Article:

Montesino-San Martin, M, Wallach, D, Olesen, JE et al. (5 more authors) (2018) Data requirements for crop modelling-Applying 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

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

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

[a] Department of Plant and Environmental Science, University of Copenhagen, Højbakkegård Allé 30, 2630 Taastrup, Denmark
[b] National Institute for Agricultural Research (INRA), UMR AGIR, Toulouse, France
[c] Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark
[d] Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK
[e] University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Grisebachstraße 6, 37077 Göttingen, Germany
[f] University of Göttingen, Centre for Biodiversity and Sustainable Land Use, Büsgenweg 1, 37077 Göttingen, Germany
[g] System Montpellier SupAgro, INRA, CIHEAM-IAMM, CIRAD, Univ Montpellier, 34060 Montpellier, France

Highlights

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

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

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

[a] Department of Plant and Environmental Science, University of Copenhagen, Højbakkegård Allé 30, 2630 Taastrup, Denmark
[b] National Institute for Agricultural Research (INRA), UMR AGIR, Toulouse, France
[c] Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark
[d] Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK
[e] University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Grisebachstraße 6, 37077 Göttingen, Germany
[f] University of Göttingen, Centre for Biodiversity and Sustainable Land Use, Büsgenweg 1, 37077 Göttingen, Germany
[g] System Montpellier SupAgro, INRA, CIHEAM-IAMM, CIRAD, Univ Montpellier, 34060 Montpellier, France

Abstract

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

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

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

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

[a] Department of Plant and Environmental Science, University of Copenhagen, Højbakkegård Allé 30, 2630 Taastrup, Denmark
[b] National Institute for Agricultural Research (INRA), UMR AGIR, Toulouse, France
[c] Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark
[d] Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK
[e] University of Göttingen, Centre for Biodiversity and Sustainable Land Use, Büsgenweg 1, 37077 Göttingen, Germany
[f] University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Grisebachstraße 6, 37077 Göttingen, Germany
[g] System Montpellier SupAgro, INRA, CIHEAM-IAMM, CIRAD, Univ Montpellier, 34060 Montpellier, France

1. Introduction

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

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

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

Drawing the learning curves requires calibrating and evaluating the model repeatedly, changing the size of the calibration dataset. This makes the process computationally demanding and data intensive. Phenology combines its relevance for yield (Craufurd and Wheeler, 2009) with its simple mathematical formulation and fast execution (e.g. Ceglar et al., 2011). Within the phenology phases, flowering is particularly critical; it is a very sensitive phase to temperature extremes (Ugarte et al., 2007) and it defines the balance between source-sink organs. Therefore, the simulation of flowering time represents a practical starting point to introduce the learning curve approach into crop modelling. Phenology modelling offers several working solutions with different mathematical formulations (Ceglar et al., 2011; Alderman and Stanfill, 2017). Learning curves are likely influenced by model structures, since prediction skills of different modelling hypotheses vary due to specific error compensations forged during calibration (Wallach et al., 2011). Hence, robust conclusions about data-model interactions with the learning curves require the assessment of multiple structures.

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

2. Methods

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

Two phenology models for simulating anthesis dates of winter wheat under climate change are considered; the Broken-Sticks (BS) and Continuous Curvilinear (CC) (Wang and Engel, 1998) models. The BS is a wide-spread practical model to simulate phenology whereas the CC model is considered a more realistic version from a biological perspective (Streck et al., 2008). Consequently, we assume that the CC model is the "*perfect model*" and the BS and the CC models are used as simulation models. Thus, two situations concerning model structures are assessed; (S1) the structure of the simulation model is an exact representation of reality (the simulation model and the "*perfect model*" are the same, both represented by the CC model), and (S2) the structure of the simulation model approximates the reality (the BS and the CC model correspond to the simulation model and the "*perfect model*" respectively). The results are used to analyse the shape of the learning curves and understand the relationships between measurements, errors and model structures.

2.1. Phenology models

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

(Fig. 1)

2.1.1. Vernalization response

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

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

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

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

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

2.1.2. Photoperiod response:

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

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

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

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

$$\omega = 1.49 - 2.96 \cdot 10^{-2} P_{base} - 1.14 \cdot 10^{-1} P_{opt} + 2.82 \cdot 10^{-3} {P_{base}}^{2} + 2.41 \cdot 10^{-3} {P_{opt}}^{2} \quad \text{(Eq. 5)}$$

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

2.1.3. Temperature response:

The response of the crop development (f_{t-BS}) to the daily air temperature (T_a) in the BS model is considered proportional when air temperatures are between the base (T_{base}) and optimum (T_{opt}) cardinal temperatures (Eq. 6). If the temperature is above the optimum, but below its critical temperature (T_{max}), the rate of development reacts inversely proportional to the difference between the air temperature and its optimum (Eq. 7). If the air temperature is below its base temperature or above its critical temperature, the daily rate of development is zero (Eq. 8).

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

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

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

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

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

$$\alpha = \frac{ln2}{ln\left[\frac{(T_{max} - T_{base})}{(T_{opt} - T_{base})}\right]} \quad \text{(Eq. 10)}$$

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

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

2.1.4. Development phase duration

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

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

In the CC model, a development stage is reached when the accumulation of daily rates (TTN) equals 1 (e.g., Eq. 13). This is achieved by using a scaling parameter (r_{max}) that represents the maximum daily development rate. The maximum development rate has an exponential form based on a parameter k (Eq. 14). Eq. 13 is an example of the computation between emergence and terminal spikelet.

