Bayesian Multi-Purpose Modelling of Plant Growth and Development Across Scales

Dissertation to obtain the doctoral degree of Agricultural Sciences (Dr. sc. agr.)

Faculty of Agricultural Sciences
University of Hohenheim
Institute of Soil Science and Land Evaluation

by
Michelle Viswanathan
from Mumbai, India

2024

This thesis was accepted as a doctoral thesis (Dissertation) in fulfillment of the regulations to acquire the doctoral degree "Doktor der Agrarwissenschaften" by the Faculty of Agricultural Sciences at University of Hohenheim on 30th June 2023

Date of oral examination:	21st December 2023
Dean of Faculty:	Prof. Dr. Ralf T. Vögele

Examination committee:

Chairperson:	Prof. Dr. Martin Hasselmann
Supervisor and Reviewer:	Prof. Dr. Thilo Streck
Co-Reviewer:	Prof. Dr. -Ing. Wolfgang Nowak
Additional Examiner:	Prof. Dr. Simone Graeff-Hönninger

Acknowledgement

I thank Prof. Dr. Thilo Streck for the opportunity to pursue my doctoral degree under his supervision, for his invaluable feedback, advice and support. I also thank my co-supervisor Dr. Tobias Weber for the discussions and guidance that helped in crystallizing my doctoral research work. I am grateful to Dr. Juliane Mai, who apart from supervising, has been a constant source of encouragement. I would also like to express my gratitude to Dr. Sebastian Gayler, Andreas Scheidegger, Dr. Anneli Guthke, Dr. Ming Han with whom it has been a great pleasure to work and I look forward to doing so in the future.

I thank the German Research Foundation (DFG) for funding my doctoral research. I would like to express my gratitude to Prof. Dr. Olaf Cirpka and my fellow IRTG members for the scientific exchange and social activities in the IRTG programme. A special thanks to my friends and colleagues at the department of Biogeophysics (and the Meiereihof) who made even the slow days enjoyable. I will always cherish our chats and discussions over coffee and beer. Ich danke Werner Scharlowski für seine Gastfreundschaft, seine Freundlichkeit und dafür, dass er seine Weisheit mitgeteilt hat. I am grateful to Sibylle Schulz and Monika Jekelius for handling all the paperwork behind the scenes.

My parents have always been a source of inspiration for hard work and creativity. I am thankful to them for making learning fun; it has stayed with me through the years. Last but not the least, I am grateful to my husband Jonathan Minz, for his patience and positivity. I look forward to our next adventure together.

Executive Summary

Crop models are invaluable tools for predicting the impact of climate change on crop production and assessing the fate of agrochemicals in the environment. To ensure robust predictions of crop yield, for example, models are usually calibrated to observations of plant growth and phenological development using different methods. However, various sources of uncertainty exist in the model inputs, parameters, equations, observations, etc., which need to be quantified, especially when model predictions influence decision-making. Bayesian inference is suitable for this purpose since it enables different uncertainties to be taken into account, while also incorporating prior knowledge. Thus, Bayesian methods are used for model calibration to improve the model and enhance prediction quality.

However, this improvement in the model and its prediction quality does not always occur due to the presence of model errors. These errors are a result of incomplete knowledge or simplifying assumptions made to reduce model complexity and computational costs. For instance, crop models are used for regional scale simulations thereby assuming that these point-based models are able to represent processes that act at regional scale. Additionally, simple statistical assumptions are made about uncertainty in model errors during Bayesian calibration. In this work, the problems arising from such applications are analysed and other Bayesian approaches are investigated as potential solutions.

A conceptually simple Bayesian approach of sequentially updating a maize phenology model, an important component in plant models, was investigated as yearly observation data were gathered. In this approach, model parameters and their uncertainty were estimated while accounting for observation uncertainty. As the model was calibrated to increasing amounts of observation data, the uncertainty in the model parameters reduced as expected. However, the prediction quality of the calibrated model did not always improve in spite of more data being available for potentially improving the model. This discrepancy was attributed to the presence of errors in the model structure, possibly due to missing environmental dependencies that were ignored during calibration.

As a potential solution, the model was calibrated using Bayesian multi-level modelling which could account for model errors. Furthermore, this approach accounted for the hierarchical data structure of cultivars nested within maize ripening groups, thus simultaneously obtaining model parameter estimates for the species, ripening groups and cultivars. Applying this approach improved the model's calibration quality and further aided in identifying possible model deficits related to temperature effects in the post-flowering phase of development and soil moisture.

As another potential solution, an alternative calibration strategy was tested which accounted for model errors by relaxing the strict statistical assumptions in classical Bayesian inference. This was done by first acknowledging that due to model errors, different data sets may yield diverse solutions to the calibration problem. Thus, instead of fitting the model to all data sets together and finding a compromise solution, a fit was found to each data set. This was implemented by modifying the likelihood, a term that accounts for information content of the data. An additive rather than the classical multiplicative strategy was used to combine likelihood values from different data sets. This approach resulted in conservative but more reliable predictions than the classical approach in most cases. The classical approach resulted in better predictions only when the prediction target represented an average of the calibration data.

The above-mentioned results show that Bayesian methods with representative error assumptions lead to improved model performance and a more realistic quantification of uncertainties. This is a step towards the effective application of process-based crop models for developing suitable adaptation and mitigation strategies.

Zusammenfassung

Pflanzenwachstumsmodelle sind wertvolle Instrumente für die Vorhersage der Auswirkungen des Klimawandels auf die Pflanzenproduktion und die Beurteilung des Verbleibs von Agrochemikalien in der Umwelt. Um zuverlässige Vorhersagen z. B. für Ernteerträge zu gewährleisten, werden die Modelle in der Regel mittels Beobachtungen des Pflanzenwachstums und der Pflanzenentwicklung kalibriert. In den Modelleingaben, Parametern, Gleichungen, Beobachtungen usw. gibt es jedoch verschiedene Quellen der Unsicherheit, die quantifiziert werden müssen, insbesondere wenn die Modellvorhersagen als Basis von Managementscheidungen dienen sollen. Die Bayes'sche Inferenz ist eine für diesen Zweck geeignete Methode, da sie ermöglicht, verschiedene Unsicherheiten zu berücksichtigen und dabei auch Vorwissen einzubeziehen. Daher werden Bayes'sche Methoden zur Modellkalibrierung eingesetzt, um das Modell zu verbessern und die Vorhersagequalität zu erhöhen.

Die erwartete Verbesserung des Modells und seiner Vorhersagequalität tritt jedoch nicht immer ein, da strukturelle Modellfehler vorhanden sein können. Diese Fehler sind das Ergebnis unvollständigen Wissens oder vereinfachender Annahmen, die zur Verringerung der Modellkomplexität und der Rechenzeit getroffen wurden. So werden beispielsweise Erntemodelle für Simulationen auf regionaler Ebene verwendet, wobei davon ausgegangen wird, dass diese punktbasierten Modelle in der Lage sind, Prozesse darzustellen, die auf regionaler Ebene ablaufen. Außerdem werden bei der Bayes'schen Kalibrierung einfache statistische Annahmen über die Unsicherheit der Modellfehler getroffen. In dieser Arbeit werden die Probleme, die sich aus solchen Anwendungen ergeben, analysiert und andere Bayes'sche Ansätze als mögliche Lösungen untersucht.

Ein konzeptionell einfacher Bayes'scher Ansatz zur sequentiellen Aktualisierung eines Phänologiemodells für Mais, einer wichtigen Komponente in Pflanzenmodellen, untersucht. Das bedeutet, dass sukzessive Beobachtungsdaten hinzugefügt wurden. Bei diesem Ansatz wurden die Modellparameter und ihre Unsicherheit unter Berücksichtigung der Beobachtungsunsicherheit geschätzt. Mit zunehmender Menge an Beobachtungsdaten wurde das Modell immer wieder re-kalibriert, wobei die Unsicherheit der Modellparameter erwartungsgemäß abnahm. Anders als erwartet, verbesserten die zusätzlichen Daten die Vorhersagequalität nicht immer. Diese Diskrepanz

wurde auf das Vorhandensein von Fehlern in der Modellstruktur zurückgeführt, die möglicherweise auf nicht berücksichtigte Abhängigkeiten von Umweltfaktoren zurückzuführen sind.

Als eine mögliche Lösung wurde das Modell mit Hilfe der Bayes'schen Multilevel-Modellierung kalibriert, einer Methode, bei der solche Modellfehler einbezogen werden können. Darüber hinaus berücksichtigt dieser Ansatz die hierarchische Struktur der Sorten, wobei die untersuchten Maissorten in Reifegruppen zusammengefasst werden. Dies ermöglichte die gleichzeitige Schätzung von Modellparametern für Mais, die Reifegruppen und die untersuchten Sorten. Dieser Ansatz verbesserte die Kalibrierung des Modells und half außerdem bei der Identifizierung möglicher Modelldefizite im Zusammenhang mit Temperatureffekten in der Entwicklungsphase nach der Blüte und der Bodenfeuchtigkeit.

Als weitere mögliche Lösung wurde eine alternative Kalibrierungsstrategie getestet, bei der Modellfehler durch Lockerung der strengen statistischen Annahmen der klassischen Bayes'schen Inferenz berücksichtigt wurden. Dabei wurde zunächst anerkannt, dass verschiedene Datensätze aufgrund von Modellfehlern unterschiedliche Lösungen für das Kalibrierungsproblem liefern können. Anstatt das Modell an alle Datensätze gemeinsam anzupassen und eine Kompromisslösung zu finden, wurde eine Anpassung für jeden Datensatz gesucht. Dazu wurde die Likelihood modifiziert, die den Informationsgehalt der Daten berücksichtigt. Um die Likelihood-Werte aus verschiedenen Datensätzen kombinieren zu können, wurde eine additive statt der klassischen multiplikativen Strategie verwendet. Dieser Ansatz führte in den meisten Fällen zu konservativeren, aber zuverlässigeren Vorhersagen als der klassische Ansatz. Der klassische Ansatz führte nur dann zu besseren Vorhersagen, wenn das Vorhersageziel einen Durchschnitt der Kalibrierungsdaten darstellte.

Die oben genannten Ergebnisse zeigen, dass Bayes'sche Methoden mit repräsentativen Fehlerannahmen zu einer verbesserten Modell-Performance und einer realistischeren Quantifizierung von Unsicherheiten führen. Dies ist ein Schritt in Richtung einer effektiven Anwendung prozessbasierter Pflanzenwachstumsmodelle bei der Entwicklung geeigneter Anpassungs- und Eindämmungsstrategien.

Table of Contents

1 Introduction 15
- 1.1 Process-based crop models . . . 17
- 1.2 Applications of Bayesian inference . . . 18
- 1.3 Research questions . . . 19
- 1.4 Research objectives and scope . . . 21

2 Theory & Methods 22
- 2.1 Bayesian inference . . . 22
  - 2.1.1 Likelihood . . . 23
- 2.2 Synthetic case study . . . 24
  - 2.2.1 Calibration approaches . . . 25
  - 2.2.2 Comparing the approaches . . . 27

3 A Bayesian sequential updating approach to predict phenology of silage maize 32
- 3.1 Introduction . . . 33
- 3.2 Materials and methods . . . 35
  - 3.2.1 Study sites and measured data . . . 35
  - 3.2.2 Soil–crop model . . . 38
  - 3.2.3 Selection of model parameters . . . 38
  - 3.2.4 Bayesian sequential updating . . . 40
    - 3.2.4.1 Likelihood function . . . 41
    - 3.2.4.2 Prior probability distribution . . . 42
    - 3.2.4.3 Posterior probability distribution . . . 42
  - 3.2.5 Performance metrics . . . 43
  - 3.2.6 Modelling cases . . . 44
    - 3.2.6.1 Synthetic sequences . . . 45
    - 3.2.6.2 True sequences . . . 46
- 3.3 Results . . . 46

3.3.1 Bayesian calibration results . . . 47
3.3.2 Parameter uncertainty . . . 48
3.3.3 Prediction quality . . . 48
3.3.3.1 Synthetic sequences . . . 48
3.3.3.2 True sequences . . . 48
3.4 Discussion . . . 52
3.4.1 Parameter uncertainty . . . 52
3.4.2 Prediction quality . . . 53
3.4.3 Limitations . . . 55
3.5 Conclusions . . . 56
3.6 Appendix A: SPASS phenology model . . . 57
3.7 Appendix B: Posterior sampling using MCMC Metropolis algorithm . . . 59
3.8 Appendix C. Single site-year calibration . . . 61
3.9 Appendix D: Platykurtic prior . . . 63

4 Bayesian multi-level calibration of a process-based maize phenology model 64
4.1 Introduction . . . 65
4.2 Materials and Methods . . . 67
4.2.1 Data . . . 67
4.2.2 Phenology model . . . 68
4.2.3 Bayesian model cases . . . 70
4.2.3.1 BM-0 . . . 72
4.2.3.2 BMM-1 . . . 72
4.2.3.3 BMM-2a . . . 72
4.2.3.4 BMM-2b . . . 73
4.2.3.5 BMM-3 . . . 73
4.2.4 Posterior sampling . . . 73
4.2.5 Calibration performance metrics . . . 74
4.3 Results . . . 74
4.3.1 Classification of site-years into environmental classes . . . 74
4.3.2 Calibration quality . . . 75
4.3.3 Phenology model parameters . . . 79
4.3.4 Environmental effects . . . 81
4.3.5 Residual uncertainty . . . 84
4.4 Discussion . . . 84
4.5 Conclusions . . . 88

4.6 Appendix A. Weather class clustering . . . 89
4.7 Appendix B. SPASS model equations . . . 90
4.8 Appendix C. Model performance metrics . . . 91
4.9 Appendix D. SPASS model parameter distributions . . . 91

5 From Bayesian Multiplicative to Additive Calibration Strategy for More Reliable Predictions - A Demonstration on Plant Phenology Modelling 94

5.1 Introduction . . . 95
5.2 Bayesian Model Calibration . . . 99
  5.2.1 Bayesian Updating . . . 99
  5.2.2 Likelihood Formulation in the Traditional Multiplicative Calibration Scheme 99
  5.2.3 Likelihood Formulation in the Additive Calibration Scheme . . . 100
  5.2.4 Skill Score Used to Evaluate Predictive Performance . . . 101
5.3 Demonstration in a Crop Phenology Modelling Case Study . . . 101
  5.3.1 Motivation and Goals . . . 101
  5.3.2 Data . . . 102
  5.3.3 Model . . . 103
  5.3.4 Calibration Schemes in the Context of Site-Years . . . 104
  5.3.5 Test Case Scenarios . . . 106
    5.3.5.1 Synthetic Case Study . . . 106
    5.3.5.2 Real-World Case Study . . . 107
  5.3.6 Numerical Implementation . . . 109
5.4 Results and Discussion . . . 110
  5.4.1 Synthetic Case Study . . . 110
  5.4.2 Real-World Case Study . . . 112
    5.4.2.1 Conservative but Reliable Predictions . . . 113
    5.4.2.2 When the Target Represents an Average Behaviour . . . 114
    5.4.2.3 Representativeness of the Calibration Data . . . 115
    5.4.2.4 Data Set Selection for a Successful Additive Strategy . . . 116
5.5 Summary, Implications and Outlook . . . 119
5.6 Appendix A: SPASS Phenology Model in R . . . 120
5.7 Appendix B: Prior Distribution . . . 120
5.8 Appendix C: Predictive Log-Score (PLS): Real-World Case Study . . . 121

6 Conclusions & Outlook 122

6.1 Summary . . . 122

6.2 Choosing an appropriate approach 123
6.3 Future scope 124
6.4 Concluding remarks 125

Supplementary Materials **139**

List of Figures

2.1 Comparison of the species parameters from different Bayesian approaches 29

2.2 Comparison of the cultivar-specific parameters from different Bayesian approaches 30

3.1 Study site locations and example of observed phenology 36

3.2 Site-years used for calibration and validation in the true and synthetic sequences 44

3.3 Phenology data used in the two synthetic sequences 45

3.4 Results of Bayesian calibration of the model to phenological development 47

3.5 Marginal parameter distributions and entropy 49

3.6 Model performance metrics in the synthetic sequences 50

3.7 Model performance metrics in the true sequences 51

3.C1 Model performance metrics after single site-year calibration 62

3.C2 Relation between median NRMSE and temperature 62

3.D1 Platykurtic probability density function 63

4.1 Graphical representation of the five Bayesian models 71

4.2 Classication of site-years into weathre classes 76

4.3 Silage maize ripening groups grown in different eco-regions across Germany . . . 77

4.4 Observed and simulated maize phenology in the pooled and full multi-level model cases 78

4.5 Comparison of RMSE and bias across model cases 79

4.6 Comparison of RMSE and bias across model cases by ripening group and weather class 80

4.7 Posterior distribution of parameters exhibiting low an high between-cultivar variability 81

4.8 Prior and posterior distributions of eco-region and weather effect parameters . . 82

4.9 Median weather effect parameters by average daily temperature during reproductive development 82

4.10 Median eco-region effect by eco-region in all multi-level model cases 83

4.11 Posterior distributions of residual uncertainty in all model cases 84
4.C1 Comparison of RMSE and bias across all model cases by eco-region and year . . 92
4.D1 Posterior distribution of process-model parameters: low between-cultivar variability 93
4.D2 Posterior distribution of process-model parameters: high between-cultivar variability 93

5.1 Prediction scenarios in the synthetic case study 107
5.2 Prediction scenarios for the real-world case study 108
5.3 Posterior parameter distributions from the synthetic case study 111
5.4 Prediction results from the synthetic case study 111
5.5 Predictive log-score from the synthetic case study 112
5.6 Prediction results from the real-world case study (low information scenario) . . . 114
5.7 Prediction results from the real-world case study (high information scenario) . . 118
5.C1 predictive log score from the real-world case study 121

List of Tables

2.1 Posterior parameter distributions obtained from the different calibration approaches 27

3.1 Site-years used for the study 37
3.2 SPASS phenology model parameters 39

4.1 Summary of site-years and cultivars used for calibration 68
4.2 Prior parameter distributions for all Bayesian model cases 70

5.1 Site-years used in the case study with ripening groups of silage maize and temperature class. 103
5.2 Prior parameter distributions 104

List of acronyms

a.s.l	Above Sea-Level
BBCH	Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie
BHM	Bayesian Hierarchical Modelling
BMM	Bayesian Multi-level Modelling
BSU	Bayesian Sequential Updating
ESS	Effective Sample Size
ET	Evapotranspiration
GR	Gelman-Rubin diagnositc
IQR	Inter-quartile range
LAI	Leaf Area Index
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
NRMSE	Normalized Root Mean Square Error
PDF	Probability Density Function
PLS	Predictive log-score
RMSE	Root Mean Square Error
SD	standard deviation
SM	Soil Moisture
TRF	Temperature Response Function
XN5	Expert-N software version 5

CHAPTER 1

Introduction

Since the Industrial Revolution, the world has seen an increase in population which, in part, has been enabled by advances in agriculture. The ability to develop large farmlands through mechanized agricultural practices, development and dissemination of new high-yielding, pest-resistant and resilient crop varieties facilitated by the Green Revolution (Pingali, 2012), along with progresses made in storage and transportation have been influential in feeding a growing population. However, this comes at a cost; with rampant deforestation, land-use change, environmental degradation due to over-fertilization and persistence of pesticides in soil and water, and greenhouse gas emissions contributing to climate change. In spite of the giant leaps made in agricultural technologies, climate change still threatens our path to a food-secure future. The influences of agricultural practices on the environment as well as the impact of climate change on food production need to be understood to implement sustainable strategies and make policy decisions.

Plants exhibit complex interactions with the environment and responses to management practices. In an attempt to comprehend these processes, mathematical models have been developed since the 1960s. These models fall into two broad categories: statistical and process-based. Statistical models consist of equations that empirically relate observed variables such as weather to target variables of interest such as crop yield. These models require large number of observations in order to establish such relationships. With increasing data availability, statistical models have found increased application in recent years (Panayi et al., 2017). On the other hand, process-based models are not as data-dependent. They represent our knowledge about the underlying biophysical processes in the form of mathematical equations (Lobell and Asseng, 2017). Following the early pioneering work of de Wit (1965) and Duncan (1971) on photosynthesis of leaf canopies, more complex crop models have been developed that can be coupled with weather, soil, and field management models for a more holistic assessment (Jones et al., 2017). The choice between these two classes of models depends on the scope of the study, which could differ in spatio-temporal scale and end-goal. For example, short-term predictions for sustainable field management decisions and yield optimization would dictate a different model choice from studies that focus on large regional-scale, multi-year predictions for climate impact assessment or water and nutrient cycling. The advantage of process-based models lies in the system-understanding that they provide, which is beneficial in exploring climate change adaptation strategies (Lobell and Asseng, 2017), for instance. The lines between these two categories of models are sometimes blurred since process-based models in use today, often contain some empirical relationships in their suite of equations (Pasquel et al., 2022). Process-based models can be combined with statistical inference to improve representation of the modelled system (Lobell and Asseng, 2017; Dietze et al., 2013).

The most common approach of combining process-based models with statistical inference is through model calibration. In the process of calibration, model simulations are compared with observations and the differences between the two are minimized. This is done by tuning parameters in the model. In process-based models, many parameters have a genetic, physiological or biophysical meaning. Their values can sometimes be determined in separate experiments (Craufurd et al., 2013), but very often they need to be estimated through calibration. Among the different calibration techniques used in the field of crop modelling, Bayesian methods have been gaining favour in recent years. The ability to include independent prior knowledge in Bayesian calibration is especially attractive, given the decades of previously-gained knowledge and insights into these models and agricultural systems. It also offers a solution for expressing different uncertainties, such as those in the model inputs, equations, parameters, observations, etc. in terms of the predicted variable of interest. This becomes important when using these models for decision-making (Porter et al., 2015; Rötter et al., 2011).

Thus, Bayesian model calibration is theoretically expected to improve representation of the underlying process by the model and accordingly, its ability to make reliable predictions. However, it does not always meet this goal due to limitations in the process-based models and a violation of statistical assumptions in calibration. In this dissertation, I address these two issues by applying different Bayesian approaches to calibrate crop models. The investigation focuses on crop models applied at regional scales, which are relevant for environmental impact assessment and regional climate impact studies.

In the following sections, the process-based crop models and the types of data used to calibrate them are discussed, with a focus on the problems that arise when such models are applied at regional scale (section 1.1). This is followed by a brief summary of different Bayesian methods used in crop modelling (section 1.2). Key research questions addressed in this work are described, considering model deficits and violated statistical assumptions (section 1.3). Finally, the research objectives are presented (section 1.4).

1.1 Process-based crop models

Process-based plant models are dynamic models which simulate important state variables of interest like crop yield, plant biomass, leaf area index (LAI), phenological development, etc. which are important from economic and ecological perspectives. Crop yield is undoubtedly important for agricultural production, while LAI, defined as the one-sided leaf area per unit area of the ground surface, is additionally used for ecosystem productivity estimation and land-surface modelling (Shi et al., 2015). Phenology defines the timing of occurrence of certain biological events in the plant's life. It is important for making field management decisions (Potgieter et al., 2021) as well as for monitoring temperature-driven changes in the environment such as climate change (Menzel et al., 2006). Furthermore, it also controls assimilate-partitioning in the plant, consequently impacting biomass development and yield. Plant models have also been coupled with soil-water and solute transport models to predict fertilizer fate (Mehdi et al., 2015) or the influence of plant growth on the water-balance (Donohue et al., 2007; Kumar et al., 2019; Zhang et al., 2020). Simulated model outputs of nitrate concentration in the soil, actual evapotranspiration (ET), and soil-moisture (SM) are investigated in such studies. Field observations and measurements of these state variables are used to calibrate such models.

Crop models are commonly calibrated to yield and phenology (Seidel et al., 2018), both of which are either available from field experiments, state administrative office databases or field-specific observation studies. LAI and biomass can be monitored using ground-based or satellite-based remote sensing measurements. Model calibration to these state variables has been shown to improve model simulations of yield (Huang et al., 2015). Coupled models have also been calibrated to satellite based SM and ET estimates (Rajib et al., 2020). These satellite-based measurements provide large spatial coverage that capture spatial heterogeneity better than point-based field measurements and are valuable for regional scale modelling.

Thus, with such new and improved data-gathering techniques we expect to obtain more accurate model parameter estimates on calibration to large good-quality data sets. However, this is usually not the case. Crop models, like most environmental models, are an imperfect representation of the true system. These imperfections hamper the parameter estimation process. Imperfect models are a result of: (a) incomplete or limited understanding of the underlying processes, (b) simplifying assumptions made in the model formulation to reduce model complexity and computation costs, and (c) extending models to well beyond their intended spatial scale of application (Pasquel et al., 2022). Uncertainty in the understanding of processes has led to the development of multiple models and the use of multi-model ensembles (Wallach et al., 2016; Rettie et al., 2022). Sometimes complex models which are more representative of the true system may be available. However, they often have many parameters that need to be determined for ensuring reliable predictions. Limited data availability for estimating these parameters through calibration makes the use of such models difficult. Additionally, complex models may also incur prohibitive computational costs. Thus, simpler models may be preferred in some cases. Many crop models were, in fact, initially developed as point-models (Pasquel et al., 2022) and are now being used at larger spatial scales (Ingwersen et al., 2018) such as in predicting water and nutrient cycles, soil-plant-atmosphere interactions and in climate impact studies. This implies that many spatial processes may not be adequately represented by the model equations (van Oijen et al., 2009). By using point-models for regional applications we make certain simplifying assumptions about the model-representativeness and system-homogeneity. In spite of these limitations, models are nonetheless considered to be valuable tools to assess agricultural processes in current and future scenarios. Therefore, attempts are made to improve model performance by calibrating them to data using methods such as Bayesian inference.

1.2 Applications of Bayesian inference

Bayesian inference has been extensively implemented in crop modelling over the past few years. Through Bayesian methods, uncertainties in inputs, observations, model parameters, and model structure have been taken into account (Ceglar et al., 2011; Iizumi et al., 2009; Sexton, 2015; Alderman and Stanfill, 2017) and expressed as uncertainty of simulated state variables. Methods such as Bayesian Model Averaging (BMA) and Bayesian Model Combination (BMC) have been used to synthesize simulations from multi-model ensembles for future predictions (Wang et al., 2017b; Wöhling et al., 2013, 2015; Gao et al., 2021). Multi-objective calibration, i.e. calibrating the model to data from multiple state-variables (Minet et al., 2015), has been performed to tackle the problem of equifinality (Mo and Beven, 2004). Furthermore, Bayesian methods like data assimilation are implemented to update models in real-time based on time-series data such as satellite-based LAI, soil moisture measurements, etc (Nearing et al., 2012; Zare et al., 2022). While Bayesian methods have yielded promising results, they may lead to erroneous inference and unreliable model predictions.

Erroneous model inference arises from the assumptions made when applying Bayesian methods. The inherent assumption in Bayesian inference is that the model is free of errors or that the errors can be perfectly described. However, as stated earlier, this is usually not the case for crop models. A violation of the assumption leads to erroneous parameter estimates and underestimation of uncertainty, which in turn lead to unreliable model predictions. Nonetheless, Bayesian inference is theoretically sound and has still been widely used since it provides a suitable framework for data-model integration.

Error-prone models and strict Bayesian assumptions may lead to poor inference. Thus, different approaches have been used to overcome these problems (see Chapter 2, section 2.1.1 for details). However, these approaches are still under-explored in the field of crop modelling. With this dissertation I aim to fill this knowledge-gap by using appropriate Bayesian methods that address these short-comings, especially when crop models are applied at regional scale. It is essential to first demonstrate the problem and then investigate different approaches to tackle it. To do so, four research questions are described in section 1.3, which I answer through this research.

1.3 Research questions

Bayesian inference can be used to update our knowledge about a given model and its parameters based on observed data. Observations are thus used to approach a 'true' model that is able to perform well in predicting the state variable of interest in new conditions. This implies that uncertainty in the parameters reduces as compared to our prior knowledge and the model should perform better at predictions, as compared to before the update. A conceptually simple consequence of this concept is to update the model parameters sequentially, as and when new data are gathered. This is referred to as Bayesian sequential updating (BSU). One would hypothesize that as the model is calibrated to more and more data, the parameter uncertainty reduces as we also progressively approach the 'true' value of the parameters and the 'true' model. Thus, predictions from this updated model are also expected to progressively improve. To test this hypothesis on crop models, the following research question is framed:

1. Does Bayesian Sequential Updating (BSU) of a crop model improve predictions?

Delving further into the fallacious assumption that the model is error-free, we inherently assume that the only sources of error are from the observations or measurement processes. Not accounting for other sources of error or lumping several sources into a single error term (Renard et al., 2010) leads to underestimation of prediction uncertainty. This is controlled by assumptions made in defining the likelihood function, a term in Bayes theorem that incorporates the information contained in the observations (see section 2.1.1). Very often simplistic, assumptions are made when defining the likelihood function due to computational costs, limited knowledge of different sources of error, or due to lack of observation data to sufficiently substantiate more complex definitions. As a consequence, these assumptions are often violated, leading to estimates of 'effective' instead of 'true' parameters (Reichert and Mieleitner, 2009). Thus, an assumption that model errors are absent or can be perfectly described, can lead to erroneous parameter estimates during calibration.

In the context of crop modelling, model errors could take the form of certain environmental dependencies which are missing in the model equations or simplifying assumptions such as different biological groups have similar growth and development, in spite of them exhibiting differences. For example, cultivars of the same species may differ in their growth and development. But they may exhibit some similarity within ripening/maturity groups to which they belong, as compared to across them. This grouping represents a hierarchical structure of cultivars nested within maturity groups of a particular species. In regional studies, however, parameter estimates for a crop species may be obtained by calibrating the model to data from different cultivars grown in contrasting environments within the target region. In other words, a common parameter set is estimated for all the data sets represented by distinct cultivar-environment combinations. By doing so, we attribute the variability in growth and development, which are due to inherent differences between cultivars and model deficits related to environmental dependencies, to random error or aleatoric uncertainty. Thus, the estimated model parameters compensate for these factors that are ignored or missing in the model during calibration. This results in a compromise solution wherein the parameters may no longer uphold their intended physiological meaning in the model. An unrealistic collapse of parameter uncertainty also ensues, since the observed variability is wrongly attributed to random error. Therefore, we arrive at wrong (effective) parameter values with high certainty, which in turn leads to unreliable model predictions.

A Bayesian multi-level modelling (BMM) approach is suitable for incorporating hierarchical structures, such as those between cultivars-ripening groups, during calibration. It can also be used to account for model errors such as those from missing environmental dependencies in the model equations (Zhang and Arhonditsis, 2009). Thus, BMM provides a method to better account for different sources of errors, rather than lumping them together. Analyses of the relationship between parameters in the BMM approach that account for these errors and environmental variables can help identify areas for model improvement. This approach is also expected to improve model calibration performance as compared to the commonly used approach of pooling all errors into a single term. The following research question is proposed to evaluate the BMM approach:

2. Can Bayesian multi-level modelling (BMM) be used to obtain reliable parameter estimates by accounting for inherent data structures and identifying model deficits?

When a common parameter set is estimated for different cultivars and environments, the model is required to fit all data sets represented by distinct cultivar-environment combinations simultaneously. This requirement is implicit in the method used to combine likelihood values from different data sets in classical Bayes (see section 2.1.1). Given the differences in data sets, this almost impossible task is only fulfilled through a compromise solution to the parameter inference problem. Thus, an alternative strategy of combining likelihoods is proposed, which relaxes the constraint of fitting all data sets simultaneously, to obtain a more representative estimate of parameter and prediction uncertainty. The following research question is framed to evaluate this alternative strategy.

3. Can an alternative strategy of combining likelihoods lead to reliable predictions?

1.4 Research objectives and scope

The following research objectives are defined so that these research questions can be addressed:

1. Evaluate model predictions upon implementing Bayesian sequential updating (BSU) to sequentially calibrate a phenology model to yearly data (Chapter 3)
2. Implement a Bayesian multi-level modelling (BMM) approach to calibrate a phenology model so that inherent data structures can be taken into account and model deficits related to environmental factors can be identified (Chapter 4)
3. Compare phenology predictions from a model calibrated using an alternate formulation of combined likelihood with those using the classical approach (Chapter 5)

In the first and third study, silage maize phenology observations made between 2010 and 2016 from two regions in southern Germany were used (Weber et al., 2022). In the second study, silage maize phenology data between 2010 and 2017 from across Germany was used (DWD Climate Data Center (CDC), 2019). The SPASS phenology model was used to simulate maize phenological development (Wang, 1997).

CHAPTER 2

Theory & Methods

The theoretical concepts used in the following chapters are explained by calibrating a simple two-parameter model to synthetic data. We start by recalling Bayes theorem, followed by a synthetic case study in which four Bayesian approaches are compared.

2.1 Bayesian inference

Bayesian inference can be used to express the parameters of a model in terms of probability distributions. According to Bayes theorem, the probability of a parameter having a certain value is conditional on the observed data. It is given by:

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)} \tag{2.1}$$

where $p(\theta|Y)$ is defined as the posterior probability of a parameter θ given the model and observations Y. It is proportional to the product of the prior probability of the parameters $p(\theta)$, and $p(Y|\theta)$ which is the probability of the observations given the model and parameter θ (commonly referred to as the *likelihood*). In case the model has multiple parameters, θ represents a parameter vector while $p(\theta)$ and $p(\theta|Y)$ are multivariate probability distributions with a dimension corresponding to each parameter. The denominator of the equation is defined as the prior predictive distribution ($p(Y) = \int p(\theta)p(Y|\theta)d\theta$). It can become mathematically intractable. Thus, sampling methods are used to estimate posterior distributions. This is the reason why there have historically been only limited applications of Bayesian inference. Computational advances have led to its revival in the past few decades. Bayesian theorem thus offers a method for updating our prior beliefs about the model parameters (from $P(\theta)$ to $P(\theta|Y)$), based on new observations. Bayes theorem can be extended to multiple models in an ensemble. However, in this dissertation the application has been limited to a single model.

2.1.1 Likelihood

An important term in Bayes theorem that deals with current observations is the likelihood. It expresses the probability of observations given the model and parameters (or likelihood of the parameters, given the observations). Consequently, it defines the distribution of the residuals (difference between the observations and simulations) around the simulated model output. Model errors are commonly assumed to be normally distributed, centred at zero, independent and identical, with the standard deviation being equal to that of the observation error (Reichert and Mieleitner, 2009). This assumption implies that the model is unbiased, the error at each observation is independent from the next, and that measurement/observation error is the only source of error. However, in models that simulate environmental processes, certain process-representations may be missing, resulting in epistemic uncertainty and error distributions that deviate from zero (Reichert and Schuwirth, 2012). Model errors may not be independent since they can be propagated through space and time, making errors at one point dependent on the previous (Reichert and Mieleitner, 2009). Furthermore, errors may not be identical (constant standard deviation) but could vary as a function of the simulated state variable (Schoups and Vrugt, 2010). Apart from errors in observations and model parameters, those in model inputs and boundary conditions contribute to the total model error. It is practically impossible to know all sources of errors and their properties, making it difficult to find a suitable expression for the likelihood function to represent this uncertainty. Thus, assumptions about model errors are usually simplistic and hence violated. Different approaches detailed in literature attempt to tackle this problem.

While there is indisputable benefit in improving process models (Wang et al., 2017a; Maiorano et al., 2017) for better system understanding and prediction capability, it is impossible to construct a perfect environmental model. Modellers have addressed this problem by defining more representative likelihood functions. Methods like Bayesian hierarchical models have been implemented, which can be used to disentangle different sources of errors that contribute to total predictive uncertainty and in-turn provide valuable insights into model limitations (Del Giudice et al., 2013). In the Bayesian total error analysis (BATEA) framework (Kavetski et al., 2006b,a), errors from forcings, response and model structure are explicitly characterized. Reichert and Mieleitner (2009) defined time-dependent parameters to account for structural uncertainty. Other studies (Weber et al. (2018) for example) have aimed at obtaining a correct statistical description of total uncertainty. Schoups and Vrugt (2010) proposed a formal generalized likelihood function to be used when residual errors are correlated, heteroscedastic, and non-Gaussian with varying degrees of kurtosis and skewness. Samadi et al. (2018) used a post-calibration error-modelling approach in which generalized additive models of location, scale and shape (GAMLSS) were applied to characterize the uncertainty. Xu and Valocchi (2015) implemented a data-driven error modelling approach in which a Gaussian process model was implemented in the Bayesian calibration framework to account for model structural error. Other methods like GLUE (Beven and Binley, 1992, 2014) provide options to avoid over-conditioning the model on observations through the use of likelihood measures, methods for likelihood combination, and exclusion of non-behavioural parameters from the posterior distribution. GLUE and its pseudo-Bayesian implementations (Mantovan and Todini, 2006; Beven et al., 2007, 2008; Vrugt et al., 2009) have been extensively applied in crop-modelling (Makowski et al., 2002; Mo and Beven, 2004; He et al., 2010; Pathak et al., 2012; Sexton et al., 2016; Gao et al., 2020). However, other methods, although well-developed in the field of hydrology, have only been applied to a limited extent in crop modelling (literature review provided in Wallach and Thorburn 2017).

In this dissertation I focus on two approaches: the first aids in identifying model deficits while providing a better representation of the likelihood; the second offers a pragmatic alternative strategy for calibration by relaxing the assumptions of the likelihood. In the following section, an example of the two approaches is provided, through a synthetic case study with a phenology model for visual depiction and comparison.

2.2 Synthetic case study

In this case study we calibrate the SPASS phenology model to observations of different maize cultivars. Synthetic observations were generated from the model such that they emulated differences in phenology between cultivars. This was done by assigning six different values to two parameters of the model (*toptv* and *toptr*), corresponding to cultivars A, B, C, D, E, and F. The remaining model parameters were fixed. The SPASS model was run using the same forcings and boundary conditions for all six cultivars. Simulated phenology at 60, 90, 120, and 150 days after sowing was recorded and a random observation error (normally distributed, centred at zero and with a standard deviation of 3 BBCH - phenology units) was added to obtain the synthetic observations.

Four Bayesian calibration approaches were used to estimate the two SPASS model parameters. The model was calibrated to observations from cultivars A, B, and C while the remaining cultivars D, E, and F were used to evaluate the model's prediction performance after calibration.

2.2.1 Calibration approaches

Let $y_{d,c}$ represent the observed phenology for cultivar $c \in \Omega$ where $\Omega = \{A, B, C\}$ at a particular day after sowing $d \in D$ where $D = \{60, 90, 120, 150\}$. Then, $\bar{y}_{d,c}(\theta, T, s)$ is the phenology on a given day d simulated by a deterministic model (SPASS in this case) as a function the model parameters θ, forcings such as temperature T and boundary conditions such as sowing date s. Let $Y = \{y_{d,c}; d \in D; c \in \Omega\}$ define the vector of phenology observations and $\bar{Y} = \{\bar{y}_{d,c}; d \in D; c \in \Omega\}$ define the corresponding vector of simulated phenology. On each day after sowing, phenology of a particular cultivar is simulated by the deterministic model such that

$$y_d = \bar{y}_d(\theta, T, s) + \epsilon_d \tag{2.2}$$

where ϵ_d is the total error or the difference between the observed and simulated value (van Oijen, 2017). We are uncertain about the error and assume that this uncertainty can be represented by a Gaussian distribution such that $\epsilon_d \sim \mathcal{N}(\mu = 0, \sigma^2)$ (i.e. centred at zero and a standard deviation σ).

Suppose a modeller is tasked with estimating the model parameters such that the calibrated model can be used for future predictions of maize phenology. Having prior knowledge about the parameters and wanting to quantify uncertainty in phenology predictions, the modeller decides to apply a Bayesian approach. For this purpose, the posterior distribution in Eq. 2.1 can be written as a proportionality by dropping the normalization constant in the denominator:

$$p(\theta \mid Y, T, s) \propto p(Y \mid \theta, T, s)\, p(\theta) \tag{2.3}$$

where $p(\theta)$ is the joint prior probability distribution of the model parameters. Based on the definition of ϵ_d in Eq. 2.2, the likelihood of the parameter given each observation can be written as

$$p(y_d \mid \theta, T, s) = \mathcal{N}(y_d - \bar{y}_d(\theta, T, s), \mu = 0, \sigma^2) \tag{2.4}$$

If we assume that the errors are uncorrelated, then the joint likelihood for all observations in a growing season is

$$p(Y \mid \theta, T, s) = \prod_{d \in D} \mathcal{N}(y_d - \bar{y}_d(\theta, T, s), \mu = 0, \sigma^2) \tag{2.5}$$

The model can be calibrated using four approaches, as described below in terms of their likelihoods.

a) Pooled

In order to simulate phenology for the maize species, the modeller may decide to implement a pooled approach by combining all data from the three cultivars together and estimating common model parameters θ_{sp} for the species. Under ideal conditions of a perfect model, the total residual error would only depend on the error in observing Y, which is known (standard deviation of measurement error is δ). In such cases the likelihood is given by:

$$p(Y \mid \theta, T, s) = \prod_{c \in \Omega} \prod_{d \in D} \mathcal{N}(y_{d,c} - \bar{y}_{d,c}(\theta_{sp}, T, s), \mu = 0, \sigma^2 = \delta^2) \tag{2.6}$$

where $\theta = \{\theta_{sp}\}$ is the estimated parameter vector.

However, the modeller may acknowledge that this assumption is unrealistic since neglecting between-cultivar differences may lead to errors, in addition to those from other error sources. The modeller chooses to estimate the total residual error as a lumped term (σ) without distinguishing between the different sources. In this case the total error constitutes that in measurements, model equations, forcings, boundary conditions, etc. In this case, the likelihood takes the same formulation but with an additional estimated parameter $\theta = \{\theta_{sp}, \sigma\}$.

b) Unpooled

On the other hand, the modeller may acknowledge that the different cultivars could vary in their biophysical parameters and calibrates the model separately to data from each of the three cultivars. This yields a cultivar-specific likelihood function expressed as:

$$p(Y_c \mid \theta, T, s) = \prod_{d \in D} \mathcal{N}(y_{d,c} - \bar{y}_{d,c}(\theta_c, T, s), \mu = 0, \sigma^2 = \delta^2) \tag{2.7}$$

where $\theta = \{\theta_c; c \in C\}$ is the estimated cultivar-specific parameter.

c) Hierarchical

As an intermediate between the pooled and unpooled cases, the modeller acknowledges that cultivars of the same maize species should share some similarities and thus incorporates this knowledge using a hierarchical expression for the parameters.

$$p(Y \mid \theta, T, s) = \prod_{c \in \Omega} \prod_{d \in D} \mathcal{N}(y_{d,c} - \bar{y}_{d,c}(\theta_{sp,c}, T, s), \mu = 0, \sigma^2) \tag{2.8}$$

where $\theta_{sp,c} = \theta_{sp} + \Delta\theta_c$ and the estimated parameter vector is $\theta = \{\theta_{sp}, \Delta\theta_c, \sigma\}$. Here, priors are specified for both the hyperparameters (θ_{sp}) and the deviation of the cultivar-specific parameters from the hyperparameters ($\Delta\theta_c$). In such an approach the cultivar-specific parameters $\theta_{sp,c}$ and the species parameters θ_{sp} can be simultaneously estimated.

d) Additive

The modeller may have an alternative perspective - since observations were made for different cultivars, not all data are equally informative for the inference problem (i.e. estimating parameters for the species). This can be accounted for by the method used to combine likelihoods. In the classical Bayesian approach, likelihoods are multiplied. By doing so we are trying to obtain a common parameter set that fits all data. By adding likelihoods, however, we identify parameters that fit any of the data-points. If probable parameter values common to all data exist in this case, they obtain a higher probability. Thus, likelihoods can be added as an alternative approach. As an extreme case, all likelihood values could be added in theory by replacing all multiplications in Eq. 2.6 by sums. However, the modeller should use system and model knowledge to define data-groups in which likelihoods within groups are multiplied and across groups are added. In this synthetic example, the modeller may make a reasonable assumption that a common parameter set should be obtained for a cultivar in one growing season. Thus, the cultivars are used to define separate data groups, such that the combined likelihood is given by:

$$p(Y \mid \theta, T, s) = \sum_{c\in\Omega} \prod_{d\in D} \mathcal{N}(y_{d,c} - \bar{y}_{d,c}(\theta_{sp}, T, s), \mu = 0, \sigma^2 = \delta^2) \tag{2.9}$$

where $\theta = \{\theta_{sp}\}$ is the estimated parameter vector. Note the difference between the Eq. 2.6 and Eq. 2.9 - the multiplication of likelihoods across cultivars has been replaced by addition. Further details of this approach are covered in Chapter 5.

2.2.2 Comparing the approaches

In these approaches, parameter estimates are obtained for the species and/or for the individual cultivars in the calibration data set (Table 2.1). In the *Pooled* case for example, parameters are only estimated for the species, while in the *Hierarchical* case, they are simultaneously estimated for the species as well as for the cultivars A, B, and C in the calibration data set.

Table 2.1: Posterior parameter distributions obtained from the different calibration approaches

	Pooled	**Hierarchical**	**Additive**	**Unpooled**
Species	yes	yes	yes	no
Cultivar	no	yes	no	yes

To assess how these approaches defer in their prediction capabilities, we analyse the posterior parameter distributions and visualize their overlap with the posterior distribution of the prediction target. The prediction targets are the cultivars which have been left out during the calibration exercise, namely D, E, and F. To obtain reference prediction target distributions, the model was calibrated to each of the individual cultivars (i.e. the unpooled approach). These distributions represent the best performance of the model, given the observed data set, in absence of any additional prior information. Target distributions for the calibration cultivars are also provided for reference. Larger the overlap between the posterior distribution from a particular calibration approach with that of the prediction target, better is that approach at prediction. In Fig. 2.1, species parameter estimates are compared with the prediction targets. In Fig. 2.2, cultivar-specific parameter estimates from the *Hierarchical* and *Unpooled* cases are provided. The grey boxes represent the parameter space formed by the two model parameters *toptr* and *toptv*. All sub-plots show the same extent of the parameter space. The posterior parameter distribution in each calibration case is shown by the coloured dot-plots of the posterior samples, where the colours indicate the posterior probability density. Contours indicate the ∼2 standard deviations of the target distributions. Markov Chain Monte Carlo (MCMC) sampling method was used for all cases except the *Additive* case in which brute-force Monte Carlo (MC) sampling was used. All parameters were assumed to have uniform prior distributions, except for $\Delta\theta_c$ in the *Hierarchical* case where a normal distribution was assumed.

Greater the overlap between the dot-plots and a prediction target contour, better will the calibrated model be at predicting that particular target cultivar. In the *Pooled* case without σ estimation (Fig. 2.1a, Eq. 2.6), the posterior distribution collapses and results in a negligible overlap with the targets, or even the cultivars used for calibration. The resultant parameters arrive at a compromise solution which will perform poorly in predicting many of the targets. This also reflects an underestimation of parameter uncertainty. In the *Pooled* case with σ estimation (Fig. 2.1b), wider parameter distributions are obtained that overlap with some targets. Conceptually, the variability between cultivars is attributed to random noise (σ), together with errors in measurements, inputs, boundary conditions, etc. While the wider parameter distributions and the estimated total error may be able to predict some targets, the predictions may not be robust due to the attribution of between-cultivar variability to aleatory instead of epistemic uncertainty. In the *Hierarchical* case (Eq. 2.8), the variability between cultivars is correctly attributed to epistemic uncertainty. This resulted in higher uncertainty in the species parameter estimates (Fig. 2.1c) that have a larger overlap with targets. Furthermore, this approach also provides parameter estimates for individual cultivars (Fig. 2.2a) used for calibration. Thus, phenology can not only be predicted in case of new cultivars or the species as a whole, but also when the same calibration cultivars are grown in the future. In the *Additive* case (Fig. 2.1b, Eq. 2.9), wider uncertainty in posterior parameter distributions is obtained which would at least be able to predict those cultivars that were used for calibration. This is apparent from the overlap with the calibration targets. The resultant cultivar-specific parameter estimates in the *Unpooled* case (Fig. 2.2b, Eq. 2.7) are useful in predicting phenology if the same cultivar is grown in the future. However, it would be a poor predictor of the species as a whole. In all cases, the posterior parameter distributions have little to no overlap with the contour for cultivar F. This is related to representativeness of the data: all methods would perform poorly if the prediction target is not well-represented in the calibration data set.

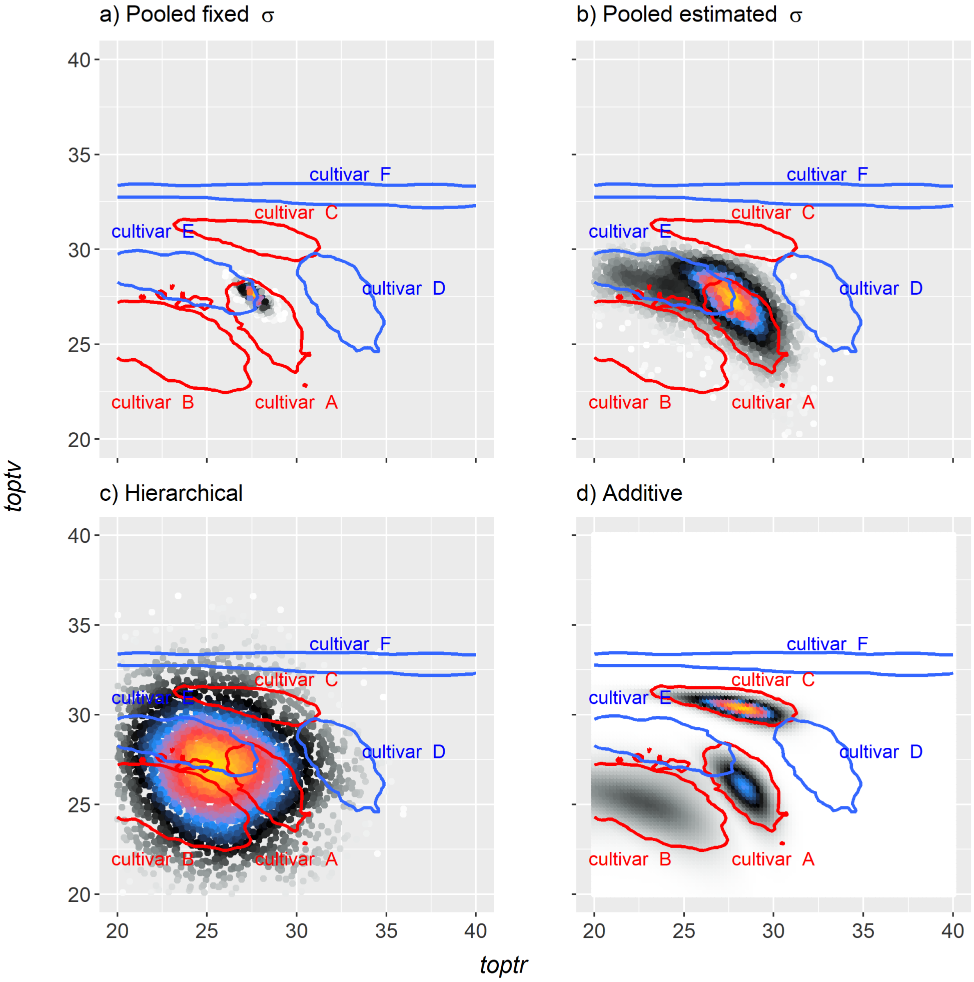


Figure 2.1: Comparison of the posterior parameter distributions for the species obtained from different Bayesian approaches. Posterior parameter distributions of *toptv* (y-axis) and *toptr* (x-axis) from the SPASS phenology model in the (a) Pooled, (b) Pooled with estimated σ, (c) Hierarchical, and (d) Additive calibration cases for the synthetic case study. The grey boxes represent the parameter space formed by the two model parameters. It shows the same extent of the parameter space in all sub-plots. The red and blue contours indicate ~ 2 SD of the posterior parameter distributions with red contours marking the cultivars that were used for calibration and blue for the prediction targets. The posterior parameter distributions are shown by the coloured dot-plot of the posterior samples, where the colours indicate sample density (posterior probability density) - black for lower and yellow for higher densities.

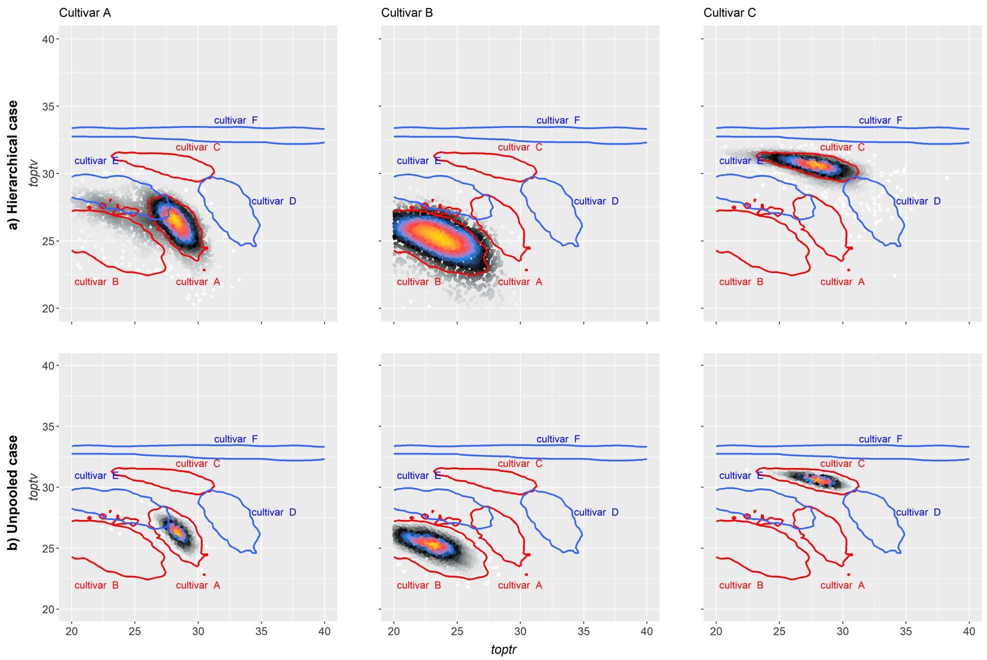


Figure 2.2: Comparison of the posterior parameter distributions for the three calibration cultivars (A, B, and C) obtained from a) Hierarchical and b) Unpooled approaches. The grey boxes represent the parameter space formed by the two SPASS phenology model parameters *toptv* (y-axis) and *toptr* (x-axis). It shows the same extent of the parameter space in all sub-plots. The red and blue contours indicate ∼ 2 SD of the posterior parameter distributions with red contours marking the cultivars that were used for calibration and blue for the prediction targets. The posterior parameter distributions are shown by the coloured dot-plot of the posterior samples, where the colours indicate sample density (posterior probability density) - black for lower and yellow for higher densities.

In the synthetic case study, the only source of model error arises from not accounting for between-cultivar variability. However, for models used with real data sets, the models may also be deficient in some process representation such as dependencies on key environmental variables. This deficit may not be known beforehand. These different errors may interact, leading to complex structures. In the following chapters, the approaches described above will be applied to real-world problems to further investigate their performance.

CHAPTER 3

A Bayesian sequential updating approach to predict phenology of silage maize

Authors: Michelle Viswanathan, Tobias K. D. Weber, Sebastian Gayler, Juliane Mai, Thilo Streck

This chapter is taken from the publication “Viswanathan et al. (2022), A Bayesian sequential updating approach to predict phenology of silage maize, Biogeosciences, 19, 2187–2209, DOI: 10.5194/bg-19-2187-2022”, with minor modifications.

Abstract

Crop models are tools used for predicting year-to-year crop development on field to regional scales. However, robust predictions are hampered by uncertainty in crop model parameters and in the data used for calibration. Bayesian calibration allows for the estimation of model parameters and quantification of uncertainties, with the consideration of prior information. In this study, we used a Bayesian sequential updating (BSU) approach to progressively incorporate additional data at a yearly time-step in order to calibrate a phenology model (SPASS) while analysing changes in parameter uncertainty and prediction quality. We used field measurements of silage maize grown between 2010 and 2016 in the regions of Kraichgau and the Swabian Alb in southwestern Germany. Parameter uncertainty and model prediction errors were expected to progressively be reduced to a final, irreducible value. Parameter uncertainty was reduced as expected with the sequential updates. For two sequences using synthetic data, one in which the model was able to accurately simulate the observations, and the other in which a single cultivar was grown under the same environmental conditions, prediction error was mostly reduced. However, in the true sequences that followed the actual chronological order of cultivation by the farmers in the two regions, prediction error increased when the calibration data were not representative of the validation data. This could be explained by differences in ripening group and temperature conditions during vegetative growth. With implications for manual and automatic data streams and model updating, our study highlights that the success of Bayesian methods for predictions depends on a comprehensive understanding of the inherent structure in the observation data and of the model limitations.

3.1 Introduction

The effects of climate change are already being felt, with increasing global temperature and frequency of extreme events (Porter et al., 2015), which will have an impact on food availability. In order to mitigate risks to food security, suitable adaptation strategies need to be devised which depend on robust model predictions of the productivity of cropping systems (Asseng et al., 2009). Soil–crop models, which are able to predict changes in crop growth and yield as a consequence of changes in model inputs such as weather, soil properties, and cultivar-specific traits, are considered suitable tools to plan for a secure future. However, achieving robust model predictions is challenging. This is because there is uncertainty in the model inputs, parameters, and process representation, as well as in the observations used to calibrate these models (Wallach and Thorburn, 2017). It is therefore essential to quantify these uncertainties.

Different interpretations of the underlying soil–crop processes have led to different representations in models of varying complexity (Wallach et al., 2016). Process model equations have parameters that represent physiological processes, but are often based on empirical relationships. These relationships describe system processes which cannot be further resolved with reasonable effort. While some parameters that represent physiological aspects of plant growth and development can be determined in dedicated experiments (Craufurd et al., 2013), many others still need to be estimated through model calibration. However, the measured parameters and state variables used for model calibration are uncertain due to errors in the measuring device or technique and due to the natural variability of the system owing to processes occurring at different spatial or temporal scales. Given the different sources of uncertainty, it is important to set up adequate workflows to enable uncertainty quantification and protocols for reporting them, especially when they influence decision-making (Rötter et al., 2011).

For this, the Bayesian approach is an elegant framework to propagate uncertainty from measurements, parameters, and models to prediction. One advantage of Bayesian inference is the use of prior information (Sexton et al., 2016). The posterior probability distribution obtained by conditioning on one dataset can then be used as a prior distribution for the next dataset in a sequential manner (Hue et al., 2008). This approach, called "Bayesian sequential updating" (BSU), would be more computationally efficient than having to re-calibrate the model to all previous datasets every time new data are available. It has been applied to big data studies in which large datasets were split to reduce computational demand and the information was sequentially incorporated (Oravecz et al., 2016). Cao et al. (2016) used BSU to analyse the evolution of the posterior parameter distribution for soil properties by incorporating data from different types of experiments. Thompson et al. (2019) applied this approach to estimate species extinction probabilities where species-siting data were sequential in time. While there are numerous examples of Bayesian methods being applied in crop modelling for uncertainty quantification and data assimilation (Alderman and Stanfill, 2017; Ceglar et al., 2011; Huang et al., 2017; Iizumi et al., 2009; Makowski, 2017; Makowski et al., 2004; Wallach et al., 2012; Wöhling et al., 2013, 2015), to the best of our knowledge, the BSU method has not been evaluated in the field of crop modelling to date.

In this study we assessed whether crop model predictions progressively improve as new information is incorporated using the BSU approach. This ascertains whether the model and parameters are both temporally and spatially transferable for a particular crop species, an important aspect for large-scale and long-term predictions. Our study focused on modelling crop phenological development.

Plant phenology is concerned with the timing of plant developmental stages such as emergence, growth, flowering, fructification, and senescence. It is controlled by environmental factors such as solar radiation, temperature, and water availability, and depends on intrinsic characteristics of the plants (Zhao et al., 2013). Phenological development is a crucial state variable in soil–crop models, since it controls many other simulated state variables such as yield, biomass, and leaf area index by influencing the timing of organ appearance and assimilate-partitioning. Phenology is not only species-specific but can also differ between cultivars of the same species (Ingwersen et al., 2018). Model parameters that influence phenology could vary depending on the cultivars (Gao et al., 2020) and possibly also on environmental conditions (Ceglar et al., 2011). Since parameter uncertainty is a major source of prediction uncertainty (Alderman and Stanfill, 2017; Gao et al., 2020), it impacts prediction quality.

