Received: 31 January 2025 | Accepted: 6 June 2025

DOI: 10.1002/agj2.70111

Agronomy Journal

ORIGINAL ARTICLE

Biometry, Modeling, and Statistics

Impact of calibration strategy and data on wheat simulation with the DSSAT-Nwheat model

Ashifur Rahman Shawon[1,4] | Ahmed Attia[1] | Jonghan Ko[2] | Emir Memic[3] | Ralf Uptmoor[4] | Bernd Hackauf[5] | Til Feike[1]

[1]Institute for Strategies and Technology Assessment, Julius Kühn Institute (JKI)–Federal Research Centre for Cultivated Plants, Kleinmachnow, Germany

[2]Department of Applied Plant Science, Chonnam National University, Gwangju, Republic of Korea

[3]Department of Agronomy, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

[4]Department of Agronomy, University of Rostock, Rostock, Germany

[5]Institute for Breeding Research on Agricultural Crops, Julius Kühn Institute (JKI)–Federal Research Centre for Cultivated Plants, Groß Lüsewitz, Germany

Correspondence

Ashifur Rahman Shawon, Institute for Strategies and Technology Assessment, Julius Kühn Institute (JKI)–Federal Research Centre for Cultivated Plants, Stahnsdorfer Damm 81, Kleinmachnow, 14532, Germany.
Email: shawon.ashifur@gmail.com, ashifur.shawon@julius-kuehn.de

Assigned to Associate Editor QuanXiao Fang.

Funding information

This research was supported by the project PRoKlima funded by the German Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) under the Climate Protection Program 2022 and by the project RYE-SUS under the ERA-NET Cofund SusCrop, Grant/Award Number: N°771134.

Abstract

Cropping system models (CSMs) are valuable tools for analyzing genotype, environment, and management (G × E × M) interactions in crop production. To apply a CSM in a new region with specific soils, climate, and cultivars, proper calibration and evaluation are required. However, calibration methods vary widely, often depending on modelers' expertise and approach. This study compares three calibration strategies for the DSSAT-Nwheat model using two datasets: one including yield components (1000-kernel mass, ears per m^2, grain number per m^2) alongside phenology and grain yield, and another excluding yield components. The datasets cover ~100 site-years of winter wheat (*Triticum aestivum*) data from German pre-registration trials and field experiments. The calibration approaches were (1) stepwise calibration of phenology, biomass, and yield, (2) simultaneous calibration of multiple genetic coefficients, and (3) a hybrid approach combining elements of both. The Time-Series cultivar coefficient estimator tool was used for implementation. Including yield component data improved model accuracy, reducing root mean square error (RMSE) by up to 10% for key variables such as phenology (3.4–5.5 days). Future wheat yield projections under selected climate scenarios varied by strategy and dataset, ranging from 6376 to 7473 kg ha^{-1} in fertile, wet soils and 6108 to 6757 kg ha^{-1} in poorer, dry soils. These results highlight the impact of calibration strategy and dataset choice on model performance. Transparent calibration practices are essential for improving CSM reliability in regional agricultural analysis under diverse environmental conditions.

Plain Language Summary

Crop models help scientists understand how wheat production might change under different environmental conditions. To use these models effectively in a new region, they must first be adjusted using local data—a process called calibration. In this study, we tested three calibration strategies for the DSSAT-Nwheat model using winter wheat data from Germany. We also compared two types of datasets: one with detailed information on yield components like grain size and number, and one without. Including this additional data made the model more accurate. We found that both the calibration method and the data used had a strong effect on future yield predictions, with differences of up to 1.4 t per ha depending on the scenario. These results highlight the need for transparent and well-documented calibration approaches to improve confidence in crop model applications, especially when used for climate impact assessments in regional agriculture.

1 | INTRODUCTION

Agricultural practices have evolved substantially over millennia, with farmers traditionally adapting their methods to local conditions based on empirical knowledge and experience (Boserup & Chambers, 2014). The selection of arable crops and cultivars was largely guided by instinct and trial-and-error, reflecting a deep understanding of the local environment (Brookfield & Padoch, 1994). However, this approach often faced challenges due to the unpredictable nature of weather and soil conditions, which substantially influence crop performance (Porter & Semenov, 2005). Despite the advancements in agricultural technology, the fundamental challenges posed by environmental variability remain pertinent (Lobell et al., 2011).

Across the past decades, an array of crop models has evolved, intricately capturing the interplay among genotype, environment, and management (G × E × M) factors. This dynamic representation of G × E × M effects has firmly established process-based crop models as a pivotal tool for a comprehensive understanding of cropping systems (Brisson et al., 2003; Keating et al., 2003; Stöckle et al., 2003). Prominent models in this category include the DSSAT modeling framework (Hoogenboom et al., 2019; Jones et al., 2003), APSIM (McCown et al., 1996), WOFOST (Diepen et al., 1989), AQUACROP (Steduto et al., 2009), STICS (Brisson et al., 2003), SIMPLACE (Gaiser et al., 2013), HERMES (Kersebaum, 2007), InfroCrop (Krishnan et al., 2016), CropSyst (Stöckle et al., 1994), SSM (Sinclair, 1986), DAISY (Hansen et al., 1991), MONICA (Nendel et al., 2011), and others. These models meticulously interconnect mathematical equations to elucidate the complex web of interactions encompassing crops, soil, and environmental factors. Although these equations are a simplification of actual processes and interconnections, they effectively capture the underlying dynamics of crop growth and development in response to different stimuli (de Wit, 1982).

The inclusion of model parameters that capture the combined effect of G × E × M factors plays a crucial role in determining the behavior of the model (Tardieu et al., 2020). These models have significantly advanced our understanding of cropping systems, but the insights they generate depend heavily on targeted parameterization, achieved through calibration and evaluation. Calibration involves adjusting model parameters to reflect G × E × M interactions, while evaluation ensures that predictions align with observed data (Wallach et al., 2006).

Reasonable ranges of parameters refer to plausible upper and lower limits for crop model parameters, informed by empirical data, expert knowledge, or literature. However, accurately defining such ranges remains challenging due to the variability introduced by diverse physiographic, climatic, and soil conditions across regions. Calibration is the primary approach for determining optimal parameter values in crop models, achieved through iterative adjustments to align simulated outputs with observed data (Wallach, Palosuo, Thorburn, Hochman, et al., 2021). In contrast, model evaluation involves validating the model's accuracy using independent datasets, ensuring robustness, and avoiding overfitting (Boote et al., 1996). In this regard the debate persists on the quantity of observed data suitable for calibration. While some experts advocate for extensive data inclusion, others caution that using too specific data can lead to overfitting, where the model becomes too tailored to the calibration data and performs poorly in other conditions (Confalonieri et al., 2016; Pasley et al., 2023).

This challenge is compounded when scaling model calibration to regional levels, where diverse environmental conditions—such as those influenced by climate change—add complexity to crop growth and yield predictions (Xiong et al., 2008). Regional applications of the cropping system model (CSM) are often required when addressing real-world agricultural challenges, which requires targeted model parameterization. Tailoring crop parameters to specific cultivars grown in the target regions is crucial to ensuring model accuracy (Rötter et al., 2012). Yet, the broader applicability of CSM is often hindered by the limited availability of empirical data, particularly regarding G × E × M interactions (Angulo et al., 2013). Genotype, often represented by pre-set coefficients, introduces specific uncertainties, and accurate calibration is vital, especially when exploring new locations or crop varieties (Balkovič et al., 2013).

Furthermore, scarcity of cultivar-specific data for calibration often restricts the ability to scale models effectively to regional or national levels.

Plant variety trials offer a potential solution to this data limitation. These trials provide multi-year, multi-location data, offering valuable insights into site-specific characteristics and cultivar traits that can inform the estimation of genetic coefficients for crop models. Multi-environment data on crop phenology, plant height, and grain yield components—readily available in performance trials—constitute valuable data enabling comprehensive model calibration. Additional datasets comprising in-season growth data, for example, on the development of leaf area and biomass over the growing season, can add value (Shawon et al., 2024). Solid crop performance data from a wide range of environments are the basis for predictions across diverse growth conditions (Hammer et al., 2010).

To effectively utilize such large multi-environment trial data, classical manual (trial and error) model calibration is not a viable option. Different automated model calibration procedures are available to comprehend such tasks, including Genotype Coefficient Calculator (GENCALC), Generalized Likelihood Uncertainty Estimation (GLUE), Time Series Estimator (TSE), Markov Chain Monte Carlo (MCMC), Parameter ESTimation (PEST) and CroptimizR (Berton Ferreira et al., 2024; Ceglar et al., 2011). However, the effectiveness of these methods depends on the choice of calibration strategy and dataset selection (Wallach et al., 2024).

Despite the availability of calibration techniques, the impact of different calibration strategies on crop model performance remains insufficiently explored. In particular, limited research has systematically evaluated how alternative calibration approaches influence model coefficient settings, performance indices, and simulated crop outcomes (Di Paola et al., 2016; Rötter et al., 2012). Understanding these effects is essential for improving model application in different environments, particularly under climate variability and changing management conditions.

This study investigates the impact of three calibration strategies using two multi-environment datasets on the parameterization and performance of the DSSAT-Nwheat model. Specifically, this study assesses the effects of these strategies on (i) cultivar coefficient settings, (ii) selected model performance indices, and (iii) simulated crop outcomes. By comparing different calibration approaches, this study aims to improve the understanding of model parameterization and its implications for crop simulations across diverse environments. Additionally, the findings contribute to refining crop model applications in novel environments, facilitating more effective collaboration among modelers, breeders, and physiologists.

Core Ideas

- This study evaluates three DSSAT-Nwheat calibration strategies, utilizing both comprehensive and limited datasets.
- The research examines how different calibration approaches affect key phenological and biomass-related parameters.
- Findings show dataset selection and calibration strategies influence crop model accuracy across environments.

2 | MATERIALS AND METHODS

2.1 | DSSAT-Nwheat model

The CSM-Nwheat was developed by Keating et al. (2001) building on the CSM-CERES (Ritchie et al., 1985). In CSM-Nwheat, crop phenology is a function of accumulated degree-days, photoperiod, and vernalization requirements. The potential production of biomass is driven by radiation use efficiency (RUE), incident solar radiation, temperature, and leaf area index (LAI) (Kassie et al., 2016). A suboptimal supply of water and nitrogen reduces the potential production of biomass. Potential crop demand for water is a function of transpiration efficiency and vapor pressure deficit. Actual water uptake depends on the uptake demand, root density, and available soil water in the soil profile. Grain yield is determined by grain quantity, grain filling, and carbohydrate remobilization. Carbohydrates available for remobilization are defined in the model as 75% of biomass growth from pre-anthesis to the beginning of the linear grain-filling phase.

The water balance in the CSM-Nwheat is shaped by soil characteristics such as lower limit (LL), drained upper limit (DUL), and saturated volumetric water content (SAT) provided as model inputs, which is similar to the CSM-CERES. The water balance model follows a cascading approach, where water flows vertically from top to bottom when the soil layer's water content surpasses field capacity. The model further considers how crop cover and surface residue impact runoff and soil evaporation. The CSM-Nwheat simulates increased atmospheric CO_2 concentration on crop growth, RUE, and transpiration efficiency using separate functions following the approach from Reyenga et al. (1999). Elevated atmospheric CO_2 concentrations affect transpiration efficiency by multiplying it by a factor that increases linearly from 1.0 to 1.37 as the CO_2 concentration rises from 350 to 700 ppm (Asseng et al., 2004).

Additional relevant adaptions from CSM-CERES to CSM-Nwheat include (i) the modification of the root-water-uptake routine based on fraction of available soil water, (ii) the crop-water-demand module linked with biomass production via transpiration efficiency, (iii) simulation of frost damage, (iv) tillering, and (v) specific leaf area. These adjustments made the CSM-Nwheat especially suitable for simulations in a broader range of environmental conditions including rising temperatures (Asseng et al., 2004), soils with low water-holding capacity (Asseng et al., 2008), and extreme heat events (Guarin et al., 2020). For this study, the CSM-Nwheat within the DSSAT platform was selected, as it is a well-tested and widely applied crop model for wheat (*Triticum aestivum*) (Kassie et al., 2016). It uses the general soil and management module of the DSSAT framework.

2.2 | Soil and weather data

The DSSAT-Nwheat requires layer wise data on soil texture, bulk density, and soil organic carbon content to define the soil's drained lower limit (DLL) and DUL and simulate soil water and soil nitrogen balance accurately. As only a very rough description on the different experiments soil type was available for the different experiments, site-specific soil data were retrieved from the European Soil Database (ESDB) v2.0, as described by Panagos et al. (2012) based on the experiment-specific geo-coordinates. This soil database is available in a 1-km^2 grid and, for Germany, builds on the 1:10^6 German soil map (Hartwich et al., 1995). Soil texture data were transformed from the German into the USDA scale (Rawls et al., 1982) using a log-log transformation using the r-package "The soil texture wizard" (Moeys, 2012).

Weather data for 1991–2019 were retrieved from the climate data center of the German weather service (Deutscher Wetterdienst, 2017), which is available on a 1 km^2 grid scale. The data include daily minimum and maximum air temperature, precipitation, global radiation, mean wind speed, and relative humidity all at 2 m above ground. The daily weather data for the test sites were 99% complete; missing or incomplete records were complemented using the WeatherMan tool in DSSAT (Pickering et al., 1994).

2.3 | Experiment data

In order to calibrate and evaluate the DSSAT-Nwheat for German growing conditions and cultivars, an extensive experimental dataset was prepared covering diverse environments that represent all agro-climatic zones of Germany. For this purpose, relevant experimental datasets were retrieved for three winter wheat cultivars representing the three (Laidig et al., 2017) quality classes: A (quality wheat, cv. Ludwig, released 1998), B (milling wheat, cv. Mulan, released 2006), and C (feed wheat, not usable for baking, cv. Winnetou, released 2002) (Table 1). According to the descriptive variety list of the Bundessortenamt (2007), the yield components of these three cultivars are different (Supporting Information 1). Plant height in Ludwig and Mulan is controlled by the GA-sensitive dwarfing gene RHt8b. Furthermore, Ludwig and Mulan carry the Ppd-D1b allele (Martynov et al., 2024) conferring a sensitive response to the photoperiod. Both cultivars are described with an earlier heading date (Score 4) as compared to Winnetou (Score 5). The dataset includes an N-increase experiment conducted at Kiel University in Northern Germany (54.313° N, 9.981° E; 30 m elevation) during 2008–2010 (Sieling et al., 2016) and a drought-stress experiment conducted at Julius Kühn-Institute, Braunschweig (52.296° N, 10.436° E; 75 m elevation) in Central Germany during 2009–2010 (Schittenhelm et al., 2014). The two datasets comprise time-series data on LAI, specific leaf area, aboveground biomass, and data on grain yield and yield components.

Complementing these detailed in-season datasets, additional cultivar-specific experimental data were retrieved from the value for cultivation use (VCU) trials of the German Federal Plant Variety Office, which are collected as preregistration data of a certain cultivar at a national scale. VCU trials are conducted in all relevant agro-climatic wheat-growing regions with new cultivars before approval and release to the market (Figure 1). Another additional data were from state variety trials (SVT) of Saxony-Anhalt state of Germany, conducted by the state authorities responsible for running post-registration trials for growers' guidance. The VCU and SVT trial dataset comprises cultivar-specific data on phenology, that is, emergence (BBCH 9), ear emergence (BBCH 59), and hard dough (BBCH 87), plus data on yield and yield components, that is, tiller number m^{-2} at harvest, 1000-kernel mass (TKM), and grain per ear at harvest. Depending on the variety, data were available during the period from 1991 until 2019. Detailed overview on the dataset used for calibration and evaluation is shown in Table 1. All datasets entail experiment-specific data on crop management, including previous crop, tillage, sowing date and density, N-fertilizer types, amounts and timing, and harvest date. Both detailed in-season datasets from Kiel and Braunschweig, plus the latest two-thirds of the variety trial data, were used for model calibration. The earliest one-third of variety trial data was retained for model evaluation, based on periodically selecting every fourth site.

