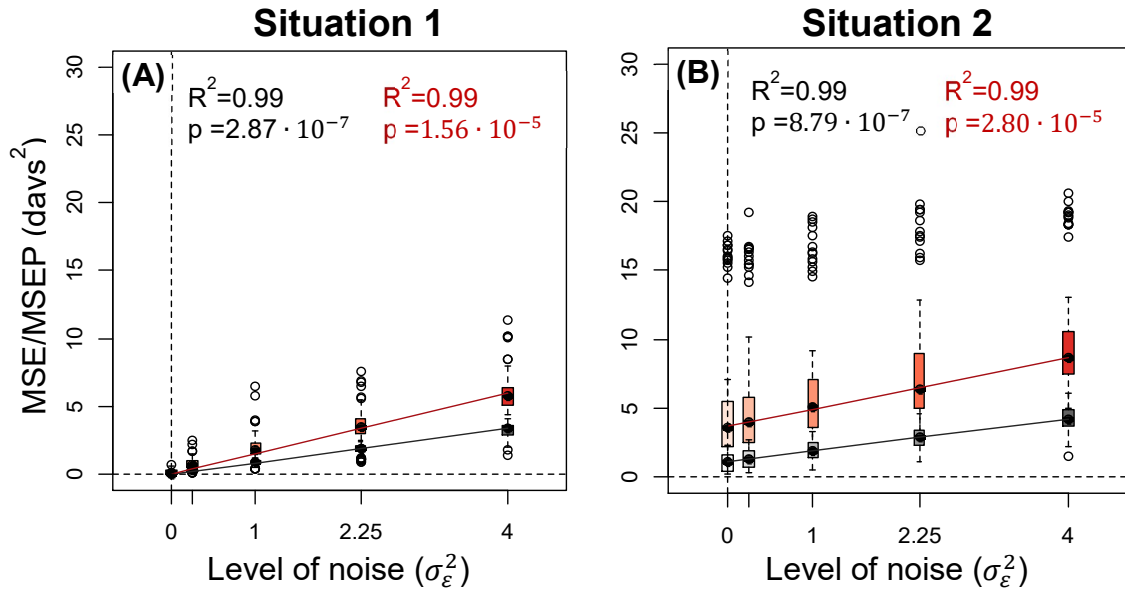
665

666 **Fig. 3: Actual (\bar{y}^{actual}) and simulated flowering dates by the “perfect model” ($y_{i-train}^{True}$**
 667 **and y_{i-test}^{True}).** The green dots represent the actual average flowering dates in 1980-2010 for
 668 winter wheat. Black crosses show the annual flowering time simulated by the Continuous
 669 Curvilinear (CC) models during baseline (1980-2010). Red circles show the annual flowering
 670 Julian days for 30 years in the decade 2050 under RCP8.5 and GCM GDFL-CM3.



672

673 **Fig. 4: Learning curves of the Continuous Curvilinear model at different levels of**
 674 **measurement error (σ_{ϵ}^2) and locations in Situation 1. The CC model is an accurate**
 675 **representation of the real system ($f^{TRUE} = f^{Sim} = CC$). Figures from the top-left to the**
 676 **bottom-right show the results for increasing levels of measurement error. Mean Square Errors**
 677 **for each location at calibration are represented by the empty grey-squared dots (MSE). Mean**
 678 **Square Errors for each location at 2050's RCP8.5 climate change Predictions are represented**
 679 **by the empty red-squared dots (MSEP). Filled dots show the median among locations. Lines**
 680 **summarize the behaviour for all locations according to the power-law (Eq. 22). The coefficients**
 681 **of determination of these lines are shown in black and red for the MSE and MSEP, respectively.**