To this end, we assessed the impact of sequentially incorporating new observations with the BSU approach on the prediction quality of phenological development. For this, we modelled phenological development of silage maize grown between 2010 and 2016 in Kraichgau and the Swabian Alb, two regions in southwestern Germany with different soil types and climatic conditions. We monitored the changes in parameter uncertainty and evaluated prediction quality by performing model validation in which simulated phenological development was compared with observations for datasets that were not used for calibration. We hypothesized that:

1. Parameter uncertainty decreases and quality of prediction improves with the sequential updates in which increasing amounts of data are used for model calibration.
2. For the first few sequential updates, the quality of prediction is variable, until the calibration samples become representative of the population.
3. The prediction error then progressively drops to an irreducible value that represents the error in inputs, measurements, model structure, and variability due to spatial heterogeneity that is below model resolution.

We tested these hypotheses by applying BSU in two modelling cases that represent ideal and real-world conditions. In the first case, we applied BSU to two *synthetic sequences*: an *ideal* sequence of observations wherein the model is able to simulate the observations accurately, and a *controlled cultivar–environment* sequence of observations which represent different growing seasons of a single cultivar grown under the same environmental conditions. In the second case, we applied the BSU to two *true sequences* that follow the actual chronological order in which different cultivars of silage maize were grown in the two regions under different environmental conditions.

With this study, we explicitly deal with a well-known problem in regional modelling, which carries particular weight in the case of maize. On a regional scale, maize cultivars may differ considerably in their phenological development, but cultivar information will rarely be available. Even if data on cultivars grown were available, phenological data on all relevant cultivars in a particular region will rarely be at hand. Consequently, model parameters are typically estimated for the crop species and not for the individual cultivars. Also, the maize cultivars of our study represent only a small subset of cultivars grown in Kraichgau and the Swabian Alb. We therefore grouped the maize cultivars into ripening groups for analysis of prediction quality.

3.2 Materials and methods

3.2.1 Study sites and measured data

The data used for the study consist of a set of measurements taken at three field sites (site 1, site 2, site 3) in Kraichgau and two field sites (site 5 and site 6) on the Swabian Alb, in southwestern Germany, between 2010 and 2016 (Fig. 3.1i) (Weber et al., 2022). The main crops in rotation were winter wheat, silage maize, winter rapeseed, and cover crops such as mustard and phacelia. Additionally, spelt and spring and winter barley were also grown on the Swabian Alb. Amongst others, continuous measurements of meteorological conditions, soil temperature, and moisture were taken. Soil profiles were sampled at the sites for characterization of soil properties.

Kraichgau and the Swabian Alb represent climatologically contrasting regions in Germany. Kraichgau is situated 100–400 m above sea level (a.s.l.) and characterized by a mild climate with a mean temperature above 9° and mean annual precipitation of 720–830 mm. It is one of the warmest regions in Germany. The Swabian Alb is located at 700–1000 m a.s.l. with a mean temperature of 6–7° and mean annual precipitation of 800–1000 mm. Kraichgau soils have often developed from several metres of Holocene loess, underlain by limestones. They are predominantly Regosols and Luvisols. The Swabian Alb has a karst landscape with clayey loam soils, often classified as Leptosols. Soils may be less than 0.3 m thick in some areas. While the soils at the sites in Kraichgau are similar, they vary across the sites on the Swabian Alb (Wizemann et al., 2015).

At every study site, which had an area of approx. 15 ha, replicate observations were made by assessing phenological development stages from maize plants in five subplots of 2 m × 2 m each. Ten maize plants were chosen from each subplot. We used the BBCH growth stage code (Meier, 1997) to define the development stages. The BBCH value of 10 marks the emergence and the start of leaf development, 30 stands for stem elongation, 50 for inflorescence, emergence or heading, 60 for flowering or anthesis, 70 for development of fruit, 80 for ripening, and 90 for senescence (Fig. 1ii). In the following sections, the individual growing seasons for silage maize are denoted by the site and year of growth, i.e. the site-year (Table 3.1). For example, silage maize grown at site 2 in Kraichgau in the year 2012 is referred to as "2_2012". The different cultivars used in the study can be grouped into three ripening or maturity groups, based on their timing of ripening. Mid-early (ME) and late (L) ripening cultivars were grown in Kraichgau, and early (E) and mid-early (ME) ripening cultivars were grown on the Swabian Alb.

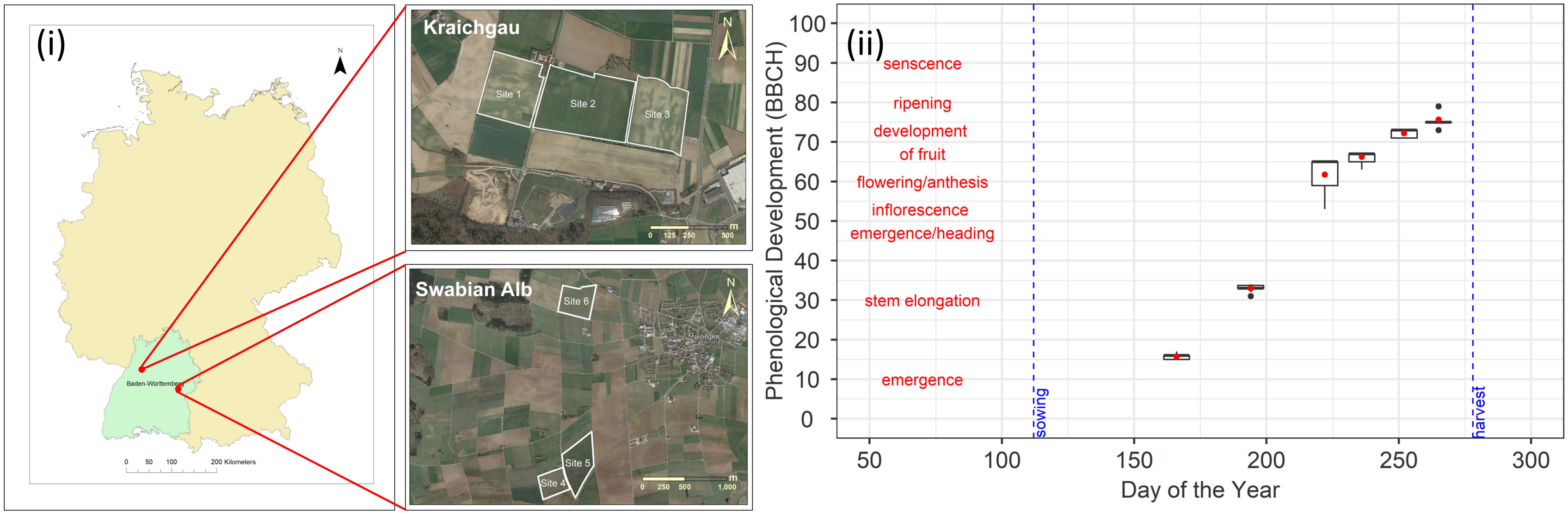


Figure 3.1: (i) Location of the sites in Kraichgau (site 1, site 2, and site 3) and the Swabian Alb (site 5 and site 6) in the state of Baden-Wuerttemberg, Germany (© Google Earth 2018 modified from Eshonkulov et al. (2019). (ii) Observations of phenological development (expressed in BBCH growth stages) of silage maize at site 6 are plotted against the day of the year in 2010. The red labels indicate important phenological development stages. The red points are means of the observations while the box and whiskers represent the range of replicate observations. The length of the box represents the inter-quartile range (IQR), whiskers extend from the box up to 1.5× IQR and values beyond this range are plotted as points. Each of the boxes and whiskers are based on 50 points corresponding to observations made on the same day, i.e. 10 maize plants at five subplots within site 6 for 1 d in 2010. In site-year 6_2010, observations were made on 6 d during the growing season.

Table 3.1: Early (E), mid-early (ME), and late (L) ripening cultivars of silage maize, with their sowing and harvest dates, grown at the study sites in Kraichgau (sites 1, 2, and 3) and the Swabian Alb (sites 5 and 6) between 2010 and 2016.

Region	**Year**	**Site**	**Site-year**	**Cultivar**	**Maturity/ ripening group**	**Sowing date (dd/mm/yyyy)**	**Harvest date (dd/mm/yyyy)**
Kraichgau	2011	3	3_2011	Canavaro	L	18/04/2011	03/10/2011
Kraichgau	2012	2	2_2012	Canavaro	L	02/05/2012	19/09/2012
Kraichgau	2014	1	1_2014	Grosso	ME	12/04/2014	09/10/2014
Kraichgau	2014	2	2_2014	Grosso	ME	11/04/2014	08/10/2014
Swabian	2010	6	6_2010	Fernandez PR 39 A 98	ME	23/04/2010	06/10/2010
Swabian	2011	5	5_2011	Agro-Yoko	ME	25/04/2011	04/10/2011
Swabian	2012	5	5_2012	Amanatidis	E	28/04/2012	07/10/2012
Swabian	2013	6	6_2013	SY Kairo & Agro Yoko	ME	26/04/2013	04/10/2013
Swabian	2015	5	5_2015	LG 30.217	E	22/04/2015	14/09/2015
Swabian	2016	5	5_2016	LG 30.217	E	07/05/2016	27/09/2016
Swabian	2016	6	6_2016	Toninio	ME	03/05/2016	23/09/2016

3.2.2 Soil–crop model

To simulate the soil–crop system, we used the SPASS crop growth model (Wang, 1997). SPASS is implemented in the Expert-N 5.0 (XN5) software package (Heinlein et al., 2017; Klein et al., 2017; Priesack, 2006). In XN5, the SPASS crop model is coupled with the Richards equation for soil–water movement as implemented in the Hydrus-1D model (Šimůnek et al., 1998). The routine uses van Genuchten–Mualem hydraulic functions (van Genuchten, 1980; Mualem, 1976) and the heat transfer scheme from the Daisy model (Hansen et al., 1990). In the SPASS model, germination to emergence (up to BBCH 10), the vegetative phase (between BBCH 10 and 60), and the generative or reproductive phase (BBCH 61 onwards) of the crop are modelled. Temperature and photoperiod are the two main factors affecting the phenological development rate (for details, refer to Appendix A: SPASS phenology model).

Daily weather data consisting of maximum and minimum temperatures were used in XN5 to calculate the air temperatures within the crop canopy. Soil properties (texture class, grain size, rock fraction, bulk density, porosity), as well as van Genuchten parameters and hydraulic properties (soil water content at wilting point, field capacity, residual and saturated water content, and saturated hydraulic conductivity), were based on soil samples taken at the sites in 2008 to characterize the soil profile. The soil horizons in the model were based on these soil profile descriptions. Initial values of soil volumetric water content were based on measurements. The simulations for each site-year were started on the harvest date of the preceding crop in the crop rotation at that site. This ensured adequate spin-up time prior to the simulation of silage maize, which was sown in April and May.

3.2.3 Selection of model parameters

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 3.2. A global sensitivity analysis using the Morris method (Morris, 1991) was then carried out to identify the sensitive parameters to be estimated through Bayesian calibration (Supplement S1). The sensitive parameters identified for calibration were: effective sowing depth (SOWDEPTH), which influences the emergence rate, and parameters affecting development in the vegetative phase (PDD1, TMINDEV1, DELTOPT1, and DELTMAX1). Parameter DELTOPT2, from the temperature response function during the reproductive phase, was estimated during calibration even though it was less sensitive. The choice of using this parameter during calibration was based on knowledge of model behaviour, so as to reduce the calibration error in the reproductive phase (Lamboni et al., 2009). Thus, out of 11 pre-selected parameters (Table 3.2), six were estimated in BSU, while the remaining parameters were fixed at their default values.

Table 3.2: SPASS model parameters for phenological development. The default values and two standard deviations (± 2 SD) were based on expert knowledge. “Status in calibration” indicates the parameters which were estimated or fixed to the default value during Bayesian calibration. Minimum (min) and maximum (max) values were set for estimated parameters to constrain the prior parameter ranges to reasonable values during calibration.

Parameter name	**Description**	**Unit**	**Default value**	**−2 SD**	**+2 SD**	**min**	**max**	**Status in calibration**
PDD1	Physiological development days from emergence to anthesis	d	45	32	60			Estimated
PDD2	Physiological development days from anthesis to maturity	d	36	25	60			Fixed
PDL	Photoperiod sensitivity factor	–	0	0	0.1			Fixed
DLOPT	Optimal photoperiod length	h	12	10	15			Fixed
TMINDEV1	Minimum temperature of vegetative development	°C	6	5	8	0	10	Estimated
DELTOPT1	Difference between optimum and minimum temperatures of vegetative development	°C	28	22	31	1	35	Estimated
DELTMAX1	Difference between maximum and optimum temperatures of vegetative development	°C	10	4	14	1	16	Estimated
TMINDEV2	Minimum temperature of reproductive development	°C	8	6	10			Fixed
DELTOPT2	Difference between optimum and minimum temperatures of reproductive development	°C	26	17	32	1	35	Estimated
DELTMAX2	Difference between maximum and optimum temperatures of reproductive development	°C	10	4	14			Fixed
SOWDEPTH	Effective sowing depth of the seeds in the soil	cm	8	5	15	1	20	Estimated

3.2.4 Bayesian sequential updating

In the BSU approach, Bayesian calibration is applied in a sequential manner. New data are used to re-calibrate the model, conditional on the prior information from previously gathered data. We describe the details of this approach here.

Bayes theorem states that the posterior probability of parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{Y}$, $P\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)$, is proportional to the product of the joint prior probability of the parameters $P\left(\boldsymbol{\theta}\right)$ and the probability of generating the observed data with the model, given the parameters $P(\boldsymbol{Y}|\boldsymbol{\theta})$. The term $P(\boldsymbol{Y}|\boldsymbol{\theta})$ is referred to as the likelihood function and is defined as the likelihood that observation $\boldsymbol{Y}$, that is observed phenological development in this study, is generated by the model using the parameter vector $\boldsymbol{\theta}$. The posterior probability distribution is obtained by normalizing this product by the prior predictive distribution (Gelman et al., 2013) or Bayesian model evidence (Schöniger et al., 2015) $P\left(\boldsymbol{Y}\right)$, which is obtained by integrating the product over the entire parameter space.

Hence, we write:

$$P\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)=\frac{P\left(\boldsymbol{\theta}\right)P(\boldsymbol{Y}|\boldsymbol{\theta})}{P(\boldsymbol{Y})}, \tag{3.1}$$

where

$$P\left(\boldsymbol{Y}\right)=\int_{\boldsymbol{\theta}}P\left(\boldsymbol{\theta}\right)P\left(\boldsymbol{Y}|\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}. \tag{3.2}$$

Equation (3.2) can become intractable, especially with a large number of parameters as this involves integrating over high-dimensional space (Schöniger et al., 2015). Instead, sampling methods such as Markov chain Monte Carlo (MCMC) are used to estimate the posterior distribution.

For one site-year sy_1 and corresponding observation vector $\boldsymbol{Y}_{\mathrm{sy}_1}$, the posterior parameter probability distribution is

$$P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_1}\right)=\frac{P\left(\boldsymbol{\theta}\right)P\left(\boldsymbol{Y}_{\mathrm{sy}_1}|\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}}P\left(\boldsymbol{\theta}\right)\,P\left(\boldsymbol{Y}_{\mathrm{sy}_1}|\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}}, \tag{3.3}$$

where $P\left(\boldsymbol{\theta}\right)$ represents the initial prior probability distribution that could be based on expert knowledge. The posterior parameter distribution $P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_1}\right)$ can now be used as a prior distribution for the next site-year sy_2. Thus, for site-year sy_n with an observation vector $\boldsymbol{Y}_{\mathrm{sy}_n}$, the posterior parameter probability distribution is

$$P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_n}\right)=\frac{P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_{(n-1)}}\right)\,P\left(\boldsymbol{Y}_{\mathrm{sy}_n}|\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}}P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_{(n-1)}}\right)\,P\left(\boldsymbol{Y}_{\mathrm{sy}_n}|\boldsymbol{\theta}\right)\,\mathrm{d}\boldsymbol{\theta}}. \tag{3.4}$$

This equation defines the BSU approach in which the model is calibrated in a sequential manner. New data from a site-year ($\boldsymbol{Y}_{\mathrm{sy}_n}$) are used to re-calibrate the model, conditional on the prior information from previous site-years. The posterior distribution obtained from the previous Bayesian calibration $P(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_{(n-1)}})$ is used as prior probability for calibration to the next site-year.

With the aim of making the computations tractable, we deviate slightly from this pure BSU approach as we do not strictly use the posterior distribution from the previous site-year as the prior distribution for the next one, but sequentially calibrate the model to data from an increasing number of site-years instead. The reason for this deviation is that in applying BSU, where the posterior parameter distribution is estimated by sampling methods, a probability density function needs to be approximated from the sample, so that it can be used as a prior probability for the subsequent site-year. This approximation introduces additional errors. Since joint inference is known to be better than sequential inference using posterior approximations (Thijssen and Wessels, 2020), Eq. (3.4) can be re-written, under the assumption that the phenology observations from all site-years are independent and identically distributed (Gelman et al., 2013), as follows:

$$P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_n}\right)=\frac{P\left(\boldsymbol{\theta}\right)\prod_{x=\mathrm{sy}_1}^{\mathrm{sy}_n}P\left(\boldsymbol{Y}_x|\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}}P\left(\boldsymbol{\theta}\right)\prod_{x=\mathrm{sy}_1}^{\mathrm{sy}_n}P\left(\boldsymbol{Y}_x|\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}}. \tag{3.5}$$

Thus, we use Eq. (3.5) to sequentially update the probability distribution of parameters by increasing the dataset size at each step through the addition of one site-year worth of new data $\boldsymbol{Y}_x$ to the previous dataset $\boldsymbol{Y}_{x-1}$.

After each inferential step, the probability of observing a certain phenology at the next site-year sy_{n+1} is predicted by

$$P\left(\boldsymbol{Y}_{\mathrm{sy}_{n+1}}|\boldsymbol{Y}_{\mathrm{sy}_n}\right)=\int P\left(\boldsymbol{Y}_{\mathrm{sy}_{n+1}}|\boldsymbol{\theta}\right)P\left(\boldsymbol{\theta}|\boldsymbol{Y}_{\mathrm{sy}_n}\right)\mathrm{d}\boldsymbol{\theta}, \tag{3.6}$$

where $P\left(\boldsymbol{Y}_{\mathrm{sy}_{n+1}}|\boldsymbol{Y}_{\mathrm{sy}_n}\right)$ is the posterior predictive distribution (Gelman et al., 2013). 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. All calculations and the BSU were carried out using the R programming language (R Core Team, 2020)(R Core Team, 2020).

In the following sections, we describe the components of Bayes formula in detail.

3.2.4.1 Likelihood function

Let $\boldsymbol{\theta}=(\varphi_1\varphi_2\varphi_3,\ldots\varphi_j)$ represent a vector of the model parameters to be estimated in this study (Table 2). Suppose $\boldsymbol{Y}=(\overline{y}_1,\overline{y}_2,\overline{y}_3,\ldots\overline{y}_d)$ is a vector of the means of observed phenological development on different days during the growing season for a particular site-year. The mean observation $\overline{y}_d$ on day d for the site-year is given by

$$\overline{y}_d=\frac{1}{P}\frac{1}{R}\sum_{p=1}^{P}\sum_{r=1}^{R}y_{r,p,d}, \tag{3.7}$$

where $y_{r,p,d}$ represents the rth replicate of observed phenological development, measured at subplot p on day d for a particular site-year, R is the total number of replicates at subplot p, and P is the total number of subplots per field.

If we assume that all replicates R in all subplots P are independent, the standard deviation of the replicate observations on day d is $\sigma_{r,p,d}=\sqrt{\sum_{p=1}^{P}\sum_{r=1}^{R}\left(y_{r,p,d}-\overline{y}_d\right)^2/(P\times R)}$. This is one source of observation error that represents the spatial variability at the study site which is below the spatial resolution of the model. We also assume an additional source of error in identification of the correct phenological stage and its exact timing of occurrence. We assume that this error is within a standard deviation of 2 BBCH ($\sigma_{\mathrm{ident},d}=2$ for each observation day d). This assumption was made because 2 is the most common difference between development stages in the phenological development of maize on the BBCH scale. Assuming that the error from replicate observations ($\sigma_{r,p,d}$) and the error in the identification of phenological stages are additive, the total observation error is $\sigma_d^2 = \left(\sigma_{r,p,d} + \sigma_{\mathrm{ident},d}\right)^2$.

The model residual $\overline{y}_d - f(\boldsymbol{\theta})_d$ is the difference between the observed $\overline{y}_d$ and the model simulated $f(\boldsymbol{\theta})_d$ phenological stage and is represented by the likelihood function. Assuming normally distributed residuals, it is given by

$$P\left(\overline{y}_d|\boldsymbol{\theta}\right) = \frac{1}{\sigma_d\sqrt{2\pi}}\, e^{-0.5\left(\frac{\overline{y}_d - f(\boldsymbol{\theta})_d}{\sigma_d}\right)^2}. \tag{3.8}$$

The likelihood values for all the observations are combined by taking the product of the likelihoods per day of observation, under the assumption of independent and identically distributed model residuals. Thus, the joint likelihood function is given by

$$P\left(\boldsymbol{Y}_x|\boldsymbol{\theta}\right) = \prod_{d=1}^{D} P\left(\overline{y}_d|\boldsymbol{\theta}\right), \tag{3.9}$$

where $\boldsymbol{Y}_x$ is the observation vector for site-year x.

3.2.4.2 Prior probability distribution

As prior information, we used a weakly informative probability distribution function (pdf) to ensure that the posterior parameter distributions are mainly determined by the data that are sequentially incorporated. For this, we used a platykurtic prior probability distribution that is a combination of a uniform and a normal distribution (Fig. 3.D1) of the form:

$$P\left(\varphi_j\right) = \left\{\begin{array}{ll} \frac{1}{c}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(\varphi_j-\mu\right)^2}{2\sigma^2}} & \text{for } a \ \leq \varphi_j < \mu - 2\sigma \\ \frac{1}{c}\frac{1}{\sigma\sqrt{2\pi}}e^{-2} & \text{for } \mu - 2\sigma \ \leq \varphi_j \leq \mu + 2\sigma \\ \frac{1}{c}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(\varphi_j-\mu\right)^2}{2\sigma^2}} & \text{for } \mu + 2\sigma \ < \varphi_j \leq b \end{array}\right). \tag{3.10}$$

where φ_j is a model parameter in the parameter vector $\boldsymbol{\theta}$, a and b are the minimum and maximum limit for the parameter, respectively, μ is the mean (default value in Table 2), and σ is the standard deviation. The normalization constant c is used to ensure that the area under the curve equals unity as required for probability density functions.

$$\begin{aligned} c = -\mathrm{erf}\left(\sqrt{2}\right) + \frac{4}{\sqrt{2\pi}}e^{-2} - \frac{1}{2}\mathrm{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) \\ + \frac{1}{2}\mathrm{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right). \end{aligned} \tag{3.11}$$

The joint prior pdf was calculated by $P\left(\boldsymbol{\theta}\right) = \prod_{j=1}^{J} P(\varphi_j)$ and the model parameters were assumed to be uncorrelated. The parameters a, b, σ, μ, of $P(\varphi_j)$ were based on expert knowledge (Table 2).

3.2.4.3 Posterior probability distribution

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 % (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 and Gelman, 1998; Gelman and Rubin, 1992). The total number of samples of the posterior distribution in each calibration case was dependent on the Gelman–Rubin diagnostic being ≤ 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.

For model validation, the posterior predictive distribution was used to simulate phenological development and compare it with observations at site-years that were not included in the calibration sequence.

3.2.5 Performance metrics

Bias and normalized root mean square error (NRMSE), as defined in Eqs. (3.12) and (3.13), for site-year sy were calculated to assess the calibration and prediction performance.

$$\mathrm{Bias}_{\mathrm{sy}}=\frac{1}{D}\sum_{d=1}^{D}\left(\overline{y}_d-f\left(\boldsymbol{\theta}_i\right)_d\right) \tag{3.12}$$

$$\mathrm{NRMSE}_{\mathrm{sy}}=\sqrt{\frac{1}{D}\sum_{d=1}^{D}\frac{\left(\overline{y}_d-f\left(\boldsymbol{\theta}_i\right)_d\right)^2}{\sigma_d^2}} \tag{3.13}$$

Here, $\boldsymbol{\theta}_i$ is the ith parameter vector, D is the total number of observation days for the particular site-year, $f(\boldsymbol{\theta}_i)_d$ is the simulated phenological development, $\overline{y}_d$ is the mean observed phenological development, and σ_d is the standard deviation of the observations (as defined in section 3.2.4.1) on day d. Under the assumption of normally distributed error, the natural logarithm of the likelihood probability is inversely proportional to the normalized mean square error: $\ln\left(P\left(\boldsymbol{Y}_{\mathrm{sy}}\mid\boldsymbol{\theta}_i\right)\right)\propto -\mathrm{NRMSE}_{\mathrm{sy}}{}^2$. The normalized bias $\mathrm{NBias}_{\mathrm{sy}}=\frac{1}{D}\sum_{d=1}^{D}\frac{\overline{y}_d-f(\boldsymbol{\theta}_i)_d}{\sigma_d}$ is also reported in some plots.

The prediction quality is good when NRMSE is low and bias is zero. Prediction performance is classified as good, moderate, or poor depending on the median NRMSE of the predictions for a site-year. We use the following categories: good performance for median NRMSE ≤ 1, moderate for $1 <$ median NRMSE ≤ 2, poor for $2 <$ median NRMSE ≤ 3 and very poor for median NRMSE > 3.

We estimated the information entropy of the posterior parameter distributions after each sequential update using the redistribution estimate equation (Beirlant et al., 1997) (Supplement S2). A change in entropy with sequential updates indicates a change in uncertainty of the parameters, where higher information entropy indicates greater uncertainty in the posterior parameters. In line with our hypotheses, we expect the entropy to decrease with sequential updates.

3.2.6 Modelling cases

The BSU approach described in the previous sections and the subsequent analysis using the performance metrics were applied to two *synthetic sequences* and two *true sequences* of site-years. The synthetic sequences were used to demonstrate the application of the BSU approach in ideal conditions, while the true sequences were used to extend the application to real-world conditions. Figure 3.2 shows the four sequences and the site-years used for calibration and validation.

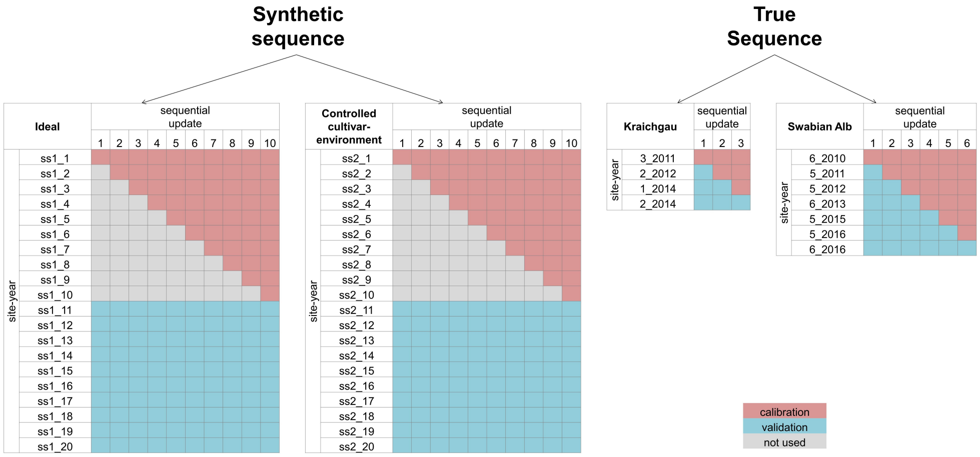


Figure 3.2: The site-years used for calibration and validation in each sequential update for the two synthetic sequences, namely ideal and controlled cultivar–environment, and the two true sequences for Kraichgau and the Swabian Alb are shown. In the synthetic sequences, a total of 10 updates were performed by sequentially adding 1 through 10 site-years to the calibration dataset. After each update, prediction quality was analysed for a set of 10 validation site-years. A total of three sequential updates in Kraichgau and six sequential updates in the Swabian Alb true sequences were analysed. In the sequential updates for the true sequences, a site-year was included for calibration, following the actual chronological order of growth. The remaining site-years grown in the region were then used for validation.

3.2.6.1 Synthetic sequences

We set up two synthetic sequences, namely *ideal* and *controlled cultivar–environment*. In each synthetic sequence, we used 10 sequential updates wherein 1 through 10 site-years were used in calibration. After each sequential update, the calibrated model was validated against a different set of 10 synthetic site-years (Fig. 3.2). Note here that the 10 site-years used for validation were the same across the sequential updates. Data from the 10 site-years used for calibration and the 10 site-years used for validation for the two synthetic sequences are shown in Fig. 3.3. Site-year 6_2010 was used to generate data for the synthetic sequences, as described here.

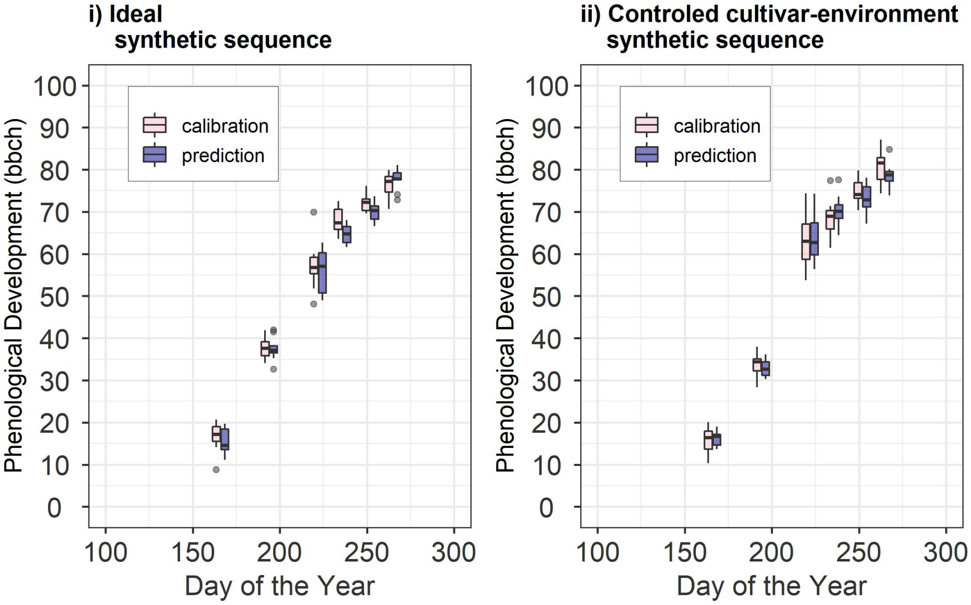


Figure 3.3: Synthetic site-year observations used for calibration and prediction in (i) the ideal and (ii) controlled cultivar–environment synthetic sequences. The pink boxes and whiskers represent the range of values for the 10 synthetic site-years used for calibration while the blue boxes and whiskers represent the range of values for the 10 site-years used for validation. The length of the box represents the inter-quartile range (IQR), whiskers extend from the box up to 1.5× IQR and values beyond this range are plotted as points.

The ideal sequence represents a case in which the model is able to accurately simulate the observations. The only sources of difference between site-years are from the spatial variability at the field site which is below model resolution and from the incorrect identification of phenological stages during field observations. To generate the ideal sequence of site-years, we first calibrated the model to phenology at 6_2010. The parameter set $\boldsymbol{\theta}_{\mathrm{MAP}}$ corresponding to the maximum a posteriori probability (MAP) estimate was used to simulate phenology and generate the synthetic dataset. To introduce inter-site-year differences, noise was added to simulated phenology $f(\boldsymbol{\theta}_{\mathrm{MAP}})_d$ on observation day d, where the noise was equal to the total observation uncertainty σ_d on that day for site-year 6_2010. Thus, for each synthetic site-year on observation day d, the phenological development was sampled from the range of total observation uncertainty σ_d at 6_2010, around simulated phenology $f(\boldsymbol{\theta}_{\mathrm{MAP}})_d$. The synthetic observations were generated for the same observation days as the actual observations at 6_2010. We ensured that phenological development stages did not decrease with time, that is $\dot{y}_d \geq \dot{y}_{d-1}$, where $\dot{y}_{d-1}$ is the sampled phenological development on the previous observation day $d-1$. Of the 20 site-years generated in this manner, 10 site-years were used for calibration while the remaining 10 were used for validation. The synthetic site-years were ordered randomly during BSU calibration.

The controlled cultivar–environment sequence represents a sequence of site-years where the same cultivar is grown under the same environmental conditions. In this case, however, the model may not accurately simulate the observations, implying the presence of model structural error (e.g. the model's inability to capture slow emergence as explained in Appendix A: SPASS phenology model). For the controlled cultivar–environment sequence, we generated the synthetic site-year data from observations of the cultivar grown at 6_2010. For each synthetic site-year, the phenological development $\dot{y}_d$ on observation day d was sampled from the range of total observation uncertainty σ_d around the observed mean $\overline{y}_d$. As in the ideal sequence, we ensured that phenological development stages did not decrease with time. Again, 10 site-years were randomly assigned for calibration.

3.2.6.2 True sequences

A total of three sequential updates in Kraichgau and six sequential updates in the Swabian Alb were analysed (Fig. 3.2). In each sequential update, an additional site-year was included in the calibration dataset, following the actual chronological order in which maize was grown in the regions. For the *Kraichgau sequence*, four site-years were available for calibration and validation (3_2011, 2_2012 1_2014, and 2_2014). The model was sequentially calibrated to phenological development of maize for site-years 3_2011, 2_2012, and 1_2014. After each update, phenological development was predicted for the subsequent site-years. For example, in the first sequential update at Kraichgau, the model was calibrated to 3_2011. The site-years 2_2012, 1_2014 ,and 2_2014 were used for validation to assess the prediction quality of the calibrated model. In the second sequential update, the model was calibrated to 3_2011 and 2_2012, while 1_2014 and 2_2014 were used for validation. Note here that the number of site-years used for validation decreases with each sequential update. In the *Swabian Alb sequence*, seven site-years were available for sequential calibration and validation (6_2010, 5_2011, 5_2012, 6_2013, 5_2015, 5_2016, and 6_2016). The sequential updates were performed in a similar manner as in Kraichgau.

3.3 Results

In this section, we first describe the results for one example of Bayesian calibration using the data from site-year 6_2010 (section 3.3.1). Here, we examine the resulting simulated phenology after calibration as well as the posterior parameter distributions. We then look at the results from the synthetic and true sequences. We first evaluate the evolution of the posterior parameter distributions with sequential updates. As an example, we analyse the marginal distributions of the individual parameters and entropy of the joint parameter distributions for the true sequences (section 3.3.2). Lastly, we report the prediction quality results for the synthetic and true sequences (section 3.3.3).

3.3.1 Bayesian calibration results

By way of example, Fig. 3.4 shows the Bayesian phenological model calibration results for silage maize for the first site-year 6_2010. Cross-plots of the posterior parameters (Fig. 3.4i) show a weak negative correlation between PDD1 and TMINDEV1 and between PDD1 and DELTOPT1, while a weak positive correlation is observed between PDD1 and DELTMAX1. The observed mean phenological development falls within the range of simulations after calibration (Fig. 3.4ii). The marginal posterior parameter distributions are narrower than the initial prior distributions (Fig. 3.4iii). A shift in parameter distribution to the margins of the prior ranges is also noteworthy.

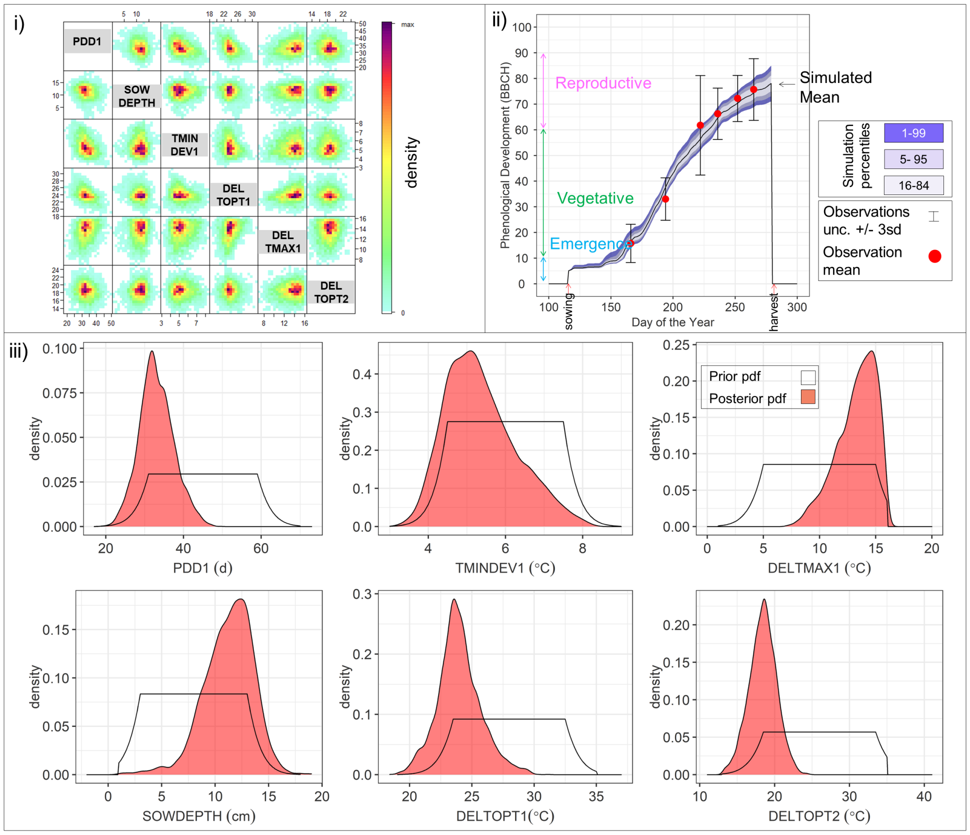


Figure 3.4: Results of Bayesian calibration of the model to phenological development (BBCH stages) for site-year 6_2010. (i) Cross-plot of the posterior samples of the six estimated parameters. Red represents high density and blue low density (IDPmisc package in R, Locher (2020)). (ii) Observed and simulated phenological development after calibration, plotted against the day of the year. The red points are the mean observations, while the black error bars indicate ± 3 SDs. The mean simulation is indicated by the continuous black line. The blue bands represent the different percentiles of simulated phenology. Note that the simulated phenology bands only represent the uncertainty in model parameters and do not include the noise term. (iii) Prior (white) and posterior (salmon) marginal parameter distributions for the six estimated parameters.

3.3.2 Parameter uncertainty

We analysed the change in posterior parameter distribution with the sequential updates. Figure 3.5i shows the marginal initial prior and posterior parameter distributions for the Swabian Alb and Kraichgau true sequences. The x-axis from left to right indicates the initial prior parameter distribution followed by the sequential calibration of the model to an increasing number of site-years. The distributions for the six estimated parameters are compared after each sequential update. The width of each box with whiskers represents the uncertainty in the parameter values. There is a clear narrowing of parameter distributions after the first sequential update from the initial prior. However, with the exception of DELTOPT2, the remaining parameters do not show a noticeable and consistent narrowing in range with sequential updates. Information entropy of the joint posterior parameter distributions in Fig. 3.5ii decreases with sequential updates and there is a large reduction in entropy with the first sequential update. In the Swabian Alb sequence (Fig. 3.5iia), entropy continues to decrease until the model is calibrated to 6_2010, 5_2011, and 5_2012, after which there is no significant reduction. In the Kraichgau sequence (Fig. 3.5iib), the inclusion of 1_2014 during calibration results in further uncertainty reduction. Similar observations were made for the synthetic sequences (Supplement S5).

3.3.3 Prediction quality

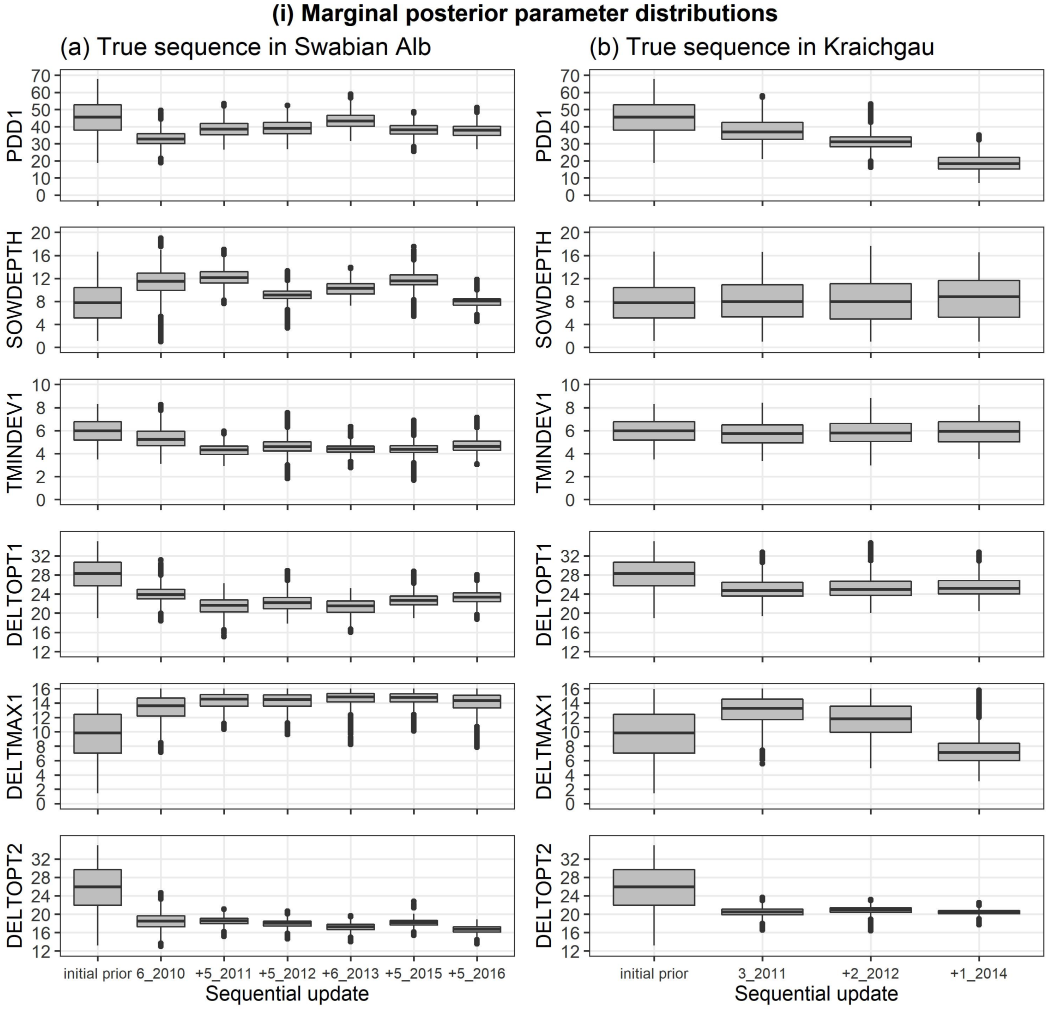
3.3.3.1 Synthetic sequences

In the synthetic sequences, we assessed the prediction quality after applying BSU to 10 synthetic site-years, while excluding model structural error and inter-site-year differences in cultivar and environmental conditions in the ideal sequence and controlled cultivar–environment sequence, respectively. In both sequences we account for identification uncertainty and spatial variability within the modelled site. Figure 3.6 shows the trend in median NRMSE and bias with the sequential updates from 1 to 10, for the two synthetic sequences. While the bias and NRMSE were calculated for all parameter vectors in the posterior sample derived from the MCMC sampling method, only the median values are plotted and analysed for simplicity.

In the ideal sequence (Fig. 3.6i), the overall median NRMSE (Fig. 3.6ia) and bias (Fig. 3.6ib) are low, with many site-years exhibiting a drop in the median NRMSE below a value of 1. However, after a few sequential updates, no further reduction is observed. In the controlled cultivar–environment sequence (Fig. 3.6ii), although most individual site-years showed a reduction in median NRMSE with the sequential updates, there were some that exhibited an increase in median NRMSE (ss2_12 and ss2_15 in Fig. 3.6iia). These site-years were also characterized by low initial median prediction bias, followed by an increase in the absolute bias with sequential updates (Fig. 3.6iib).

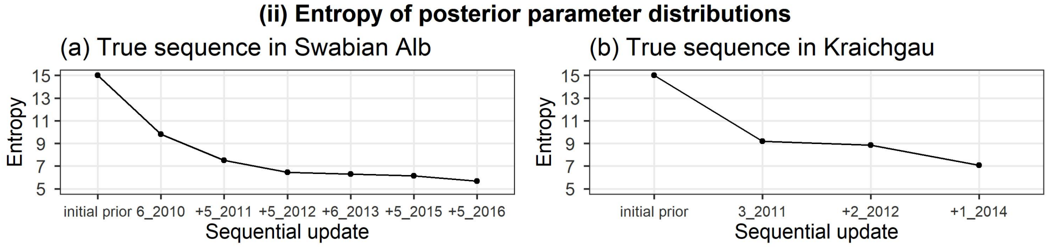
3.3.3.2 True sequences

Because fewer site-years were used for validation in the true sequence as compared to the synthetic sequence, we analysed the prediction quality for each validation site-year individually, with the sequential


(i) Marginal posterior parameter distributions

(a) True sequence in Swabian Alb

(b) True sequence in Kraichgau


(ii) Entropy of posterior parameter distributions

(a) True sequence in Swabian Alb

(b) True sequence in Kraichgau

Figure 3.5: (i) Marginal initial prior and posterior parameter distributions of the six estimated parameters plotted against the calibration site-years, after BSU was applied to a true sequence (a) on the Swabian Alb and (b) in Kraichgau. The SPASS model was calibrated to observed phenological development (BBCH). (ii) Information entropy of the joint posterior parameter distributions plotted against the calibration site-years, after BSU was applied to the true sequences. The x-axis labels from left to right indicate the initial prior parameter distribution followed by the sequential calibration of the model to an increasing number of site-years. The "+" symbol before the site-year label on the x-axis indicates the new site-year that was included in the sequential calibration. The length of the box in (i) represents the inter-quartile range (IQR), whiskers extend from the boxes up to 1.5× IQR and values beyond this range are plotted as points.

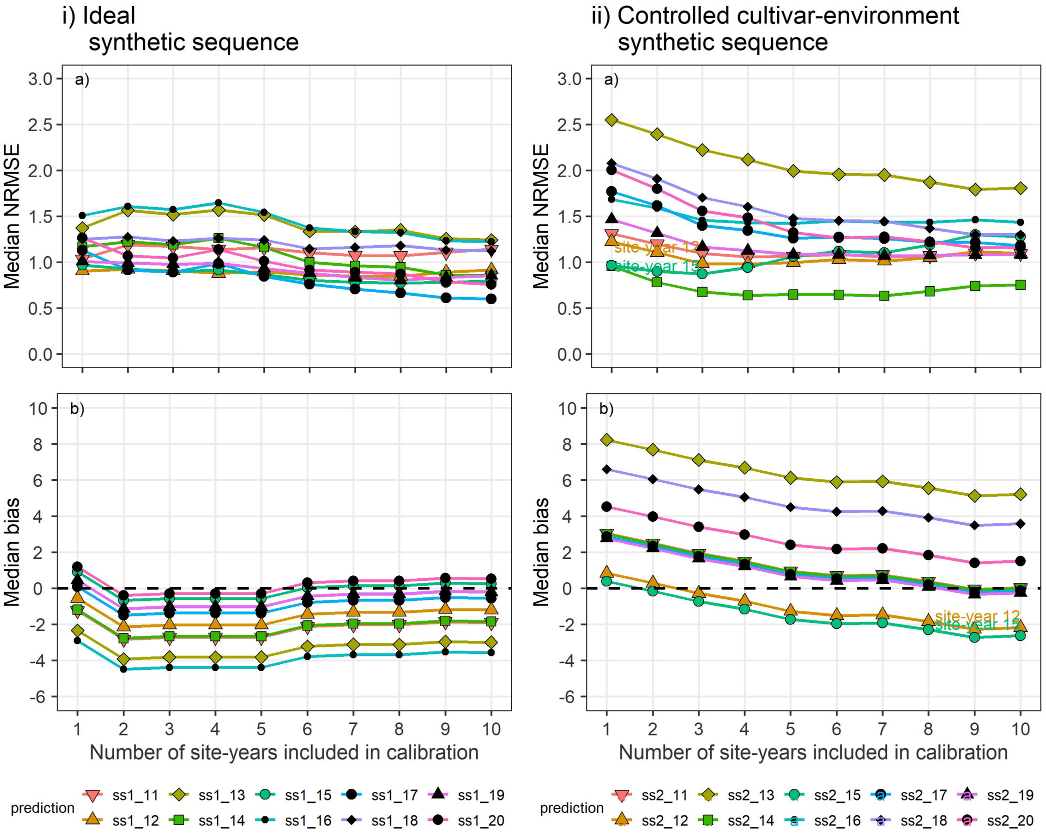


Figure 3.6: (a) Median NRMSE and (b) median bias of prediction for the 10 validation site-years, after BSU was applied to the ideal (i) and controlled cultivar–environment (ii) synthetic sequences. The number of site-years used for calibration is shown on the x-axis and represents the sequential updates from 1 to 10. The SPASS model was calibrated to phenological development (BBCH). The lines and points correspond to the 10 synthetic validation site-years: ss1_11-ss1_20 from the ideal sequence and ss2_11-ss2_20 from the controlled cultivar–environment sequence.

updates. Figure 3.7 shows the prediction quality (i.e. NRMSE and bias for all the posterior predictive samples) of the model after BSU was applied to the true sequence of site-years in Kraichgau (Fig. 3.7i–iii) and on the Swabian Alb (Fig. 3.7iv–ix). For each site-year, we plot the quality of prediction, after calibration to all preceding site-years. For example, Fig. 3.7vi shows the performance metric for site-year 6_2013 after the model was calibrated first to 6_2010, then to 6_2010 and 5_2011, and finally to 6_2010, 5_ 2011, and 5_2012, respectively (blue box-plots from left to right). As a reference, the performance metric derived from calibrating the model to the target site-year, namely 6_2013 in Fig. 3.7vi, is shown as the leftmost result (grey box-plot) of each sequence. It is clear that this calibration always yields the best performance metrics for the given data. While the NBias was calculated for all parameter vectors in the posterior MCMC sample, only the median values of the absolute NBias are also plotted to compare the trends between NRMSE and NBias with the sequential updates.

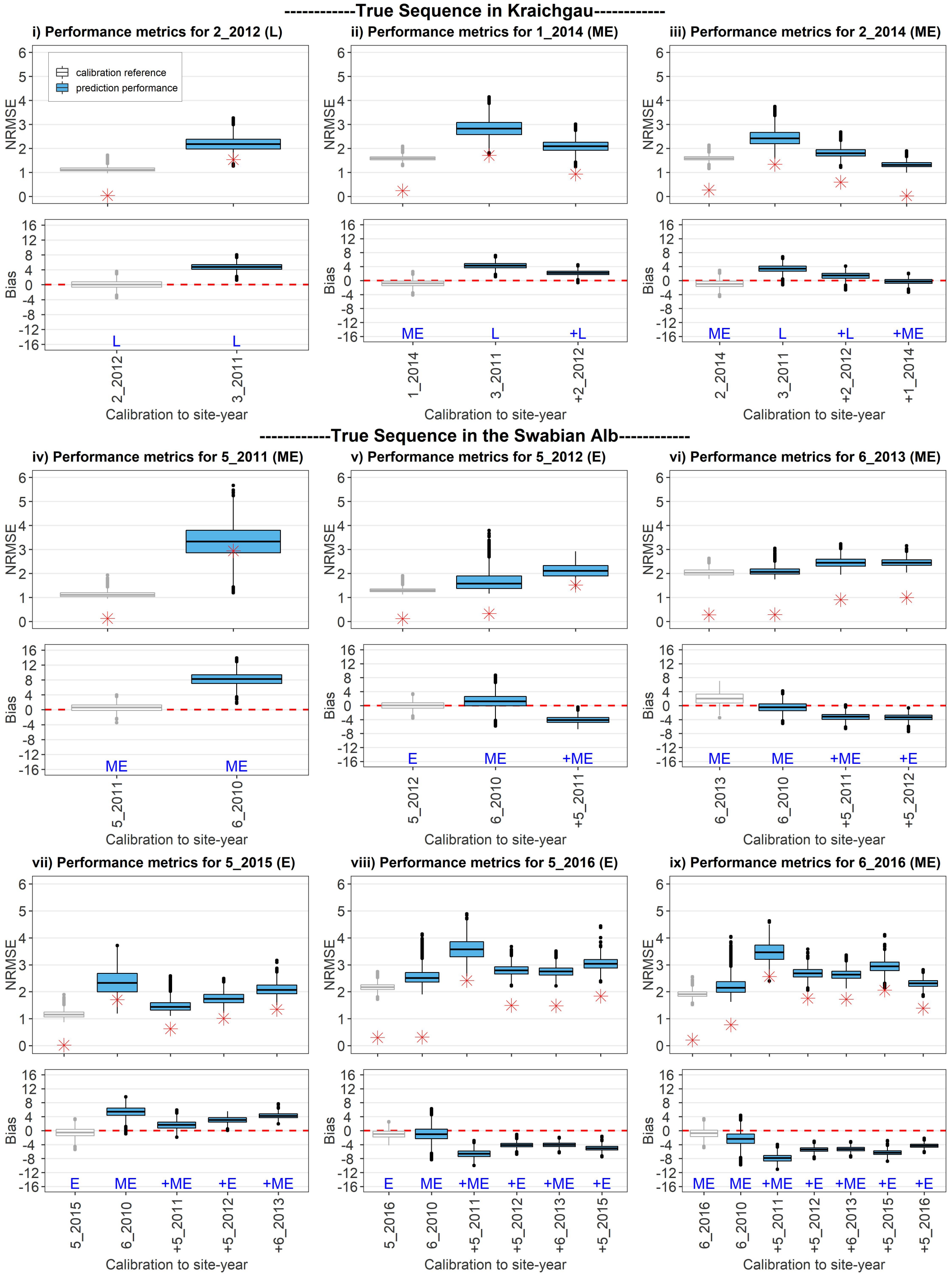


Figure 3.7: Performance metrics for site-years in Kraichgau (i–iii) and on the Swabian Alb (iv–ix), after applying BSU to the two true sequences. The SPASS model was calibrated to observed phenological development (BBCH). NRMSE and bias are plotted against the site-years used in calibration. In each sub-plot, the grey box-plot represents the calibration performance metric, i.e. when the model is calibrated to the site-year of interest. The blue box-plots represent the prediction performance metrics when the model is calibrated (from left to right) to an increasing number of preceding site-years. L, ME, and E indicate the maturity group of the cultivars: late, mid-early, and early, respectively. The "+" symbol before the site-year label on the x-axis and before the maturity group label indicates the new site-year that was included in the sequential calibration. The length of the box represents the inter-quartile range (IQR), whiskers extend from the box up to 1.5× IQR and values beyond this range are plotted as points. The zero bias is indicated by a red dashed line in the bias plots. The median values of the absolute NBias are represented by red asterisks (*) in the NRMSE plots.

The NRMSE is expected to decrease with the inclusion of more site-years for calibration. This holds true in the case of Kraichgau, where mid-early cultivars were grown (Fig. 3.7ii, iii), but in hardly any case on the Swabian Alb (Fig. 3.7iv–ix). We also expected the prediction quality to improve when a calibration sequence is made up of the same cultivar or ripening group. Note, however, the poor prediction quality in Fig. 3.7iv and the increase in NRMSE with the inclusion of 5_2011 in the calibration sequence in Fig. 3.7ix. Additionally, the prediction quality for the early cultivar at 5_2016 (Fig. 3.7viii) deteriorates upon the inclusion of the same cultivar grown at 5_2015 in the calibration sequence. In all predictions, the absolute NBias follows a similar trend as the NRMSE. Note that there is a difference in the performance metrics between the different site-years when the model is directly calibrated to the target site-year (grey box-plots in Fig. 3.7). The three site-years in Kraichgau and site-years 5_2011, 5_2012, 5_2015, and 6_2016 in the Swabian Alb exhibit good-to-moderate calibration quality, while 6_2013 and 5_2016 have moderate-to-poor calibration quality.

3.4 Discussion

In this study, we aimed to analyse whether progressively incorporating more data through BSU reduces model parameter uncertainty and produces robust parameter estimates for predicting phenology of silage maize.

3.4.1 Parameter uncertainty

Bayesian calibration resulted in reduced posterior parameter uncertainty in comparison to the initial prior ranges that were guided by expert knowledge (Fig. 3.4iii). The uncertainty in parameter DELTOPT2 decreased, as seen from the narrowing of the marginal posterior distributions (Fig. 3.5). The remaining parameters did not show a consistent progressive reduction in uncertainty with the sequential updates. They also had a relatively higher correlation with the other parameters (Fig. 3.4i). The lack of uncertainty reduction may be due to equifinality, meaning that multiple parameter combinations produce the same output (Adnan et al., 2020; He et al., 2017a; Lamsal et al., 2018). The reduction in information entropy of the posterior parameter distributions after the sequential updates (Fig. 3.5ii) confirms the reduction in overall parameter uncertainty.

The optimum temperatures for vegetative ($\mathrm{TOPTDEV1} = \mathrm{TMINDEV1} + \mathrm{DELTOPT1}$) and reproductive ($\mathrm{TOPTDEV2} = \mathrm{TMINDEV2} + \mathrm{DELTOPT2}$) development are lower than our prior belief. The effective sowing depth (SOWDEPTH) is higher than the actual sowing depth of 3–5 cm, 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. 3.5ib), 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. 3.5ia), 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 since parameters may compensate for model structural errors and some parameters are correlated (Alderman and Stanfill, 2017).

3.4.2 Prediction quality

We analysed synthetic sequences to assess whether a consistent reduction in prediction error is achieved when more site-years are available for calibration, in the absence of model structural errors (ideal sequence), and in the absence of inter-site-year differences due to cultivars and environmental conditions (controlled cultivar–environment sequence). For the ideal sequence we used simulated phenology and added a random noise term that represents spatial variability and identification error. For the controlled cultivar–environment sequence we used the observations instead of simulated phenology to generate the dataset. Hence, in the latter sequence, there is not only random noise but also a model structural error component. As the noise and model error components cannot be resolved, the estimated model parameters compensate for both, leading to larger prediction errors (Fig. 3.6ii).

In the ideal sequence, the model was able to accurately simulate the observations, the only source of between-site-year variability being within-site spatial variability and identification uncertainty. The overall initial prediction quality was moderate to good, indicating that when there was no model structural error, the calibrated model was able to predict moderately well in spite of some observational variability (Fig. 3.6i). The progressive drop in median NRMSE to a value of 1 indicated that the calibrated model was able to explain all other variability apart from that arising from the total observation uncertainty. Thus, with this sequence, we demonstrated the successful application of the BSU approach in ideal conditions.

In the controlled cultivar–environment sequence, the same cultivar was grown in the same environmental conditions across the site-years. With this sequence, we tested the success of the BSU approach when model structural errors could exist in addition to between-site-year variability as in the ideal sequence. The overall change in prediction error decreased with the sequential updates, as it possibly approaches an irreducible value. This is seen from the convergence of the different lines corresponding to the prediction site-years in Fig. 3.6iia. However, this irreducible value is higher than an NRMSE of 1 due to model structural error. Prediction error for most individual site-years decreased with the sequential updates. However, there were two site-years where the error increased (ss2_12 and ss2_15). These two site-years initially exhibited a low positive prediction bias that progressively became negative with the sequential updates (Fig. 3.6iib). This can be attributed to representativeness of the calibration data (Wallach et al., 2021a). The two prediction site-years were more similar to the initial few site-years than the later site-years in the calibration sequence.

