Referee comments 1

The authors present a very interesting approach to Bayesian model calibration that has been under-exploited within the crop modeling community. I very much enjoyed reading it. The topic and its treatment in this manuscript are compelling and likely of interest to the Biogeosciences readership and crop modeling community more generally. The results and discussion presented are interesting, but the sampling approach and checks for convergence were not documented well enough for me to evaluate whether the results presented were valid. Further documentation is needed before the manuscript can be reconsidered for publication.

We would like to thank you for your feedback and comments that have made us critically review our work and helped improve the manuscript. We are happy to hear that you enjoyed reading it. We address your comments in detail below. We have added more details regarding the sampling approach in the manuscript as well as in Supplementary Materials.

Our responses are in bold. Additions to the manuscript are marked as *MS, line…:* in *italics*. Additions to Supplementary Materials are marked as *Supp:* in *italics.*

Multiple details of the sampling approach used in this study remain unclear. The authors provide equations 3 & 4 as a formal expression of Bayesian sequential updating (BSU) in which the prior is defined based on a priori beliefs and the likelihood is derived from first site-year of data. Equation 4 indicates that the prior for the second site-year would then be the posterior distribution sampled using equation 3. The prior for the third site-year would be the posterior of the second site year, and so on. However, if I understand correctly, BSU is not the approach used in this study. Instead, the prior remains fixed across all site-year combinations and only the quantity of data used for the likelihood calculation increases with each subsequent site-year. This approach is fine and perhaps is better than the BSU approach in that over time the evidentiary basis for posterior inference is broader and increasingly more likely to encompass the full range of environments over which prediction can be accurately performed. However, can this second approach be accurately termed BSU? I would suggest using an alternate term for this approach (at least something like "approximate BSU") and adjusting the title accordingly.

It is correct that in the actual implementation of BSU in this study, we keep the prior fixed across all the site-year combinations and only update the data in the likelihood estimation. Although this methodology is not strictly BSU, the results are a much better representation of the true approach than using posterior as the next prior. This is due to problems arising from approximating the posterior density in the strictly BSU approach. In fact, the results from using the strictly BSU approach would lead to larger approximations of the real results than the current approach. We therefore refrain from using the term 'approximate' in this case and are of the opinion that a change in terminology is unwarranted. We have, however, added the following sentence in the manuscript.

MS, line 197: We refer to the current methodology as BSU, although it is not strictly so, for reasons of simplicity and the formal similarity of our approach.

Still, that is a relatively minor point. The greater issue is the number of questions remaining on the how this general approach was implemented. For example:

We have added details of the approach in the manuscript and provided supporting plots in the Supplementary Materials that have also been included below. We have also addressed each question individually in the sections that follow.

We have added the following paragraph to the manuscript:

MS, line 231: The posterior parameter distribution was sampled using the Markov Chain Monte Carlo method – Metropolis algorithm (Metropolis et al., 1953) (for details refer to Appendix B: Posterior sampling using MCMC Metropolis algorithm). Three chains were run in parallel. A normal distribution was chosen as the transition kernel. The jump size was adapted so that the acceptance rate would be between 25% and 35% (A. Gelman et al., 1996; Tautenhahn et al., 2012). For each sequential update calibration case, when a new site-year was added to the calibration sequence, the three chains were re-initialized and the transition kernel was re-tuned. A preliminary calibration test case, in which the model was calibrated to site-year 6_2010, was used to generate the starting points of the chains for each of the calibration cases. The starting points were randomly sampled from the posterior parameter range of the calibrated test case. This was done to reduce the time to

convergence. For the test case calibration, the starting points of the chains were randomly sampled from the prior range. The number of iterations for adapting the transition kernel varied between the different calibration cases. This number was low for some of the calibration cases because we set the initial pre-adaptation value for the standard deviation of the transition kernel, so that the acceptance rate would be between 25% and 35%. This initial value was based on knowledge gained from preliminary calibration test simulations. Convergence of the chains after jump adaptation was checked using the Gelman-Rubin convergence diagnostic (Brooks & Gelman, 1998; Andrew Gelman & Rubin, 1992). The total number of samples of the posterior distribution in each calibration case was dependent on when the Gelman-Rubin diagnostic was <=1.1, while ensuring a minimum of 500 accepted samples per chain, that is, a minimum of 1500 samples across the three chains. In effect, the total number of samples per calibration case was greater than 1500. The burn-in was variable and depended on the jump-adaptation. Only the iterations from the jump-adaptation step were discarded as burn-in. Parameter mixing was evaluated using trace-plots.

We have included the following details and table in Supplementary Materials S7.

Supp: *The sequential update calibration cases for the true sequences in the Swabian Alb and in Kraichgau are listed in Table S7-1. The number of iterations required to adapt the jump-size (A) were variable (20-580) and dependent on the calibration case. In some cases this number was low because we set the initial pre-adaptation value for the standard deviation of the transition kernel so that the acceptance rate would be between 25% and 35%. This initial value was based on knowledge gained from preliminary calibration test runs. The jump adjustment factor (f) in Table S7-1 influences the standard deviation of the transition kernel (V) such that $V = \frac{sd}{f}$ where sd is the standard deviation of the prior parameter distributions taken from Table 2 in the main text. With N being the total number of iterations per chain, the total number of iterations across the three chains after burn-in is given by $T = (N - A) \times 3$.*

On adding a new site-year, the chains were re-initialized and the transition kernel was re-tuned. New data was added to the dataset and the chains were allowed to adapt. The burn-in was variable and dependent of the jump-size adaptation. We ensured that a minimum of 500 accepted samples were generated per chain, that is, a minimum of 1500 total samples across chains were drawn. However, the actual number of samples drawn (T) was higher and dependent of when the Gelman-Rubin convergence diagnostic was <=1.1.

To assess parameter mixing, trace-plots were analysed (examples provided in Figure S27 and Figure S28). Additionally, auto-correlation plots (Figure S29, Figure S30) are provided (coda package in R (Plummer et al., 2006)) and effective sample size (ESS in Table S7-1) were calculated (mcmcse package in R (Flegal et al., 2021), (Vats et al., 2019)). Parameter DELTOPT2 generally showed good mixing and low auto-correlation. The effective sample size between 145 and 332, together with the Gelman-Rubin convergence diagnostic (<=1.1), provide sufficiently reliable posterior statistics for this study.