682

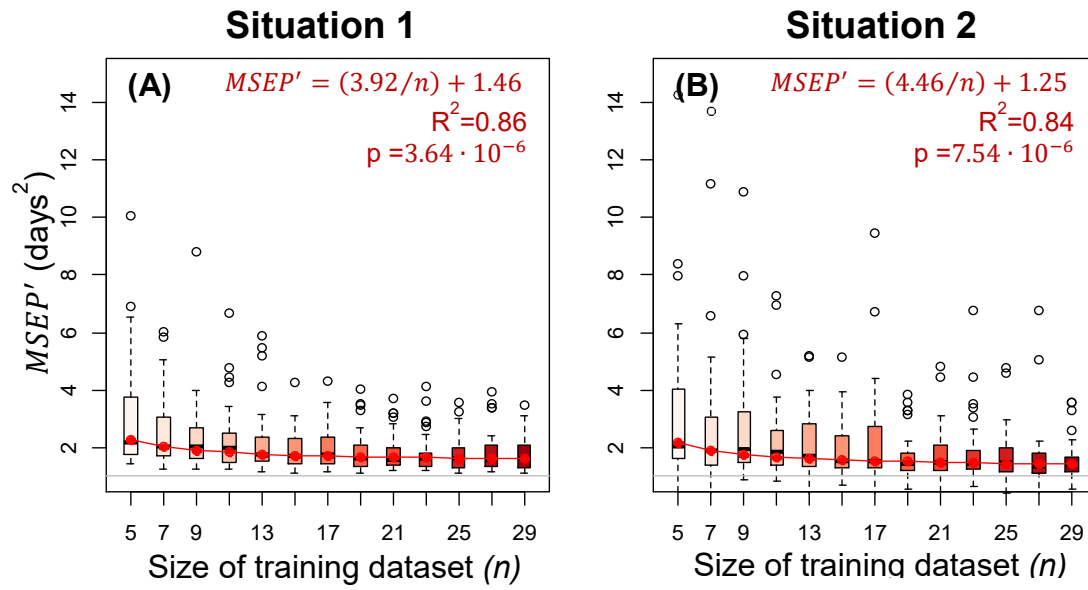
683 **Fig. 5: Squared error of simulation ($MSE/MSEP$) depending on measurement error (σ_ϵ^2).**

684 The boxes show the range of $MSEs$ (grey scale) and $MSEPs$ (red scale) obtained with different

685 sizes of datasets (n). The solid black and red lines represent the linear response of MSE and

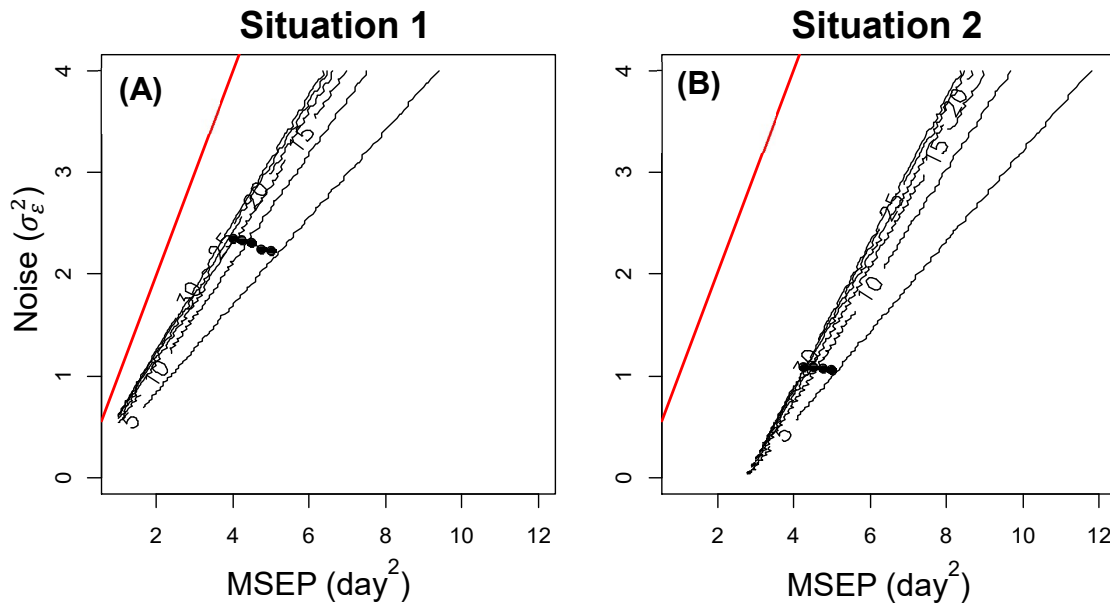
686 $MSEP$, respectively, to measurement error. Graph A and B show the results for the Situation

