From Bayesian “AND´´ to “OR´´ Calibration Strategy For More Reliable Predictions - A Demonstration on Plant Phenology Modelling

Michelle Viswanathan[1], Tobias K D Weber[1], and Anneli Guthke[2]

[1]University of Hohenheim
[2]Universität Stuttgart

November 22, 2022

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 (AND calibration strategy), it is inherently assumed that the model is error-free, and that different data sets are similarly informative for the inference problem. However, practically every model applied to real-world case studies suffers from model-structural 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 propose an alternative OR calibration strategy which allows the model to fit distinct data sets, individually. To demonstrate the effect of choosing between the traditional AND and the proposed OR strategy, we present a case study of calibrating a plant phenology model to observations of the maize crop grown in southwestern Germany between 2010 and 2016. We demonstrate that the OR strategy results in conservative but more reliable predictions than the AND strategy when the behaviour of the target prediction does not represent an average of all data sets. Further, an expert knowledge-based combination of AND-OR could be useful; however, selection of representative calibration data sets is not trivial. We expect our proposed strategy to improve the predictive reliability of imperfect, dynamic models in general, by a more realistic formulation of the likelihood function in the “perfect model setting” of Bayesian updating.

From Bayesian "AND" to "OR" Calibration Strategy for More Reliable Predictions - A Demonstration on Plant Phenology Modelling

Michelle Viswanathan[1], Tobias K.D. Weber[1], Anneli Guthke[2]

[1]Institute of Soil Science and Land Evaluation, Biogeophysics, University of Hohenheim, Stuttgart, Germany

[2]Stuttgart Center for Simulation Science, Cluster of Excellence EXC 2075, University of Stuttgart, Stuttgart, Germany

Key Points:

- Due to model errors, traditional Bayesian calibration on large/combined data sets typically leads to a sub-optimal compromised fit
- We propose an alternative strategy for combining data sets in Bayesian calibration to overcome this problem
- Our strategy estimates uncertainties more realistically leading to more reliable predictions

Corresponding author: Michelle Viswanathan, `michelle.viswanathan@uni-hohenheim.de`

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 (*AND calibration strategy*), it is inherently assumed that the model is error-free, and that different data sets are similarly informative for the inference problem. However, practically every model applied to real-world case studies suffers from model-structural 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 propose an alternative *OR calibration strategy* which allows the model to fit distinct data sets, individually. To demonstrate the effect of choosing between the traditional AND and the proposed OR strategy, we present a case study of calibrating a plant phenology model to observations of the maize crop grown in southwestern Germany between 2010 and 2016. We demonstrate that the OR strategy results in conservative but more reliable predictions than the AND strategy when the behaviour of the target prediction does not represent an average of all data sets. Further, an expert knowledge-based combination of AND-OR could be useful; however, selection of representative calibration data sets is not trivial. We expect our proposed strategy to improve the predictive reliability of imperfect, dynamic models in general, by a more realistic formulation of the likelihood function in the "perfect model setting" of Bayesian updating.

Plain Language Summary

Model parameters can be estimated though a process of calibration to observed data. Bayesian inference is commonly used for parameter estimation since it accounts for prior information and is able to account for different sources of uncertainty. Resultant parameter estimates and subsequent model predictions are expressed as probability distributions which are important while using these models for decision-making. However, the assumption in Bayesian inference, that the model is without errors, is usually not fulfilled, leading to an underestimation of uncertainty and wrong predictions. Part of the problem can be solved when formulating the so-called likelihood function in a different way: we propose an alternative strategy of combining the information in several data sets (e.g., different data types, different time periods with varying system conditions, etc.)

that relaxes this fundamental assumption. We compare the traditional and the alternative strategy in a case study where we calibrate a plant phenology model to observations from maize grown in southwestern Germany. The proposed alternative resulted in more reliable predictions than the traditional strategy when the data-to-be-predicted did not represent the average behaviour of all data sets used for calibration, and when the calibration data and conditions were representative of those in prediction.

Keywords: Bayesian calibration, maize phenology, model errors, model validation, prediction uncertainty, Bayesian modelling

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., 2006; Ajami et al., 2007), net ecosystem exchange (Weber et al., 2018), and crop modelling (e.g., Dumont et al., 2014; Wöhling et al., 2015; Gao et al., 2021; Viswanathan et al., 2022).

However, the fundamental assumption of Bayes theorem is that the underlying model structure is true, or when considering several models, that the true model is in this set. 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. In other words, model deficits exist that are expressed as errors in prediction (e.g., Wöhling et al., 2013; Viswanathan et al., 2022). 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).

Since there is no other theoretically satisfying and pragmatic alternative to the Bayesian approach, it is used despite the fact that the assumption of a true model is not fulfilled. The result is overconfident and biased parameter estimates and prediction intervals (Brynjarsdóttir & O'Hagan, 2014; Xu & Valocchi, 2015). One possible strategy to address this problem is to try and account for model error in the 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 & Valocchi, 2015; Makowski, 2017; Reichert et al., 2021). However, these approaches may incur high computational costs and are prone to parameter identifiability problems. As a somewhat ad-hoc alternative, it has been proposed to rather use shorter data sets for Bayesian calibration, in order to avoid the extreme narrowing of the posterior distribution (e.g., 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.

To overcome this situation, 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, i.e., combined via a logical "OR", not a logical "AND" as traditionally done for the full data set. 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 structural 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 proposed OR case, each sub-period for calibration might favour its own parameter sets, and these are combined to reflect the model's struggle with the varying boundary conditions and observed data more realistically.

Hence, our approach can be understood as an attempt to make Bayesian updating aware of model errors. It 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 contextualize the existing calibration technique mathematically, to compare the mathematical formulation of our proposed approach with the traditional approach, and make modelers aware of how their calibration decisions affect the model performance.

An evaluation of the impact of different likelihood combinations and functions on the result of crop model parameter estimation has been provided by He et al. (2010). Since they performed synthetic experiments without introducing model errors, the true model was in the set of possible model outcomes. This is exactly why they found that the AND strategy performs well in reducing posterior uncertainty the most. The problem emerges when we consider real-world modelling case studies with imperfect models, and this is the challenge we tackle here.