2.4 | Calibration tool and strategy

Boote (1999) proposed a systematic methodological procedure for calibrating a CSM for a new crop, cultivar, or dataset utilizing high-quality phenological observation,

TABLE 1 Overview of experimental data available for model calibration and evaluation.

Cultivar	Site-years	Sites	Calibration + evaluation dataset	Site-years extensive data	Site-years LSV-trials[a]	Site-years BSA-trials[b]
				(No.)		
Ludwig	86	54	65 + 21	4	31	51
Mulan	92	47	69 + 23	13	43	46
Winnetou	61	37	50 + 11	12	2	47

[a]Landessortenversuche (State variety trials).
[b]Bundessortenamtsversuche (Federal Plant Variety Office trials).

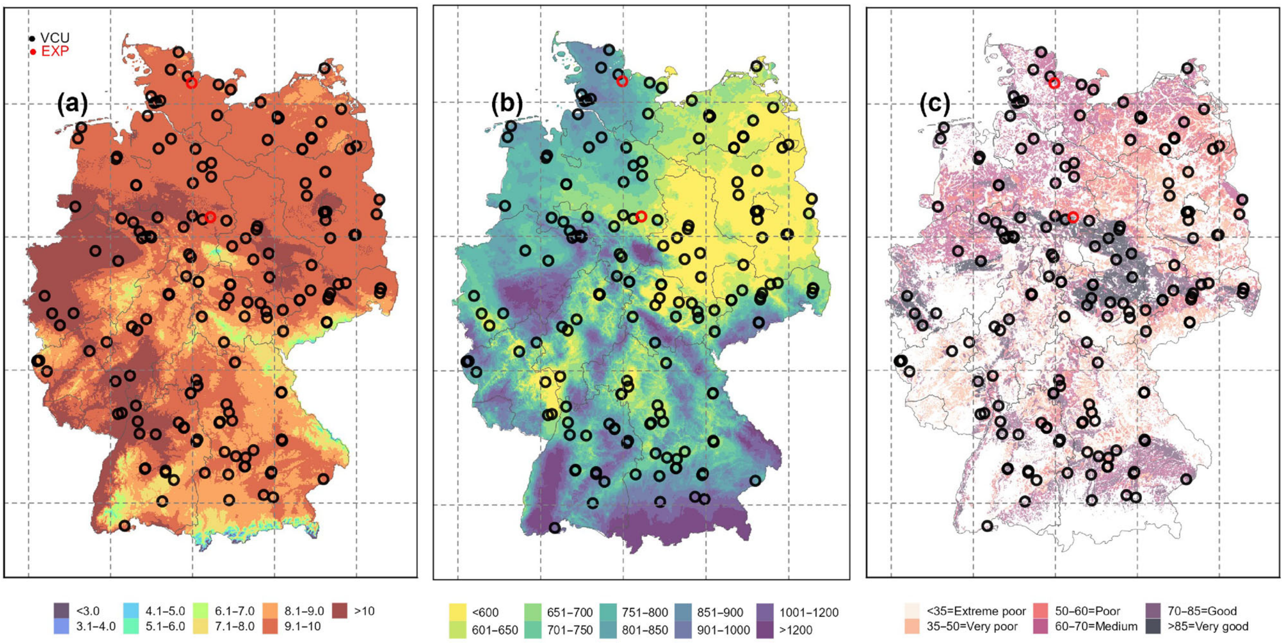

FIGURE 1 Experiment location, value for cultivation use (VCU) from plant variety trials and EXP from extensive phenotypic trials.

in-season growth analysis, and weather and soil data. Crop phenology and growth data are required for optimizing parameters that control phenology and dry matter development in the model, which ensure accurate model performance simulating these variables. In contrast, calibrating the model based only on final yield entails the risk of getting "the correct answer for the wrong reason" since various combinations of different parameters' values may result in similar final yield, while crop growth and dry matter partitioning may differ substantially. Thus, a conceptual approach in model calibration following crop physiological principles as suggested by Boote (1999) that takes into account crop phenological development, biomass growth, and yield formation in a systematic way is likely to reduce the uncertainty of simulation results. However, such comprehensive datasets are not always available, and there might be cases, depending on the objectives of the modeling work, where an accurate prediction of final yield is satisfactory regardless of crop growth. Therefore, the objective was to assess the impact of the availability of specific observation data from experiments during model calibration on final model performance. The dataset included the above-described data on phenology, aboveground biomass, grain yield, and yield components. This dataset is referred as the full dataset. A restricted version was derived by excluding data on aboveground biomass and yield components, that is, TKM, grain per m^2, and grain per ear. This dataset referred as the restricted dataset. One needs to note that the two datasets are identical with regard to the number of site-years, but they differ in the number of variables considered in the calibration.

For calibration, TSE (https://github.com/memicemir/TSE) was utilized, a python-based tool initially developed as a potential external DSSAT plug-in (Memic et al., 2021; Röll et al., 2020). The module can work as a standalone suite accessing externally the functional experiment input file in the DSSAT model, such as crop management, soil, weather, and observed data phenology and growth-related data (included in DSSAT File-A and File-T input files). File-A contains end-season data, such as harvested yield, while File-T includes in-season time-series data, such as LAI, biomass, and phenological development over time. TSE employs a generic algorithm to suggest the best model fit by optimizing selected parameters/coefficients over multi-year experimental data of multiple sites, minimizing normalized root mean square error (nRMSE) between simulated and observed time-series data. It allows the estimation of up to seven cultivar coefficients for up to four optimization variables simultaneously and helps to explore their intermingled effects. The relative weight-nRMSE method implemented in the TSE further allows to assign priority ranks to target variables, to enable prioritizing specific variables during calibration. This ranking approach facilitates the examination of specific coefficients' effects on various CSM processes, such as phenology's impact on final grain yield. By assigning ranks based on relative importance, the calibration process ensures that higher-priority variables (e.g., phenology with rank 1) are modeled with stricter accuracy thresholds, such as an nRMSE below 1%, while lower-priority variables (e.g., grain yield with rank 4) may allow a higher nRMSE, such as 4%. For example, in the calibration strategies, phenology was prioritized to ensure its precise representation in the model, while grain yield is given slightly more flexibility, allowing the calibration to focus on critical processes while maintaining overall model balance.

Accurate yield prediction can often suffice for modeling goals, even if some aspects of crop growth are overlooked. To evaluate the influence of different crop variables on model performance, the dataset was partitioned into two subsets: a full dataset containing phenology, growth, yield components, and final yield, and a restricted dataset containing only phenology and yield data. Both datasets shared the same observations and site-years but differed in the range of included variables. Starting from the default parameter settings (Table 2), specific calibration strategies were applied to each dataset to assess their impacts on model accuracy.

Calibration strategies are as follows:

- Strategy One—Stepwise: This sequential calibration approach focuses on a single target variable at each crop growth stage. Calibration begins with phenology-related parameters, then moves to growth, and finally to yield and its components. At each stage, the most relevant variable is selected for calibration, for instance, LAI during the growth stage or grain yield in the yield stage. This strategy ensures that calibration prioritizes the most critical variable for each developmental phase of the crop, particularly when data availability is limited. The only difference between the restricted (Figure 2a) and full dataset (Figure 2b) approaches lies in the choice of calibration targets during the yield stage: the restricted dataset uses total grain yield, whereas the full dataset includes individual yield components—tiller number m^{-2}, grain number m^{-2}, and TKM.
- Strategy Two—Hybrid: Similar to Strategy One, but two target variables are calibrated simultaneously at each stage. The primary target variable is ranked as 1 (highest priority), while the secondary variable is ranked as 2 (lower priority). For example, during the phenology stage, parameters related to emergence, anthesis, and physiological maturity are prioritized, with phenology as the primary target and yield treated as a secondary consideration. A similar distinction applies between the restricted and full versions: the restricted dataset (Figure 2c) prioritizes phenology as the primary target and yield as secondary, whereas the full dataset (Figure 2d) simultaneously calibrates yield and its components, assigning priority ranks to reflect their relative importance.
- Strategy Three—Altogether: This approach involves calibrating all relevant parameters simultaneously, using up to seven cultivar coefficients and a maximum of four target variables—the operational limits of TSE. In the restricted dataset (Figure 2e), phenology and final grain yield are selected due to limited data availability. In contrast, the full dataset (Figure 2f) allows joint optimization of phenology, grain yield, and key yield components such as grain number m^{-2} and TKM, within the four-variable limit. This comprehensive calibration enables more accurate predictions across multiple crop variables, even under data constraints.

2.5 | Statistical evaluation

The performance of the six DSSAT-Nwheat parameterizations derived from the three different calibration approaches was assessed by comparing the simulated and observed phenological stages, LAI, grain yield, and yield components (only for Strategy Two). For the statistical evaluation, the root mean square error (RMSE; Equation 1), the nRMSE (Equation 2), mean absolute error (MAE; Equation 3), the agreement index (d-index; Equation 4), Nash Sutcliffe efficiency (NSE; Equation 5) were used. The RMSE measures the variance between the simulated (S_i) and observed (O_i) values, providing an indication of how well the model's predictions match the actual data. A smaller RMSE value indicates a better fit, with an ideal RMSE close to 0. The nRMSE normalizes this error by the observed values, expressed as a percentage, offering a relative measure of accuracy. The MAE quantifies the average magnitude of errors in a set of predictions, without considering their direction, while the NSE evaluates how well the model's predictions perform compared to the mean of the observed data. Finally, the d-index measures the model's agreement with the observed data, where a value close to 1.0 indicates a near-perfect match.

These metrics were calculated separately for each site-year and the results were summarized using averages, consistent with approaches used in prior studies (Memic et al., 2021).

TABLE 2 List of considered model coefficient for parameterization.

Parameter	Description	Effect in model Direct	Effect in model Indirect	Range	Based on
VSEN	Vernalization sensitivity	Phenology	Biomass, yield	1–5	Keating et al. (2001)
PPSEN	Photoperiod sensitivity	Phenology	Biomass, yield	1–5	Keating et al. (2001)
P1	Thermal time from emergence to end of juvenile phase	Phenology	Biomass, yield	300–550	Established cultivars
PHINT	Phyllochron interval	Phenology, leaf addition	Biomass, yield	110	Established cultivars
P5	Thermal time (base 0°C) from beginning of grain fill to maturity	Maturity		500–700	Established cultivars
GRNO	Coefficient of kernel number per stem weight at beginning of grain filling (kernels [g stem^{-1}])	Yield		20–35	Established cultivars
MXFIL	Potential kernel growth rate (mg kernel^{-1} day^{-1})	Grain filling/grain yield		1–3	Cultivar file documentation
STMMX	Potential final dry mass of single tiller (excluding grain) (g stem^{-1})	Biomass growth/yield		1–3	Cultivar file documentation
SLAP1	Leaf area to mass ratio at emergence (cm^2 g^{-1})	Slope of specific leaf area		200–300	Max range of established cultivars
SLAP2	Leaf area to mass ratio at end of leaf growth (cm^2 g^{-1})	Leaf area index and biomass		200–280	Max range of established cultivars

While some indices, like NSE, can yield negative values that complicate averaging, others (e.g., RMSE, nRMSE, d-index) are more suitable for this approach due to their linear and bounded nature. Direct calculation of indices across all experiments was not possible, as it requires aggregating datasets, which would alter the interpretation of the metrics.

Root mean square error single model (RMSE):

$$\mathrm{RMSE_m} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(S_i - O_i\right)^2\right]^{0.5} \tag{1}$$

Normalized root mean square error (nRMSE):

$$\mathrm{nRMSE_n} = \frac{\mathrm{RMSE_m}}{\mathrm{O}}100 \tag{2}$$

Mean absolute error (MAE):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\lfloor S_i - O_i\rfloor}{\mathrm{n}} \tag{3}$$

Index of agreement single model (d-index):

$$\mathrm{d-index} = 1 - \left[\frac{\sum_{i=1}^{n}\left(S_i - O_i\right)^2}{\sum_{i=1}^{n}\left(\left|S_i - O\right| + \left|O_i - O\right|\right)^2}\right] \tag{4}$$

Nash Sutcliffe efficiency (NSE):

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{m}\left(S_i - O_i\right)^2}{\sum_{i=1}^{m}\left(O_i - O\right)^2} \tag{5}$$

2.6 | Model application for uncertainty analysis

To analyze the uncertainty in future grain yield predictions across three calibration strategies and two datasets, a model application scenario was developed based on two point-based locations in Germany (Appendix 2). These locations were selected based on their distinct differences in annual precipitation and soil quality, measured by the percentage of sand in the topsoil (0–60 cm). The first location, Thyrow, is situated in the southwest of Brandenburg, Germany, with an average annual precipitation of 552 mm and a topsoil sand content of 75.5%. The second location, Feldkirchen, lies in eastern Bavaria, with higher annual precipitation of 756 mm and a sand content of 13.1%. For the climate scenarios, 12 distinct scenarios were selected (Appendix 1) from the DWD core ensemble data at a 5 km × 5 km resolution (www.dwd.de), in combination with two representative concentration pathways (RCP 4.5 and RCP 8.5). The simulations were conducted for

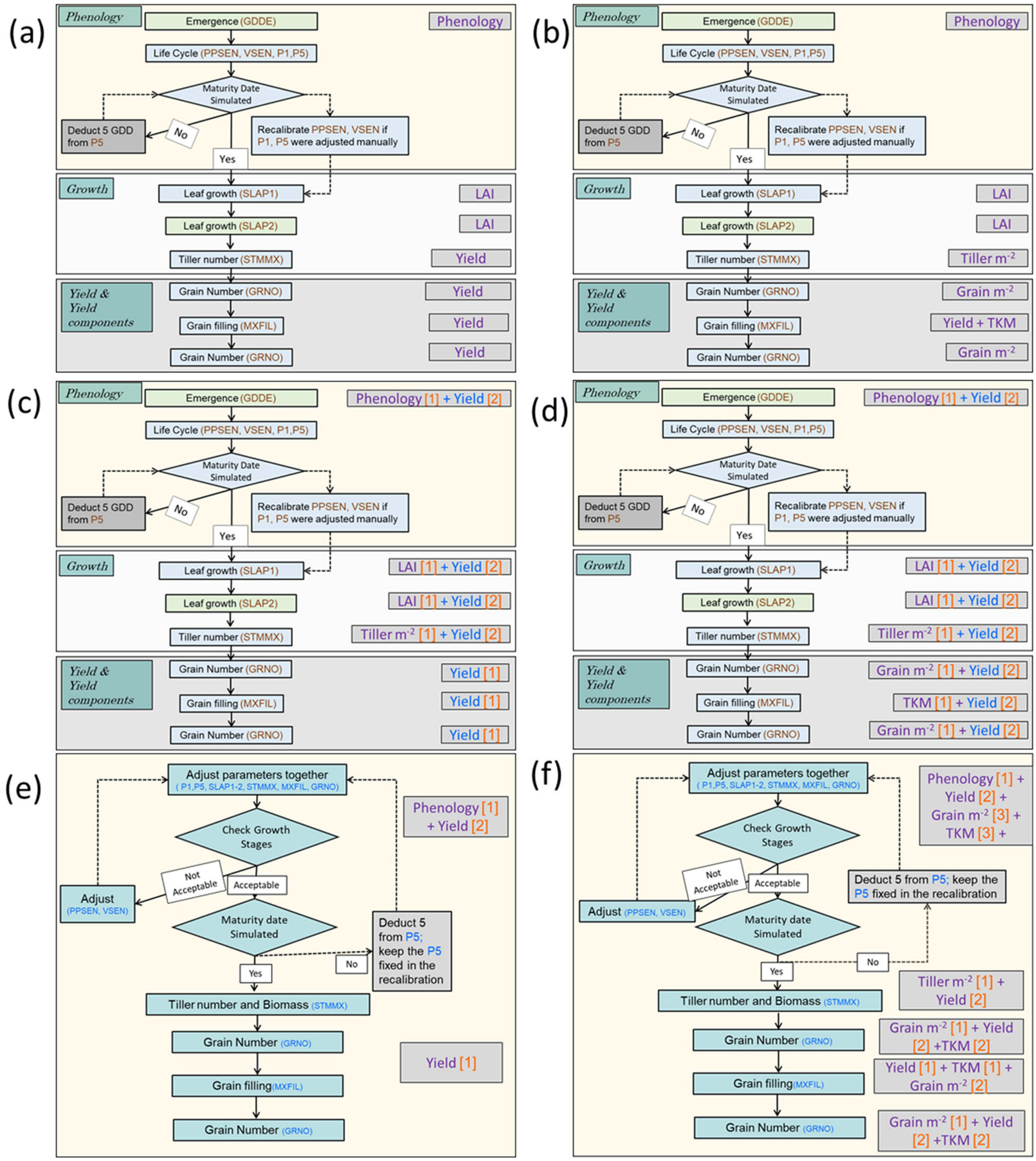

FIGURE 2 Formulated strategies for model parameterization based on selection of variables for error reductions, subparts (a), (c), and (e) correspond to strategies One, Two, and Three, respectively, from the restricted dataset, while (b), (d), and (f) correspond to strategies One, Two, and Three, respectively, from the full dataset. LAI, leaf area index; TKM, 1000-kernal mass.