687 S1 ($f^{TRUE} = f^{Sim} = CC$) and Situation S2 ($CC = f^{TRUE} \neq f^{Sim} = BS$), respectively.



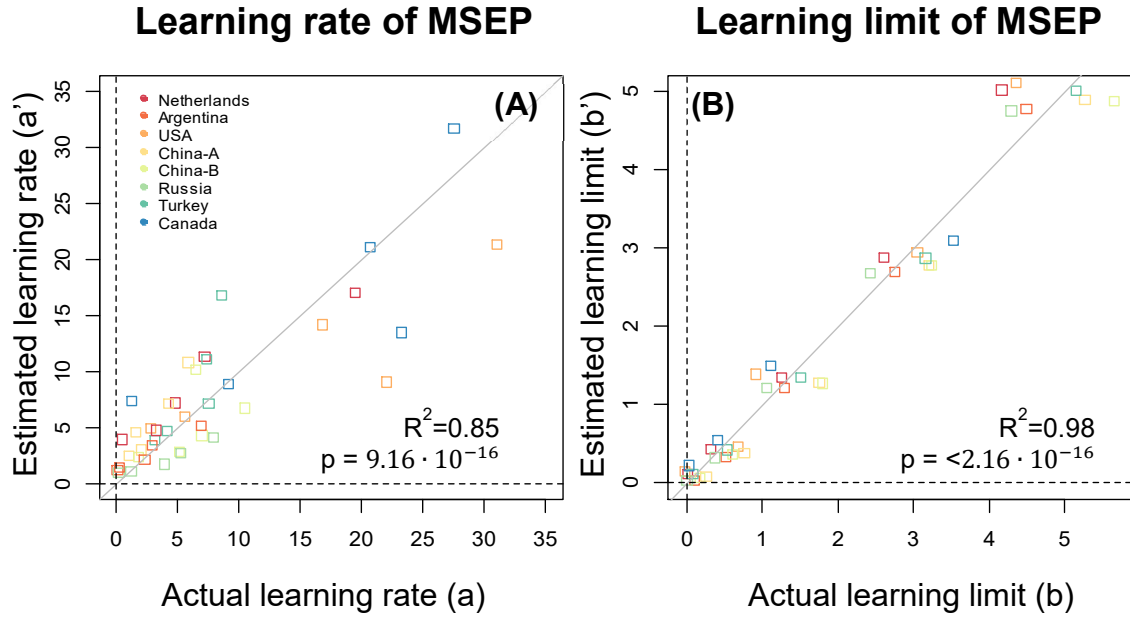
688

689 **Fig. 6: Transformed squared error of simulation ($MSEP'$) depending on the size of the**
 690 **training dataset (n).** The boxes show the range of $MSEP$ s obtained in both situations. The
 691 solid red line is the power-law curve representing the response of $MSEP$ to n . Graph A and B
 692 show the results for the Situation S1 ($f^{TRUE} = f^{Sim} = CC$) and Situation S2 ($CC = f^{TRUE} \neq$
 693 $f^{Sim} = BS$), respectively.