Instead of 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 AND vs. OR 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) and in distributed hydrological models (Immerzeel & Droogers, 2008).

We illustrate the performance of both the traditional AND and the new proposed OR calibration approach on the example of crop phenology modelling. Crop models at regional scales can be used for climate impact assessment, future crop production and food security evaluation as well as for investigating the fate of agrochemicals in the environment (Chenu et al., 2017). An important state variable in these crop models is phenological development which influences other state variables such as plant biomass, leaf area index (LAI), and yield. Phenological development depends on environmental drivers and does not only differ between crop species (such as maize vs. wheat) but also between cultivars of the same species and the ripening groups to which these cultivars belong. In regional simulations, where we would like to draw inferences for the crop species as a whole, it is important to account for uncertainty about the predicted ripening groups. So a modeler might decide to gather all the information they have in the form of observed data from different ripening groups, combine them into one big data set, and perform Bayesian calibration on it - with the goal of preparing the model for "anything that could happen". Unfortunately, this decision is tragically wrong, because the outcome is an extremely narrow posterior predictive distribution that is likely to not have any (substantial) overlap with what is happening in the real system.

So what has gone wrong? By trying to fit all 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 AND 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 modeler 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 at all and practically no uncertainty left in the model parameters, which in reality are still quite uncertain.

We will first theoretically demonstrate that, in the typical likelihood formulation, the logical AND is the source of this problem and show how such a multi-data set calibration task may be framed mathematically with a more adequate OR calibration scheme. Secondly, we demonstrate the differences between both approaches in a real-world case study. We calibrate a plant phenology model using the traditional AND and the proposed OR approaches. We use phenology observations of silage maize which was grown in two regions in southwestern Germany between 2009 and 2016. 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 case study with known model deficits, we will derive recommendations on when the traditional AND strategy should be applied, when the proposed OR strategy is more appropriate for more reliable predictions, and when an in-between AND-OR strategy might be useful.

This article is structured as follows: We start by recalling Bayesian updating in Section 2.1 and the reasoning behind the traditional AND Bayesian likelihood formulation in Section 2.2. Then, we present the alternative OR strategy based on predefined subsets of a calibration data set in Section 2.3, and the mixed specification of AND-OR in Section 2.4. We explain the skill score used to compare both approaches in Section 2.5. Section 3 features the phenology modelling case study. Results of the different calibration strategies are discussed in Section 4. General conclusions and an outlook towards further potential adaptations of our proposed approach are given in Section 5.

2 Bayesian Model Calibration

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 $\boldsymbol{y^o}$:

$$p\left(\boldsymbol{\theta}|M,\boldsymbol{y^o}\right)=\frac{p\left(\boldsymbol{y^o}|M,\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|M\right)}{p\left(\boldsymbol{y^o}|M\right)}. \tag{1}$$

For the sake of brevity, we omit the notation $(\cdot|M)$ (conditional on model M) in the following, 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.

2.2 Likelihood Formulation in the Traditional AND 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:

$$p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{AND} = p\left(y^{o,1} \cap y^{o,2} \cap \ldots \cap y^{o,N_o}|\boldsymbol{\theta}\right) = \prod_{j=1}^{N_o} p\left(y^{o,j}|\boldsymbol{\theta}\right) \tag{2}$$

The notation of Eq. 2 explicitly shows 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.

2.3 Likelihood Formulation in the Proposed OR Calibration Scheme

Instead of the traditional AND 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 with an OR-condition. Mathematically this is achieved by replacing the product with a sum in the equation. Here, we show the extreme case of subdividing into individual data points for the ease of notation:

$$p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{OR} = p\left(y^{o,1} \cup y^{o,2} \cup \ldots \cup y^{o,N_o}|\boldsymbol{\theta}\right) = \sum_{j=1}^{N_o} p\left(y^{o,j}|\boldsymbol{\theta}\right). \tag{3}$$

This can be interpreted as requiring the model to fit *either* data $y^{o,1}$ *or* data $y^{o,2}$ *or* data $y^{o,3}$, and so on. Through the sum over all data values, a parameter sample will score a high likelihood if it fits one data value extremely well, or many data values sufficiently well. Badly predicted values will reduce the score, but not to the extreme extent as in the traditional AND scheme. Additionally, if any likelihood $p\left(y^{o,j}|\boldsymbol{\theta}\right) = 0$, the combined likelihood $p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{OR}$ does not necessarily equal zero, as it would in case of $p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{AND}$.

In actual applications, one would select data subsets that contain several values, since the calibration effect of a single data point is very weak. Selecting an ideal length of the subsets can be a challenge - the periods should be long enough to achieve a “healthy” calibration effect on that data, but short enough (time-wise) or specific enough (data type-wise) to assume constant system conditions for the model to mimic (see the related discussion of Hsueh et al. (2022) on the choice of an optimal window length for time-windowed Bayesian model error analysis). When using data subsets (instead of individual data points) for the OR calibration scheme, this could be named an AND-OR strategy in a strict sense (see Section 2.4).

2.4 Likelihood Formulation in an AND-OR Calibration Scheme

We now turn to a mixture between the two previously described schemes which 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. In such a scenario that is typical of real-world conditions, we propose to use an AND-OR calibration strategy:

$$p\left(\boldsymbol{y^o}|\boldsymbol{\theta}\right)_{AND-OR} = p\left(\boldsymbol{y_1^o} \cup \boldsymbol{y_2^o} \ldots \cup \boldsymbol{y_{N_s}^o}|\boldsymbol{\theta}\right) = \sum_{s=1}^{N_s} p\left(\boldsymbol{y_s^o}|\boldsymbol{\theta}\right), \tag{4}$$

with N_s subsets of data. Within each subset s, the traditional AND scheme is used to determine the joint likelihood of the N_d data values:

$$p\left(\boldsymbol{y_s^o}|\boldsymbol{\theta}\right) = p\left(y_s^{o,1} \cap y_s^{o,2} \cap \ldots \cap y_s^{o,N_d}|\boldsymbol{\theta}\right) = \prod_{j=1}^{N_d} p\left(y_s^{o,j}|\boldsymbol{\theta}\right) \tag{5}$$

2.5 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:

$$PLS = \sum_{j=1}^{N_t} \log p(y^{t,j}|\boldsymbol{\theta}, \boldsymbol{y^o}) \tag{6}$$

Note, that we do not specify how the calibration on $\boldsymbol{y^o}$ was performed (AND vs. OR vs. AND-OR), 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 an AND 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 AND 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 3); however, our proposed calibration scheme is independent of the chosen metric such that modelers could decide to use other skill scores to reflect their individual modelling goals.