Table S7-1: MCMC sampling details for True sequence calibration cases in Kraichgau and the Swabian Alb

Sequence	*Calibration case*	*Number of accepted runs per chain during jump adaptation (A)*	*Jump adjustment factor (f)*	*Total accepted samples per chain (N)*	*Total samples after burn-in in all chains (T)* $= (N - A) \times 3$	*ESS*
True sequence Swabian Alb	*6_2010*	*20*	*3*	*1480*	*4380*	*236*

	6_2010, 5_2011	*580*	*3.97*	*1100*	*1560*	*332*
	6_2010, 5_2011, 5_2012	*20*	*5*	*800*	*2340*	*145*
	6_2010, 5_2011, 5_2012, 6_2013	*40*	*4.95*	*820*	*2340*	*167*
	6_2010, 5_2011, 5_2012, 6_2013, 5_2015	*20*	*5*	*620*	*1800*	*196*
	6_2010, 5_2011, 5_2012, 6_2013, 5_2015, 5_2016	*240*	*6.7*	*1400*	*3480*	*159*
True Sequence Kraichgau	*3_2011*	*60*	*5.005*	*3280*	*9660*	*153*
	3_2011, 2_2012	*20*	*5*	*4480*	*13380*	*163*
	3_2011, 2_2012, 1_2014	*20*	*7.7*	*5100*	*15240*	*168*

- How were chains initialized? Randomly sampling the prior? (The effectiveness of the Gelman-Rubin diagnostic generally depends on the starting points for multiple chains be overdispersed with respect to the posterior.)

MS, line 236: ***A preliminary calibration test case, in which the model was calibrated to site-year 6_2010, was used to generate the starting points of the chains for each of the calibration cases. The starting points were randomly sampled from the posterior parameter range of the calibrated test case. This was done to reduce the time to convergence. For the test case calibration, the starting points of the chains were randomly sampled from the prior range.***

- How many iterations were used for adapting the jump-size/transition kernel?

MS, line 239: ***The number of iterations for adapting the transition kernel varied between the different calibration cases. This number was low for some of the calibration cases because we set the initial pre-adaptation value for the standard deviation of the transition kernel, so that the acceptance rate would be between 25% and 35%. This initial value was based on knowledge gained from preliminary calibration test simulations.***

- When adding a new site-year, how were the chains handled? Were they re-initialized (along with retuning the transition kernel)? Was new data simply added to the dataset and chains allowed to adapt?

MS, line 234: ***For each sequential update calibration case, when a new site-year was added to the calibration sequence, the three chains were re-initialized and the transition kernel was re-tuned.***

- How long was the warmup/burn-in? Was this variable?

MS, line 247: ***The burn-in was variable and depended on the jump-adaptation. Only the iterations from the jump-adaptation step were discarded as burn-in.***

The iterations from the jump adaptation step varied by calibration case and are provided in Table S7-1 of the Supplementary Materials (see above) in column 'Number of accepted runs per chain during jump adaptation (A)'.

- How many samples were generated after warmup? I see a number of 500 in Appendix B. That seems very low.

The number of accepted samples (T) after warm-up varied by calibration case and are provided in Table S7-1 of the Supplementary Materials (see above) in column 'Total samples after burn-in in all chains (T)= $(N - A) \times 3$'.

MS, line 244: ***The total number of samples of the posterior distribution in each calibration case was dependent on when the Gelman-Rubin diagnostic was <=1.1, while ensuring a minimum of 500 accepted samples per chain, that is, a minimum of 1500 samples across the three chains. In effect, the total number of samples per calibration case was greater than 1500.***

- How was parameter mixing evaluated?

MS, line 248: ***Parameter mixing was evaluated using trace-plots.***

Please find an excerpt from the Supplementary Materials S7 below with example trace-plots:

Supp: ***To assess parameter mixing, trace-plots were analysed (examples provided in Figure S27 and Figure S28). Additionally, auto-correlation plots (Figure S29, Figure S30) are provided (coda package in R (Plummer et al., 2006)) and effective sample size (ESS in Table S7-1) were calculated (mcmcse package in R (Flegal et al., 2021), (Vats et al., 2019)).***

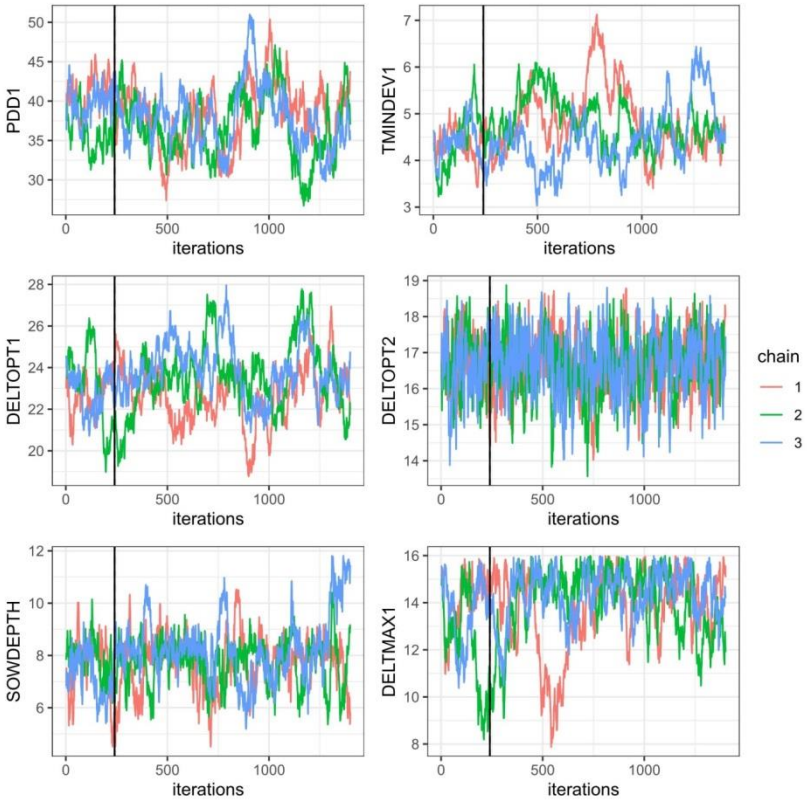


Figure S27: Trace-plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in the Swabian Alb at 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016. The x-axis is the number of iterations and y-axis is the parameter. The colours indicate the three chains. The black solid vertical line indicates the burn-in phase during which the transition kernel was adapted.