120 years, allowing us to assess trends over time. The historical baseline period was defined as 1981–2023, reflecting the availability of observed climate data and consistency with common reference periods. The future projection period was defined as 2024–2099, in line with scenario-based climate modeling frameworks.

2.7 | Grain yield trend analysis with linear mixed model

To evaluate the impact of the calibration strategy, dataset, locations, and climate scenarios on grain yield prediction, a generalized linear mixed model was utilized based on the glmmTMB package in R (Brooks et al., 2017),

$$Y_{ijkl} = \mu + C_i + D_j + L_k + R_l + (CD)_{ij} + (CL)_{ij} + (CR)_{ik} + (DL)_{jk} + (DR)_{jl} + (LR)_{kl} + (CDL)_{ijk} + \epsilon_{ijkl} \quad (6)$$

where Y_{ijkl} is the predicted grain yield for the ith calibration strategy, the jth dataset, the kth location, and the lth climate scenario. μ is the overall mean, C_i is the main effect of the ith calibration strategy, D_j is the main effect of the jth dataset, L_k is the main effect of the kth location, and R_l is the main effect of the lth climate scenario. The effects $(CD)_{ij}$, $(CL)_{ik}$, $(CR)_{il}$, $(DL)_{jk}$, $(DR)_{jl}$, $(LR)_{kl}$, and $(CDL)_{ijk}$ are the interaction effects between the corresponding main effects. ϵ_{ijkl} is the residual error term, assumed to be independent and normally distributed with constant variance.

All effects except μ are treated as fixed effects. Additionally, a temporal autocorrelation of order AR(1) was included to account for dependencies in the data over time. This model setup allows for the investigation of how each factor and their interactions influence grain yield predictions under different environmental and climate conditions.

3 | RESULTS

3.1 | Effects of calibration strategy

3.1.1 | Calibration strategy effect on phenology

The default coefficient settings for simulating wheat phenology generally overestimated developmental stages, with RMSE values ranging from 8.7 to 10.2 days across the selected cultivars (see Figure 3). Application of the calibration strategies led to a reduction in RMSE values, indicating improved agreement between simulated and observed phenology, with reductions of up to 10% depending on the cultivar. For example, in the Ludwig cultivar, the initial RMSE of 10.2 days was reduced to approximately 8.1 days after calibration (Tables 3–5).

These improvements in model accuracy were accompanied by distinct adjustments to key phenological parameters across strategies and cultivars. The calibration strategies led to varied adjustments in key phenological coefficients such as vernalization sensitivity (VSEN), photoperiod sensitivity (PPSEN), and thermal time (P1 and P5). For instance, in the cultivar Ludwig, Strategy One maintained VSEN at 1.50 across both datasets, while Strategy Three adjusted it to 2.00 in the full dataset and 1.25 in the restricted dataset, indicating differences in vernalization requirements based on the dataset used. In the Mulan cultivar, Strategy Two increased VSEN to 4.25, compared to 3.75 in Strategy One, suggesting a stronger vernalization response in Strategy Two (Table 6).

These parameter adjustments were reflected in model performance metrics, particularly RMSE, across the three strategies. Across all cultivars, Strategy One generally resulted in lower RMSE values compared to Strategies Two and Three (Tables 3–5). For example, in cultivar Ludwig, using the full dataset, Strategy One achieved an RMSE of 4.5 during calibration and 5.4 during validation, with similar trends observed for Mulan and Winnetou. In terms of PPSEN, Mulan retained a PPSEN of 3.00 in Strategy One for both datasets, while Strategy Three lowered it to 2.6 in the full dataset and 2.33 in the restricted dataset, reflecting a reduced sensitivity to day length. For Winnetou, Strategy One used a PPSEN of 3.85, whereas Strategy Two decreased it to 1.75, reflecting differing assumptions regarding the crop's photoperiod response (Table 7).

In addition to VSEN and PPSEN, adjustments in thermal time parameters (P1 and P5) further contributed to the observed differences among strategies. The dataset influenced strategy performance differently across the cultivars. Thermal time (P1 and P5) adjustments also varied among strategies. In cultivar Ludwig, Strategy One increased P1 from the default 400 to 431, while Strategy Two further adjusted it to 437. Whereas for cultivar Mulan, Strategy Three raised P1 to 500 in the restricted dataset, suggesting the need for a longer developmental phase. In contrast, Winnetou, which does not carry the Ppd-D1b allele, showed smaller adjustments in P1 across all strategies, reflecting its photoperiod insensitivity.

Variations in P5 values followed similar trends and further illustrate cultivar-specific responses. In cultivar Ludwig, Strategy One kept P5 at 710, while Strategy Two reduced it to 688, possibly reflecting a shorter grain-filling period. Mulan's Strategy Three raised P5 to 670 in the restricted dataset, while Strategy One held it constant at 670, reflecting a more consistent grain-filling period with Strategy One. Winnetou demonstrated relatively stable P5 adjustments across strategies, aligning with its different genetic background.

Overall, Strategy One consistently produced more accurate phenology simulations, as seen in the RMSE values for BBCH stages. However, when additional variables such as grain weight were included in the optimization process (as in Strategy Two), the accuracy of the phenology simulations sometimes decreased, likely due to the trade-off required to optimize multiple components simultaneously (as shown in Tables 3–5).

3.1.2 | Calibration strategy effect on biomass and tillers m^{-2} at harvest

Before calibration, the default parameter settings were expected to result in discrepancies between simulated and observed values, as they do not account for cultivar-specific differences. For example, in cultivar Ludwig, NSE values for tillers m^{-2}, LAI, and aboveground dry mass (AGDM) were −1.6, −0.433, and −0.477, respectively, with nRMSE values of 160.2, 119, and 112.5. Similarly, the cultivar Mulan showed an NSE of −1.33 for tillers m^{-2} and an nRMSE of 151.7. While the cultivar Winnetou had particularly large errors for tillers m^{-2}, with an NSE of −28.1 and an nRMSE of 535.

Following calibration, model performance improved substantially, as adjustments to key parameters aligned the simulations more closely with observed data. These improvements were particularly evident in variables such as tiller density and aboveground biomass. For instance, in cultivar Ludwig, the nRMSE for tillers m^{-2} was reduced by 29.2% (from 160.2 to 113.5), and the NSE improved from −1.6 to 0.50. Similar improvements were observed in Mulan, where the nRMSE for AGDM decreased by 62.3% (from 137 to 51.6), and in Winnetou, where the nRMSE for tillers m^{-2} dropped by 62.2% (from 535 to 202.1).

These gains in simulation accuracy were primarily driven by targeted adjustments to phenology and biomass-related parameters. For example, in cultivar Ludwig, Strategy Three increased P1 (thermal time from sowing to end of juvenile phase) to 475 in the full dataset, compared to 431 in Strategy One, which contributed to greater vegetative growth and biomass accumulation. Conversely, P5 (thermal time from anthesis to maturity) was reduced to 650 in Strategy Three, compared to 710 in Strategy One, indicating a shorter grain-filling period. Similarly, Strategy Three adjusted biomass-specific coefficients, such as SLAP1 and SLAP2 (leaf area to mass ratios), to 220 and 205, respectively, compared to default values of 230 and 240, thereby optimizing light interception and biomass production.

Similar patterns were observed for other cultivars, where Strategy Three systematically refined key growth parameters to improve biomass-related outputs. In Mulan, Strategy Three's adjustments to STMMX (potential final dry mass of a single tiller) and SLAP1/SLAP2 improved alignment between simulated and observed biomass and tiller counts. STMMX was reduced to 1.86, and SLAP1/SLAP2 were optimized to 220 and 260, respectively, ensuring better simulation of biomass accumulation and distribution. Winnetou showed similar trends, with Strategy Three refining SLAP1 and SLAP2 to 230 and 270, while also recalibrating P1 and P5 to better match the cultivar's unique developmental and biomass allocation patterns. By tailoring these parameters to the specific characteristics of each cultivar, the calibration process effectively reduced errors and improved the model's ability to simulate biomass and tillers m^{-2} across varying environmental conditions.

3.1.3 | Calibration strategy effects on yield and yield components

The analysis reveals that Strategy One consistently performed as the most accurate calibration method for simulating grain yield, grain number m^{-2}, and TKM across the three cultivars (Ludwig, Mulan, and Winnetou). In Ludwig, Strategy One achieved an nRMSE of 134.8 for grain yield, 173.7 for grain m^{-2}, and 152.4 for TKM, outperforming both Strategies Two and Three (Table 3). Compared to Strategy Two, Strategy One reduced nRMSE by 2.44% for grain yield, 5.36% for grain m^{-2}, and 2.16% for TKM. Similarly, compared to Strategy Three, Strategy One showed greater improvements, with reductions of 15.06% for grain yield, 18.53% for grain m^{-2}, and 0.78% for TKM.

This performance advantage of Strategy One was also evident in Mulan. Strategy One yielded an nRMSE of 127.1 for grain yield, 140.6 for grain m^{-2}, and 130.5 for TKM, surpassing both Strategies Two and Three (Table 4). Strategy One reduced nRMSE for grain yield by 20.04% and for TKM by 11.42% compared to Strategy Two, and it also improved by 9.72% for grain yield and 8.45% for TKM when compared to Strategy Three.

A similar trend was observed in Winnetou, and Strategy One continued to deliver the most accurate results, with an nRMSE of 137.35 for grain yield, 186.45 for grain m^{-2}, and 119.1 for TKM (Table 5). The reductions in nRMSE were 10.17% for grain yield, 2.03% for grain m^{-2}, and 4.77% for TKM when compared to Strategy Two, and 9.16%, 3.28%, and 8.92%, respectively, when compared to Strategy Three. Strategy One's superiority is also evident in simulations of grain yield at harvest, where the restricted dataset consistently showed the lowest RMSE among all strategies. Similarly, the difference between simulated and observed means in Strategy One from the restricted dataset was about 6%, the smallest difference across all strategies.

These differences in performance among strategies can be traced to key parameter settings related to yield formation. MXFIL, which governs grain growth rate, and GRNO, controlling grain number per unit stem weight, play crucial roles in determining yield outcomes. In Ludwig, Strategy One set MXFIL at 1.57, while Strategy Three increased it to 1.88, which may have resulted in an overestimation of grain filling under Strategy Three. Similarly, GRNO was set at 23.61 in Strategy One and adjusted to 23.3 in Strategy Three, reflecting slight variations in grain number calibration. P5, which influences the duration of grain fill, was set at 710 in Strategy One and reduced to 650 in Strategy Three, potentially contributing to differences in the accuracy of yield predictions.

Despite improvements through calibration, a tendency to underestimate grain yield remained across all strategies.

TABLE 3 Performance analysis of model calibration strategies for cultivar Ludwig in crop modeling: A comparison of full and restricted dataset approaches.

Database	Strategy	Cal/Eval	Variables	No obs	RMSE	nRMSE	MAE	d-index	Pearson correlation	NSE
Full dataset with TKM-tiller m^{-2}-grain m^{-2}	One	Cal	BBCH	312	4.5	15.3	3.3	0.99	0.99	0.98
		Eval	BBCH	116	5.4	20.1	4.0	0.99	0.98	0.96
		Cal	AGDM	15	3966.3	58.0	1972.4	0.90	0.82	0.64
		Cal	LAI	96	1.4	113.7	1.1	0.70	0.56	−0.31
		Cal	Tillers m^{-2}	149	147.4	136.7	71.5	0.47	0.22	−0.88
		Eval	Tillers m^{-2}	59	596.2	580.8	166.7	0.05	−0.01	−33.32
		Cal	Grain m^{-2}	68	3815.1	168.7	2616.2	0.39	0.12	−1.89
		Eval	Grain m^{-2}	20	1831.4	178.7	1476.7	0.35	0.14	−2.36
		Cal	TKM	68	10.6	201.7	7.9	0.36	0.10	−3.13
		Eval	TKM	20	3.4	103.0	3.0	0.66	0.54	−0.12
		Cal	Grain yield	63	2219.4	160.8	1627.1	0.53	0.25	−1.63
		Eval	Grain yield	21	885.2	108.7	666.4	0.51	0.19	−0.24
	Two	Cal	BBCH	312	4.5	15.5	3.4	0.99	0.99	0.98
		Eval	BBCH	116	5.1	18.9	3.7	0.99	0.98	0.96
		Cal	AGDM	15	3908.2	57.1	2003.7	0.90	0.82	0.65
		Cal	LAI	96	1.2	92.9	0.9	0.74	0.56	0.13
		Cal	Tillers m^{-2}	149	146.0	135.3	73.2	0.44	0.14	−0.84
		Eval	Tillers m^{-2}	59	121.5	118.3	66.4	0.50	0.20	−0.42
		Cal	Grain m^{-2}	68	3842.1	169.9	2661.6	0.39	0.13	−1.93
		Eval	Grain m^{-2}	20	2009.5	196.1	1605.9	0.37	0.21	−3.05
		Cal	TKM	68	11.3	213.8	8.2	0.35	0.11	−3.64
		Eval	TKM	20	3.1	93.7	2.6	0.72	0.52	0.08
		Cal	Grain yield	63	2328.6	168.7	1683.3	0.50	0.24	−1.89
		Eval	Grain yield	21	1171.6	143.9	918.0	0.47	0.13	−1.18
	Three	Cal	BBCH	312	4.7	16.2	3.4	0.99	0.99	0.97
		Eval	BBCH	116	6.0	22.5	4.4	0.99	0.98	0.95
		Cal	AGDM	15	3889.9	56.9	1974.5	0.90	0.82	0.65
		Cal	LAI	96	1.2	93.1	0.9	0.75	0.57	0.12
		Cal	Tillers m^{-2}	149	122.4	113.5	65.6	0.50	0.24	−0.30
		Eval	Tillers m^{-2}	59	161.9	157.7	79.3	0.48	0.25	−1.53
		Cal	Grain m^{-2}	68	3904.5	172.6	2712.8	0.38	0.12	−2.02
		Eval	Grain m^{-2}	20	2076.1	202.6	1754.8	0.34	0.23	−3.32
		Cal	TKM	68	11.4	215.6	8.2	0.31	0.04	−3.72
		Eval	TKM	20	2.9	86.3	2.5	0.74	0.55	0.22
		Cal	Grain yield	63	2439.9	176.7	1742.9	0.48	0.22	−2.17
		Eval	Grain yield	21	1151.9	141.5	897.6	0.49	0.20	−1.10
Restricted dataset without TKM-tiller m^{-2}-grain m^{-2}	One	Cal	BBCH	312	4.5	15.3	3.3	0.99	0.99	0.98
		Eval	BBCH	116	5.4	20.1	4.0	0.99	0.98	0.96
		Cal	AGDM	15	4085.4	59.7	2061.4	0.90	0.82	0.62
		Cal	LAI	96	1.8	147.3	1.5	0.61	0.53	−1.19
		Cal	Tillers m^{-2}	149	156.5	145.1	73.6	0.45	0.21	−1.12
		Eval	Tillers m^{-2}	59	623.0	606.9	174.7	0.05	−0.01	−36.46