We applied the BSU approach to real-world conditions represented by the true sequences of silage maize grown in Kraichgau and on the Swabian Alb (Fig. 3.7). In Kraichgau, the prediction quality improved with sequential updates as expected. However, it deteriorated for many site-years on the Swabian Alb. This is again attributed to the representativeness of the calibration data as seen in the controlled cultivar–environment sequence. To understand this behaviour we carried out single site-year calibration and predictions, i.e. calibrating the model to individual site-years and predicting the remaining site-years (for details, refer to Appendix C. Single site-year calibration). As parameter estimates may vary by ripening group or cultivar, we analysed the prediction results within these classes. Calibrating the model to a site-year from the same ripening group or even the same cultivar as the prediction target site-year did not always result in the best prediction quality. Within the mid-early and early ripening groups, prediction quality showed a correlation with the difference in average temperature during the vegetative phase, between the calibration and prediction target site-year. This correlation indicated that the best predictions of phenology for a particular site-year would be achieved when the model is calibrated to a cultivar from the same ripening group and grown under the same temperature conditions during the vegetative phase. The calibration quality for the individual site-years represented by grey box-plots in Fig. 3.7 shows that the model is able to simulate some site-years better than others. Residual analysis (Supplement S3) revealed that the model was unable to capture the slow development during the vegetative phase for these site-years with poorer calibration quality. This could be due to model limitations (i.e. model equations or hard-coded parameters) and could explain the correlation between temperature similarity and prediction quality.

The single site-year predictions showed that site-years 1_2014 and 2_2014, where the same mid-early cultivar was grown, were the best predictors of each other and their prediction by the late cultivar at 3_2011 was poorer. Therefore, in the case of the Kraichgau sequence (Fig. 3.7ii–iii), we observed a decrease in prediction error as we progressively calibrated the model to 3_2011, to 3_2011 and 2_2012, and to 3_2011, 2_2012 and 1_2014. In the Swabian Alb sequence (Fig. 3.7iv–ix) where mid-early and early cultivars are grown, the effect of different ripening groups and temperatures caused an increase in prediction error.

In real-world conditions represented by the true sequences, the prediction quality thus depends on the interplay between model limitations and inherent data structures presented in the differences between maturity group and cultivars. Since the model calibration and prediction quality varies with environmental factors, it highlights the need to better account for the influence of these environmental drivers in the model. This would increase model transferability to other sites. This could be best achieved by improving the process representation in the model and by including the uncertainty in forcings during calibration. An alternative approach would be to define separate cultivar- and environment-specific parameter distributions. It is common practice to determine cultivar-specific parameters in crop modelling (Gao et al., 2020). He et al. (2017b) found that data from different weather and site conditions are required to obtain a good calibrated parameter set for a particular cultivar. Improved crop model performance has been reported upon the inclusion of environment-specific parameters in calibration (Coelho et al., 2020). Cultivar- or genotype- and environment-specific parameters already exist in some models (Jones et al., 2003; Wang et al., 2019). However, these genotype parameters have also been found to vary with the environment, indicating that they may represent genotype × environment interactions and not fundamental genetic traits (Lamsal et al., 2018). Further analysis of calibrated model parameters and model performance metrics with respect to environmental variables would provide insights into areas for model improvement. Nonetheless, the cultivar and environmental dependency of parameters is a major drawback for large-scale model applications and long-term predictions, as information on crop cultivars is usually not available on regional scales and specific characteristics of future cultivated varieties are currently unknown. It is essential to collect cultivar and maturity group information in official surveys. Furthermore, other Bayesian approaches such as hierarchical Bayes, which allow for the incorporation of this information during calibration, should be explored. Model calibration in a Bayesian hierarchical framework would enable inherent data structures, represented by the cultivars within ripening groups of a particular species, to be accounted for. Additionally, differences in environmental conditions can also be represented. On regional scales, where information about maturity groups and cultivars is unavailable, accounting for environmental effects alone may still prove to be beneficial. A Bayesian hierarchical approach could even be applied to predict the growth of current as well as future cultivars.

3.4.3 Limitations

We would like to draw attention to the three assumptions in the current study which might cause an underestimation of uncertainties. First, the standard deviation of the likelihood model was not estimated, but assumed to be known and equal to the sum of observed spatial variability and identification error. It represents the minimum error and is equal to the total error only if there are no differences in environmental conditions and cultivars across the site-years. Second, the likelihood model was assumed to be centred at 0, which only holds true when there are no structural errors. In most cases, however, model structural errors and other systematic errors will exist, which may result in much larger errors than what was assumed. Third, the errors are assumed to be independent and identically distributed. A violation of this assumption can lead to underestimation of uncertainty in the parameters and the output state variable (Wallach et al., 2017). In the residual analysis of the sequential updates with three or more site-years, a slight deviation from a Gaussian distribution was observed (Supplement S3). This skewness was caused due to model limitations, that is its inability to capture the slow development observed during the vegetative phase in some site-years. Autocorrelation of errors can exist for state variables such as phenology that are based on cumulative sums. However, based on the limited dataset, an autocorrelation in the errors could not be substantiated and an in-depth analysis is far beyond the scope of this study.

We observed that the posterior parameter distributions were at the margins of the initial prior distribution ranges, for which this study now provides a basis to update this prior belief. This considerable update of the parameter prior indicates that either the prior ranges are not suitable for the cultivars in this study or that the parameters are compensating for structural limitations of the model. Further in-depth investigation of their potential contributions could only be achieved with datasets that are much larger than the one employed here.

3.5 Conclusions

Through a Bayesian sequential updating (BSU) approach, we extended a classical application of Bayesian inference through time to analyse its effectiveness in the calibration and prediction of a crop phenology model. We assessed whether BSU of the SPASS model parameters, based on new observations made in different years, progressively improves prediction of the phenological development of silage maize.

We applied BSU to synthetic sequences and true sequences. As expected, the parameter uncertainty decreased in all sequences. The prediction errors decreased in most cases in the synthetic sequences, where we had an ideal model that was able to accurately simulate observations, and where the model could contain structural errors but the dataset contained only a single maize cultivar grown under the same environmental conditions. In the ideal synthetic sequence, the prediction quality was variable for the first few sequential updates. The prediction error then decreased in both synthetic sequences until it approached an irreducible value. In the true sequences, however, which included cultivars from different ripening groups and environmental conditions, the prediction quality deteriorated in most cases. Differences in ripening group and temperature during the vegetative phase of growth between the calibration and prediction site-years influenced prediction quality.

With an increasing amount of data being gathered and with improvements in data-gathering techniques, there is a drive to use all available data for model calibration. However, our study shows that a simplistic approach of updating the model parameter estimates without accounting for model limitations and inherent differences between datasets can lead to unsatisfactory predictions. To obtain robust parameter estimates for crop models applied on a large scale, the Bayesian approach needs to account for differences not only in maturity groups and cultivars but also in environment. This could be achieved by applying Bayesian inference in a hierarchical framework, which will be the subject of future work.

3.6 Appendix A: SPASS phenology model

In the following paragraphs we describe the equations in the SPASS phenology model (Wang, 1997). The model parameters are indicated by words with all capitalized letters (e.g. SOWDEPTH, PDD1 etc.).

The crop passes through four main stages: sowing (stage −1.0), germination (stage −0.5), anthesis (stage 1.0, end of the vegetative phase and beginning of reproductive phase), and maturity (stage 2.0). Temperature and photoperiod are the two main factors affecting phenological development rate. The impact of water availability on germination is also reflected in the SPASS model.

For germination, soil moisture is the limiting factor. Germination occurs when

$$\theta_{\mathrm{act}(i_{\mathrm{s}})} > \theta_{\mathrm{pwp}(i_{\mathrm{s}})} \tag{3.A1}$$

or

$$\begin{aligned} 0.02 \leq\ & 0.65\left[\theta_{\mathrm{act}(i_{\mathrm{s}})} - \theta_{\mathrm{pwp}(i_{\mathrm{s}})}\right] \\ & + 0.35\left[\theta_{\mathrm{act}(i_{\mathrm{s}}+1)} - \theta_{\mathrm{pwp}(i_{\mathrm{s}}+1)}\right] \end{aligned},$$

where $\theta_{\mathrm{act}(i_{\mathrm{s}})}$ is the actual volumetric water content of the seed soil layer i_{s} and $\theta_{\mathrm{pwp}}(i_{\mathrm{s}})$ is the volumetric water content in the seed soil layer at permanent wilting point. If these conditions are not met within 40 d of sowing, crop failure is assumed.

The development rate from germination to emergence ($R_{\mathrm{dev,emerg}}$) (d^{-1}) is controlled by air temperature:

$$R_{\mathrm{dev,emerg}} = (T_{\mathrm{avg}} - T_{\mathrm{base}}) \times 0.5/\Sigma T, \tag{3.A2}$$

where, T_{avg} (°C) is the daily average air temperature and T_{base} (°C) is the base temperature set to 10 °C for maize. The term ΣT (°C) is the temperature sum needed for emergence:

$$\Sigma T = 15.0 + 6.0 \times \mathrm{SOWDEPTH}, \tag{3.A3}$$

where SOWDEPTH (cm) is the sowing depth of the seed.

After emergence, the development rate in the vegetative phase $R_{\mathrm{dev,v}}$ (d^{-1}) depends on temperature and photoperiod:

$$R_{\mathrm{dev,v}} = R\mathrm{max}_{\mathrm{dev,v}} f_{T,\mathrm{v}}(T) f\left(h_{\mathrm{php}}\right) \tag{3.A4}$$

where $R\mathrm{max}_{\mathrm{dev,v}} = 1/\mathrm{PDD1}$ is the maximum development rate in the vegetative phase (d^{-1}), PDD1 is the number of physiological development days from emergence to anthesis (d), $f(h_{\mathrm{php}})$ is the photoperiod factor, and $f_{T,\mathrm{v}}(T)$ is the temperature response function (TRF) for the vegetative phase. The photoperiod factor is expressed as

$$f\left(h_{\mathrm{php}}\right)=1-e^{\frac{-4\left(h_{\mathrm{php}}-dl\,\mathrm{min}\right)}{\mathrm{DLOPT}-dl\,\mathrm{min}}} \tag{3.A5}$$

where

$$\mathrm{dlmin}=\mathrm{DLOPT}+4/\mathrm{PDL}$$

h_{php} (h) is the photoperiod length, that is the amount of time between the beginning of the civil twilight before sunrise and the end of the civil twilight after sunset (the time when the true position of the centre of the sun is 4° below the horizon), PDL $(-)$ is the photoperiod sensitivity, and DLOPT (h) is the optimum daylength for a particular cultivar.

The development rate in the generative or reproductive phase $(R_{\mathrm{dev},r})$ (d^{-1}) only depends on temperature such that:

$$R_{\mathrm{dev,r}}=R\mathrm{max}_{\mathrm{dev,r}}f_{T,\mathrm{r}}(T) \tag{3.A6}$$

where $R\mathrm{max}_{\mathrm{dev,r}}= 1/\mathrm{PDD2}$ is the maximum development rate in the reproductive phase (d^{-1}), PDD2 is the number of physiological development days from anthesis to maturity (d), and $f_{T,\mathrm{r}}(T)$ is the temperature response function (TRF) for the reproductive phase.

The temperature response function f_T has cardinal temperatures: minimum temperature, $T_{\min}$ (°C), optimum temperature, T_{opt} (°C), and maximum temperature, $T_{\max}$ (°C):

$$f_T\left(T,T_{\min},T_{\mathrm{opt}},T_{\max}\right) = \begin{cases} \frac{2(T-T_{\min})^{\alpha}\cdot\left(T_{\mathrm{opt}}-T_{\min}\right)^{\alpha}-(T-T_{\min})^{2\alpha}}{\left(T_{\mathrm{opt}}-T_{\min}\right)^{2\alpha}} & \text{if } T_{\min}\leq T\leq T_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{3.A7}$$

where

$$\alpha=\frac{\ln 2}{\ln\left(\frac{T_{\max}-T_{\min}}{T_{\mathrm{opt}}-T_{\min}}\right)}$$

As the TRF is phase-specific, the cardinal temperatures are also phase-specific. For $f_{T,\mathrm{v}}$, the cardinal temperatures are $T_{\min}=\mathrm{TMINDEV1}, T_{\mathrm{opt}}=\mathrm{TOPTDEV1}, T_{\max}=\mathrm{TMAXDEV1}$, while for $f_{T,\mathrm{r}}$, the cardinal temperatures are $T_{\min}=\mathrm{TMINDEV2}, T_{\mathrm{opt}}=\mathrm{TOPTDEV2}, T_{\max}=\mathrm{TMAXDEV2}$.

The development stages after germination (S_{dev}) are calculated in daily time steps as

$$S_{\mathrm{dev}}=\sum_{d=d_{\mathrm{germ}}}^{n} R_{\mathrm{dev}}-0.5, \tag{3.A8}$$

where d_{germ} is the day on which seed germination occurs and n is the number of days after germination:

$$R_{\mathrm{dev}}=\begin{cases} R_{\mathrm{dev,emerg}} & \text{if } -0.5\leq S_{\mathrm{dev}}<0.0 \\ R_{\mathrm{dev,v}} & \text{if } 0.0\leq S_{\mathrm{dev}}<1.0 \\ R_{\mathrm{dev,r}} & \text{if } 1.0\leq S_{\mathrm{dev}}<2.0. \end{cases} \tag{3.A9}$$

Finally, the SPASS development stages ($-0.5 \leq S_{\text{dev}} \leq 2$) are converted to BBCH development stages ($0 \leq \text{BBCH} \leq 95$). Here, $S_{\text{dev}} = 0$ corresponds to BBCH = 10 (emergence and start of the vegetative phase), $S_{\text{dev}} = 0.4$ to BBCH = 31, and $S_{\text{dev}} = 1$ to BBCH = 61 (start of the generative or reproductive phase).

Preliminary simulations showed that the model was unable to capture the slow rate of emergence after sowing, as seen in the observations, when the true sowing depth for maize was used. This could be due to uncertainty in the hard-coded parameters in the emergence rate Eq. (3.A2) which were not estimated in this study. This is an example of structural error in the model. In order to simulate this slow emergence, an effective sowing depth (SOWDEPTH) was set, which is deeper than the actual sowing depth range for maize (3–5 cm). Another example of model structural error would be missing factors, which play a role in phenological development. SPASS assumes that phenological development depends only on temperature and daylength. Other factors such as water stress, nitrogen deficiencies, and high ozone concentrations could also play a role but are ignored. Moreover, the shape of the temperature response function could be inadequate in capturing the plant's true response to temperature.

In the case of the cardinal temperatures for the vegetative and reproductive phases, the parameters DELTOPT and DELTMAX were introduced instead of TOPTDEV and TMAXDEV during sensitivity analysis and MCMC sampling, to ensure that during parameter sampling TMINDEV < TOPTDEV < TMAXDEV. Thus, TMINDEV, DELTOPT, and DELTMAX were used to parameterize the temperature response function during calibration, where TOPTDEV =TMINDEV+DELTOPT and TMAXDEV=TOPTDEV+DELTMAX.

3.7 Appendix B: Posterior sampling using MCMC Metropolis algorithm

The posterior parameter distribution was sampled using a Markov chain Monte Carlo (MCMC) method based on the Metropolis algorithm (Iizumi et al., 2009; Metropolis et al., 1953). Three Markov chains were run in parallel using the foreach (Microsoft and Weston, 2020) and doParallel (Microsoft and Weston, 2019) packages in R (R Core Team, 2020). First, initial parameter vectors were selected as a starting point for each chain. Then, the size of the transition kernel used to propose new candidate parameter vectors in the chain was adapted, based on the acceptance rate, to improve the efficiency of the MCMC algorithm (Gelman et al., 1996). After the adaptation, the Markov chains were run until the Gelman–Rubin convergence diagnostic for the posterior parameter distribution was ≤ 1.1 (Brooks and Gelman, 1998; Gelman and Rubin, 1992). The detailed steps are given here.

First sample

Step 1: Let $\boldsymbol{\theta}_1$ be an arbitrary initial parameter vector in a chain, selected from within the parameter ranges provided by the expert. This method of selection was used for the Bayesian calibration of site-year 6_2010. For the other calibration cases, the initial parameter vectors were obtained by sampling from the range of the posterior parameter distribution after calibration to 6_2010. This was done to reduce the time to convergence as it is expected that the posterior parameter distributions for the other calibration cases would be in the vicinity of the posterior distribution obtained after calibration to 6_2010. Bayes theorem is estimated as

$$P\left(\boldsymbol{\theta}_1 \,|\, \boldsymbol{Y}\right) \propto P\left(\boldsymbol{\theta}_1\right) P\left(\boldsymbol{Y}|\boldsymbol{\theta}_1\right), \tag{3.B1}$$

where$P(\boldsymbol{Y}|\boldsymbol{\theta}_1)$ and $P\left(\boldsymbol{\theta}_1\right)$ are calculated using Eqs. (3.9) and (3.10), respectively. The error function in Eq. (3.11) required for $P\left(\boldsymbol{\theta}_1\right)$ was calculated using the pracma package (Borchers, 2020).

Jump adaptation

A symmetrical transition kernel or jump distribution is used to select the next candidate parameter vector. The transition kernel is a normal distribution that is centred at the current parameter vector, and has a variance vector $\boldsymbol{V}^2$. The off-diagonal elements of the variance–covariance matrix are 0.

Step 2: The transition kernel centred at $\boldsymbol{\theta}_{t-1}$ is used to propose a new candidate parameter vector $\boldsymbol{\theta}_t^*$.

Step 3: The model is simulated using parameter vector $\boldsymbol{\theta}_t^*$ and the numerator of Bayes theorem is calculated using the prior and likelihood as per Eq. (3.B1).

Step 4: The acceptance ratio (r) for a proposed candidate parameter vector is

$$r = \frac{P\left(\boldsymbol{\theta}_t^*\right) P\left(\boldsymbol{Y}|\boldsymbol{\theta}_t^*\right)}{P\left(\boldsymbol{\theta}_{t-1}\right) P\left(\boldsymbol{Y}|\boldsymbol{\theta}_{t-1}\right)}. \tag{3.B2}$$

Step 5: The candidate parameter vector $\boldsymbol{\theta}_t^*$ is either accepted or rejected as the new parameter vector $\boldsymbol{\theta}_t$ based on the condition

$$\boldsymbol{\theta}_t = \left\{ \begin{array}{ll} \boldsymbol{\theta}_t^* & r > u \\ \boldsymbol{\theta}_{t-1} & r \leq u \end{array} \right), \tag{3.B3}$$

where $u \sim U(0,1)$ is a random sample from a uniform distribution between 0 and 1. Proposals of parameters which were outside the bounds of the prior and likelihood result in a zero in the numerator of Eq. 3.B2. These parameters are rejected and discarded. The next proposal is generated with the jump distribution centred at the last accepted parameter vector, until the next proposal is accepted.

Step 6: After 20 accepted parameter vectors per chain, the acceptance rate ar = acc/tot is calculated across the chains, where acc represents the number of accepted vectors (i.e. 20 accepted runs per chain ×3 chains in this case) and tot represents the total vectors proposed. Based on the acceptance rate (ar), the standard deviation V of the transition kernel, which controls the jump size, is adapted as per the condition in Eq. 3.B4, so that the acceptance rate is between 25 % and 35 % (Gelman et al., 1996;

Tautenhahn et al., 2012):

$$V = \begin{cases} V \times 1.01 & \text{ar} \geq 0.35 \\ V \times 0.99 & \text{ar} \leq 0.25 \\ V & 0.25 < \text{ar} < 0.35. \end{cases} \tag{3.B4}$$

If the acceptance rate ar is between 25 % and 35 %, we proceed to the main set of runs to obtain the posterior parameter distributions.

Main runs

In the main runs, steps 2–5 are repeated with the final jump distribution achieved at the end of the jump adaptation steps.

Step 7: The convergence of the chains after jump adaptation is checked using the Gelman–Rubin convergence criteria (GR). The gelman.diag function from the coda package in R (Plummer et al., 2006) was used to evaluate the GR diagnostic after every 20 accepted parameter vectors in each chain. As per the GR diagnostic criteria, the Markov chains have converged to represent a stable posterior distribution if within-chain variance is approximately equal to between-chain variance. The MCMC chains are stopped if there are a minimum of 500 accepted runs per chain and if GR ≤ 1.1 (Brooks and Gelman, 1998) for each parameter.

Step 8: In the final step, all the runs from the jump adaptation phase are discarded as burn-in. Parameters from the remaining accepted runs define the posterior distribution.

3.8 Appendix C. Single site-year calibration

In order to better understand the results of the true sequences, single site-year calibration and predictions were made within and across the two regions. Since calibration yields the best performance metrics, we analysed the median NRMSE ratio for each prediction-target site-year, i.e. the ratio between the median NRMSE of prediction and the median NRMSE of calibration to the prediction target (Fig. 3.C1). We expect that the model predicts best, i.e. with a low median NRMSE ratio, when it is calibrated to the same cultivar or ripening group. However, we found that this was not always the case. This is a result of careful analyses of calibration–prediction performance, detailed here.

The mid-early cultivar at 5_2011 was poorly predicted by all mid-early cultivars, but was better predicted by early cultivars. Site-years 1_2014 and 2_2014 in Kraichgau, where the mid-early cultivar Grosso was grown, were the best predictors of each other. However, even though the early cultivar LG 30.217 was grown at 5_2015 and 5_2016, these two site-years were not the best predictors of each other. Similarly, site-years 2_2012 and 3_2011, where the late cultivar Canavaro was grown, were also not the best predictors of each other. In predictions for mid-early cultivars, a spread in median NRMSE ratio was seen when the model was calibrated to other mid-early cultivars. The mid-early cultivar at 1_2014 and 2_2014 in Kraichgau had a comparable prediction quality when the model was calibrated to the late cultivar grown in Kraichgau or to the mid-early cultivars grown on the Swabian Alb.

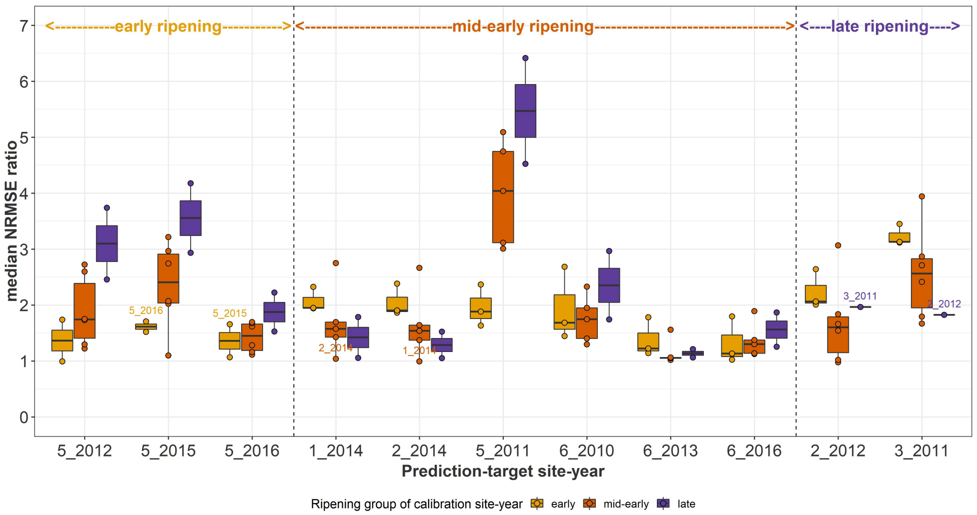


Figure 3.C1: Median NRMSE ratio for prediction-target site-years after single site-year calibration of the SPASS model to observed phenological development (BBCH). The median NRMSE ratio on the y-axis is the ratio between the median NRMSE of prediction and the median NRMSE of calibration to the prediction-target site-year. Each point represents the median NRMSE ratio of prediction of the site-year on the x-axis when the model was calibrated to phenology from every other site-year separately (single site-year calibration). The points are grouped and coloured by ripening group of the calibration site-year while the ripening group of the prediction target site-years are indicated on the top of the plot. The box and whiskers show the spread in median NRMSE ratio of predicting a particular site-year after the model was separately calibrated to site-years from a particular ripening group. Calibration site-year points from the same cultivar as the prediction site-year are labelled.

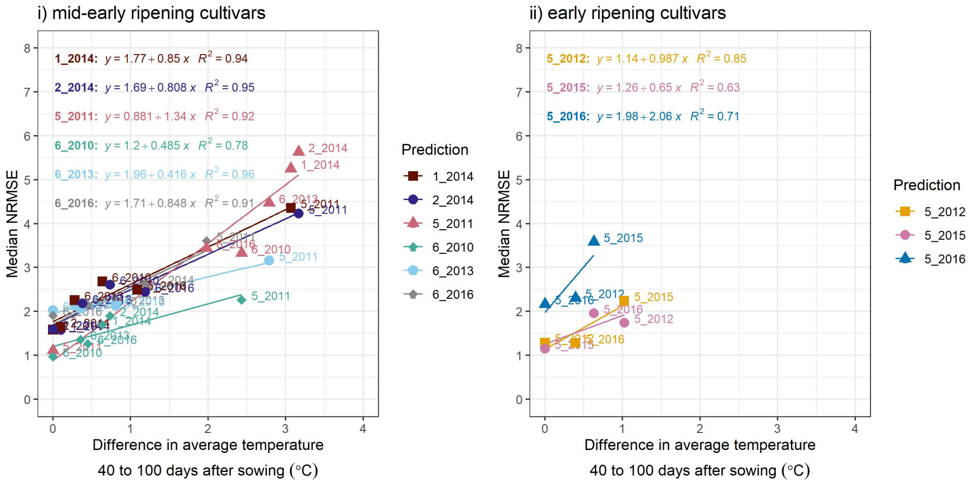


Figure 3.C2: A cross-plot between the performance metric median NRMSE and the absolute difference in temperature between the site-year used for calibration and the prediction-target site-year, averaged over 40–100 d after sowing, for (i) mid-early and (ii) early ripening cultivars. Colours of the best-fit lines and points indicate the prediction-target site-year. Median NRMSE points at 0 °C on the x-axis are calibration performance metrics for the target site-year while the remaining are prediction performance metrics. Point labels indicate the site-years to which the model was calibrated. The SPASS model was calibrated to observed phenological development (BBCH).

To explain the spread in prediction NRMSE within ripening groups, we examined the relationship between NRMSE and the difference in average temperature between the site-year used for calibration and the predicted or target site-year. The temperature was averaged over an interval of 40–100 d after sowing (i.e. approximate vegetative phase of development). For the mid-early ripening cultivars (Fig. 3.C2i), the median NRMSE shows a clear correlation. Albeit tested with a limited number of site-years, early-ripening cultivars (Fig. 3.C2ii) show a similar trend.

3.9 Appendix D: Platykurtic prior

An example of a platykurtic probability density function which is used as a weakly informative prior for the model parameters is shown in Fig. 3.D1. It is a combination of a uniform and normal distribution. The default, minimum, maximum, and standard deviation values from Table 3.2 were used in Eq. (3.10) to obtain the prior probability distribution for the estimated parameters.

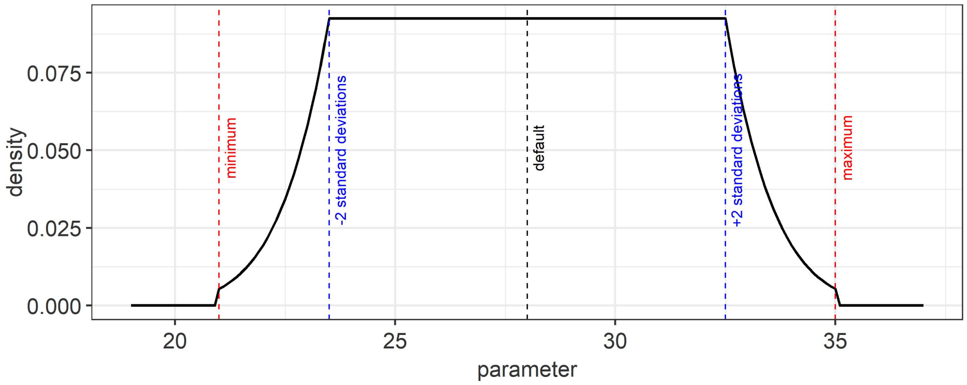


Figure 3.D1: An example of the platykurtic probability density function that was used as a prior for the model parameters. The default, minimum, maximum, and standard deviation values for the parameter are used to define this function.

CHAPTER 4

Bayesian multi-level calibration of a process-based maize phenology model

Authors: Michelle Viswanathan, Andreas Scheidegger, Thilo Streck, Sebastian Gayler, Tobias K. D. Weber

This chapter is taken from the publication “Viswanathan et al. (2022), Bayesian multi-level calibration of a process-based maize phenology model, Ecological Modelling, Volume 474, 110154, DOI: 10.1016/j.ecolmodel.2022.110154”, with minor modifications.

Abstract

Plant phenology models are important components in process-based crop models, which are used to assess the impact of climate change on food production. For reliable model predictions, parameters in phenology models have to be accurately known. They are usually estimated by calibrating the model to observations. However, at regional scales in which different cultivars of a crop species may be grown, not accounting for inherent differences in phenological development between cultivars in the model and the presence of model deficits lead to inaccurate parameter estimates. To account for inherent differences between cultivars and to identify model deficits, we used a Bayesian multi-level approach to calibrate a phenology model (SPASS) to observations of silage maize grown across Germany between 2009 and 2017. We evaluated four multi-level models of increasing complexity, where we accounted for different combinations of ecological, weather, and year effects, as well as the hierarchical classification of cultivars nested within ripening groups of the maize species. We compared the calibration quality from this approach to the commonly used pooled approach in which none of these factors are considered. The pooled model led to over-confident process model parameter estimates and comparatively poor calibration quality. The mean value of the unexplained residual error standard deviation reduced from 5.5 BBCH (phenological development units) in the pooled model case (BM-0) to 5.3 BBCH when eco-region and year effects (BMM-1) were considered. Additionally accounting for weather effects (BMM-2a) resulted in a mean value of 5.2 BBCH. Calibration quality especially improved when the hierarchical classification of cultivars within ripening groups of maize was incorporated. Including the hierarchical

classification with eco-region and year effects (BMM-2b) led to a mean residual error of 4.4 BBCH while additionally considering weather effects in the full model case (BMM-3) resulted in a value of 4.3 BBCH. Our findings have implications for regional model calibration and data-gathering studies, since it emphasizes that ripening group and cultivar information is essential. Furthermore, we found that if this information is not available, at least weather, eco-region and year effects should be taken into account. Accounting for only the eco-region and year effects led to parameter-compensation of the missing weather effects. Our results can facilitate model improvement studies since we identified possible model limitations related to temperature effects in the reproductive (post-flowering) phase and to soil-moisture. We demonstrate that Bayesian multi-level calibration of a phenology model facilitates the incorporation of hierarchical dependencies and the identification of model limitations. Our approach can be extended to full crop models at different spatial scales.

4.1 Introduction

Plant phenology plays an important role when assessing the impact of climate change and evaluating crop production (Menzel et al., 2006; Siebert and Ewert, 2012; Zhao et al., 2013; Wittich and Liedtke, 2015; He et al., 2017b; Wallach et al., 2021a). It is controlled by environmental variables and determines the timing of plant organ development and the distribution of the products of photosynthesis, such as sugars, to different parts of the plant. Thus, predictions of phenological development are essential for evaluating crop growth and yield, and for supporting field management decisions such as the timing of fertilizer application (Potgieter et al., 2021). These phenology predictions are made possible by using numerical models.

Phenology models are in turn important components of crop models, which are used for simulating crop growth and development, and yield. Besides data-driven statistical models, it is process-based models which enable a thorough understanding of the underlying processes for evaluating potential policy interventions and adaptation to climate change (Lobell and Asseng, 2017). In these process models, phenology is simulated as a parametric function of environmental variables such as temperature and photoperiod. Parameters of these models have to be determined accurately to ensure reliable predictions.

Since model parameters often cannot be measured directly, they need to be estimated by comparing model outputs with observed data using methods such as Bayesian inference. Bayesian calibration provides a framework to quantify different sources of uncertainty, which is essential for better predictions, with the added value of being able to include prior information (Makowski et al., 2002). To this end, Bayesian methods have been applied in numerous crop model calibration studies (Makowski et al., 2006; Iizumi et al., 2009; Sexton et al., 2016; Alderman and Stanfill, 2017; He et al., 2017b; Gao et al., 2020).

During the calibration of phenology models, cultivar-specific parameters are usually estimated (Gao et al., 2020). This is because phenological traits differ markedly, not only between species and between ripening or maturity groups of crop species such as maize (Oluwaranti et al., 2015), but also between cultivars within these ripening groups. Phenological development of a cultivar is also dependent on the environment and reflects genotype × environment interactions. Thus, methods such as selecting cultivar observations from contrasting environments for calibration (He et al., 2017b) and using cross-validation tests while evaluating environmental responses of the cultivar (Fukui et al., 2015) are suggested for determining these cultivar-specific parameters.

However at regional scales, where many cultivars of a particular species are grown together, cultivar-specific parameters may not be suitable. In such calibration studies, region-specific model parameter estimates are obtained for the crop species (Iizumi et al., 2009; Therond et al., 2011; Angulo et al., 2013; Soltani et al., 2016), but differences between cultivars grown in the region are usually not taken into account. The resultant estimates are a compromised solution for all the cultivars grown in different environments represented by the calibration data set.

Furthermore, models may not represent the underlying processes accurately. Commonly, environmental interactions are incompletely or poorly understood, leading to conceptual uncertainty. This is reflected in multiple model formulations to represent the same process (Kumudini et al., 2014; Wang et al., 2015; Wu et al., 2017). Consequently, models may have structural deficits. But the implicit assumption in Bayesian inference is that the model is without errors or that all errors are perfectly described (Hsueh et al., 2022). During calibration, the estimated parameters may compensate for model limitations (Wallach, 2011). As a consequence, even parameters which are meant to be cultivar-specific have been found to vary with the environment (Ceglar et al., 2011), thus often loosing their original physiological meaning (Lamsal et al., 2018).

Ignoring inherent data structures and the presence of model deficits result in inaccurate parameter estimates. When data structures such as the hierarchical classification of cultivars nested within ripening groups of a species are ignored, the uncertainty in the resultant ‘effective’ parameters are underestimated. Furthermore, indiscriminate use of large amounts of data to calibrate imperfect models leads to an overconfidence in erroneous parameter estimates (Motavita et al., 2019), which in turn has been shown to result in erroneous model predictions (Viswanathan et al., 2022b). Thus, it is important to account for these data structures and model deficits during parameter estimation.

Therefore, we propose a Bayesian multi-level calibration of a process-based plant phenology model to account for inherent data structures and to identify model deficits. Bayesian multi-level modelling (BMM) has been widely applied in ecological modelling (Clark, 2003; Li et al., 2015; Thomas et al., 2017; Tian et al., 2020), and has more recently been applied to plant models. For example, Patrick et al. (2009) applied a hierarchical Bayesian approach to estimate parameters of the Farquhar photosynthesis model. Jarquín et al. (2016) used a hierarchical Bayesian formulation of a linear-bilinear model to investigate genotype × environment (G × E) interactions of maize from breeding trials. Fer et al. (2021) applied hierarchical Bayes to a dynamic vegetation model in conjunction with a Bayesian model emulator. Senf et al. (2017) applied Bayesian hierarchical modelling to a satellite-based data-driven phenology model to account for spatial and temporal variation in phenology. Qiu et al. (2020) developed a Bayesian hierarchical space–time model to study the impact of climate change and extreme events on phenological development. To the best of our knowledge, the BMM approach has not been applied to calibrate a process-based phenology model on a regional scale. By applying the BMM approach, we can honour the hierarchical classification of cultivars nested within ripening groups of a crop species. Thus, species-, ripening group-, and cultivar-specific parameters can be simultaneously estimated (Van Oijen and Höglind, 2016). We can also account for phenological development that depends on additional environmental factors which are not already captured in the model equations (Del Giudice et al., 2013). Methods such as cross-calibration have been used to determine crop model parameters for representative crop cultivars grown in different agro-ecological sub-zones (Xiong et al., 2008). But parameter estimates from such an approach would have limited applications when the phenological development of a new cultivar belonging to a different ripening group is to be predicted. Parameter estimates from the BMM approach can be used for such applications. The advantage of BMM lies in its borrowing strength (Zhang and Arhonditsis, 2009), where parameter estimates for data-limited cultivars can benefit from data-rich ones. Additionally, the appropriate depiction of data groups results in a representative quantification of prediction uncertainty (Gelman, 2006a).

We tested the proposed approach by evaluating four BMM cases of increasing complexity in which we calibrated the SPASS (Wang, 1997; Wang and Engel, 1998, 2000) phenology model to observations of silage maize grown across Germany from 2009 to 2017. The SPASS model has proven to be successful for different crop species (Gayler et al., 2002; Priesack et al., 2006; Biernath et al., 2011), including maize, and was one of the well-performing models in the Agricultural Model Intercomparison and Improvement Project (AgMIP) studies (Bassu et al., 2014; Durand et al., 2018; Kimball et al., 2019). It works well for cultivar-specific crop simulations. However, crop species simulations suffer when ripening group and cultivar information is not incorporated during calibration, a problem that is not unique to the SPASS model. Furthermore, a previous study (Viswanathan et al., 2022b) highlighted possible environment-related deficits in the SPASS phenology model which needed systematic evaluation. Thus, in the four BMM cases, we accounted for different combinations of yearly variability, environmental effects arising from growth in different ecological regions and weather conditions, and the classification of cultivars into ripening groups. With these cases we assessed the importance of including cultivar information in regional calibration studies. We evaluated the BMM approach by comparing calibration results from the four cases with the commonly used *pooled* approach, where a set of model parameters was estimated for silage maize grown across all environments. We also analysed trends between environmental effect parameters and environmental variables to identify possible model deficits. The findings of our study are expected to have implications for regional calibration and model improvement studies.

4.2 Materials and Methods

4.2.1 Data

We used phenology observations of silage maize grown between 2009 and 2017 at locations across Germany, collected by the German National Meteorological Service (Deutsche Wetterdienst-DWD) (DWD Climate Data Center (CDC), 2019). The observers reported the date of the first detected occurrence of maize phenological development stages, namely, 10, 31, 53, 61, 75, 83, and 87 on the BBCH scale (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) (Meier, 2018), in the assigned observation area. Corresponding cultivars and ripening groups were also reported. We refer to data sets by the site and year in which silage maize was cultivated, that is, “site-year”. We removed site-years for which sowing and harvest dates were not reported and in which the BBCH data was not strictly monotonically increasing with time. Additionally, observations that fell outside the range of sowing and harvest dates were discarded. We note that not all of the seven above-mentioned phenological stages were available in all site-years.

Minimum and maximum daily air temperatures were used as inputs to the SPASS phenology model. Weather data from the DWD stations were not available at all plant observation sites. Therefore, temperature data were extracted at all sites from the ERA5-Land re-analysis gridded data set (Muñoz Sabater, 2019). This data set has a spatial resolution of 0.1°× 0.1° and hourly temporal resolution. The hourly data were aggregated to daily values. We consider the re-analysis data to be a better spatial representation than point measurements at the weather stations.

To assess model limitations related to temperature and precipitation, site-years were classified into ten weather classes based on average temperature and cumulative precipitation between April and June, and between July and September (periods during which maize usually undergoes vegetative and reproductive development, respectively). A K-means clustering algorithm was applied to define the weather classes (details in Appendix A. Weather class clustering). Site-years were also grouped into nine ecological regions based on the classification provided by the BfN (Bundesamt für Naturschutz) (2017). For computational reasons, a subset of 100 site-years out of 3004 was randomly selected for calibration where it was ensured that at least one site-year was selected from each of the four ripening groups, nine ecological regions, ten weather classes and nine years. The calibration data set consisted of 66 cultivars from the four ripening groups (Table 4.1). It was also ensured that the relative proportions of site-years from the different ripening groups in the full data set were maintained in the calibration subset (early: 34%, mid-early: 54%, mid-late: 11% and late: 0.4% in the full data set). The 100 site-years used for calibration contained 604 phenology observations.

Table 4.1: Summary of site-years and cultivars used for calibration

ripening group	**early**	**mid-early**	**mid-late**	**late**	**Total**
number of cultivars	25	33	7	1	66
number of site-years	35	55	9	1	100

4.2.2 Phenology model

Air temperature, site-latitude, sowing and harvest dates are required as inputs to the SPASS phenology model. The model has nine parameters, seven of which were estimated during calibration (Table 4.2) and the remaining two were fixed at default values. The model equations and details are given in

Appendix B. SPASS model equations. We provide a brief summary below.

Three main phases of development are defined in the model: emergence, vegetative and reproductive phases. Emergence is dependent on the sowing depth, assumed to be fixed for all the site-years at 3 cm, and on the temperature above a minimum value (*emt*). The development rate during the vegetative and reproductive phases is dependent on the number of physiological development days at optimum temperature (*pdd1* for vegetative and *pdd2* for reproductive) and on the Temperature Response Function (TRF). The TRF is defined by phase-specific minimum, optimum and maximum cardinal temperatures (*tminv*, *toptv* and *tmaxv*, respectively, for vegetative and *tminr*, *toptr* and *tmaxr*, respectively for reproductive). The SPASS phenology model as described in Wang (1997) was implemented in our study with the following modifications: (a) the photoperiod effect on the vegetative phase was not considered, (b) no soil water-limiting effect on germination was assumed and germination occurs instantaneously after sowing, and (c) for numerical reasons the transition between emergence and vegetative phases was defined by a sigmoidal function instead of the original step function.

The parameters for physiological development days at optimum temperature for the vegetative (*pdd1*) and reproductive phases (*pdd2*) were estimated in the study. Additionally, minimum and optimum temperature for vegetative (*tminv* and *toptv*, where $toptv = tmaxv - dtoptv$) and reproductive (*tminr* and *toptr*, where $toptr = tmaxr - dtoptr$) phases, as well as the minimum temperature required for emergence (*emt*), were estimated. However, the parameters *tmaxv* and *tmaxr* were not estimated. The range of average daily temperatures during the growing season at the study site-years were between -6 and 31 °C. This is usually expected to be at or lower than the optimal temperatures for maize, a warm-weather plant. The lack of observations in the supra-optimal temperature range would make constraining *tmax* difficult (Wang et al., 2015) and is expected to incur problems of equifinality. To avoid these problems, the values of *tmaxv* and *tmaxr* were fixed at 44 °C.

We used Bayesian inference to determine the posterior probability of the model parameters. Let ϕ_d represent the given phenology observation on day d. The phenology $\bar{\phi}_d(\boldsymbol{T}, \boldsymbol{\theta})$ at day d, simulated by the SPASS model with a parameter vector $\boldsymbol{\theta}$, is dependent on air temperatures $\boldsymbol{T}$ from the date of germination to the day d. The phenology observations are available at days D, so the parameters are conditioned on $\Phi = \{\phi_d; d \in D\}$ through the likelihood function $p(\Phi \mid \boldsymbol{\theta}, T)$. The posterior parameter distribution is given by

$$p(\boldsymbol{\theta} \mid \Phi, \boldsymbol{T}) \propto p(\Phi \mid \boldsymbol{\theta}, \boldsymbol{T})\, p(\boldsymbol{\theta}) \tag{4.1}$$

where $p(\boldsymbol{\theta})$ is the joint prior probability distribution of the parameters.

Table 4.2: Prior distributions for estimated parameters in the process-based SPASS phenology model and the multi-level model cases. The mean and standard deviation (SD) are specified for normal distributions while the minimum (min.) and maximum (max.) are specified for uniform distributions.

Parameter	**unit**	**Description**	**Distribution**	**mean/ min.**	**SD/ max.**
emt	°C	Base temperature for emergence	Normal	10	1
pdd1	day	Physiological development days-vegetative phase	Normal	45	7
tminv	°C	Minimum temperature- vegetative phase	Normal	6	0.7[a]
dtoptv	°C	Difference between maximum and optimum temperature - vegetative phase	Normal	10	1.5
pdd2	day	Physiological development days - reproductive phase	Normal	36	10
tminr	°C	Minimum temperature - reproductive phase	Normal	8	1[b]
dtoptr	°C	Difference between maximum and optimum temperature - reproductive phase	Normal	10	1.5
σ	BBCH	Standard deviation of model residual error	Uniform	0	10
δ_w	BBCH	Weather effect by weather class w	Normal	0	5
γ_e	BBCH	Eco-region effect by eco-region e	Normal	0	5
λ	BBCH	Standard deviation of year effects τ_y	Uniform	0	5

[a] SD = 1.5 for $\Delta\theta_r$ and $\Delta\theta_{r,c}$; [b] SD = 2 for $\Delta\theta_r$ and $\Delta\theta_{r,c}$

4.2.3 Bayesian model cases

We describe five Bayesian model cases (one pooled and four multi-level model cases as seen in Fig. 4.1) in terms of their likelihood functions in the following sections. The prior distributions for all the estimated parameters are provided in Table 4.2.

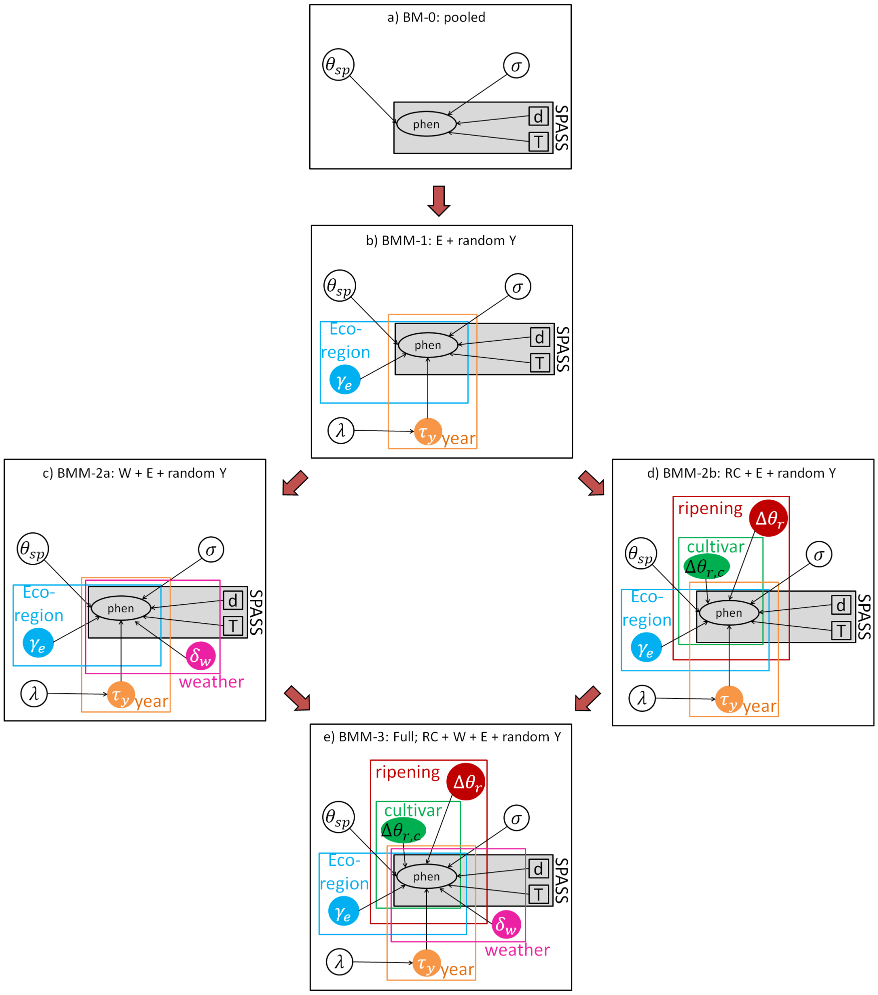


Figure 4.1: Graphical representation of the five Bayesian models: (a) BM-0: pooled, (b) BMM-1: eco-region effects with random year effects, (c) BMM-2a: weather and eco-regions effects with random year effects, (d) BMM-2b: hierarchical classification of cultivars into ripening groups with eco-regions effects and random year effects, and (e) BMM-3 or full model: hierarchical classification of cultivars into ripening groups, eco-region and weather effects with random year effects. In the SPASS phenology model, phenological development on a given day (d) is a function of air temperatures ($\boldsymbol{T}$) from the date of germination to that day. $\boldsymbol{\theta}_{sp}$ is the maize species-level parameter vector, $\Delta\boldsymbol{\theta}_r$ is the difference between the ripening group-level parameter $\boldsymbol{\theta}_{sp,r}$ and $\boldsymbol{\theta}_{sp}$, $\Delta\boldsymbol{\theta}_{r,c}$ is the difference between the cultivar-level parameter $\boldsymbol{\theta}_{sp,r,c}$ and $\boldsymbol{\theta}_{sp,r}$, and σ is the standard deviation of the likelihood function. γ_e represents the eco-region effect, δ_w, the weather class effect, τ_y the year effect and λ is the standard deviation of the year effect. E=eco-regions, Y = year, W = weather class, R = ripening group, C = cultivar. The red arrows outside the model sketches represent model extensions. During calibration, $\boldsymbol{\theta}_{sp,r}$ is estimated for the 4 ripening groups, $\boldsymbol{\theta}_{sp,r,c}$ for 66 cultivars, δ_w for 10 weather classes, γ_e for 9 eco-regions, and τ_y for 9 years.

4.2.3.1 BM-0: Pooled model

The pooled model is the most commonly used calibration setup for regional scale studies, where a common parameter set is estimated for all cultivars grown in different environmental conditions in the region. Assuming independent Gaussian observation errors, the likelihood function for the pooled model is given by

$$p(\Phi \mid \boldsymbol{\theta}, \boldsymbol{T}) = \prod_{d \in D} \mathcal{N}(\bar{\phi}_d(\boldsymbol{\theta}_{sp}, \boldsymbol{T}),\, \sigma^2) \tag{4.2}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{sp}, \sigma\}$, $\boldsymbol{\theta}_{sp}$ represents the maize species (sp) parameters, and $\mathcal{N}(\bar{\phi}_d(\boldsymbol{\theta}_{sp}, \boldsymbol{T}), \sigma^2)$ the density of a normal distribution with mean equal to the simulated phenology $\bar{\phi}_d$ and standard deviation σ. The pooled model case is shown in Fig. 4.1a. The joint prior probability $p(\boldsymbol{\theta}_{sp}) \sim \mathcal{N}(\boldsymbol{\mu}_{sp}, \boldsymbol{\Sigma}_{sp})$ was represented by a multivariate normal distribution with seven dimensions corresponding to the SPASS model parameters (*emt, pdd1, tminv, dtoptv, pdd2, tminr, dtoptr*). The mean vector of the distribution ($\boldsymbol{\mu}_{sp}$) and main diagonal elements of the variance–covariance matrix ($\boldsymbol{\Sigma}_{sp}$) are defined in Table 4.2 (mean and squared standard deviation, respectively) while the off-diagonal elements are zero.

4.2.3.2 BMM-1: Fixed eco-region effects and random year effects

We expect that the different eco-regions and years in which silage maize was grown in Germany influence phenology. We analysed this effect with the BMM-1 model (Fig. 4.1b), where we accounted for fixed effects due to the different eco-regions and random effects arising from variability between the years. If the different eco-regions and years are represented by $e \in E$ and $y \in Y$, respectively, then

$$p(\Phi \mid \boldsymbol{\theta}, \boldsymbol{T}) = \prod_{e \in E} \prod_{y \in Y} \prod_{d \in D} \mathcal{N}(\bar{\phi}_d(\boldsymbol{\theta}_{sp}, \boldsymbol{T}) + \gamma_e + \tau_y,\, \sigma^2) \tag{4.3}$$

where parameters γ_e and $\tau_y \sim \mathcal{N}(0, \lambda)$ represent the effects by eco-region e and year y, respectively, $E = 9$ is the total number of eco-regions, $Y = 9$ is the total number of years, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{sp}, \gamma_e, \lambda, \sigma\}$. A uniform prior density was assumed for the standard deviation of the year effects λ as per Gelman (2006b) (Table 4.2).

4.2.3.3 BMM-2a: Fixed eco-region and weather effects and random year effects

Although the SPASS model accounts for the effect of temperature on phenological development, there could be other weather conditions (i.e. temperature and precipitation during specific phases) that are important but not adequately captured in the model. Also, differences in weather conditions could result in a perceived variability between eco-regions and between years. In the BMM-2a model (Fig. 4.1c), we additionally accounted for the effects due to the different weather classes. If the different weather classes are represented by $w \in W$ and parameter δ_w represents the effects by weather class, then

$$p(\Phi \mid \boldsymbol{\theta}, \boldsymbol{T}) = \prod_{w \in W} \prod_{e \in E} \prod_{y \in Y} \prod_{d \in D} \mathcal{N}(\bar{\phi}_d(\boldsymbol{\theta}_{sp}, \boldsymbol{T}) + \gamma_e + \tau_y + \delta_w,\, \sigma^2) \tag{4.4}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{sp}, \delta_w, \gamma_e, \lambda, \sigma\}$ and $W = 10$ is the total number of weather classes.

4.2.3.4 BMM-2b: Fixed ripening, cultivar, eco-region effects and random year effects

As a modification from BMM-1, we also accounted for the inherent structure in the data in BMM-2b (Fig. 4.1d) wherein the cultivars c, are nested within ripening groups r of the maize species sp.

$$p(\Phi \mid \boldsymbol{\theta}, \boldsymbol{T}) = \prod_{e \in E} \prod_{y \in Y} \prod_{d \in D} \mathcal{N}(\bar{\phi}_d(\boldsymbol{\theta}_{sp,r,c}, \boldsymbol{T}) + \gamma_e + \tau_y,\, \sigma^2) \tag{4.5}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{sp,r,c}, \gamma_e, \lambda, \sigma\}$ and $\boldsymbol{\theta}_{sp,r,c}$ represents the joint probability distribution of all the estimated cultivar-level parameters in the hierarchy. It can be expressed as $\boldsymbol{\theta}_{sp,r,c} = \boldsymbol{\theta}_{sp} + \Delta\boldsymbol{\theta}_r + \Delta\boldsymbol{\theta}_{r,c}$, where $\Delta\boldsymbol{\theta}_r = \boldsymbol{\theta}_{sp,r} - \boldsymbol{\theta}_{sp}$ is the difference between the species-level parameters ($\boldsymbol{\theta}_{sp}$) and the ripening group-level parameters ($\boldsymbol{\theta}_{sp,r}$), and $\Delta\boldsymbol{\theta}_{r,c} = \boldsymbol{\theta}_{sp,r,c} - \boldsymbol{\theta}_{sp,r}$ is the difference between the ripening group-level and the cultivar-level parameters. Thus, the SPASS model parameters corresponding to 66 cultivars (cultivar-level) and 4 ripening groups (ripening group-level) in the calibration data set are estimated. Their prior probability $p(\Delta\boldsymbol{\theta}_r) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_r)$ and $p(\Delta\boldsymbol{\theta}_{r,c}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{r,c})$ were represented by multivariate normal distributions, centred at zero. Their variance–covariance matrices ($\boldsymbol{\Sigma}_r$, $\boldsymbol{\Sigma}_{r,c}$) were equivalent to that of the species-level prior $\boldsymbol{\Sigma}_{sp}$ for all parameters except *tminv* and *tminr* (note different standard deviation in the footnote of Table 4.2).

4.2.3.5 BMM-3: Full model

Finally, in the full model (Fig. 4.1e), we accounted for the inherent hierarchical data structure, eco-regions and weather effects, as well as year effects.

$$p(\Phi \mid \boldsymbol{\theta}, \boldsymbol{T}) = \prod_{w \in W} \prod_{e \in E} \prod_{y \in Y} \prod_{d \in D} \mathcal{N}(\bar{\phi}_d(\boldsymbol{\theta}_{sp,r,c}, \boldsymbol{T}) + \gamma_e + \tau_y + \delta_w,\, \sigma^2) \tag{4.6}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{sp,r,c}, \delta_w, \gamma_e, \lambda, \sigma\}$ and $\boldsymbol{\theta}_{sp,r,c} = \boldsymbol{\theta}_{sp} + \Delta\boldsymbol{\theta}_r + \Delta\boldsymbol{\theta}_{r,c}$.

4.2.4 Posterior sampling

Markov Chain Monte Carlo sampling of the posterior parameter distributions was performed using the Gibbs algorithm from the Jags software (Plummer, 2003) implementation in R2jags (Su and Yajima, 2020) and jagsUI (Kellner, 2021) packages in R (R Core Team, 2020). For the model cases BM-0, BMM-1, BMM-2a and BMM-2b, 500 runs were used for adaptation. Three chains were run and 5000 iterations were run per chain until the Gelman Rubin convergence diagnostic was ≤ 1.1. Of these iterations, every 5th parameter vector (thinning = 5) was stored, resulting in a total of 3000 samples that were used for generating the posterior parameter distributions and simulated phenology described in the results. For BMM-3, 100 runs were used for adaptation. Three chains were run and 3600 iterations per chain were run until the Gelman Rubin convergence diagnostic was ≤ 1.1. All the samples (total of 10,800) were used for the plots. Diagnostic plots for the MCMC samples are provided in Supplement S6.

4.2.5 Calibration performance metrics

Calibration quality was assessed for each of the 100 site-years used for calibration and for each of the five model cases by estimating the expected value ($\pi(\bar{\Phi})$) of a loss function $L(\Phi, \bar{\Phi})$ where $\bar{\Phi} = \{\bar{\phi_d}; d \in D\}$ is a vector of phenological development simulated by the model.

$$\pi(\bar{\Phi}) = \int_{\boldsymbol{\theta}} L(\Phi, \bar{\Phi}) p(\boldsymbol{\theta} \mid \Phi, T) d\boldsymbol{\theta} \tag{4.7}$$

where $L(\Phi, \bar{\Phi})$ is either RMSE $= \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\phi_d - \bar{\phi_d})^2}$ or bias $= \frac{1}{D}\sum_{d=1}^{D}(\phi_d - \bar{\phi_d})$.

4.3 Results

We first provide results from the classification of 3004 site-years into environmental classes (section 4.3.1). We then describe the calibration quality of the SPASS model in the different Bayesian model cases (section 4.3.2), followed by an analysis of the posterior distributions of the SPASS model parameters (section 4.3.3), the environmental effects (section 4.3.4) and residual uncertainty (section 4.3.5). All figures were made using the ggplot2 (Wickham, 2016) package in R. We note here again that out of the 3004 site-years, 100 were used for calibration.

4.3.1 Classification of site-years into environmental classes

All site-years were classified into ten weather classes (Fig. 4.2). The weather classes were based on the average temperatures and cumulative precipitation between April and June and between July and September. Silage maize cultivated across Germany generally undergoes vegetative development from April to June and reproductive development from July to September. Phenological development during the vegetative and reproductive phases are dependent on temperature. The relationship between temperature and phenological development is usually represented by equations in phenological models, including SPASS. However, existing model equations may not accurately capture this temperature response. Additionally, the influence of factors like precipitation that are known to influence phenology in some plant species (Moore and Lauenroth, 2017), could also have either a direct influence on maize phenology or an indirect effect by influencing temperatures within the crop canopy. However, these effects are not represented in the SPASS phenology model. Thus, site-years were classified into the weather classes to assess model limitations related to temperature and precipitation.

All site-years were also classified into nine ecological regions (Fig. 4.3). Eco-region 0 includes the Alpine foreland and foothills characterized by flysch and molasse deposits, as well as glacial moraines. This eco-region has experienced on average higher rainfall (average cumulative precipitation of ∼700 mm, based on site-years in the full data set) than the other eco-regions during the maize growing season

(April–September between 2009 and 2017). Eco-region 1 comprises the Swabian and Franconian Alb, and the Black Forest (average elevation ∼400 m a.s.l.). It is marked by sedimentary deposits and the development of loess and loamy soils. The Rhine river plain is in eco-region 2. Eco-regions 3 and 4 consist of the middle highlands (average elevation of 300–400 m a.s.l.) of the Thuringian Forest and Harz Mountains. The region is characterized by sedimentary deposits with some metamorphic rocks and the development of loess loam. The remaining eco-regions consist of the northern lowlands, typified by moraine deposits. Eco-region 7 includes the Mecklenburg Lake District and consists of young moraine deposits. Eco-regions 6 and 8 largely consist of old moraine deposits. Eco-region 8 has been, on average, hotter and drier (average daily temperature of 15–16 °C and average cumulative precipitation of ∼500 mm) than the other eco-regions during the growing season (please refer to Figs.S7.2 and S7.3 in Supplement S7.2 for details). Note that since the 100 site-years used for calibration were randomly sampled, the calibration data set contained only a few site-years from ecological regions in the northeastern part of Germany.