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

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

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

2.1.5. Phenology model parameters

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

2.2. Perfect models and artificial flowering date records

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

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

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

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

(Table 1)

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

(Table 2)

2.3. Steps to generate the learning curves

The models were recalibrated (Fig. 2) using OLS (Eq. 17) and *n* randomly sampled seasons from the training dataset (Eq. 16). The resulting model ($f^{Sim}(\hat{\theta}, x_i)$) was used to simulate the *n* seasons of the calibration dataset (baseline) and the 30 seasons of the testing dataset (i.e. 2050's anthesis dates under RCP8.5). The assessment of the performance of $f^{Sim}(\hat{\theta}, x_i)$ was based on its Mean Square Error (MSE) (Eq. 18) and the Mean Square Error of Prediction (*MSEP*) (Eq. 20).

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

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

(Fig. 2)

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

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

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

Phenology models in climate impact assessments contradict both premises; (1) they are far from linear and (2) the baseline (training datasets) and future climate flowering dates (testing dataset) represent different populations. Instead of Eq. 21, the relationship will be expressed according to the power law (Eq. 22). In Eq. 22, *a* and *b* represent the learning rate and learning limit, respectively. The learning rate (*a*) represent the portions of the *MSEP* that is reducible with larger training datasets (n). Conversely, the learning limit (*b*) constitutes the unreducible part of *MSEP*. Eq. 22 is a more general form of Eq. 21 since a and b can adopt the values $a = p\sigma_\varepsilon^2$ and $b = \sigma_\varepsilon^2$.

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

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

3. Results

3.1. "Perfect model" calibration, training and testing datasets

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

(Fig. 3)

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

Several calibrations and evaluations of the CC model were carried out following the algorithm described above. The calibration dataset was changed with respect to the number of seasons (n) and levels of noise (σ_ε^2) and the model performance was tested in terms of mean squared errors (*MSE* and *MSEP*). The squared errors of the *CC* models can be seen in Figs. 4-5. In general, Fig. 4 shows an increase of *MSE* and a decrease of *MSEP* with greater sizes of the calibration dataset (n). The *MSE* and *MSEP* tend to the variance of noise, i.e. 0.25, 1, 2.25, and 4 days^2, without reaching it for the range of n explored. It should be noted that the graphs differ in the range of squared errors displayed on the y-axis for visualization purposes. Results show that prediction performance (*MSEP*) worsens proportionally with the level of measurement error in both calibration and evaluation (R^2=0.99) (Fig. 5a).

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

(Fig. 4)

(Fig. 5)

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

(Fig. 6)

A larger n and higher σ_ε^2 had positive and negative impacts, respectively, on the prediction performance (Fig. 4-5a). To investigate the compensations between n and σ_ε^2 we rearranged Eq. 21-22 to calculate the n required to reach a specific *MSEP* ($n = \hat{a}/(MSEP/\sigma_\varepsilon^2) - \hat{b}$). Combined sequences of *MSEP* and σ_ε^2 were fed into the equation to build the response surfaces seen in Fig. 7a. The graph shows the n (z-axis) depending on the *MSEP* (x-axis) and the σ_ε^2 (y-axis). The non-equidistant contour lines in Fig. 8a depict the non-linearities between *MSEP* and n captured in Eq. 21 and 22. The straightness of the contour lines reflects the linear relationship between *MSEP* and σ_ε^2 represented in Eq. 21. We inspected whether larger but less precise datasets could lead lower *MSEPs* than smaller but more precise datasets. The dashed black line in Fig. 7a shows one case where the *MSEP* is reduced from 5 day^2 to 4 day^2 (in steps of 0.25 day^2) by using training datasets with size n equal to 4, 6, 9, 13 and 30 and noise levels equal to 2.22, 2.25, 2.37, 2.41 and 2.51 days^2, respectively. Eqs. 22-23 and Fig. 7a confirm that it is possible in theory to compensate the lack of precision in the measurements with more seasons observed. However, the equations and the results in Fig. 7a highlight two major limitations for this type of compensations; (1) the noise imposes a minimum limit of the *MSEP* ($\lim_{n\to\infty} MSEP = \sigma_\varepsilon^2$) and (2) n changes very quickly with *MSEP* and σ_ε^2 (n = $\mathrm{a}/(MSEP - b)$), becoming rapidly very large and practically unfeasible.

(Fig. 7)

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

(Table 3)

Every square dot in Fig. 4 represents the squared error (*MSE/MSEP*) of a particular location. The dispersion of the *MSEP* values reveals that the variation between locations is large. To explore the reasons behind these differences, Eq. 22 was adjusted independently for the results of each location. We inspected whether the variance of the training population (flowering dates 1980-2010) might be behind the differences in the location-specific learning rates (a) and limits (b) of the *MSEPs*. Fig. 8 displays the a and b obtained from the *MSEPs* for each location and noise level on the x-axis. On the y-axis, the graph shows the *a'* and *b'* obtained from a regression based on noise (σ_ε^2) and the variance of the training dataset (σ_T^2). We found that the variance of the training dataset and the variance of noise in the measurements explained most of the variability in the learning rates (Fig. 8a). The regression of *a'* based on σ_ε^2 and σ_T^2 shows a good fit between the actual and the estimated learning rates (R^2=0.85). The variance of training dataset and its product with the variance of noise ($\sigma_T^2 \cdot \sigma_\varepsilon^2$) were highly significant ($p < 0.01$) to explain the variations in learning rates. The variability in b' (Fig. 8b) was only significantly explained ($p < 0.01$) by the noise (R^2=0.98).