3 Demonstration in a Crop Phenology modelling Case Study

3.1 Motivation and Goals

We apply and compare the traditional AND calibration strategy with our proposed OR and AND-OR strategies 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 an important state variable in crop models as it influences the appearance of different plant organs and partitioning of assimilates. 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 AND calibration strategy results in a compromised solution that may not always lead to reliable predictions (Viswanathan et al., 2022).

The proposed OR calibration strategy has the potential to improve predictions by relaxing the model's prediction intervals and allowing the model to fit each data set individually. To assess the prediction performance with the OR 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 compare the calibrated model's prediction performance from the two strategies using the predictive log-score (PLS) (Section 2.5).

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. (2022) 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 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 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 classified into three groups: (1) low (≤15.4°C), (2) mid (>15.4°C and ≥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 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

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, Palosuo, Thorburn, Gourdain, et al., 2021; Wallach, Palosuo, Thorburn, Hochman, et al., 2021) 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, 2022) 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 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 (*pddv* and *pddr*, 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).

The six model parameters estimated during calibration were: effective sowing depth (*sowdepth*), physiological development days at optimum temperature (*pddv*, *pddr*), 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 2. Ranges for the estimated SPASS model parameters used to define weakly informative prior distributions.

Parameter	Description	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

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

$$p(\boldsymbol{\theta}|\boldsymbol{y_s^o})_{AND} = \frac{p(\boldsymbol{\theta}) \cdot \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \cdot \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{7}$$

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 AND 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 modeler 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 AND strategy.

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

$$p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}})_{AND} = \frac{p(\boldsymbol{\theta}) \cdot \prod_{s=1}^{N_s-1} \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \cdot \prod_{s=1}^{N_s-1} \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{8}$$

The alternative OR 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 *OR* case, the posterior probability of the parameters is given by

$$p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}})_{OR} = \frac{p(\boldsymbol{\theta}) \cdot \sum_{s=1}^{N_s-1} \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \cdot \sum_{s=1}^{N_s-1} \prod_{d=1}^{N_d} p(y_s^{o,d}|\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{9}$$

Note the subtle difference between Eqs. 8 and 9: in Eq. 8 a double product is used, while Eq. 9 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 (OR). Strictly speaking, this is already an instance of the AND-OR strategy (Section 2.4). However, in the context of this case study, we distinguish between AND and OR with respect to how data from different site-years are treated. In principle, the AND combination within a single site-year across different development phases (emergence, vegetative and reproductive) could be questioned and changed into OR or AND-OR as well. 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

$$p(\boldsymbol{y^o_{N_s}}|\boldsymbol{y^o_{1:N_s-1}}) = \int p(\boldsymbol{y^o_{N_s}}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}}) d\boldsymbol{\theta}, \tag{10}$$

with the posterior parameter distributions $p(\boldsymbol{\theta}|\boldsymbol{y^o_{1:N_s-1}})$ obtained from either the AND (Eq. 8) or the OR case (Eq. 9).

3.5 Test Case Scenarios

We compare the AND and OR calibration strategies using the predictive log-score (PLS) in predicting phenology at all 11 site-years (Table 1). For each prediction target site-year, the SPASS phenology model was calibrated to the 10 remaining site-years (leave-one-site-year-out). We also test the AND-OR scenario, using a selected subset of site-years for calibration in which we combine likelihoods from site-years within the same group using AND and across groups using OR. The test case scenarios are summarized in Fig. 1.

In the AND scenario, likelihood values from the calibration site-years are combined using Eq. 8 while in the OR scenario they are combined using Eq. 9. For the AND-OR scenario, we subdivide the data based on knowledge about the model's performance. A previous study (Viswanathan et al., 2022) 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 AND-OR scenario, only site-years which were from the same vegetative temperature class (Table 1) as the prediction target site-year were used for calibration. Knowledge about the cropping system was then used 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 using AND (Eq. 8) and across ripening groups were combined using OR (Eq. 9). For example, in the AND-OR 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. Likelihoods from site-years 1-2014 and 2-2014 in the mid-early ripening group were combined using AND. This was then combined using OR with the likelihood from 2-2012 in the late ripening group and the likelihood from 5-2015 in the early ripening group. Note, that there is no test case for predicting 5-2011 in the AND-OR scenario as there were no other site-years from the same temperature class.

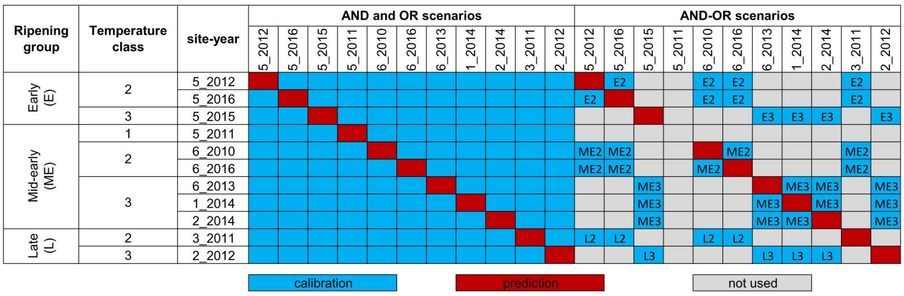

Figure 1. AND, OR, and AND-OR test case scenarios. For each case represented by a vertical column, the prediction target site-year is marked in red while the site-years used for calibration are marked in blue. For the AND-OR cases, site-years not used for calibration are in grey, while those site-years that were used for calibration are labeled with their respective ripening group 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 using AND strategy. Likelihoods across ripening groups were combined using OR strategy.