4.3.2 Calibration quality

As an example, we compare observed and simulated phenology for silage maize grown at a site in the state of Bavaria in 2009 from the pooled model case (BM-0) and the full model case (BMM-3) (Fig. 4.4). The blue bands show the 5–95th percentile of simulated phenology that account for uncertainty from model parameters ($\boldsymbol{\theta}_{sp}$ for BM-0 in Fig. 4.4a and $\boldsymbol{\theta}_{sp,r,c}$ for BMM-3 in Fig. 4.4b) while the red bands additionally account for environmental effects ($\delta_w, \gamma_e, \tau_y$ in Fig. 4.4b). The grey bands show the 5–95th percentile of simulated phenology that additionally account for the unresolved residual error (σ). There is a reduction in bias in BMM-3 as compared to BM-0. Furthermore, there is an overall reduction in unresolved residual error in BMM-3 with the model parameters and environmental effects accounting for a large share of the error variance. We also note that the blue bands in Fig. 4.4a have collapsed around the mean simulated phenology.

The model performance represented by the mean RMSE and bias for the 100 calibration site-years (Fig. 4.5) improved with model complexity from the pooled model to the full model (BM-0, BMM-1, BMM-2a, BMM-2b, BMM-3). This is evident from the reduction in mean RMSE and shrinkage of the mean bias towards zero. In Fig. 4.6 the model performance from the five cases were analysed by ripening group and weather class. Across the plots, the two cases BMM-2b and BMM-3 that account for the ripening group-cultivar hierarchy generally exhibit a lower bias and RMSE as compared to the others. Across the four ripening groups (Fig. 4.6a), the mean bias is closer to zero with increasing model complexity, as seen in Fig. 4.5. A decrease in mean bias and RMSE occurs on the inclusion of cultivar-ripening group information through the hierarchy in BMM-2b and BMM-3. The single site-year from the late ripening cultivar included in calibration also exhibits a clear improvement in RMSE and bias. While the inclusion of weather effects (Fig. 4.6b) in the model cases BMM-2a and BMM-3 result in smaller mean RMSE and bias only in some weather classes, the inclusion of cultivar-ripening group hierarchy results in an improvement in most classes. Although the inclusion of eco-regions and year

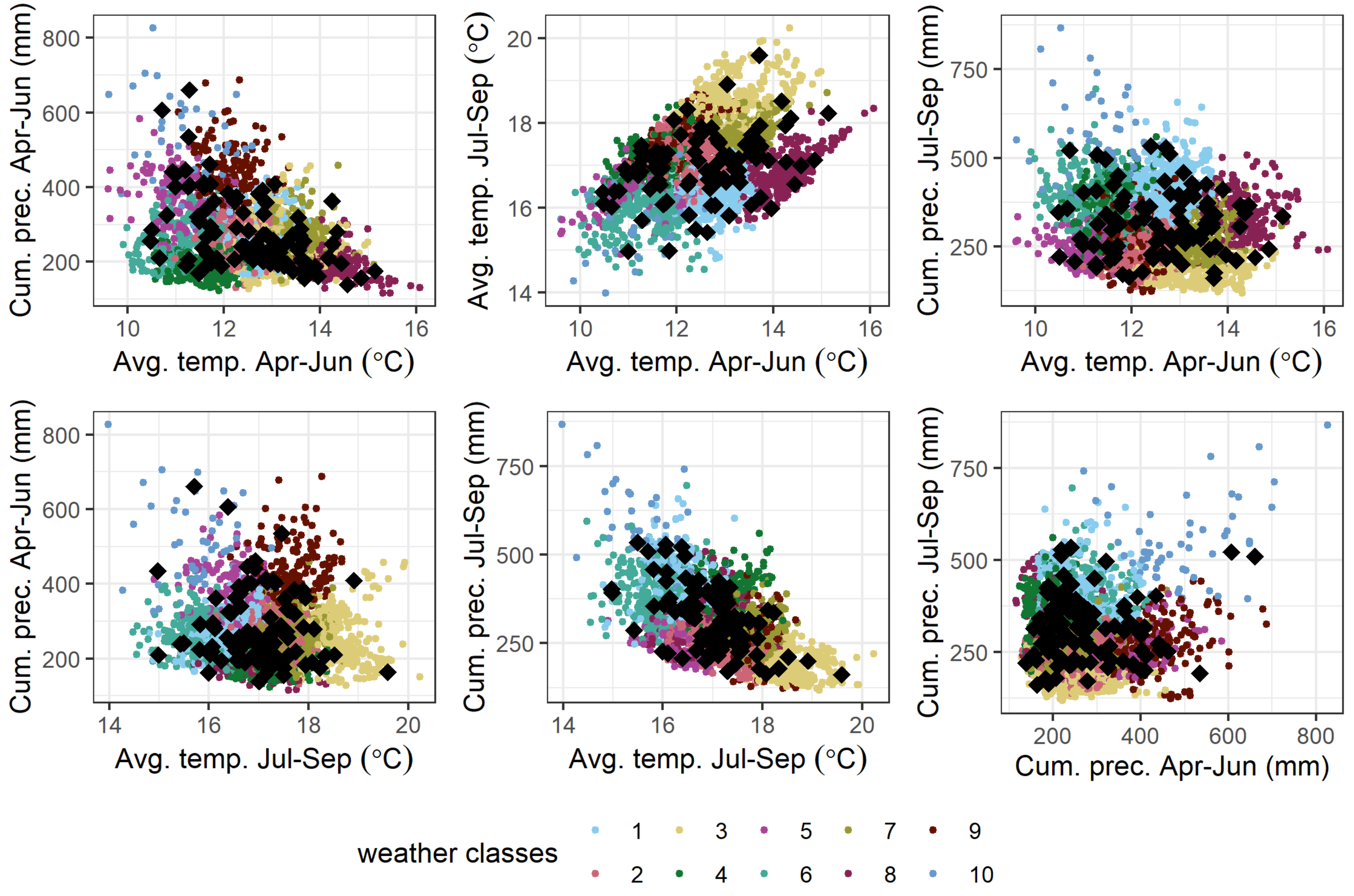


Figure 4.2: Site-years of silage maize grown across Germany were classified into ten weather classes based on average temperature (avg. temp.) and cumulative precipitation (cum. prec.) during April–June (Apr–Jun) and July-September (Jul–Sep). Black diamonds are the 100 site-year samples selected for calibration from the overall data set of site-years indicated by the coloured points. The colours indicate the ten identified weather classes.

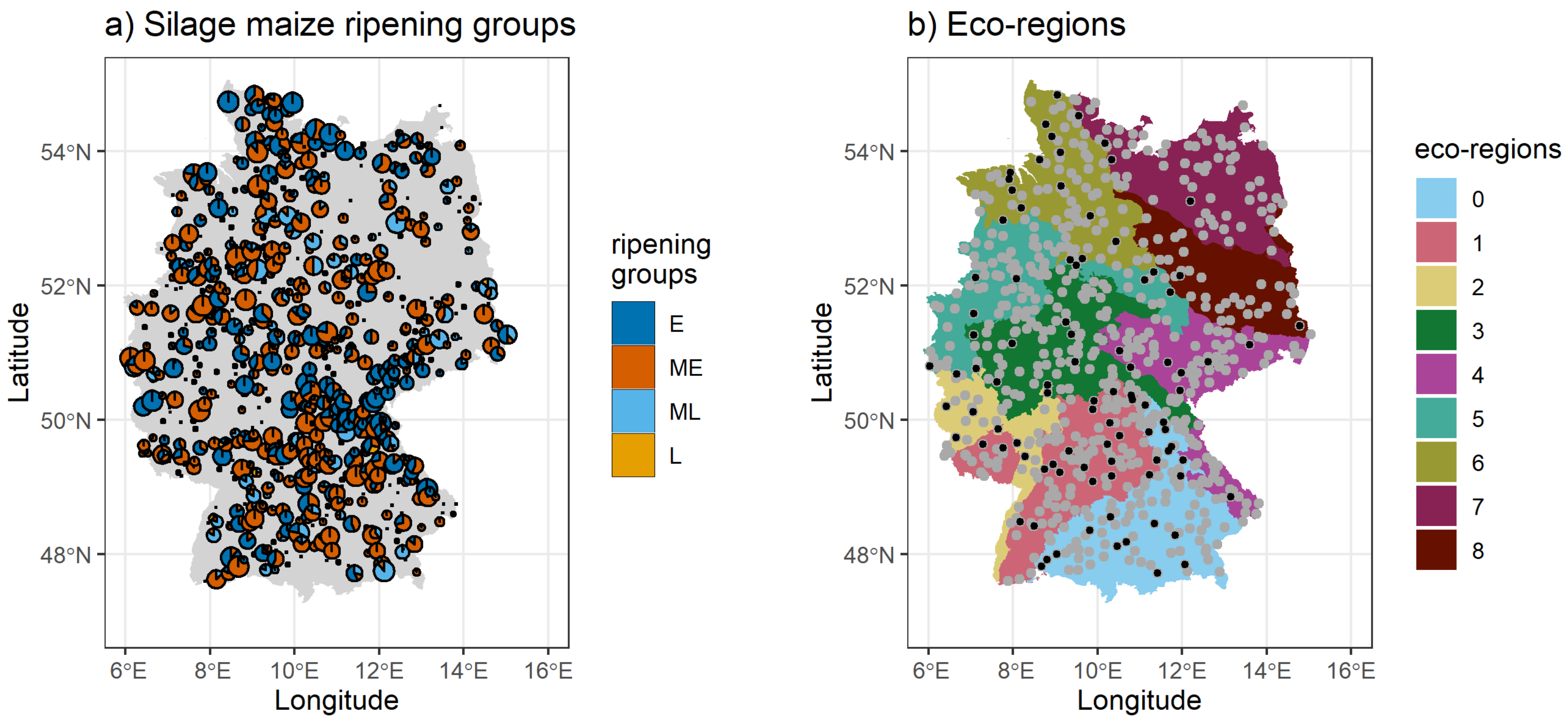


Figure 4.3: Ripening groups of silage maize cultivated in nine eco-regions across Germany from 2009 to 2017. (a) Pie-charts of silage maize ripening groups (E=early, ME=mid-early, ML=mid-late, L=late) grown at different sites in Germany. (b) Location of different sites (points) where silage maize was grown across the nine eco-regions (coloured polygons). The black points are site locations of the 100 site-year samples used for calibration. Projection system: DHDN 3 Degree Gauss Zone 3.

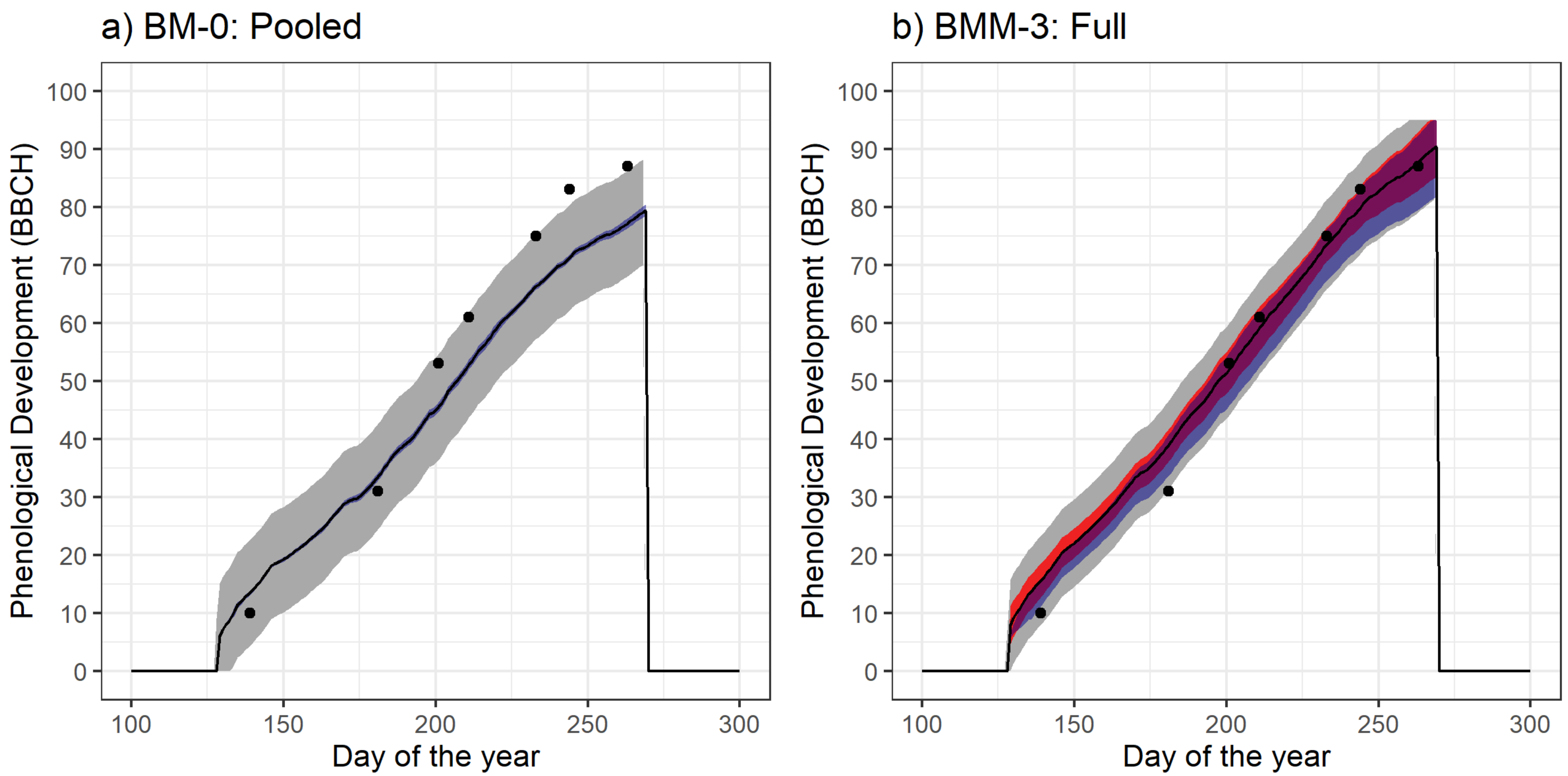


Figure 4.4: Observed and simulated phenology of silage maize (cultivar: Amatus) at Benediktbeuern in the state of Bavaria in 2009 in (a) the pooled model case (BM-0) and (b) full model case (BMM-3). The black points are the observations. The solid black line is the mean simulated phenology, while the coloured bands represent 5–95th percentile of simulated phenology: The blue bands represent the uncertainty originating from process model parameters, while the red band in (b) represents the uncertainty originating from process model parameters and environmental effects (weather, eco-region and year). The grey bands additionally account for the unresolved residual error.

effects (BMM-1 to BMM-3) (Fig. 4.C1 in the Appendix C. Model performance metrics) improves RMSE and bias in some eco-regions and years, a clear trend across all the classes cannot be identified.

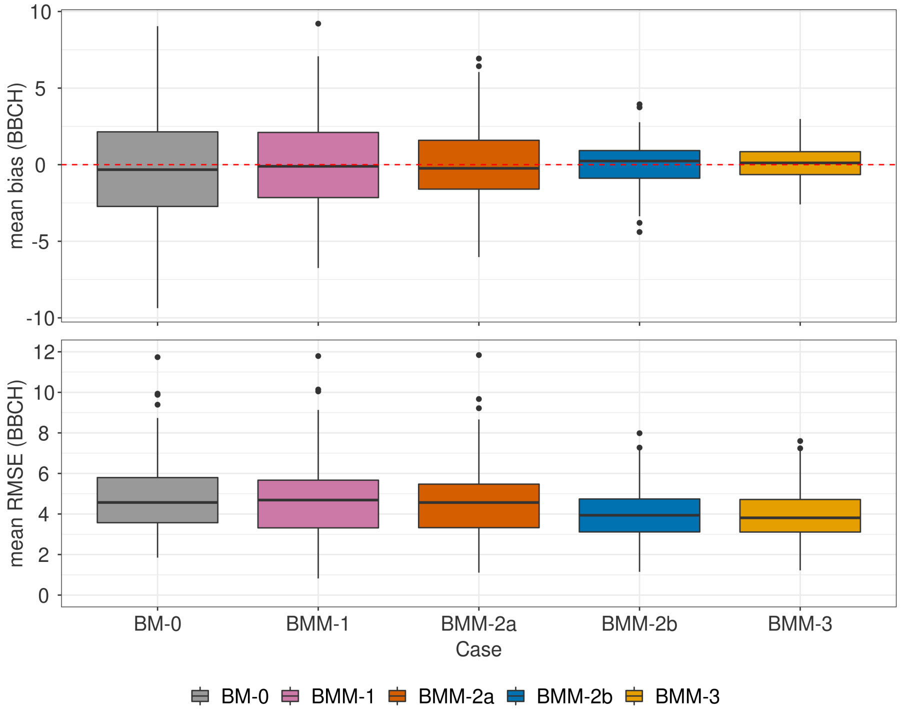


Figure 4.5: Box-plots of the mean RMSE and bias for each calibration site-year in the five model cases, BM-0 (pooled), BMM-1 (eco-regions, random year effects), BMM-2a (eco-regions, weather class, random year effects), BMM-2b (ripening group-cultivar hierarchy, eco-region, random year effects), BMM-3 (full model with cultivar-ripening group hierarchy, eco-regions, weather class, and random year effects). Each box-plot represents the 100 site-years used for calibration. Hinges of the box-plot represent the inter-quartile range (IQR), whiskers extend from the hinges up to 1.5×IQR and values beyond this range are plotted as points. Note that the phenology simulations used to estimate RMSE and bias do not take the σ parameter uncertainty into account.

4.3.3 Phenology model parameters

The marginal posterior distributions of the SPASS model parameters were analysed for the full model (BMM-3) to investigate differences between cultivar, ripening group and maize species parameter estimates after the environmental effects are taken into account. Figure 4.7 shows the posterior parameter distribution by species ($\boldsymbol{\theta}_{sp}$), ripening groups ($\boldsymbol{\theta}_{sp,r} = \boldsymbol{\theta}_{sp} + \Delta\boldsymbol{\theta}_r$) and cultivars ($\boldsymbol{\theta}_{sp,r,c} = \boldsymbol{\theta}_{sp} + \Delta\boldsymbol{\theta}_r + \Delta\boldsymbol{\theta}_{r,c}$) for two parameters *tminv* and *pdd1*. Parameter *tminv* shows low variability between cultivars of the same ripening group while *pdd1* shows high variability. A similar visual inspection of other parameters showed that they could be classified into the categories of low (*tminv, tminr, toptr*) and high (*pdd1, pdd2, toptv, emt*) between-cultivar variability (Figs.4.D1, 4.D2 in Appendix D. SPASS model parameter distributions).

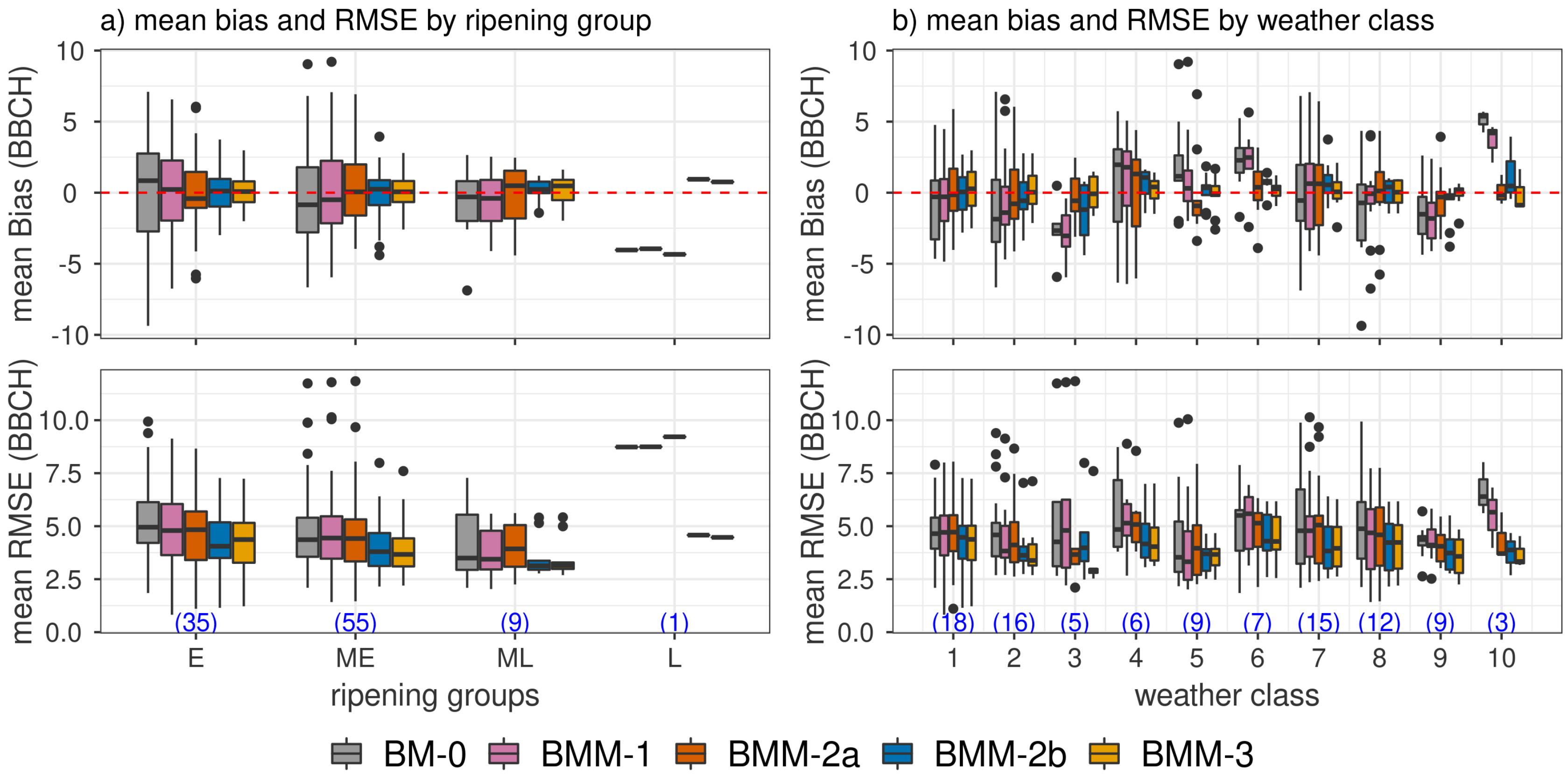


Figure 4.6: Box-plots of the mean RMSE and bias by site-year within ripening groups (a), weather class (b) for each of the five model cases, BM-0 (pooled), BMM-1 (eco-regions, random year effects), BMM-2a (eco-regions, weather class, random year effects), BMM-2b (ripening group-cultivar hierarchy, eco-region, random year effects), BMM-3 (full model with cultivar-ripening group hierarchy, eco-regions, weather class, and random year effects). The numbers in blue at the bottom of each plot indicate the number of site-years and consequently the number of points in the groups defined on the x-axis that were used to obtain the box-plots. Hinges of the box-plot represent the inter-quartile range (IQR), whiskers extend from the hinges up to 1.5×IQR and values beyond this range are plotted as points. Note that the phenology simulations used to estimate RMSE and bias do not take the σ parameter uncertainty into account.

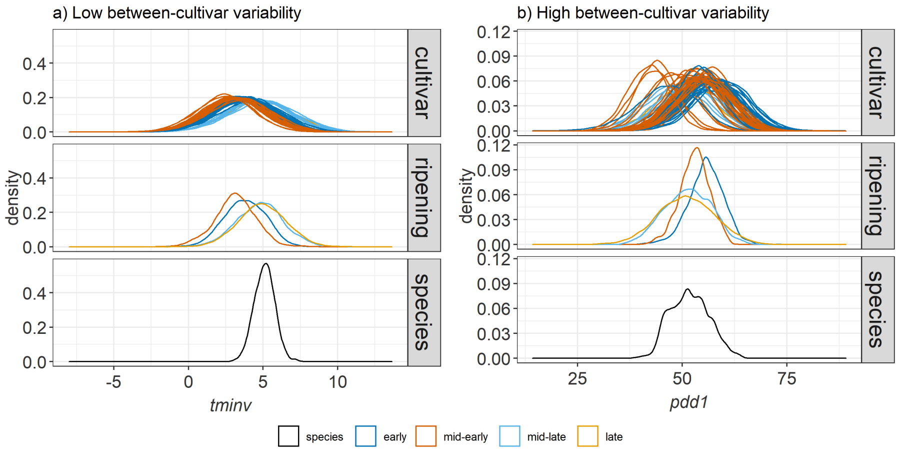


Figure 4.7: Posterior distribution of parameters in the full model case BMM-3: minimum temperature for development in the vegetative phase (*tminv*) and physiological development days for the vegetative phase at optimum temperature (*pdd1*). The distributions are provided for the species ($\boldsymbol{\theta}_{sp}$), ripening group ($\boldsymbol{\theta}_{sp,r}$) and cultivar ($\boldsymbol{\theta}_{sp,r,c}$) levels of the hierarchy. Colours of cultivar distributions correspond to their respective ripening groups.

4.3.4 Environmental effects

The prior and posterior parameter distributions of the weather effects (δ_w) and eco-region effects (γ_e) were analysed with respect to their corresponding classes in the four multi-level model cases (Fig.4.8). Negative effects indicate an overestimation, while positive effects indicate underestimation of phenology by the SPASS model and the remaining effects that were considered in the particular BMM case. This over/underestimation is corrected by the corresponding environmental effect parameter to improve the model's fit to the data in that BMM case. The posterior parameter distributions deviate from the prior which is normally distributed around zero and are narrower than the prior. The eco-region 2 (Fig. 4.8a) exhibits a negative effect in all model cases. Eco-regions 0 and 1 exhibit similar effects and so do eco-regions 3 and 4, in all model cases. The parameter distributions of the weather effects (Fig. 4.8b) are only shown for BMM-2a and BMM-3 since these effects are taken into account only in these two cases. Weather classes 6 and 10 have similar posterior parameter distributions for weather effects. These classes have different average cumulative precipitation but similar average temperatures (Supplement S7.3). This indicates that temperatures have a larger influence than precipitation on these weather effects.

To identify possible model deficits, we analysed trends between the median value of the weather effects parameters and mean of the average daily temperature and cumulative precipitation of the weather classes from April to June and July to September for the cases BMM-2a and BMM-3. A high correlation coefficient is seen between the median weather effect per class and the mean of the average daily temperature from July to September (Fig. 4.9). The correlation coefficient reduces from -0.87 in BMM-2a to -0.64 in the full model BMM-3 where the cultivar-ripening group hierarchy is considered. This is also accompanied by a widening in confidence intervals of the linear regression line.

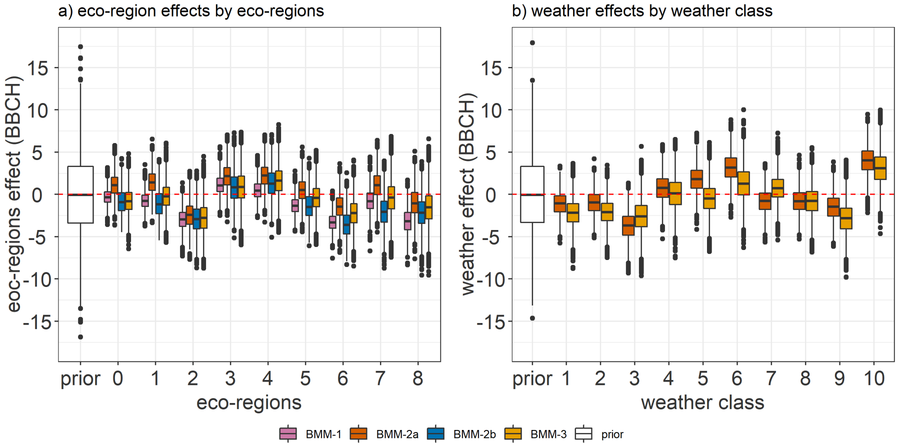


Figure 4.8: The prior and posterior parameter distributions of (a) eco-region effects for the nine eco-regions and (b) weather effects for the ten weather classes. The distributions of the eco-region effects (γ_e) and weather effects (δ_w) (y-axis) are plotted against their corresponding classes (x-axis) in the four Bayesian multi-level model cases. The Bayesian multi-level models are: BMM-1: eco-region with random year effects; BMM-2a: weather, eco-region with random year effects; BMM-2b: cultivar-ripening group hierarchy, eco-region and random year effects; BMM-3: cultivar-ripening group hierarchy, weather, eco-region random year effects. Hinges of the box-plot represent the inter-quartile range (IQR), whiskers extend from the hinges up to 1.5×IQR and values beyond this range are plotted as points.

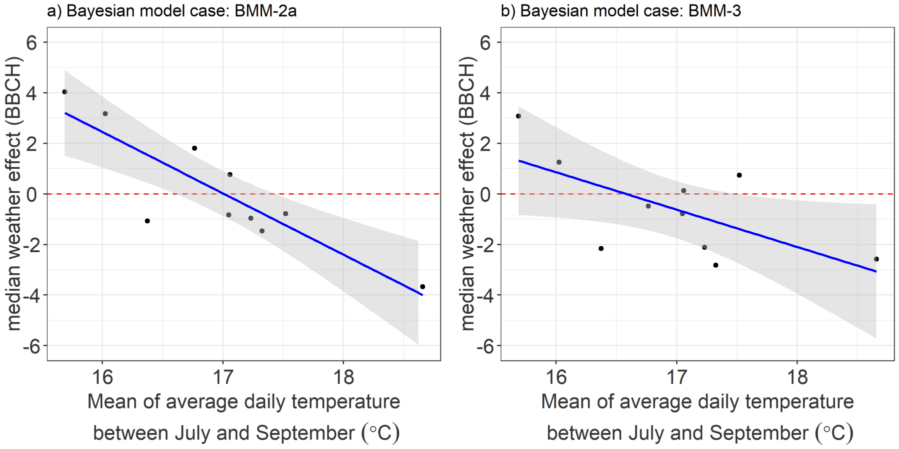


Figure 4.9: Median value of the weather effects parameters plotted against the mean of the average daily temperature between June and September for the 10 weather classes for the Bayesian model case (a) BMM-2a (eco-regions, weather class, random year effects) and (b) BMM-3 (full model with cultivar-ripening group hierarchy, eco-regions, weather class, and random year effects). The ten points correspond to the ten weather classes. The grey bands represent 95% confidence interval of the regression line.

The eco-region effect (γ_e) was included in all the Bayesian multi-level model cases. Figure 4.10 shows a comparison between the median eco-regions effects for the four multi-level models. A negative eco-region effect indicates an overestimation of phenology by the SPASS model and the other effects that were accounted for in the particular BMM case. This overestimation was corrected by the eco-region effect parameter. Conversely, a positive eco-region effect indicates an underestimation of phenology by the SPASS model and the other effects. Eco-regions 6 and 8 have similar median eco-region effects, and so do 3 and 4. Eco-regions 2, 6, and 8 show a negative eco-region effect while 3 and 4 show a positive effect irrespective of the model case. A comparison of BMM-1 (Fig. 4.10a) with BMM-2a (Fig. 4.10b) and BMM-2b (Fig. 4.10c) shows that the inclusion of weather effects (BMM-2a) results in a positive eco-region effect in most regions, while the inclusion of cultivar-ripening group hierarchy (BMM-2b) results in negative eco-region effects. However, this tendency is not seen when both weather effects and cultivar-ripening group hierarchy are included in BMM-3 (Fig. 4.10d).

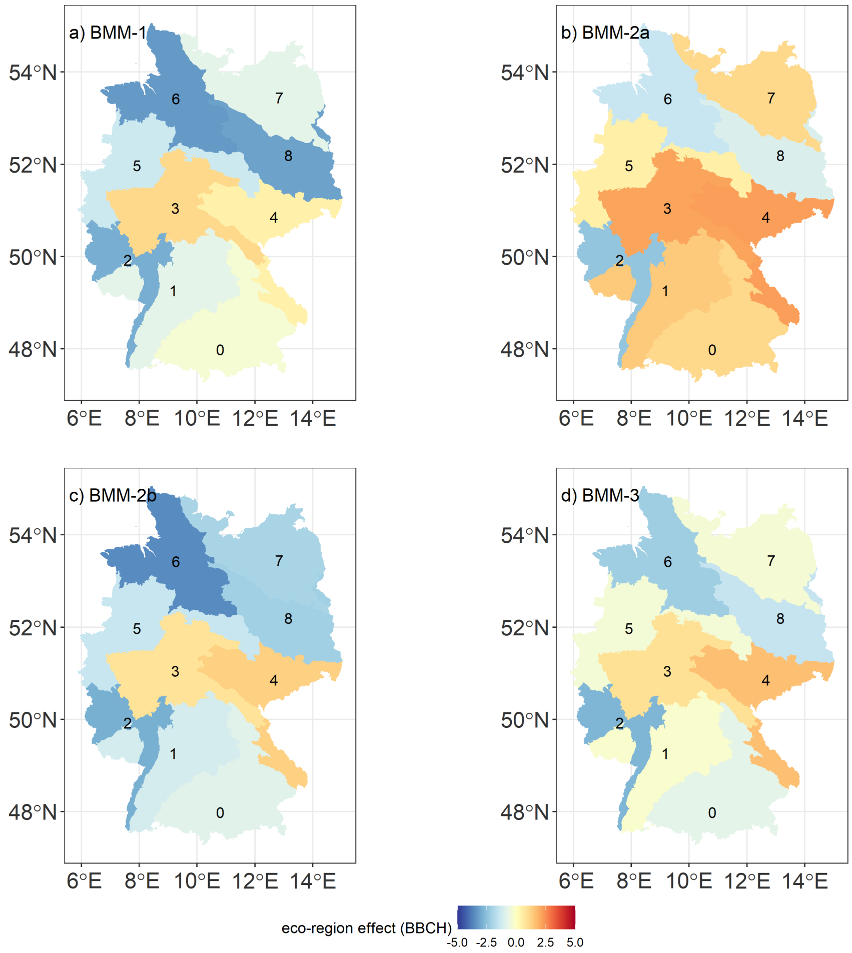


Figure 4.10: Median value of the eco-region effects parameters plotted for the nine eco-regions across Germany for the Bayesian multi-level models (a) BMM-1 (eco-regions, random year effects), (b) BMM-2a (eco-regions, weather class, random year effects), (c) BMM-2b (ripening group-cultivar hierarchy, eco-region, random year effects), and (d) BMM-3 (full model with cultivar-ripening group hierarchy, eco-regions, weather class, and random year effects). The numbers indicate the nine eco-regions. (Projection system: DHDN 3 Degree Gauss Zone 3).

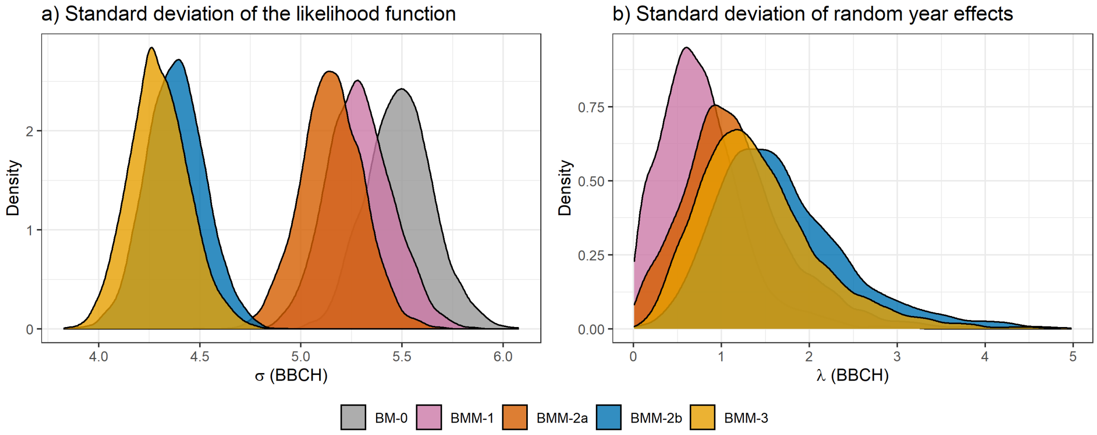


Figure 4.11: Posterior parameter distributions of (a) standard deviation of the likelihood function (σ) and (b) standard deviation of the random year effects (λ), for the Bayesian models (BM-0 to BMM-3), expressed in units of phenological development (BBCH).

4.3.5 Residual uncertainty

The standard deviation (σ) of the likelihood function and standard deviation (λ) of the year effect τ_y, represent the unresolved and resolved components of residual uncertainty, respectively. Figure 4.11a shows the posterior parameter distribution of σ from the five Bayesian models. There is a reduction in unresolved residual uncertainty with increasing model complexity from BMM-0 to BMM-3. A large reduction is seen on the inclusion of ripening-cultivar hierarchy in BMM-2b and BMM-3. The standard deviation of the year effect (λ) (Fig. 4.11b) shows an increase from BMM-1, BMM-2a to BMM-2b, followed by a slight decrease in BMM-3.

4.4 Discussion

The Bayesian multi-level models were able to improve calibration by partitioning some of the residual uncertainty into those arising from the hierarchical classification of cultivars and differences in eco-regions, weather conditions and year of growth. As seen in Fig. 4.4, the unresolved residual uncertainty in simulated phenology was smaller in the full multi-level model case (BMM-3 in Fig. 4.4b) than in the pooled case (BM-0 in Fig. 4.4a). This was also shown by the reduction in the standard deviation of the likelihood function (σ) in Fig. 4.11a. Moreover, the multi-level models reduced bias and RMSE (Fig. 4.5), especially in the cases of BMM-2b and BMM-3. The differences between different cultivars and ripening groups were taken into account by the hierarchical structure of cultivars nested within ripening groups in BMM-2b and BMM-3 (Fig. 4.6a). As a result, the uncertainty in simulated phenology (Fig. 4.4) originating from the process model parameters did not collapse as seen in the pooled case (BM-0), due to a collapse of the posterior parameter distribution (not shown). The over-confidence in the parameters of the pooled case can lead to poor predictions (Motavita et al., 2019) because most of the variability between site-years, which can be attributed to cultivar and ripening group differences, are attributed to random noise (σ). The additional inclusion of the weather effects (from BMM-2b to BMM-3) further improved calibration quality (Fig. 4.6b). The full multi-level model BMM-3 allowed for a more representative estimate of SPASS model parameter uncertainty. This was seen from the wider ranges (blue bands) of the resultant simulated phenology in Fig. 4.4b as compared to Fig. 4.4a.

Based on their between-cultivar variability, the posterior distributions of SPASS model parameters were grouped into: cultivar-specific (high variability) and ripening group-specific (low variability) parameters (Fig. 4.7). The parameters that exhibited low between-cultivar variability such as *tminv*, *tminr*, and *toptr* (Fig. 4.D1) define the Temperature Response Function (TRF). In general, cardinal temperatures are expected to be ripening group-specific, while parameters such as *pdd1* and *pdd2* (Fig. 4.D2), should be cultivar-specific because they represent traits that would be optimized in different cultivars by plant breeders (Parent et al., 2018; Zheng et al., 2012; Challinor et al., 2016). Most model parameters exhibited this expected behaviour, with the exception of *toptv*. The posterior distribution of parameter *toptv*, which influences vegetative development, may not only represent the ripening group-specific optimum temperature but may have also compensated for a missing cultivar-specific effect. We assumed no photoperiod effect on vegetative development in the model since this effect is small for maize grown in temperate regions (van Bussel et al., 2015). Nonetheless, such an effect could exist and be cultivar-dependent. This missing effect could have been compensated by *toptv* in the model. The base temperature for emergence *emt* also exhibited some between-cultivar variability. This parameter could have compensated for the effects of some cultivar-specific hard-coded parameters in the emergence equation (Eq. 4.B1). Furthermore, the ripening group-level distributions of parameter *tminv* (Fig. 4.7a) showed that, as expected, early/mid-early ripening groups on average have a lower minimum temperature requirement for vegetative development than the mid-late/late ripening groups. Overall, the match between the behaviour of calibrated parameters and theoretical expectation highlights the model's robustness if weather and eco-region effects are also accounted for through the multi-level modelling approach.

Analyses of the weather effects parameters highlight model deficits related to temperature effects during reproductive development. Firstly, the non-zero values of the posterior parameter distributions indicate that weather effects are influential (Fig. 4.8). Also, the narrow posterior distributions as compared to the prior show that the parameter values are informed by the observations. Furthermore, there was a high correlation between median weather effects and average reproductive phase temperature by weather class (Fig. 4.9). These weather effects indicate that the model overestimates phenological development at higher late summer temperatures and underestimates at lower temperatures. This trend reduced with the introduction of the cultivar-ripening group hierarchy. Certain cultivars may have been selected by the farmers based on their performance in the local environmental conditions such as temperature (Siebert and Ewert, 2012; Parker et al., 2017; Parent et al., 2018). Introducing the ripening groups as a level in the hierarchy could account for the differences between the groups in reaching maturity. This differentiation of ripening groups becomes evident in the reproductive phase of development during the late summer months. In the absence of cultivar-ripening group hierarchy in BMM-2a, the weather effects captured these differences, as shown in Fig. 4.9a. This has implications in cases where ripening group and cultivar information are not available. Regional data sets, such as those from the state office of agricultural statistics, may not contain information about the cultivars grown, but information about environmental conditions are usually more readily available. In the absence of ripening group and cultivar information, accounting for weather effects will still result in some degree of improvement in model calibration quality. Although weakened in BMM-3, the weather effect-temperature trend was not completely removed (Fig. 4.9b). This indicates that model deficits related to temperature effects persist in the reproductive phase. Thus, accounting for weather effects in addition to the hierarchy in BMM-3 resulted in improved site-year calibration quality in many weather classes as compared to BMM-2b (Fig. 4.6b). The current TRF in the reproductive phase may not sufficiently capture the true development behaviour. Different TRFs for maize phenology should be evaluated using this approach with the aim of identifying a better representation of the underlying processes.

Analyses of eco-region effects parameters point to model deficits related to soil moisture. Since eco-region effects were accounted for in all the four multi-level model cases, we first provide a detailed discussion of the results here. The median values of the eco-region effects exhibited a positive increase in many regions on the inclusion of weather effects (Fig. 4.10b), and a negative increase on the inclusion of cultivar-ripening group hierarchy (Fig. 4.10c) as compared to the BMM-1 (Fig. 4.10a). Accounting for either one of these factors (weather effects or cultivar-ripening group hierarchy) without the other possibly resulted in the eco-region effects compensating for these missing factors. The neighbouring eco-regions 6 and 8 (northern Lowlands), and 2 (Rhine plain), as well as 3 and 4 (central Uplands) exhibited similar trends in eco-region effects, irrespective of the model case. Eco-regions 6 and 8 and eco-regions 3 and 4 can be further grouped, thus allowing for model simplification. The distinct differences between the Lowlands, the Rhine plains (negative eco-region effect), and the Uplands (positive eco-region effect) could be due to distinct climatological or pedological features. The northern Lowlands are characterized by moraines that have high groundwater levels. The Rhine plain also exhibits higher groundwater productivity than the central Uplands (Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), 1993). In the full model (BMM-3), phenological development was overestimated in the northern Lowlands and Rhine plain, even after the process-model output was corrected for weather and year effects. This overestimation was then corrected by the negative eco-region effects in these regions. This could indicate a possible influence of water-logging which slows phenological development and is not accounted for in the SPASS model. Liu et al. (2021) found that accounting for water-logging stress on phenology and yield in the APSIM model improved model performance for barley. A similar consideration in the SPASS model equations may be required to account for this stress. This effect could also occur because high soil–water content or water-logged soils might lower temperatures within the crop canopy through the cooling effect of evaporation, changes in albedo at the soil surface, enhanced soil heat capacity, and heat transport resulting in a heat transfer away from the soil surface. None of these effects are accounted for in the model. An alternate formulation of the multi-level model that separately accounts for factors like soil, climate and topography instead of eco-regions is suggested to aid further investigation. By separating the effects of ecological regions from the process model parameters we were able to gain further insights into the model and identify possible model deficits.

As is expected, the unresolved model residual error represented by the standard deviation of the likelihood function reduced with increasing model complexity. A large reduction was seen due to the inclusion of the ripening group-cultivar hierarchy. This could have been a consequence of including a large number of parameters corresponding to 4 ripening groups and 66 cultivars in models BMM-2b and BMM-3. The other nuisance parameter, the standard deviation of the random year effects, increased with increasing model complexity from BMM-1, BMM-2a to BMM-2b, followed by a slight reduction in BMM-3. Accounting for other effects resulted in a better estimation of the random year effects that were previously attributed to the unresolved residual error.

While Bayesian multi-level modelling improved calibration, we acknowledge the limitations of our approach in using limited data and excluding input uncertainty. Although the inclusion of only eco-region, weather and year effects do account for some improvement in model performance (Fig. 4.5), this was not clearly evident when model performance was analysed by eco-region and weather classes (Fig. 4.C1). These effects may be convoluted by a possibly stronger effect of cultivar-ripening group hierarchy, through compensation when it is not explicitly taken into account. Although, more observation data would help in disentangling the different effects, they may not be available. Additionally, as noted in this study, using more data is also accompanied by high computation costs, especially for modelling cases with higher complexity. Importantly, the uncertainty in phenology observations and in the model inputs like temperature or reported sowing and harvest dates were not considered and could further confound results. Also, since the observations were made for fixed BBCH stages, we recommend that the model and likelihood function should be reformulated to represent the uncertainty in the days rather than phenology stages, for future work.

The resultant posterior distributions from the full model case (BMM-3) facilitate phenology predictions for cultivars, ripening groups and for the maize species grown in different environmental conditions in Germany. The phenological development of a current cultivar which has been used for calibration, can be predicted in a different ecological region or weather class. In this case, the posterior parameter distribution of the cultivar can be used to simulate phenology using the SPASS model, while the corresponding parameters of the ecological region and weather effects can be used to correct the model simulations for structural deficits. The random year effects and unresolved residuals are added to represent the total uncertainty in predictions. Furthermore, cultivar-level parameters within a given ripening group can be used to simulate phenology for new cultivars in that group, while all cultivar-level parameters can be used to predict phenological development of silage maize grown in Germany in future scenarios. Although multi-level models are expected to improve prediction quality (Gelman, 2006a), this may not always occur. Fer et al. (2021) showed that Bayesian hierarchical modelling does not always lead to best predictions. Instead, it may result in a better representation of prediction uncertainty. This lack of improvement in prediction quality could be attributed to bias–variance trade-off, in which we introduce more explanatory parameters in the multi-level models at the cost of over-fitting. Although there is always a danger of over-fitting, we justify the complexity of the multi-level models since we employed a systematic approach to increase complexity based on system knowledge. While assessing prediction quality of the models is important, this is beyond the current scope of gaining a better understanding of the phenology model and identifying model deficits in its application to regional studies.

The BMM approach can be used as a diagnostic tool to guide model improvement efforts. For example, the possible influence of water-logging on phenological development in maize, as hypothesized from the resultant eco-region effects, emphasizes the need for field experiments to verify and investigate its impact (Liu et al., 2020). These experiments can then be used to formulate and parameterize water-logging stress in the process model equations (Liu et al., 2021). Wang et al. (2017a) showed that improved temperature response functions (TRFs) led to reduced uncertainty in wheat yield projections. Further studies should focus on comparing the performance of alternate TRFs during reproductive (post-flowering) development in maize, against experimental data. The BMM approach can also be applied to process-based crop models, wherein these point-based models can be spatialized (Pasquel et al., 2022). Additionally, gene-based models can be integrated with crop models to determine more representative genotype-specific parameters (Wallach et al., 2018; Casadebaig et al., 2020; Oliveira et al., 2021). BMM can be used to calibrate such models to data from multi-location breeding trails, so that genotype-dependent parameters and their environmental interactions can be disentangled.

4.5 Conclusions

In this study, we demonstrated that Bayesian multi-level modelling (BMM) is a suitable approach to account for the hierarchical structure of cultivars nested within ripening groups of a crop species, while simultaneously providing insights into model deficits related to environmental factors. The pooled model case (BM-0) led to an over-confidence in the process model parameters and comparatively poor calibration quality. While accounting for the eco-region and year effects (BMM-1) improved calibration quality, the eco-region effects possibly compensated for the missing weather effects. Estimating eco-region, weather, and year effects (BMM-2a) highlighted temperature-related model deficits during reproductive development. It also showed that in the absence of cultivar-ripening group information, the weather effects were able to capture their missing effect to some extent. However, accounting for cultivar-ripening group information (BMM-2b and BMM-3) led to more representative estimates of parameter uncertainty and clearly improved calibration performance. In the full model case (BMM-3) with the additional inclusion of the weather effects, the eco-region effects did not compensate for the missing weather effects (as they did in BMM-2b). Eco-region effects could possibly be linked to water-logging stress on phenology which is not represented in the process-based model. Furthermore, between-cultivar variability in posterior parameter distributions matched theoretical expectations, thereby emphasizing the strength of the full multi-level model. Thus, accounting for the eco-regions, weather and year effects, and specifically the hierarchical classification of cultivars and maize ripening groups led to better calibration and representation of parameter uncertainty as compared to the commonly used pooled approach.

With our approach, models that have been primarily used for field scale or cultivar-specific studies can be extended to regional scales. Although we do not explicitly correct the process model equations, we account for effects of model deficits related to environmental conditions. This approach also highlights model deficiencies which can facilitate model improvement. These findings can be used to design dedicated experiments and data-gathering procedures to support the refinement of model equations. BMM could also be applied to small-scale studies to account for between-farm and within-farm variability. BMM is a valuable tool in the Bayesian tool-box that should be implemented in crop model calibration studies.

4.6 Appendix A. Weather class clustering

The 3004 site-years available for the study, were classified into ten weather classes based on average temperatures and cumulative precipitation between April and June, and between July and September. K-means clustering was used to define the weather classes. The K-means algorithm generates clusters by minimizing within-cluster variance that is based on Euclidean distances (Hartigan and Wong, 1979). First, the average of the mean daily temperature ($T_{sy,s}$) from April to June and from July to September were calculated for each site-year (sy) as follows:

$$T_{sy,s} = \frac{1}{N_s} \sum_{n=1}^{N_s} T_{sy,n} \tag{4.A1}$$

where s represents the season from April to June or from July to September, N_s is the total number of days in that season, and $T_{sy,n}$ is the mean temperature (°C) on a given day n at site-year sy. Similarly, cumulative precipitation ($P_{sy,s}$) was also calculated from April to June, and from July to September as follows:

$$P_{sy,s} = \sum_{n=1}^{N_s} P_{sy,n} \tag{4.A2}$$

where $P_{sy,n}$ is the precipitation (mm) on a given day n at site-year sy.

The values for each of the four factors ($T_{sy,Apr-Jun}$, $T_{sy,Jul-Sep}$, $P_{sy,Apr-Jun}$, $P_{sy,Jul-Sep}$) per site-year were then normalized by scaling the values between 0 and 1.

$$\bar{k}_{sy} = \frac{k_{sy} - k_{min}}{k_{max} - k_{min}} \tag{4.A3}$$

where k_{sy} represents the factor, $\bar{k}_{sy}$ is its normalized value, and $k_{min} = \min(k_1, k_2, \ldots, k_{SY})$ and $k_{max} = \max(k_1, k_2, \ldots, k_{SY})$ are the minimum and maximum values of the particular factor across all the site-years ($SY = 3004$), respectively. The kmeans function from the stats package in R was run to generate ten clusters. To ensure stability of the resultant clusters, 100 starting points were set. The maximum number of iterations was set to 1000 and 10 clusters were specified to generate 10 weather classes.

4.7 Appendix B. SPASS model equations

Phenological development in the SPASS model occurs in three main phases: emergence, vegetative phase and reproductive phase. The rate of phenological development during emergence or the emergence rate is a function of sowing-depth (*Sdep*) in cm and a minimum or base temperature (*emt*) in °C requirement only above which emergence occurs. For a particular day d between sowing and harvest, if T_d is the temperature, then the emergence rate Re_d (d^{-1}) is given by

$$Re_d = \max\left(0, \frac{0.5 \times (T_d - emt)}{15.0 + 6 \times Sdep}\right) \tag{4.B1}$$

The temperature response function (TRF) defines the phenological development rate during the vegetative and reproductive phases, as a function of temperature. Development occurs only between the minimum (*tmin*) and maximum (*tmax*) temperature defined by the TRF, with maximum development occurring at the optimum temperature (*topt*). The TRF is given by

$$f(T_d, tmin,\ topt,\ tmax) = \begin{cases} \frac{2(T_d - tmin)^{\alpha} \cdot (topt - tmin)^{\alpha} - (T_d - tmin)^{2\alpha}}{(topt - tmin)^{2\alpha}} & \text{if } tmin \leq T_d \leq tmax \\ 0 & \text{otherwise} \end{cases} \tag{4.B2}$$

where

$$\alpha = \frac{\ln 2}{\ln\left(\frac{tmax - tmin}{topt - tmin}\right)} \tag{4.B3}$$

These cardinal temperatures are expressed in °C and are development phase-specific. Thus, for the vegetative phase, the development rate Rv_d (d^{-1}) is dependent on the phase-specific TRF (*fv*) (-) and the maximum development rate (1/*pdd1*) for the vegetative phase at optimum temperature (d^{-1}). The TRF scales between 0 and 1 and acts as a reduction factor on the maximum development rate when the temperature is not at the optimum.

$$Rv_d = \frac{fv(T_d, tminv, toptv, tmaxv)}{pdd1} \tag{4.B4}$$

Similarly, for the reproductive phase, if 1/*pdd2* (d^{-1}) is the maximum development rate at optimum temperature and *fr* (-) is the TRF, then the development rate Rr_d (d^{-1}) is given by

$$Rr_d = \frac{fr(T_d, tminr, toptr, tmaxr)}{pdd2} \tag{4.B5}$$

The phenological development rate R_d (d^{-1}) at a given day d is given by

$$R_d = \begin{cases} (1 - \psi_{d-1}) \cdot Re_d + (\psi_{d-1}) \cdot Rv_d & \text{if } -0.5 \leq Sdev_{d-1} < 1.0 \\ Rr_d & \text{if } 1.0 \leq Sdev_{d-1} < 2.0 \end{cases} \tag{4.B6}$$

where

$$\psi_{d-1} = \frac{1}{1 + e^{-100(Sdev_{d-1})}} \tag{4.B7}$$

In Eq. 4.B6 we introduced a slight modification for the original SPASS model equations (Wang, 1997). We defined a sigmoid (ψ_{d-1} in Eq. 4.B7) instead of a step function for the transition between emergence rate and vegetative development rate.

The internal development stage $Sdev_d$ is given by

$$Sdev_d = \sum_{d=germ}^{D} R_d - 0.5 \tag{4.B8}$$

where *germ* is the date of germination or, in this case, date of sowing since germination is assumed to be instantaneous.

The internal development stages are converted to BBCH stages $bbch_d$ by

$$bbch_d = \begin{cases} 10 \times (Sdev_d + 1) & \text{if } Sdev_d < 0.0 \\ 10 \times (1 + \frac{Sdev_d}{0.2}) & \text{if } 0.0 \leq Sdev_d < 1.0 \\ 10 \times (6 + \frac{Sdev_d - 1}{0.28}) & \text{if } 1.0 \leq Sdev_d \end{cases} \tag{4.B9}$$

In the R and Jags implementation of the model, phenological development rate is calculated at time-step of 0.1 day. Air temperatures are first interpolated at this time-step and used as model inputs. The development rate Eqs. 4.B1, 4.B4 and 4.B5 in units of d^{-1} are multiplied by 0.1. Phenological development is calculated from the first time-step of the sowing day to the end on the harvest day.

4.8 Appendix C. Model performance metrics

The model calibration performance represented by the mean RMSE and bias from the 100 site-years used for calibration (Fig. 4.C1), was analysed by eco-region and year for the five model cases BM-0, BMM-1, BMM-2a, BMM-2b, BMM-3. Overall, the ripening-cultivar hierarchy in BMM-2b and BMM-3 show a clear improvement in model calibration quality. The inclusion of eco-region and year effects from BMM-1 onwards, does not show a clear improvement trend across all classes.

4.9 Appendix D. SPASS model parameter distributions

Figures 4.D1 and 4.D2 show posterior distributions of parameters that exhibit low and high between-cultivar variability, respectively. In general, parameters that exhibit low between-cultivar variability, correspond to the temperature response function (TRF) in the vegetative and reproductive phases.

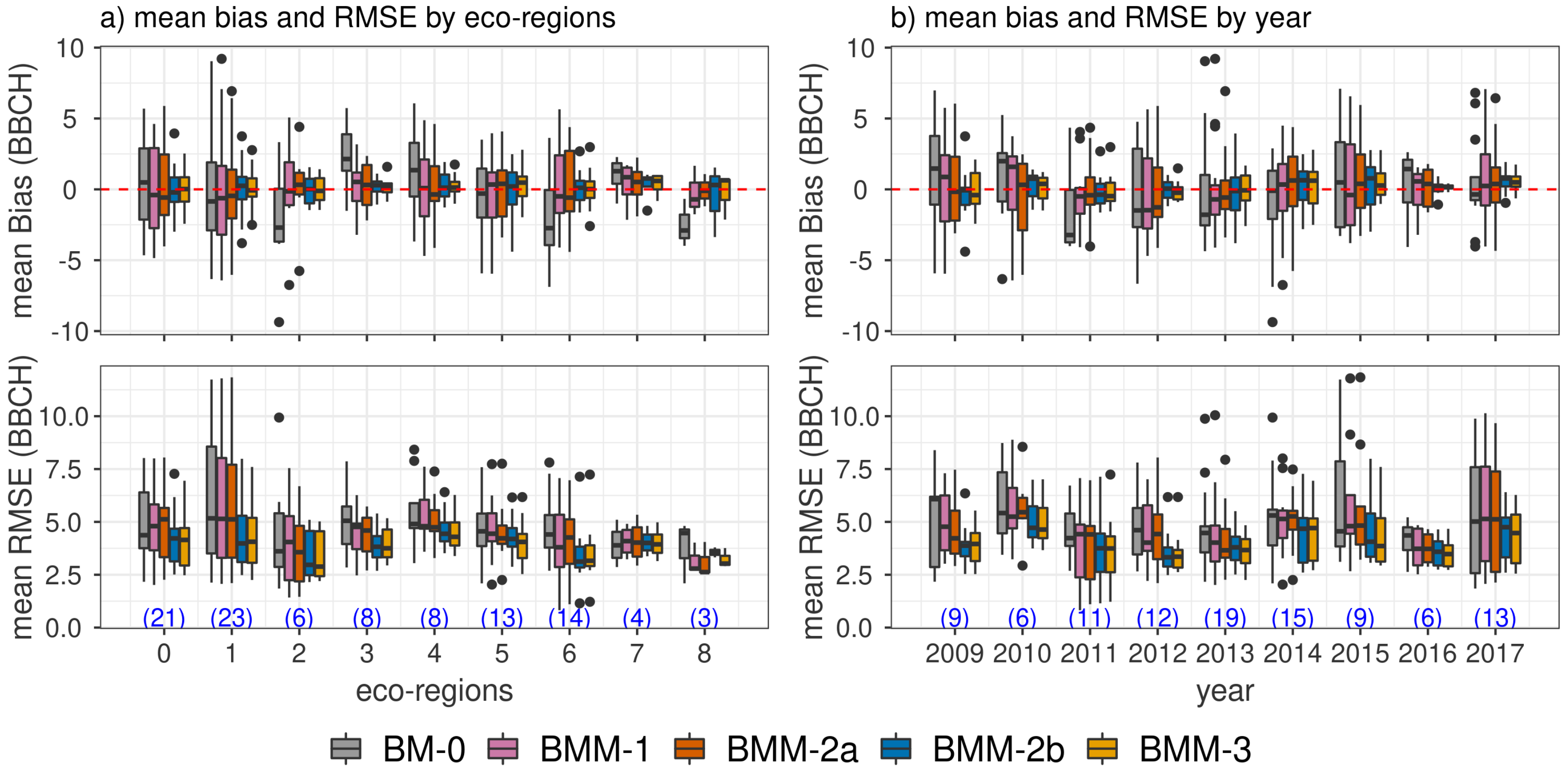


Figure 4.C1: Box-plots of the mean RMSE and bias by site-year within (a) eco-regions and (b) years for each of the five model cases: BM-0 (pooled), BMM-1 (eco-regions and random year effects), BMM-2a (eco-regions, weather class, and random year effects), BMM-2b (cultivar-ripening group hierarchy, eco-region, and random year effects), BMM-3 (full model with cultivar-ripening group hierarchy, eco-regions, weather class, and random year effects). The numbers in blue at the bottom of each plot indicate the number of site-years and consequently the number of points in the groups defined on the x-axis that were used to obtain the box-plots. Hinges of the box-plot represent the inter-quartile range (IQR), whiskers extend from the hinges up to 1.5×IQR and values beyond this range are plotted as points. Note that the phenology simulations used to estimate RMSE and bias do not take the σ parameter uncertainty into account.