(Continues)

TABLE 3 (Continued)

Database	Strategy	Cal/Eval	Variables	No obs	RMSE	nRMSE	MAE	d-index	Pearson correlation	NSE
		Cal	Grain m^{-2}	68	4044.0	178.8	2892.2	0.41	0.12	−2.24
		Eval	Grain m^{-2}	20	1563.8	152.6	1221.9	0.31	0.11	−1.45
		Cal	TKM	68	11.5	218.1	8.6	0.35	0.11	−3.83
		Eval	TKM	20	3.4	101.0	2.9	0.71	0.56	−0.07
		Cal	Grain yield	63	2192.6	158.8	1596.2	0.54	0.26	−1.56
		Eval	Grain yield	21	955.5	117.4	716.9	0.48	0.13	−0.45
	Two	Cal	BBCH	312	4.5	15.5	3.4	0.99	0.99	0.98
		Eval	BBCH	116	5.1	18.9	3.7	0.99	0.98	0.96
		Cal	AGDM	15	4025.8	58.8	2020.8	0.90	0.82	0.63
		Cal	LAI	96	1.4	110.7	1.1	0.69	0.53	−0.24
		Cal	Tillers m^{-2}	149	168.2	155.9	80.0	0.41	0.13	−1.45
		Eval	Tillers m^{-2}	59	143.6	139.9	76.0	0.45	0.17	−0.99
		Cal	Grain m^{-2}	68	4658.6	206.0	3679.6	0.41	0.13	−3.31
		Eval	Grain m^{-2}	20	2183.0	213.1	1877.4	0.35	0.16	−3.78
		Cal	TKM	68	12.8	242.4	9.8	0.33	0.07	−4.97
		Eval	TKM	20	4.9	146.9	4.0	0.57	0.44	−1.27
		Cal	Grain yield	63	2262.4	163.9	1647.1	0.53	0.25	−1.73
		Eval	Grain yield	21	839.1	103.1	653.6	0.52	0.23	−0.12
	Three	Cal	BBCH	312	4.6	15.9	3.4	0.99	0.99	0.97
		Eval	BBCH	116	6.7	25.2	5.0	0.98	0.98	0.94
		Cal	AGDM	15	4062.1	59.4	2013.2	0.90	0.82	0.62
		Cal	LAI	96	1.9	151.6	1.6	0.61	0.52	−1.32
		Cal	Tillers m^{-2}	149	409.9	380.0	133.0	0.16	0.09	−13.54
		Eval	Tillers m^{-2}	59	187.0	182.1	85.5	0.45	0.27	−2.37
		Cal	Grain m^{-2}	68	3883.3	171.7	2654.9	0.38	0.13	−1.99
		Eval	Grain m^{-2}	20	1941.6	189.5	1530.7	0.35	0.18	−2.78
		Cal	TKM	68	10.1	191.9	7.3	0.36	0.11	−2.74
		Eval	TKM	20	3.0	91.0	2.4	0.68	0.50	0.13
		Cal	Grain yield	63	2186.8	158.4	1595.1	0.53	0.28	−1.55
		Eval	Grain yield	21	1044.7	128.4	820.7	0.47	0.14	−0.73

Abbreviations: MAE, mean absolute error; nRMSE, normalized root mean square error; NSE, Nash Sutcliffe efficiency; RMSE, normalized root mean square error; TKM, 1000-kernal mass.

Simulated grain yield was generally underestimated, with values ranging between 525.4 and 1242.2 kg ha^{-1}. The correlation between observed and simulated yield values was weak (ranging from 0.11 to 0.15), as noted in Figure 3, which aligns with prior findings that highlight the model's insensitivity to certain regression coefficients (Yang et al., 2014).

3.2 | Effect of database

3.2.1 | Database effect on phenology

The choice of database did not significantly influence the accuracy of phenology simulations across the three calibration strategies. However, differences among strategies remained apparent regardless of the dataset used. When comparing the impact of the database on different cultivars, it is clear that Strategy One consistently produced the most accurate alignment between simulated and observed phenological stages (BBCH). For example, in the Ludwig cultivar, the RMSE values during the calibration phase remained stable, ranging between 4.5 and 5.4 days, demonstrating reliable simulation performance. Similarly, in the Mulan and Winnetou cultivars, Strategy One yielded RMSE values between 3.4 and 5.5 days and 3.1 and 4.8 days, respectively, confirming its robustness across datasets. In contrast, Strategy Three generally resulted in slightly higher RMSE values, indicating less precision in simulating phenological events, although this

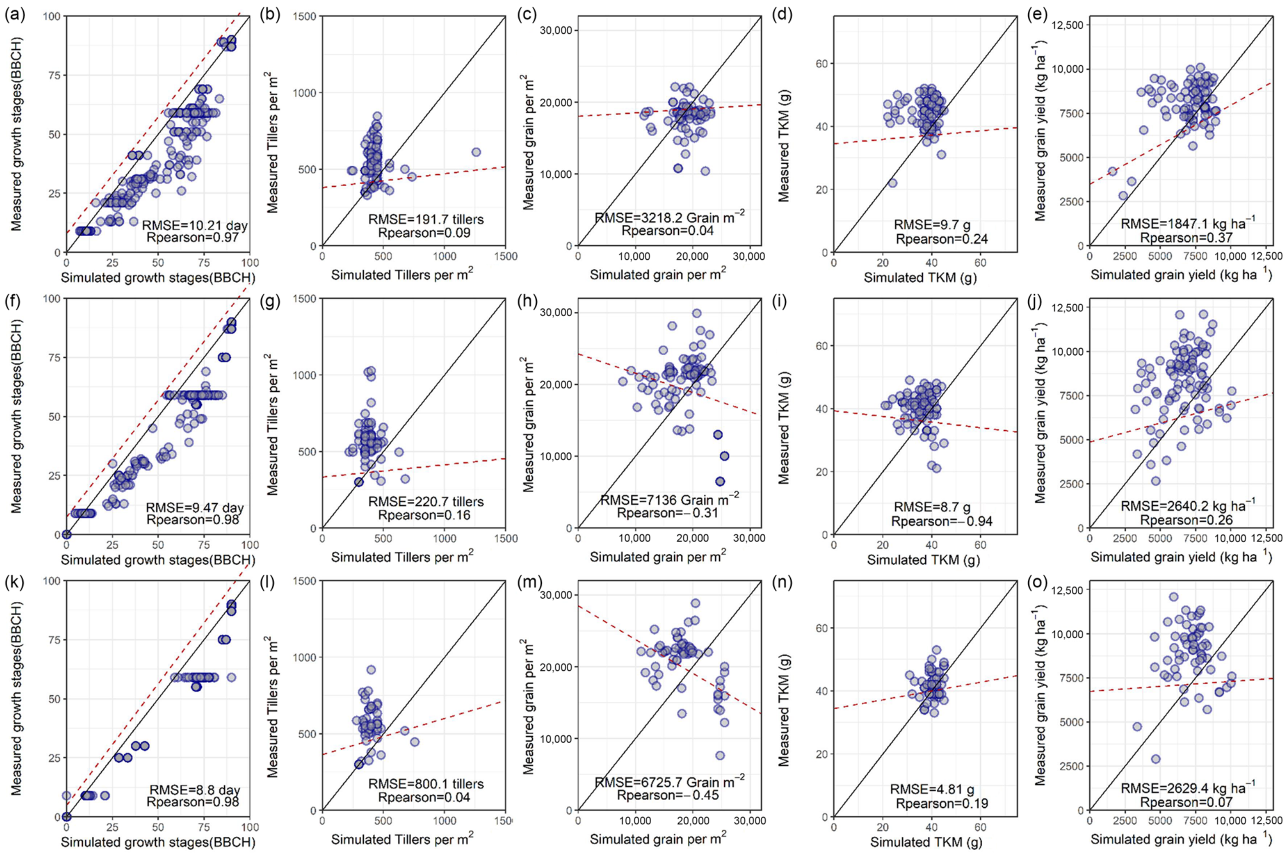

FIGURE 3 Comparative analysis of simulated and observed data for growth stages (BBCH), tillers m^{-2} at harvest, grain m^{-2} at harvest, 1000-kernel mass (TKM), and grain yield (kg ha^{-1}) among three selected wheat cultivars using default settings (before calibration). The data are organized from left to right for three different cultivar settings: Ludwig default (‘a,’ ‘b,’ ‘c,’ ‘d,’ ‘e’), Mulan default (‘f,’ ‘g,’ ‘h,’ ‘i,’ ‘j’), and Winnetou default (‘k,’ ‘l,’ ‘m,’ ‘n,’ ‘o’).

variation was more attributable to calibration strategy than the database itself.

3.2.2 | Database effect on biomass and tiller m^{-2}

Considering the full dataset has improved biomass simulation overall accuracy by 3.5%, 0.3%, and 0.5%, respectively, for Ludwig, Mulan, and Winnetou in terms of RMSE. With the restricted dataset, the model underestimated by 78.7 kg ha^{-1}, whereas with the full dataset, there is an overestimation of 107 kg ha^{-1} for the cultivar Ludwig. In Mulan, overestimation increased by 15.1 kg, whereas in Winnetou, overestimation decreased by 12.4 kg ha^{-1}. Inclusion or exclusion of growth and yield component data does not show any effect on the final biomass at the harvest, assuming adjusting light interception will not improve carbon assimilation after a certain point.

The inclusion of growth data, however, substantially improved LAI simulations across all three cultivars. RMSE was decreased by 26.8%, 6.2%, and 9.4% for Ludwig, Mulan, and Winnetou, respectively. In terms of other statistical indices, like the d-index, a similar Pearson correlation trend was visible.

Similarly, adding growth and yield component data significantly enhanced the accuracy of tiller m^{-2} simulations. Across the reference three cultivars, improvements were observed in all strategies: Strategy One improved by 1.41%, Strategy Two improved by 3.40%, and Strategy Three by 49.14%, highlighting the particularly large benefit of additional growth data for Strategy Three.

3.2.3 | Database effect on yield and yield components

In cultivars Ludwig, TKM was improved slightly by 0.7, 1.5, and 1.5 mg seed^{-1} for Strategies One, Two, and Three, respectively. Tiller m^{-2} by 16.1, 22.1, and 22.1 by the same strategies, respectively. For all three strategies, grain no RMSE was improved by 166.8, 744.7, and 744.7, respectively. Overall grain yield RMSE was decreased by 14.4–96.3 kg ha^{-1} for all three strategies.

TABLE 4 Performance analysis of model calibration strategies for cultivar Mulan in crop modeling: A comparison of full and restricted dataset approaches.

Database	Strategy	Cal/Eval	Variables	No obs	RMSE	nRMSE	MAE	d-index	Pearson correlation	NSE
Full dataset with TKM-tiller m^{-2}-grain m^{-2}	One	Cal	BBCH	383	3.4	10.2	2.3	1.00	1.00	0.99
		Eval	BBCH	132	5.5	19.9	4.2	0.99	0.99	0.96
		Cal	AGDM	80	4367.7	53.6	2260.1	0.94	0.90	0.71
		Cal	LAI	504	2.9	220.8	2.3	0.52	0.55	−3.89
		Cal	Tillers m^{-2}	132	177.5	134.8	88.6	0.56	0.34	−0.83
		EEval	Tillers m^{-2}	63	159.0	97.7	87.5	0.67	0.49	0.03
		Cal	Grain m^{-2}	106	6462.6	144.6	5181.4	0.29	−0.21	−1.11
		Eval	Grain m^{-2}	24	3900.4	136.6	3193.2	0.58	0.52	−0.95
		Cal	TKM	106	8.2	159.4	6.2	0.32	−0.03	−1.56
		Eval	TKM	24	3.9	101.5	3.0	0.49	0.30	−0.07
		Cal	Grain yield	70	2639.6	137.5	2262.0	0.36	−0.02	−0.92
		Eval	Grain yield	23	1862.8	116.6	1645.6	0.68	0.71	−0.42
	Two	Cal	BBCH	377	3.7	10.9	2.6	1.00	1.00	0.99
		Eval	BBCH	132	5.1	18.3	4.0	0.99	0.99	0.97
		Cal	AGDM	80	4254.4	52.2	2197.9	0.94	0.90	0.72
		Cal	LAI	504	3.3	250.7	2.6	0.49	0.54	−5.30
		Cal	Tillers m^{-2}	130	174.7	132.6	76.6	0.63	0.47	−0.77
		Eval	Tillers m^{-2}	63	369.6	227.1	147.0	0.33	0.20	−4.24
		Cal	Grain m^{-2}	105	6035.5	135.4	5226.8	0.22	−0.22	−0.85
		Eval	Grain m^{-2}	24	5080.0	177.9	4538.3	0.50	0.53	−2.30
		Cal	TKM	105	8.1	158.5	6.1	0.32	−0.04	−1.54
		Eval	TKM	24	4.6	120.4	3.8	0.48	0.31	−0.51
		Cal	Grain yield	69	2965.7	155.3	2542.6	0.39	−0.04	−1.45
		Eval	Grain yield	23	2609.9	163.4	2418.0	0.58	0.71	−1.79
	Three	Cal	BBCH	377	3.3	9.8	2.4	1.00	1.00	0.99
		Eval	BBCH	132	6.1	22.0	4.6	0.99	0.99	0.95
		Cal	AGDM	80	4204.3	51.6	2176.8	0.94	0.90	0.73
		Cal	LAI	504	2.6	201.0	2.1	0.55	0.54	−3.05
		Cal	Tillers m^{-2}	130	189.4	143.8	73.6	0.58	0.42	−1.08
		Eval	Tillers m^{-2}	63	175.9	108.1	94.3	0.48	0.22	−0.19
		Cal	Grain m^{-2}	105	6150.8	138.0	5208.7	0.26	−0.22	−0.92
		Eval	Grain m^{-2}	24	4724.4	165.4	4230.6	0.51	0.51	−1.85
		Cal	TKM	105	8.0	156.6	5.8	0.34	−0.02	−1.48
		Eval	TKM	24	4.7	121.5	4.1	0.56	0.46	−0.54
		Cal	Grain yield	69	2900.0	151.9	2514.2	0.38	−0.03	−1.34
		Eval	Grain yield	23	2599.5	162.7	2413.7	0.57	0.70	−1.77
Restricted dataset without TKM-tiller m^{-2}-grain m^{-2}	One	Cal	BBCH	383	3.4	10.2	2.3	1.00	1.00	0.99
		Eval	BBCH	132	5.5	19.9	4.2	0.99	0.99	0.96
		Cal	AGDM	80	4336.2	53.3	2244.8	0.94	0.90	0.71
		Cal	LAI	504	2.9	220.9	2.3	0.52	0.55	−3.89
		Cal	Tillers m^{-2}	132	177.5	134.8	88.6	0.56	0.34	−0.83
		Eval	Tillers m^{-2}	63	159.0	97.7	87.5	0.67	0.49	0.03
		Cal	Grain m^{-2}	106	8902.8	199.3	6702.1	0.32	−0.21	−3.01
		Eval	Grain m^{-2}	24	3320.4	116.3	2492.1	0.66	0.52	−0.41