(Fig. 8)

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

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

(Fig. 9)

The impact of structural error on *MSEP* was removed by subtracting the location-specific minimum prediction error obtained with zero noise training datasets (k_{loc}). As in S1, the differences among *MSEPs* caused by noise were eliminated by dividing *MSEP* by σ_ε^2 ($MSEP' = (MSEP - k_{loc})/\sigma_\varepsilon^2$)). We adjusted Eq. 22 to $MSEP'$ by calibrating a and b (Fig. 6b). The model was significant ($p - value = 7.54 \cdot 10^{-6}$) and explained a high portion of the variability in $MSEP'$ (R^2 = 0.84). The estimated values of the coefficients $\hat{a}$ and $\hat{b}$ were 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 to S1 and its theoretical value. Therefore, the model structure hampered the parameter estimation, since $\hat{a}$ is the portion of *MSEP* attributed to parameter estimation error.

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

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

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

(Fig. 10)

4. Discussion

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

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

We estimated that 5-7 observations of flowering dates were enough to conduct impact assessments under 2050's climate change conditions. These results correspond to 0.25 day^2 measurement error and perfect model structures. However, model structures are known to be imperfect representations of the agricultural systems (Rötter et al., 2011). Therefore, S2 is more realistic representation of the situation in crop modelling. In our experiment, structural approximations (S2) translated into an increase of prediction error. The error increase was specific to each model and location. Structural errors also interfered with parameter estimation, increasing the data requirements. Therefore, moving from S1 to S2 caused an increase of data requirements to 7-9 with 0.25 day^2 of measurement error. The number of field measurements (years) usually available to compare observations and simulation ranges from 5 to 10 before the cultivar becomes obsolete. This number of measurements is around the recommended minimum number estimated in our analysis. However, noise in field observations is likely larger than 0.25 days^2. To get more measurements in the same time period, multi-environmental trials or experiments with multiple sowing dates have to be conducted, which goes in line with recommendations by He et al. (2017). Strictly, neither structures can be considered correct, nor are parameter values true. For these reasons, the results obtained with this kind of assessment are merely theoretical and advisory. These recommendations can vary among locations: the data required depends on the learning rate and results show that it varies with the inter-annual flowering variability of the training population (Fig. 8 and 10). Therefore, the suitable size of the dataset could be larger in places where there is greater variability among seasons.

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

5. Conclusions

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

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

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

Acknowledgements

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

References:

Alderman, P. D., & Stanfill, B. (2017). Quantifying model-structure-and parameter-driven uncertainties in spring wheat phenology prediction with Bayesian analysis. European Journal of Agronomy, 88, 1-9.

Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S., & Ewert, F. (2013). Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe. Agricultural and Forest Meteorology, 170, 32-46.

Asseng, S., et al. (2015). Rising temperatures reduce global wheat production. Nature Climate Change, 5(2), 143-147.

Bogard, M., et al. (2014). Predictions of heading date in bread wheat (Triticum aestivum L.) using QTL-based parameters of an ecophysiological model. Journal of experimental botany, eru328.

Boote, K.J. 1999. Data requirements for model evaluation and techniques for sampling crop growth and development. In: G. Hoogenboom, P.W. Wilkens, and G.Y. Tsuji, editors, DSSAT Version 3. A decision support system for agrotechnology transfer. Vol. 4. University of Hawaii, Honolulu. p. 201–216.

Boote, K. J., et al., (2016). Sentinel site data for crop model improvement—definition and characterization. Improving Modeling Tools to Assess Climate Change Effects on Crop Response, (advagricsystmodel7), 125-158.

Ceglar, A., Črepinšek, Z., Kajfež-Bogataj, L., & Pogačar, T. (2011). The simulation of phenological development in dynamic crop model: the Bayesian comparison of different methods. Agricultural and Forest Meteorology, 151(1), 101-115.

Confalonieri, R., Bregaglio, S., & Acutis, M. (2016). Quantifying uncertainty in crop model predictions due to the uncertainty in the observations used for calibration. Ecological Modelling, 328, 72-77.

Craufurd, P. Q., & Wheeler, T. R. (2009). Climate change and the flowering time of annual crops. Journal of Experimental Botany, 60(9), 2529-2539.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wang, E., & Engel, T. (1998). Simulation of phenological development of wheat crops. Agricultural systems, 58(1), 1-24.

Weir, A. H., Bragg, P. L., Porter, J. R., & Rayner, J. H. (1984). A winter wheat crop simulation model without water or nutrient limitations. *The Journal of Agricultural Science*, *102*(2), 371-382.

White, J. W., et al., (2013). Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. Computers and electronics in agriculture, 96, 1-12.