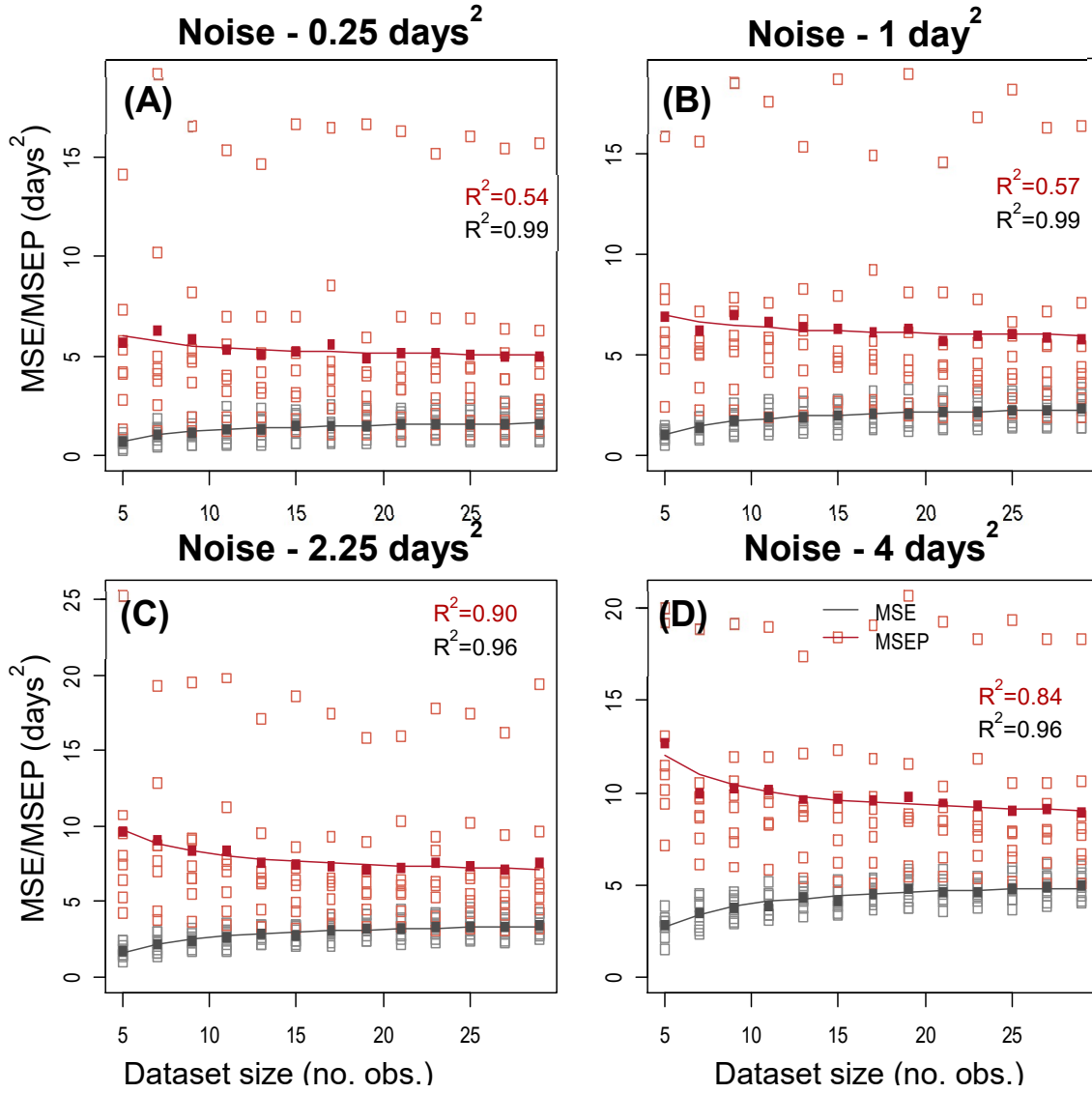
694

695 **Fig. 7: Response surface of the number of observations required (n) to reach a specific**
 696 **Mean Square Error of Prediction (MSEP, x-axis) with noise (σ_ε^2 , y-axis) in S1(A) and S2**
 697 **(B).** Contour lines show changes in n for every 5 observations, from $n = 5$ to $n = 30$. The red
 698 thick line is the minimum limit of $MSEP$ that can be achieved with a specific noise level
 699 ($\min(MSEP) = \sigma_\varepsilon^2$). The black dots represent the paths to improve the prediction skills of the
 700 models (decreasing $MSEP$) by using less precise (i.e., higher σ_ε^2) but larger datasets (i.e.,
 701 greater n).