(Continues)

TABLE 4 (Continued)

Database	Strategy	Cal/Eval	Variables	No obs	RMSE	nRMSE	MAE	d-index	Pearson correlation	NSE
		Cal	TKM	106	10.6	206.3	8.5	0.36	−0.04	−3.30
		Eval	TKM	24	8.7	227.1	8.3	0.41	0.44	−4.38
		Cal	Grain yield	70	2584.4	134.6	2221.7	0.39	0.03	−0.84
		Eval	Grain yield	23	1856.1	116.2	1645.4	0.69	0.71	−0.41
	Two	Cal	BBCH	377	3.7	10.9	2.6	1.00	1.00	0.99
		Eval	BBCH	132	5.1	18.3	4.0	0.99	0.99	0.97
		Cal	AGDM	80	4243.9	52.1	2193.1	0.94	0.90	0.72
		Cal	LAI	504	3.3	250.7	2.6	0.49	0.54	−5.30
		Cal	Tillers m^{-2}	130	149.7	113.6	83.4	0.61	0.39	−0.30
		Eval	Tillers m^{-2}	63	257.8	158.4	126.5	0.41	0.15	−1.55
		Cal	Grain m^{-2}	105	11,687.9	262.3	9377.5	0.30	−0.22	−5.95
		Eval	Grain m^{-2}	24	5694.6	199.4	5045.3	0.49	0.53	−3.15
		Cal	TKM	105	15.3	298.1	14.1	0.32	−0.05	−7.97
		Eval	TKM	24	15.4	399.5	14.9	0.28	0.31	−15.65
		Cal	Grain yield	69	2798.2	146.5	2415.1	0.40	0.01	−1.18
		Eval	Grain yield	23	2395.9	150.0	2197.7	0.61	0.71	−1.35
	Three	Cal	BBCH	377	3.7	10.8	2.4	1.00	0.99	0.99
		Eval	BBCH	132	4.4	15.9	3.4	0.99	0.99	0.97
		Cal	AGDM	80	4283.8	52.6	2218.5	0.94	0.90	0.72
		Cal	LAI	504	3.2	245.2	2.6	0.50	0.59	−5.03
		Cal	Tillers m^{-2}	130	782.3	593.8	161.2	0.15	0.25	−34.53
		Eval	Tillers m^{-2}	63	245.5	150.9	116.1	0.42	0.18	−1.31
		Cal	Grain m^{-2}	105	6297.3	141.3	5529.8	0.25	−0.25	−1.02
		Eval	Grain m^{-2}	24	7030.8	246.2	6556.3	0.43	0.57	−5.32
		Cal	TKM	105	12.0	233.3	10.6	0.30	−0.12	−4.49
		Eval	TKM	24	9.0	233.5	7.6	0.39	0.07	−4.69
		Cal	Grain yield	69	2980.6	156.1	2509.5	0.32	−0.11	−1.47
		Eval	Grain yield	23	2038.4	127.6	1815.2	0.66	0.72	−0.70

Abbreviations: MAE, mean absolute error; nRMSE, normalized root mean square error; NSE, Nash Sutcliffe efficiency; RMSE, normalized root mean square error; TKM, 1000-kernal mass.

In contrast, in the pervious cultivar, grain yield RMSE was increased by 45.6, 177.7, and 177.7 kg ha^{-1} for the three strategies, respectively. However, grain m^{-2} improved significantly from 2093 to 4957.5, with TKM ranging from 2.7 to 7.7 mg $seed^{-1}$ and tiller m^{-2} at harvest of 63.4.

In the case of Winnetou, grain yield for Strategy One improved by 34.9 kg ha^{-1} with the effect of the database. However, this was accompanied by greater error in grain m^{-2} for Strategy One, while Strategies Two and Three showed improvement of 78.4 grain m^{-2}. TKM was increased similarly to other cultivars by 6.6–8.8 mg $seed^{-1}$ and a tiller m^{-2} a harvest by 29.0 m^{-2} for Strategies Two and Three. Strategy One did not show any difference in the database. Notably, Strategy One showed no additional improvements beyond grain yield.

3.3 | Model application

The grain yield predictions varied significantly depending on the calibration strategy, dataset used, and future climate scenario (RCP 4.5 and RCP 8.5) across the two locations, Feldkirchen and Thyrow (Figures 4–6).

At Feldkirchen, where conditions are favorable with better soil and higher precipitation, grain yield predictions showed less variation across calibration strategies and datasets. Using Strategy One with the full dataset, the predicted grain yield under historical conditions was 6719 kg ha^{-1}, while predictions using the restricted dataset were 6376 kg ha^{-1}. Under future climate scenarios, the predicted yields increased, with 7026 kg ha^{-1} under RCP 4.5 and 6937 kg ha^{-1} under RCP 8.5 for the full dataset. Predictions using the restricted dataset

TABLE 5 Performance analysis of model calibration strategies for cultivar Winnetou in crop modeling: A comparison of full and restricted dataset approaches.

Database	Strategy	Cal/Eval	Variables	No obs	RMSE	nRMSE	MAE	d-index	Pearson correlation	NSE
Full dataset with TKM-tiller m^{-2}-grain m^{-2}	One	Cal	BBCH	337	3.6	10.8	2.2	1.00	0.99	0.99
		Eval	BBCH	33	3.1	9.3	2.7	1.00	1.00	0.99
		Cal	AGDM	80	3951.7	45.7	1978.8	0.95	0.91	0.79
		Cal	LAI	512	2.0	147.2	1.6	0.65	0.60	−1.17
		Cal	Tillers m^{-2}	105	312.7	241.0	102.0	0.31	0.20	−4.86
		Eval	Tillers m^{-2}	22	208.2	147.9	106.0	0.34	−0.25	−1.29
		Cal	Grain m^{-2}	85	6471.6	156.1	5083.3	0.35	−0.15	−1.47
		Eval	Grain m^{-2}	11	5850.5	216.8	4927.7	0.36	−0.17	−4.17
		Cal	TKM	85	5.7	126.1	4.8	0.42	0.10	−0.61
		Eval	TKM	11	4.5	112.1	3.6	0.46	0.07	−0.38
		Cal	Grain yield	50	1982.6	112.5	1690.6	0.41	0.10	−0.29
		Eval	Grain yield	11	1807.7	162.2	1427.1	0.44	0.06	−1.89
	Two	Cal	BBCH	337	4.0	12.0	2.6	1.00	0.99	0.99
		Eval	BBCH	33	4.8	14.6	3.3	0.99	0.99	0.98
		Cal	AGDM	80	3980.9	46.0	1989.2	0.95	0.91	0.79
		Cal	LAI	512	2.2	161.7	1.7	0.63	0.59	−1.62
		Cal	Tillers m^{-2}	105	511.3	394.0	120.2	0.15	0.14	−14.68
		Eval	Tillers m^{-2}	22	152.8	108.5	74.8	0.57	0.30	−0.23
		Cal	Grain m^{-2}	85	5781.2	139.5	4749.7	0.34	−0.11	−0.97
		Eval	Grain m^{-2}	11	6499.2	240.9	5747.5	0.37	−0.06	−5.38
		Cal	TKM	85	4.9	109.1	3.8	0.52	0.26	−0.20
		Eval	TKM	11	3.9	98.4	2.7	0.51	0.25	−0.07
		Cal	Grain yield	50	2034.3	115.4	1780.4	0.46	0.16	−0.36
		Eval	Grain yield	11	2335.8	209.6	2131.9	0.45	0.45	−3.83
	Three	Cal	BBCH	337	4.1	12.4	2.6	1.00	0.99	0.98
		Eval	BBCH	33	4.9	15.0	3.3	0.99	0.99	0.98
		Cal	AGDM	80	3767.4	43.6	1870.7	0.95	0.91	0.81
		Cal	LAI	512	2.6	192.5	2.1	0.58	0.64	−2.71
		Cal	Tillers m^{-2}	105	427.8	329.7	118.0	0.20	0.17	−9.98
		Eval	Tillers m^{-2}	22	163.0	115.8	88.8	0.46	0.13	−0.40
		Cal	Grain m^{-2}	85	6051.5	146.0	4836.1	0.35	−0.12	−1.16
		Eval	Grain m^{-2}	11	5937.3	220.0	5048.6	0.36	−0.21	−4.33
		Cal	TKM	85	4.9	109.9	3.9	0.45	0.15	−0.22
		Eval	TKM	11	4.4	110.0	3.5	0.62	0.43	−0.33
		Cal	Grain yield	50	1986.7	112.7	1741.4	0.45	0.15	−0.30
		Eval	Grain yield	11	2241.4	201.1	2005.5	0.46	0.40	−3.45
Restricted dataset without TKM-tiller m^{-2}-grain m^{-2}	One	Cal	BBCH	337	3.6	10.8	2.2	1.00	0.99	0.99
		Eval	BBCH	33	3.1	9.3	2.7	1.00	1.00	0.99
		Cal	AGDM	80	3951.7	45.7	1978.8	0.95	0.91	0.79
		Cal	LAI	512	2.0	147.2	1.6	0.65	0.60	−1.17
		Cal	Tillers m^{-2}	105	312.7	241.0	102.0	0.31	0.20	−4.86
		Eval	Tillers m^{-2}	22	208.2	147.9	106.0	0.34	−0.25	−1.29
		Cal	Grain m^{-2}	85	5506.2	132.8	4806.5	0.31	−0.15	−0.79
		Eval	Grain m^{-2}	11	9632.9	357.0	9162.5	0.29	−0.17	−13.02

(Continues)

TABLE 5 (Continued)

Database	Strategy	Cal/Eval	Variables	No obs	RMSE	nRMSE	MAE	d-index	Pearson correlation	NSE
		Cal	TKM	85	14.3	317.2	13.4	0.32	0.11	−9.18
		Eval	TKM	11	14.7	369.7	14.2	0.28	NA	−14.03
		Cal	Grain yield	50	1968.9	111.7	1654.6	0.48	0.17	−0.27
		Eval	Grain yield	11	2067.9	185.6	1627.0	0.40	−0.10	−2.79
	Two	Cal	BBCH	337	4.0	12.0	2.6	1.00	0.99	0.99
		Eval	BBCH	33	4.8	14.6	3.3	0.99	0.99	0.98
		Cal	AGDM	80	4019.4	46.5	1991.7	0.95	0.91	0.78
		Cal	LAI	512	2.6	191.6	2.1	0.58	0.61	−2.68
		Cal	Tillers m^{-2}	105	543.4	418.8	123.6	0.14	0.14	−16.71
		Eval	Tillers m^{-2}	22	154.2	109.6	79.5	0.58	0.30	−0.26
		Cal	Grain m^{-2}	85	5363.6	129.4	4612.9	0.32	−0.11	−0.69
		Eval	Grain m^{-2}	11	9288.8	344.3	8787.2	0.29	−0.22	−12.04
		Cal	TKM	85	11.4	252.2	10.1	0.37	0.17	−5.44
		Eval	TKM	11	12.1	305.6	11.5	0.34	0.42	−9.27
		Cal	Grain yield	50	1927.1	109.4	1669.4	0.47	0.20	−0.22
		Eval	Grain yield	11	2037.3	182.8	1775.3	0.47	0.38	−2.68
	Three	Cal	BBCH	337	4.8	14.6	3.0	0.99	0.99	0.98
		Eval	BBCH	33	5.7	17.4	3.6	0.99	0.98	0.97
		Cal	AGDM	80	3783.6	43.8	1870.7	0.95	0.91	0.81
		Cal	LAI	512	2.9	214.7	2.4	0.55	0.63	−3.62
		Cal	Tillers m^{-2}	105	262.2	202.1	101.8	0.43	0.31	−3.12
		Eval	Tillers m^{-2}	22	165.4	117.5	91.9	0.45	0.11	−0.45
		Cal	Grain m^{-2}	85	121,20.8	292.4	10,390.6	0.32	−0.08	−7.65
		Eval	Grain m^{-2}	11	3950.7	146.4	3339.8	0.30	−0.18	−1.36
		Cal	TKM	85	10.9	243.1	9.9	0.37	0.09	−4.98
		Eval	TKM	11	11.4	286.3	10.6	0.38	0.47	−8.01
		Cal	Grain yield	50	1841.0	104.5	1616.5	0.47	0.21	−0.11
		Eval	Grain yield	11	2016.8	181.0	1730.2	0.51	0.41	−2.60

Abbreviations: MAE, mean absolute error; nRMSE, normalized root mean square error; NSE, Nash Sutcliffe efficiency; RMSE, normalized root mean square error; TKM, 1000-kernal mass.

followed a similar pattern, with 6976 kg ha^{-1} under RCP 4.5 and 7473 kg ha^{-1} under RCP 8.5 (Figure 6).

The impact of the calibration strategy and dataset choice became more evident when the model was applied to different locations and climate scenarios. Feldkirchen, with its better soil and precipitation, showed more consistent predictions, with smaller differences across strategies. In this location, the strategy × data interaction had less influence on the yield variability, likely because favorable growing conditions helped to buffer potential discrepancies between strategies.

In Thyrow, which has poorer soil and limited precipitation, predictions showed greater variability across calibration strategies and datasets. For Strategy One with the full dataset, the predicted grain yield was 6520 kg ha^{-1} under historical conditions, rising to 6796 kg ha^{-1} under RCP 4.5 and 6707 kg ha^{-1} under RCP 8.5. The restricted dataset predicted lower yields of 6108 kg ha^{-1} under historical conditions, with increases to 6454 kg ha^{-1} under RCP 4.5 and 6757 kg ha^{-1} under RCP 8.5 (Figures 5 and 6).

The strategy × data interaction was especially pronounced in Thyrow, a location with harsher growing conditions. Model performance varied substantially depending on the calibration strategy used and the amount of data available. With its poorer soil and limited water availability, Thyrow showed significant yield differences, especially under Strategy Three, which combined multiple variables and produced higher variability in predictions. This suggests that more complex strategies, such as Strategy Three, may introduce greater uncertainty in marginal environments.

Strategy Two, which prioritized dual variables (phenology and grain yield), displayed similar trends. At Feldkirchen, historical predictions were 6520 kg ha^{-1} with higher

TABLE 6 Values of coefficients estimated based on strategies and dataset.

Cultivar	Strategy	Dataset	Coefficients VSEN	PPSEN	P1	P5	SLAP1	SLAP2	STMMX	MXFIL	GRNO
Mulan	One	Full	3.75	3	300	670	300	245	3	1.29	24.67
		Restricted	3.75	3	300	670	300	245	3	1	29.79
	Two	Full	4.25	3.14	300	538	300	280	2.14	1.57	22.5
		Restricted	4.25	3.14	300	538	300	280	3	1	34
	Three	Full	2.6	2.6	344	600	220	260	1.86	1.25	23.41
		Restricted	1.89	2.33	500	600	300	280	3	3	20
Winnetou	One	Full	1.57	3.85	357	750	300	200	3	1.29	25.84
		Restricted	1.57	3.85	357	750	300	200	3	3	20
	Two	Full	2.5	1.75	450	632	290	210	3	1.57	24.51
		Restricted	2.5	1.75	450	632	300	245	3	3	20
	Three	Full	2.5	1.5	488	670	230	270	3	1.22	25
		Restricted	3.25	1.56	484	678	300	265	3	1	35
Ludwig	One	Full	1.5	3.14	431	710	230	240	3	1.57	23.61
		Restricted	1.5	3.14	431	710	300	270	3	1.29	24.69
	Two	Full	1.5	3.25	437	688	220	200	3	1.57	23.8
		Restricted	1.5	3.25	437	688	300	206	3	1.86	27.5
	Three	Full	2	2	475	650	220	205	3	1.88	23.3
		Restricted	1.25	2.25	500	631	300	280	3	1.57	22.6

TABLE 7 Error statistics from default setting against the observed data.