3.6 Graphical Illustration of AND vs. OR vs. AND-OR Strategy

We illustrate the concept behind the AND and OR scenarios with a Venn diagram (Fig. 2) in the context of the maize phenology model. Note that the site-years discussed in Fig. 2 are only for illustration. The squares represent the parameter space formed by uniformly distributed priors of two parameters that are plotted on the horizontal and vertical margins. The three circles X, Y and Z represent the posterior parameter space if the model were calibrated individually to three data sets corresponding to the maize cultivar A grown in site-year 1-2004, 2-2004 and cultivar B grown in site-year 4-2008, respectively. In this theoretical example, the degree of overlap between the circles is representative of the similarity between the data sets. Thus, the information in site-years 1-2004 and 2-2004 is assumed more alike than it is similar to 4-2008. The red-shaded area represents the resultant posterior probability densities arising from calibration if the individual likelihoods for these site-years were combined as per the traditional AND (Fig. 2a), AND-OR (Fig. 2b), and OR (Fig. 2c) scenarios. The AND and OR scenarios represent two extremes on this spectrum. In the AND scenario (Fig. 2a), the three site-years are assumed to provide similar information. Therefore, the likelihoods are combined as $p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta}) \cap p(\boldsymbol{Y}|\boldsymbol{\theta}) \cap p(\boldsymbol{Z}|\boldsymbol{\theta})$, where $p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})$ represents the probability of the data $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ corresponding to the three site-years, given the parameter vector $\boldsymbol{\theta}$. Since the site-year data sets are not very similar and lead to a limited overlap in acceptable parameter sets, we observe a collapse of the posterior parameter distribution represented by the red-shaded area.

On the other hand, all three site-years are considered to be distinct in the OR scenario (Fig. 2c), and to provide complementary information for parameter estimation. Here the likelihoods are combined as $p(\boldsymbol{X}|\boldsymbol{\theta}) \cup p(\boldsymbol{Y}|\boldsymbol{\theta}) \cup p(\boldsymbol{Z}|\boldsymbol{\theta})$. The resultant posterior parameter distribution encompasses the total area occupied by the three individual circles.

If, however, knowledge of the cropping system tells us that the cultivar A in year 2004 at sites 1 and 2 would have similar phenological development, then we can choose to combine their likelihoods using the AND strategy while the data from cultivar B is combined to them using the OR strategy as $(p(\boldsymbol{X}|\boldsymbol{\theta}) \cap p(\boldsymbol{Y}|\boldsymbol{\theta})) \cup p(\boldsymbol{Z}|\boldsymbol{\theta})$. This special case is referred to as the AND-OR scenario (Fig. 2 2b) which can be interpreted as an intermediate solution between the two extremes.

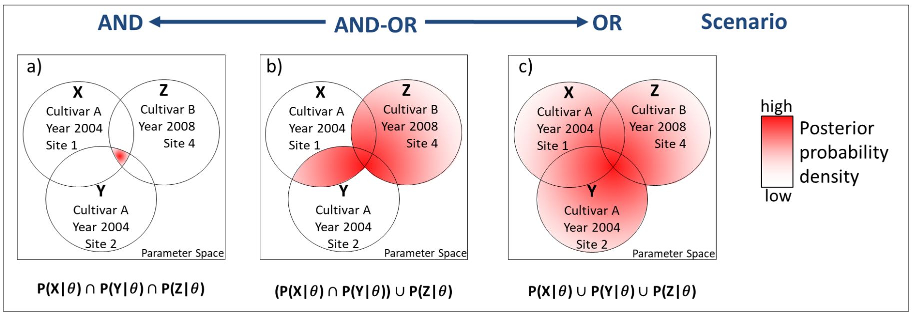

Figure 2. Venn diagram to illustrate the (a) AND calibration strategy, the (b) OR strategy, and an example of the (c) AND-OR strategy. The squares represent the uniform prior parameter space formed by two parameters. The three circles represent the posterior parameter space when the model is calibrated individually to data X and Y from cultivar A in site-year 1-2004 and 2-2004, respectively, and data Z from cultivar B in 4-2008. The shades of red indicate the resultant posterior parameter density when using the AND, OR, and AND-OR strategies to combine the likelihood values from the three site-years.

3.7 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, $f(\boldsymbol{\theta}_i)$, $i = 1 \ldots N_{MC}$, 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).

Considering the shape of the likelihood function, we assumed that the standardized 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.

$$p(y_s^{o,d}|\boldsymbol{\theta}) = \frac{1}{\sigma_s^d \sqrt{2\pi}} \exp\left(- \frac{y_s^{o,d} - f(\boldsymbol{\theta})_s^d}{2\sigma_s^d} \right)^2 \tag{11}$$

The Effective Sample Size (ESS, Liu (2008)) was estimated to ensure that a large enough number of ensemble members contribute to posterior statistics. Obtained ESS values range from < 10 for the AND scenario to 2,000 $<$ ESS $<$ 4,000 for the OR scenario with $N_s{-}1$ calibration site-years. The ESS starts to drop below 20 in the AND scenario 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 (cf. also the visual illustration of the very small posterior parameter space in Fig. 2a). Hence, the reliability of these AND prediction results is questionable, but we still show them for discussion.

In contrast, the ESS values in the OR calibration strategy show that this sampling problem can be mitigated by our proposed approach because the sampling method does not have to struggle as hard to find suitable parameter values. In the AND-OR scenario in which only a selected subset of site-years is used for calibration, the ESS ranges between 200 $<$ ESS $<$ 1,500. Here, the sampling problem is mitigated due to both, data set selection as well as the AND-OR strategy. For comparison, in those cases of selected subsets of calibration site-years, the ESS ranges between $50 < \text{ESS} < 200$ in the AND scenario and $1{,}000 < \text{ESS} < 2{,}000$ in the OR scenario. 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 $500 < \text{ESS} < 2{,}000$ (900 on average).