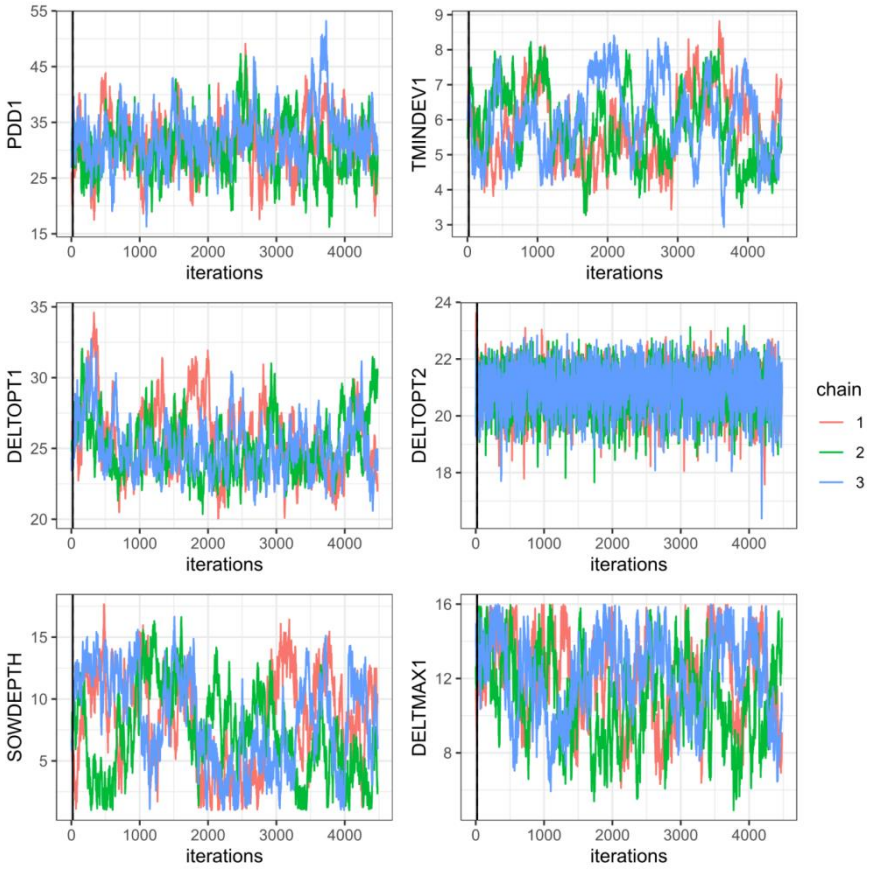


Figure S28: Trace-plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in Kraichgau at 3_2011 and 2_2012. The x-axis is the number of iterations and y-axis is the parameter. The colours indicate the three chains. The black solid vertical line indicates the burn-in phase during which the transition kernel was adapted.

How did the traceplots look? (Consider including representative traceplots in manuscript or supplementary methods)

Please see the response above.

- How did you check for auto-correlation in samples?

Auto-correlations plots have now been included in Supplementary Materials S7. Please find an excerpt from the Supplementary Materials S7 below:

Supp: ***Additionally, auto-correlation plots (Figure S29, Figure S30) are provided (coda package in R (Plummer et al., 2006)) and effective sample size (ESS in Table S7-1) were calculated (mcmcse package in R (Flegal et al., 2021), (Vats et al., 2019)). Parameter DELTOPT2 generally showed good mixing and low auto-correlation.***

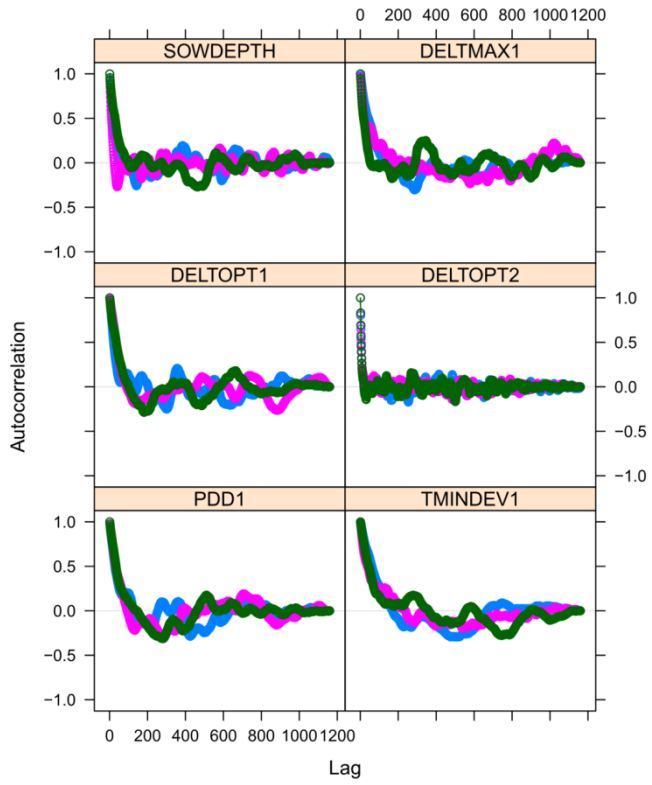


Figure S29: Auto-correlation plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in the Swabian Alb at 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016. The x-axis is the lag distance and y-axis is the auto-correlation. The colours indicate the three chains.

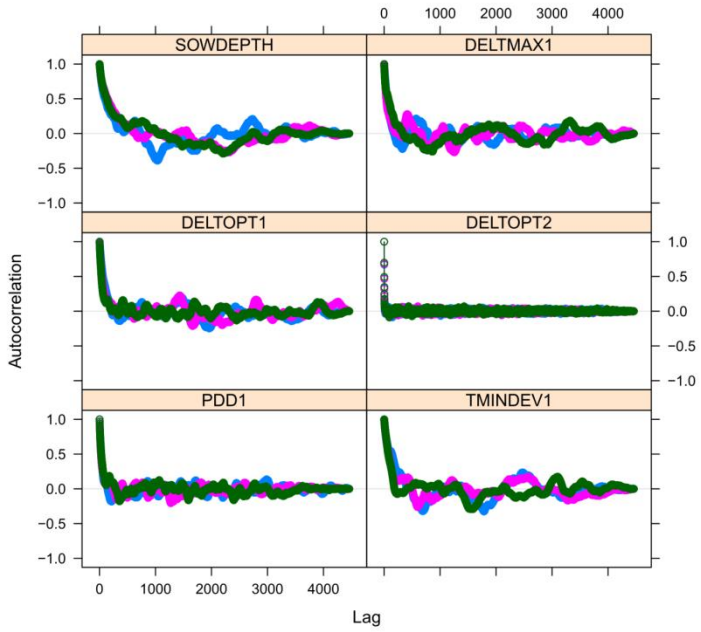


Figure S30: Auto-correlation plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in Kraichgau at 3_2011 and 2_2012. The x-axis is the lag distance and y-axis is the auto-correlation. The colours indicate the three chains.

What were the numbers of effective samples (e.g. see https://mc-stan.org/docs/2_28/reference-manual/effective-sample-size.html)?

Please find an excerpt from the Supplementary Materials S7 below. The effective sample size estimates are provided in Table S7-1 in the sections above.

Supp: ***…effective sample size (ESS in Table S7-1) were calculated (mcmcse package in R (Flegal et al., 2021), (Vats et al., 2019))…The effective sample size between 145 and 332, together with the Gelman-Rubin convergence diagnostic (<=1.1), provide sufficiently reliable posterior statistics for this study.***

It is essential that these questions be addressed and I think doing so should not require more than adding a paragraph or two of text and possibly a supporting figure.