Cultivars	Error statistics	BBCH	Tiller m^{-2}	Grain m^{-2}	TKM	Grain yield	LAI	Top
Ludwig (n = 86) 54 sites	RMSE	10.21	191.7	3218	9.73	1847	1.37	5595
	NSE	0.868	−1.60	−1.44	−1.02	−1.11	−0.433	−0.477
	Pearson	0.970	0.096	0.043	0.236	0.370	0.159	0.122
	d-index	0.968	0.433	0.326	0.476	0.590	0.470	0.544
	MAE	7.71	153.0	2444	7.49	1407	1.21	4366
	nRMSE	36.30	160.2	155.3	141.2	144.5	119.0	112.5
Mulan (n = 93) 48 sites	RMSE	9.48	220.7	7136	8.67	2640	2.10	6857
	NSE	0.901	−1.33	−1.28	−1.78	−0.970	−2.35	−0.937
	Pearson	0.978	0.169	−0.318	−0.094	0.260	−0.093	−0.050
	d-index	0.976	0.448	0.191	0.268	0.510	0.356	0.409
	MAE	7.19	179.1	5204	7.02	2283	1.78	5645
	nRMSE	31.40	151.7	150.1	165.8	139.7	182.7	137.0
Winnetou (n = 61) 37 sites	RMSE	8.76	800.2	6726	4.81	2629	1.90	4050
	NSE	0.929	−28.1	−2.33	−0.394	−1.43	−0.918	0.778
	rPearson	0.985	0.044	−0.451	0.189	0.073	0.250	0.895
	d-index	0.983	0.074	0.140	0.477	0.448	0.570	0.944
	MAE	6.37	251.2	5944	3.77	2309	1.53	2080
	nRMSE	26.60	535.0	181.0	117.1	154.7	138.4	46.8

Abbreviations: LAI, leaf area index; MAE, mean absolute error; nRMSE, normalized root mean square error; NSE, Nash Sutcliffe efficiency; RMSE, normalized root mean square error; TKM, 1000-kernal mass.

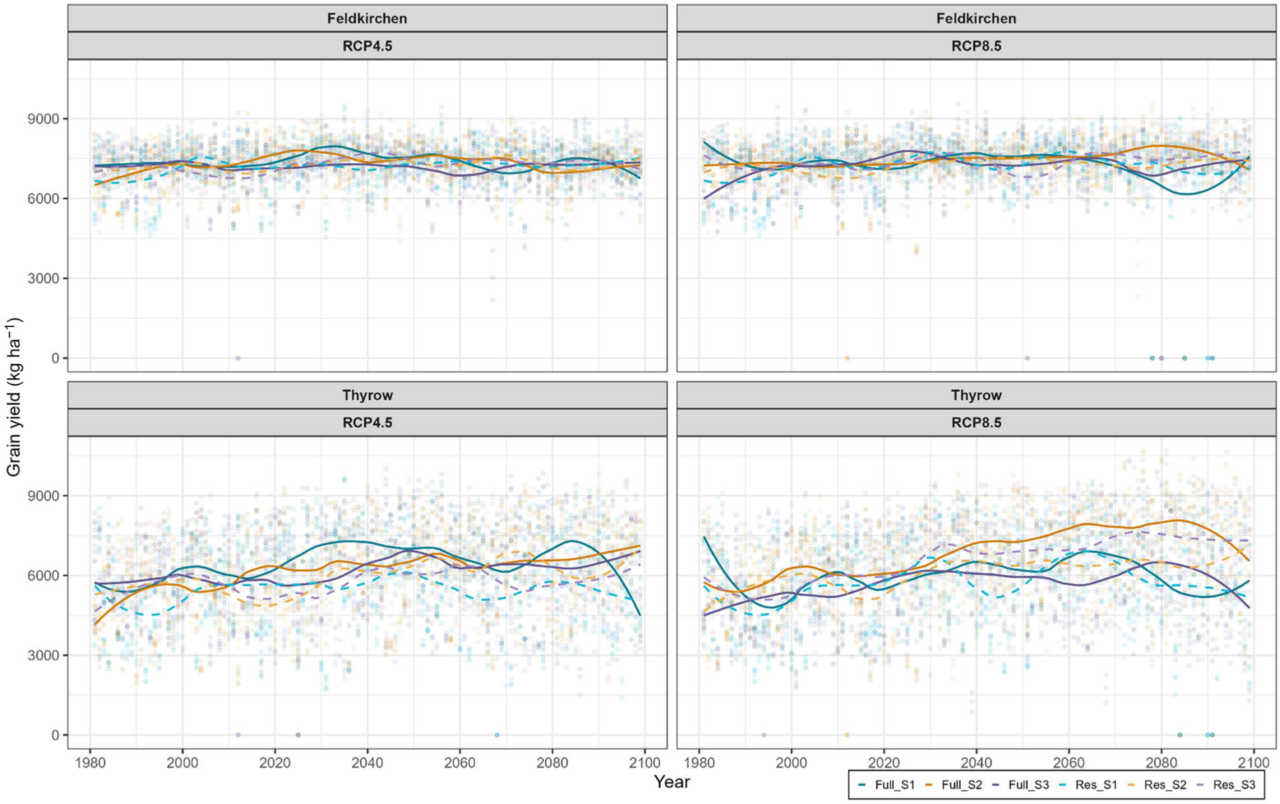

FIGURE 4 Projected grain yield for Feldkirchen and Thyrow under representative concentration pathways (RCP) 4.5 and RCP 8.5 using different calibration strategies and dataset selections. Solid lines represent "full dataset"; dashed lines "restricted dataset". Colors indicate calibration scenarios: Strategy One/S1 (blue), Strategy Two/S2 (orange), Strategy Three/S3 (purple).

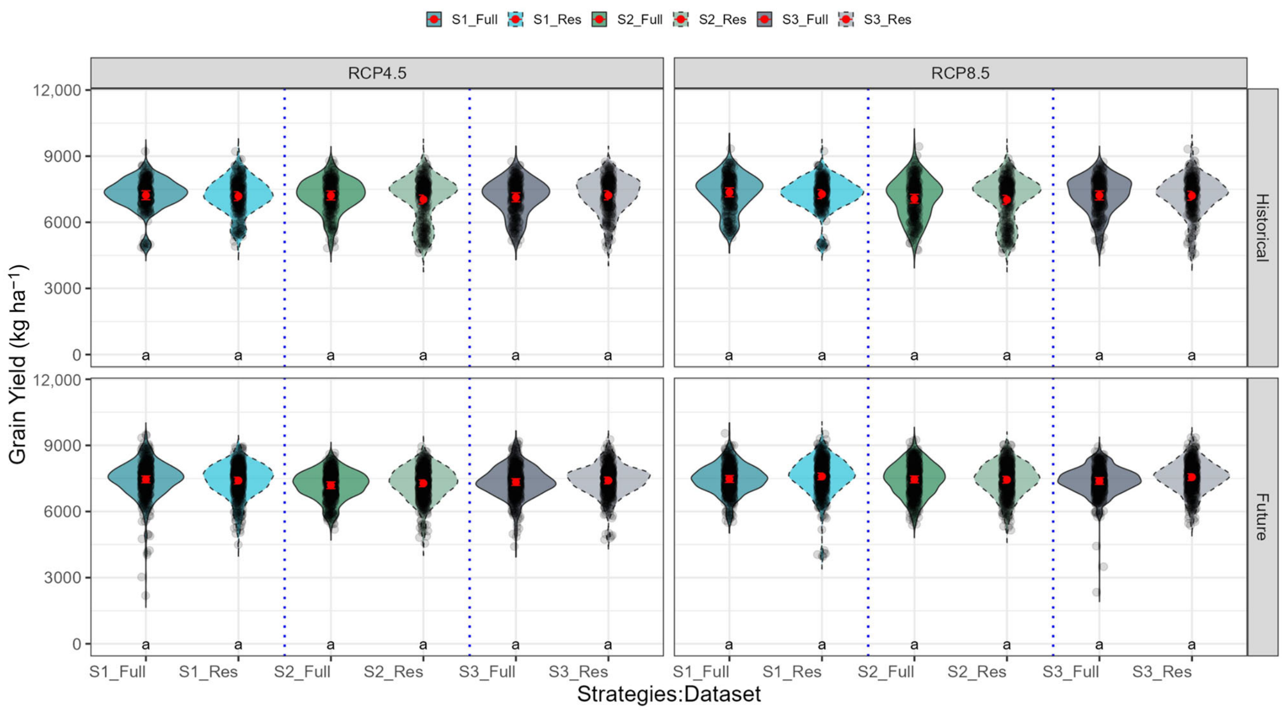

FIGURE 5 Comparison of calibration strategies for grain yield predictions in Feldkirchen under representative concentration pathway (RCP) 4.5 and RCP 8.5.

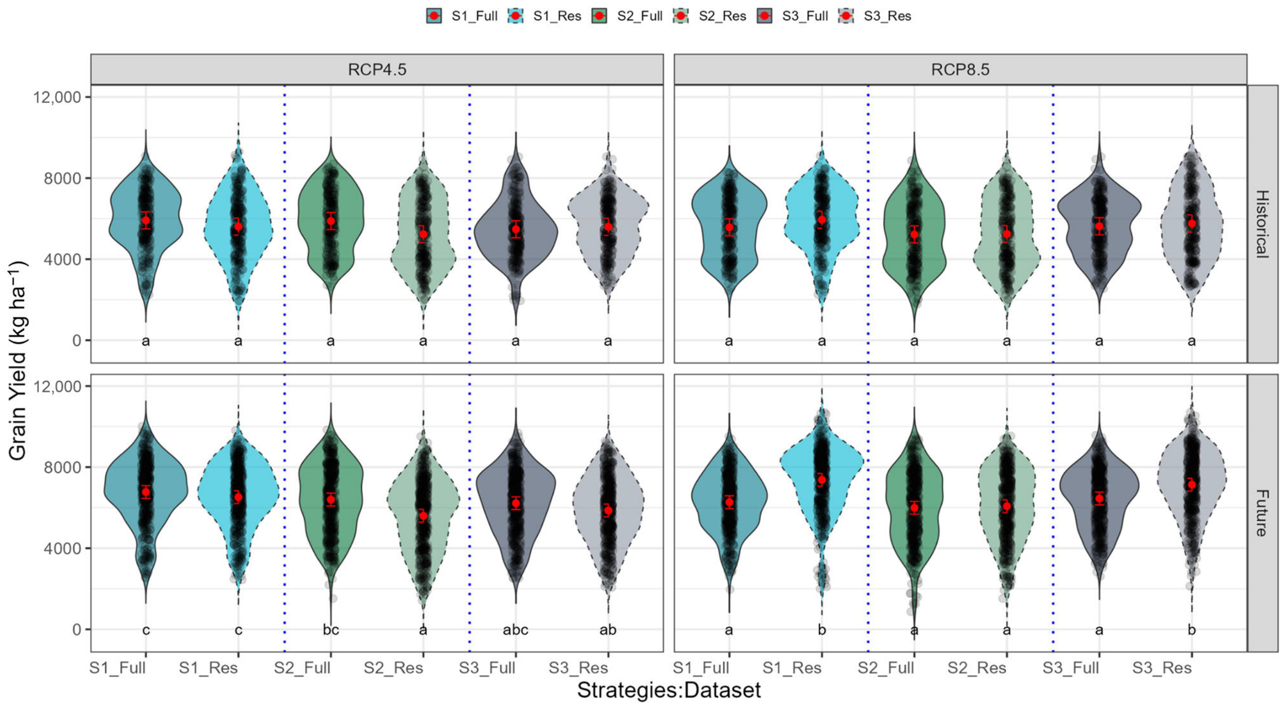

FIGURE 6 Comparison of calibration strategies for grain yield predictions in Thyrow under representative concentration pathway (RCP) 4.5 and RCP 8.5.

values under future climate conditions (Figures 4 and 6). At Thyrow, historical predictions were 6174 kg ha^{-1}, increasing to 6707 kg ha^{-1} under RCP 8.5. The restricted dataset for Strategy Two also showed lower historical predictions but moderate increases under future scenarios.

In comparison, Strategy Three, which considered all variables together, exhibited more variability, particularly in Thyrow. Using the restricted dataset, Strategy Three predicted 6295 kg ha^{-1} under historical conditions and 6718 kg ha^{-1} under RCP 4.5, with the highest prediction of 7269 kg ha^{-1} under RCP 8.5.

Across both locations, the predictions under future climate conditions (RCP 4.5 and RCP 8.5) were generally higher compared to historical conditions, with Thyrow showing more pronounced variation between strategies and datasets. Strategy Three, especially with the restricted dataset, resulted in the highest yield predictions in Thyrow under RCP 8.5.

4 | DISCUSSION

4.1 | Evaluating calibration strategies and their impact on crop model performance

In this study, we applied TSE and distinct datasets to calibrate the CSM-Nwheat model using three calibration strategies. The results highlight how dataset selection and calibration strategies influence model coefficient settings, performance indices, and crop model application in different environmental conditions.

Wallach et al. (2018) emphasize that model calibration is inherently imperfect; the objective is not to find a “true” parameter set but one that is least wrong and physiologically plausible. Our results showed that no single calibration method consistently outperformed overall performance measures. Each strategy emphasized different facets of the DSSAT-Nwheat model. Notably, Strategy Two, which prioritized dual variables, demonstrated the best overall performance. However, it displayed one of the least favorable outcomes in grain yield simulation when utilizing the full dataset. This reflects the trade-offs inherent in multi-objective calibration, where optimizing for certain variables may degrade performance in others (Wallach, Palosuo, Thorburn, Gourdain, et al., 2021).

The model application extended our analysis by exploring how these calibration strategies interact with different environmental conditions and climate scenarios across two locations: Thyrow and Feldkirchen. Thyrow, characterized by poorer soil and drier conditions, showed more pronounced yield variations under the future RCP 8.5 scenario, particularly for Strategy One and Strategy Three, where yields increased from 6373 (historical) to 6937 kg ha^{-1} (future RCP 8.5). Feldkirchen, with better soil and wetter conditions, demonstrated more stable predictions across all strategies, suggesting that future climate impacts may affect yield predictions more in marginal environments.

However, one limitation of our study is the absence of measured water and nitrogen stress data, which restricts our ability to evaluate the accuracy of the simulated stress impacts. While the DSSAT model internally simulates stress responses, their validation under extreme events such as drought and heat remains challenging. This is particularly important under future climate scenarios where the frequency and intensity of extreme events are projected to increase (Trnka et al., 2014; Zhao et al., 2016).

Intriguingly, our study found that the incorporation of dynamic variables did not universally enhance model predictions. In fact, using full dataset-based strategies resulted in higher errors for grain yield and tiller density. However, Strategy Three, based on a restricted dataset, exhibited significantly higher standard deviations for specific plant processes, such as tiller production. (Supporting Information 1 and 2). This can be linked to parameter equifinality, where different parameter combinations yield similar outcomes but represent distinct physiological assumptions (He et al., 2017).

