Diagnosing similarities in probabilistic multi-model ensembles: an application to soil–plant-growth-modeling

Aline Schäfer Rodrigues Silva · Tobias K. D. Weber · Sebastian Gayler · Anneli Guthke · Marvin Höge · Wolfgang Nowak · Thilo Streck

Received: 10 December 2021 / Accepted: 9 May 2022 / Published online: 19 June 2022
© The Author(s) 2022

Abstract
There has been an increasing interest in using multi-model ensembles over the past decade. While it has been shown that ensembles often outperform individual models, there is still a lack of methods that guide the choice of the ensemble members. Previous studies found that model similarity is crucial for this choice. Therefore, we introduce a method that quantifies similarities between models based on so-called energy statistics. This method can also be used to assess the goodness-of-fit to noisy or deterministic measurements. To guide the interpretation of the results, we combine different visualization techniques, which reveal different insights and thereby support the model development. We demonstrate the proposed workflow on a case study of soil–plant-growth modeling, comparing three models from the Expert-N library. Results show that model similarity and goodness-of-fit vary depending on the quantity of interest. This confirms previous studies that found that “there is no single best model” and hence, combining several models into an ensemble can yield more robust results.

Keywords Multi-model ensembles · Energy statistics · Model set visualization · Crop modeling

Introduction

Multi-model approaches
Multi-model ensembles have received increasing interest in crop-modeling over the last decade (Palosuo et al. 2011; Asseng et al. 2013, 2015; Martre et al. 2015; Wöhling et al. 2015; Yun et al. 2017; Makowski 2017; Wallach et al. 2018). While multi-model approaches can serve different purposes (Höge et al. 2019; Minka 2002), the main focus in the crop-modeling community has been to improve the accuracy of predictions (Asseng et al. 2015; Martre et al. 2015) or to estimate the uncertainty due to model choice, often referred to as conceptual uncertainty (Asseng et al. 2013). Please note that no multi-model method can quantify the conceptual uncertainty on an absolute level (Nearing and Gupta 2018). The reason for this is evident: In practice, there is no way to create and sample from an exhaustive list of all plausible

1 Department of Stochastic Simulation and Safety Research for Hydrosystems, Institute for Modelling Hydraulic and Environmental Systems/Cluster of Excellence “Data-Integrated Simulation Science”, University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany
2 Department of Biogeophysics, Institute of Soil Science and Land Evaluation, University of Hohenheim, Emil-Wolff-Straße 27, 70593 Stuttgart, Germany
3 Junior Research Group for Statistical Model-Data Integration, Cluster of Excellence “Data-Integrated Simulation Science”, University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany
4 Department of Systems Analysis, Integrated Assessment and Modelling, Eawag-Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, Dübendorf 8600, Zurich, Switzerland
models that cover the entire range of possible outcomes (Nearing and Gupta 2018; Vehtari and Ojanen 2012; Höge et al. 2019; Ferré 2017).

Therefore, Nearing and Gupta (2018) suggest understanding multi-model methods rather as sensitivity analyses. From this point of view, multi-model methods are tools that make modelers aware of how much predictions may differ within the model set depending on the choice for a certain conceptual model. This follows the line of thought of Ferré (2017) who introduced the idea of a multi-model ensemble as a “team of rivals”, which provides competing views of a system. If the competing models agree on a certain prediction, this increases the decision makers’ confidence, while disagreement indicates the need for further investigation (Ferré 2017).

When different models are merged to improve predictions, modelers hope for two effects that make the ensemble more skillful than its individual members: (1) the errors of the individual models cancel one another. This requires the individual models to be independent, i.e. different in their assumptions and conceptualizations (e.g. Abramowitz and Gupta 2008; Abramowitz 2010; Evans et al. 2013; Sanderson et al. 2015a, b; Abramowitz et al. 2018; Enemark et al. 2019). However, this assumption is often not met (Abramowitz and Gupta 2008; Abramowitz 2010; Bishop and Abramowitz 2013; Evans et al. 2013; Sanderson et al. 2015a, b; Knutti et al. 2017; Abramowitz et al. 2018). (2) The ensemble covers a broad spectrum of possible system behavior (Enemark et al. 2019) and thus compensates for over-confident individual models (e.g. Fritsch 2000; Doblas-Reyes et al. 2005; Weigel et al. 2008).