4 Results and Discussion

For the purpose of discussion, we present selected results of the leave-one-site-year-out cross-validation exercise. AND and OR scenarios are shown for predictions of the early cultivar at 5-2012 (Fig. 3a), the mid-early cultivar at 6-2010 (Fig. 3b), and the late cultivar at 3-2011 (Fig. 3c). We also present the results of the AND-OR scenario applied to predictions of site-years 2-2014 (Fig. 4a) and 6-2016 (Fig. 4b). The PLS of all other investigated cases are summarized in Fig. C1 in Appendix C.

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 only. 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 AND-OR scenario test cases, we additionally present results from the AND and OR scenarios where only the selected subsets of site-years were used as opposed to all N_s-1 remaining site-years. 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 AND-OR strategy across ripening groups can further improve performance. To distinguish the AND and OR cases from these additional scenarios, we will label the AND and OR cases based on N_s-1 site-years as AND_all and OR_all, respectively.

4.1 OR Strategy is Conservative but Reliable

For all three target site-years shown in Fig. 3, the idealized case of calibrating on the target site-year only (first column in Fig. 3) yields accurate mean predictions and tight credible intervals, with observation uncertainty being partly larger than model parameter and model error uncertainty.

The traditional AND_all calibration strategy (second column), however, performs very differently, depending on the analyzed target site-year. For site-year 5-2012 (Fig. 3a), the prediction interval in the AND_all scenario 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, different cultivars, different temperature classes) 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.

The proposed OR_all strategy (third column in Fig. 3), 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 OR_all scenario, and this is why the calibration is less strong and allows for more variability in the posterior credible intervals. The OR_all intervals succeed in capturing all target data points. This is also reflected in the PLS values (fourth column in Fig. 3) with that of the OR_all scenario being higher than the AND_all scenario. Compared to the idealized case of calibration on this site-year only, the OR_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 OR_all calibration strategy leads to conservative but more reliable prediction results than the AND_all strategy. The is also observed for the prediction of phenology at site-years 5-2015, 6-2013, 5-2011, and 2-2014 (Fig. C1).

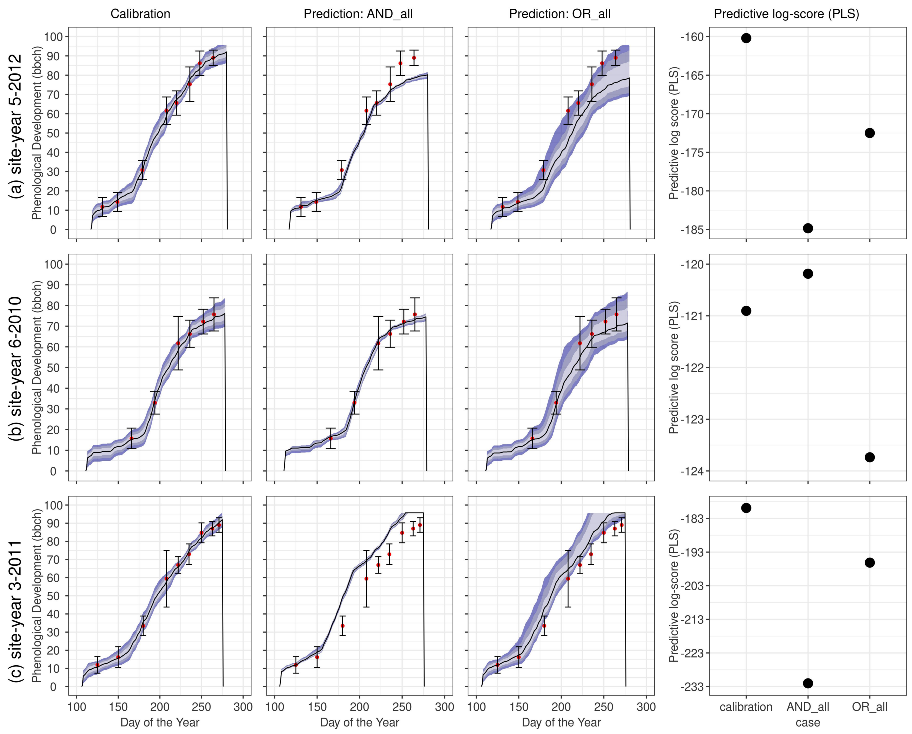

Figure 3. Observed and simulated phenology at site-years (a) 5-2012, (b) 6-2010, and (c) 3-2011. First column shows posterior credible intervals obtained from calibration on the target site-year only; second and third columns show posterior credible intervals from AND_all and OR_all calibration scenarios, 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.

4.2 AND Strategy Succeeds when the Target Represents an Average Behaviour

In the prediction of phenology at site-year 6-2010 (Fig. 3b), the OR_all scenario performs worse than the AND_all scenario due a special feature of maize phenological development. Here, the AND_all scenario prediction performs really well and captures the data points even better than the calibration reference as shown by the PLS values. The AND_all scenario 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 behavior. While the PLS value of the prediction in the AND_all scenario 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 AND_all scenario. 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 OR_all scenario yielded wider prediction intervals, but this time the loss of precision resulted in a lower PLS value than the AND_all scenario. This is because the AND_all scenario 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 AND_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 AND_all scenario, 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 AND_all scenario already performed very well, the relaxation of the prediction bands in the OR_all scenario led to poorer predictions. Similarly, prediction with the AND_all scenario was better than the OR_all scenario for the mid-early cultivars at 6-2016 and 1-2014 (the interested reader is referred to Fig. C1 in Appendix C).