On the other hand, Strategy Two predicted the dynamic variables more accurately than Strategies One and Three (but not all cases) based on both restricted and full datasets. This implies that if the aim is to accurately simulate final outputs like yield, calibration focused on those variables may be sufficient. However, if intermediate traits are important—for example, for stress sensitivity—those variables should be explicitly included in the calibration procedure (Guillaume et al., 2011; Thorp et al., 2008).

In our calibration strategies, there are some interesting patterns that were noticed. Strategies One and Two always came up with the same coefficients for modeling phenology. But Strategy Three indicated that the Ludwig cultivar should have a longer vernalization (VSEN) and a shorter photoperiod (PPSEN). This is consistent with physiological theory, which suggests that increasing PPSEN and decreasing P1 can delay flowering when photoperiod exerts a stronger control on development (Adams et al., 2001). These new ideas show how important it is to make sure that calibration strategies take into account the unique phenological traits of each crop.

The results also show that using all relevant coefficients together with all possible target variables and user-specified calibration priorities outperforms the expert system, which is designed for the specific model under consideration and requires user intervention throughout the entire calibration process. However, it should be noted that Strategy Three, based on generic selection of coefficient, can also take a significant amount of time while calibrating, because it will iterate through many possible combinations to find a best match based on the lowest sum squared error.

First, as previously stated, user intervention is critical for model calibration to a specific application. Second, critical and sound crop physiological knowledge is required to evaluate data quality, model setup, and model calibration. To avoid the calibration routine providing parameter estimates that compensate for errors rather than optimizing model performance, data and model errors should be properly taken into account. For example, because of the inherent uncertainties in measuring catchment average tiller density, attempting to obtain a perfect grain yield fit may be questionable. When weighing objectives, one should prioritize other factors within certain limits, say 5%–10%.

We also found that using grain yield as a secondary criterion during the calibration steps can improve estimates of the final grain yield. This dual-criteria approach makes it easier for the model to match up with more aspects of crop growth and yield, which could lead to more accurate results. This finding aligns with recommendations by H. Madsen (2003), who emphasized the benefits of multi-objective calibration for model identifiability.

The results of our study have implications for crop modeling in general. It is clear that using a large dataset and a sophisticated calibration strategy are important for making crop models more accurate. This supports previous findings that multi-environment calibration using process-based variables improves model reliability under diverse scenarios (Huffman et al., 2014; Liu et al., 2020). Additionally, our study shows how important it is to think about phenological characteristics when coming up with calibration strategies. The crop's phenology is a key factor in determining its growth and yield, and a model's accuracy depends on how well it represents these changes over time.

When comparing various calibration algorithms, it is generally prudent to consider their efficiency. Strategies One, Two and Three required between 15,000 and 300,000 (Supporting Information 4) model evaluations in the current instance, whereas the expert system could be significantly more effective and only required fifty model evaluations. A threshold of 1000–10,000 model evaluations may be deemed unsuitable for more expensive models. While the generic search routines discussed in this article may still be applicable in such circumstances, their effectiveness must be compromised. Utilizing prior knowledge regarding the calibration parameters, such as realistic limit specifications and correlation structures, as well as understanding the effects of parameter changes, will, on average, improve the performance of an automatic calibration scheme (Wallach et al., 2001).

A problem that is intricately linked to the dimensions of efficiency and effectiveness is that of parameter insensitivity. The identification of insensitive parameters through sensitivity analysis, such as the Monte Carlo-based procedure suggested by Spear and Hornberger (1980), reduces the quantity of free model parameters that require calibration. Some of these parameters can only be inferred through output calibration to observed system data, which may conflict with the fact that a given model output may be affected by the interaction among several parameters (Attia et al., 2021).

Therefore, this type of analysis should be approached with caution when interpreting its results. The analysis fails to adequately consider parameter correlations, which suggests that seemingly insensitive parameters might have significant correlations with other parameters that are critical for the behavior of the model (H. O. Madsen, 2000). Additionally, it is worth mentioning that the incorporation of multiple objectives during the calibration procedure results in parameters that are easier to identify and a model structure that is more well-posed (H. O. Madsen, 2000).

In the analysis, Strategy Three, which incorporated phenology, growth, and yield traits, provided broader flexibility in simulating plant responses. However, this also introduced higher uncertainty in parameter estimates, particularly for tiller density. In contrast, Strategy One, which focused on fewer objectives, showed more stable and consistent parameter identification, particularly in phenology simulation, but it lacked accuracy in predicting growth-related variables. This suggests that a more comprehensive model structure does not necessarily guarantee better predictive performance and may lead to overfitting for specific traits.

4.2 | Challenges of addressing spatial variability and secondary dataset in process-based crop models

Process-based crop models were developed primarily with point-based considerations, assuming high-quality soil, weather, and crop growth data for calibration, which often results in excellent predictive accuracy under such controlled conditions (Jones et al., 2003). However, when models rely on secondary dataset like plant variety trials, uncertainties inherent in the data, particularly regarding soil and location information, can introduce significant variability in simulations (Carcedo et al., 2023). For instance, in this study, the BÜK2000 soil dataset—a generalized polygon-based national inventory—was used to estimate soil parameters. Locations in the plant variety trials dataset lacked precise geo-coordinates, requiring estimation based on contextual information such as soil quality ratings, precipitation, and expert knowledge, which may lead to approximations. Furthermore, the spatial variability of soil types within a single estimated location, coupled with changes in trial field locations due to crop rotation, adds additional layers of uncertainty (Laidig et al., 2022). These challenges are compounded by the diversity of soil types (from light sandy to heavy organic soils) and climatic conditions (ranging from extreme drought to excessive rainfall) included in the study. Such variability often leads to higher RMSE and lower d-index in model simulations when broad environmental conditions are considered, as opposed to smaller, site-specific dataset where model parameterization can closely mirror observed conditions. Experimental errors, deviations in soil data, and unaccounted biotic or abiotic stresses further highlight the trade-off between model accuracy and robustness (Grassini et al., 2015). Consequently, the need for robust models that account for G × E × M interactions, while managing uncertainties and spatial variability, becomes a strategic necessity to ensure reliable predictions across diverse agro-environments.

5 | CONCLUSION

The findings of our study indicate that the selection of the dataset significantly impacts the model's ability to accurately simulate various crop variables. The utilization of the complete dataset resulted in significantly improved accuracy in the simulations of biomass, LAI, and tiller m^{-2} at harvest. On the contrary, the restricted dataset exhibited an unexpected overestimation of the grain yield. The observed disparity can likely be attributed to the enhanced level of granularity present in the comprehensive dataset, thereby facilitating a more comprehensive understanding of the intricate interconnections among various aspects of crop development. Furthermore, it was observed that the three calibration strategies exhibited varying performance across different crop variables. Among the three strategies considered, it was found that Strategy One exhibited the highest level of accuracy in simulating grain yield, grain number, and TKM for all three cultivars. Among the strategies employed, it was found that Strategy Two yielded the highest level of accuracy in simulating LAI. On the other hand, Strategy Three demonstrated the greatest accuracy in simulating biomass and tiller density per m^{-2} at the time of harvest. These findings indicate that the selection of a calibration strategy should be determined individually for each case, taking into account the particular crop variable being simulated. It is generally advisable to utilize the complete dataset; however, the limited dataset may be employed if it proves to be more feasible or cost-efficient. In general, this study offers significant contributions to the understanding and refinement of crop models' calibration. The findings can be utilized to enhance the precision of agricultural models and to facilitate more informed choices regarding crop management.

AUTHOR CONTRIBUTIONS

Ashifur Rahman Shawon: Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing—original draft; writing—review and editing. **Ahmed Attia**: Writing—review and editing. **Jonghan Ko**: Writing—review and editing. **Emir Memic**: Software; writing—review and editing. **Ralf Uptmoor**: Writing—review and editing. **Bernd Hackauf**:

Funding acquisition; project administration; supervision; writing—review and editing. **Til Feike**: Funding acquisition; methodology; project administration; resources; supervision; writing—review and editing.

ACKNOWLEDGMENTS

The authors would like to thank the Federal Plant Variety Office (Bundessortenamt, Hannover, Germany) for kindly providing the variety trial data used in this study, and the German Weather Service (DWD) for supplying the daily weather data. Further gratitude is extended to Dr. Siegfried Schittenhelm and Dr. Klaus Sieling for generously providing detailed wheat phenotyping data. The authors also acknowledge the funding received by the German Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) through the PRoKlima project as part of the German Climate Protection Program 2022 and financial support received through the RYE-SUS project funded by the European Union's Horizon 2020 research and innovation programme, implemented under the ERA-NET Cofund SusCrop (Grant No. 771134) as part of the Joint Programming Initiative on Agriculture, Food Security and Climate Change (FACCE-JPI).

Open access funding enabled and organized by Projekt DEAL.

ORCID

Ashifur Rahman Shawon https://orcid.org/0000-0002-0276-7497
Ahmed Attia https://orcid.org/0000-0002-6751-6584
Jonghan Ko https://orcid.org/0000-0001-7974-3808
Emir Memic https://orcid.org/0000-0002-0184-5481

REFERENCES

Adams, S. R., Pearson, S., & Hadley, P. (2001). Improving quantitative flowering models through a better understanding of the phases of photoperiod sensitivity. *Journal of Experimental Botany*, *52*(357), 655–662. https://doi.org/10.1093/jexbot/52.357.655

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. https://doi.org/10.1016/j.agrformet.2012.11.017

Asseng, S., Jamieson, P. D., Kimball, B. A., Pinter, P. J., Sayre, K. D., Bowden, J. W., & Howden, M. S. (2004). Simulated wheat growth affected by rising temperature, increased water deficit and elevated atmospheric CO_2. *Field Crops Research*, *85*(2–3), 85–102. https://doi.org/10.1016/S0378-4290(03)00154-0

Asseng, S., Milroy, S. P., & Poole, M. L. (2008). Systems analysis of wheat production on low water-holding soils in a Mediterranean-type environment. *Field Crops Research*, *105*(1–2), 97–106. https://doi.org/10.1016/j.fcr.2007.08.003

Attia, A., El-Hendawy, S., Al-Suhaibani, N., Tahir, M. U., Mubushar, M., Vianna, M. D. S., Ullah, H., Mansour, E., & Datta, A. (2021). Sensitivity of the DSSAT model in simulating maize yield and soil carbon dynamics in arid Mediterranean climate: Effect of soil, genotype and crop management. *Field Crops Research*, *260*, 107981. https://doi.org/10.1016/j.fcr.2020.107981

Balkovič, J., van der Velde, M., Schmid, E., Skalský, R., Khabarov, N., Obersteiner, M., Stürmer, B., & Xiong, W. (2013). Pan-European crop modelling with EPIC: Implementation, up-scaling and regional crop yield validation. *Agricultural Systems*, *120*, 61–75. https://doi.org/10.1016/j.agsy.2013.05.008

Berton Ferreira, T., Shelia, V., Porter, C., Moreno Cadena, P., Salmeron Cortasa, M., Sohail Khan, M., Pavan, W., & Hoogenboom, G. (2024). Enhancing crop model parameter estimation across computing environments: Utilizing the GLUE method and parallel computing for determining genetic coefficients. *Computers and Electronics in Agriculture*, *227*, 109513. https://doi.org/10.1016/j.compag.2024.109513

Boote, K. J. (1999). Concepts for calibrating crop growth models. In G. Hoogenboom, P. W. Wilkens, and G. Y. Tsuji (Eds.), *DSSAT version 3* (Vol. 4, pp. 179–200). University of Hawaii.

Boote, K. J., Jones, J. W., & Pickering, N. B. (1996). Potential uses and limitations of crop models. *Agronomy Journal*, *88*(5), 704–716. https://doi.org/10.2134/agronj1996.00021962008800050005x

Boserup, E., & Chambers, R. (2014). *The conditions of agricultural growth*. Routledge. https://doi.org/10.4324/9781315070360

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche-Wachter, D., Zimmer, D., Sierra, J. O., Bertuzzi, P., Burger, P., Bussière, F., Cabidoche, Y. M., Cellier, P., Debaeke, P., Gaudillère, J. P., Hénault, C., Maraux, F., Seguin, B., & Sinoquet, H. (2003). An overview of the crop model STICS. *European Journal of Agronomy*, *18*(3–4), 309–332. https://doi.org/10.1016/S1161-0301(02)00110-7

Brookfield, H., & Padoch, C. (1994). Appreciating agrodiversity: A look at the dynamism and diversity of indigenous farming practices. *Environment: Science and Policy for Sustainable Development*, *36*(5), 6–45. https://doi.org/10.1080/00139157.1994.9929164

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Bundessortenamt. (2007). *Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen*. Bundessortenamt, Hannover. https://www.bundessortenamt.de/bsa/pruefungsansteller/richtlinien/durchfuehrung/

Carcedo, A. J., Vieira Junior, N., Marziotte, L., Correndo, A. A., Araya, A., Prasad, P. V., Min, D., Stewart, Z. P., Faye, A., & Ciampitti, I. A. (2023). The urgency for investment on local data for advancing food assessments in Africa: A review case study for APSIM crop modeling. *Environmental Modelling & Software*, *161*, 105633. https://doi.org/10.1016/j.envsoft.2023.105633

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. https://doi.org/10.1016/j.agrformet.2010.09.007

Confalonieri, R., Bregaglio, S., & Acutis, M. (2016). Quantifying uncertainty in crop model predictions due to the uncertainty in the observations used for calibration. *Ecological Modelling*, *328*, 72–77. https://doi.org/10.1016/j.ecolmodel.2016.02.013

de Wit, C. T. (1982). Simulation of living systems. In F. W. T. Penning de Vries & H. H. van Laar (Eds.), *Simulation of plant growth and crop production* (pp. 3–8). Centre for Agricultural Pub. and Documentation.