I also have several other specific suggestions that I think would improve the manuscript:

- line 140 Please indicate the identity of the expert (possibly in the Table 2 caption?) Citing as personal communication?

Since the expert is one of the co-authors, (Sebastian Gayler), we rephrased this sentence:

MS, line 148: ***Parameters were pre-selected (Hue et al., 2008; Makowski et al., 2006) based on expert knowledge. The prior default values and uncertainty ranges are given in Table 2.***

- line 176-189 A flow chart to show sequence of steps described would be very helpful.

We have included a flow-chart in Supplementary Materials S8. Please find the excerpt below:

Supp: ***Figure S31 explains the concept of Bayesian Sequential Updating and the methodology used to implement it in this study. For the first site-year, a prior based on expert knowledge (initial prior) is used. In the next sequential update with site-year 2 data, the parameter posterior probability distribution after model calibration to site-year 1 can be used as a prior distribution. This can be repeated for n site-years. In this study, however, instead of using the previous site-year as prior for the next update, we use the initial prior and only update the likelihood function with new data.***

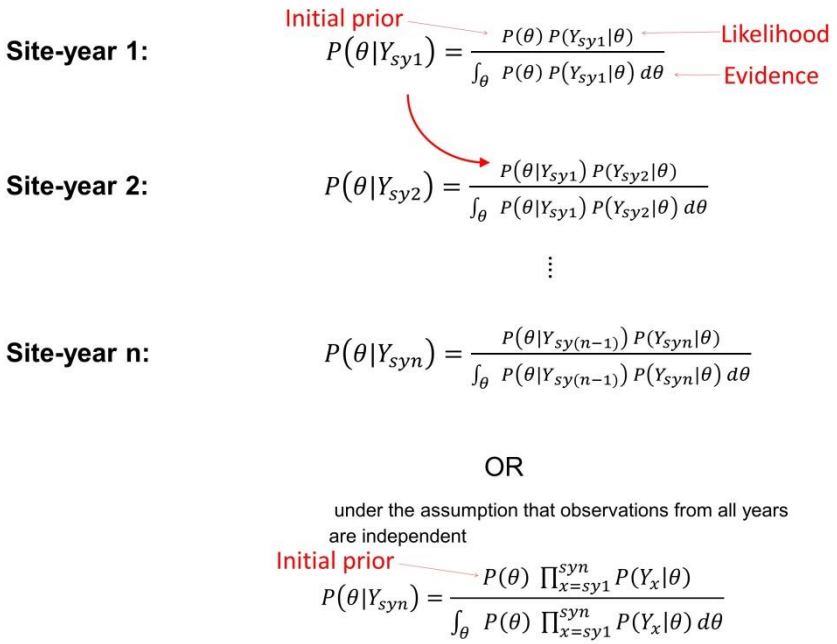


Figure S31: A schematic sketch to explain the concept of Bayesian Sequential Updating (BSU) and its implementation in this study

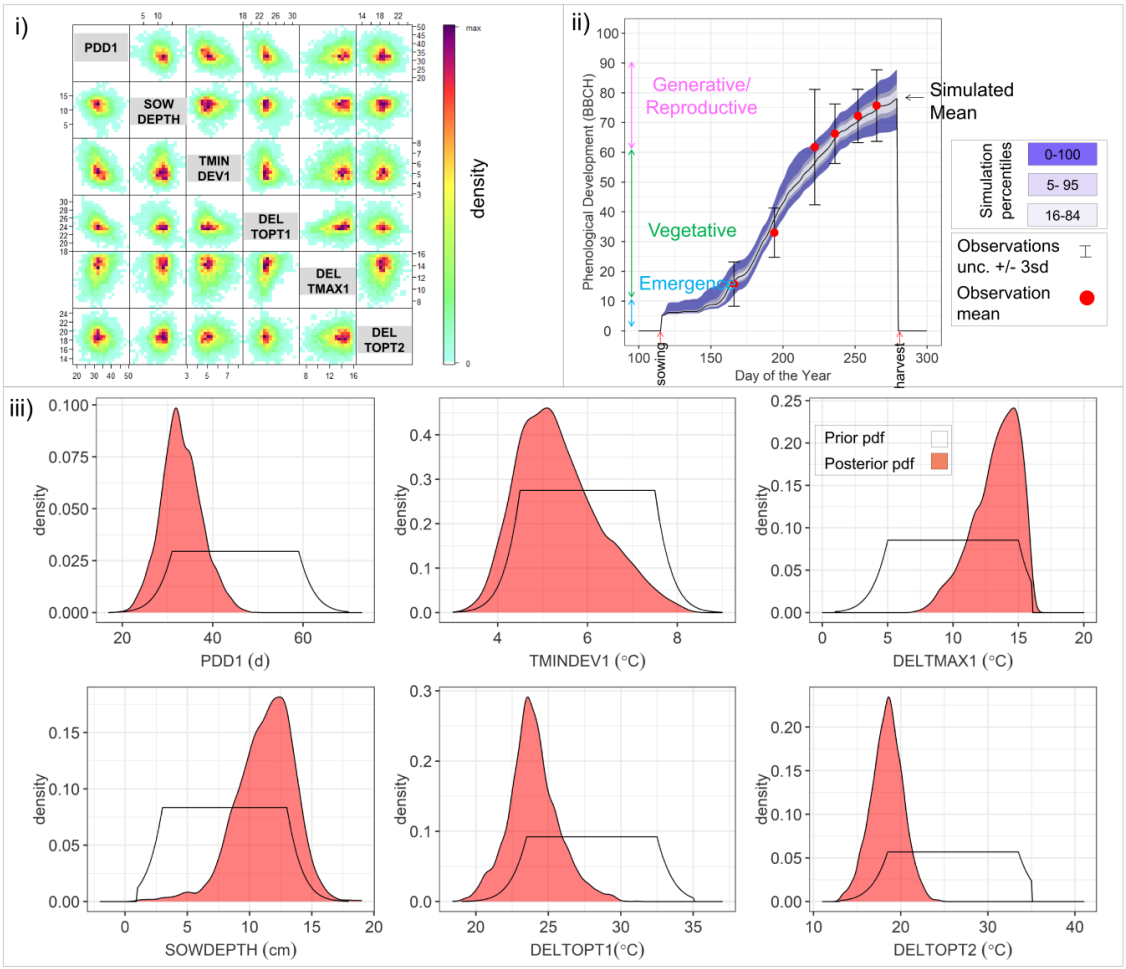
- line 210-215 A figure visualizing the shape of eq 10 and 11 would be very helpful.

The shape of the prior distribution is seen in Figure 4 (iii). We have included a reference to this figure at this point in the text.