Figures

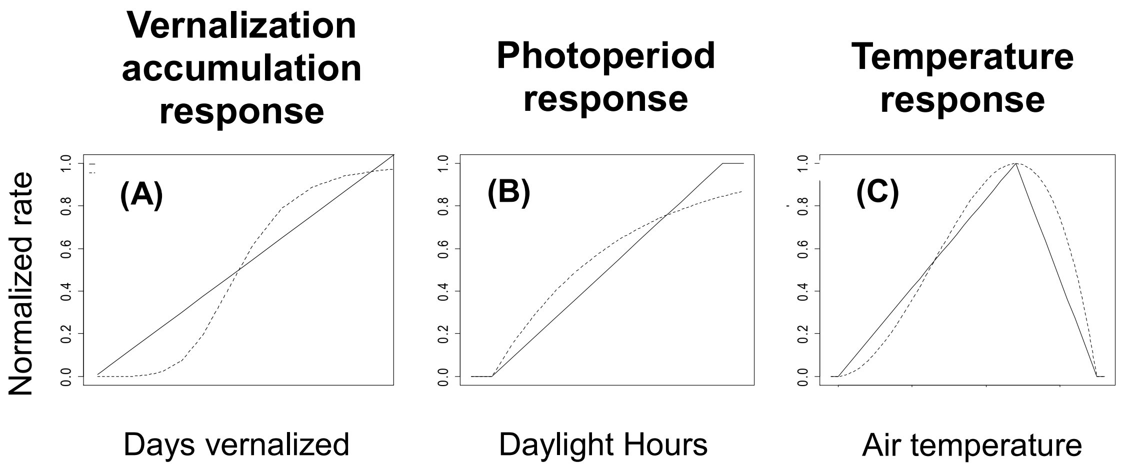

Fig. 1: Normalized responses of crop development to vernalization (A), photoperiod (B) and temperatures (C) simulated by the Broken-Sticks Model (solid line) and the Continuous Curvilinear Model (dashed line)

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

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

d. Calibrate the model by OLS using $calibration\ dataset$

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

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

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

f. Build the testing dataset

$$\text{testing} = \{\text{y}_{1-\text{tes}}^{\text{Measure}}, \dots, \text{y}_{30-\text{test}}^{\text{Measure}}\} = \{\text{y}_{1-\text{test}}^{\text{True}} + \varepsilon_1, \dots, \text{y}_{\text{n}-\text{test}}^{\text{True}} + \varepsilon_\text{n}\} =$$
$$= \{\text{f}^{\text{True}}(\theta^{\text{True}}, \text{x}_{1-\text{test}}) + \varepsilon_1, \dots, \text{f}^{\text{True}}(\theta^{\text{True}}, \text{x}_{30-\text{tes}}\ ) + \varepsilon_{30}\} \qquad \text{(Eq.19)}$$

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

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

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.

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

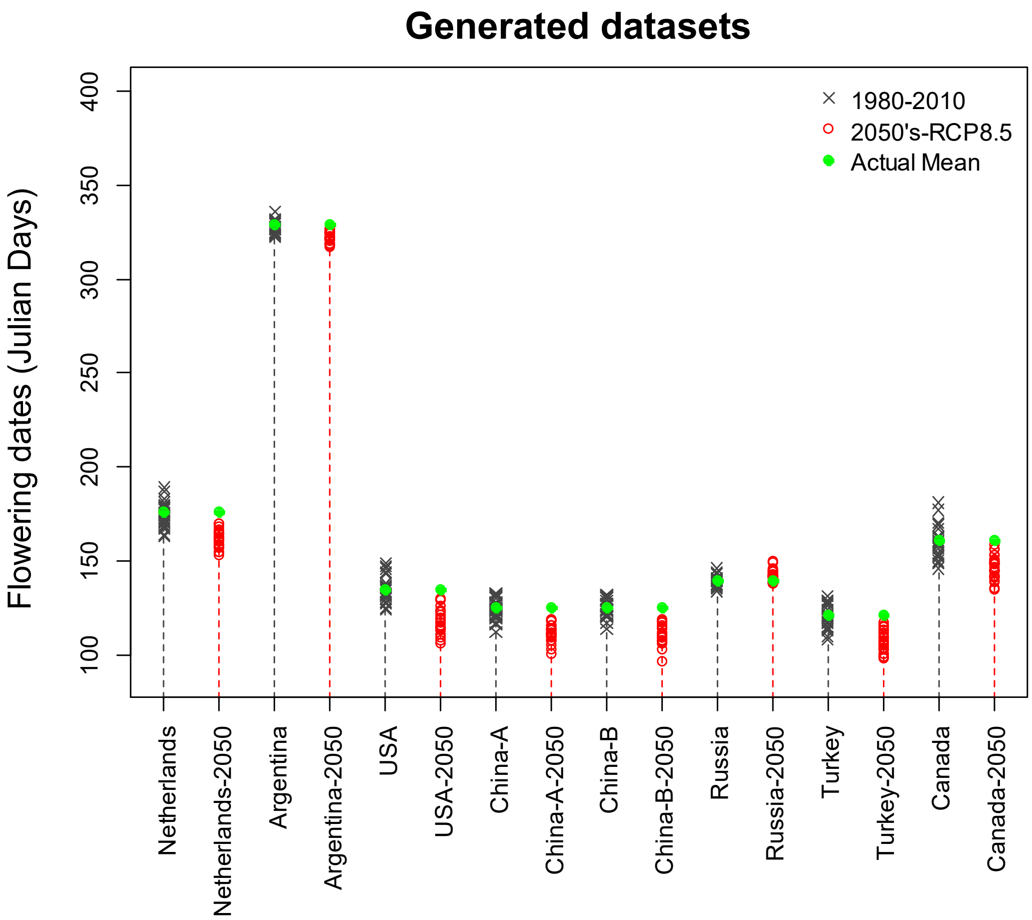

Fig. 3: Actual ($\bar{y}^{actual}$) and simulated flowering dates by the "perfect model" ($y^{True}_{i-train}$ and y^{True}_{i-test}). The green dots represent the actual average flowering dates in 1980-2010 for winter wheat. Black crosses show the annual flowering time simulated by the Continuous Curvilinear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