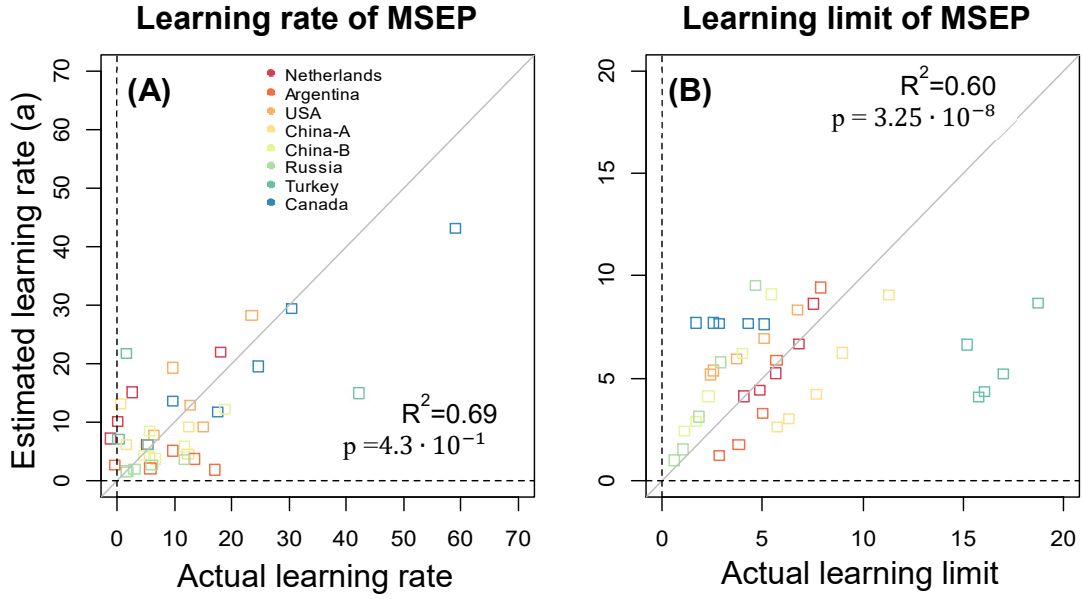
702

703 **Fig. 8: Exploring the location-specific learning curves and their dependence on the**
 704 **variance of the target population in Situation S1.** The graph on the left (A) and the right (B)
 705 show the learning rates (a) and the learning limits (b) for all location and noise levels. The x-
 706 axis represents the actual values derived from fitting Eq. 22 to the results in Fig. 4 for each
 707 location. The y-axis shows the estimated coefficients from the equations; $a' = 0.03\sigma_\varepsilon^2 +$
 708 $0.11\sigma_T^2 + 0.1(\sigma_\varepsilon^2 \cdot \sigma_T^2)$ and $b' = 1.16\sigma_\varepsilon^2 + 0.003\sigma_T^2 + 0.001(\sigma_\varepsilon^2 \cdot \sigma_T^2)$. Locations are represented by
 709 different colours.



710

711 **Fig. 9: Learning curves of the Broken-Stick model at different levels of measurement**
 712 **error (σ_{ϵ}^2) and locations in Situation S2. The model BS is an approximate representation**
 713 **of the real system ($f^{True} = CC$; $f^{Sim} = BS$).** Figures from the top-left to the bottom-right
 714 show the results for increasing levels of measurement error. Mean Square Errors for each
 715 location at calibration are represented by the empty grey-squared dots (*MSE*). Mean Square
 716 Errors for each location at 2050's RCP8.5 climate change predictions are represented by the
 717 empty red-squared dots (*MSEP*). Filled squares show the median among locations. Lines
 718 summarize the behaviour for all locations according to the power-law (Eq. 22). The coefficients
 719 of determination of these lines are shown in black and red for the MSE and MSEP, respectively.