Deutscher Wetterdienst. (2017). *CDC (climate data center)*. https://cdc.dwd.de/portal/

Diepen, C. A., Wolf, J., van Keulen, H. V., & Rappoldt, C. (1989). WOFOST: A simulation model of crop production. *Soil Use and Management*, *5*(1), 16–24. https://doi.org/10.1111/j.1475-2743.1989.tb00755.x

Di Paola, A., Valentini, R., & Santini, M. (2016). An overview of available crop growth and yield models for studies and assessments in agriculture. *Journal of the Science of Food and Agriculture*, *96*(3), 709–714. https://doi.org/10.1002/jsfa.7359

Gaiser, T., Perkons, U., Küpper, P. M., Kautz, T., Uteau-Puschmann, D., Ewert, F., Enders, A., & Krauss, G. (2013). Modeling biopore effects on root growth and biomass production on soils with pronounced subsoil clay accumulation. *Ecological Modelling*, *256*, 6–15. https://doi.org/10.1016/j.ecolmodel.2013.02.016

Grassini, P., van Bussel, L. G., van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., Groot, H. d., van Ittersum, M. K., & Cassman, K. G. (2015). How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Research*, *177*, 49–63. https://doi.org/10.1016/j.fcr.2015.03.004

Guarin, J. R., Asseng, S., Martre, P., & Bliznyuk, N. (2020). Testing a crop model with extreme low yields from historical district records. *Field Crops Research*, *249*, 107269. https://doi.org/10.1016/j.fcr.2018.03.006

Guillaume, S., Bergez, J.-E., Wallach, D., & Justes, E. (2011). Methodological comparison of calibration procedures for durum wheat parameters in the STICS model. *European Journal of Agronomy*, *35*(3), 115–126. https://doi.org/10.1016/j.eja.2011.05.003

Hammer, G. L., van Oosterom, E., McLean, G., Chapman, S. C., Broad, I., Harland, P., & Muchow, R. C. (2010). Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. *Journal of Experimental Botany*, *61*(8), 2185–2202. https://doi.org/10.1093/jxb/erq095

Hansen, S. L., Jensen, H. E., Nielsen, N. E., & Svendsen, H. (1991). Simulation of nitrogen dynamics and biomass production in winter wheat using the Danish simulation model DAISY. *Nutrient Cycling in Agroecosystems*, *27*(2–3), 245–259. https://doi.org/10.1007/BF01051131

Hartwich, R., Adler, G. H., Behrens, J., Eckelmann, W., & Richter, A. (1995). The 1:1000 000 soil map of the Federal Republic of Germany. *Zeitschrift fur Angewandte Geologie*, *41*(1), 31–36. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0029500251&partnerID=40&md5=8a98572ba1d83055a35569708b81dbf4

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

Hoogenboom, G., Porter, C. H., Boote, K. J., Shelia, V., Wilkens, P. W., Singh, U., White, J. W., Asseng, S., Lizaso, J. I., Moreno-Cadena, L. P., Pavan, W., Ogoshi, R. M., Hunt, L. A., Tsuji, G. Y., & Jones, J. W. (2019). The DSSAT crop modeling ecosystem. In K. J. Boote (Ed.), *Advances in crop modelling for a sustainable agriculture* (pp. 173–216). Burleigh Dodds Science Publishing. https://doi.org/10.19103/AS.2019.0061.10

Huffman, T., Qian, B., Jong, R. D., Liu, J., Wang, H., McConkey, B., Brierley, T., & Yang, J. J. (2014). Upscaling modelled crop yields to regional scale: A case study using DSSAT for spring wheat on the Canadian prairies. *Canadian Journal of Soil Science*, *95*(1), 49–61. https://doi.org/10.4141/cjss-2014-076

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A., & Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, *18*(3–4), 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7

Kassie, B. T., Asseng, S., Porter, C. H., & Royce, F. S. (2016). Performance of DSSAT-Nwheat across a wide range of current and future growing conditions. *European Journal of Agronomy*, *81*, 27–36. https://doi.org/10.1016/j.eja.2016.08.012

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D. P., Huth, N., Hargreaves, J. N., Meinke, H. H., Hochman, Z., McLean, G. B., Verburg, K., Snow, V., Dimes, J. P., Silburn, M., Wang, E., Brown, S. D., Bristow, K. L., Asseng, S., … Smith, C. J. (2003). An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, *18*(3–4), 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9

Keating, B. A., Meinke, H., Probert, M. E., Huth, N., & Hills, I. G. (2001). *NWheat: Documentation and performance of a wheat module for APSIM*. CSIRO.

Kersebaum, K. C. (2007). Modelling nitrogen dynamics in soil–crop systems with HERMES. *Nutrient Cycling in Agroecosystems*, *77*(1), 39–52. https://doi.org/10.1007/s10705-006-9044-8

Krishnan, P., Sharma, R. K., Dass, A., Kukreja, A., Srivastav, R., Singhal, R. J., Bandyopadhyay, K. K., Lal, K., Manjaiah, K., Chhokar, R. S., & Gill, S. C. (2016). Web-based crop model: Web InfoCrop—Wheat to simulate the growth and yield of wheat. *Computers and Electronics in Agriculture*, *127*, 324–335. https://doi.org/10.1016/j.compag.2016.06.008

Laidig, F., Hüsken, A., Rentel, D., & Piepho, H.-P. (2022). Protein use efficiency and stability of baking quality in winter wheat based on the relation of loaf volume and grain protein content. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *135*(4), 1331–1343. https://doi.org/10.1007/s00122-022-04034-x

Laidig, F., Piepho, H.-P., Rentel, D., Drobek, T., Meyer, U., & Huesken, A. (2017). Breeding progress, environmental variation and correlation of winter wheat yield and quality traits in German official variety trials and on-farm during 1983-2014. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *130*(1), 223–245. https://doi.org/10.1007/s00122-016-2810-3

Liu, H., Pequeno, D. N. L., Hernández-Ochoa, I. M., Krupnik, T. J., Sonder, K., Xiong, W., & Xu, Y. (2020). A consistent calibration across three wheat models to simulate wheat yield and phenology in China. *Ecological Modelling*, *430*, 109132. https://doi.org/10.1016/j.ecolmodel.2020.109132

Lobell, D. B., Schlenker, W., & Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science*, *333*(6042), 616–620. https://doi.org/10.1126/science.1204531

Madsen, H. (2003). Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources*, *26*(2), 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1

Madsen, H. O. (2000). Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology*, *235*(3), 276–288. https://doi.org/10.1016/S0022-1694(00)00279-1

Martynov, S., Dobrotvorskaya, T., & Dobrotvorskiy, D. (2024). *Genetic resources information system for wheat and triticale*. GRIS. http://wheatpedigree.net

McCown, R. L., Hammer, G. L., Hargreaves, J. N., Holzworth, D. P., & Freebairn, D. M. (1996). APSIM: A novel software system for model development, model testing and simulation in agricultural systems research. *Agricultural Systems*, *50*(3), 255–271. https://doi.org/10.1016/0308-521X(94)00055-V

Memic, E., Graeff-Hönninger, S., Boote, K. J., Hensel, O., & Hoogenboom, G. (2021). Cultivar coefficient estimator for the cropping system model based on time-series data: A case study for soybean. *Transactions of the ASABE*, *64*(4), 1391–1402. https://doi.org/10.13031/trans.14432

Moeys, J. (2012). *The soil texture wizard: R Functions for plotting, classifying, transforming and exploring soil texture data*. https://cran.r-project.org/web/packages/soiltexture/

Nendel, C., Berg, M. C., Kersebaum, K. C., Mirschel, W., Specka, X., Wegehenkel, M., Wenkel, K. O., & Wieland, R. (2011). The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. *Ecological Modelling*, *222*(9), 1614–1625. https://doi.org/10.1016/j.ecolmodel.2011.02.018

Panagos, P., van Liedekerke, M., Jones, A., & Montanarella, L. (2012). European soil data centre: Response to European policy support and public data requirements. *Land Use Policy*, *29*(2), 329–338. https://doi.org/10.1016/j.landusepol.2011.07.003

Pasley, H., Brown, H., Holzworth, D. P., Whish, J., Bell, L., & Huth, N. (2023). How to build a crop model. A review. *Agronomy for Sustainable Development*, *43*(1), Article 2. https://doi.org/10.1007/s13593-022-00854-9

Pickering, N. B., Hansen, J. W., Jones, J. W., Wells, C. M., Chan, V. K., & Godwin, D. C. (1994). WeatherMan: A utility for managing and generating daily weather data. *Agronomy Journal*, *86*(2), 332–337. https://doi.org/10.2134/agronj1994.00021962008600020023x

Porter, J. R., & Semenov, M. A. (2005). Crop responses to climatic variation. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, *360*(1463), 2021–2035. https://doi.org/10.1098/rstb.2005.1752

Rawls, W. J., Brakensiek, D. L., & Saxtonn, K. E. (1982). Estimation of soil water properties. *Transactions of the ASAE*, *25*(5), 1316–1320. https://doi.org/10.13031/2013.33720

Reyenga, P. J., Howden, S. M., Meinke, H., & McKeon, G. M. (1999). Modelling global change impacts on wheat cropping in south-east Queensland, Australia. *Environmental Modelling & Software*, *14*(4), 297–306. https://doi.org/10.1016/S1364-8152(98)00081-4

Ritchie, J. T., Godwin, D. C., & Otter-Nacke, S. (1985). *CERES-wheat: A user-oriented wheat yield model preliminary documentation* (AGRISTARS Publication No. YM-U3-04442-JSC-18892). Michigan State University.

Röll, G., Memic, E., & Graeff-Hönninger, S. (2020). Implementation of an automatic time-series calibration method for the DSSAT wheat models to enhance multi-model approaches. *Agronomy Journal*, *112*(5), 3891–3912. https://doi.org/10.1002/agj2.20328

Rötter, R. P., Palosuo, T., Kersebaum, K. C., Angulo, C., Bindi, M., Ewert, F., Ferrise, R., Hlavinka, P., Moriondo, M., Nendel, C., Olesen, J. E., Patil, R. H., Ruget, F., Takáč, J., & Trnka, M. (2012). Simulation of spring barley yield in different climatic zones of Northern and Central Europe: A comparison of nine crop models. *Field Crops Research*, *133*, 23–36. https://doi.org/10.1016/j.fcr.2012.03.016

Schittenhelm, S., Kraft, M., & Wittich, K.-P. (2014). Performance of winter cereals grown on field-stored soil moisture only. *European Journal of Agronomy*, *52*, 247–258. https://doi.org/10.1016/j.eja.2013.08.010

Shawon, A. R., Memic, E., Kottmann, L., Uptmoor, R., Hackauf, B., & Feike, T. (2024). Comprehensive evaluation of the DSSAT-CSM-CERES-Wheat for simulating winter rye against multi-environment data in Germany. *Agronomy Journal*, *116*(4), 1844–1868. https://doi.org/10.1002/agj2.21590

Sieling, K., Böttcher, U., & Kage, H. (2016). Canopy traits in rye, triticale and wheat under varying N supply. *Agronomy Research*, *14*(4), 1467–1485. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84995471844&partnerID=40&md5=fcfb694ae3589f2a9e46c959daae9d59

Sinclair, T. R. (1986). Water and nitrogen limitations in soybean grain production I. Model development. *Field Crops Research*, *15*(2), 125–141. https://doi.org/10.1016/0378-4290(86)90082-1

Spear, R. C., & Hornberger, G. M. (1980). Eutrophication in peel inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. *Water Research*, *14*(1), 43–49. https://doi.org/10.1016/0043-1354(80)90040-8

Steduto, P., Hsiao, T. C., Raes, D., & Fereres, E. (2009). AquaCrop—The FAO crop model to simulate yield response to water: I. Concepts and underlying principles. *Agronomy Journal*, *101*(3), 426–437. https://doi.org/10.2134/agronj2008.0139s

Stöckle, C. O., Donatelli, M., & Nelson, R. (2003). CropSyst, a cropping systems simulation model. *European Journal of Agronomy*, *18*(3–4), 289–307. https://doi.org/10.1016/S1161-0301(02)00109-0

Stöckle, C. O., Martin, S. A., & Campbell, G. S. (1994). CropSyst, a cropping systems simulation model: Water/nitrogen budgets and crop yield. *Agricultural Systems*, *46*(3), 335–359. https://doi.org/10.1016/0308-521X(94)90006-2

Tardieu, F., Granato, Í. S. C., van Oosterom, E. J., Parent, B., & Hammer, G. L. (2020). Are crop and detailed physiological models equally 'mechanistic' for predicting the genetic variability of whole-plant behaviour? The nexus between mechanisms and adaptive strategies. *in silico Plants*, *2*(1), Article diaa011. https://doi.org/10.1093/insilicoplants/diaa011

Thorp, K. R., DeJonge, K. C., Kaleita, A. L., Batchelor, W. D., & Paz, J. O. (2008). Methodology for the use of DSSAT models for precision agriculture decision support. *Computers and Electronics in Agriculture*, *64*(2), 276–285. https://doi.org/10.1016/j.compag.2008.05.022

Trnka, M., Rötter, R. P., Ruiz-Ramos, M., Kersebaum, K. C., Olesen, J. E., Žalud, Z., & Semenov, M. A. (2014). Adverse weather conditions for European wheat production will become more frequent with climate change. *Nature Climate Change*, *4*(7), 637–643. https://doi.org/10.1038/nclimate2242

Wallach, D., Buis, S., Seserman, D.-M., Palosuo, T., Thorburn, P. J., Mielenz, H., Justes, E., Kersebaum, K.-C., Dumont, B., Launay, M., & Seidel, S. J. (2024). A calibration protocol for soil-crop models. *Environmental Modelling & Software*, *180*, 106147. https://doi.org/10.1016/j.envsoft.2024.106147

Wallach, D., Goffinet, B., Bergez, J.-E., Debaeke, P., Leenhardt, D., & Aubertot, J.-N. (2001). Parameter estimation for crop models. *Agronomy Journal*, *93*(4), 757–766. https://doi.org/10.2134/agronj2001.934757x

Wallach, D., Makowski, D., & Jones, J. W. (2006). *Working with dynamic crop models: Evaluation, analysis, parameterization, and applications*. Elsevier.

Wallach, D., Makowski, D., Jones, J. W., & Brun, F. (2018). *Working with dynamic crop models: Methods, tools and examples for agriculture* *and environment* (3rd ed.). Elsevier. https://doi.org/10.1016/C2016-0-01552-8

Wallach, D., Palosuo, T., Thorburn, P. J., 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., Hoogenboom, G., Horan, H., … 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. https://doi.org/10.1016/j.eja.2020.126195

Wallach, D., Palosuo, T., Thorburn, P. J., 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., … Seidel, S. J. (2021). The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. *Environmental Modelling & Software*, *145*, 105206. https://doi.org/10.1016/j.envsoft.2021.105206

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*(3–4), 365–380. https://doi.org/10.1016/j.ecolmodel.2008.01.005

Yang, J. M., Yang, J. Y., Liu, S., & Hoogenboom, G. (2014). An evaluation of the statistical methods for testing the performance of crop models with observed data. *Agricultural Systems*, *127*, 81–89. https://doi.org/10.1016/j.agsy.2014.01.008

Zhao, C., Piao, S., Huang, Y., Wang, X., Ciais, P., Huang, M., Zeng, Z., & Peng, S. (2016). Field warming experiments shed light on the wheat yield response to temperature in China. *Nature Communications*, *7*, Article 13530. https://doi.org/10.1038/ncomms13530

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shawon, A. R., Attia, A., Ko, J., Memic, E., Uptmoor, R., Hackauf, B., & Feike, T. (2025). Impact of calibration strategy and data on wheat simulation with the DSSAT-Nwheat model. *Agronomy Journal*, *117*, e70111. https://doi.org/10.1002/agj2.70111

APPENDIX 1: List of global climate model/global circulation model-regional climate model (GCM-RCM).

GCM/RCM SL	GCM (RCP 4.5)	RCM (RCP 4.5)	GCM (RCP 8.5)	RCM (RCP 8.5)
1	ICHEC—EC-EARTH (r1)	KNMI-RACMO22E	ICHEC—EC-EARTH (r1)	KNMI-RACMO22E
2	ICHEC—EC-EARTH (r12)	KNMI-RACMO22E	CCCma-CanESM2 (r1)	CLMcom-CCLM4-8-17
3	ICHEC—EC-EARTH (r12)	SMHI-RCA4	MOHC—HadGEM-ES (r1)	CLMcom-CCLM4-8-17
4	MOHC—HadGEM-ES (r1)	CLMcom-CCLM4-8-17	MIROC—MIROC 5(r1)	GERICS-REMO2015
5	MPI-M—MPI-ESM-LR (r1)	MPI—CSC-REMO2009	MPI-M—MPI-ESM-LR (r1)	UHOH-WRF361H
6	MPI-M—MPI-ESM-LR (r2)	MPI—CSC-REMO2009	MPI-M—MPI-ESM-LR (r2)	MPI—CSC-REMO2009

Abbreviation: RCP, representative concentration pathway.

APPENDIX 2: Soil properties of experimental locations used for model application.

Location	Depth (cm)	Soil type	% Organic matter	Clay (%)	Silt (%)	Sand (%)
Feldkirchen	0–30	Silt loam	1.74	20	67.2	12.8
	30–35	Silt loam		14.5	69	16.5
	35–40	Silt loam		14.5	69	16.5
	40–110	Silty clay loam		28	63.9	8.1
	110–200	Silt loam		14.5	69	16.5
Thyrow	0–30	Loamy sand	0.87	6.5	16.3	77.2
	30–50	Loamy sand		6.5	16.3	77.2
	50–60	Sandy loam		14.5	23.3	62.2
	60–80	Loamy sand		6.5	16.3	77.2
	80–200	Loam		21	32.6	46.4