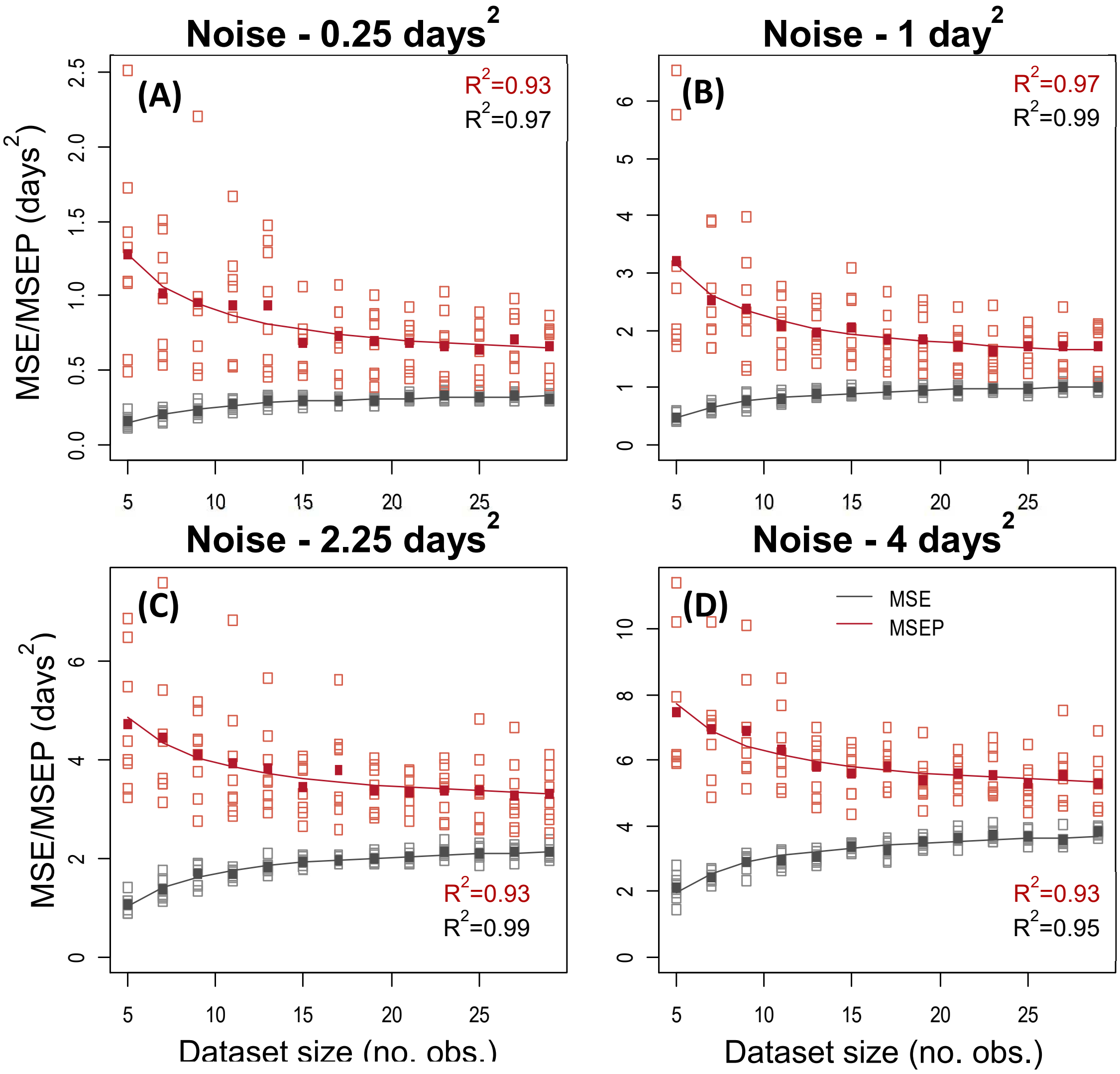

Fig. 4: Learning curves of the Continuous Curvilinear model at different levels of measurement error (σ_ε^2) and locations in Situation 1. The CC model is an accurate representation of the real system ($f^{TRUE} = f^{Sim} = CC$). Figures from the top-left to the bottom-right show the results for increasing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