MS, line 224: ***…prior probability distribution that is a convolution of a uniform and a normal distribution (Fig. 4 iii) of the form:…***

- In Figure 4(ii) What is meant by the term "Generative"? Is that referring to the "Reproductive" phase of growth (i.e. post-flowering)?

Yes. We have updated the figure and renamed the phase to 'Generative/ Reproductive' so as to avoid confusion. Please refer to the updated figure below.



- I suggest adding some more discussion of the posterior distribution of parameter values presented in Figure 5. For example, why are there differences in parameter values across the two sites? Why do PDD1 and DELTMAX1 both decrease when the sequential years are added? Also, the shifts in distribution from prior to posterior indicate learning from the data. What do those shifts tell you about the cropping/soil system that was not known beforehand?

We thank you for this crucial comment. We have included the following paragraph in the discussion section:

MS, line 437: ***The optimum temperatures for vegetative (TOPTDEV1= TMINDEV1 + DELTOPT1) and reproductive (TOPTDEV2 = 8 + DELTOPT2) development are lower than our prior belief. The effective sowing depth (SOWDEPTH) is higher than the actual sowing depth of 3-5cm as the model cannot capture slow emergence (as discussed in the Appendix A: SPASS phenology model). In Kraichgau, the posterior distributions for SOWDEPTH and minimum temperature for vegetative development (TMINDEV1) did not change significantly as compared to the prior, indicating that the model did not learn much from the data. These parameters, however, show a change from the prior in the Swabian Alb. Kraichgau is warmer than the Swabian Alb. On most days, temperatures in Kraichgau are above the minimum temperature for vegetative development (TMINDEV1), resulting in limited learning. A similar reasoning applies to SOWDEPTH which is a proxy parameter that impacts emergence rate. Emergence occurs only above a certain threshold***

temperature which is hard-coded in the model. Temperatures in Kraichgau are mostly above this threshold temperature for emergence, resulting in limited learning and insignificant change from the prior distribution. In the Kraichgau sequence (Fig. 5i-b), PDD1 and DELTMAX1 decrease when site-year 1_2014 is added to the calibration sequence. Both parameters cause a faster development rate during the vegetative phase. This faster vegetative development results in earlier initiation of the reproductive phase, as seen in the mid-early ripening cultivar 1_2014 as compared to the late cultivars 3_2011 and 2_2012. In the Swabian Alb sequence (Fig. 5i-a), inclusion of early cultivars at 5_2012 and 5_2016 results in shallower SOWDEPTH and consequently, faster emergence. However, whether this early emergence is truly a feature of early cultivars or a consequence of the timing of first observations in the growing season cannot be satisfactorily distinguished with the available data. The physiological development days at optimum vegetative phase temperature (PDD1) were also lower than our initial prior belief. We, however, interpret these results with caution as parameters may compensate for model structural errors and some parameters are correlated (Alderman & Stanfill, 2017).

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. https://doi.org/10.1016/j.eja.2016.09.016

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., & Maji, U. (2021). mcmcse: Monte Carlo Standard Errors for MCMC. Riverside, CA, and Kanpur, India.

Gelman, A., Roberts, G. O., & Gilks, R. W. (1996). Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics* (Vol. 5, pp. 599–608). Oxford University Press.

Gelman, Andrew, & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511. Retrieved from https://projecteuclid.org/euclid.ss/1177011136

Hue, C., Tremblay, M., & Wallach, D. (2008). A bayesian approach to crop Model calibration under unknown error covariance. *Journal of Agricultural, Biological, and Environmental Statistics*, *13*(3), 355–365. https://doi.org/10.1198/108571108X335855

Makowski, D., Hillier, J., Wallach, D., Andrieu, B., & Jeuffroy, M. H. (2006). Parameter Estimation for Crop Models. In *Working with Dynamic Crop Models*. Elsevier.

Metropolis, N., Rosenbluth, A. ., Rosenbluth, M. ., & Teller, A. . (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6). Retrieved from https://bayes.wustl.edu/Manual/EquationOfState.pdf

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

Tautenhahn, S., Heilmeier, H., Jung, M., Kahl, A., Kattge, J., Moffat, A., & Wirth, C. (2012). Beyond distance-invariant survival in inverse recruitment modeling: A case study in Siberian Pinus sylvestris forests. *Ecological Modelling*, *233*, 90–103. https://doi.org/10.1016/j.ecolmodel.2012.03.009

Vats, D., Flegal, J. M., & Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, *106*(2), 321–337. https://doi.org/10.1093/biomet/asz002

S7.MCMC diagnostics

The sequential update calibration cases for the true sequences in the Swabian Alb and in Kraichgau are listed in Table S7-1. The number of iterations required to adapt the jump-size (A) were variable (20-580) and dependent on the calibration case. In some cases this number was low because we set the initial pre-adaptation value for the standard deviation of the transition kernel so that the acceptance rate would be between 25% and 35%. This initial value was based on knowledge gained from preliminary calibration test runs. The jump adjustment factor (f) in Table S7-1 influences the standard deviation of the transition kernel (V) such that $\mathrm{V} = \frac{\mathrm{sd}}{\mathrm{f}}$ where sd is the standard deviation of the prior parameter distributions taken from Table 2 in the main text. With N being the total number of iterations per chain, the total number of iterations across the three chains after burn-in is given by $\mathrm{T} = (\mathrm{N} - \mathrm{A}) \times 3$.

On adding a new site-year, the chains were re-initialized and the transition kernel was re-tuned. New data was added to the dataset and the chains were allowed to adapt. The burn-in was variable and dependent of the jump-size adaptation. We ensured that a minimum of 500 accepted samples were generated per chain, that is, a minimum of 1500 total samples across chains were drawn. However, the actual number of samples drawn (T) was higher and dependent of when the Gelman-Rubin convergence diagnostic was <=1.1.

To assess parameter mixing, trace-plots were analysed (examples provided in Figure S27 and Figure S28). Additionally, auto-correlation plots (Figure S29, Figure S30) are provided (coda package in R (Plummer et al., 2006)) and effective sample size (ESS in Table S7-1) were calculated (mcmcse package in R (Flegal et al., 2021), (Vats et al., 2019)). Parameter DELTOPT2 generally showed good mixing and low auto-correlation. The effective sample size between 145 and 332, together with the Gelman-Rubin convergence diagnostic (<=1.1), provide sufficiently reliable posterior statistics for this study.