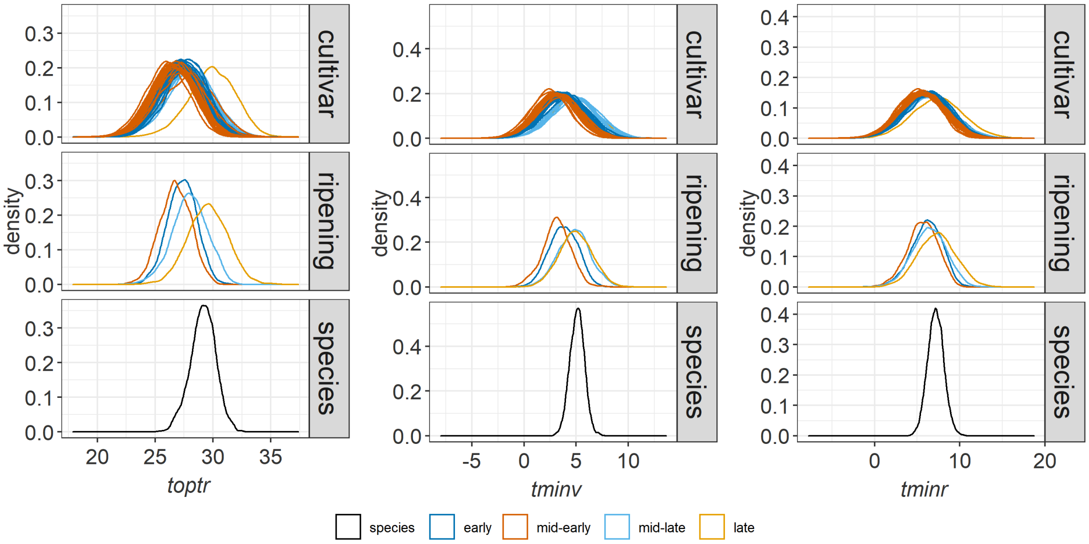


Figure 4.D1: Posterior distributions of parameters that exhibit low between-cultivar variability. Parameters include: minimum and optimum temperatures for development in the reproductive phase (*tminr* and *toptr*, respectively) and minimum temperature for development in the vegetative phase (*tminv*) for the Bayesian full model case BMM-3. The distributions are provided for the species, ripening group and cultivar levels of the hierarchy. The cultivar distributions are coloured by their corresponding ripening groups.

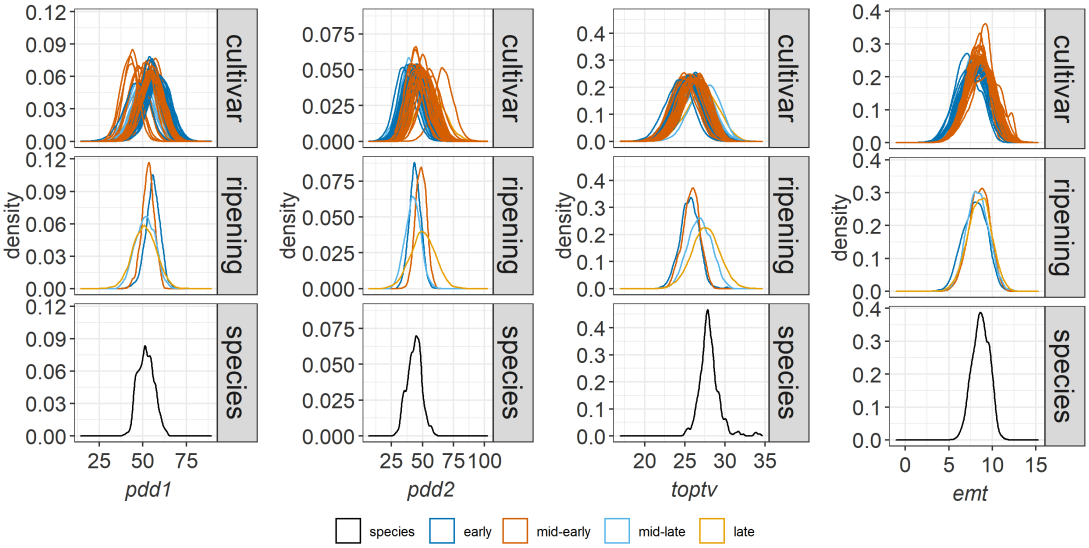


Figure 4.D2: Posterior distributions of parameters that exhibit high between-cultivar variability. Parameters include: phyisiological development days for vegetative (*pdd1*) and generative (*pdd2*) phases, optimum temperature for development in the vegetative phase (*toptv*), base temperature for emergence (*emt*) for the Bayesian full model case BMM-3. The distributions are provided for the species, ripening group and cultivar levels of the hierarchy. The cultivar distributions are coloured by their corresponding ripening groups.

CHAPTER 5

From Bayesian Multiplicative to Additive Calibration Strategy for More Reliable Predictions - A Demonstration on Plant Phenology Modelling

Authors: Michelle Viswanathan, Tobias K. D. Weber, Anneli Guthke

Abstract

Bayesian inference of the most plausible parameter values during model calibration is influenced by the method used to combine likelihood values from different observation data sets. In the traditional method of combining likelihood values (*multiplicative calibration strategy*), it is inherently assumed that the model is true, and that different data sets are similarly informative for the inference problem. However, practically every model applied to real-world case studies suffers from (misrepresented) model errors. Forcing an imperfect model to describe all data sets simultaneously inevitably leads to a compromised solution. As a result, biased and overconfident predictions hinder responsible risk management and any other prediction-based decisions. To overcome this problem, we recommend to use the alternative *additive calibration strategy* that allows the model to fit distinct data sets individually. To demonstrate the effect of choosing between the traditional multiplicative and the alternative additive strategy, we present a synthetic and real-world case study of calibrating a plant phenology model to observations of the maize crop. We demonstrate that the additive strategy results in conservative but more reliable predictions than the multiplicative strategy when the behaviour of the prediction target does not represent an average of all data sets. Further, expert knowledge-based data-grouping could be useful; however, selection of representative calibration data sets is not trivial. We expect the additive strategy to improve the predictive reliability of imperfect dynamic models in general, by a more realistic formulation of the likelihood function instead of assuming a "perfect model setting" in Bayesian updating.

5.1 Introduction

Hydrological models for water resources research suffer from diverse sources of uncertainty, such as sparse and noisy observations of input and output data, limited knowledge of heterogeneously distributed parameter values, and competing hypotheses about relevant processes at different spatial and temporal scales (Renard et al., 2010; McMillan et al., 2018). These uncertainties also exist in distributed plant and crop models, which may be coupled to hydrological models to account for vegetation-water interactions (Siad et al., 2019). The Bayesian framework allows to quantitatively consider these different sources of uncertainty during calibration (Bayesian updating), which makes it a popular approach for training simulation models under uncertainty, e.g. in the fields of rainfall-runoff (Kavetski et al., 2006b; Ajami et al., 2007), net ecosystem exchange (Weber et al., 2018), and crop modelling (Dumont et al., 2014; Wöhling et al., 2015; Gao et al., 2021; Viswanathan et al., 2022b).

However, a fundamental assumption that is commonly made when applying Bayes theorem is that the underlying model structure is true, or when considering several models, that the true model is in this set. Simplistic assumptions are made regarding model residual errors and their distribution, and they are often assumed to only arise from uncertainty in observations. This means that with regard to the example of parameter inference, if the analyzed model is true, Bayesian updating will identify the true system's parameter values in the limit of infinite calibration data. In real-world applications, the assumption of a true model is always violated, because the chosen model will be a coarse abstraction of the natural system. Model errors may exist due to simplified or erroneous model equations (structural errors), errors in the model inputs or boundary conditions, etc. In other words, model deficits exist that are expressed as errors in prediction (Wöhling et al., 2013; Viswanathan et al., 2022b). Several model deficits with respect to different processes might interact and produce complicated patterns of model error that depend on simulation period-specific boundary conditions, acting processes, amongst others (Hsueh et al., 2022). Thus, it is practically impossible to perfectly describe all errors, and simplistic assumptions about them do not hold true in most real-world applications. Despite the fact that these assumptions are not fulfilled, they are still made when applying the Bayesian approach for pragmatic reasons and to maintain simplicity. This typically results in overconfident and biased parameter estimates and prediction intervals (Brynjarsdóttir and O'Hagan, 2014; Xu and Valocchi, 2015).

One possible strategy to address this problem is to try and account for model error in Bayesian analysis either within the model structure or by an end-of-pipe statistical model error description (Kuczera et al., 2006; Del Giudice et al., 2013; Xu and Valocchi, 2015; Makowski, 2017; Reichert et al., 2021). However, these approaches require sufficiently large data sets for defining suitable error models. These methods may also incur high computational costs with increasing number of parameters to be estimated, especially when the underlying process model is already computationally expensive.

As a somewhat ad-hoc solution, it has been proposed to use smaller data sets for Bayesian calibration, in order to avoid extreme narrowing of the posterior distribution (Motavita et al., 2019). By using less data, the assumption of the model being quasi-true is more likely to be met (Hsueh et al., 2022). Although this is a valid recommendation, it is scientifically unsatisfying to discard information just because the updating procedure is not adequately tailored to the problem. Along the same lines of using smaller data sets but maintaining the information of the full data set, we propose to divide the available data into subsets based on expert knowledge, and then to perform Bayesian calibration individually on each subset. By doing so, we reduce the degree of violation of the fundamental Bayesian assumption. Finally, the obtained posterior distributions from all subsets are averaged in an additive unweighted likelihood combination scheme (Zak et al., 1997), henceforth referred to as the "additive likelihood approach". The interpretation of the proposed routine is that the model is required to fit certain segments of a data set (e.g., a time series period that represents a certain hydrological condition, or one growing season of a specific crop, etc.), but not several segments of different conditions simultaneously, i.e., with the *same* parameter set.

We do not believe that a model is generally able to simultaneously fit various conditions of the natural system without changing model parameters because of the deficits mentioned above. Instead, model parameters are forced to compensate for model errors during calibration, leading to biased parameter distributions with misquantified uncertainties. In the traditional case, parameter sets are estimated that fit well in a compromise sense to the full data set. This is nearly impossible (and often physically implausible), and explains the typical collapse of the posterior predictive distribution to very narrow intervals. In the additive strategy proposed in this study, each sub-period for calibration might favour its own parameter sets, and these are combined to reflect the model's struggle to match the observed data more realistically, given the varying boundary conditions.

In a similar vein, Bayesian hierarchical modeling (BHM) has been applied to increase the model's generalizability in such situations (Kuczera et al., 2006; Viswanathan et al., 2022a). In contrast to the simple unweighted averaging of the additive likelihood approach, BHM requires the specification of hyperparameters and hyperpriors and a smart adaptation of the sampling scheme, which may sometimes hinder its application in real-world scenarios.

In contrast to the approach taken by Hsueh et al. (2022), who propose a moving time-window concept for model error diagnosis in a Bayesian framework, we consider expert-elicited sub-data sets (not necessarily consecutive in time, could also be data sets from different spatial regions, or different data types, etc.), and contrast the effects of the traditional Bayesian multiplicative (*mult*) vs. the alternative additive (*add*) calibration strategy in their respective predictive performances. We note that this type of sub-setting and differential treatment of data groups is archetypal for crop model calibration strategies (Wöhling et al., 2013), in soil-water (Vrugt et al., 2001) and hydrological models (Razavi and Tolson, 2013; Motavita et al., 2019). This is handled in the additive strategy through an alternative method of combining likelihoods within and across data-subsets or groups, thus influencing the probability of data being generated by the given model and parameter set.

The proposed approach of subdividing the available calibration data in view of varying system conditions and applying the additive calibration scheme in Bayesian updating mitigates known problems of overconfident and biased posterior distributions, which often spoil probabilistic model predictions for practical purposes such as resources management, risk assessment, or climate change impact assessment. The goal of this study is to compare the mathematical formulation of the alternative additive likelihood approach with the traditional multiplicative likelihood approach in Bayesian updating, and make modellers aware of how their calibration decisions affect the model performance.

He et al. (2010) have evaluated the impact of different likelihood measures (formal and informal likelihoods) and combinations on crop model parameter estimation within the GLUE framework (Beven and Binley, 1992, 2014). Since they generated synthetic data without introducing model errors, the true model was in the set of possible model outcomes. This is exactly why they found that the multiplicative strategy performs well in reducing posterior uncertainty the most. The problem emerges when we consider real-world modelling case studies with imperfect models (Beven et al., 2008), and this is the challenge we tackle here. We investigate the additive and multiplicative likelihood combination strategies in the presence of model structural errors. In such real-world conditions, we offer a structured perspective on how to handle multiple data sets and analyse the effect this has on the resulting posterior. This choice is relevant both in formal Bayesian approaches and in informal GLUE approaches (Mantovan and Todini, 2006; Beven et al., 2007). We recommend relying on the Bayesian approach (with formal likelihoods, as used in this study), because it yields proper probability density functions (PDFs) which are invaluable for decision-making.

We illustrate the performance of both the traditional multiplicative and the alternative additive calibration approach on the example of a crop phenology model. Phenology is an important state variable in crop models. It influences plant biomass, leaf area index (LAI), and yield. Crop models, in turn, are applied at regional scale for several purposes such as climate impact assessment, yield projection and food security evaluation as well as for investigating the fate of agrochemicals in the environment (Chenu et al., 2017). Phenological observations of crops are made during field-visits, and can be used to calibrate crop models. Phenological development depends on environmental drivers and does not only differ between crop species (such as maize vs. wheat) but also between cultivars. Cultivars of a species may be grouped into ripening groups based on similar phenological development in response to environmental drivers. In regional simulations, where we would like to draw inferences for the crop species as a whole, it is important to account for uncertainty arising from differences between ripening groups or cultivars. Since phenology model equations do not account for these differences, this limitation manifests as structural deficits when these models are applied at regional scales. Ideally, such an application would be a suitable candidate for BHM (Viswanathan et al., 2022a). However, cultivar and ripening group information may not be available in regional data sets (Teixeira et al., 2017). Field-based phenology observations may also be limited and insufficient for estimating all hierarchical model parameters.

In such a situation, a modeller might decide to proceed with a “pooled” approach by gathering all available observed data over all ripening groups, combine them into one big data set, and perform Bayesian calibration on it to obtain a common parameter set - with the goal of preparing the model for “anything that could happen”. By doing so, the modeller inherently and erroneously assumes that the model is able to capture differences between cultivars and ripening groups. Unfortunately, this decision is tragically wrong, because the outcome is an extremely narrow posterior predictive distribution that is unlikely to have any (substantial) overlap with what is happening in the real system.

So what has gone wrong? By trying to fit different data sets that reflect diverse system conditions (ripening groups and also soil conditions, weather inputs, etc.), the model struggles to the extent that numerical sampling might simply fail to find a single parameter set that can predict the full data set with acceptable accuracy. The traditional multiplicative likelihood-based Bayesian updating routine will then yield a collapse of the posterior ensemble. So instead of adequately representing the uncertainty about the ripening group to be predicted, the modeller has posed an impossible task. The model will become unusable because its predictions have collapsed to a best-compromise solution with possibly no physical interpretation and practically no uncertainty left in the model parameters, which in reality are still quite uncertain.

We will first theoretically demonstrate that the multiplication of likelihoods is the source of this problem and show how such a multi-data set calibration task may be framed mathematically with the more appropriate additive calibration scheme. Then, we demonstrate the differences between both approaches in a synthetic and real-world case study. With the synthetic case study, we generate synthetic data and introduce model structural errors to demonstrate general properties of the additive calibration strategy. In the real-world case study we investigate the performance of the two strategies in low data conditions (field-based observations) and low information (no cultivar or ripening group information) scenarios which are common limitations, as highlighted earlier. We calibrate a plant phenology model using the traditional multiplicative and the alternative additive approaches. We use phenology observations of silage maize which was grown in two regions in southwestern Germany between 2009 and 2016 (Weber et al., 2022). During this time period, different cultivars of silage maize belonging to different ripening groups were grown in different environmental conditions. Furthermore, as in the case of most environmental models, the phenology model is known to contain model deficits. By investigating different combinations of calibration data sets and prediction targets in a real-world case study with known model deficits, we will derive recommendations on when the traditional multiplicative strategy should be applied and when the additive strategy is more appropriate for reliable predictions.

This article is structured as follows: We start by recalling Bayesian updating in Section 5.2.1 and the reasoning behind the traditional multiplicative Bayesian likelihood formulation in Section 5.2.2. Then, we present the alternative additive strategy based on predefined subsets of calibration data in Section 5.2.3. We explain the skill score used to compare both approaches in Section 5.2.4. Section 5.3 features the phenology modelling in synthetic and real-world case studies. Results of the calibration strategies are discussed in Section 5.4. General conclusions and an outlook towards further potential adaptations of our proposed approach are given in Section 5.5.

5.2 Bayesian Model Calibration

5.2.1 Bayesian Updating

Model calibration via Bayesian updating defines the posterior probability $p\left(\boldsymbol{\theta}|M,\boldsymbol{y^o}\right)$ of a parameter set $\boldsymbol{\theta}$ given a specific model structure M as the product of its prior $p\left(\boldsymbol{\theta}|M\right)$ and the likelihood $p\left(\boldsymbol{y^o}|M,\boldsymbol{\theta}\right)$ to have produced the observed data $\boldsymbol{y^o}$:

$$p\left(\boldsymbol{\theta}|M,\boldsymbol{y^o}\right)=\frac{p\left(\boldsymbol{y^o}|M,\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|M\right)}{p\left(\boldsymbol{y^o}|M\right)} \tag{5.1}$$

For the sake of brevity, we omit the notation $(\cdot|M)$ (conditional on model M) from now on, since we are not concerned with comparing the calibration of competing models, but with comparing alternative calibration strategies to condition one individual model.

The data used for Bayesian updating, $\boldsymbol{y^o}$, typically comprises either all available data, or the fraction of it devoted to calibration when the remaining fraction is withheld for validation and/or testing. We will denote the calibration data set length with N_o. Through the likelihood function, the goodness-of-fit between model predictions as a function of model parameters, $\boldsymbol{y}=f(\boldsymbol{\theta})$, and observed data $\boldsymbol{y^o}$ is assessed and used to identify the most-likely regions of the parameter space. The strength of the calibration effect depends on the exact formulation of the likelihood function. We note that the informativeness of the prior may also play an important role, but is not investigated here. We focus on the specific question of how data sets of different types (be it different seasons, different hydrological conditions, different observed state variables, etc.) can be combined into a formal likelihood function.

5.2.2 Likelihood Formulation in the Traditional Multiplicative Calibration Scheme

Traditionally, a joint likelihood for all data points is formulated. If we assume measurement errors to be independent, the likelihood simplifies to the product of univariate likelihood functions - an assumption frequently made in environmental modelling:

$$p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{mult}=\prod_{j=1}^{N_o}p\left(y^{o,j}|\boldsymbol{\theta}\right) \tag{5.2}$$

Equation 5.2 implies that the calibration requires each individual parameter set to fit data $y^{o,1}$ *and* data $y^{o,2}$ *and* data $y^{o,3}$, and so on. If even one of the data points has a very low likelihood, the overall product of likelihoods will be very low, and in the extreme case will be zero. This also becomes obvious from the equivalence of the product of likelihoods with the sum of the log-likelihoods. The logarithm places a large importance on small values, so the overall likelihood will be dominated by badly predicted individual data points. This reveals the difficulty of achieving high (not close-to-zero) likelihoods for large data sets that cover different conditions/states of a natural system with an imperfect model.

In the context of numerical evaluation, this means that we seek individual parameter sets that fit all data points sufficiently well - a very small number of random samples will prove to be “good enough” in the usually quite vast parameter space of the model. More precisely, the overlap of the extremely sharp posterior with the typically rather wide prior is so small, that numerical sampling schemes are pushed to their limits. This difficulty exists no matter which numerical method is used, but of course the methods differ in accuracy and efficiency. Popular approaches are Monte Carlo simulations with different types of sampling schemes, such as posterior sampling (Markov chain Monte Carlo, see e.g. Hastings (1970)), or prior sampling (brute-force Monte Carlo, see e.g. Schöniger et al. (2014)). It is important to point out that the problem of inefficient search for the high-likelihood region of the model increases with larger model errors. In other words, the inability of the model to fit all data types simultaneously and/or larger data sets increases concomitantly, simply because the chance to achieve a high likelihood at each data point decreases.

5.2.3 Likelihood Formulation in the Additive Calibration Scheme

Instead of the traditional multiplicative calibration scheme that rests on a joint likelihood formulation for all data points, we propose to subdivide the calibration data set into meaningful subsets and combine their likelihoods by addition. Mathematically this is achieved by combining likelihoods within subsets using a product (traditional multiplicative scheme), and across subsets using a sum. With N_s subsets of data and each subset being denoted by s, containing N_d data points:

$$p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{add} = \sum_{s=1}^{N_s} \prod_{j=1}^{N_d} p\left(y_s^{o,j}|\boldsymbol{\theta}\right) \tag{5.3}$$

Through the sum over all data groups, a parameter sample will score a high likelihood if it fits one data group extremely well, or many data groups sufficiently well. Badly predicted values will reduce the score, but not to the extreme extent as in the traditional *mult* scheme. Additionally, if any likelihood $p\left(y_s^{o,j}|\boldsymbol{\theta}\right) = 0$, the combined likelihood $p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{add}$ does not necessarily equal zero, as it would in case of $p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{mult}$.

Selecting an ideal length of the subsets can be a challenge - the periods should be long enough to achieve a “healthy” calibration effect by constraining parameter values and reducing uncertainty, but short enough (time-wise) or specific enough (data type-wise) to assume constant system conditions for the model to simulate (see the related discussion of Hsueh et al. (2022) on the choice of an optimal window length for time-windowed Bayesian model error analysis).

The data subsets may be motivated by expert knowledge, for example. It may be possible to define subsets of the available calibration data based on very similar system conditions. These subsets could be used to group calibration data such that the model should be able to fit all groups equally well with the *same* parameter sets. Other groupings may reflect different system states. Acknowledging that parameters tend to compensate for model errors, we should aim to identify parameter sets that fit at least *either one* of the different data groups.

5.2.4 Skill Score Used to Evaluate Predictive Performance

Our goal is to achieve a more realistic estimate of uncertainty in predictions that are informed by a combination of various data sets. Hence, we are interested in how well future data points are covered by the posterior predictive distribution. This information is quantified by the predictive density of the data. We use the predictive log-score (PLS) (Good, 1952) to multiply the densities of all N_t target data points, or equivalently, sum over their log-densities:

$$PLS = \sum_{j=1}^{N_t} \log p(y^{t,j} | \boldsymbol{\theta}, \boldsymbol{y^o}) \tag{5.4}$$

Note, that we do not specify how the calibration on $\boldsymbol{y^o}$ was performed (multiplicative vs. additive), because this skill score evaluates the performance on the validation (target) data set independent from the chosen method for calibration.

While using this skill score seems similar to using a multiplicative scheme for performance evaluation, there is a fundamental difference: at each data point, the full predictive distribution is taken into account, which means that different parameter sets can be the best ones for different data points. In contrast, in the multiplicative calibration case, individual parameter sets are required to fit *all* data points simultaneously.

We choose the PLS because it is an adequate measure to rank the quality of the predictive distributions in our application (see Section 5.3); however, our proposed calibration scheme is independent of the chosen metric such that modellers could decide to use other skill scores to reflect their individual modelling goals.

5.3 Demonstration in a Crop Phenology Modelling Case Study

5.3.1 Motivation and Goals

We applied and compared the traditional multiplicative calibration strategy with the alternate additive strategy on a case study of crop phenology modelling. Phenology defines the timing of plant developmental stages like emergence, stem elongation, flowering, development of fruit, and senescence. It is controlled by environmental factors such as temperature, photoperiod, water availability, and also depends on intrinsic plant characteristics (Zhao et al., 2013).

As mentioned earlier, the influence of these environmental factors on phenological development is not only species-specific (for example, difference between the species of maize and wheat), but also differs between ripening groups and cultivars of the same species. This can be modelled using equations with ripening group- or cultivar-specific parameters. However, for regional-scale modelling studies, where cultivars belonging to different ripening groups of a crop species are grown, it may be necessary to determine a common parameter estimate for the species, in order to predict future production.

Since these models are usually not error-free, because not all environmental interactions are adequately taken into account in the model equations, estimating common parameter sets for different ripening groups grown in different environments with the traditional multiplicative calibration strategy results in a compromised solution that may not always lead to reliable predictions (Viswanathan et al., 2022b).

The alternative additive calibration strategy has the potential to improve predictions by relaxing the model's prediction intervals and allowing the model to fit each predefined subset of data, individually. To assess the prediction performance with the additive calibration strategy, we used both strategies to calibrate a silage maize phenology model, to phenology observations made in southwestern Germany between 2010 and 2016. We compared the calibrated model's prediction performance from the two strategies using the predictive log-score (PLS) (Section 5.2.4).

5.3.2 Data

The data used for the study consist of phenology observations and temperature measurements from three field sites (site 1, site 2, site 3) in Kraichgau and two field sites (site 5 and site 6) on the Swabian Alb, taken between 2010 and 2016 (Weber et al., 2022). At each study site and year combination (called "site-year" in the following sections), phenological development stages were observed in five subplots where ten maize plants in each sub-plot were monitored. The BBCH growth stage code (Meier, 2018) was used to define the development stages.

We calculated arithmetic means of the ten replicates in the five subplots (5 × 10) for every day of observation. These mean observations were used in model calibration $\boldsymbol{y_s^o} = \{y_s^{o,1}, y_s^{o,2} ... y_s^{o,N_d}\}$. The total observation uncertainty δ_s^d was calculated as detailed in Viswanathan et al. (2022b) for a site-year s on a given day of observation d. It was assumed to represent both the uncertainty in identification of the correct phenological development stages and the spatial variability within the field.

The cultivars grown at the study sites belong to early (E), mid-early (ME), and late (L) ripening groups. Ripening groups indicate differences in the timing required by the maize cultivars in reaching maturity, for example: the early ripening cultivars mature the earliest, followed by the mid-early and then the late ones. Data from 11 site-years were used for the study (Table 5.1). Based on the average of daily temperatures between 40 and 100 days after sowing, which is the approximate time during which vegetative development (phenological development between emergence and flowering) occurs, the site-years were grouped into three temperature classes: (1) low ($\leq$15.4°C), (2) mid ($>$15.4°C and $\geq$16.6°C), and (3) high ($>$16.6°C). For example, site-years 3-2011 and 6-2010 are in the *mid* temperature class and thus maize crops grown there experienced similar average temperatures (15.4-16.6 °C) between 40-100 days after sowing.

Table 5.1: Site-years used in the case study with ripening groups of silage maize and temperature class.

Region	**site-year**	**site**	**year**	**ripening group**	**temperature class**
Kraichgau	3-2011	3	2011	late	(2) mid
Kraichgau	2-2012	2	2012	late	(3) high
Kraichgau	1-2014	1	2014	mid-early	(3) high
Kraichgau	2-2014	2	2014	mid-early	(3) high
Swabian Alb	6-2010	6	2010	mid-early	(2) mid
Swabian Alb	5-2011	5	2011	mid-early	(1) low
Swabian Alb	5-2012	5	2012	early	(2) mid
Swabian Alb	6-2013	6	2013	mid-early	(3) high
Swabian Alb	5-2015	5	2015	early	(3) high
Swabian Alb	5-2016	5	2016	early	(2) mid
Swabian Alb	6-2016	6	2016	mid-early	(2) mid

5.3.3 Model

The SPASS crop growth model (Wang, 1997) has been part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Bassu et al., 2014; Durand et al., 2018; Falconnier et al., 2020; Kimball et al., 2019; Wallach et al., 2021a,b) and has been among the well-performing models. It is implemented in the Expert-N 5.0 (XN5) software package (Heinlein et al., 2017; Klein et al., 2017; Priesack, 2006). In this study, we implemented the SPASS phenology sub-model in the R programming language (R Core Team, 2020) and used it to simulate phenological development of silage maize grown at the 11 site-years.

The SPASS phenology model contains 12 parameters, of which 6 were estimated while the remaining were fixed at their default values (Table 4.2). We modelled three main development phases, emergence (up to BBCH 10), vegetative (between BBCH 10 and 61) and reproductive (BBCH 61 onwards). Emergence is a function of the sowing depth (*sowdepth*) and a certain minimum or base temperature requirement (*emt*). The development rate during the vegetative and reproductive phases are dependent on the number of physiological development days at optimum temperature (*pdd1* and *pdd2*, respectively) and on the Temperature Response Function (TRF). The TRF is defined by phase-specific minimum (*tminv*, *tminr*), optimum (*toptv*, *toptr*), and maximum (*tmaxv*, *tmaxr*) cardinal temperatures for the vegetative and reproductive phases, respectively. The values of the TRF lie between 0 and 1, with the highest development rate occurring at optimum temperature. The internal development stages are a cumulative sum of development rates during the three main phases. Finally, the internal development stages in SPASS are converted to BBCH stages based on conversion relationships (for details please see Appendix A: SPASS Phenology Model in R).

The six model parameters estimated during calibration were: effective sowing depth (*sowdepth*), physiological development days at optimum temperature (*pdd1*, *pdd2*), the optimum temperatures (*toptv* = *tmaxv* − *dtoptv*, *toptr* = *tmaxr* − *dtoptr*) for respective vegetative and reproductive phases, and the BBCH stage corresponding to the internal development stage of 0.4 (*convert*). The remaining parameters were fixed at their default values: *tminv* = 6°C, *tmaxv* = 44°C, *tminr* = 8°C, *tmaxr* = 44°C, *pdl* = 0 (photoperiod sensitivity).

Table 5.2: Ranges for the estimated SPASS model parameters used to define weakly informative prior distributions.

Parameter	**Description**	**Unit**	**Mean**	**SD**	**Min**	**Max**
pdd1	physiological development days - vegetative phase	day	45	7	15	70
pdd2	physiological development days - reproductive phase	day	36	8.75	5	70
dtoptv	Difference between maximum and optimum temperature - vegetative phase	°C	10	1.5	5	20
dtoptr	Difference between maximum and optimum temperature - reproductive phase	°C	10	1.5	5	20
convert	equivalent bbch stage for 0.4 internal phenology stage	BBCH	30	7.5	11	59
sowdepth	effective sowing depth	cm	8	2.5	1	20

5.3.4 Calibration Schemes in the Context of Site-Years

Let $\boldsymbol{\theta}$ represent the vector of uncertain model parameters and $\boldsymbol{y_s^o}$ represent the vector of observations $y_s^{o,1}, y_s^{o,2}, \ldots, y_s^{o,N_d}$ at N_d days for the s^{th} site-year. The probability of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y_s^o}$ as per Bayes theorem is

$$p(\boldsymbol{\theta}|\boldsymbol{y_s^o})_{mult} = \frac{p(\boldsymbol{\theta}) \cdot \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \cdot \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{5.5}$$

where $p(\boldsymbol{\theta})$ is the prior probability of the parameter vector and $p(y_s^{o,d}|\boldsymbol{\theta})$ represents the likelihood of observing one data point $y_s^{o,d}$, given the parameter set $\boldsymbol{\theta}$. By multiplying the individual likelihoods, $\prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})$, we assume that the observations are independent from each other (no correlation in measurement errors over time), and we require the model and its parameter vector to fit the whole time-series simultaneously (traditional multiplicative strategy). This seems justifiable for observations made within a site-year since a single cultivar is grown within a field site in a given year. Therefore, the parameters of the model, which are based on plant characteristics, are not expected to vary within a single growing season.

Since data from N_s site-years are available ($N_o = N_s \times N_d$), we wish to calibrate our model on this collection of data sets, by following the general modeller intuition of “using all information we have”. For testing and evaluation purposes, we keep one site-year for validation and exclude it from the calibration data. To avoid artefacts in our conclusions stemming from distinct site-year characteristics, we systematically investigate predictive skill for all N_s site-years when calibrating on the data from the remaining $N_s - 1$ site-years (leave-one-site-year-out cross-validation).

The maize crop exhibits differences in phenological development between different ripening groups (Oluwaranti et al., 2015) as well as between cultivars (Gao et al., 2020) within these ripening groups. Furthermore, these cultivars also exhibit differences in development as a function of the environment (Lamsal et al., 2018). Ideally, models are expected to capture these environmental dependencies so as to make them transferable to new environments. However, cultivar-specific parameters are often found to vary with environmental conditions (Ceglar et al., 2011), indicating possible model structural limitations in capturing these environmental interactions. When a common parameter set is estimated for such a model by using all the site-years for calibration, irrespective of ripening group, cultivar or environmental conditions during growth, the resultant parameter set is a compromised solution. This corresponds to the traditional multiplicative strategy.

With the case study-specific notation introduced here, the posterior probability of the parameters in the *mult* case is given by

$$p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}})_{mult} = \frac{p(\boldsymbol{\theta}) \cdot \prod_{s=1}^{N_s-1} \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \cdot \prod_{s=1}^{N_s-1} \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{5.6}$$

The alternative *add* strategy, which allows the model to fit data sets from each individual site-year, would account for the differences between data sets arising from distinct ripening groups, cultivars, and environmental conditions. In this sense, it would make use of all information in the observations. The differences between the site-years are translated into wider posterior parameter distributions. As the posterior parameter distributions then better reflect the variable characteristics of the calibration site-years, it increases the probability of reliably predicting a new target site-year.

In this *add* case, the posterior probability of the parameters is given by

$$p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}})_{add} = \frac{p(\boldsymbol{\theta}) \cdot \sum_{s=1}^{N_s-1} \frac{\prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}}}{\int p(\boldsymbol{\theta}) \cdot \sum_{s=1}^{N_s-1} \frac{\prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}} d\boldsymbol{\theta}} \tag{5.7}$$

Note the subtle difference between Eqs. 5.6 and 5.7: in Eq. 5.6 a double product is used, while Eq. 5.7 combines the data within one site-year using a product as per the traditional joint likelihood formulation, but the likelihoods of multiple site-years are summed up (*add*). In principle, the multiplicative combination within a single site-year across different development phases (emergence, vegetative and reproductive) could be questioned and data points could be regrouped based on development phases.

This would require a detailed insight into model structural errors as a function of plant growth which is beyond the scope of this study.

The posterior predictive distribution, that is, the probability of observing $\boldsymbol{y^o_{N_s}}$ given the observations from the $N_s - 1$ site-years is expressed as

$$p(\boldsymbol{y^o_{N_s}}|\boldsymbol{y^o_{1:N_s-1}}) = \int p(\boldsymbol{y^o_{N_s}}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}}) d\boldsymbol{\theta}, \tag{5.8}$$

with the posterior parameter distributions $p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}})$ obtained from either the *mult* (Eq. 5.6) or the *add* case (Eq. 5.7).

5.3.5 Test Case Scenarios

We applied the *add* and the *mult* calibration strategies in a synthetic and a real-world case study.

5.3.5.1 Synthetic Case Study

We developed a synthetic case study for an intuitive visual comparison of the resultant posterior parameter distributions in the two strategies and for demonstrating properties of the *add* strategy. In the synthetic case study, we calibrated the phenology model to synthetic data sets representing different cultivars. For our illustrative example, only two model parameters (*toptv* and *toptr*) were estimated during model calibration, while the other parameters were assumed to be known and fixed. Six synthetic data sets were generated using the SPASS model, based on predefined values of *toptv* and *toptr*, representing differences that could exist between cultivars. Since the phenology model equations are not able to account for between-cultivar differences, this limitation is considered to be a representation of model structural error. For simplicity, the six cultivars were assumed to be grown at the same site 6 in the year 2010, i.e. the same environmental conditions. For each of the six synthetic cultivars, phenology observations were made 60, 120, 150, and 180 days after sowing. A measurement uncertainty of 3 BBCH standard deviation was assumed and this random error was added to the simulated phenology from the model to generate the synthetic observations.

The phenology model was calibrated to three cultivars using the two strategies, while the remaining three cultivars were used for validation. To demonstrate the general properties of the additive calibration strategy and to contrast it with those of the multiplicative strategy, we defined four prediction scenarios by reinterpreting the M-settings defined by Höge et al. (2020) in terms of parameter space and calibration-prediction scenarios. The four scenarios represent different degrees to which the prediction data set is from the same population (Wallach et al., 2021a) as the calibration data sets. In each of the four scenarios, we compared the prediction quality on using the two calibration strategies. In Figure 5.1, each box represents the parameter space formed by the two estimated model parameters as the axes. The ovals represent a projection of the posterior parameter distributions if the model were calibrated to each cultivar individually using the traditional multiplicative strategy (Eq. 5.5). The blue ovals represent the three cultivars used for calibration and green are those used for prediction in the four M-settings. In the M-closed setting, the prediction site-year has exactly the same cultivar grown as one of the site-years in the calibration data set. In this scenario the same cultivar A that is in our calibration data set is predicted. In the Quasi-M-closed setting, the prediction target (cultivar D) is somewhat similar to one of the cultivars (cultivar C) in the calibration data set. This is indicated by the overlap between the ovals which represent their posterior distributions when the model is calibrated individually to them. In the M-complete setting, the prediction target (cultivar E) represents an average behaviour of all the cultivars in the calibration data set since it lies in-between the three calibration cultivars. In the M-open setting the prediction target (cultivar F) is not well represented by members of the calibration data set since it lies away from any of the calibration cultivars.

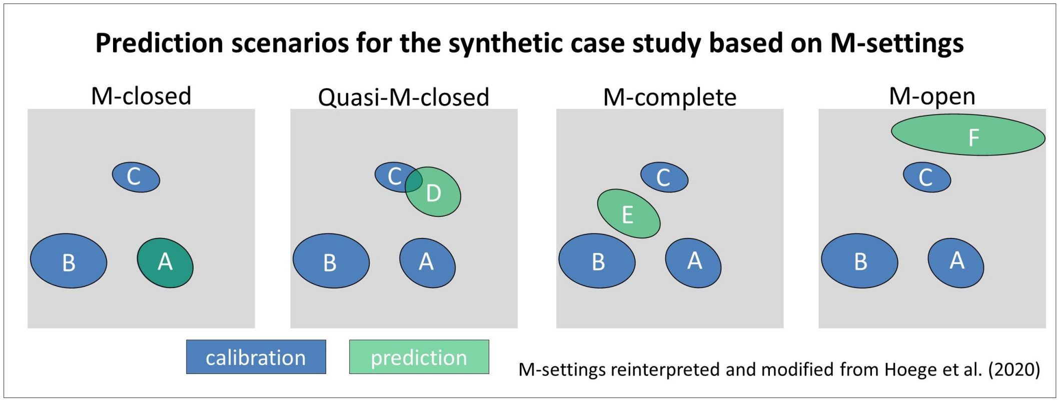
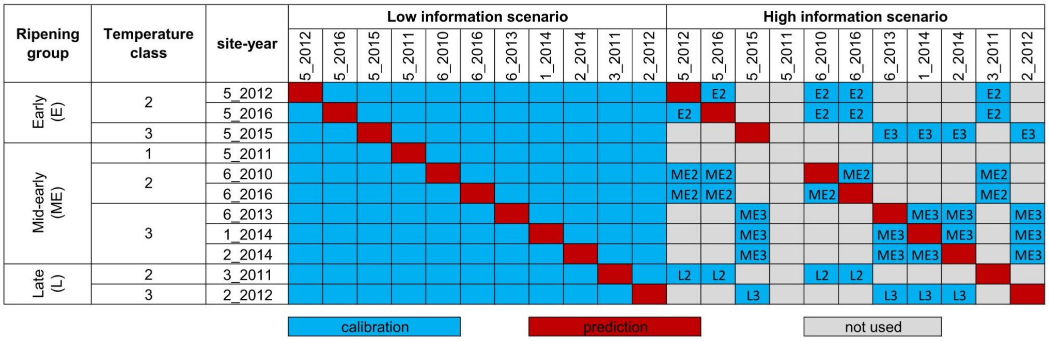

Prediction scenarios for the synthetic case study based on M-settings
M-closed
Quasi-M-closed
M-complete
M-open
C
B
A
C
D
B
A
C
E
B
A
F
C
B
A
calibration
prediction
M-settings reinterpreted and modified from Hoege et al. (2020)

Figure 5.1: Prediction scenarios based on M-settings (Höge et al., 2020). The grey box represents the model parameter space formed by the estimated parameters. It spans the same parameter space in each of the four scenarios with the ovals representing the posterior parameter distribution on calibration to individual cultivars. Labels in the ovals indicate the cultivar names. Blue represents the cultivars used for calibration (cultivars A, B, and C) while green the prediction targets (cultivars A, D, E, and F).

5.3.5.2 Real-World Case Study

In this case study, we compared the *mult* and *add* calibration strategies in predicting phenology at all 11 site-years (Table 5.1). For each prediction target site-year, the SPASS phenology model was calibrated to the 10 remaining site-years (leave-one-site-year-out). In this low information scenario, we were blind to the information about the ripening groups and temperature classes provided in Table 5.1 during calibration. We also tested a high information scenario, in which we used this additional information to select a subset of site-years for calibration and defined data-groups where likelihoods from site-years within the same group were combined using a product and across groups using a sum. The test case scenarios in the real-world case study are summarized in Fig. 5.2.

In the low information scenario, likelihood values from the calibration site-years were combined using Eq. 5.6 for the *mult* strategy. In the *add* calibration strategy, likelihood values of data-points within each site-year were combined by a product while likelihoods across site-years were combined by a sum as shown in Eq. 5.7. For the high information scenario, we subdivided the data based on knowledge about the model's performance. A previous study (Viswanathan et al., 2022b) showed that the SPASS phenology model was able to predict better when the prediction site-years had the same average temperature during vegetative development as the calibration site-year. Therefore, in the high information scenario, only site-years which were from the same vegetative temperature class (Table 5.1) as the prediction target site-year were used for calibration with the *mult* and *add* strategies. Knowledge about the cropping system was then used in the *add* scheme to define the likelihood combination strategy. Cultivars from the same ripening group are expected to exhibit similarities in phenological development. Therefore, likelihoods from the same ripening group were combined by a product and across ripening groups by a sum (Eq. 5.7). For example, in the high information scenario prediction of site-year 6-2013, only site-years in the same temperature class 3 (high average temperature during vegetative development) as the target, namely 5-2015, 1-2014, 2-2014, and 2-2012 were used for calibration. For the *mult* strategy, the likelihoods of all data points within this selection of site-years were combined by a product (Eq. 5.6), while in the *add* strategy, likelihoods from site-years 1-2014 and 2-2014 in the mid-early ripening group were combined using a product. This was then combined with the likelihood from 2-2012 in the late ripening group and the likelihood from 5-2015 in the early ripening group using a sum. Note, that there is no test case for predicting 5-2011 in the high information scenario as there were no other site-years from the same temperature class.



Ripening group	Temperature class	site-year	Low information scenario											High information scenario										
			5_2012	5_2016	5_2015	5_2011	6_2010	6_2016	6_2013	1_2014	2_2014	3_2011	2_2012	5_2012	5_2016	5_2015	5_2011	6_2010	6_2016	6_2013	1_2014	2_2014	3_2011	2_2012
Early (E)	2	5_2012	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	E2	⬜	⬜	E2	E2	⬜	⬜	⬜	E2	⬜
		5_2016	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	E2	🟥	⬜	⬜	E2	E2	⬜	⬜	⬜	E2	⬜
	3	5_2015	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	⬜	⬜	🟥	⬜	⬜	⬜	E3	E3	E3	⬜	E3
Mid-early (ME)	1	5_2011	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	2	6_2010	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	ME2	ME2	⬜	⬜	🟥	ME2	⬜	⬜	⬜	ME2	⬜
		6_2016	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	ME2	ME2	⬜	⬜	ME2	🟥	⬜	⬜	⬜	ME2	⬜
	3	6_2013	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	⬜	⬜	ME3	⬜	⬜	⬜	🟥	ME3	ME3	⬜	ME3
		1_2014	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	⬜	⬜	ME3	⬜	⬜	⬜	ME3	🟥	ME3	⬜	ME3
		2_2014	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	⬜	⬜	ME3	⬜	⬜	⬜	ME3	ME3	🟥	⬜	ME3
Late (L)	2	3_2011	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	L2	L2	⬜	⬜	L2	L2	⬜	⬜	⬜	🟥	⬜
	3	2_2012	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	⬜	⬜	L3	⬜	⬜	⬜	L3	L3	L3	⬜	🟥

🟦 calibration
🟥 prediction
⬜ not used

Figure 5.2: The low and high information scenarios, on which the multiplicative and additive calibration strategies were demonstrated. For each case represented by a vertical column, the prediction target site-year is marked in red while the site-years used for calibration are marked in blue. In the low information scenarios, ripening group and temperature class information was not considered for calibration and all remaining site-years except the prediction target were used. For the high information scenario, where ripening group and temperature class information was used to select site-years for calibration and to define data-groups, site-years excluded from calibration are in grey, while those site-years that were used for calibration are labeled with their respective ripening group (E = early, ME = mid-early, L = late) and temperature class (1 = low, 2 = mid, 3 = high). All likelihoods from site-years with the same label belonged to the same ripening group and were combined by a product in the additive strategy. Likelihoods across ripening groups were combined by a sum in this strategy.

5.3.6 Numerical Implementation

Since different versions of likelihood formulation are straightforward to implement in brute-force Monte Carlo sampling, we chose this numerical approach to obtain posterior parameter distributions. Alternatively, we could have used, e.g., an MCMC method, but would have had to rerun the MCMC for each prediction scenario, since the objective function changes with the considered calibration data sets. This would have caused a tremendous computational effort. For Monte Carlo sampling, in contrast, the effort was in creating the prior ensemble once, while likelihoods for different test case scenarios were obtained in the form of less-expensive post-processing.

The Monte Carlo ensemble consists of $N_{MC} = 511,000$ samples of the six parameters $\boldsymbol{\theta} = \{\phi_1, \phi_2 ..., \phi_6\}$. Maize phenology is simulated as a function of each parameter realization, $f(\boldsymbol{\theta}_i)$, $i = 1 \dots N_{MC}$, for $N_s = 11$ site-years. A weakly informative parameter prior $p(\boldsymbol{\theta})$, defined by a platykurtic distribution, is prescribed (details can be found in Appendix B: Prior Distribution).

Considering the shape of the likelihood function, we assumed that the residuals followed a normal distribution with a fixed standard deviation $\sigma_s^d = \sqrt{\delta_s^{d\,2} + \omega^2}$ where δ_s^d is a combined measure for the uncertainty in the measurement stemming from the observation process of BBCH and spatial heterogeneity in the field. The additional variance of $\omega^2 = 4$ represents a lumped model error term.

$$p(y_s^{o,d}|\boldsymbol{\theta}) = \frac{1}{\sigma_s^d \sqrt{2\pi}} \exp\left(- \frac{y_s^{o,d} - f(\boldsymbol{\theta})_s^d}{2\sigma_s^d} \right)^2 \tag{5.9}$$

The Effective Sample Size (ESS, Liu (2008)) was estimated to ensure that a large enough number of ensemble members contribute to posterior statistics.

In our low information scenario, the ESS values range from < 10 for the *mult* strategy to 2,000 < ESS < 4,000 for the *add* strategy with $N_s - 1$ calibration site-years. The ESS starts to drop below 20 in the *mult* strategy after using four or more site-years for calibration. This demonstrates the ensemble collapse that is often observed in Bayesian calibration on large data sets that contain a lot of non-redundant information. In contrast, the ESS values in the *add* calibration strategy show that this is not a problem in our proposed approach because the sampling method does not have to struggle as hard to find suitable parameter values. While ESS values for the *mult* strategy would undoubtedly improve with a better sampling algorithm, we still report these results here to highlight the unrealistic peakedness of the posterior distribution.

In the high information scenario in which only a selected subset of site-years is used for calibration, the ESS values for the *add* strategy range between 200 < ESS < 1,500. Here, the sampling problem is mitigated due to both, data set selection as well as ripening group based data-groups in the *add* strategy. For comparison, the *mult* strategy in the high information scenario yielded ESS ranging between 50 < ESS < 200. As a reference for these values, when the model was only calibrated to data from the prediction target site-year, the range of ESS is 500 < ESS < 2,000 (900 on average).

5.4 Results and Discussion

We first present the results of the synthetic case study, followed by the real-world case study and discuss conditions in which the *add* or *mult* strategy results in better prediction.

5.4.1 Synthetic Case Study

Figure 5.3 shows the posterior parameter distributions of the two parameters (*toptv* and *toptr*) in the synthetic case study. The parameter space is defined by the margins of the box. Here we see a striking difference between the additive and multiplicative calibration strategies. Figure 5.3a shows the posterior distribution when the model is calibrated separately to each of the three calibration cultivars A, B, and C. Figures 5.3b and c show the posterior distribution in the *add* and *mult* calibration strategies, respectively. The posterior distribution is wider and multi-modal in the *add* case since the parameter uncertainty accounts for the differences in the individual cultivars. In the *mult* case, the posterior distribution has collapsed and has resulted in a compromised solution that is not representative of any single one of the cultivars considered for calibration.

Figure 5.4a shows the posterior parameter distribution if the model were calibrated separately to each of the prediction target cultivars in the respective M-setting. Figure 5.4b shows the prediction results from the *mult* and *add* calibration strategies. The predictive log scores (PLS) for the four M-setting scenarios are provided in the Fig. 5.5. The M-closed (cultivar A) and Quasi-M-closed (cultivar D) cases show that the *mult* strategy performs poorly even in predicting a cultivar that is well-represented in the calibration data set (Fig. 5.4). This can also be seen from the lower PLS values (Fig. 5.5) for the *mult* strategy in predicting all calibration cultivars A, B, and C (M-closed setting). In the Quasi-M-closed setting, the *mult* strategy resulted in overestimated phenological development in the later part of the growing season. However, these observations are captured within the wider prediction intervals of the *add* strategy. In the M-complete setting, the *mult* strategy is able to capture observed phenological behaviour. It performs almost as well as the *add* strategy (note similar PLS values in Fig. 5.5) when the prediction target represents this average behaviour as in the M-complete case. In fact, it has the potential to perform even better than the *add* strategy by reducing variance in cases where the bias is already low. The M-open case highlights the problem of representativeness of the calibration data set, where none of the approaches can perform well. From the predictive-log-scores (PLS) (Fig. 5.5) we see that the *add* strategy performs better than the *mult* in all M-setting scenarios of our synthetic example.

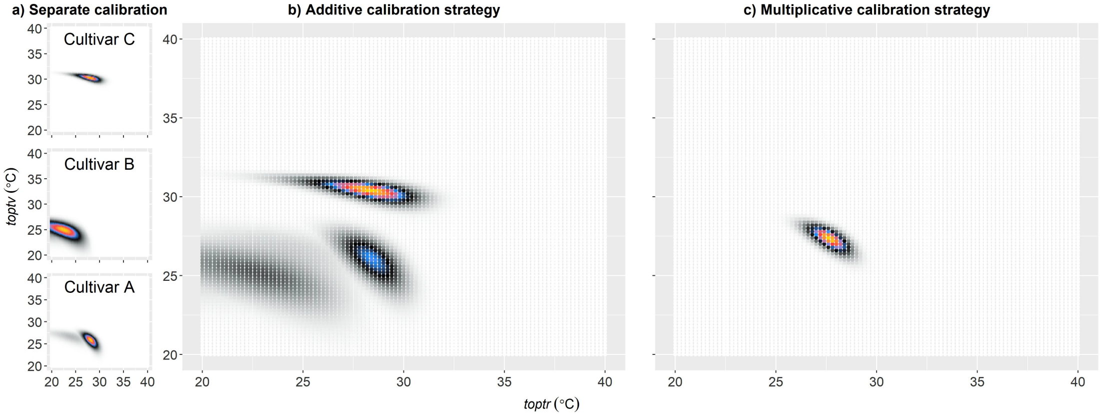


Figure 5.3: Posterior parameter distributions of the two parameters (*toptv* and *toptr*) within the uniform prior parameter space defined by the margins of the box. (a) Posterior distribution when the model is calibrated individually to each of three cultivars A, B, and C. Posterior distribution in the (b) additive and (c) multiplicative calibration strategies. The axes of all the graphs are at the same scale for ease of comparison. The colours indicate probability density and have been re-scaled to the minimum and maximum values in each sub-plot. Yellow and red colours indicate higher probabilities.

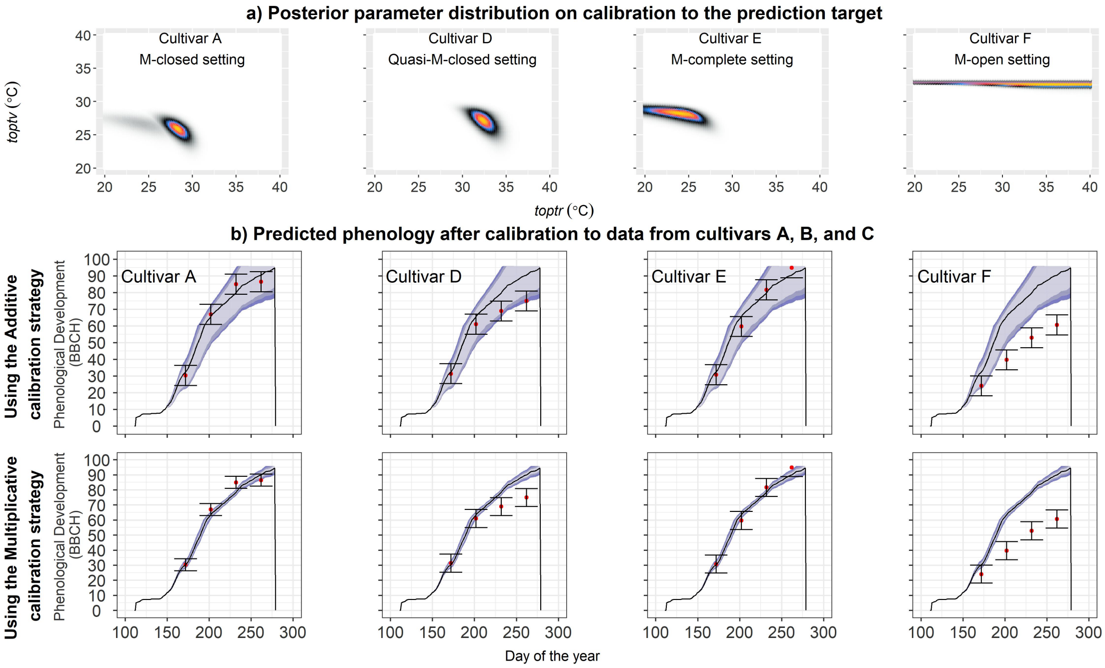


Figure 5.4: a) Posterior parameter distributions if the model were calibrated separately to each of the prediction target cultivars in the four different M-settings, i.e. cultivars A, D, E, and F. (b) Prediction results from the multiplicative and additive calibration strategies. The red points represent the mean of the observed phenology while the error bars represent two standard deviations of observation uncertainty. The coloured bands represent the different percentiles of simulated phenology (1 SD, 5-95, 1-99) using the SPASS phenology model, consisting of model parameter uncertainty only.

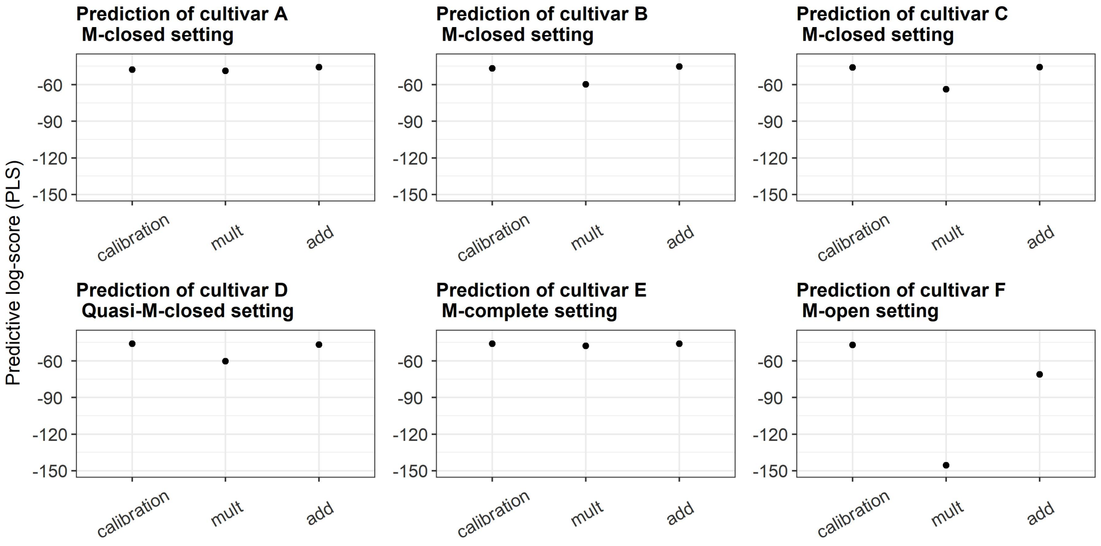


Figure 5.5: The predictive log-score (PLS) for calibration and prediction results for M-settings in the synthetic case study. The predictions in the *mult* and *add* strategies were made after calibrating the model to cultivars A, B, and C. The calibration result in which the model was calibrated to the prediction target cultivar is also provided as reference.

5.4.2 Real-World Case Study

For the purpose of discussion, we present selected real-world case study results of the low information scenario. *Mult* and *add* strategy results are shown for predictions of the early cultivar at 5-2012 (Fig. 5.6a), the mid-early cultivar at 6-2010 (Fig. 5.6b), and the late cultivar at 3-2011 (Fig. 5.6c). We also present the results in the high information scenario for prediction of site-years 2-2014 (Fig. 5.7a) and 6-2016 (Fig. 5.7b). The PLS of all other investigated cases are summarized in Fig. 5.C1 in Appendix C: Predictive Log-Score (PLS): Real-World Case Study.

As a reference, we also show calibration results for the prediction target site-year, where the model was calibrated to the data set from this target site-year alone. This can be understood as an idealized case, because we use exactly the data to be predicted for constraining the model's parameter distributions. Hence, prediction intervals should be tight around the data values. When calibrating on other site-years (realistic case), we would expect an inferior prediction performance, and wish to identify the calibration strategy that brings prediction intervals as close to the target data as possible.

For the high information scenario, apart from the *mult* (MULT) and *add* (ADD) cases as described in section 5.3, we additionally present results from the ADD_sy case in which data-groups of the selected subset of site-years were only based on site-year information, while ripening group information was ignored (i.e. only selected subset of site-years where used where likelihood values within a site-year were combined by multiplication while those across site-years by a sum). The motivation is to understand whether simply excluding site-years with a different temperature class than that of the prediction target is beneficial, and to what extent the additional ripening group information can further improve performance. The *mult* and *add* strategies using $N_s - 1$ site-years are also provided for reference, and are labeled as MULT_all and ADD_all, respectively.