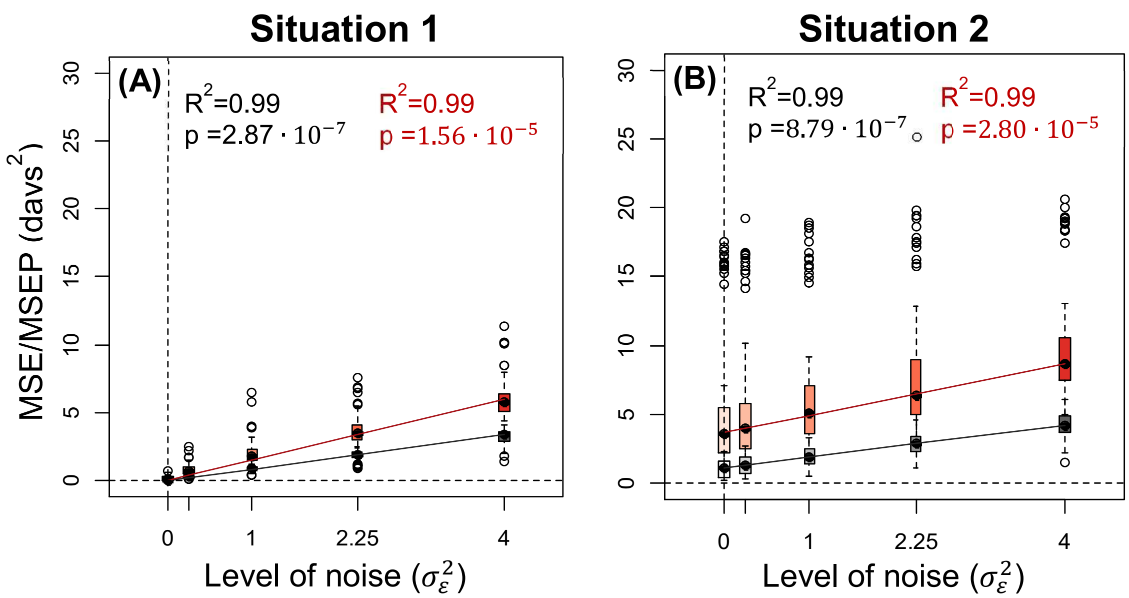

Fig. 5: Squared error of simulation (*MSE/MSEP*) depending on measurement error (σ_ε^2). The boxes show the range of *MSEs* (grey scale) and *MSEPs* (red scale) obtained with different sizes of datasets *(n)*. The solid black and red lines represent the linear response of *MSE* and *MSEP*, respectively, to measurement error. Graph A and B show the results for the Situation S1 ($f^{TRUE} = f^{Sim} = CC$) and Situation S2 ($CC = f^{TRUE} \neq f^{Sim} = BS$), respectively.

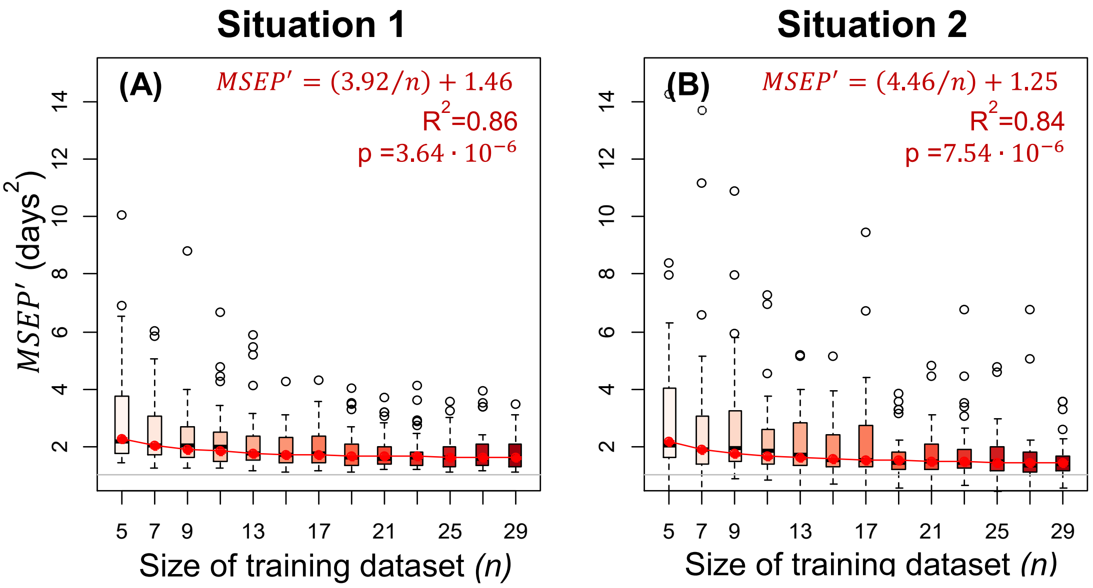

Fig. 6: Transformed squared error of simulation (*MSEP*') depending on the size of the training dataset (n). The boxes show the range of *MSEPs* obtained in both situations. The solid red line is the power-law curve representing the response of *MSEP* to *n*. Graph A and B show the results for the Situation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