Table S7-1: MCMC sampling details for True sequence calibration cases in Kraichgau and the Swabian Alb

Sequence	**Calibration case**	**Number of accepted runs per chain during jump adaptation (A)**	**Jump adjustment factor (f)**	**Total accepted samples per chain (N)**	**Total samples after burn-in in all chains (T)** $= (\boldsymbol{N} - \boldsymbol{A}) \times \mathbf{3}$	**ESS**
True sequence Swabian Alb	6_2010	20	3	1480	4380	236
	6_2010, 5_2011	580	3.97	1100	1560	332
	6_2010, 5_2011, 5_2012	20	5	800	2340	145
	6_2010, 5_2011, 5_2012, 6_2013	40	4.95	820	2340	167
	6_2010, 5_2011, 5_2012, 6_2013, 5_2015	20	5	620	1800	196
	6_2010, 5_2011, 5_2012, 6_2013, 5_2015, 5_2016	240	6.7	1400	3480	159
True Sequence Kraichgau	3_2011	60	5.005	3280	9660	153
	3_2011, 2_2012	20	5	4480	13380	163
	3_2011, 2_2012, 1_2014	20	7.7	5100	15240	168

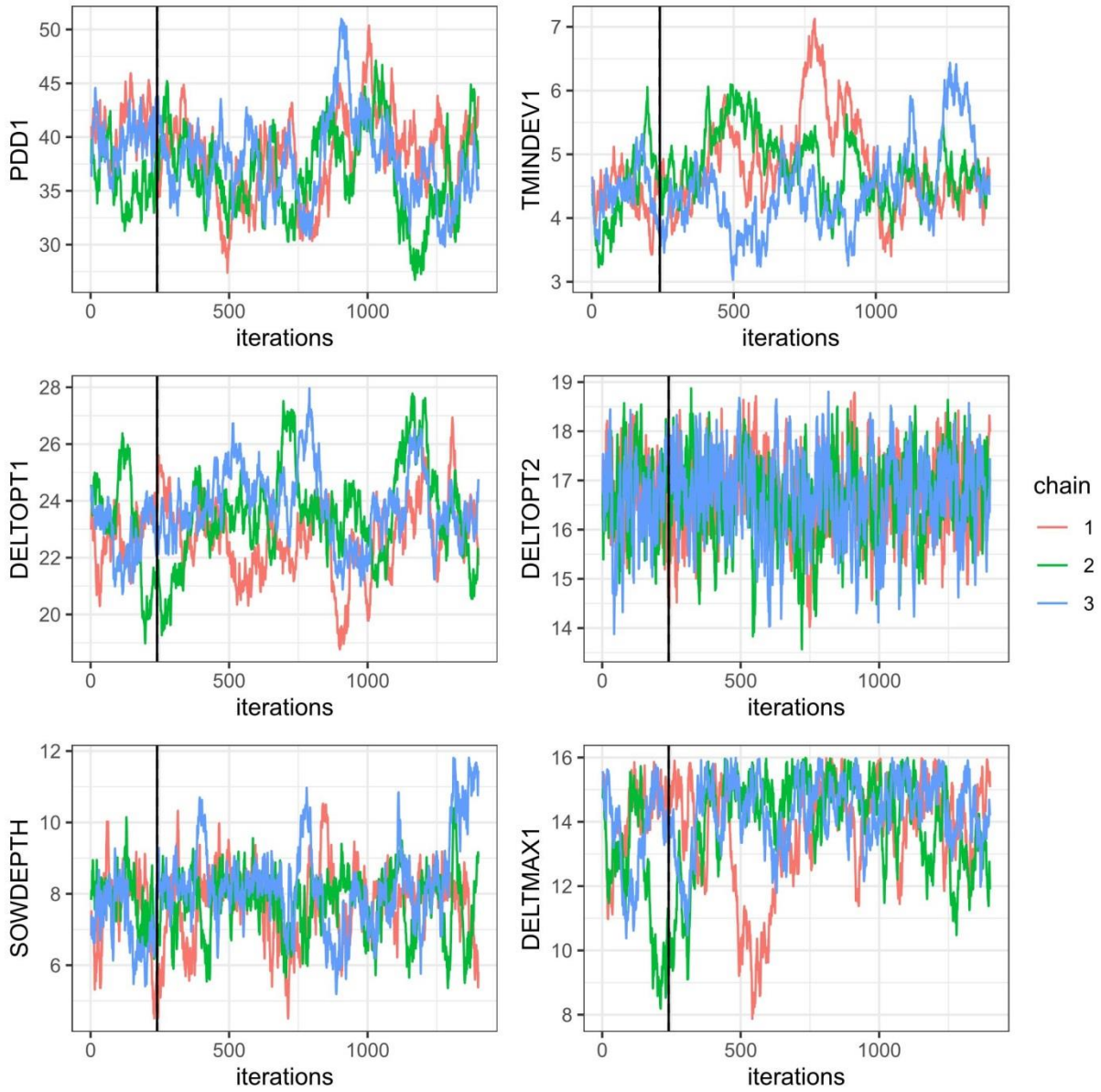


Figure S27 Trace-plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in the Swabian Alb at 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016. The x-axis is the number of iterations and y-axis is the parameter. The colours indicate the three chains. The black solid vertical line indicates the burn-in phase during which the transition kernel was adapted.