5.4.2.1 The Additive Strategy is Conservative but Reliable

For all three target site-years shown in Fig. 5.6, the idealized case of calibrating on the target site-year only (first column in Fig. 5.6) yields accurate mean predictions and tight credible intervals, with observation uncertainty being partly larger than model parameter and model error uncertainty.

The traditional MULT_all calibration strategy (second column), however, performs very differently, depending on the analysed target site-year. For site-year 5-2012 (Fig. 5.6a), the prediction interval in the MULT_all case is even narrower than the calibration reference, and fails to cover many observations in the later phenological development stages. This result clearly demonstrates that combining large data sets representing different system conditions (here: different sites, cultivars, temperature classes, etc.) via a joint likelihood function leads to overconfident and biased predictions. Hence, the traditional approach of using all available site-years, and thereby assuming that maize has similar phenological development irrespective of differences in ripening group and environmental conditions during development, fails. The narrow posterior interval reveals that only very few parameter samples could be found that belong to the "not-close-to-zero likelihood region" of the model. This is reflected in the ESS value which is as low as 5, and thereby results would be deemed numerically unreliable. Since the sampling effort to achieve a certain convergence increases exponentially in MC, a drastic extension of the ensemble would be needed to lift ESS up to reassuring values. In order to keep computational costs within reasonable limits we did not increase the sample size, but recommend that better sampling methods be implemented for the *mult* strategy.

The proposed ADD_all strategy (third column in Fig. 5.6), in contrast, produces a much wider credible interval that relies on a comfortable ESS of 2,790. Maize phenological development is assumed to be distinct between the site-years in the ADD_all strategy, and this is why the calibration is not as strong and allows for more variability in the posterior credible intervals. The ADD_all intervals succeed in capturing all target data points. This is also reflected in the PLS values (fourth column in Fig. 5.6) with that of the ADD_all strategy being higher than the MULT_all strategy. As compared to the idealized case of calibration on this site-year only, the ADD_all intervals are much wider, and hence the predictive density of the individual data points is lower, leading to (as expected) a worse PLS as compared to this idealized reference.

In summary, for this specific prediction site-year, the ADD_all calibration strategy leads to conservative but more reliable prediction results than the MULT_all strategy. This is also observed for the prediction of phenology at site-years 5-2015, 6-2013, 5-2011, and 2-2014 (Fig. 5.C1). These results are similar to the M-closed and Quasi-M-closed scenarios in the synthetic case study.

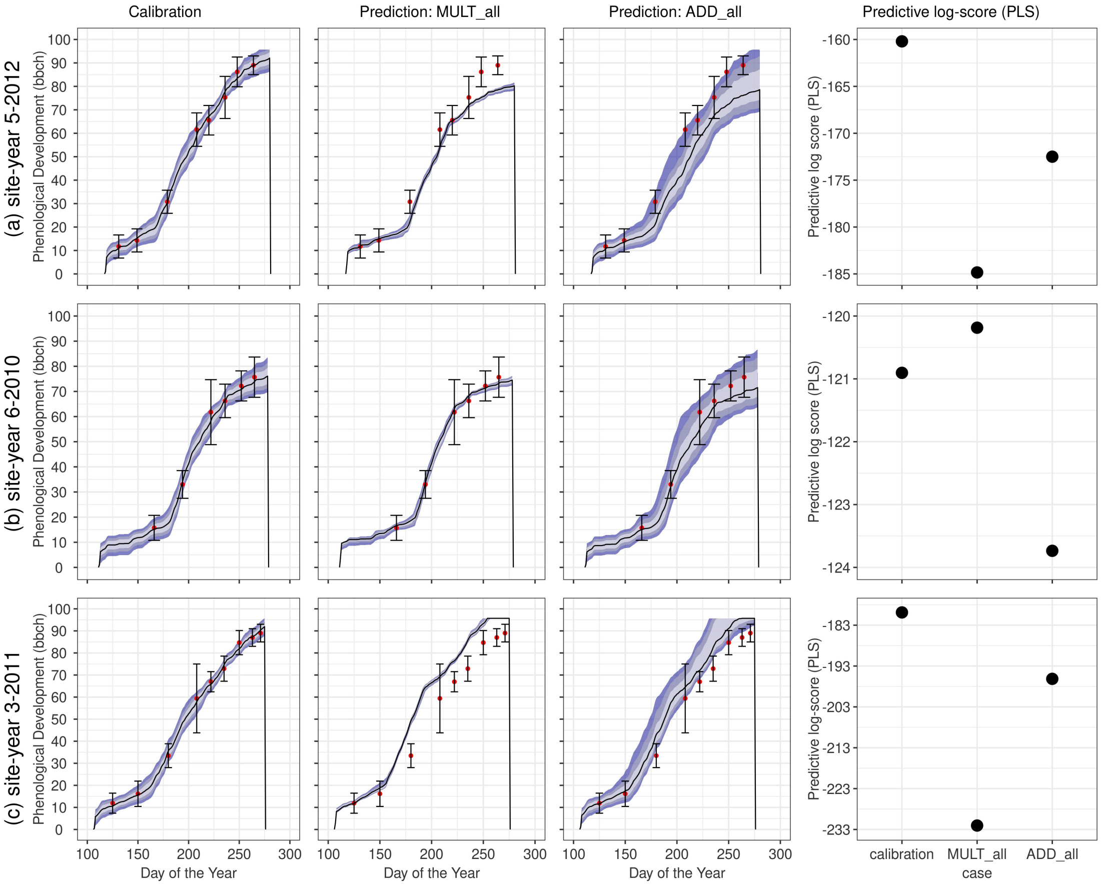


Figure 5.6: Observed and simulated phenology at site-years (a) 5-2012, (b) 6-2010, and (c) 3-2011 in the low information scenario. First column shows posterior credible intervals obtained from calibration on the target site-year only; second and third columns show posterior credible intervals from MULT_all and ADD_all calibration strategies, respectively; fourth column summarizes the predictive log-score for the three cases. The red points represent the mean of the observed phenology while the error bars represent two standard deviations of the observation uncertainty. The coloured bands represent the different percentiles of simulated phenology (1 SD, 5-95, 1-99) using the SPASS phenology model, consisting of model parameter uncertainty and a model error term. The solid line represents the posterior mean of the simulations.

5.4.2.2 The Multiplicative Strategy Succeeds when the Target Represents an Average Behaviour

In the prediction of phenology at site-year 6-2010 (Fig. 5.6b), the ADD_all strategy performs worse than the MULT_all strategy due a special feature of maize phenological development. Here, the MULT_all strategy prediction performs really well and captures the data points even better than the calibration reference as shown by the PLS values. The MULT_all case demonstrates what we would ideally like to achieve through calibration: with more and more data added (here: ten site-years instead of just the target one), model predictions should converge toward the observed system behaviour. While the PLS value of prediction in the MULT_all case might seem only slightly higher than the PLS of the calibration reference, we find that important phenological development stages like the ones around flowering (60

BBCH) exhibit a narrower range of uncertainty in the MULT_all case. Predicting the number of days after sowing that are required to reach this development stage is important for making field management decisions such as the timing of fertilizer applications.

Again, the ADD_all strategy yielded wider prediction intervals, but this time the loss of precision resulted in a lower PLS value than the MULT_all strategy. This is because the MULT_all strategy achieves a high precision paired with a very low bias, which is optimal for predicting each data value with a high predictive density.

The exceptionally good performance of the MULT_all strategy in this test case can be explained by the characteristic development behaviour of the three ripening groups. As indicated by the name, mid-early ripening cultivars generally mature earlier than the late ripening cultivars, but later than the early ripening cultivars. Although deviations occur due to environmental conditions and field management decisions, this general pattern can still be observed. Thus, the phenological development of mid-early cultivars, like the one at site-year 6-2010, represents an average behaviour of the three ripening groups. In the MULT_all strategy, the resultant compromised solution for phenology predictions after calibrating the model to data sets from the three ripening groups closely matched the observed development at 6-2010. Since the MULT_all strategy already performed very well, the relaxation of the prediction bands in the ADD_all strategy led to poorer predictions. Similarly, prediction with the MULT_all strategy was better than the ADD_all strategy for the mid-early cultivars at 6-2016 and 1-2014 (the interested reader is referred to Fig. 5.C1 in Appendix C: Predictive Log-Score (PLS): Real-World Case Study). These results are similar to the M-complete scenario in the synthetic case study.

5.4.2.3 Representativeness of the Calibration Data Plays a Role

In the case of site-year 3-2011 (Fig. 5.6c), the MULT_all strategy results in poor predictions and the ADD_all strategy yields only a marginal improvement as the wider prediction intervals still do not fully capture many of the observations. This is attributed to the representativeness of the calibration data (Wallach et al., 2021a) as observed in the M-open scenario of the synthetic case study. The calibration data consists of only one site-year from the same cultivar as the prediction target site-year but this cultivar was grown under different temperature conditions. However, even though the same cultivar was grown at 2-2012, the MULT_all calibration strategy was better than the ADD_all strategy at prediction (Fig. 5.C1). This site-year falls in the ‘high’ temperature class (Table 5.1) to which many calibration site-years belong and thus has representative site-years in the calibration data set. The high temperature results in earlier phenological development of this cultivar even though it belongs to the late ripening group, thus representing an average behaviour (Section 5.4.2.2). On the other hand, even though 5-2011 is a mid-early ripening cultivar, the ADD_all strategy performs better than the MULT_all. This is because there are no other site-years that lie within the same temperature class, and thus does not represent an average behaviour like the other mid-early cultivars.

In studies where data availability is not a limitation, we would only choose representative data for calibration, e.g. site-years from the same ripening group or cultivar, or those from the same environmental conditions as the prediction site-year. However, in regional studies with an aim to forecast a particular species where different cultivars and ripening groups are grown in different conditions, the ADD_all strategy enables us to account for the differences in data sets when estimating model parameters and uncertainty, resulting in a more conservative and reliable prediction outcome.

5.4.2.4 Data Set Selection for a Successful Additive Strategy is no Trivial Exercise

To test the potential of expert knowledge-based combination of selected site-years for calibration (high information scenario), only site-years 5-2015, 6-2013, 1-2014, and 2-2012 (all temperature class 3, cf. Fig. 5.2) were used for calibration in order to predict phenology at site-year 2-2014 (Fig. 5.7a). Recall from Section 5.3.5.2 that, in this approach, we combined site-years of the same ripening group by a product, and used a sum across different ripening groups (ADD case in Fig. 5.7). For comparison, we also show predictions with the *mult* strategy (MULT), and the *add* strategy without ripening group information (ADD_sy, i.e. data-groups based on site-years). Additionally, we provide prediction results from the low information scenarios where $N_s - 1 = 10$ non-target site-years were used for calibration (MULT_all vs. ADD_all).

The traditional MULT_all case (Fig. 5.7a-v) leads to overconfident prediction intervals for this predicted site-year, and the ADD_all case (Fig. 5.7a-vi) improves on that with wider intervals that succeed in capturing all target data points. The question whether this uncertainty can be reduced again without making overconfident and biased predictions via the *add* strategy can be answered with yes in this case: the ADD prediction interval has become narrower without losing any data points (Fig. 5.7a-ii). This is also obvious from the increase in PLS (Fig. 5.7a-vii). This effect can be caused by either the mere selection of site-years (as opposed to taking all available data independent of their representativeness, cf. Section 5.4.2.3) and/or due to the incorporation of ripening group information in defining data-subsets. We find that the mere selection of site-years improves over the $N_s - 1$ cases (the PLS increases for MULT vs. MULT_all and ADD_sy vs. ADD_all). But the ADD case, in which we additionally use the ripening group information, indeed performs best (second after calibration on the target site-year only).

However, for the *add* strategy with high information to succeed, a good understanding of model limitations and knowledge about data groups are needed. In the prediction of phenology at site-year 6-2016 (Fig. 5.7b), the site-year selection resulted in a lower PLS in the MULT case than in the MULT_all case in which all the remaining 10 site-years were used for calibration, because the MULT_all case yields very confident prediction intervals with relatively low bias. Naturally, calibrating on less data in the MULT case then leads to a weaker calibration effect and a lower PLS. The ADD_sy case with site-year-based data-groups resulted in a marginal improvement in PLS as compared to the MULT case (the wider intervals of ADD_sy now cover e.g. the last data value of the season better), while the ADD case with the ripening group-based data-subsets performs worse. Yet, in the ADD and ADD_sy cases, all observations and their measurement uncertainty ranges are covered by the high-probability region of the predictive interval, which is not observed in the other calibration cases. Thus, when aiming at reliable predictions and rather accepting variance than bias, the *add* strategy (ADD, ADD_all, ADD_sy) is better suited than the traditional MULT_all case.

The prediction results in the additive strategy could potentially benefit from expert knowledge-based plausibility constraints or by implementing a data-driven approach for defining data groups, e.g., informed by model deficits which can be evaluated using calibration performance indicators such as residuals. Defining data subsets based on the highest information content with respect to specific parameters (Vrugt et al., 2001) would also be insightful. Although we use a Brute-force MC sampling method, we suggest that MCMC algorithms which have been developed for high-dimensional multimodal posterior probability density functions (for example DREAM (Laloy and Vrugt, 2012)) should be tested in future applications.

The additional information of ripening groups could be taken into account using a Bayesian Hierarchical Modelling (BHM) approach to calibration. However, BHM requires the definition of priors for all hyperparameters, which is not a trivial task. The additive calibration strategy does not require this step. In practical applications our proposed approach may be preferred over BHM because of its ease of implementation and clear assumptions, while it still ensures reliable predictions. In fact, it would be similar to combining posterior probability distributions obtained from an unpooled calibration approach i.e. after calibrating the model separately to each data-group (compare Figs. 5.3a and b). Van Oijen and Höglind (2016) combined plant cultivar-specific parameter distributions to capture genetic variations. The model was first individually calibrated to two cultivars and then a beta distribution was defined with the same mean and variance as the union of the cultivar posterior distributions. A combined parameter distribution which is conservative but reliable can be achieved in a single step with the additive strategy, without the need for making further assumptions and methodological choices.

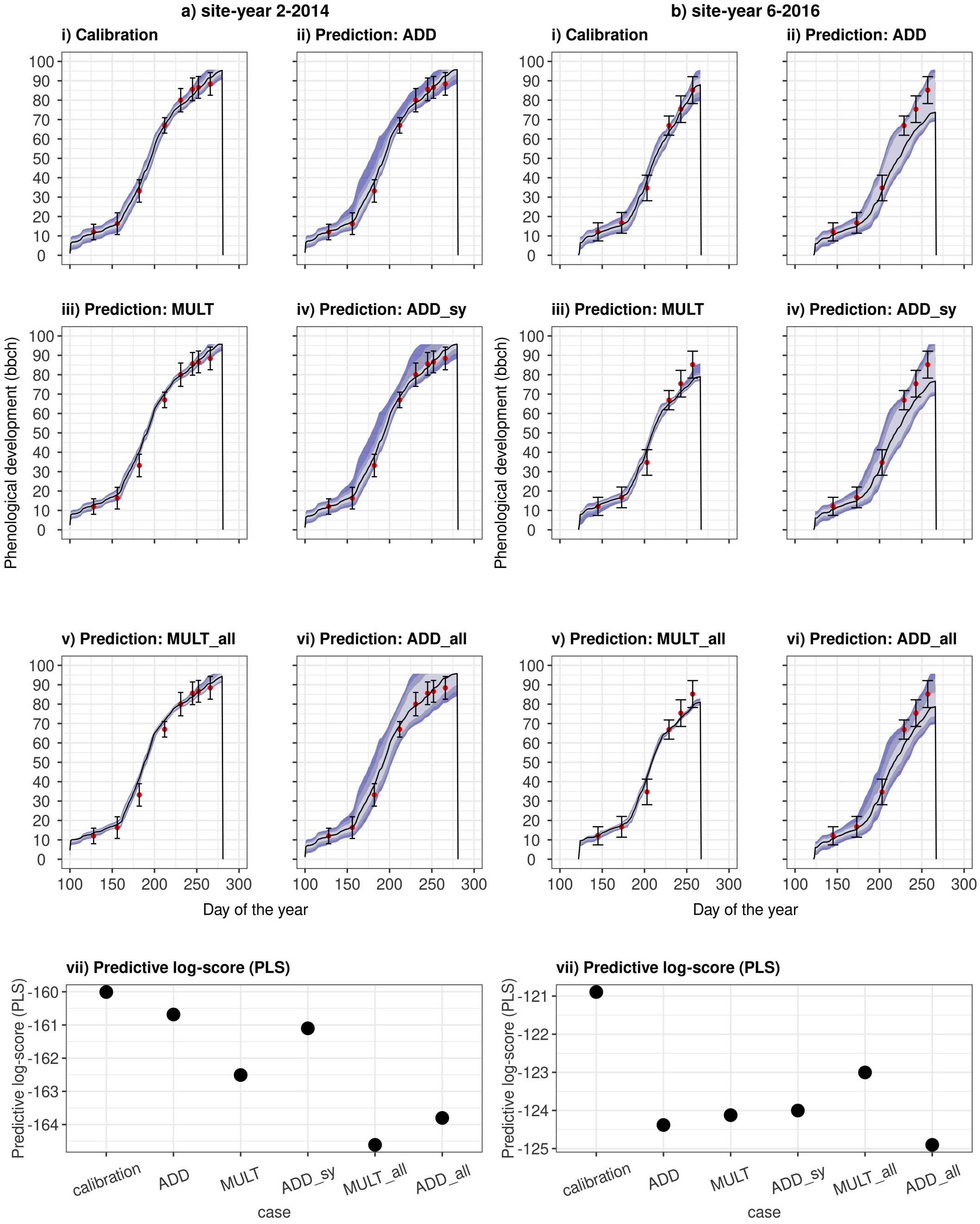


Figure 5.7: Observed and simulated phenology at site-years (a) 2-2014 and (b) 6-2016. Posterior credible intervals obtained from i) calibration on the target site-year only, ii) ADD , iii) MULT calibration cases in the high information scenario, iv) ADD_sy case with selected subset of calibration site-years but site-year based grouping, v) MULT_all and vi) ADD_all cases in the low information scenario; vii) summarizes the predictive log-score for all cases. The red points represent the mean of the observed phenology while the error bars represent two standard deviations of observation uncertainty. The coloured bands represent the different percentiles of simulated phenology (1 SD, 5-95, 1-99) using the SPASS phenology model, consisting of model parameter uncertainty and a model error term. The solid line represents the mean of the simulations.

5.5 Summary, Implications and Outlook

With this contribution, we tackle the problem that traditional Bayesian calibration on large, mixed data sets often leads to overconfident and biased predictions. The reason is the implicit assumption of Bayesian updating that the model is error-free, or with the errors known and adequately described, thereby assuming that any data set is similarly informative for the inference problem. However, practically every model applied to real-world case studies suffers from model structural errors, not all of which are necessarily known beforehand. Forcing an imperfect model to fit diverse data sets simultaneously (what we call the *multiplicative calibration strategy*) inevitably leads to a compromised solution to the parameter estimation problem, and triggers unreliable predictions. To overcome this problem, we have proposed that an alternative *additive calibration strategy* should be used which allows the model to fit distinct data sets individually. The posterior distributions resulting from calibration on the individual data sets are then combined (averaged) to reflect the remaining uncertainty after calibration. The proposed approach therefore represents one possible way forward to relax the assumption of a true model in Bayesian updating, and to obtain more realistic predictive uncertainty intervals in the presence of model errors.

First, we discussed the mathematical framework in which both strategies are embedded, which clearly points out the decisive differences in the formulation of the likelihood function. Secondly, we have compared the performance of the traditional multiplicative and the alternative additive strategies in a synthetic and real-world case study where a plant phenology model was calibrated to silage maize observations. The model's performance in predicting a data set that was not used during calibration was compared using the predictive log-score (PLS) as a metric. This metric directly evaluates the predictive density of observed data values, and thus accounts for both bias and variance in the posterior distributions. We found that the additive strategy resulted in higher scores when the predicted data set did not represent an average behaviour of the calibration data sets (e.g., with respect to temperature class or ripening group). As a special case, we also tested a high information scenario in which ripening group information of the maize crops was used to define data groups. Additionally, only those data sets from the same temperature class as the prediction target were used for calibration. In this scenario, likelihoods within groups were combined by a product and across groups were combined by a sum. While superior to the MULT and ADD_sy (site-year-based data-subsets) strategies in some cases, we found that the additive strategy with ripening group-based data-subsets (ADD) requires a more refined definition of data-groups based on expert elicitation.

Our proposed method generally applies to mathematical models where diverse data sets (comprising different state variables, periods of different system conditions, etc.) are used for model calibration. This approach can also be applied in multi-objective calibration studies, by combining likelihoods of different objectives using the additive strategy. Testing this approach on different types of models and data sets and in different applications is recommended for future work. Further, the prediction results in the additive strategy could be improved by defining data-groups that also account for the model's calibration performance and information content of data with respect to model parameters. We expect such advances to be very useful for environmental modelling studies where model structural errors are ubiquitous.

5.6 Appendix A: SPASS Phenology Model in R

The SPASS phenology model used for the study was implemented in R based on the implementation in the ExpertN-5 (Heinlein et al., 2017) modelling software and as described in (Wang, 1997), with some modifications: (a) No water-limiting conditions were considered for germination, i.e. germination occurred instantaneously upon sowing; (b) Photoperiod effect on the vegetative phase of development was not considered; (c) The phenological development stage in BBCH (*convert*) that corresponds to the internal development stage of 0.4 was included as a parameter in the model. In the SPASS model, the internal development stage ($Sdev_d$) on a given day d is converted to BBCH stage ($bbch_d$) as follows:

$$bbch_d = \begin{cases} 10(Sdev_d + 1) & \text{if } Sdev_d < 0.0 \\ (\frac{1}{0.4}(convert - 10))Sdev_d + 10 & \text{if } 0.0 \leq Sdev_d < 0.4 \\ \frac{1}{0.6}((60 - convert)Sdev_d + (-24 + convert)) & \text{if } 0.4 \leq Sdev_d < 1.0 \\ 10(6 + \frac{Sdev_d - 1}{0.28}) & \text{if } 1.0 \leq Sdev_d \end{cases} \tag{5.A1}$$

The conversion equations for phenological development stages are equivalent to the those described in Wang (1997); Viswanathan et al. (2022b) when $convert = 30$.

5.7 Appendix B: Prior Distribution

A weakly informative prior parameter probability $p(\boldsymbol{\theta})$, defined by a platykurtic distribution (Viswanathan et al., 2022b) was assumed for each parameter ϕ_h:

$$p(\boldsymbol{\theta}) = \prod_{h=1}^{6} p(\phi_h), \tag{5.B1}$$

where

$$p(\phi_h) = \begin{cases} \frac{1}{c_h} \frac{1}{\gamma_h \sqrt{2\pi}} \exp - \frac{(\phi_h - \mu_h)^2}{2{\gamma_h}^2}, & \text{if } a_h \leq \phi_h < \mu_h - 2\gamma_h \\ \frac{1}{c_h} \frac{1}{\gamma_h \sqrt{2\pi}} \exp -2, & \text{if } \mu_h - 2\gamma_h \leq \phi_h \leq \mu_h + 2\gamma_h \\ \frac{1}{c_h} \frac{1}{\gamma_h \sqrt{2\pi}} \exp - \frac{(\phi_h - \mu_h)^2}{2{\gamma_h}^2}, & \text{if } \mu_h + 2\gamma_h < \phi_h \leq b_h. \end{cases} \tag{5.B2}$$

Parameters of the platykurtic probability density function a_h, b_h, μ_h and γ_h are the minimum (Min), maximum (Max), mean (default), and standard deviation (SD), respectively, of a parameter ϕ_h based on expert knowledge (Table 5.2) and c_h is the normalization constant:

$$c_h = -\mathrm{erf}(\sqrt{2}) + \frac{4}{\sqrt{2\pi}} \exp -2 - \frac{1}{2}\mathrm{erf}\left(\frac{a_h - \mu_h}{\gamma_h \sqrt{2}}\right) + \frac{1}{2}\mathrm{erf}\left(\frac{b_h - \mu_h}{\gamma_h \sqrt{2}}\right). \qquad (5.B3)$$

5.8 Appendix C: Predictive Log-Score (PLS): Real-World Case Study

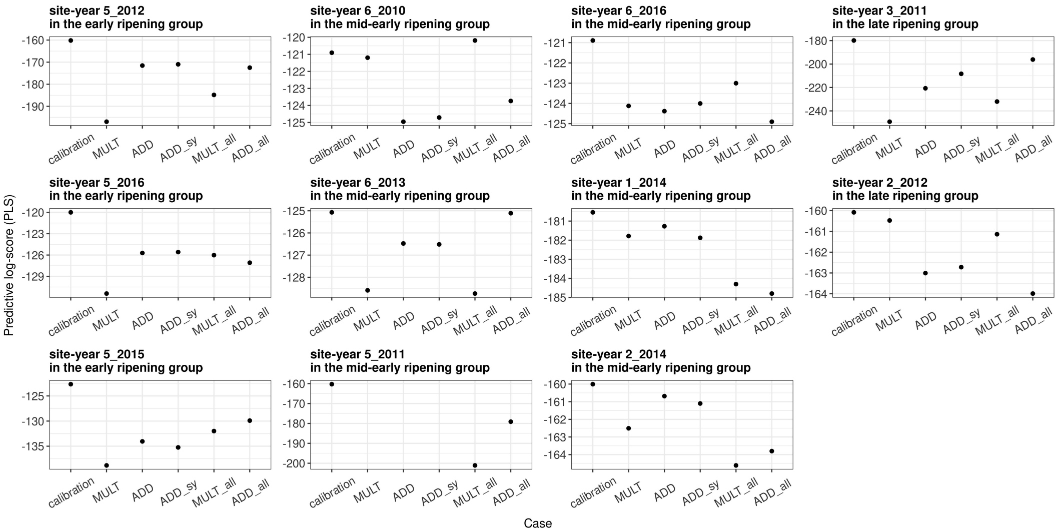


Figure 5.C1: The predictive log-score (PLS) for calibration and prediction results. The predictions in the MULT, ADD, ADD_sy strategies were made after calibrating the model to a selection of site-years. The predictions in the MULT_all and ADD_all scenarios were made after calibrating the model to all remaining site-years, excluding the prediction target.

CHAPTER 6

Conclusions & Outlook

6.1 Summary

Reiterating from the first chapter, the research objectives were to first identify the problems faced when applying simple Bayesian updating to plant models, and then to investigate the application of other suitable Bayesian methods to solve these problems.

We have seen that, *although Bayes theorem is a suitable framework for incorporating prior information and uncertainty quantification during process-model calibration, it leads to unreliable and erroneous results when model errors are present and statistical assumptions about the errors are violated.* This was demonstrated in Chapter 3 where Bayesian inference was applied to sequentially calibrate a maize phenology model to data on a yearly basis. Although parameter uncertainty reduced with increasing amounts of data, the prediction quality did not improve. These problems can be tackled (1) by improving the process model representation and (2) through better statistical representation of errors. Process model improvement is imperative to have better long term forecasts of crop production and to understand underlying mechanisms for making predictions in future scenarios and conditions which may be different from those seen today. However, it must be acknowledged that there will always be some errors in such environmental models where we try to simulate complex systems and their interactions. Therefore, the approaches implemented in this dissertation focus on improving statistical representation of errors while providing some insights into how model deficits can be identified so that studies addressing process-model improvement could be designed.

Although stated separately, these two solutions are in fact inter-related. This is well demonstrated in Chapter 4 in which a Bayesian multi-level modelling (BMM) approach was implemented. *Accounting for the hierarchical structure inherent in the data and environmental effects using BMM led to improved model calibration as compared to the commonly used pooled approach of lumping all errors into one term. Model deficits related to soil moisture and temperature effects during reproductive development (post-flowering) were identified.* These indicators of model deficits can be used to guide and focus model improvement initiatives. Although the process model equations are not corrected, the BMM approach does compensate for these model deficits to some extent, thus improving the overall model performance. Furthermore, instead of treating the residual error as random in the pooled approach, it was resolved into components arising from different environmental factors in the BMM approach, thus also ensuring better statistical representation of errors. This point was also addressed in Chapter 5 by relaxing the strict assumptions of Bayesian inference so as to account for model deficits. This relaxation of assumptions took the form of an alternative likelihood combination strategy applied during calibration. With this strategy the model is allowed to find a fit to each data set rather than finding a common fit to all data sets. By doing so, the model parameter estimates account for the differences in data sets caused by missing process representation and simplified process-model and statistical-model assumptions. As a result a more representative prediction uncertainty is obtained. Staying true to the 'Bayesian spirit', prior information (expert knowledge) can be used to guide data-groups within which the likelihoods are combined by a product and across which, by a sum. *The alternative strategy performed better in predictions as compared to the classical strategy of a product-based likelihood combination, except when the observations represented an average behaviour of all the calibration data sets.*

Thus, ignoring model structural errors leads to unreliable parameter estimates and consequently, predictions. It has been demonstrated that better-suited Bayesian methods exist and should be applied when calibrating crop models to improve prediction quality.

6.2 Choosing an appropriate approach

The choice of method depends on the goal of a study and other project management constraints (e.g. trade-off between quality, time, and cost (Fu and Zhang, 2016)). I have outlined qualitative recommendations, in terms of goal, time, computational constraints, to guide such decisions. While the additive calibration strategy (Chapter 5) provides a better representation of prediction uncertainty, BMM (Chapter 4) provides insights into model structure and enables identification of model deficits. By detangling the different sources of model structural error, it provides valuable information for data-gathering to facilitate uncertainty reduction. It may also result in more representative model parameter estimates. However, the computational costs could be high depending on the number of parameters as seen when the full model is applied (Chapter 4). Furthermore, the cost of analyses and interpretation could also be high, especially when error interactions are taken into account. The additive calibration strategy works well when the objective is to obtain a more representative prediction uncertainty. With minimum system and model knowledge the alternative strategy is recommended for conservative but reliable predictions at relatively low computational and analytical costs. In case of more system and model performance knowledge, better-informed data-groups can be defined. However, caution should be exercised since its success is believed to be highly dependent on the definition of these groups. Detailed likelihood modelling has a higher data requirement and is best-suited for big-data studies. It can be computationally and analytically demanding but would provide a less conservative and more representative estimate of prediction uncertainty. Additionally, separating out uncertainty into its components can help in improving model predictions in new environmental conditions.

6.3 Future scope

Further work should focus on implementing Bayesian methods to provide insights for improving process-model representation. Although this dissertation does not directly deal with process-model improvement, we have seen that approaches such as BMM can be used for this purpose. As the next step in this direction, a method to quantitatively compare existing models with respect to their deficits is needed. To achieve this, BMM could be applied to other phenology models. The calibration results can then be used for model comparison by setting up a coordinate system, i.e. a 'model-deficit' space. This will provide a map of models' performance that can guide model improvement efforts through targeted research and experimentation. Furthermore, these results can also be used as a guide for model weights in multi-model ensembles, when applying these models in specific environmental conditions. Additionally, extending this application for full crop models and incorporating gene-based modules to simulate existing processes in crop-simulation models could be a promising next step (Oliveira et al., 2021; Wallach et al., 2018; Vallejos et al., 2022; Peng et al., 2020).

With increasing data-availability, data-driven machine learning can further augment Bayesian model inference. As a concrete example, unsupervised machine-learning (ML) algorithms for cluster analysis (k-means clustering, hierarchical clustering, etc.) can be implemented to define data-groups based not only on observations, model forcings, system knowledge, but also on model calibration performance as expressed by the residuals. With systematic definition of groups that also incorporate model performance information, the additive calibration strategy is expected to lead to better predictions by also reducing variance in predictions. ML can also be used to link models at different scales (Alber et al., 2019). McCormick et al. (2021) combined knowledge-based (process) model and data-driven modelling for predicting soybean phenology in the Americas, by using predictions from process-models as additional features in the machine-learning (LSTM) models. On the other hand, Droutsas et al. (2022) incorporated the ML algorithm into a process-based crop model. Such applications should be investigated further with large data sets such as those from precision agriculture (Sharma et al., 2021).

The approaches implemented here should also be extended to multi-objective calibration studies. Bayesian multi-objective calibration has been attempted in hydrological modelling (Tang et al., 2018), but it has only been applied to a limited extent in crop modelling (Minet et al., 2015). A common practice in the crop modelling community is to first calibrate the model to phenology and then to other state variables (Seidel et al., 2018). Wöhling et al. (2012) attempted multi-objective optimization of four soil-crop models to LAI, soil moisture and actual evapotranspiration. They found that in some models multi-objective optimization led to compromised parameter estimates and poor calibration performance for all state variables of interest, than if they were individually calibrated. The additive calibration strategy could be promising in such studies. Data-groups for combining likelihoods could be based on the state variable. It is expected to yield better calibration results for the individual state variables than the classical strategy.

Residual analysis could be used to select a suitable likelihood model especially when satellite-based *measurements of LAI, biomass, etc. are used to calibrate crop models.* Certain problems with erroneous likelihood assumptions become more relevant when temporally and spatially distributed measurements, such as those from satellites, are used for calibration. For example, model errors are often assumed to be independent. But they may be auto-correlated in time (Schoups and Vrugt, 2010) and space (Pasquel et al., 2022). Based on the nature of some model equations, these model error behaviours are theoretically expected to exist, for instance, the presence of storage terms in a hydrological model which would exhibit 'memory' effects (Evin et al., 2014) or equations based on cumulative sums in crop models. Furthermore, model errors are commonly assumed to be identically distributed, but they may in fact change as a function of the simulated state variable (heteroscadasticity) (Weber et al., 2018). However, these model errors cannot be adequately specified with the sparse data that are usually available for crop model calibration. But with the increasing availability of satellite-based measurements these trends may become readily apparent through residuals analysis (Supplement S3). This analysis can then be used to validate simplistic assumptions about the errors being Gaussian, independent, and identically distributed and subsequently update the likelihood model. Selecting a suitable likelihood model is expected to result in representative prediction uncertainty. Although such likelihood modelling approaches are commonly applied in the field of hydrology, there are limited examples (Tang et al., 2018, 2019) in crop, agroecosystem or vegetation models calibrated to plant measurements.

6.4 Concluding remarks

In this dissertation a multi-purpose application of Bayesian inference from calibration and prediction, to identifying model errors was demonstrated at field, regional, and country scale. When properly applied, Bayesian methods offer the opportunity to express uncertainties in simulated model outputs which may be state variables of interest such as phenology, biomass or yield. Quantifying these uncertainties are essential for decision analysis (Howard, 1988; Troost and Berger, 2014). For example, monetary investments for data-collection should target uncertainty reduction of the more uncertain factors that have a high impact on the state variable of interest. Accurate uncertainty assessment also plays an important role in agricultural insurance schemes. They offer a means of protection to farmers against financial losses in case of an event, for a fee. If the risk of a devastating event (for example, floods damaging crops) is wrongly assessed as too high, insurance costs may be prohibitively high. Farmers would have to pay a high fee to protect against an event that is unlikely to occur in reality. On the other hand, if such events are inaccurately assessed as low risk, farmers would have to bear much of the financial burden due to inaccurate estimates of maximum possible losses. Thus, accurately quantifying uncertainty using appropriate Bayesian methods in crop modelling has broad applications in developing suitable climate change adaptation strategies. Ultimately, as aptly stated by Bickel and Bratvold (2008), "Uncertainty quantification creates value only to the extent that it holds the possibility of changing a decision that would otherwise be made differently".

Bibliography

Adnan, A.A., Diels, J., Jibrin, J.M., Kamara, A.Y., Shaibu, A.S., Craufurd, P., Menkir, A., 2020. CERES-Maize model for simulating genotype-by-environment interaction of maize and its stability in the dry and wet savannas of Nigeria. Field Crops Research 253. doi:`10.1016/j.fcr.2020.107826`.

Ajami, N.K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research 43. doi:`10.1029/2005WR004745`.

Alber, M., Buganza Tepole, A., Cannon, W.R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W.W., Perdikaris, P., Petzold, L., Kuhl, E., 2019. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. npj Digital Medicine 2. doi:`10.1038/s41746-019-0193-y`, `arXiv:1910.01258`.

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. doi:`10.1016/j.eja.2016.09.016`.

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. doi:`10.1016/j.agrformet.2012.11.017`.

Asseng, S., Cao, W., Zhang, W., Ludwig, F., 2009. Chapter 20 - crop physiology, modelling and climate change: Impact and adaptation strategies, in: Sadras, V., Calderini, D. (Eds.), Crop Physiology. Academic Press, San Diego, pp. 511–543. doi:`10.1016/B978-0-12-374431-9.00020-7`.

Bassu, S., Brisson, N., Durand, J.L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.H., Kumar, N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? Global Change Biology 20, 2301–2320. doi:`10.1111/gcb.12520`.

Beirlant, J., Dudewicz, E., Györfi, L., van der Meulen, E., 1997. Nonparametric entropy estimation. An overview. International Journal of Mathematical and Statistical Sciences 6, 17–39.

Beven, K., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6, 279–298. doi:`10.1002/hyp.3360060305`.

Beven, K., Binley, A., 2014. GLUE: 20 years on. Hydrological Processes 28, 5897–5918. doi:`10.1002/hyp.10082`.

Beven, K., Smith, P., Freer, J., 2007. Comment on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todini. Journal of Hydrology 338, 315–318. doi:`10.1016/j.jhydrol.2007.02.023`.

Beven, K.J., Smith, P.J., Freer, J.E., 2008. So just why would a modeller choose to be incoherent? Journal of Hydrology 354, 15–32. doi:`10.1016/j.jhydrol.2008.02.007`.

BfN (Bundesamt für Naturschutz), 2017. Naturräumliche Gliederung Deutschlands (nach Meynen/Schmithüsen).

Bickel, J.E., Bratvold, R.B., 2008. From uncertainty quantification to decision making in the oil and gas industry. Energy Exploration and Exploitation 26, 311–325. doi:`10.1260/014459808787945344`.

Biernath, C., Gayler, S., Bittner, S., Klein, C., Högy, P., Fangmeier, A., Priesack, E., 2011. Evaluating the ability of four crop models to predict different environmental impacts on spring wheat grown in open-top chambers. European Journal of Agronomy 35, 71–82. doi:`10.1016/j.eja.2011.04.001`.

Borchers, H.W., 2020. pracma: Practical Numerical Math Functions. URL: `https://CRAN.R-project.org/package=pracma`. r package version 2.2.9.

Brooks, S.P., Gelman, A., 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7, 434–455. doi:`10.1080/10618600.1998.10474787`.

Brynjarsdóttir, J., O'Hagan, A., 2014. Learning about physical parameters: the importance of model discrepancy. Inverse Problems 30, 114007. doi:`10.1088/0266-5611/30/11/114007`.

Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), 1993. Ergiebigkeit der Grundwasservorkommen (ERGW1000). URL: `https://www.deutsche-rohstoffagentur.de/DE/Themen/Wasser/Projekte/abgeschlossen/Beratung/Had/had_projektbeschr.html?nn=1557832`.

van Bussel, L.G.J., Stehfest, E., Siebert, S., Müller, C., Ewert, F., 2015. Simulation of the phenological development of wheat and maize at the global scale. Global Ecology and Biogeography 24, 1018–1029. doi:`10.1111/geb.12351`.

Cao, Z.J., Wang, Y., Li, D.Q., 2016. Site-specific characterization of soil properties using multiple measurements from different test procedures at different locations – A Bayesian sequential updating approach. Engineering Geology , 150–161.doi:`10.1016/j.enggeo.2016.06.021`.

Casadebaig, P., Debaeke, P., Wallach, D., 2020. A new approach to crop model calibration: Phenotyping plus post-processing. Crop Science 60. doi:`10.1002/csc2.20016`.

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, 101–115. doi:`10.1016/j.agrformet.2010.09.007`.

Challinor, A.J., Koehler, A.K., Ramirez-Villegas, J., Whitfield, S., Das, B., 2016. Current warming will reduce yields unless maize breeding and seed systems adapt immediately. Nature Climate Change 6, 954–958. doi:`10.1038/nclimate3061`.

Chenu, K., Porter, J.R., Martre, P., Basso, B., Chapman, S.C., Ewert, F., Bindi, M., Asseng, S., 2017. Contribution of crop models to adaptation in wheat. Trends in Plant Science 22, 472–490. doi:`10.1016/j.tplants.2017.02.003`.

Clark, J.S., 2003. Uncertainty and variability in demography and population growth: A hierarchical approach. Ecology 84, 1370–1381. doi:`10.1890/0012-9658(2003)084[1370:UAVIDA]2.0.CO;2`.

Coelho, A.P., Dalri, A.B., Fischer Filho, J.A., de Faria, R.T., Silva, L.S., Gomes, R.P., 2020. Calibration and evaluation of the DSSAT/Canegro model for sugarcane cultivars under irrigation managements. Revista Brasileira de Engenharia Agricola e Ambiental 24, 52–58. doi:`10.1590/1807-1929/agriambi.v24n1p52-58`.

Craufurd, P.Q., Vadez, V., Jagadish, S.K., Prasad, P.V., Zaman-Allah, M., 2013. Crop science experiments designed to inform crop modeling. Agricultural and Forest Meteorology 170, 8–18. doi:`10.1016/j.agrformet.2011.09.003`.

Cuntz, M., Mai, J., Zink, M., Thober, S., Kumar, R., Schäfer, D., Schrön, M., Craven, J., Rakovec, O., Spieler, D., Prykhodko, V., Dalmasso, G., Musuuza, J., Langenberg, B., Attinger, S., Samaniego, L., 2015. Computationally inexpensive identification of noninformative model parameters by sequential screening. Water Resources Research 51, 6417–6441. doi:`10.1002/2015WR016907`.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., Rieckermann, J., 2013. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. Hydrology and Earth System Sciences 17, 4209–4225. doi:`10.5194/hess-17-4209-2013`.

Dietze, M.C., Lebauer, D.S., Kooper, R., 2013. On improving the communication between models and data. Plant, Cell and Environment 36, 1575–1585. doi:`10.1111/pce.12043`.

Donohue, R.J., Roderick, M.L., McVicar, T.R., 2007. On the importance of including vegetation dynamics in Budyko's hydrological model. Hydrology and Earth System Sciences 11, 983–995. doi:`10.5194/hess-11-983-2007`.

Droutsas, I., Challinor, A.J., Deva, C.R., Wang, E., 2022. Integration of machine learning into process-based modelling to improve simulation of complex crop responses. in silico Plants 4, 1–16. doi:10.1093/insilicoplants/diac017.

Dumont, B., Leemans, V., Mansouri, M., Bodson, B., Destain, J.P., Destain, M.F., 2014. Parameter identification of the stics crop model, using an accelerated formal mcmc approach. Environmental Modelling & Software 52, 121–135. doi:10.1016/j.envsoft.2013.10.022.

Duncan, W.G., 1971. Leaf Angles, Leaf Area, and Canopy Photosynthesis 1. Crop Science 11, 482–485. doi:10.2135/cropsci1971.0011183x001100040006x.

Duong, T., 2021. ks: Kernel Smoothing. URL: https://CRAN.R-project.org/package=ks. r package version 1.12.0.

Durand, J.L., Delusca, K., Boote, K., Lizaso, J., Manderscheid, R., Weigel, H.J., Ruane, A.C., Rosenzweig, C., Jones, J., Ahuja, L., Anapalli, S., Basso, B., Baron, C., Bertuzzi, P., Biernath, C., Deryng, D., Ewert, F., Gaiser, T., Gayler, S., Heinlein, F., Kersebaum, K.C., Kim, S.H., Müller, C., Nendel, C., Olioso, A., Priesack, E., Villegas, J.R., Ripoche, D., Rötter, R.P., Seidel, S.I., Srivastava, A., Tao, F., Timlin, D., Twine, T., Wang, E., Webber, H., Zhao, Z., 2018. How accurately do maize crop models simulate the interactions of atmospheric CO2 concentration levels with limited water supply on water use and yield? European Journal of Agronomy 100, 67–75. doi:10.1016/j.eja.2017.01.002.

DWD Climate Data Center (CDC), 2019. Phenological observations of crops from sowing to harvest (immediate reporters, historical), Version v006. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/phenology/immediate_reporters/crops/historical/DESCRIPTION_obsgermany-phenology-immediate_reporters-crops-historical.pdf.

Eshonkulov, R., Poyda, A., Ingwersen, J., Wizemann, H.D., Weber, T.K.D., Kremer, P., Högy, P., Pulatov, A., Streck, T., 2019. Evaluating multi-year, multi-site data on the energy balance closure of eddy-covariance flux measurements at cropland sites in southwestern Germany. Biogeosciences 16, 521–540. URL: https://bg.copernicus.org/articles/16/521/2019/, doi:10.5194/bg-16-521-2019.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., Kuczera, G., 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resources Research 50, 2350–2375. doi:10.1002/2013WR014185.

Falconnier, G.N., Corbeels, M., Boote, K.J., Affholder, F., Adam, M., MacCarthy, D.S., Ruane, A.C., Nendel, C., Whitbread, A.M., Justes, r., Ahuja, L.R., Akinseye, F.M., Alou, I.N., Amouzou, K.A., Anapalli, S.S., Baron, C., Basso, B., Baudron, F., Bertuzzi, P., Challinor, A.J., Chen, Y., Deryng, D., Elsayed, M.L., Faye, B., Gaiser, T., Galdos, M., Gayler, S., Gerardeaux, E., Giner, M., Grant, B., Hoogenboom, G., Ibrahim, E.S., Kamali, B., Kersebaum, K.C., Kim, S.H., van der Laan, M., Leroux, L., Lizaso, J.I., Maestrini, B., Meier, E.A., Mequanint, F., Ndoli, A., Porter, C.H., Priesack, E., Ripoche, D., Sida, T.S., Singh, U., Smith, W.N., Srivastava, A., Sinha, S., Tao, F., Thorburn, P.J., Timlin, D., Traore, B., Twine, T., Webber, H., 2020. Modelling climate change impacts on maize yields under low nitrogen input conditions in sub-saharan africa. Global Change Biology 26, 5942–5964. doi:10.1111/gcb.15261.

Fer, I., Shiklomanov, A., Novick, K.A., Gough, C.M., Altaf Arain, M., Chen, J., Murphy, B., Desai, A.R., Dietze, M.C., 2021. Capturing site-to-site variability through hierarchical bayesian calibration of a process-based dynamic vegetation model. bioRxiv URL: https://www.biorxiv.org/content/early/2021/04/29/2021.04.28.441243, doi:10.1101/2021.04.28.441243.

Fu, F., Zhang, T., 2016. A new model for solving time-cost-quality trade-off problems in construction. PLoS ONE 11, 1–15. doi:10.1371/journal.pone.0167142.

Fukui, S., Ishigooka, Y., Kuwagata, T., Hasegawa, T., 2015. A methodology for estimating phenological parameters of rice cultivars utilizing data from common variety trials. Journal of Agricultural Meteorology 71, 77–89. doi:10.2480/agrmet.D-14-00042.

Gao, Y., Wallach, D., Hasegawa, T., Tang, L., Zhang, R., Asseng, S., Kahveci, T., Liu, L., He, J., Hoogenboom, G., 2021. Evaluation of crop model prediction and uncertainty using bayesian parameter estimation and bayesian model averaging. Agricultural and Forest Meteorology 311, 108686. doi:10.1016/j.agrformet.2021.108686.

Gao, Y., Wallach, D., Liu, B., Dingkuhn, M., Boote, K.J., Singh, U., Asseng, S., Kahveci, T., He, J., Zhang, R., Confalonieri, R., Hoogenboom, G., 2020. Comparison of three calibration methods for modeling rice phenology. Agricultural and Forest Meteorology 280, 107785. doi:`10.1016/j.agrformet.2019.107785`.

Gayler, S., Wang, E., Priesack, E., Schaaf, T., Maidl, F.X., 2002. Modeling biomass growth, N-uptake and phenological development of potato crop. Geoderma 105, 367–383. doi:`10.1016/S0016-7061(01)00113-6`.

Gelman, A., 2006a. Multilevel (hierarchical) modeling: What it can and cannot do. Technometrics 48, 432–435. doi:`10.1198/004017005000000661`.

Gelman, A., 2006b. Prior distributions for variance parameters in hierarchical models. Bayesian analysis 1, 515–534. doi:`10.1214/06-BA117A`.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis (Tests in Statistical Science). 3 ed., Chapman & Hal l/CRC. doi:`10.1201/b16018`.

Gelman, A., Roberts, G.O., Gilks, R.W., 1996. Efficient Metropolis jumping rules, in: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics. Oxford University Press, pp. 599–607.

Gelman, A., Rubin, D., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7, 457–472. doi:`10.1214/ss/1177011136`.

van Genuchten, M.T., 1980. A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. Soil Science Society of America Journal 44, 892–898. doi:`10.2136/sssaj1980.03615995004400050002x`.

Good, I.J., 1952. Rational Decisions. Journal of the Royal Statistical Society. Series B (Methodological) 14, 107–114. doi:`10.1111/j.2517-6161.1952.tb00104.x`.

Hansen, S., Jensen, H., Nielsen, N., Svendsen, H., 1990. DAISY-Soil Plant Atmosphere System Model. Technical Report. The National Agency for Environmental Protection. Copenhagen, Denmark.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 100–108. doi:`10.2307/2346830`.

Hastings, W.K., 1970. Monte carlo sampling methods using markov chains and their applications. Biometrika 57, 97–109. doi:`10.1093/biomet/57.1.97`.

He, D., Wang, E., Wang, J., Lilley, J., Luo, Z., Pan, X., Pan, Z., Yang, N., 2017a. Uncertainty in canola phenology modelling induced by cultivar parameterization and its impact on simulated yield. Agricultural and Forest Meteorology 232, 163–175. doi:`10.1016/j.agrformet.2016.08.013`.

He, D., Wang, E., Wang, J., Robertson, M.J., 2017b. Data requirement for effective calibration of process-based crop models. Agricultural and Forest Meteorology 234-235, 136–148. doi:`10.1016/j.agrformet.2016.12.015`.

He, J., Jones, J.W., Graham, W.D., Dukes, M.D., 2010. Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method. Agricultural Systems 103, 256–264. doi:`10.1016/j.agsy.2010.01.006`.

Heinlein, F., Biernath, C., Klein, C., Thieme, C., Priesack, E., 2017. Evaluation of Simulated Transpiration from Maize Plants on Lysimeters. Vadose Zone Journal 16, vzj2016.05.0042. doi:`10.2136/vzj2016.05.0042`.

Höge, M., Guthke, A., Nowak, W., 2020. Bayesian model weighting: The many faces of model averaging. Water 12. doi:`10.3390/w12020309`.

Howard, R.A., 1988. Uncertainty about Probability: A Decision Analysis Perspective. Risk Analysis 8, 91–98. doi:`10.1111/j.1539-6924.1988.tb01156.x`.

Hsueh, H.F., Guthke, A., Wöhling, T., Nowak, W., 2022. Diagnosis of model errors with a sliding time-window bayesian analysis. Water Resources Research 58, e2021WR030590. doi:`10.1029/2021WR030590`.

Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W., 2015. Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. Agricultural and Forest Meteorology 204, 106–121. doi:`10.1016/j.agrformet.2015.02.001`.

Huang, X., Huang, G., Yu, C., Ni, S., Yu, L., 2017. A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging. Field Crops Research 211, 114–124. doi:`10.1016/j.fcr.2017.06.011`.

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, 355–365. doi:`10.1198/108571108X335855`.

Iizumi, T., Yokozawa, M., Nishimori, M., 2009. Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach. Agricultural and Forest Meteorology 149, 333–348. doi:`10.1016/j.agrformet.2008.08.015`.

Ingwersen, J., Högy, P., Wizemann, H., Warrach-Sagi, K., Streck, T., 2018. Coupling the land surface model Noah-MP with the generic crop growth model Gecros: Model description, calibration and validation. Agricultural and Forest Meteorology 262, 322–339. doi:`10.1016/j.agrformet.2018.06.023`.

Iooss, B., Veiga, S.D., Janon, A., Pujol, G., with contributions from Baptiste Broto, Boumhaout, K., Delage, T., Amri, R.E., Fruth, J., Gilquin, L., Guillaume, J., Idrissi, M.I., Le Gratiet, L., Lemaitre, P., Marrel, A., Meynaoui, A., Nelson, B.L., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati, T., Weber, F., 2021. sensitivity: Global Sensitivity Analysis of Model Outputs. URL: `https://CRAN.R-project.org/package=sensitivity`. r package version 1.24.0.

Jarquín, D., Pérez-Elizalde, S., Burgueño, J., Crossa, J., 2016. A hierarchical bayesian estimation model for multienvironment plant breeding trials in successive years. Crop Science 56, 2260–2276. doi:`10.2135/cropsci2015.08.0475`.

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., Ritchie, J., 2003. The dssat cropping system model. European Journal of Agronomy 18, 235–265. doi:`10.1016/S1161-0301(02)00107-7`.

Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017. Brief history of agricultural systems modeling. Agricultural Systems 155, 240–254. doi:`10.1016/j.agsy.2016.05.014`.

Kavetski, D., Kuczera, G., Franks, S.W., 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research 42. doi:`10.1029/2005WR004368`.

Kavetski, D., Kuczera, G., Franks, S.W., 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resources Research 42. doi:`10.1029/2005WR004376`.

Kellner, K., 2021. jagsui: A wrapper around rjags to streamline jags analyses. URL: `https://CRAN.R-project.org/package=jagsUI`. r package version 1.5.2.

Kimball, B.A., Boote, K.J., Hatfield, J.L., Ahuja, L.R., Stockle, C., Archontoulis, S., Baron, C., Basso, B., Bertuzzi, P., Constantin, J., Deryng, D., Dumont, B., Durand, J.L., Ewert, F., Gaiser, T., Gayler, S., Hoffmann, M.P., Jiang, Q., Kim, S.H., Lizaso, J., Moulin, S., Nendel, C., Parker, P., Palosuo, T., Priesack, E., Qi, Z., Srivastava, A., Stella, T., Tao, F., Thorp, K.R., Timlin, D., Twine, T.E., Webber, H., Willaume, M., Williams, K., 2019. Simulation of maize evapotranspiration: An inter-comparison among 29 maize models. Agricultural and Forest Meteorology 271, 264–284. doi:`10.1016/j.agrformet.2019.02.037`.

Klein, C., Biernath, C., Heinlein, F., Thieme, C., Gilgen, A.K., Zeeman, M., Priesack, E., 2017. Vegetation Growth Models Improve Surface Layer Flux Simulations of a Temperate Grassland. Vadose Zone Journal 16, 1–19. URL: `https://onlinelibrary.wiley.com/doi/abs/10.2136/vzj2017.03.0052`, doi:`10.2136/vzj2017.03.0052`.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology 331, 161–177. doi:`10.1016/j.jhydrol.2006.05.010`.

Kumar, S.V., Mocko, D.M., Wang, S., Peters-Lidard, C.D., Borak, J., 2019. Assimilation of Remotely Sensed Leaf Area Index into the Noah-MP Land Surface Model: Impacts on Water and Carbon Fluxes and States over the Continental United States. Journal of Hydrometeorology 20, 1359–1377. doi:`10.1175/JHM-D-18-0237.1`.

Kumudini, S., Andrade, F.H., Boote, K.J., Brown, G.A., Dzotsi, K.A., Edmeades, G.O., Gocken, T., Goodwin, M., Halter, A.L., Hammer, G.L., Hatfield, J.L., Jones, J.W., Kemanian, A.R., Kim, S.H., Kiniry, J., Lizaso, J.I., Nendel, C., Nielsen, R.L., Parent, B., Stoeckle, C.O., Tardieu, F., Thomison, P.R., Timlin, D.J., Vyn, T.J., Wallach, D., Yang, H.S., Tollenaar, M., 2014. Predicting maize phenology: Intercomparison of functions for developmental response to temperature. Agronomy Journal 106, 2087–2097. doi:`10.2134/agronj14.0200`.

Laloy, E., Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. Water Resources Research 48. doi:`10.1029/2011WR010608`.

Lamboni, M., Makowski, D., Lehuger, S., Gabrielle, B., Monod, H., 2009. Multivariate global sensitivity analysis for dynamic crop models. Field Crops Research 113, 312–320. doi:`10.1016/j.fcr.2009.06.007`.

Lamsal, A., Welch, S.M., White, J.W., Thorp, K.R., Bello, N.M., 2018. Estimating parametric phenotypes that determine anthesis date in Zea mays: Challenges in combining ecophysiological models with genetics. PLoS ONE 13, 1–23. doi:`10.1371/journal.pone.0195841`.

Li, X., Yeluripati, J., Jones, E.O., Uchida, Y., Hatano, R., 2015. Hierarchical bayesian calibration of nitrous oxide (n2o) and nitrogen monoxide (no) flux module of an agro-ecosystem model: Ecosse. Ecological Modelling 316, 14–27. doi:`10.1016/j.ecolmodel.2015.07.020`.

Liu, J.S., 2008. Monte Carlo Strategies in Scientific Computing. Springer Science & Business Media. URL: `https://link.springer.com/book/10.1007/978-0-387-76371-2`.

Liu, K., Harrison, M.T., Archontoulis, S.V., Huth, N., Yang, R., Liu, D.L., Yan, H., Meinke, H., Huber, I., Feng, P., Ibrahim, A., Zhang, Y., Tian, X., Zhou, M., 2021. Climate change shifts forward flowering and reduces crop waterlogging stress. Environmental Research Letters 16, 094017. doi:`10.1088/1748-9326/ac1b5a`.

Liu, K., Harrison, M.T., Ibrahim, A., Manik, S.M., Johnson, P., Tian, X., Meinke, H., Zhou, M., 2020. Genetic factors increasing barley grain yields under soil waterlogging. Food and Energy Security 9, 1–12. doi:`10.1002/fes3.238`.

Lobell, D.B., Asseng, S., 2017. Comparing estimates of climate change impacts from process- based and statistical crop models. Environmental Research Letters 12. doi:`10.1088/1748-9326/aa518a`.

Locher, R., 2020. IDPmisc: Utilities of Institute of Data Analyses and Process Design. URL: `https://CRAN.R-project.org/package=IDPmisc`. r package version 1.1.20.

Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R.P., Ruane, A.C., Semenov, M.A., Wallach, D., Wang, E., Alderman, P.D., Kassie, B.T., Biernath, C., Basso, B., Cammarano, D., Challinor, A.J., Doltra, J., Dumont, B., Rezaei, E.E., Gayler, S., Kersebaum, K.C., Kimball, B.A., Koehler, A.K., Liu, B., O'Leary, G.J., Olesen, J.E., Ottman, M.J., Priesack, E., Reynolds, M., Stratonovitch, P., Streck, T., Thorburn, P.J., Waha, K., Wall, G.W., White, J.W., Zhao, Z., Zhu, Y., 2017. Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. Field Crops Research 202, 5–20. doi:`10.1016/j.fcr.2016.05.001`.

Makowski, D., 2017. A simple Bayesian method for adjusting ensemble of crop model outputs to yield observations. European Journal of Agronomy 88, 76–83. doi:`10.1016/j.eja.2015.12.012`.

Makowski, D., Hillier, J., Wallach, D., Andrieu, B., Jeuffroy, M.H., 2006. Parameter Estimation for Crop Models, in: Working with Dynamic Crop Models. Elsevier.

Makowski, D., Jeuffroy, M.H., Guérif, M., 2004. Bayesian methods for updating crop-model predictions, applications for predicting biomass and grain protein content, in: van Boekel, M., Stein, A., van Bruggen, A. (Eds.), Volume 3 Bayesian Statistics and Quality Modelling in the Agro-Food Production Chain. Frontis.

Makowski, D., Wallach, D., Tremblay, M., 2002. Using a Bayesian approach to parameter estimation; comparison of the GLUE and MCMC methods. Agronomie 22, 191 – 203. doi:`10.1051/agro:2002007`.

Mantovan, P., Todini, E., 2006. Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. Journal of Hydrology 330, 368–381. doi:`10.1016/j.jhydrol.2006.04.046`.