Various studies have compared the predictive performance of multi-model ensembles to the ones of the individual ensemble members (e.g. Krishnamurti et al. 2000; Georgakakos et al. 2004; Doblas-Reyes et al. 2005; Hagedorn et al. 2005; Weigel et al. 2008; Diks and Vrugt 2010; Arsenault et al. 2015; Yun et al. 2017; Christiansen 2018). They found the following aspects to be crucial for the performance of ensemble predictions:

1. the applied combination/averaging_WEIGHTING method (Krishnamurti et al. 2000; Doblas-Reyes et al. 2005; Diks and Vrugt 2010; Arsenault et al. 2015; Höge et al. 2019, 2020),
2. the performance measure used for assessing the predictive skills (Hagedorn et al. 2005; Weigel et al. 2008; Diks and Vrugt 2010),
3. the spread (variability) of the individual models (Fritsch 2000; Doblas-Reyes et al. 2005; Weigel et al. 2008), and
4. the similarity of the individual models regarding their structure and their predictions (Tebaldi and Knutti 2007; Abramowitz and Gupta 2008; Abramowitz 2010; Winter and Nychka 2010; Bishop and Abramowitz 2013; Evans et al. 2013; Sanderson et al. 2015a, b; Abramowitz et al. 2018; Enemark et al. 2019).

In the present study, we focus on the last two of these aspects: the similarity and the spread of the individual models. We show how the similarities of (or distances between) the models can be quantified while accounting for the spread in their predictions.

Referring to the often claimed superiority of ensemble predictions, Hagedorn et al. (2005) argue that the question of whether the ensemble has higher predictive skill than the best individual model is posed wrongly because there is often no “best” individual model. They show that it is usually not the same model that performs best considering all quantities of interest or under all conditions. Rather, what is typically identified as the “best” model looking at a particular aspect of the simulated system might be a weak model considering another aspect. Therefore, Hagedorn et al. (2005) conclude that the ensemble is superior to single models because its predictions are more robust, i.e. better over a broad range of predicted variables and modeling periods. We set up the present analysis accordingly, such that it enables a model comparison considering various quantities of interest.

**Ensemble predictions in crop modeling**

In crop modeling, a systematic assessment of multi-model ensembles was initiated within the Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al. 2013). In the wheat pilot study of that project, Asseng et al. (2013) compared 27 wheat models by analyzing the predicted grain yield under climate change conditions. They found that predictions vary significantly between different models. Thus, there is considerable uncertainty concerning model choice when predicting yields under climate change conditions. In an earlier comparison of yield predictions from eight crop models, Palosuo et al. (2011) also found great differences between the individual models. This study showed that none of the models was able to outperform its competitors across different environmental conditions and for different variables. In addition, the comparison of four crop models by Wöhling et al. (2015) had similar findings that support the argument of Hagedorn et al. (2005) that there is no “single best model”.

While Asseng et al. (2013) focused on the “end-of-season variable” grain yield, Martre et al. (2015) compared the same 27 models also regarding grain protein concentration and “in-season variables” (leaf area index, plant-available soil water, total aboveground biomass, total above-ground nitrogen, and nitrogen nutrition index), with all models being calibrated on phenology data. The authors found that the ensemble predictions are more reliable and attributed this improvement to the different process descriptions providing...
a wide range of plausible system behavior. The study also reports that some models had rather small errors for the end-of-season variables yield and grain protein concentration while showing large errors for in-season variables. Therefore, Martre et al. (2015) emphasize that for comparing the performance of different models, it is important to consider several variables as a model might perform well regarding a certain variable, but poorly regarding others.

Another ubiquitous issue when comparing models is calibration. In recent studies, Wallach et al. (2020) discuss the “chaos in calibrating crop models”, i.e. the lack of a unified calibration procedure in the crop modeling community. Martre et al. (2015) emphasize that for comparing the performance of different models, it is important to consider several variables as a model might perform well regarding a certain variable, but poorly regarding others.