4.3 Representativeness of the Calibration Data Plays a Role

In the case of site-year 3-2011 (Fig. 3c), the AND_all scenario results in poor predictions and the OR_all scenario 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, Palosuo, Thorburn, Gourdain, et al., 2021). 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. Yet, even though the same cultivar was grown at 2-2012, the AND_all calibration strategy was better than the OR_all strategy at prediction (Fig. C1). This site-year falls in the 'high' temperature class (Table 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 4.2). On the other hand, even though 5-2011 is a mid-early ripening cultivar, the OR strategy performs better than the AND. 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 OR_all scenario 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.

4.4 Data Set Selection for a Successful AND-OR Strategy is no Trivial Exercise

To test the potential of expert knowledge-based combination of selected site-years for calibration, only site-years 5-2015, 6-2013, 1-2014, and 2-2012 (all temperature class 3, cf. Fig. 1) were used for calibration with the AND-OR scheme in order to predict phenology at site-year 2-2014 (Fig. 4a). Recall that, in this approach, we combined site-years of the same ripening group by AND, and used OR across different ripening groups (Section 3.5). For comparison, we also show predictions of AND vs. OR scenarios *with only those site-years* (AND vs. OR scenarios), while all $N_s - 1 = 10$ non-target site-years were used for calibration in the AND_all vs. OR_all scenarios.

The traditional AND_all scenario leads to overconfident prediction intervals for this predicted site-year (Fig. 4v), and the OR_all case improves on that with wider intervals that succeed to capture all target data points. The question whether this uncertainty can be reduced again without making overconfident and biased predictions via the AND-OR scenario can be answered with yes in this case: the AND-OR prediction interval has become narrower without losing any data points (Fig. 4ii). This is also obvious from the increase in PLS (Fig. 4vii). 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 4.3) and/or by the combination of AND with OR. We find that the mere selection of site-years improves over the $N_s - 1$ cases (the PLS increases for AND vs. AND_all and OR vs. OR_all), but the AND-OR case indeed performs best (second after calibration on the target site-year only).

However, for the AND-OR scenario 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. 4b), the site-year selection resulted in a lower PLS in the AND case than in the AND_all case in which all the remaining 10 site-years were used for calibration, because the AND_all case yields very confident prediction intervals with relatively low bias. Naturally, calibrating on less data in the AND case then leads to a weaker calibration effect and a lower PLS. The OR case resulted in a marginal improvement in PLS as compared to the AND case (the wider intervals of OR now cover e.g. the last data value of the season better), while the AND-OR case performs worse. Yet, in the AND-OR and OR cases, all observations and their measurement uncertainty range is covered by the high-probability region of the predictive interval, which is not the case in the other calibration scenarios. Thus, when aiming at reliable predictions and rather accepting variance than bias, these strategies are better suited than the traditional AND_all case.