McCormick, R.F., Truong, S.K., Rotundo, J., Gaspar, A.P., Kyle, D., Van Eeuwijk, F., Messina, C.D., 2021. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. In Silico Plants 3. doi:`10.1093/insilicoplants/diab004`.

McMillan, H.K., Westerberg, I.K., Krueger, T., 2018. Hydrological data uncertainty and its implications. Wiley Interdisciplinary Reviews: Water 5. doi:`10.1002/wat2.1319`.

Mehdi, B., Ludwig, R., Lehner, B., 2015. Evaluating the impacts of climate change and crop land use change on streamflow, nitrates and phosphorus: A modeling study in Bavaria. Journal of Hydrology: Regional Studies 4 Part B, 60–90. doi:`10.1016/j.ejrh.2015.04.009`.

Meier, U., 1997. Growth Stages of Mono- and Dicotyledonous Plants. Open Agrar Repositorium, Julius Kühn Institute,Quedlinburg. doi:`10.5073/20180906-074619`.

Meier, U., 2018. Growth stages of mono- and dicotyledonous plants: Bbch monograph. doi:`10.5073/20180906-074619`.

Menzel, A., Sparks, T.H., Estrella, N., Roy, D.B., 2006. Altered geographic and temporal variability in phenology in response to climate change. Global Ecology and Biogeography 15, 498–504. doi:`10.1111/j.1466-822X.2006.00247.x`.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21. URL: `https://bayes.wustl.edu/Manual/EquationOfState.pdf`.

Microsoft, Weston, S., 2019. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. URL: `https://CRAN.R-project.org/package=doParallel`. r package version 1.0.15.

Microsoft, Weston, S., 2020. foreach: Provides Foreach Looping Construct. URL: `https://CRAN.R-project.org/package=foreach`. r package version 1.5.0.

Minet, J., Laloy, E., Tychon, B., François, L., 2015. Bayesian inversions of a dynamic vegetation model at four European grassland sites. Biogeosciences 12, 2809–2829. doi:`10.5194/bg-12-2809-2015`.

Mo, X., Beven, K., 2004. Multi-objective parameter conditioning of a three-source wheat canopy model. Agricultural and Forest Meteorology 122, 39–63. doi:`10.1016/j.agrformet.2003.09.009`.

Moore, L.M., Lauenroth, W.K., 2017. Differential effects of temperature and precipitation on early- vs. late-flowering species. Ecosphere 8, e01819. doi:`10.1002/ecs2.1819`.

Morris, M., 1991. Factorial Sampling Plans for Preliminary Computational Experiments. Technometrics 33, 161–174. doi:`10.2307/1269043`.

Motavita, D., Chow, R., Guthke, A., Nowak, W., 2019. The comprehensive differential split-sample test: A stress-test for hydrological model robustness under climate variability. Journal of Hydrology 573, 501–515. doi:`10.1016/j.jhydrol.2019.03.054`.

Mualem, Y., 1976. A New Model for Predicting the Hydraulic Conductivity of Unsaturated Porous Media. Water Resources Research 12, 513–522. doi:`10.1029/WR012i003p00513`.

Muñoz Sabater, J., 2019. ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 04-Nov-2020). URL: `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview`, doi:`10.24381/cds.e2161bac`.

Nearing, G.S., Crow, W.T., Thorp, K.R., Moran, M.S., Reichle, R.H., Gupta, H.V., 2012. Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. Water Resources Research 48. doi:`10.1029/2011WR011420`.

van Oijen, M., 2017. Bayesian Methods for Quantifying and Reducing Uncertainty and Error in Forest Models. Current Forestry Reports 3, 269–280. doi:`10.1007/s40725-017-0069-9`.

van Oijen, M., Thomson, A., Ewert, F., 2009. Spatial upscaling of process-based vegetation models : An overview of common methods and a case-study for the U.K. StatGIS 2009 .

Oliveira, F.A., Jones, J.W., Pavan, W., Bhakta, M., Vallejos, C.E., Correll, M.J., Boote, K.J., Fernandes, J.M., Hölbig, C.A., Hoogenboom, G., 2021. Incorporating a dynamic gene-based process module into a crop simulation model. In Silico Plants 3. doi:`10.1093/insilicoplants/diab011`.

Oluwaranti, A., Fakorede, M.A.B., Adeboye, F.A., 2015. Maturity groups and phenology of maize in a rainforest location. International Journal of Agriculture Innovations and Research 4, 124–127. URL: `https://ijair.org/administrator/components/com_jresearch/files/publications/IJAIR_1486_Final.pdf`.

Oravecz, Z., Huentelman, M., Vandekerckhove, J., 2016. Sequential Bayesian Updating for Big Data, in: Jones, M.N. (Ed.), Big Data in Cognitive Science. Psychology Press, New York. chapter 2, p. 21. doi:`10.4324/9781315413570`.

Panayi, E., Peters, G.W., Kyriakides, G., 2017. Statistical modelling for precision agriculture: A case study in optimal environmental schedules for Agaricus Bisporus production via variable domain functional regression. PLoS ONE 12. doi:`10.1371/journal.pone.0181921`.

Parent, B., Leclere, M., Lacube, S., Semenov, M.A., Welcker, C., Martre, P., Tardieu, F., 2018. Maize yields over Europe may increase in spite of climate change, with an appropriate use of the genetic variability of flowering time. Proceedings of the National Academy of Sciences of the United States of America 115, 10642–10647. doi:`10.1073/pnas.1720716115`.

Parker, P.S., Shonkwiler, J.S., Aurbacher, J., 2017. Cause and consequence in maize planting dates in germany. Journal of Agronomy and Crop Science 203, 227–240. doi:`10.1111/jac.12182`.

Pasquel, D., Roux, S., Richetti, J., Cammarano, D., Tisseyre, B., Taylor, J.A., 2022. A review of methods to evaluate crop model performance at multiple and changing spatial scales. Precision Agriculture 23, 1489—-1513. doi:`10.1007/s11119-022-09885-4`.

Pathak, T.B., Jones, J.W., Fraisse, C.W., Wright, D., Hoogenboom, G., 2012. Uncertainty analysis and parameter estimation for the CSM-CROPGRO-cotton model. Agronomy Journal 104, 1363–1373. doi:`10.2134/agronj2011.0349`.

Patrick, L.D., Ogle, K., Tissue, D.T., 2009. A hierarchical bayesian approach for estimation of photosynthetic parameters of c3 plants. Plant, Cell & Environment 32, 1695–1709. doi:`10.1111/j.1365-3040.2009.02029.x`.

Peng, B., Guan, K., Tang, J., Ainsworth, E.A., Asseng, S., Bernacchi, C.J., Cooper, M., Delucia, E.H., Elliott, J.W., Ewert, F., Grant, R.F., Gustafson, D.I., Hammer, G.L., Jin, Z., Jones, J.W., Kimm, H., Lawrence, D.M., Li, Y., Lombardozzi, D.L., Marshall-Colon, A., Messina, C.D., Ort, D.R., Schnable, J.C., Vallejos, C.E., Wu, A., Yin, X., Zhou, W., 2020. Towards a multiscale crop modelling framework for climate change adaptation assessment. Nature Plants 6, 338–348. doi:`10.1038/s41477-020-0625-3`.

Pingali, P.L., 2012. Green revolution: Impacts, limits, andthe path ahead. Proceedings of the National Academy of Sciences of the United States of America 109, 12302–12308. doi:`10.1073/pnas.0912953109`.

Plummer, M., 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. Coda: Convergence diagnosis and output analysis for mcmc. R News 6, 7–11. URL: `https://journal.r-project.org/archive/`.

Porter, J.R., Xie, L., Challinor, A.J., Cochrane, K., Howden, S.M., Iqbal, M.M., Lobell, D.B., Travasso, M.I., Aggarwal, P., Hakala, K., Jordan, J., 2015. Food Security and Food Production Systems, in: Field, C.B., Barros, V.R., Dokken, D.J., Mach, K.J., Mastrandrea, M.D. (Eds.), Climate Change 2014 Impacts, Adaptation, and Vulnerability. Cambridge University Press, Cambridge, pp. 485–534. doi:`10.1017/CBO9781107415379.012`.

Potgieter, A.B., Zhao, Y., Zarco-Tejada, P.J., Chenu, K., Zhang, Y., Porker, K., Biddulph, B., Dang, Y.P., Neale, T., Roosta, F., Chapman, S., 2021. Evolution and application of digital technologies to predict crop type and crop phenology in agriculture. In Silico Plants 3. doi:`10.1093/insilicoplants/diab017`.

Priesack, E., 2006. Expert-N Dokumentation der Modellbibliothek FAM – Bericht 60. Technical Report. GSF-Forschungszentrum fuer Umwelt und Gesundheit; Institut fuer Bodenoekologie.

Priesack, E., Gayler, S., Hartmann, H.P., 2006. The impact of crop growth sub-model choice on simulated water and nitrogen balances. Nutrient Cycling in Agroecosystems 75, 1–13. doi:`10.1007/s10705-006-9006-1`.

Qiu, T., Song, C., Clark, J.S., Seyednasrollah, B., Rathnayaka, N., Li, J., 2020. Understanding the continuous phenological development at daily time step with a bayesian hierarchical space-time model: impacts of climate change and extreme weather events. Remote Sensing of Environment 247, 111956. doi:`10.1016/j.rse.2020.111956`.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rajib, A., Kim, I.L., Golden, H.E., Lane, C.R., Kumar, S.V., Yu, Z., Jeyalakshmi, S., 2020. Watershed modeling with remotely sensed big data: Modis leaf area index improves hydrology and water quality predictions. Remote Sensing 12. doi:`10.3390/rs12132148`.

Razavi, S., Tolson, B.A., 2013. An efficient framework for hydrologic model calibration on long data periods. Water Resources Research 49, 8418–8431. doi:`10.1002/2012WR013442`.

Reichert, P., Ammann, L., Fenicia, F., 2021. Potential and challenges of investigating intrinsic uncertainty of hydrological models with stochastic, time-dependent parameters. Water Resources Research 57, e2020WR028400. doi:`10.1029/2020WR028400`.

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. Water Resources Research 45. doi:`10.1029/2009WR007814`.

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resources Research 48. doi:`10.1029/2011WR011391`.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46. doi:`10.1029/2009WR008328`.

Rettie, F.M., Gayler, S., Weber, T.K., Tesfaye, K., Streck, T., 2022. Climate change impact on wheat and maize growth in Ethiopia: A multi-model uncertainty analysis. PLoS ONE 17. doi:`10.1371/journal.pone.0262951`.

Rötter, R.P., Carter, T.R., Olesen, J.E., Porter, J.R., 2011. Crop–climate models need an overhaul. Nature Climate Change 1, 175–177. doi:`10.1038/nclimate1152`.

Samadi, S., Tufford, D.L., Carbone, G.J., 2018. Estimating hydrologic model uncertainty in the presence of complex residual error structures. Stochastic Environmental Research and Risk Assessment 32, 1259–1281. doi:`10.1007/s00477-017-1489-6`.

Schöniger, A., Wöhling, T., Nowak, W., 2015. A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. Water Resources Research 51, 7524–7546. doi:`10.1002/2015WR016918`.

Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W., 2014. Model selection on solid ground: Rigorous comparison of nine ways to evaluate b ayesian model evidence. Water resources research 50, 9484–9513. doi:`10.1002/2014WR016062`.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resources Research 46. doi:`10.1029/2009WR008933`.

Seidel, S.J., Palosuo, T., Thorburn, P., Wallach, D., 2018. Towards improved calibration of crop models – Where are we now and where should we go? European Journal of Agronomy 94, 25–35. doi:`10.1016/j.eja.2018.01.006`.

Senf, C., Pflugmacher, D., Heurich, M., Krueger, T., 2017. A Bayesian hierarchical model for estimating spatial and temporal variation in vegetation phenology from Landsat time series. Remote Sensing of Environment 194, 155–160. doi:`10.1016/j.rse.2017.03.020`.

Sexton, J., Everingham, Y., Inman-Bamber, G., 2016. A theoretical and real world evaluation of two Bayesian techniques for the calibration of variety parameters in a sugarcane crop model. Environmental Modelling & Software 83, 126–142. doi:`10.1016/j.envsoft.2016.05.014`.

Sexton, J.D., 2015. Bayesian Statistical Calibration of Variety Parameters in a Sugarcane Crop Model. Ph.D. thesis. URL: `http://researchonline.jcu.edu.au/41338/1/41338-sexton-2015-thesis.pdf`.

Sharma, A., Jain, A., Gupta, P., Chowdary, V., 2021. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. IEEE Access 9, 4843–4873. doi:`10.1109/ACCESS.2020.3048415`.

Shi, Y., Wang, J., Qin, J., Qu, Y., 2015. An upscaling algorithm to obtain the representative ground truth of LAI time series in heterogeneous land surface. Remote Sensing 7, 12887–12908. doi:`10.3390/rs71012887`.

Siad, S.M., Iacobellis, V., Zdruli, P., Gioia, A., Stavi, I., Hoogenboom, G., 2019. A review of coupled hydrologic and crop growth models. Agricultural Water Management 224, 105746. doi:`10.1016/j.agwat.2019.105746`.

Siebert, S., Ewert, F., 2012. Spatio-temporal patterns of phenological development in Germany in relation to temperature and day length. Agricultural and Forest Meteorology 152, 44–57. doi:`10.1016/j.agrformet.2011.08.007`.

Sievert, C., 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC. URL: `https://plotly-r.com`.

Šimůnek, Šejna, J.M., van Genuchten, M.T., 1998. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. Version 1.0. Agricultural Research Service, US Department of Agriculture .

Soltani, A., Bakker, M.M., Veldkamp, A., Stoorvogel, J.J., 2016. Comparison of three modelling approaches to simulate regional crop yield: A case study of winter wheat yield in western Germany. Journal of Agricultural Science and Technology 18, 191–206. URL: `http://jast.modares.ac.ir/article-23-2579-en.html`.

Su, Y.S., Yajima, M., 2020. R2jags: Using r to run jags. URL: `https://CRAN.R-project.org/package=R2jags`. r package version 0.6-1.

Tang, Y., Marshall, L., Sharma, A., Ajami, H., 2018. A Bayesian alternative for multi-objective ecohydrological model specification. Journal of Hydrology 556, 25–38. doi:`10.1016/j.jhydrol.2017.07.040`.

Tang, Y., Marshall, L., Sharma, A., Ajami, H., Nott, D.J., 2019. Ecohydrologic Error Models for Improved Bayesian Inference in Remotely Sensed Catchments. Water Resources Research 55, 4533–4549. doi:`10.1029/2019WR025055`.

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. doi:`10.1016/j.ecolmodel.2012.03.009`.

Teixeira, E.I., Zhao, G., de Ruiter, J., Brown, H., Ausseil, A.G., Meenken, E., Ewert, F., 2017. The interactions between genotype, management and environment in regional crop modelling. European Journal of Agronomy 88, 106–115. doi:`10.1016/j.eja.2016.05.005`.

Therond, O., Hengsdijk, H., Casellas, E., Wallach, D., Adam, M., Belhouchette, H., Oomen, R., Russell, G., Ewert, F., Bergez, J.E., Janssen, S., Wery, J., Van Ittersum, M.K., 2011. Using a cropping system model at regional scale: Low-data approaches for crop management information and model calibration. Agriculture, Ecosystems & Environment 142, 85–94. doi:`10.1016/j.agee.2010.05.007`.

Thijssen, B., Wessels, L.F.A., 2020. Approximating multivariate posterior distribution functions from Monte Carlo samples for sequential Bayesian inference. PLOS ONE 15, e0230101. doi:`10.1371/journal.pone.0230101`, `arXiv:1712.04200`.

Thomas, R.Q., Brooks, E.B., Jersild, A.L., Ward, E.J., Wynne, R.H., Albaugh, T.J., Dinon-Aldridge, H., Burkhart, H.E., Domec, J.C., Fox, T.R., Gonzalez-Benecke, C.A., Martin, T.A., Noormets, A., Sampson, D.A., Teskey, R.O., 2017. Leveraging 35 years of *Pinus taeda* research in the southeastern us to constrain forest carbon cycle predictions: regional data assimilation using ecosystem experiments. Biogeosciences 14, 3525–3547. doi:`10.5194/bg-14-3525-2017`.

Thompson, C.J., Kodikara, S., Burgman, M.A., Demirhan, H., Stone, L., 2019. Bayesian updating to estimate extinction from sequential observation data. Biological Conservation 229, 26–29. doi:`10.1016/j.biocon.2018.11.003`.

Tian, X., Minunno, F., Cao, T., Peltoniemi, M., Kalliokoski, T., Mäkelä, A., 2020. Extending the range of applicability of the semi-empirical ecosystem flux model preles for varying forest types and climate. Global Change Biology 26, 2923–2943. doi:`10.1111/gcb.14992`.

Troost, C., Berger, T., 2014. Dealing with uncertainty in agent-based simulation: Farm-level modeling of adaptation to climate change in southwest Germany. American Journal of Agricultural Economics 97, 833–854. doi:`10.1093/ajae/aau076`.

Vallejos, C.E., Jones, J.W., Bhakta, M.S., Gezan, S.A., Correll, M.J., 2022. Dynamic QTL-based ecophysiological models to predict phenotype from genotype and environment data. BMC Plant Biology 22. doi:`10.1186/s12870-022-03624-7`.

Van Oijen, M., Höglind, M., 2016. Toward a Bayesian procedure for using process-based models in plant breeding, with application to ideotype design. Euphytica 207, 627–643. doi:`10.1007/s10681-015-1562-5`.

Viswanathan, M., Scheidegger, A., Streck, T., Gayler, S., Weber, T.K., 2022a. Bayesian multi-level calibration of a process-based maize phenology model. Ecological Modelling 474, 110154. doi:`10.1016/j.ecolmodel.2022.110154`.

Viswanathan, M., Weber, T.K.D., Gayler, S., Mai, J., Streck, T., 2022b. A bayesian sequential updating approach to predict phenology of silage maize. Biogeosciences 19, 2187–2209. doi:`10.5194/bg-19-2187-2022`.

Vrugt, J., Bouten, W., Weerts, A., 2001. Information Content of Data for Identifying Soil Hydraulic Parameters from Outflow Experiments. Soil Science Society of America Journal 65, 19–27. doi:`10.2136/sssaj2001.65119x`.

Vrugt, J.A., ter Braak, C.J., Gupta, H.V., Robinson, B.A., 2009. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic Environmental Research and Risk Assessment 23, 1011–1026. doi:`10.1007/s00477-008-0274-y`.

Wallach, D., 2011. Crop model calibration: A statistical perspective. Agronomy Journal 103, 1144–1151. doi:`10.2134/agronj2010.0432`.

Wallach, D., Hwang, C., Correll, M.J., Jones, J.W., Boote, K., Hoogenboom, G., Gezan, S., Bhakta, M., Vallejos, C.E., 2018. A dynamic model with QTL covariables for predicting flowering time of common bean (Phaseolus vulgaris) genotypes. European Journal of Agronomy 101, 200–209. doi:`10.1016/j.eja.2018.10.003`.

Wallach, D., Keussayan, N., Brun, F., Lacroix, B., Bergez, J.E., 2012. Assessing the Uncertainty when Using a Model to Compare Irrigation Strategies. Agronomy Journal 104, 1274–1283. doi:`10.2134/agronj2012.0038`.

Wallach, D., Mearns, L.O., Ruane, A.C., Rötter, R.P., Asseng, S., 2016. Lessons from climate modeling on the design and use of ensembles for crop modeling. Climatic Change 139, 551–564. doi:`10.1007/s10584-016-1803-1`.

Wallach, D., Nissanka, S.P., Karunaratne, A.S., Weerakoon, 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. doi:`10.1016/j.eja.2016.05.013`.

Wallach, D., Palosuo, T., Thorburn, P., Gourdain, E., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hochman, Z., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S., 2021a. How well do crop modeling groups predict wheat phenology, given calibration data from the target population? European Journal of Agronomy 124, 126195. doi:`10.1016/j.eja.2020.126195`.

Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hiremath, S., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M., Jansson, P.E., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Lewan, E., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Padovan, G., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S.J., 2021b. The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environmental Modelling & Software 145, 105206. doi:`10.1016/j.envsoft.2021.105206`.

Wallach, D., Thorburn, P.J., 2017. Estimating uncertainty in crop model predictions: Current situation and future prospects. European Journal of Agronomy 88, A1–A7. doi:`10.1016/j.eja.2017.06.001`.

Wang, E., 1997. Development of a Generic Process-Oriented Model for Simulation of Crop Growth. Ph.D. thesis. Technische Universität München, Germany.

Wang, E., Brown, H.E., Rebetzke, G.J., Zhao, Z., Zheng, B., Chapman, S.C., 2019. Improving process-based crop models to better capture genotype×environment×management interactions. Journal of Experimental Botany 70, 2389–2401. doi:`10.1093/jxb/erz092`.

Wang, E., Engel, T., 1998. Simulation of phenological development of wheat crops. Agricultural Systems 58, 1–24. doi:`10.1016/S0308-521X(98)00028-6`.

Wang, E., Engel, T., 2000. SPASS: A generic process-oriented crop model with versatile windows interfaces. Environmental Modelling and Software 15, 179–188. doi:`10.1016/S1364-8152(99)00033-X`.

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R.P., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., Reynolds, M.P., Alderman, P.D., Aggarwal, P.K., Anothai, J., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G., Doltra, J., Fereres, E., Garcia-Vila, M., Gayler, S., Hoogenboom, G., Hunt, L.A., Izaurralde, R.C., Jabloun, M., Jones, C.D., Kersebaum, K.C., Koehler, A.K., Liu, L., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Palosuo, T., Priesack, E., Eyshi Rezaei, E., Ripoche, D., Ruane, A.C., Semenov, M.A., Shcherbak, I., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Thorburn, P., Waha, K., Wallach, D., Wang, Z., Wolf, J., Zhu, Y., Asseng, S., 2017a. The uncertainty of crop yield projections is reduced by improved temperature response functions. Nature Plants 3, 17102. URL: `10.1038/nplants.2017.102`, doi:`10.1038/nplants.2017.102`.

Wang, N., Wang, J., Wang, E., Yu, Q., Shi, Y., He, D., 2015. Increased uncertainty in simulated maize phenology with more frequent supra-optimal temperature under climate warming. European Journal of Agronomy 71, 19–33. doi:`10.1016/j.eja.2015.08.005`.

Wang, W., Ding, Y., Shao, Q., Xu, J., Jiao, X., Luo, Y., Yu, Z., 2017b. Bayesian multi-model projection of irrigation requirement and water use efficiency in three typical rice plantation region of China based on CMIP5. Agricultural and Forest Meteorology 232, 89–105. doi:`10.1016/j.agrformet.2016.08.008`.

Weber, T.K., Gerling, L., Reineke, D., Weber, S., Durner, W., Iden, S.C., 2018. Robust Inverse Modeling of Growing Season Net Ecosystem Exchange in a Mountainous Peatland: Influence of Distributional Assumptions on Estimated Parameters and Total Carbon Fluxes. Journal of Advances in Modeling Earth Systems 10, 1319–1336. doi:`10.1029/2017MS001044`.

Weber, T.K.D., Ingwersen, J., Högy, P., Poyda, A., Wizemann, H.D., Demyan, M.S., Bohm, K., Eshonkulov, R., Gayler, S., Kremer, P., Laub, M., Nkwain, Y.F., Troost, C., Witte, I., Reichenau, T., Berger, T., Cadisch, G., Müller, T., Fangmeier, A., Wulfmeyer, V., Streck, T., 2022. Multi-site, multi-crop measurements in the soil–vegetation–atmosphere continuum: a comprehensive dataset from two climatically contrasting regions in southwestern germany for the period 2009–2018. Earth System Science Data 14, 1153–1181. doi:`10.5194/essd-14-1153-2022`.

Wei, T., Simko, V., 2017. R package "corrplot": Visualization of a Correlation Matrix. URL: `https://github.com/taiyun/corrplot`. (Version 0.84).

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL: `https://ggplot2.tidyverse.org`.

de Wit, C., 1965. Photosynthesis of leaf canopies. PUDOC, Wageningen. URL: `https://library.wur.nl/WebQuery/wurpubs/fulltext/187115`.

Wittich, K.P., Liedtke, M., 2015. Shifts in plant phenology: A look at the sensitivity of seasonal phenophases to temperature in Germany. International Journal of Climatology 35, 3991–4000. doi:`10.1002/joc.4262`.

Wizemann, H.D., Ingwersen, J., Högy, P., Warrach-Sagi, K., Streck, T., Wulfmeyer, V., 2015. Three year observations of water vapor and energy fluxes over agricultural crops in two regional climates of Southwest Germany. Meteorologische Zeitschrift 24, 39–59. doi:`10.1127/metz/2014/0618`.

Wöhling, T., Gayler, S., Ingwersen, J., Streck, T., Vrugt, J.A., Priesack, E., 2012. Multiobjective calibration of coupled soil-vegetation-atmosphere models. IAHS-AISH Publication 355, 357–363.

Wöhling, T., Gayler, S., Priesack, E., Ingwersen, J., Wizemann, H.D., Högy, P., Cuntz, M., Attinger, S., Wulfmeyer, V., Streck, T., 2013. Multiresponse, multiobjective calibration as a diagnostic tool to compare accuracy and structural limitations of five coupled soil-plant models and clm3. 5. Water Resources Research 49, 8200–8221. doi:`10.1002/2013WR014536`.

Wöhling, T., Geiges, A., Nowak, W., Gayler, S., Högy, P., Wizemann, H., 2013. Towards Optimizing Experiments for Maximum-confidence Model Selection between Different Soil-plant Models. Procedia Environmental Sciences 19, 514–523. doi:`10.1016/j.proenv.2013.06.058`.

Wöhling, T., Schöniger, A., Gayler, S., Nowak, W., 2015. Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. Water Resources Research 51, 2825–2846. doi:`10.1002/2014WR016292`.

Wu, L., Feng, L., Zhang, Y., Gao, J., Wang, J., 2017. Comparison of five wheat models simulating phenology under different sowing dates and varieties. Agronomy Journal 109, 1280–1293. doi:`10.2134/agronj2016.10.0619`.

Xiong, W., Holman, I., Conway, D., Lin, E., Li, Y., 2008. A crop model cross calibration for use in regional climate impacts studies. Ecological Modelling 213, 365–380. doi:`10.1016/j.ecolmodel.2008.01.005`.

Xu, T., Valocchi, A.J., 2015. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. Water Resources Research 51, 9290–9311. doi:`10.1002/2015WR017912`.

Zak, S.K., Beven, K., Reynolds, B., 1997. Uncertainty in the estimation of critical loads: a practical methodology. Water, Air, and Soil Pollution 98, 297–316. doi:`10.1007/BF02047040`.

Zare, H., Weber, T.K., Ingwersen, J., Nowak, W., Gayler, S., Streck, T., 2022. Combining Crop Modeling with Remote Sensing Data Using a Particle Filtering Technique to Produce Real-Time Forecasts of Winter Wheat Yields under Uncertain Boundary Conditions. Remote Sensing 14. doi:`10.3390/rs14061360`.

Zhang, W., Arhonditsis, G.B., 2009. A Bayesian hierarchical framework for calibrating aquatic biogeochemical models. Ecological Modelling 220, 2142–2161. doi:`10.1016/j.ecolmodel.2009.05.023`.

Zhang, X., Maggioni, V., Rahman, A., Houser, P., Xue, Y., Sauer, T., Kumar, S., Mocko, D., 2020. The influence of assimilating leaf area index in a land surface model on global water fluxes and storages. Hydrology and Earth System Sciences 24, 3775–3788. doi:`10.5194/hess-24-3775-2020`.

Zhao, M., Peng, C., Xiang, W., Deng, X., Tian, D., Zhou, X., Yu, G., He, H., Zhao, Z., 2013. Plant phenological modeling and its application in global climate change research: overview and future challenges. Environmental Reviews , 1–14doi:`10.1139/er-2012-0036`.

Zheng, B., Chenu, K., Fernanda Dreccer, M., Chapman, S.C., 2012. Breeding for the future: What are the potential impacts of future frost and heat events on sowing and flowering time requirements for Australian bread wheat (Triticum aestivium) varieties? Global Change Biology 18, 2899–2914. doi:`10.1111/j.1365-2486.2012.02724.x`.

Supplementary Materials

Selected supplementary materials that have been referenced within the chapters are provided below.

S1 Sensitivity Analysis (chapter 3)

The Morris or elementary effects screening method (Morris, 1991) was used to conduct a qualitative global sensitivity analysis on phenological development of maize. Sensitivity analysis was only performed for site-year 6_2010 under the assumption that ranks of the most sensitive parameters would not change significantly due do the different weather and initial conditions in Kraichgau and the Swabian Alb. The sensitivity package in R (Iooss et al., 2021) was used. The one-at-a-time (OAT) design in the morris function was used to define the parameter vectors. A total of 11 parameters that influence phenological development in the SPASS model were pre-selected based on expert knowledge. Uniform parameter distributions with a range equal to three standard deviations from the expected value were used. It is noted that different distributions have been used for Bayesian calibration (platykurtic prior distribution) and sensitivity analysis (uniform distribution). However, this is assumed to have a limited influence in identifying the most sensitive parameters. Settings to the morris function were provided: 1000 samples, 10 levels and a grid jump-size of 2 units. Phenology was simulated using the SPASS model in XN5 software for all the proposed parameter vectors. The *morris* function was then used to estimate elementary effects (Cuntz et al., 2015; Morris, 1991) of phenological development at an interval of every 5 days within the growing season. The sensitivity measures, namely, the mean (μ^*) of the absolute value of the elementary effects of a parameter and the standard deviation (σ) were calculated on these days to evaluate parameter sensitivity over the growing season.

$$\mu^*_{\theta_i} = \frac{1}{N}\sum_{n=1}^{N} \mid ee_{n,\theta_i} \mid \tag{S1.1}$$

$$\sigma_{\theta_i} = \sqrt{\frac{\sum_{n=1}^{N}(ee_{n,\theta_i} - \mu_{\theta_i})^2}{N}} \tag{S1.2}$$

where $\mu^*_{\theta_i}$ and σ_{θ_i} are the μ^* and σ sensitivity measures for the i th parameter in the parameter vector θ, ee_n is the elementary effects for the n th parameter vector, N are the total parameter vectors and

μ_{θ_i} is given by:

$$\mu_{\theta_i} = \frac{1}{N}\sum_{n=1}^{N} ee_{n,\theta_i} \tag{S1.3}$$

Based on μ^*, the effective sowing depth (SOWDEPTH) was the most and only sensitive parameter during emergence, which is intuitive as the other parameters influence development after emergence (Fig.S1.1). Then the relative importance of parameters that define the cardinal temperatures (DELTMAX1, DELTOPT1 and TMINDEV1) and the physiological development days (PDD1) of the vegetative phase increased. These parameters continued to be the most influential parameters even through the generative phase of development. Even though DELTOPT2 and PDD2 are important parameters for the generative phase of development, their influence was small and over-shadowed by the influence of the vegetative phase parameters.

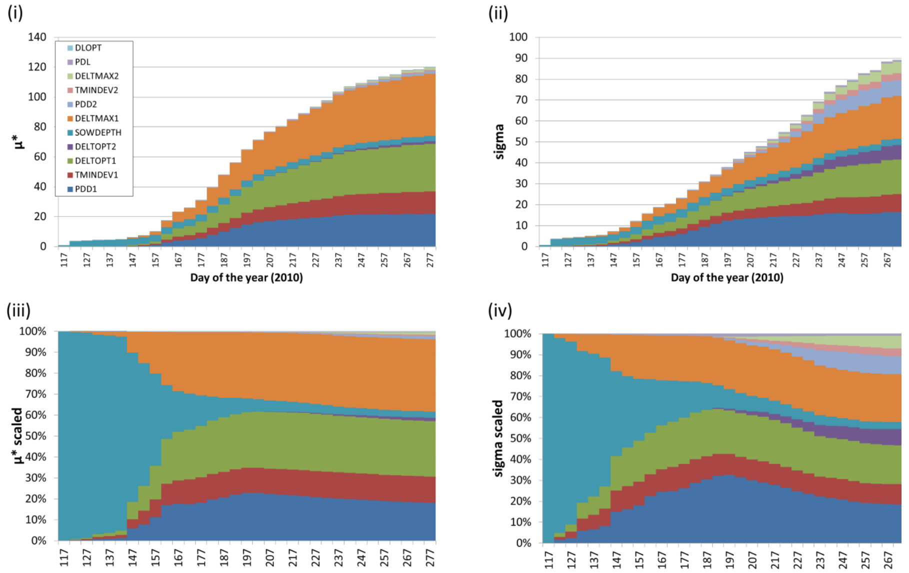


Figure S1.1: Plots of (i) μ^* and (ii) sigma of elementary effects calculated for simulated phenological development at an interval of 5 days over the growing season of silage maize (between sowing day 112 and harvest day 278 of the year) at site 6 in the year 2010. The parameters that influence phenological development in the SPASS model are listed in the legend. Plots (iii) and (iv) are the normalized μ^* and sigma values per day, respectively, expressed as a percentages.

S2 Estimation of information entropy (chapter 3)

Information entropy (H) for a continuous distribution is given by:

$$H = -\int f(\theta)\ln(f(\theta))d\theta \tag{S2.1}$$

where $f(\theta)$ is the probability density function of θ. Information entropy estimates of the posterior parameter distributions were obtained using the redistribution estimate equation (Beirlant et al., 1997):

$$H_n = -\frac{1}{n}\sum_{i=1}^{n} \ln f_n(\theta_i) \tag{S2.2}$$

where H_n is the estimate of information entropy, f_n is the Kernel Density Estimate (KDE) and $\theta_1, \ldots \theta_n$ are independent and identically distributed (i.i.d.) parameter vector samples from the posterior distribution. The KDE was obtained by using the kde function from the *ks* package in R (Duong, 2021). Least Squares Cross-Validation (LSCV) was used for bandwidth selection.

S3 Residual Analysis (chapter 3)

Residuals were analysed for the synthetic and true sequences for simulated phenology at the maximum a posteriori probability (MAP) estimate of the model parameters. The residual plots provided in the following sections have been separated into the synthetic sequences (section S3.1), Swabian Alb true sequence (section S3.2), and Kraichgau true sequence (section S3.3). Homoscedasticity was checked by plotting the residuals against days-after-sowing and simulated phenology (Figs.S3.1, S3.2, S3.7 – S3.12, S3.17 – S3.19). In general, heteroscedasticity was not observed. Normal assumption of the error model was verified by plotting histograms of the residuals and quantile-quantile plots (Figs.S3.3, S3.4, S3.13 – S3.15, S3.20). For the first few sequential updates, the number of observations were limited making a thorough analysis difficult. For the latter few sequential updates, the residuals were found to be nearly normal.

In the synthetic sequences, the residual error distribution was nearly normal (Figs.S3.3, S3.4). The slight skewness is attributed to model limitations (controlled cultivar-environment sequence) and specific site-years that had a different phenological development as compared to the remaining site-years in the calibration sequence (both synthetic sequences).

The slight skewness observed in the true sequence is attributed to model limitations where the model is unable to capture the slow development during the vegetative phase that was observed at a few site-years like 6_2013 (Figs.S3.13, S3.14, S3.15) and 5_2016 (Fig.S3.15). Autocorrelation was estimated after padding the dataset as the observations are not at regular time-intervals. Therefore, there is no ACF estimated at some lags. Figure S3.16 contains the autocorrelation (ACF) plot of the residuals after the model is calibrated to data from site-years 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016. Based on the limited data with unequal lags, no autocorrelation was detected. However, it is suspected that with state variables like phenology, which are based on cumulative sums, autocorrelation of errors could theoretically exist. However, due to data limitations, error modelling would be limited in its scope for improving the results.

S3.1 Synthetic sequences

In the ideal sequence where there is no model structural error, the skewness in the residual distribution (Fig.S3.3) is caused by site-year 2. This site-year exhibits a different development-behaviour as compared to other site-years in the calibration sequence (Fig.S3.5). In the controlled cultivar-environment sequence the slight skewness (Fig.S3.4) in the distribution of the residuals are caused due to two reasons. The site-year 9 exhibits a different phenological development-behaviour as compared to other site-years in the calibration sequence (Fig.S3.6). Additionally, the model is unable to capture the rapid growth seen in site-years 3, 4, 5, 8 and 9 between 82 and 110 days after sowing.

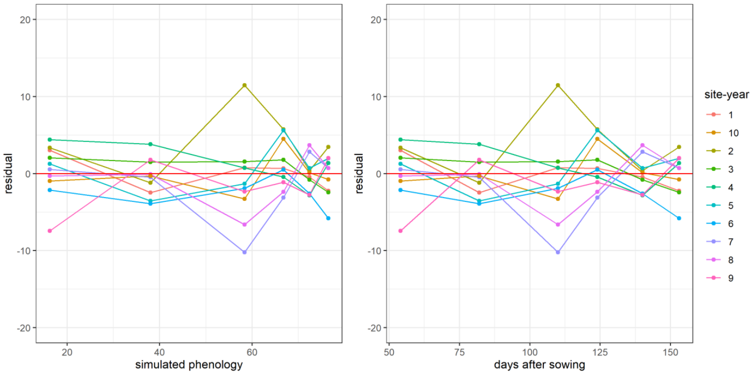


Figure S3.1: Residuals vs simulated phenology and days after sowing after calibration to 10 site-years in the ideal synthetic sequence

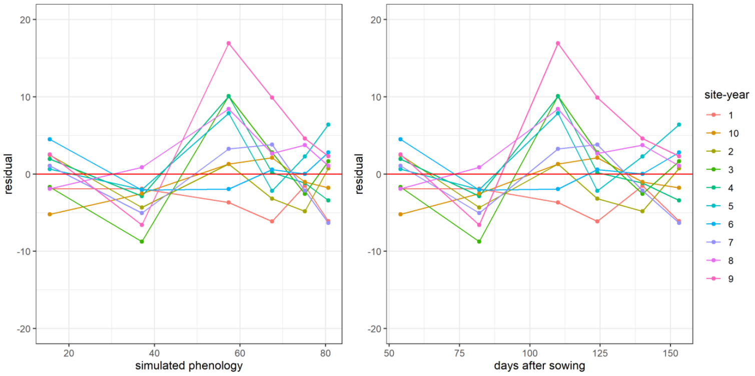


Figure S3.2: Residuals vs simulated phenology and days after sowing after calibration to 10 site-years in the controlled cultivar-environment synthetic sequence

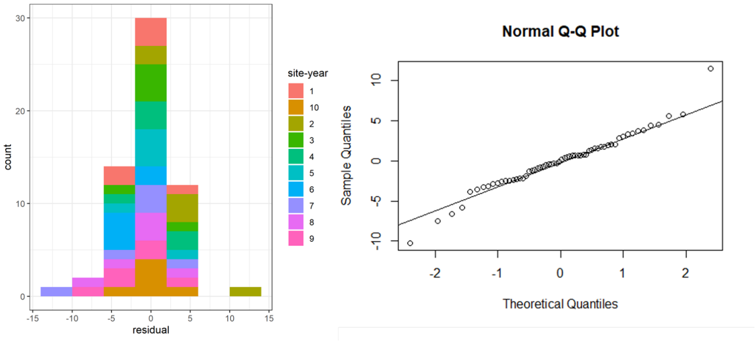


Figure S3.3: Histogram and quantile-quantile plots of the residuals after calibration to 10 site-years of the ideal synthetic sequence

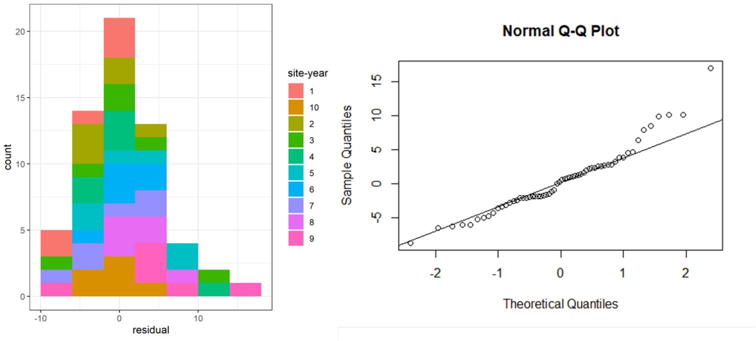


Figure S3.4: Histogram and quantile-quantile plots of the residuals after calibration to 10 site-years of the controlled cultivar-environment synthetic sequence

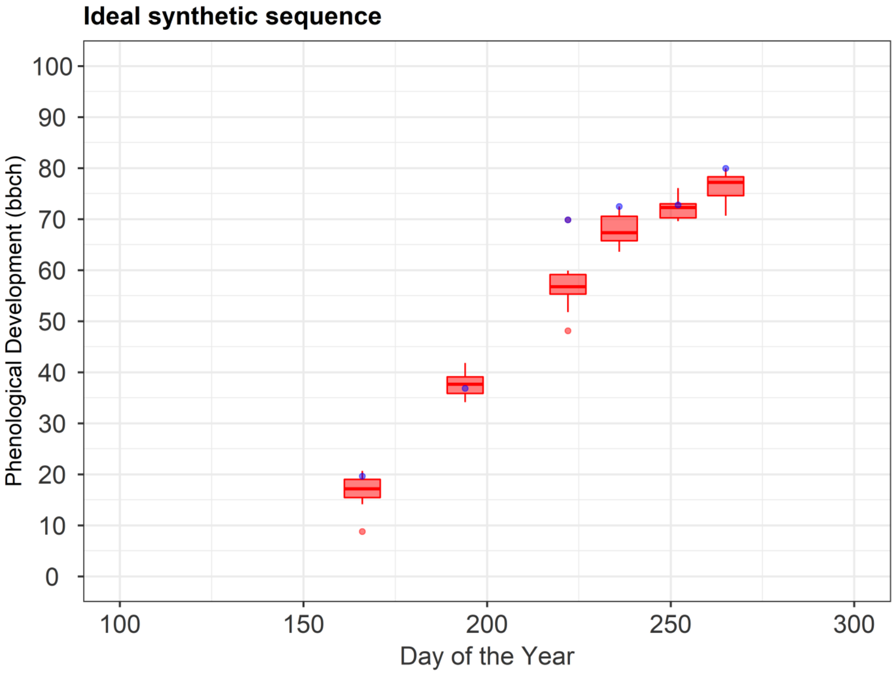


Figure S3.5: The boxplots show the phenological development (BBCH) of all the site-years used in calibration in the ideal synthetic sequence. The blue point corresponds to the phenological development (BBCH) for site-year 2.

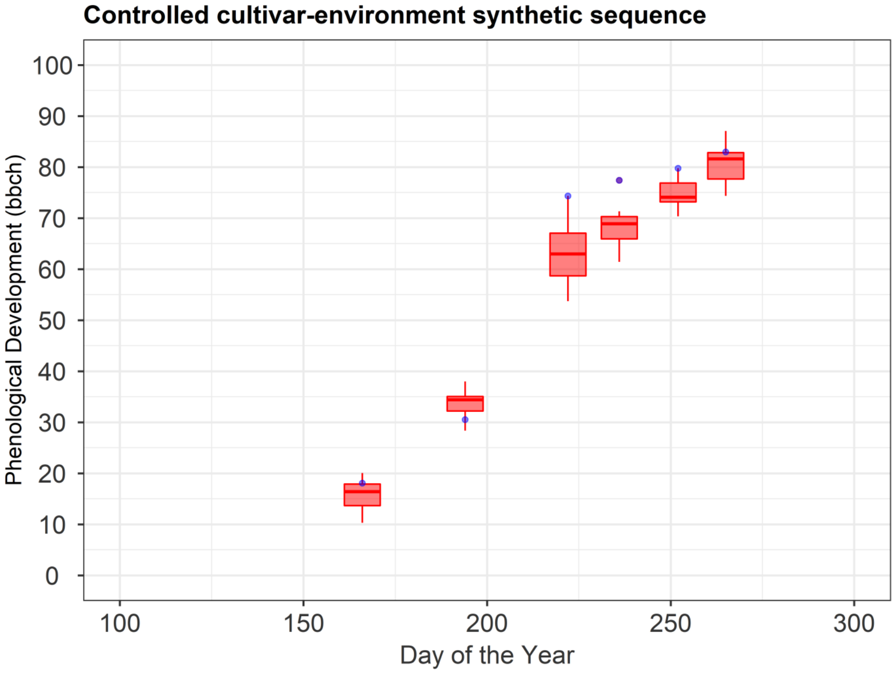


Figure S3.6: The boxplots show the phenological development (BBCH) of all the site-years used in calibration in the controlled cultivar-environment synthetic sequence. The blue point corresponds to the phenological development (BBCH) for site-year 9.

S3.2 True sequence in Swabian Alb

The residual plots for the sequential updates with greater than 3 calibration site-years show high residuals in the vegetative phase (simulated phenology¡61BBCH) (Figs.S3.10, S3.11, S3.12). Residuals from site-years 6_2013 and 5_2016 cause this skewness in the distribution of the residuals (Figs. S3.13, S3.14, S3.15). This behaviour is attributed to the model's inability to capture the slow development seen in these site-years as evident from the single-site-year calibration results in Fig.S4.1.

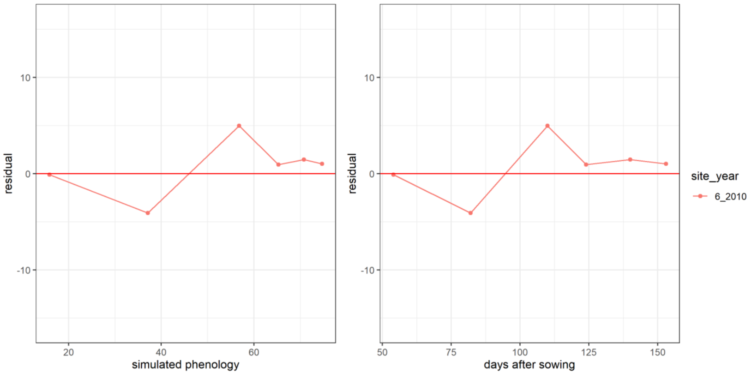


Figure S3.7: Residuals vs simulated phenology and days after sowing after calibration to site-year 6_2010

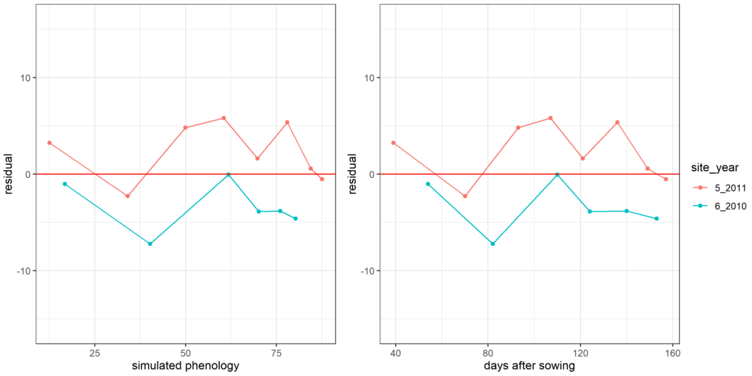


Figure S3.8: Residuals vs simulated phenology and days after sowing after calibration to site-years 6_2010 and 5_2011

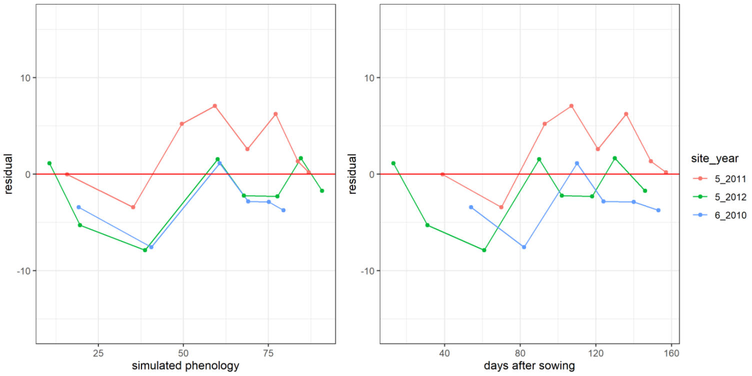


Figure S3.9: Residuals vs simulated phenology and days after sowing after calibration to site-years 6_2010, 5_2011, and 5_2012

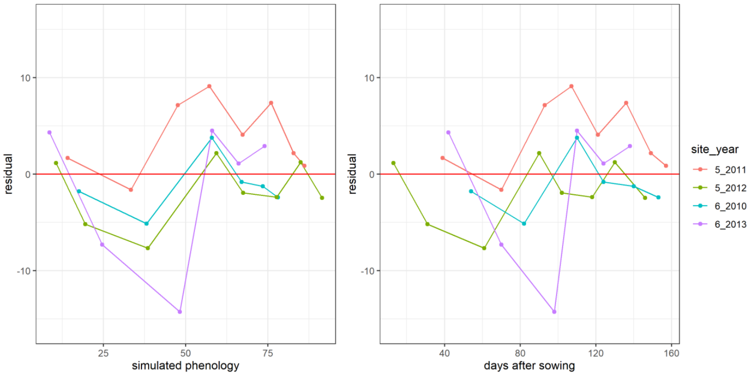


Figure S3.10: Residuals vs simulated phenology and days after sowing after calibration to site-years 6_2010, 5_2011, 5_2012, and 6_2013

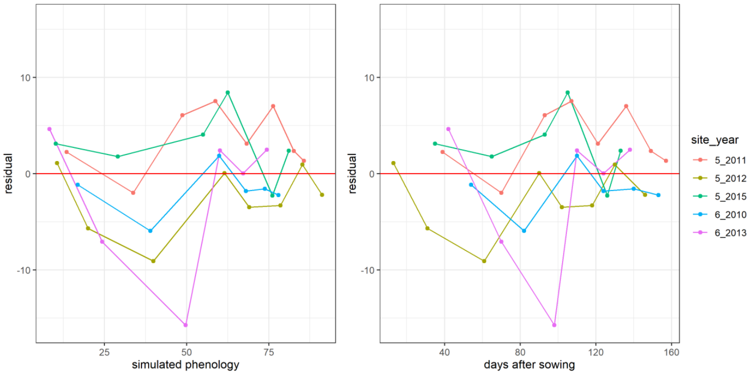


Figure S3.11: Residuals vs simulated phenology and days after sowing after calibration to site-years 6_2010, 5_2011, 5_2012, 6_2013, and 5_2015

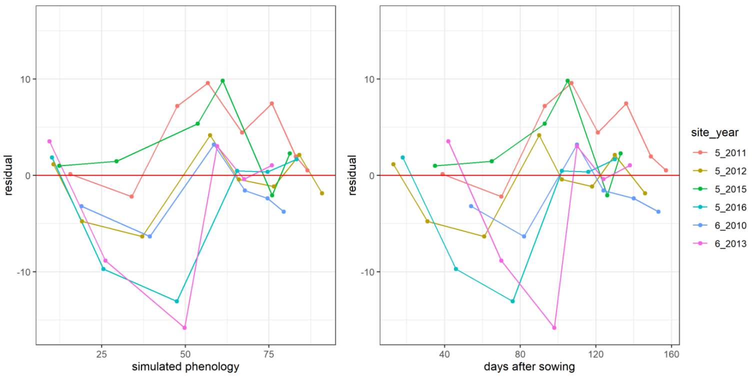


Figure S3.12: Residuals vs simulated phenology and days after sowing after calibration to site-years 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016

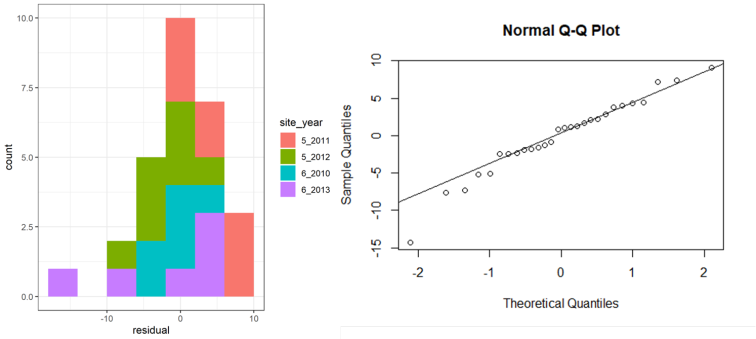


Figure S3.13: Histogram and quantile-quantile plots of the residuals after calibration to site-years 6_2010, 5_2011, 5_2012, and 6_2013

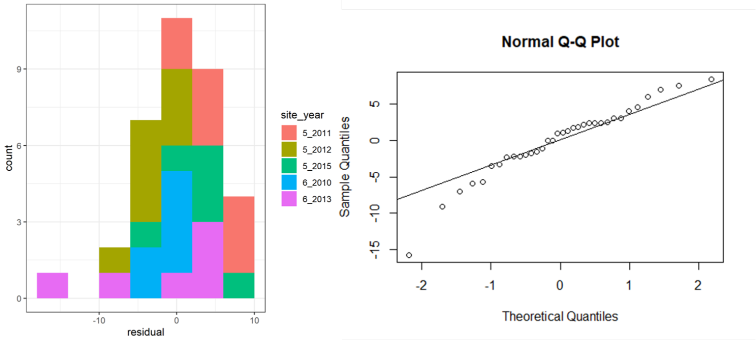


Figure S3.14: Histogram and quantile-quantile plots of the residuals after calibration to site-years 6_2010, 5_2011, 5_2012, 6_2013, and 5_2015

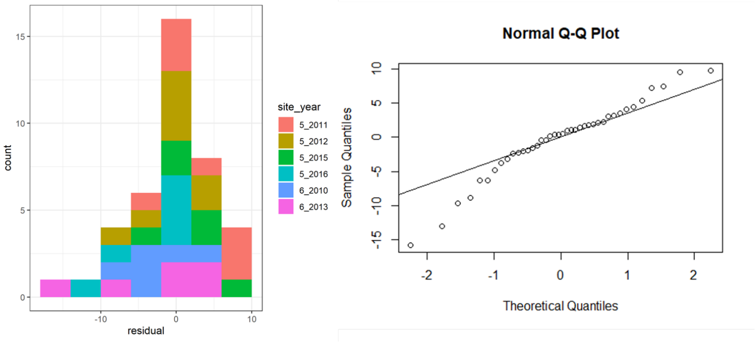


Figure S3.15: Histogram and quantile-quantile plots of the residuals after calibration to site-years 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016

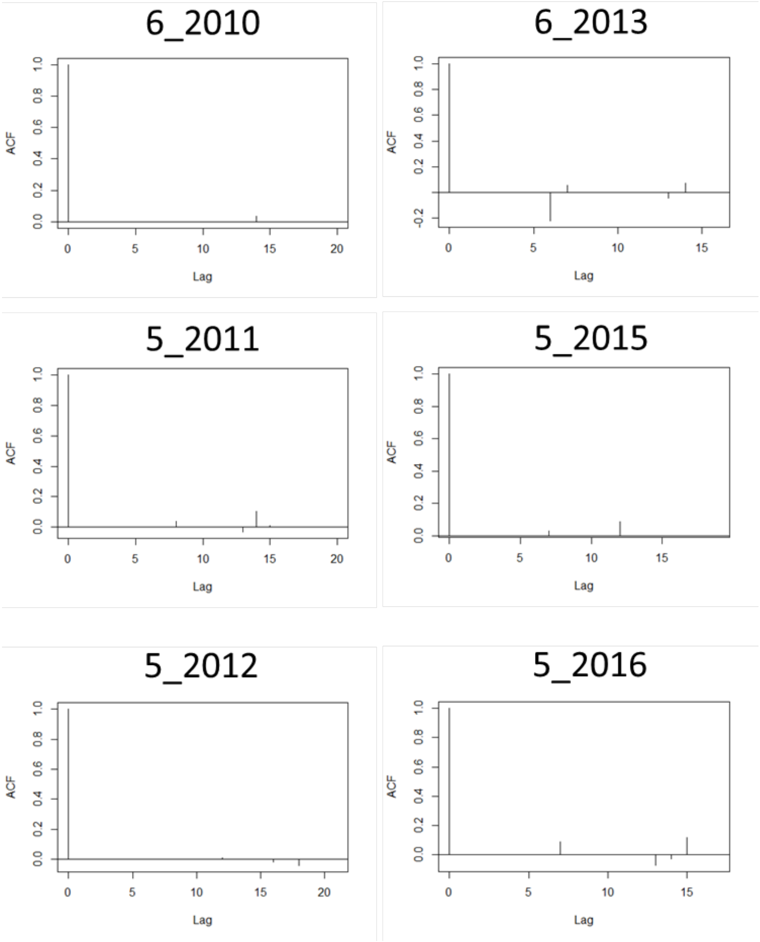


Figure S3.16: ACF (auto-correlation function) plots of the residuals after calibration to site-years 6_2010, 5_2011, 5_2012, 6_2013, 5_2015, and 5_2016

S3.3 True sequence in Kraichgau

The residual plots for Kraichgau with limited observations show no evidence of heteroscedasticity (Figs.S3.17, S3.18, S3.19) and a nearly normal distribution (Fig.S3.20).

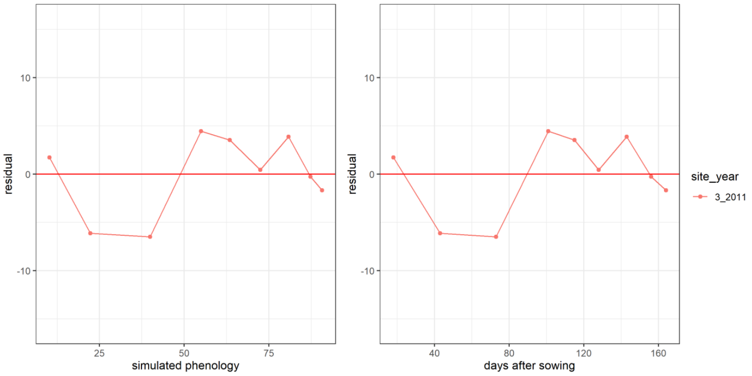


Figure S3.17: Residuals vs simulated phenology and days after sowing after calibration to site-years 3_2011

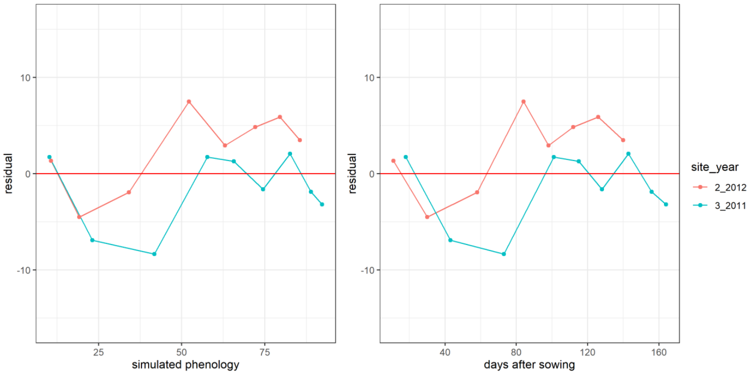


Figure S3.18: Residuals vs simulated phenology and days after sowing after calibration to site-years 3_2011 and 2_2012

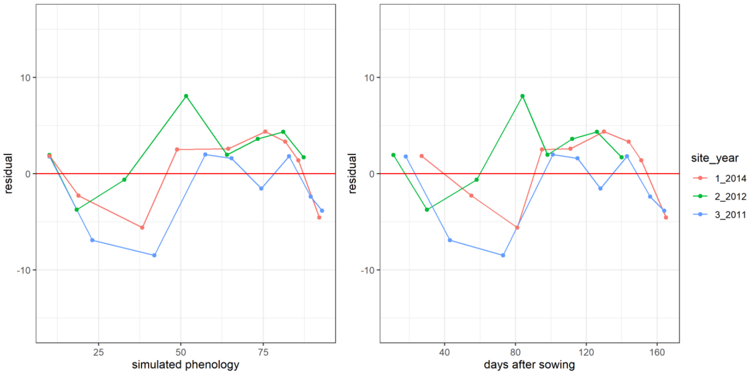


Figure S3.19: Residuals vs simulated phenology and days after sowing after calibration to site-years 3_2011, 2_2012, and 1_2014

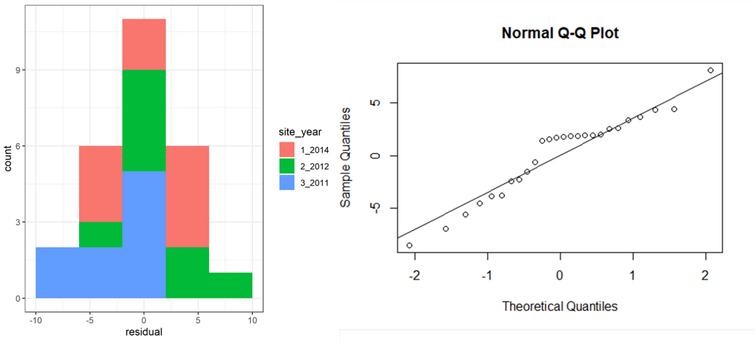


Figure S3.20: Histogram and quantile-quantile plots of the residuals after calibration to site-years 3_2011, 2_2012, and 1_2014

S4 Single-site-year calibration results (chapter 3)

Observed and simulated phenology, after the SPASS model was calibrated individually to the site-years in the study, are plotted in Fig.S4.1.