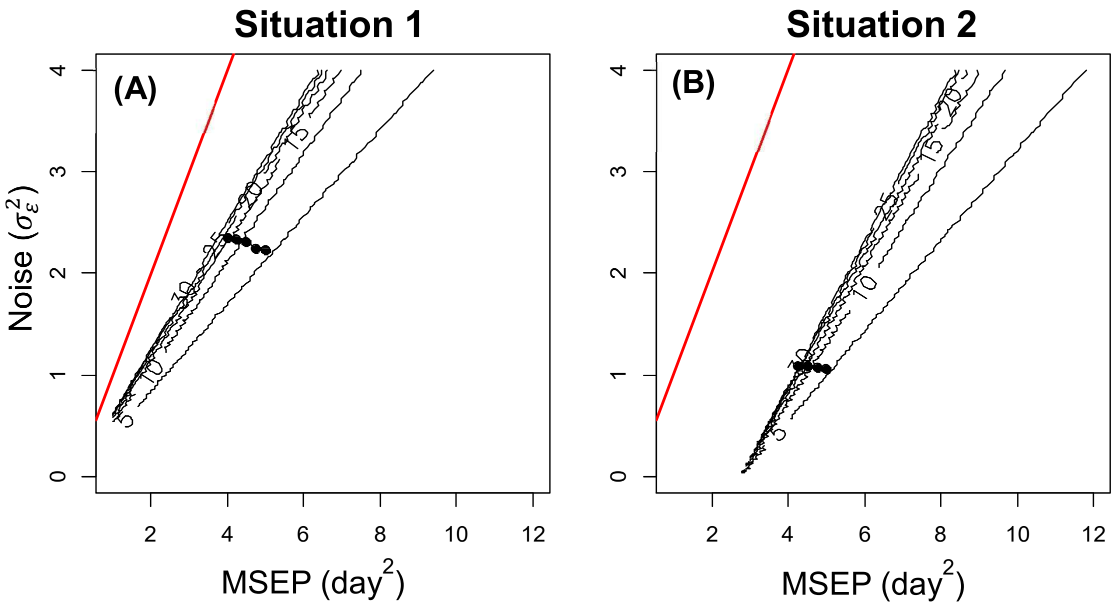

Fig. 7: Response surface of the number of observations required (*n*) to reach a specific Mean Square Error of Prediction (MSEP, x-axis) with noise (σ_ε^2, y-axis) in S1(A) and S2 (B). Contour lines show changes in *n* for every 5 observations, from *n* = 5 to *n* = 30. The red thick line is the minimum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

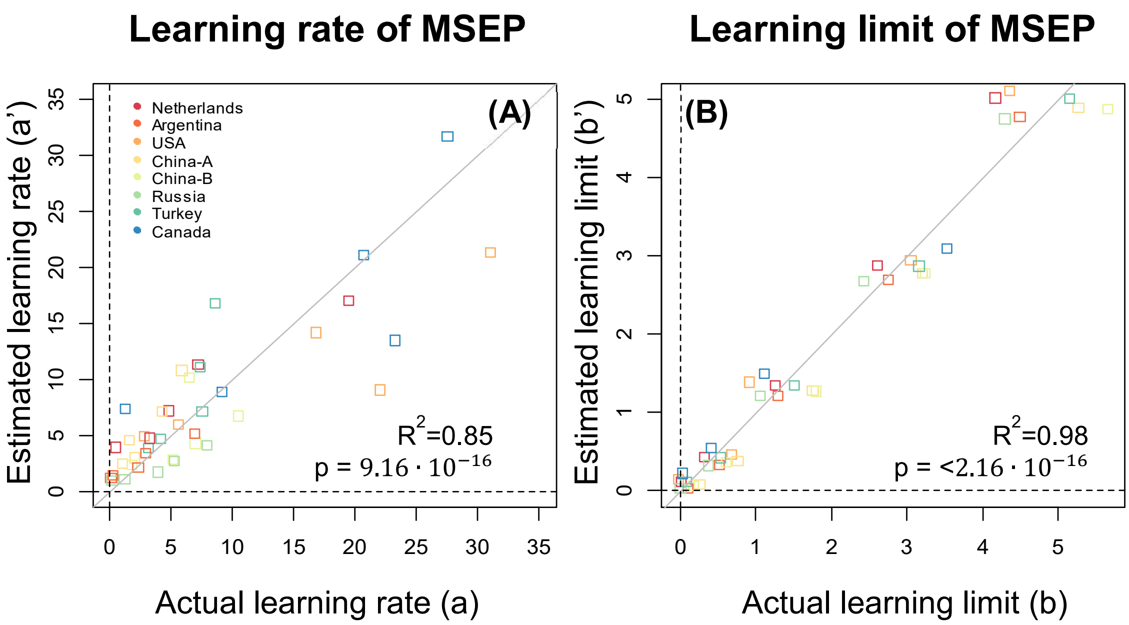

Fig. 8: Exploring the location-specific learning curves and their dependence on the variance of the target population in Situation S1. The graph on the left (A) and the right (B) show the learning rates (a) and the learning limits (b) for all location and noise levels. The x-axis represents the actual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