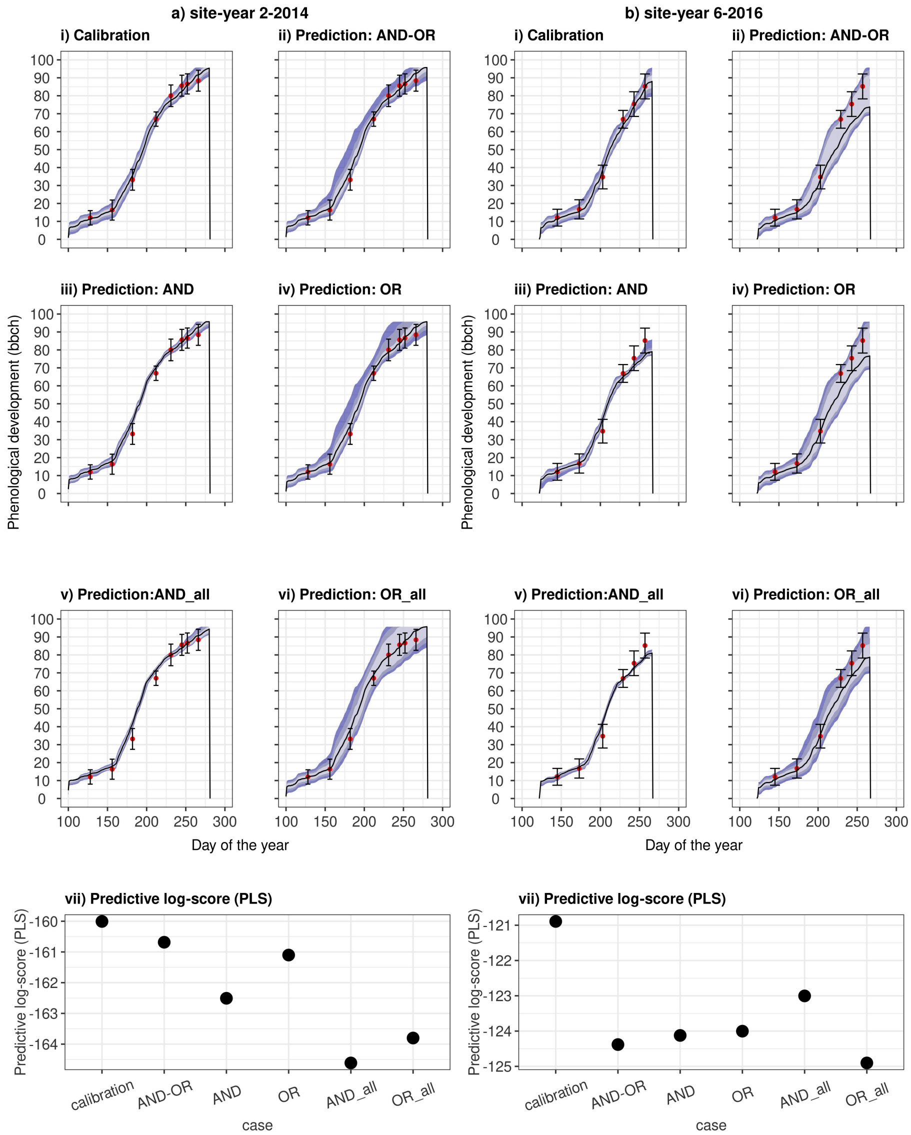

Figure 4. Observed and simulated phenology at site-years (a) 2-2014 and (b) 6-2016. Posterior credible intervals obtained from i) calibration on the target site-year only, ii) AND-OR calibration scenario, iii) AND scenario; iv) OR scenario, v) AND_all scenario, vi) OR_all scenario, and vii) summarizes 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 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 true (error-free), and hence 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. Forcing an imperfect model to fit diverse data sets simultaneously (what we call the *AND calibration strategy*) inevitably leads to a compromised solution to the parameter estimation problem, and triggers unreliable predictions. To overcome this problem, we have proposed an alternative *OR calibration strategy* 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 have 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 AND and the alternative OR strategies in a real-world case study where a plant phenology model was calibrated to silage maize observations from southwestern Germany. The model's performance in predicting a data set that was not used during calibration (leave-one-site-year-out cross-validation) 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 OR strategy resulted in higher scores when the predicted data set did not represent an average behavior of the calibration data sets (e.g., with respect to temperature class or ripening group). As a special case, we also tested a combined AND-OR strategy. To this end, only those data sets from the same temperature class as the prediction target were used for calibration. These data sets were then grouped by ripening group, wherein likelihoods within groups were combined with AND and across groups were combined using OR. While superior to the AND and OR strategies in some cases, we found that the AND-OR strategy requires a fine-grained 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 OR or AND-OR strategy. Testing this approach on different types of models and data sets and in different application scenarios is recommended for future work. Further, the prediction results in the AND-OR strategy could potentially benefit from implementing a data-driven approach to define the data groups in addition to expert knowledge, e.g., informed by model deficits which can be evaluated using calibration performance indicators such as residuals. We expect such advances to be very useful for environmental modelling studies where model structural errors are ubiquitous.

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 occured instantenously 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:

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

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

Appendix B Prior Distribution

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

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

where

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

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 2) and c_h is the normalization constant:

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

Appendix C Predictive Log-Score (PLS) for All Cases

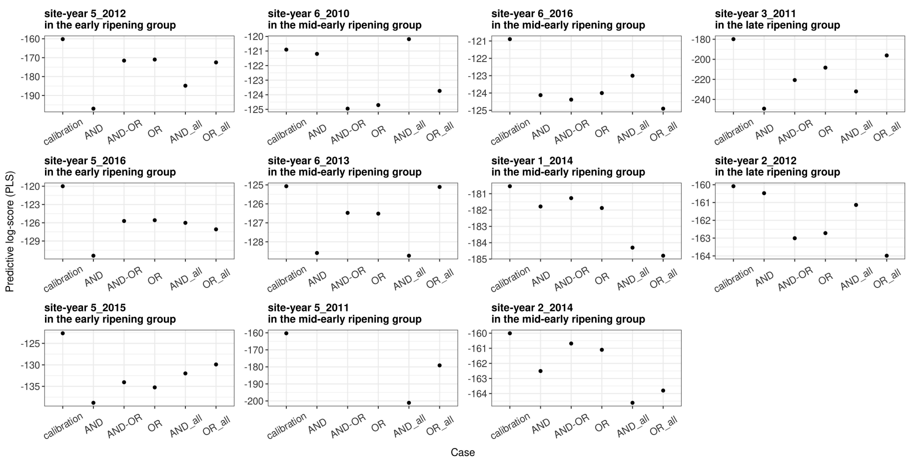

Figure C1. The predictive log-score (PLS) for calibration and prediction results. The predictions in the AND, ANDOR, and OR scenarios were made after calibrating the model to a selection of site-years for calibration. The predictions in the AND_all and OR_all scenarios were made after calibrating the model to all remaining site-years.

Data availability

All observational data used for the study are publicly available in (Weber et al., 2022).

Code availability

The R codes used for the study are available at (A link to the Zenodo repository will be provided upon acceptance).

Acknowledgments

The contribution of Michelle Viswanathan was made possible through the Integrated Hydrosystem Modelling Research Training Group, funded by the German Research Foundation (DFG, GRK 1829). The contribution of Tobias K.D. Weber was possible through the Collaborative Research Centre 1253 CAMPOS (Project 7: Stochastic Modelling Framework), funded by the German Research Foundation (DFG, Grant Agreement SFB 1253/1 2017). Anneli Guthke would like to thank the DFG for financial support of the project within the Cluster of Excellence EXC 2075 "Data-integrated Simulation Science (SimTech)" (project number 390740016). The authors thank Thilo Streck for supervision and the provision of resources for this study. The authors further acknowledge support by the state of Baden-Wuerttemberg through the HPC cluster bwUniCluster (2.0).

References

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*(1).

Bassu, S., Brisson, N., Durand, J. L., Boote, K., Lizaso, J., Jones, J. W., … Waha, K. (2014). How do various maize crop models vary in their responses to climate change factors? *Global Change Biology*, *20*(7), 2301–2320. doi: 10.1111/gcb.12520

Brynjarsdóttir, J., & O'Hagan, A. (2014, oct). Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, *30*(11), 114007. Retrieved from `https://doi.org/10.1088/0266-5611/30/11/114007` doi: 10.1088/0266-5611/30/11/114007

Ceglar, A., Črepinšek, Z., Kajfež-Bogataj, L., & Pogačar, T. (2011). The simulation of phenological development in dynamic crop model: The bayesian comparison of different methods. *Agricultural and Forest Meteorology*, *151*(1), 101-115. Retrieved from https://www.sciencedirect.com/science/article/pii/S0168192310002534 doi: https://doi.org/10.1016/j.agrformet.2010.09.007

Chenu, K., Porter, J. R., Martre, P., Basso, B., Chapman, S. C., Ewert, F., … Asseng, S. (2017). Contribution of crop models to adaptation in wheat. *Trends in Plant Science*, *22*(6), 472-490. Retrieved from https://www.sciencedirect.com/science/article/pii/S1360138517300389 doi: https://doi.org/10.1016/j.tplants.2017.02.003

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*(10), 4209–4225.