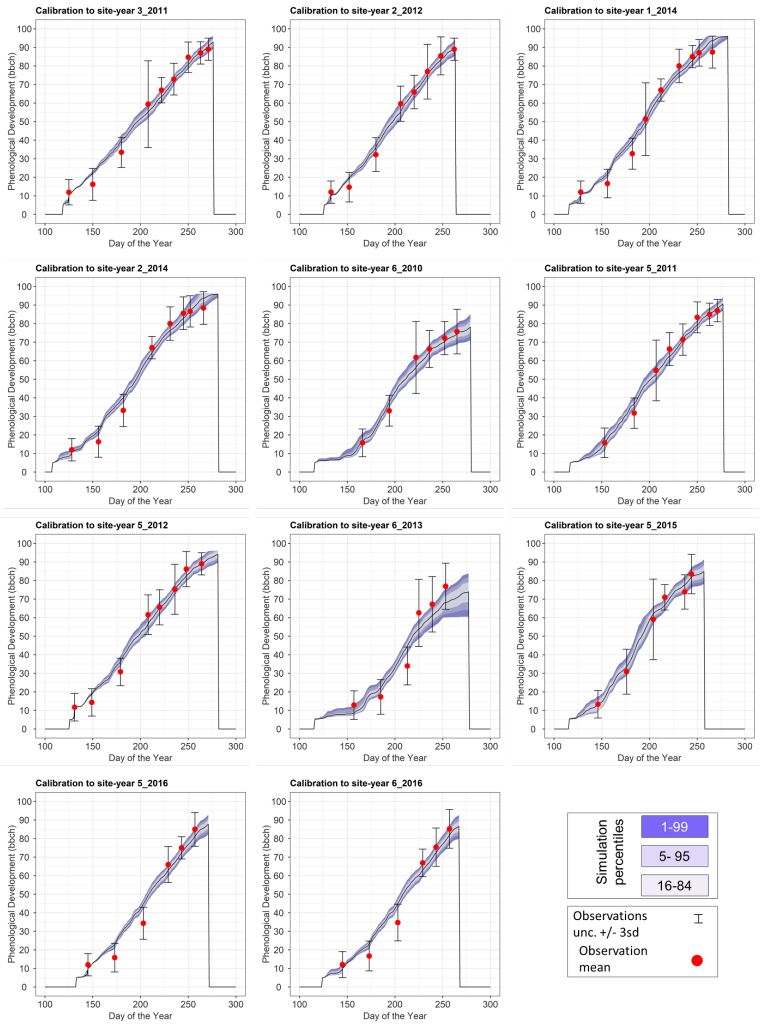


Figure S4.1: Observed and simulated phenological development after calibration, plotted against the day of the year. The red points are the mean observations, while the black error bars indicate +/- 3 standard deviations. The mean simulation is indicated by the continuous black line. The green bands represent the different percentiles of simulated phenology. It is noted that for some site-years, the calibrated model is unable to capture the slow development rate during the vegetative phase.

S5 Parameter distributions and entropy: synthetic sequences (chapter 3)

Marginal prior and posterior distributions for the 6 estimated parameters of the SPASS phenology model and the entropy estimates are plotted for the ideal (Fig.S5.1) and controlled cultivar-environment (Fig.S5.2) synthetic sequences.

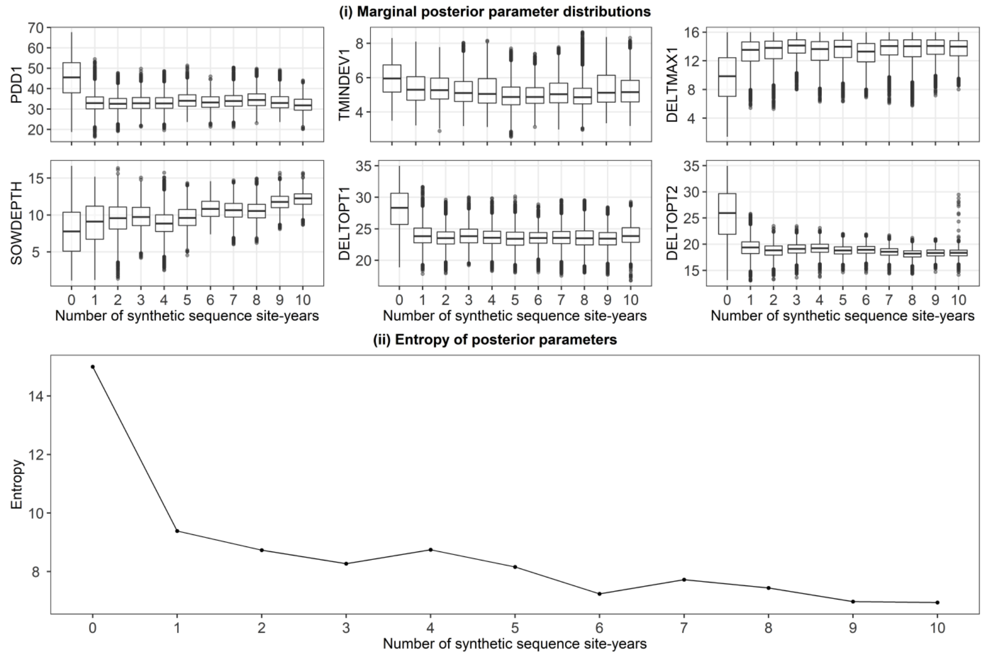
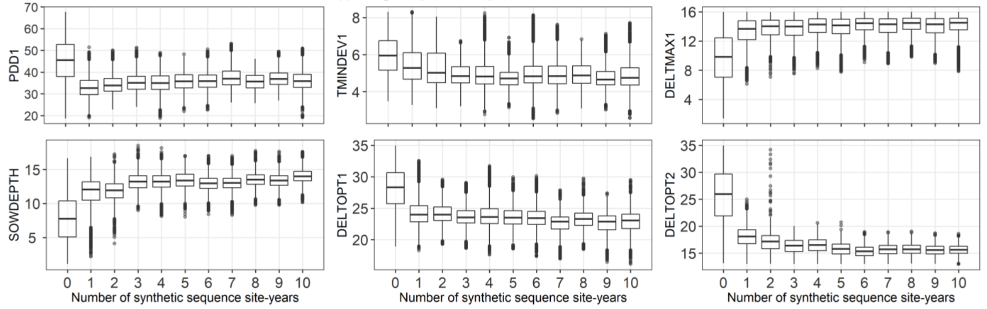


Figure S5.1: (i) Marginal prior and posterior parameter distributions of the 6 estimated parameters after BSU in the ideal synthetic sequence. Marginal posterior parameter values (y-axis) is plotted against the number of site-years used for calibration (x-axis), starting with the initial prior (0 on x-axis). (ii) Information entropy of the posterior parameter distributions after BSU was applied to the synthetic sequence. Length of the box represents the inter-quartile range (IQR), whiskers extend from the boxes up to 1.5 × IQR and values beyond this range are plotted as points. The ranges for parameters SOWDEPTH and DELTOPT2 narrowed through the sequential updates while the remaining parameters do not show a noticeable narrowing in range.

(i) Marginal posterior parameter distributions



(ii) Entropy of posterior parameters

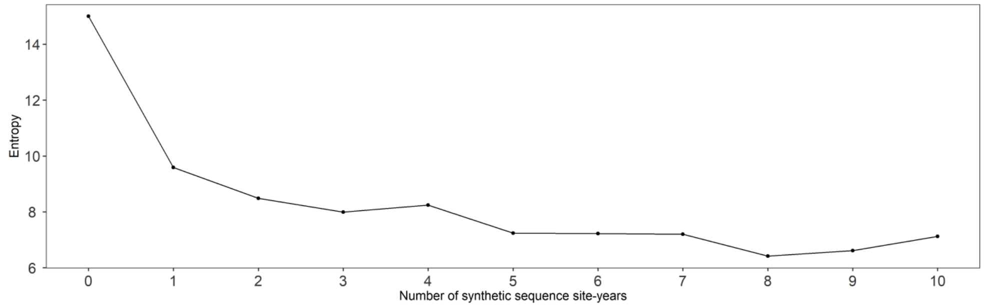


Figure S5.2: (i) Marginal prior and posterior parameter distributions of the 6 estimated parameters after BSU in the controlled cultivar-environment synthetic sequence. Marginal posterior parameter values (y-axis) is plotted against the number of site-years used for calibration (x-axis), starting with the initial prior (0 on x-axis). (ii) Information entropy of the posterior parameter distributions after BSU was applied to the synthetic sequence. Length of the box represents the inter-quartile range (IQR), whiskers extend from the boxes up to 1.5 × IQR and values beyond this range are plotted as points. The ranges for parameters SOWDEPTH and DELTOPT2 narrowed through the sequential updates while the remaining parameters do not show a noticeable narrowing in range.

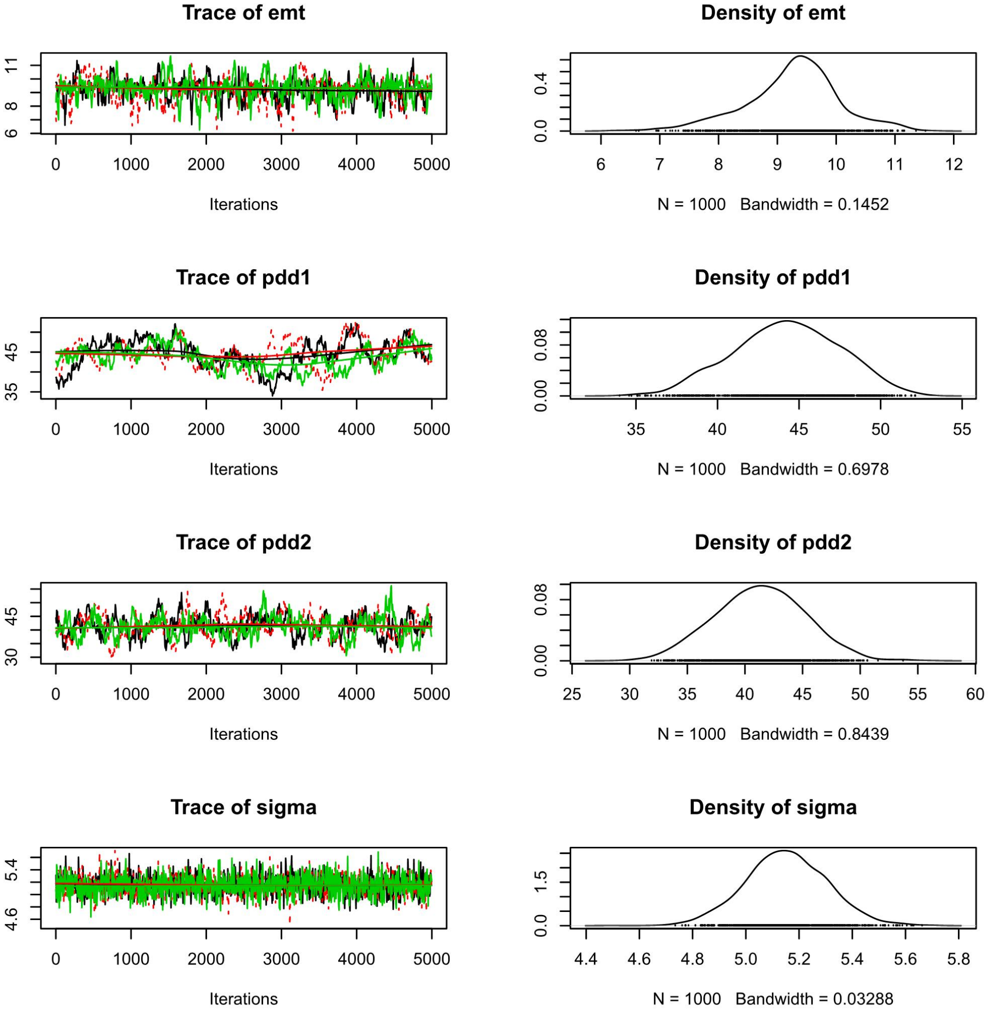
S6 MCMC posterior samples (chapter 4)

Markov Chain Monte Carlo (MCMC) sampling of the posterior parameter distributions was performed using the R2jags (Su and Yajima, 2020) (for cases BM-0, BMM-1, BMM-2a and BMM-2b) and jagsUI (Kellner, 2021) packages (for BMM-3) in R. In the following section, diagnostic plots are provided for some MCMC parameter samples from the five model cases. The SPASS model parameters at the species level (θ_{sp}) of the hierarchy are emt_sp, pdd1_sp, tminv_sp, toptv_sp, pdd2_sp, tminr_sp, and toptr_sp for model cases BM-2b and BMM-3, and emt, pdd1, tminv, toptv, pdd2, tminr, and toptr for model cases BM-0, BMM-1, and BMM-2a. The standard deviation of the likelihood function is sigma (σ). Parameters weath, eco, and year correspond to the weather effects (δ_w), eco-region effects (γ_e), and year effects (τ_y), respectively.

S6.1 Trace-plots

The trace-plots and density plots (coda package (Plummer et al., 2006) in R (R Core Team, 2020)) for some parameters from the BMM-2a case are provided as an example. In Fig. S6.1 parameter sigma

(σ) shows good mixing across the three chains, while parameters pdd1 and pdd2 show relatively poor mixing. The poor mixing is attributed to the parameter correlations (section S6.4). Parameters weath (γ_w) (Fig. S6.2), eco (γ_e) (Fig. S6.3), and year (τ_y) (Fig. S6.4) show good mixing.


Trace of emt

Density of emt

Trace of pdd1

Density of pdd1

Trace of pdd2

Density of pdd2

Trace of sigma

Density of sigma

Figure S6.1: Trace-plots (left) and density plots (right) for some of the species level SPASS model parameters and the standard deviation of the likelihood function from the BMM-2a case.

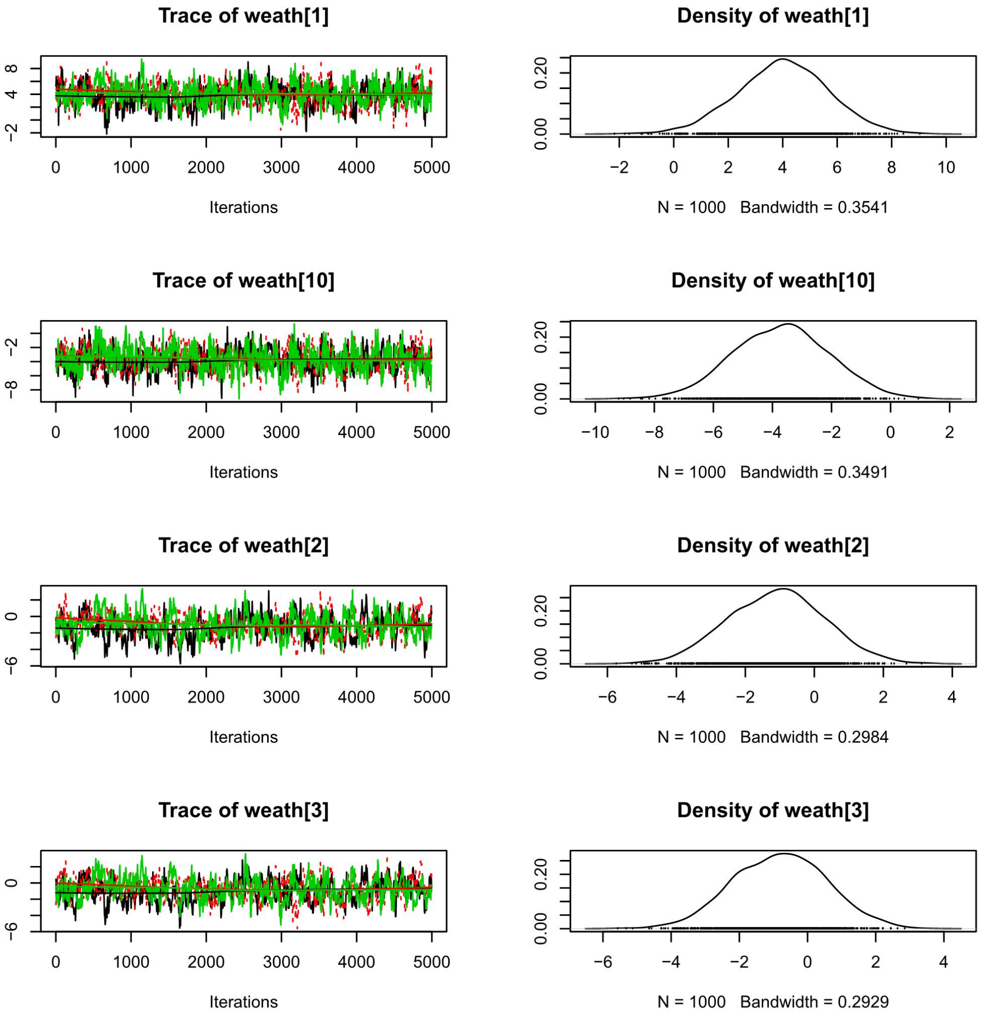


Figure S6.2: Trace-plots (left) and density plots (right) for weather effect parameters for some of the weather classes from the BMM-2a case.

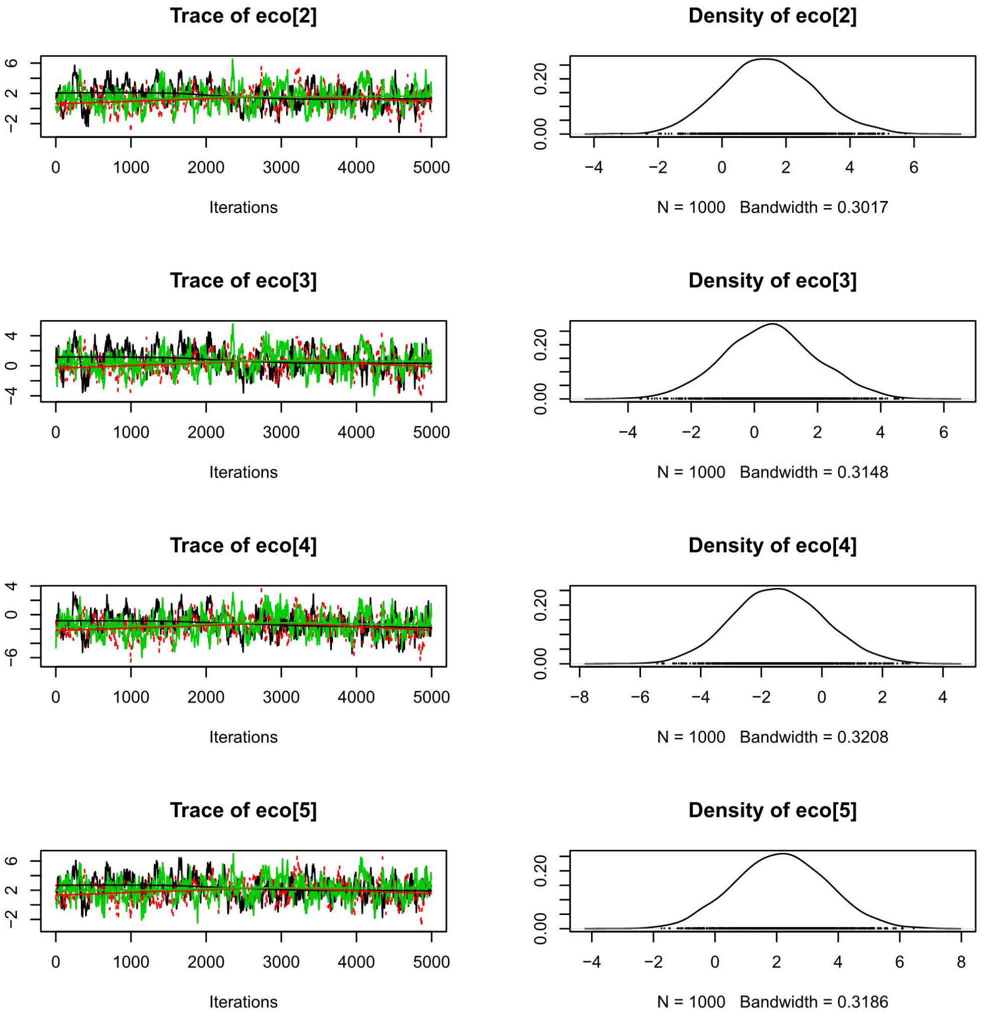


Figure S6.3: Trace-plots (left) and density plots (right) for the eco-region effect parameters for some of the eco-regions from the BMM-2a case.

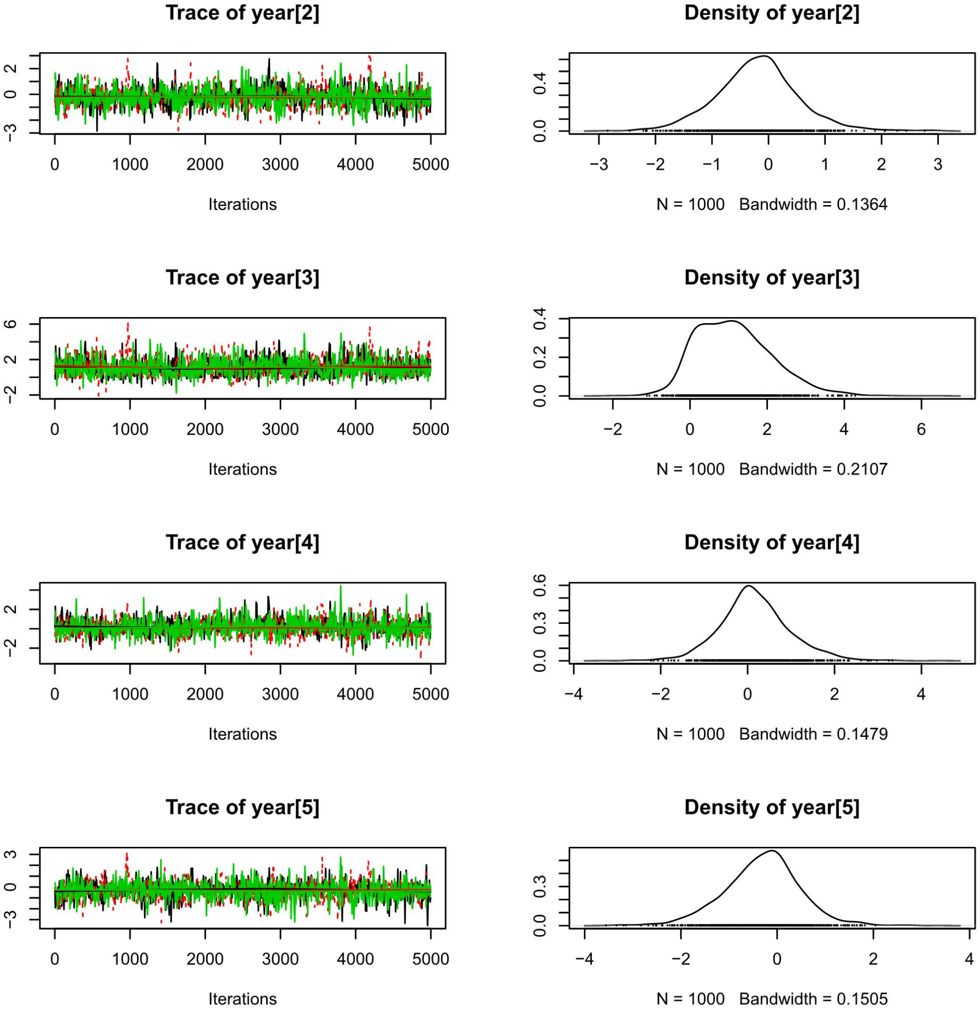


Figure S6.4: Trace-plots (left) and density plots (right) for the year effect parameters for some of the years from the BMM-2a case.

S6.2 Convergence diagnostic

The three chains for the MCMC algorithm were run until the Gelman-Rubin convergence diagnostic was ≤ 1.1. In Fig. S6.5 we provide a plot of the shrink factor or the convergence diagnostic (gelman.plot in coda package (Plummer et al., 2006)) for some of the parameters from the BMM-3 case. It can be seen that the parameters have converged after 3,500 iterations.

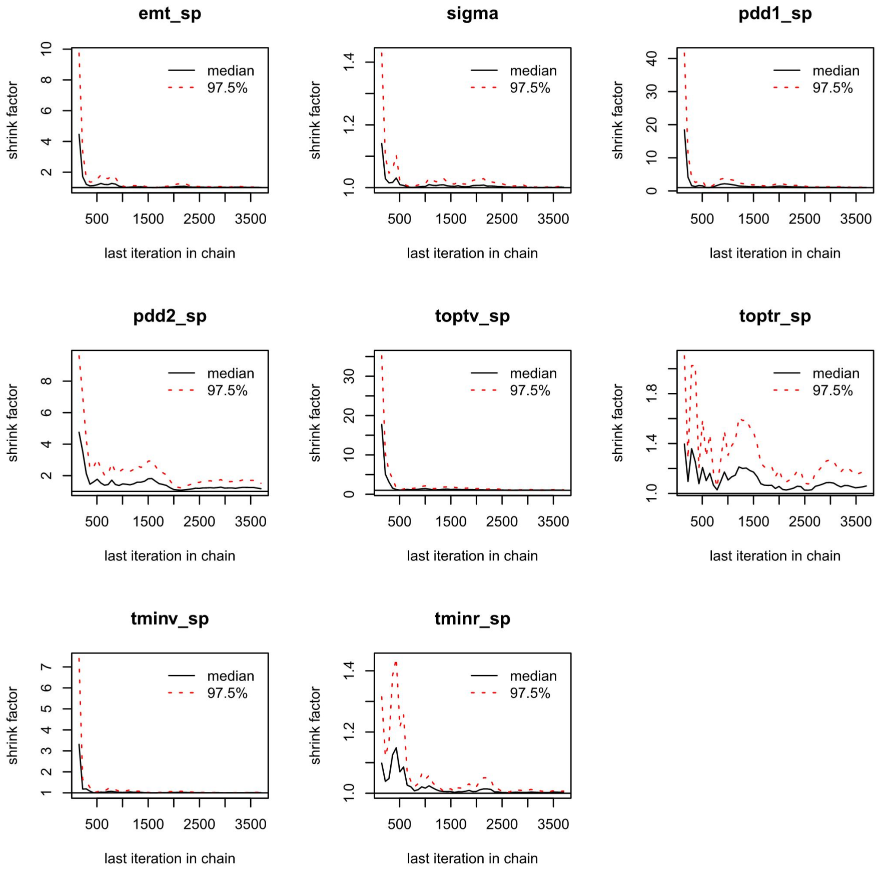


Figure S6.5: Evolution of Gelman-Rubin shrink factor (y-axis) for some species-level parameters (θ_{sp}) in BMM-3 case with an increase in the number of iterations (x-axis)

S6.3 Auto-correlation plots

The auto-correlation plots (acfplot in coda package (Plummer et al., 2006)) are provided for the species-level parameters (θ_{sp}) and sigma (σ) in the BM-0 (Fig. S6.6) and BMM-3 (Fig. S6.7) cases. These show the auto-correlation of the parameters within the chain. The auto-correlation decreases to zero with greater lag. Note that in BM-0 the 5,000 posterior samples were thinned to obtain a final set of 1000 samples. The samples were not thinned in BMM-3. Parameters toptv_sp ,toptr_sp, pdd1_sp, and pdd2_sp exhibit a higher auto-correlation which can be attributed to between-parameter correlations (section S6.4).

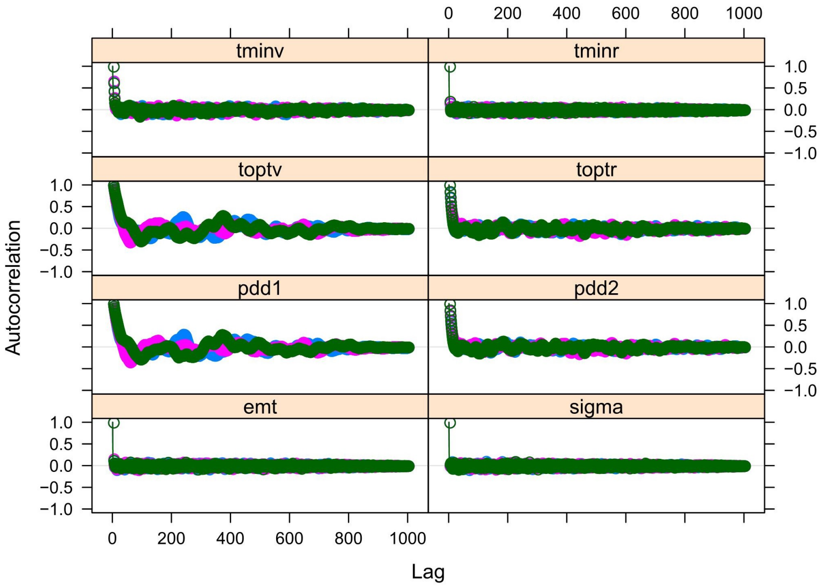


Figure S6.6: Auto-correlation plots for some parameters in the BM-0 case.

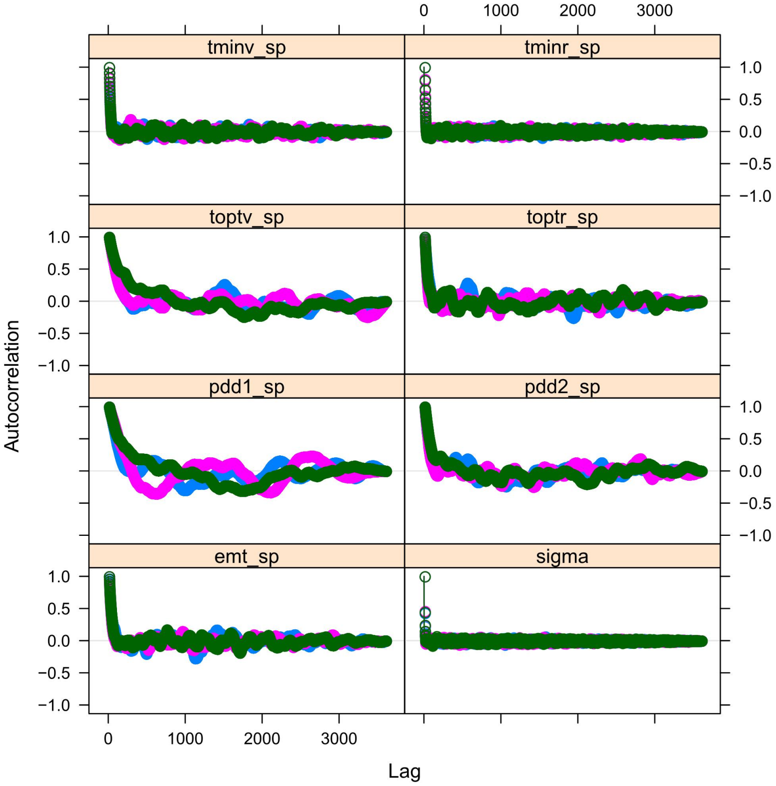


Figure S6.7: Auto-correlation plots for some parameters in the BMM-3 case.

S6.4 Correlation plots

Correlation plots (ipairs function in IDPmisc package Locher (2020)) of the species-level parameters (θ_{sp}) in the hierarchy and sigma (σ) of the likelihood function for cases BM-0 and BMM-2b are provided in Fig. S6.8 and S6.9. Additionally, correlation coefficient plots (corrplot function in corrplot package (Wei and Simko, 2017)) of some of the parameters are provided for all the cases (Fig. S6.10, S6.11, S6.12, S6.13, S6.14). The colours of the ellipse indicate positive (blue) or negative correlation (red), while the colour intensity and shape of the ellipse indicates the value. A correlation coefficient of one is a diagonal line, while no correlation is represented by a white circle. There is strong negative correlation between parameters pdd1 and toptv as well as between pdd2 and toptr in BM-0 (Fig. S6.8), but this is not seen in BMM-2b(Fig. S6.9). The correlation coefficient plots show that in BM-0 (Fig. S6.10), BMM-1 (Fig. S6.11), and BMM-2a (Fig. S6.12), the correlation exists but is not seen in BMM-2b (Fig. S6.13) and BMM-3 (Fig. S6.14) where the ripening and cultivar hierarchy is introduced. In these two cases, these correlations are seen in the ripening- and cultivar-specific parameters (not shown). The eco-regions effect parameters (eco) are positively correlated with each other and also to the base temperature for emergence (emt) (Fig. S6.11). The weather effect parameters (weath) in BMM-2a (Fig. S6.12) are also positively correlated with each other and negatively correlated with the eco-region effect parameters (eco). This correlation between eco-region and weather effect parameters could be because there is a very likely overlap between the weather classes and eco-region class as the eco-regions are also based on climatological characteristics.

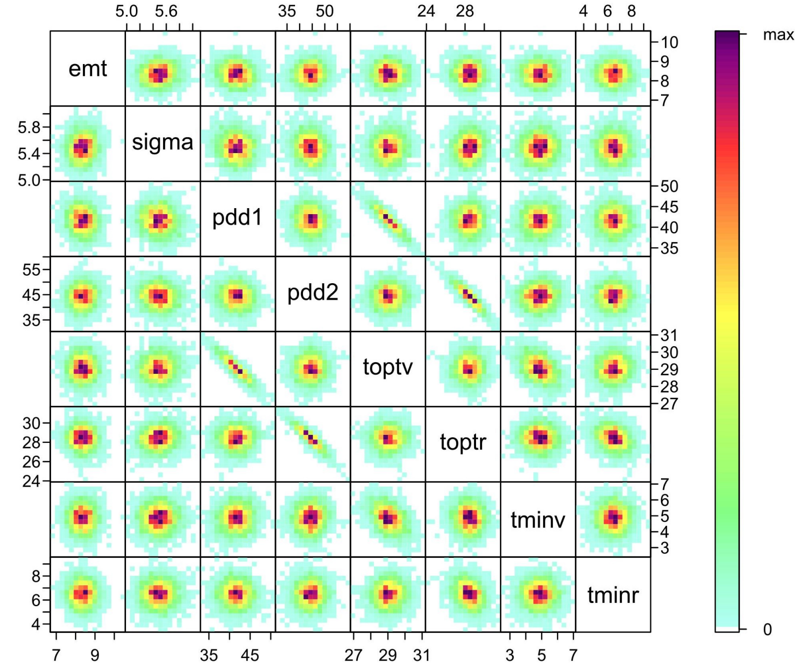


Figure S6.8: Cross-plot of the posterior samples of the some estimated parameters in the BM-0 case. Red represents high density and blue low density.

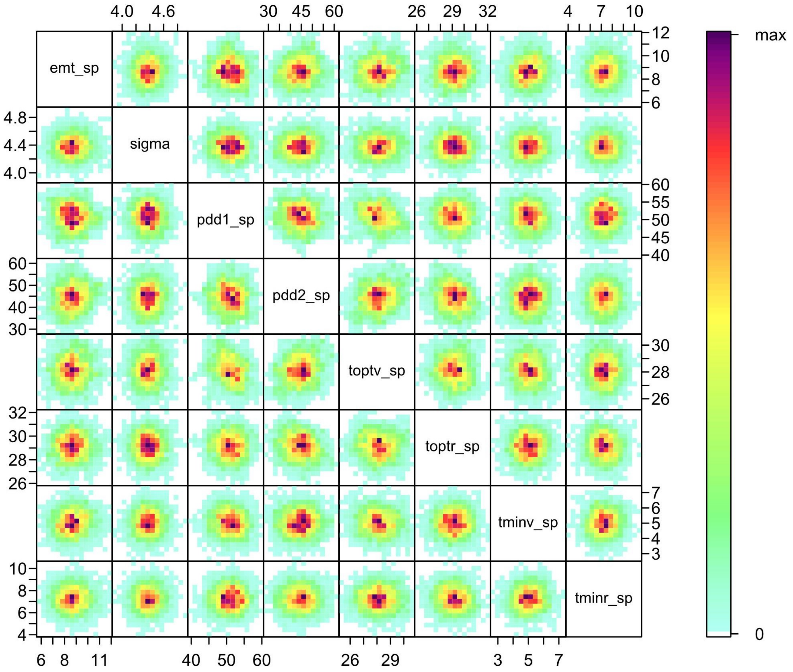


Figure S6.9: Cross-plot of the posterior samples of the some estimated parameters in the BMM-2b case. Red represents high density and blue low density.

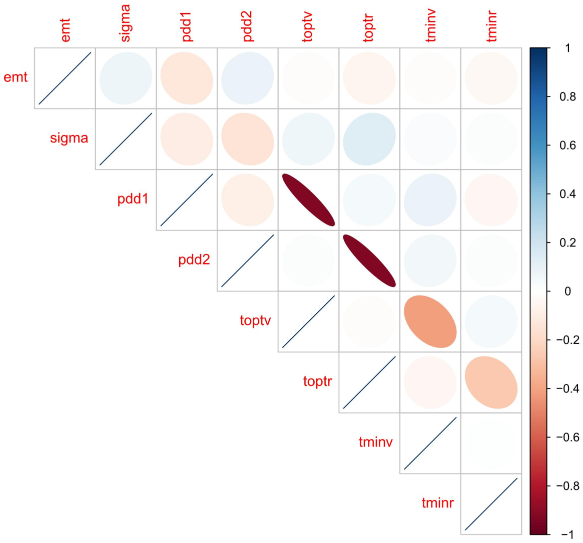


Figure S6.10: A plot of the correlation coefficients between some of the parameters in the BM-0 case. The colours of the ellipse indicate positive (blue) or negative correlation (red), while the colour intensity and shape of the ellipse indicates the value. A correlation coefficient of one is a diagonal line, while no correlation is represented by a white circle.

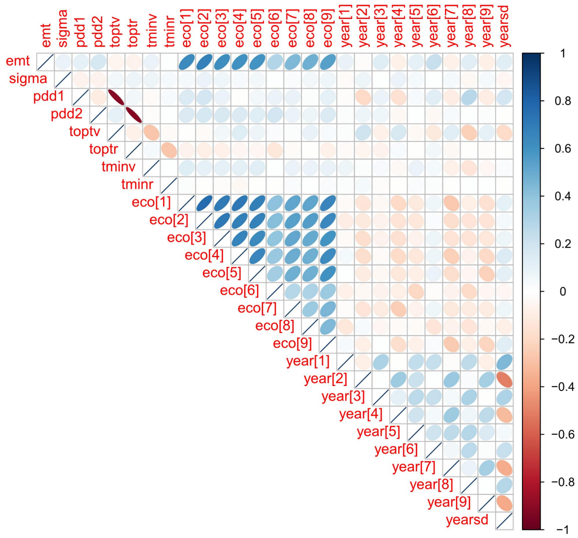


Figure S6.11: A plot of the correlation coefficients between some of the parameters in the BMM-1 case. The colours of the ellipse indicate positive (blue) or negative correlation (red), while the colour intensity and shape of the ellipse indicates the value. A correlation coefficient of one is a diagonal line, while no correlation is represented by a white circle.

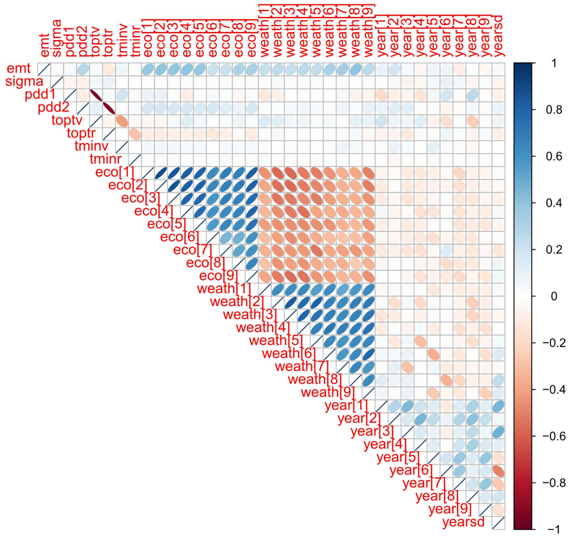


Figure S6.12: A plot of the correlation coefficients between some of the parameters in the BMM-2a case. The colours of the ellipse indicate positive (blue) or negative correlation (red), while the colour intensity and shape of the ellipse indicates the value. A correlation coefficient of one is a diagonal line, while no correlation is represented by a white circle.

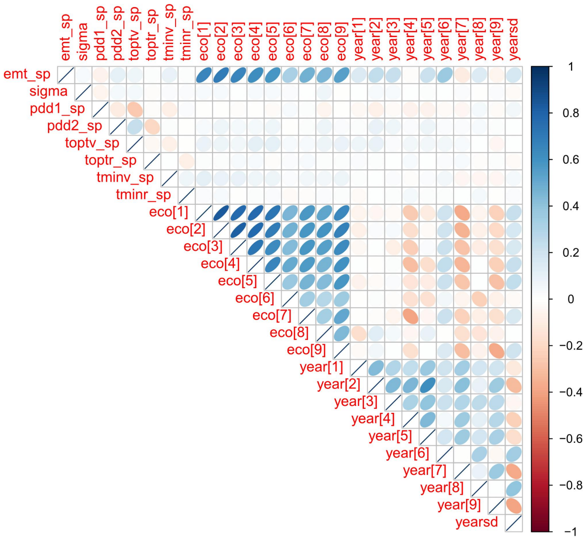


Figure S6.13: A plot of the correlation coefficients between some of the parameters in the BMM-2b case. The colours of the ellipse indicate positive (blue) or negative correlation (red), while the colour intensity and shape of the ellipse indicates the value. A correlation coefficient of one is a diagonal line, while no correlation is represented by a white circle.

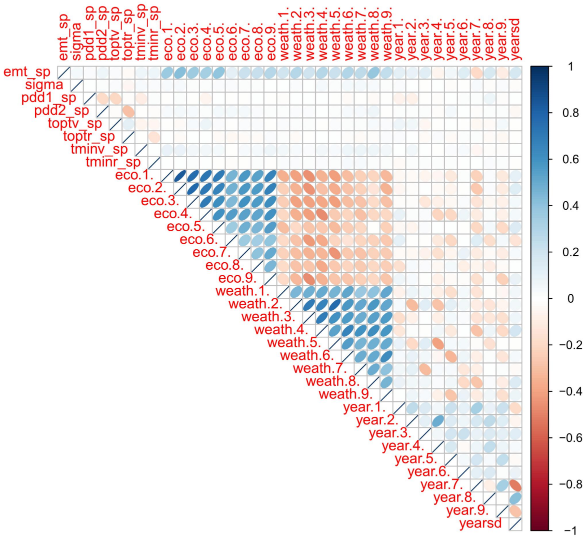


Figure S6.14: A plot of the correlation coefficients between some of the parameters in the BMM-3 case. The colours of the ellipse indicate positive (blue) or negative correlation (red), while the colour intensity and shape of the ellipse indicates the value. A correlation coefficient of one is a diagonal line, while no correlation is represented by a white circle.

S7 Data set information (chapter 4)

S7.1 Ripening groups and cultivars

Figure S7.1 shows (plotly package (Sievert, 2020)) the names of the different cultivars within the four ripening groups that were used for calibration. The circle represents 100 site-years used for calibration. The late ripening group (L) contains only one site-year belonging to cultivar MAS 40F.

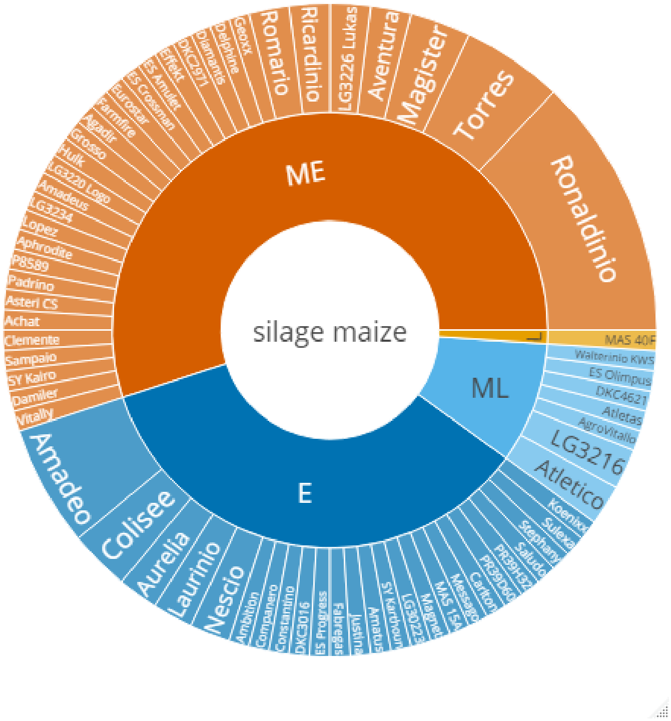
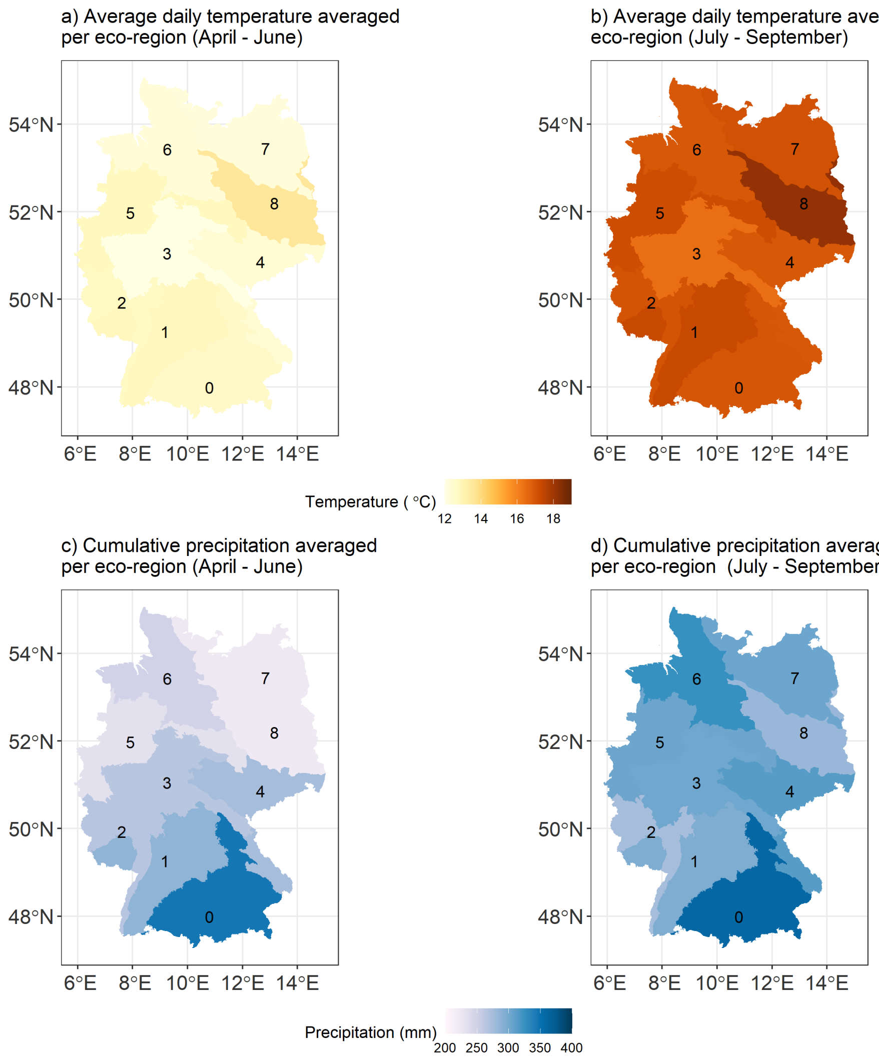


Figure S7.1: Cultivars and ripening groups of the 100 site-years that were used for calibration. The colours represent the ripening groups. E: early, ME: mid-early, ML: mid-late, and L: late. The circle represents 100 site-years. The cultivar MAS 40F in the late ripening group has only one site-year.

S7.2 Eco-regions

Figure S7.2 shows the average daily temperature and average cumulative precipitation per eco-region for April-June and July-September, based on all 3,004 site-years. The months of April-June and July-September correspond to the time when vegetative and reproductive phenological development usually occurs for silage maize grown in Germany. These averages are based on 689 locations across Germany and nine years (2009-2017). Eco-region 8 (eastern part of the northern plains) has the highest temperature during April-June and July-September. In general, the southern regions received more precipitation in April-June than the northern regions. Eco-region 0 (region to the north of the Alps) received the most precipitation while eco-region 8 received the least in both periods.

Figure S7.3 shows the soil regions (Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), 1993) across the nine eco-regions of Germany. Eco-region 0 consists of carbonates and moraine deposits. Eco-regions 1, 3, and 4 consist of sedimentary deposits such as sandstone, siltstone, carbonates, claystones and marls with loess and loamy soils developed in some areas. The Rhine river plain is in eco-region 2 consists of fluvial landscapes. The northern eco-regions 5-8 consist of moraine deposits.


a) Average daily temperature averaged per eco-region (April - June)

b) Average daily temperature averaged eco-region (July - September)

c) Cumulative precipitation averaged per eco-region (April - June)

d) Cumulative precipitation averaged per eco-region (July - September)

Figure S7.2: Average daily temperature (a, b) and average cumulative precipitation (c, d) per eco-region for April-June and July-September, based on all 3004 site-years. The numbers indicate the eco-regions based on the classification provided by the BfN (Bundesamt für Naturschutz) (2017).

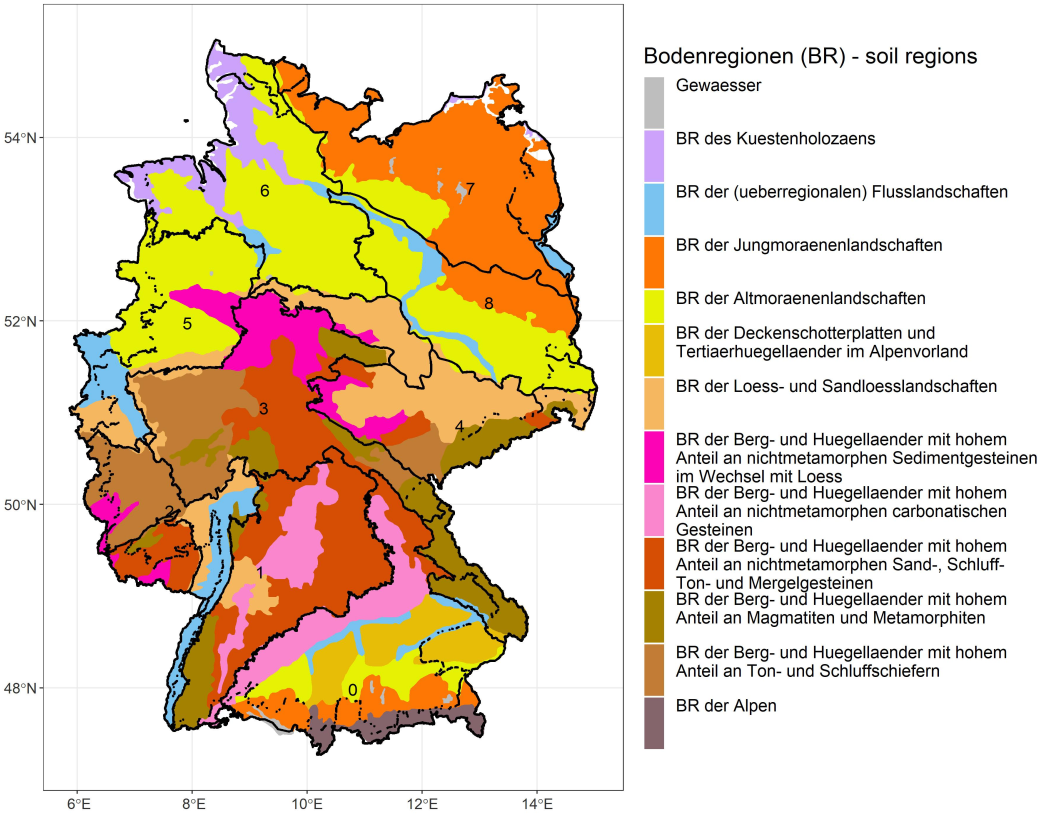

Bodenregionen (BR) - soil regions
Gewaesser
BR des Kuestenholozaens
BR der (ueberregionalen) Flusslandschaften
BR der Jungmoraenenlandschaften
BR der Altmoraenenlandschaften
BR der Deckenschotterplatten und Tertiaerhuegellaender im Alpenvorland
BR der Loess- und Sandloesslandschaften
BR der Berg- und Huegellaender mit hohem Anteil an nichtmetamorphen Sedimentgesteinen im Wechsel mit Loess
BR der Berg- und Huegellaender mit hohem Anteil an nichtmetamorphen carbonatischen Gesteinen
BR der Berg- und Huegellaender mit hohem Anteil an nichtmetamorphen Sand-, Schluff-Ton- und Mergelgesteinen
BR der Berg- und Huegellaender mit hohem Anteil an Magmatiten und Metamorphiten
BR der Berg- und Huegellaender mit hohem Anteil an Ton- und Schluffschiefern
BR der Alpen
54°N
52°N
50°N
48°N
6°E
8°E
10°E
12°E
14°E
6
7
8
5
3
4
2
1
0

Figure S7.3: The soil regions across Germany (data source: data set classification ©Bundesanstalt für Geowissenschaften und Rohstoffe (BGR) (1993)-BGL5000V2.0) within the nine eco-regions (data source: BfN (Bundesamt für Naturschutz) (2017)). The colours indicate the soil regions while the black outline demarcates the nine eco-regions. The numbers indicate the eco-regions.

S7.3 Weather classes

Weather classes were defined using k-means clustering, based on the average daily temperature and cumulative precipitation at the 3,004 site-years in the full data set. Figure S7.4 shows the minimum, mean and maximum values by weather class in the full data set and for the 100 site-years in the calibration data set. In most cases, the mean temperature and cumulative precipitation per weather class in the calibration data set are close to those in the full data set. Figure S7.5 shows the distribution of weather classes by eco-region in the full data set and calibration data set. In the full data set, the eco-regions 0 and 1 to the south of Germany have more number of site-years in weather classes 8, 9, and 10 as compared to the other eco-regions. These weather classes (especially 9 and 10) are characterized by high precipitation. Eco-region 8 is dominated by weather classes 3 and 7 which have high late summer temperatures.

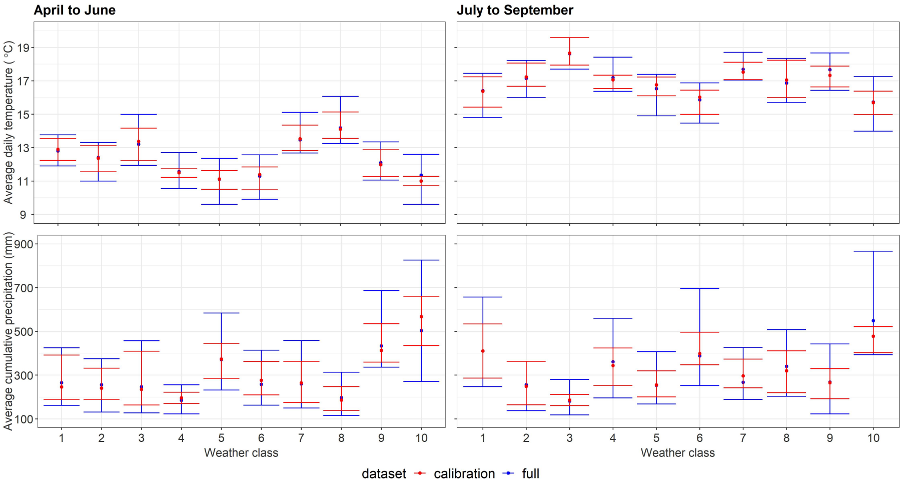


Figure S7.4: Summary statistics of the average daily temperature (°C) and cumulative precipitation (mm) per weather class in the two periods of April-June and July-September, for the full data set (3,004 site-years) and calibration data set (100 site-years). The points indicate the mean value while the error bars indicate the minimum and maximum values. Note that the k-means clustering used to generate the weather classes was based on the full data set.

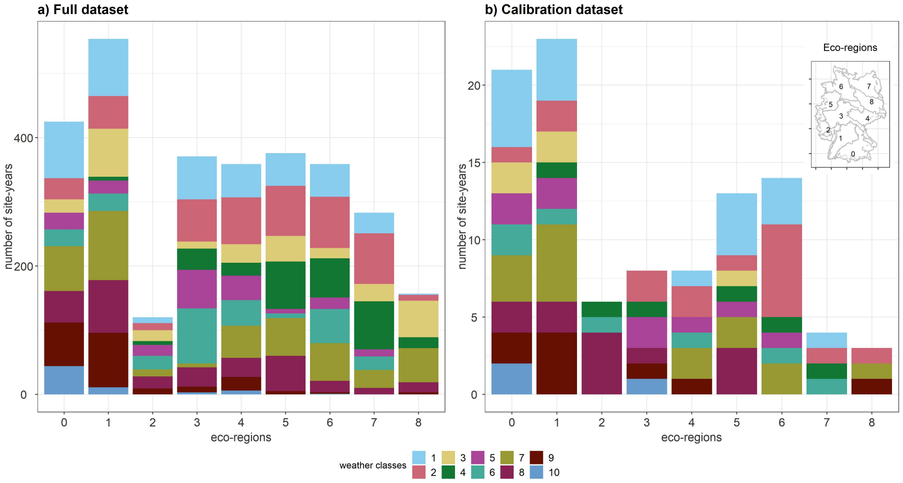


Figure S7.5: Weather class representation by eco-region for the full data set (a) and the calibration data set (b). The x-axes are the eco-regions in Germany. The y-axes are the number of site-years. The colours indicate the weather classes. The inset map shows the locations of the eco-regions (data source: BfN (Bundesamt für Naturschutz) (2017)).

Curriculum Vitae

Name	**Michelle Viswanathan**

Education

2007 – 2009	**M.Sc. Applied Geology** Indian Institute of Technology Roorkee, India
2004 – 2007	**B.Sc. Geology** St. Xavier's College - University of Mumbai, India

Work Experience

Oct 2022 – present	**Research Associate** Julius Kühn Institute, Kleinmachnow, Germany
Sep 2018 – Jul 2022	**Research Assistant** University of Hohenheim, Germany
Nov 2009 – May 2018	**Production Geologist** Shell India Markets Pvt. Ltd., India
May 2008	**Intern** Physical Research Laboratory, Ahmedabad, India

Presentations & Posters

EGU 2022 (*Presentation*)	An alternative strategy for combining likelihood values in Bayesian calibration to improve model predictions
AGU 2021 (*Poster*)	Improving water-balance of a coupled vegetation-hydrological model through Bayesian calibration to satellite-based Leaf Area Index
EGU 2021 (*Presentation*)	A Bayesian hierarchical approach to improve model parameter estimates and predictions of silage maize phenology in Germany
CAMPOS Conference 2020 (*Poster*)	Crop model predictions: The impact of environment-dependent parameters
ICROPM 2020 (*Presentation*)	Bayesian sequential updating of crop yield prediction in soil-crop-atmosphere systems
Agricultural Land Use and Feedbacks in a Changing Climate 2019 (*Poster*)	Sequential Bayesian updating for parameter estimation in soil-crop-atmosphere systems