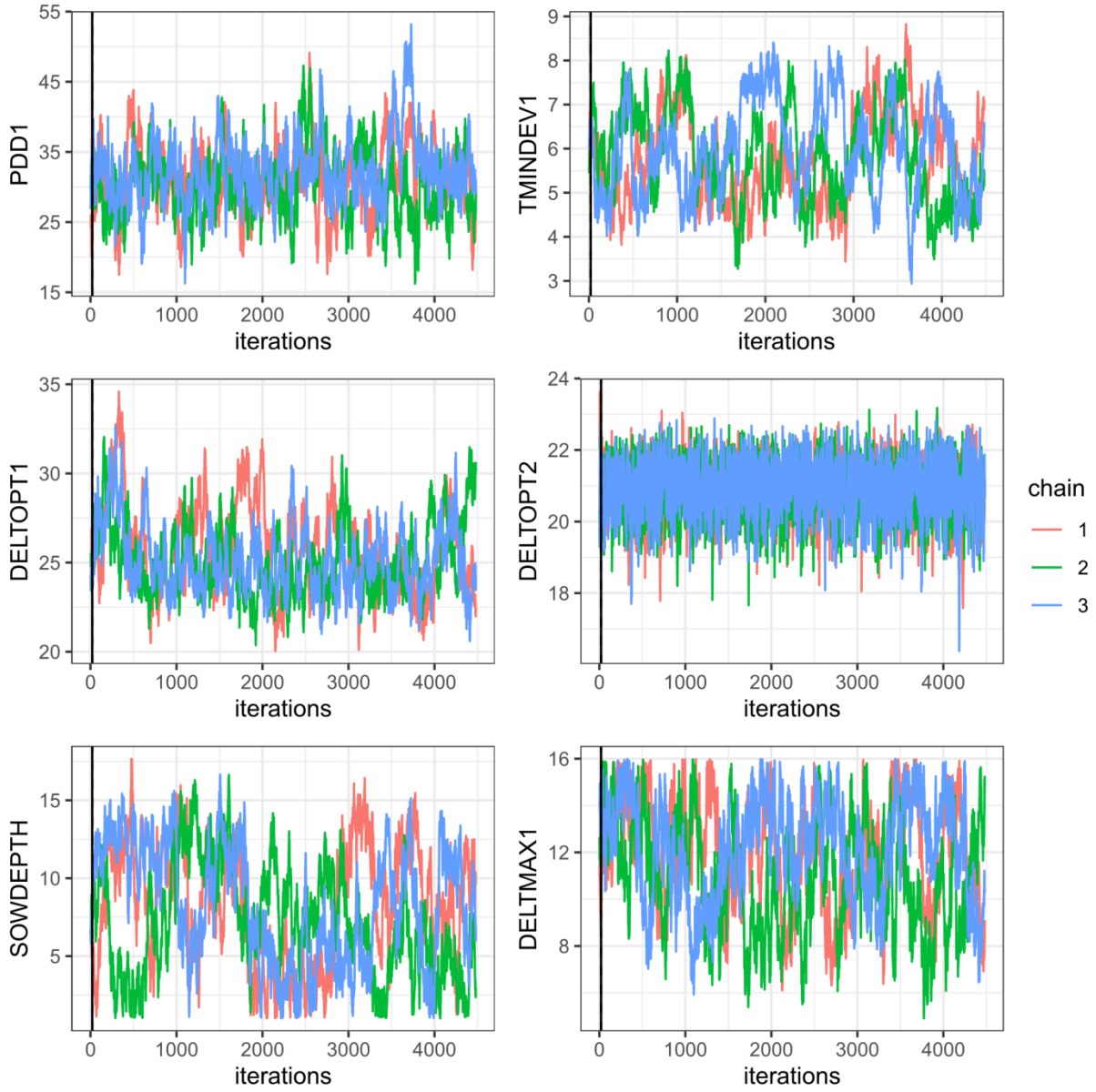


Figure S28 Trace-plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in Kraichgau at 3_2011 and 2_2012. The x-axis is the number of iterations and y-axis is the parameter. The colours indicate the three chains. The black solid vertical line indicates the burn-in phase during which the transition kernel was adapted.

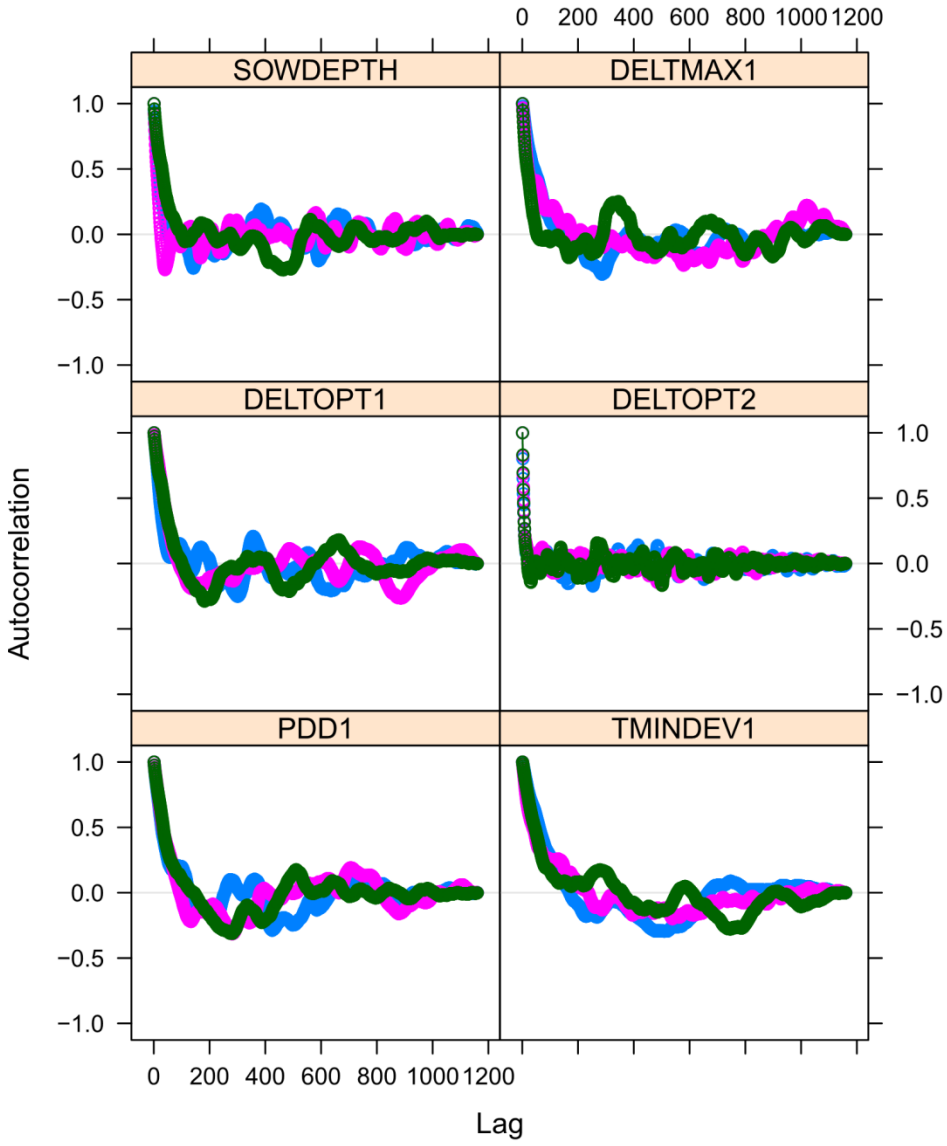


Figure S29 Auto-correlation plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in the Swabian Alb at 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016. The x-axis is the lag distance and y-axis is the auto-correlation. The colours indicate the three chains.