Another ubiquitous issue when comparing models is calibration. In a recent study, Wallach et al. (2020) discuss the “chaos in calibrating crop models”, i.e. the lack of a unified calibration procedure in the crop modeling community. In an earlier study, Wallach (2011) stated that in the calibration of crop models, often model structural errors are compensated by specifying non-physical parameter values. As a result, a model might perform well for the quantity of interest it has been calibrated on, but poorly for others. This effect is more severe, the more parameters a model has (e.g. Jeffereys and Berger 1992; Lever et al. 2016). Therefore, Vogel and Sankarasubramanian (2003) recommend checking model adequacy in an uncalibrated state. We follow that recommendation and implement our analysis in a Monte Carlo framework, sampling prior parameter distributions. This allows us to evaluate the model performance independent of a specific parameter choice and we avoid to cloud model errors by assigning parameter values that compensate for structural deficiencies. As in any Bayesian framework, a subjective choice of prior distributions based on expert knowledge is needed. In fact, in the Bayesian setting, plausible ranges and assumed distributions are part of the model, just as a fixed parameter assumption (or the decision that a parameter is free for calibration) would be part of a model in deterministic modelling. Different choices among these options makeup different models.

Goal and approach of this study

The main goal of this study is to quantify similarities between probabilistic model predictions and visualize them intuitively. The proposed methods help modelers to gain deeper insights into the model set and to choose a suitable multi-model strategy accordingly. Our approach is to use a statistical metric, the so-called energy distance (Rizzo and Székely 2016) to quantify the (dis-)similarity between probabilistic model predictions and noisy measurements. Metrics, in general, are distance measures. Statistical metrics measure how close statistical objects such as probability distributions are, i.e. they take probability densities into account. We use the energy distance as a metric to compare predictive distributions that are generated by sampling from the prior parameter distributions of each model in a Monte Carlo framework. This enables us to take parametric uncertainty into account and to compare the models independent of a specific parameter choice. Thus, for comparing the models, we calculate the energy distance between their predictive distributions.

With the same method, we can also assess model performance by calculating the energy distance between model predictions and noisy observations. For this, we fit a probability density function to replicate measurements. From this distribution, we draw samples to calculate the energy distance in a Monte Carlo framework. If no distribution for the measurement errors can be defined (e.g. because no replicate measurements are available and no assumptions about measurement noise seem defendable), we can use the deterministic counterpart of the energy distance: The so-called energy score (Gneiting and Raftery 2007) compares probabilistic distributions to deterministic values and is directly related to the energy distance. This makes the concept of energy statistics widely applicable for rating probabilistic models.

The proposed method fulfills the following properties:

1. It can act on multivariate model predictions and thus reflect “overall” model characteristics.
2. It quantifies the similarity between pairs of models in the same way as the similarity between models and observations. Thus, it can assess the similarity between model predictions as well as model performance given observed data.
3. It can be used for comparing probabilistic model predictions to both noisy observations (by using the energy distance) and to deterministic observations (by using the energy score).
4. It acts on prior predictive distributions and thus accounts for parametric uncertainty in each model.

For guiding the multi-model process, we need an intuitive way to visualize the quantified similarities among the
models and measurements. Therefore, we suggest different methods for visualizing similarities among models, which highlight different aspects of similarity and, when combined, provide a detailed overview for interpreting the model set.

The paper is structured as follows: First, we present the mathematical methods, i.e. the “Energy distance” and the “Energy score”, and visualization techniques in the section “Visualizing predictive similarity”. Experimental data are presented in “Field experiments” and we introduce the model set in “Model description”. This is followed by “Results and discussion”. We summarize our findings and provide conclusions in the section “Summary and conclusions”.

Methods

Energy distance

In this section, we describe how the (dis-)similarity between two probabilistic models or between a probabilistic model and noisy measurements can be quantified and visualized based on Monte Carlo samples of the models’ predictive distributions.

Well-known distance measures like the Euclidean or Manhattan distance ($L_p$-metrics) are based on the coordinates of points in the Euclidean space. These distances do not take the density of probability distributions into account. In contrast, statistical distances (also known as probability metrics) measure the distance between two statistical objects such as random variables, probability distributions, or data samples (e.g. Deza and Deza 2016) and include information about probability densities.

Rizzo and Székely (2016) introduced the energy distance as a metric that measures the distance between two random vectors $X, Y$ in $\mathbb{R}^D$. It is called energy distance because of the analogy to the potential energy between objects (Rizzo and Székely 2016). It satisfies all axioms of a distance metric (non-negativity, identity of indiscernibles, triangle inequality) (e.g. Deza and Deza 2016). The squared energy distance $D^2$ between the distributions $F(X)$ and $G(Y)$ is defined as

$$D^2(F,G) = 2\mathbb{E}[|X - Y|_2] - \mathbb{E}[|X - X'|_2] - \mathbb{E}[|Y - Y'|_2] \geq 0,$$

(1)

where $X \sim F$, $y \sim G$, and $N_{\text{MC}}$ being the number of Monte Carlo samples.