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

Durand, J.-L., Delusca, K., Boote, K., Lizaso, J., Manderscheid, R., Weigel, H. J., … 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. Retrieved from https://www.sciencedirect.com/science/article/pii/S1161030117300084 (Recent advances in crop modelling to support sustainable agricultural production and food security under global change) doi: https://doi.org/10.1016/j.eja.2017.01.002

Falconnier, G. N., Corbeels, M., Boote, K. J., Affholder, F., Adam, M., MacCarthy, D. S., … Webber, H. (2020). Modelling climate change impacts on maize yields under low nitrogen input conditions in sub-saharan africa. *Global Change Biology*, *26*(10), 5942-5964. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15261 doi: https://doi.org/10.1111/gcb.15261

Gao, Y., Wallach, D., Hasegawa, T., Tang, L., Zhang, R., Asseng, S., … 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., … Hoogenboom, G. (2020). Comparison of three calibration methods for modeling rice phenology. *Agricultural and Forest Meteorology*, *280*(September 2019), 107785. Retrieved from `https://doi.org/10.1016/j.agrformet.2019.107785` doi: 10.1016/j.agrformet.2019.107785

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 107–114.

Hastings, W. K. (1970, 04). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97-109. Retrieved from `https://doi.org/10.1093/biomet/57.1.97` doi: 10.1093/biomet/57.1.97

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*(5), 256-264. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0308521X1000017X` doi: https://doi.org/10.1016/j.agsy.2010.01.006

Heinlein, F., Biernath, C., Klein, C., Thieme, C., & Priesack, E. (2017, jan). Evaluation of Simulated Transpiration from Maize Plants on Lysimeters. *Vadose Zone Journal*, *16*(1), vzj2016.05.0042. Retrieved from `http://doi.wiley.com/10.2136/vzj2016.05.0042` doi: 10.2136/vzj2016.05.0042

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*(2), e2021WR030590.

Immerzeel, W., & Droogers, P. (2008). Calibration of a distributed hydrological model based on satellite evapotranspiration. *Journal of hydrology*, *349*(3-4), 411-424.

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. application. *Water resources research*, *42*(3).

Kimball, B. A., Boote, K. J., Hatfield, J. L., Ahuja, L. R., Stockle, C., Archontoulis, S., … Williams, K. (2019). Simulation of maize evapotranspiration:

An inter-comparison among 29 maize models. *Agricultural and Forest Meteorology*, *271*, 264-284. Retrieved from https://www.sciencedirect.com/science/article/pii/S0168192319300966 doi: https://doi.org/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*(13), vzj2017.03.0052. Retrieved from https://acsess.onlinelibrary.wiley.com/doi/abs/10.2136/vzj2017.03.0052 doi: https://doi.org/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*(1-2), 161–177.

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*(4), 1–23. doi: 10.1371/journal.pone.0195841

Liu, J. S. (2008). *Monte carlo strategies in scientific computing.* Springer Science & Business Media.

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

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water*, *5*(6), e1319.

Meier, U. (2018). *Growth stages of mono- and dicotyledonous plants: Bbch monograph.* Quedlinburg: Open Agrar Repositorium. Retrieved from https://www.openagrar.de/receive/openagrar_mods_00042351 doi: 10.5073/20180906-074619

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.

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*(1), 124–127. Retrieved from https://ijair.org/index.php/component/jresearch/?view=publication&task=show&id=591&Itemid=166

Priesack, E. (2006). *Expert-N Dokumentation der Modellbibliothek FAM – Bericht 60* (Tech. Rep.). Munich, Germany: GSF-Forschungszentrum fuer Umwelt und Gesundheit; Institut fuer Bodenoekologie.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

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*(3), e2020WR028400.

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*(5).

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*(12), 9484–9513.

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

Viswanathan, M., Weber, T. K. D., Gayler, S., Mai, J., & Streck, T. (2022). A bayesian sequential updating approach to predict phenology of silage maize. *Biogeosciences*, *19*(8), 2187–2209. Retrieved from https://bg.copernicus.org/articles/19/2187/2022/ doi: 10.5194/bg-19-2187-2022

Wallach, D., Palosuo, T., Thorburn, P., Gourdain, E., Asseng, S., Basso, B., . . . Seidel, S. J. (2021). How well do crop modeling groups predict wheat phenology, given calibration data from the target population? *European Journal of Agronomy*, *124*, 126195. Retrieved from https://www.sciencedirect.com/science/article/pii/S1161030120302021 doi: https://doi.org/10.1016/j.eja.2020.126195

Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., ... Seidel, S. J. (2021). The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. *Environmental Modelling & Software*, *145*, 105206. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1364815221002486` doi: https://doi.org/10.1016/j.envsoft.2021.105206

Wang, E. (1997). *Development of a generic process-oriented model for simulation of crop growth* (Unpublished doctoral dissertation). Technische Universität München, Germany.

Weber, T. K. D., 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*(6), 1319–1336. doi: 10.1029/2017MS001044

Weber, T. K. D., Ingwersen, J., Högy, P., Poyda, A., Wizemann, H.-D., Demyan, M. S., ... 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*(3), 1153–1181. Retrieved from `https://essd.copernicus.org/articles/14/1153/2022/` doi: 10.5194/essd-14-1153-2022

Wöhling, T., Gayler, S., Priesack, E., Ingwersen, J., Wizemann, H.-D., Högy, P., ... 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*(12), 8200–8221.

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*(4), 2825–2846. doi: 10.1002/2014WR016292

Xu, T., & Valocchi, A. J. (2015). A bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, *51*(11), 9290-9311. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017912` doi: https://doi.org/10.1002/2015WR017912

Zhao, M., Peng, C., Xiang, W., Deng, X., Tian, D., Zhou, X., ... Zhao, Z. (2013, mar). Plant phenological modeling and its application in global climate change research: overview and future challenges. *Environmental Reviews*, *21*(1), 1–14. Retrieved from http://www.nrcresearchpress.com/doi/10.1139/er-2012-0036 doi: 10.1139/er-2012-0036