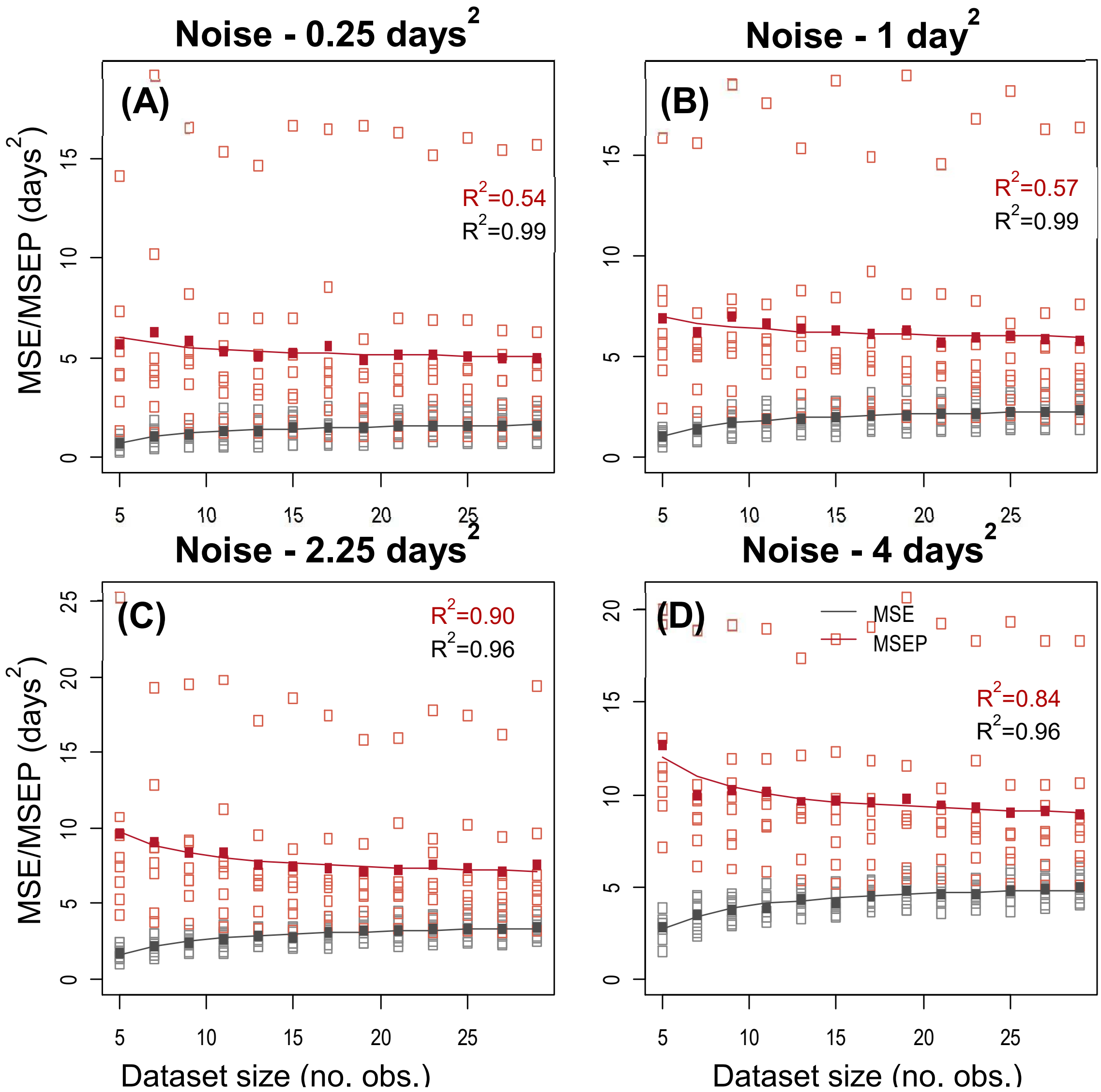

Fig. 9: Learning curves of the Broken-Stick model at different levels of measurement error (σ_ε^2) and locations in Situation S2. The model BS is an approximate representation of the real system ($f^{True} = CC$; $f^{Sim} = BS$). Figures from the top-left to the bottom-right show the results for increasing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

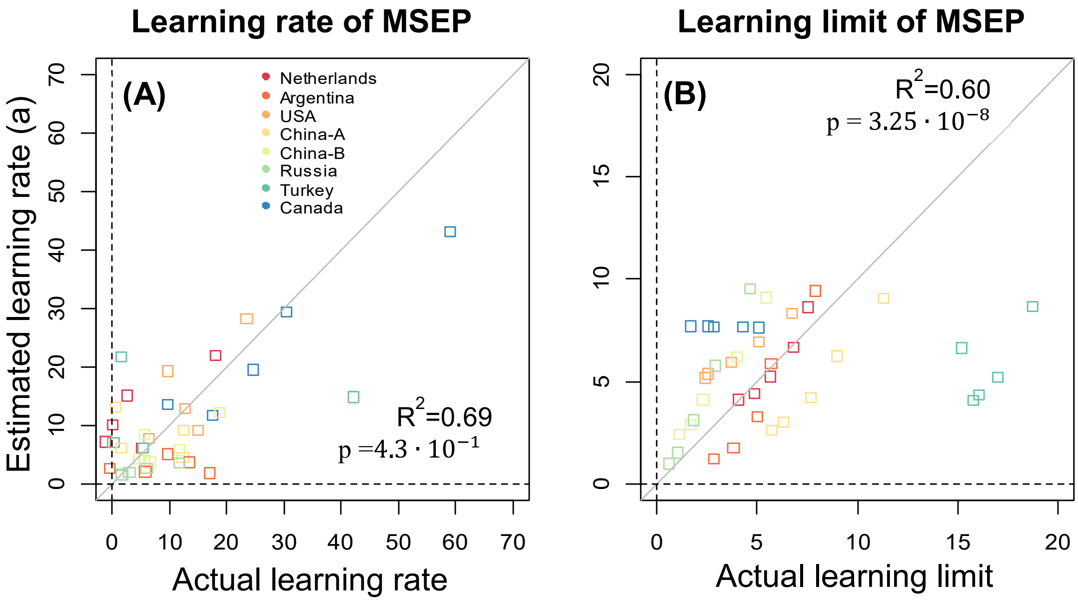

Fig. 10: Exploring the location-specific learning curves and their dependence on the variance of the target population in Situation S2. The graph on the left (A) and right (B) show the learning rates (a) and the learning limits (b) for all location and noise levels. The x-axis represents the actual values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

Tables

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

Location	Country	Latitude (º)	Sowing (JD)	Anthesis (JD)	$\bar{y}_{BS}^{actual}$ (JD)	$\sigma_{\bar{y}}$ (JD)	$\bar{y}_{CC}^{actual}$ (JD)	$\sigma_{\bar{y}}$ (JD)	ΔT (ºC)
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

Table 2: List of all the datasets generated with the perfect model. The level of noise or measurement error is represented by σ_ε^2. The maximum number of observations in the dataset 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

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