Comparing Fig. 1a, d as well as (b) and (e) shows that keeping the same mean and increasing the variance of distribution $G$ decreases the energy distance $d(F, G)$ between both distributions. Subfigure (c) shows that for two identical distributions, the energy distance becomes zero, while the expected value of the Euclidean norm $\mathbb{E}[|X - Y|_2]$ is not equal to 0. Subfigure (f) illustrates the energy distance between two distributions with the same mean but different variances.

Energy score

When working with real and error-prone data, we do not have access to the full distribution of the data (i.e. a “true” value and a distribution function of errors) but only to the measured instances thereof, i.e. our observations. In some cases, these measurements suffice for estimating the underlying distribution reasonably well. If this is not the case (e.g. if there are only a few measurements available), we need an alternative for rating probabilistic predictions given deterministic measurements.

In deterministic modeling, the performance of a model is usually evaluated by an error measure between the model’s best estimate $\hat{y}_k$ and the observations $y_{\text{meas}}$. Different models are then rated based on the achieved best estimate error, e.g. a root mean square error (RMSE) or the mean absolute error (MAE).

In probabilistic modeling, model rating is based on so-called scoring rules (Gneiting and Raftery 2007). These scores account for the entire predictive distribution of the model instead of only the best estimate. Many different scores exist (Gneiting and Raftery 2007; Yao et al. 2018). Among these, the energy score is directly related to the above-introduced energy distance (Székely and Rizzo 2013), i.e. it resembles the one-sided version of the energy distance (Ziel and Berk 2019). The energy score $\text{ES}$ for the model predictive distribution $G$ and observations $y_{\text{meas}}$ writes as:

$$\text{ES}(G, y_{\text{meas}}) = \frac{1}{2} \mathbb{E}[|Y - Y'|_2] - \mathbb{E}[|Y - y_{\text{meas}}|_2].$$

(3)
with $\beta \in (0, 2)$. We choose $\beta = 1$ as it is a standard choice for distributions that are not heavily tailed (Ziel and Berk 2019). For $\beta = 2$, the energy score is equal to the negative squared error (Gneiting and Raftery 2007).

In cases we cannot assume a reasonable distribution based on replicate measurements, we will use the energy score instead of energy distance. We want both quantities to act on the same scale, so they are directly comparable. Therefore, we use $d(F, G) = \sqrt{D^2(F, G)}$ and $\sqrt{ES}$ for the analysis.

Visualizing predictive similarity

We want to visualize the (dis-)similarity of the model predictions to get an intuitive understanding of the diversity in the considered model set. At the same time, we want to visualize how well the predictions match the measurements. Therefore, we treat both the models and the measurements as objects in a common model predictions-observations space, which we call “quantity of interest space”. Representing the similarities of $N$ objects (models and observed data) leads to $n_{comb} = N \cdot (N - 1)/2$ combinations. While in our application, this number (three models and one measurement data set, hence, six combinations) is comparatively small, visualization of model similarity in two dimensions is already not a straightforward task. Clearly, the number of models to be compared can become much higher in extensive multi-model ensembles. Therefore, the methods we propose for visualization are also suitable for larger model sets.

Each of these objects (models and observed data) consists of $n_{qoi}$ variables (quantities of interest). In the case of probabilistic modeling, each variable is assigned a probability distribution. Therefore, we have to deal with high-dimensional data, and regarding its two-dimensional visualization, we have to balance the interpretability and the preservation of the original structure in the applied projection (Liu et al. 2017), which is a typical problem in the visualization of high-dimensional data.