720

721 **Fig. 10: Exploring the location-specific learning curves and their dependence on the**
 722 **variance of the target population in Situation S2.** The graph on the left (A) and right (B)
 723 show the learning rates (a) and the learning limits (b) for all location and noise levels. The x-
 724 axis represents the actual values derived from fitting Eq. 22 to the results in Fig. 9 for each
 725 location. The y-axis show the estimated coefficients from the equations: $a' = -0.53\sigma_\varepsilon^2 +$
 726 $0.18\sigma_T^2 - 0.13(\sigma_\varepsilon^2 \cdot \sigma_T^2)$ and $b' = 2.45\sigma_\varepsilon^2 + 0.12\sigma_T^2 - 0.03(\sigma_\varepsilon^2 \cdot \sigma_T^2)$. Locations are represented by
 727 different colours.

728 **Tables**

729 **Table 1: Details of the locations used in the analysis. Dates of sowing and anthesis are**
 730 **shown as Julian Days (JD). $\bar{y}_{BS/CC}^{actual}$ and $\sigma_{\bar{y}}$ represent the average anthesis dates between**
 731 **1980 and 2100 and their standard deviations simulated by the BS and CC perfect models.**
 732 **ΔT is the projected increase in local temperature from baseline (1980-2010) to projected**
 733 **climate change (2050's).**

Location	Country	Latitude (°)	Sowing (JD)	Anthesis (JD)	\bar{y}_{BS}^{actual} (JD)	$\sigma_{\bar{y}}$ (JD)	\bar{y}_{CC}^{actual} (JD)	$\sigma_{\bar{y}}$ (JD)	ΔT (°C)
Wageningen	Netherlands	51.97	309	176	176	4.25	176	6.09	2.83
Balcarce	Argentina	-37.75	217	329	328	2.21	329	3.17	1.66
Manhattan	USA	43.03	274	135	136	5.1	135	6.38	4.58
Nanjing	China (A)	32.03	278	125	125	3.76	125	4.70	3.24
Luancheng	China (B)	37.53	278	125	126	3.91	125	4.47	3.46
Krasnodar	Russia	45.02	258	140	140	2.36	140	2.80	-0.76
Izmir	Turkey	38.60	319	121	122	4.49	121	6.06	2.82
Lethbridge	Canada	49.70	253	161	161	6.33	161	8.15	4.44

734

735 **Table 2: List of all the datasets generated with the perfect model. The level of noise or**
 736 **measurement error is represented by σ_ε^2 . The maximum number of observations in the dataset**
 737 **is represented by n_{max} .**

Purpose	Period	Perfect model	Noise - σ_ε^2	n_{max}
Training	1980-2010	CC	0.00	30
Training	1980-2010	CC	0.25	30
Training	1980-2010	CC	1.00	30
Training	1980-2010	CC	2.25	30
Training	1980-2010	CC	4.00	30
Testing	2050's - RCP8.5	CC	0.00	30
Testing	2050's - RCP8.5	CC	0.25	30
Testing	2050's - RCP8.5	CC	1.00	30
Testing	2050's - RCP8.5	CC	2.25	30
Testing	2050's - RCP8.5	CC	4.00	30

738

739 **Table 3: Data required (n) for both the CC and the BS model under situations S1 and S2**
740 **to reach the point where additional data did not imply relevant improvements of the**
741 **prediction skills**

Level of noise (σ_ε^2)	Situation 1	Situation 2
0.25	6(± 1)	8(± 1)
1.00	11(± 1)	16(± 2)
2.25	17(± 2)	23(± 4)
4.00	23(± 3)	31(± 5)

742