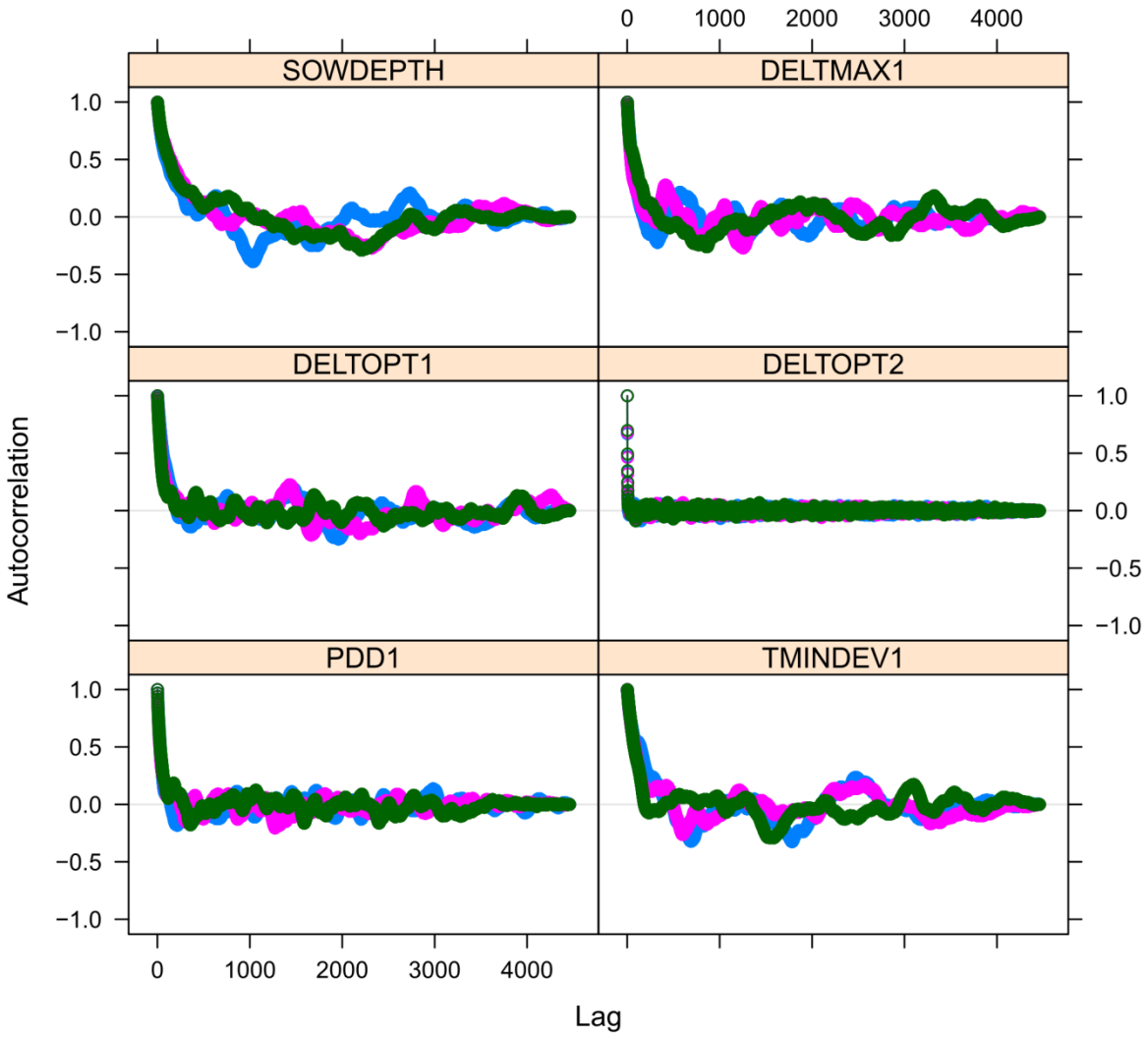


Figure S30 Auto-correlation plots of 6 estimated parameters for the true sequence calibration of SPASS to phenology grown in Kraichgau at 3_2011 and 2_2012. The x-axis is the lag distance and y-axis is the auto-correlation. The colours indicate the three chains.

S8.BSU implementation

Figure S31 explains the concept of Bayesian Sequential Updating and the methodology used to implement it in this study. For the first site-year, a prior based on expert knowledge (initial prior) is used. In the next sequential update with site-year 2 data, the parameter posterior probability distribution after model calibration to site-year 1 can be used as a prior distribution. This can be repeated for n site-years. In this study, however, instead of using the previous site-year as prior for the next update, we use the initial prior and only update the likelihood function with new data.

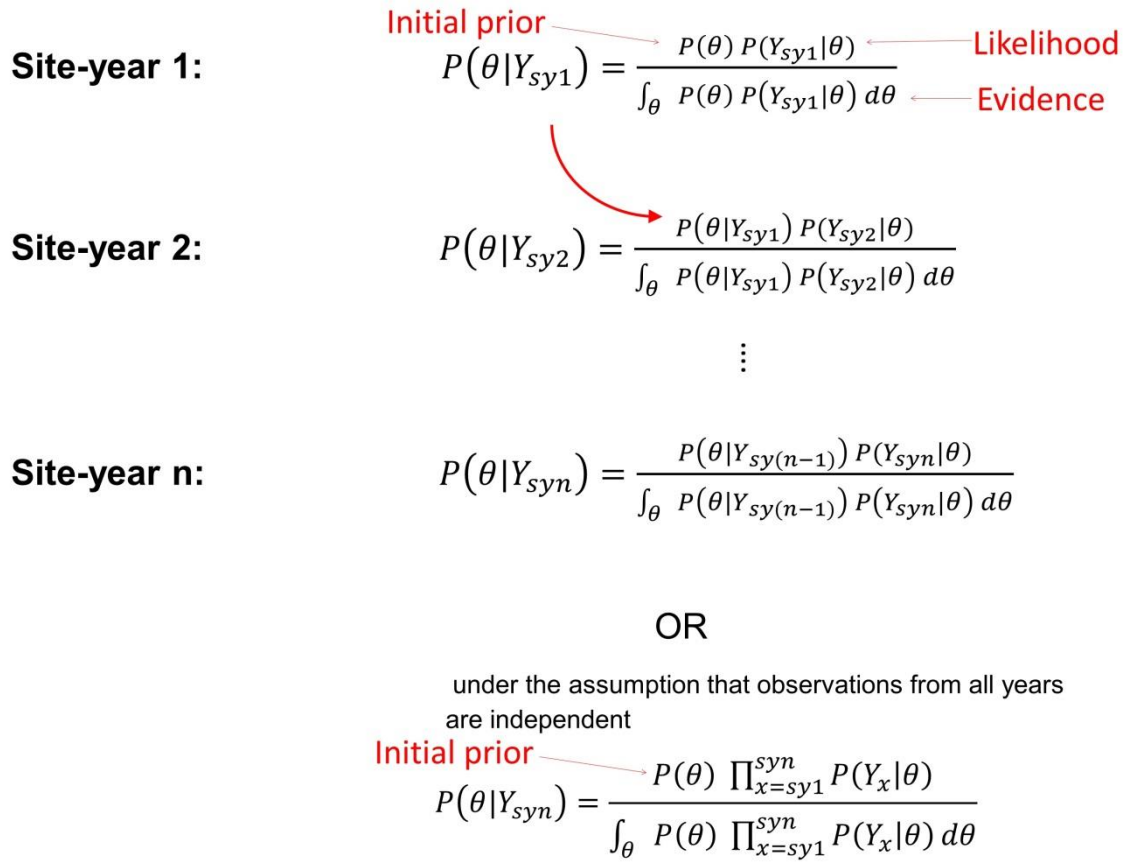


Figure S31 A schematic sketch to explain the concept of Bayesian Sequential Updating (BSU) and its implementation in this study