We make use of different techniques for visualizing the similarity of two objects (model-model, model-data) under different conditions. Each visualization method highlights a different aspect, so which method is the most insightful one depends on the specific question we ask about the model set:

1. Heatmaps: In a matrix, the distances between all pairs of objects are visualized through varying colors or intensity (e.g. Nandi and Sharma 2020).
2. Radar charts: Several axes are arranged radially starting from a common center. Each axis represents a certain quantity of interest, i.e. a different variable or the same variable under different boundary conditions. Each value (here, the distance between two objects) is plotted along one axis. This is repeated for all axes. Finally, all values are connected to a polygon, representing one object (e.g. Nandi and Sharma 2020).
3. Dendrograms: Dendrograms are tree-like diagrams that are typically used for visualizing hierarchical structures. A dendrogram consists of branches that connect objects.
depending on their similarity. The height at which two objects are joined together represents the distance between these objects (e.g. Nandi and Sharma 2020). For creating such a diagram, we use an agglomerative hierarchical clustering approach (Xu and Wunsch 2008): An algorithm identifies pairs of clusters with minimal distance in-between and merges them. This merging procedure is repeated until all data points are finally in one overarching group (Xu and Wunsch 2008). The merging depends on the chosen linkage method, i.e. the definition of the distance between two clusters. For the present study, we chose a linkage that uses the average distance between data points in two clusters.

Case study description

Central to our study is the simulation of wheat growth, energy and water fluxes in six agricultural fields in two regions during several years (2010–2015). These fields feature slightly contrasting meteorological conditions at two different sets of sites (1–3 and 4–6), with soils being only similar in sites 1–3. The study is based on an extensive data set enabling coupled soil–plant-growth modeling and comparing simulation results to high-quality measurements. The field data is summarized in “Field experiments” and the participating models of the ensemble in “Model description”.

Field experiments

The data set used in this study is a subset of a multi-site, multi-year, and multi-crop data set that contains extensive characterization of soil properties and states, plant growth and yield, management, and soil–atmosphere fluxes of energy, water and carbon dioxide. It was obtained from measurement campaigns in intensively managed agricultural fields of local farmers. We will limit the description of the data set to a minimum since it was published alongside a manuscript with full methodological details (Weber et al. 2021).

The data were collected between May 2009 and September 2018. In this study, we use a subset that covers the sites and years in which winter wheat was cultivated from 2010–2015 (year of harvest). The combination of a site at which and a year during which winter wheat was cultivated are reported as site-years. For example, winter wheat grown at site 1 from November 2014 to July 2015 is denoted as site 1, year 2015. In total, we analyze data of 14 site-years. Details of the two regions can be found in Weber et al. (2021), and their soil properties are summarized in Table 1.

All required meteorological forcings were measured at half-hourly time intervals and data gaps were filled. In the data set, grain yield was reported both by the farmer as a field average and by extrapolation of the plot sampling to the field. Phenology and leaf area index were measured at least biweekly during the main vegetation season (April to mid July).

| Site | Depth | No. of simulation layers | Organic matter (g g⁻¹ · 100) | Clay | Silt | Sand | Bulk density (g cm⁻³) | θr | θs |
|------|-------|--------------------------|-------------------------------|------|------|------|----------------------|----|----|
| 1    | 0–30  | 6                        | 1.75                          | 18.2 | 79.4 | 2.5             | 1.37                         | 7.5 | 46.0  |
|      | 30–45 | 3                        | 0.61                          | 18.8 | 79.2 | 2.0             | 1.51                         | 7.2 | 42.9  |
|      | 45–165| 24                       | 0.42                          | 18.7 | 80.4 | 0.9             | 1.48                         | 7.3 | 43.9  |
| 2    | 0–30  | 6                        | 1.53                          | 17.9 | 79.5 | 2.6             | 1.33                         | 7.6 | 47.0  |
|      | 30–45 | 3                        | 0.52                          | 20.1 | 77.0 | 2.9             | 1.46                         | 7.5 | 43.9  |
|      | 45–165| 24                       | 0.34                          | 18.7 | 79.7 | 1.6             | 1.53                         | 7.1 | 42.5  |
| 3    | 0–30  | 6                        | 1.64                          | 17.1 | 81.1 | 1.8             | 1.37                         | 7.4 | 46.2  |
|      | 30–45 | 3                        | 0.83                          | 18.7 | 80.4 | 1.0             | 1.50                         | 7.3 | 43.4  |
|      | 45–165| 24                       | 0.63                          | 16.1 | 83.0 | 0.8             | 1.51                         | 6.9 | 43.2  |
| 4    | 0–20  | 4                        | 4.35                          | 37.8 | 56.0 | 6.2             | 1.31                         | 9.5 | 49.2  |
|      | 20–30 | 2                        | 2.13                          | 38.6 | 52.5 | 8.9             | 1.34                         | 9.4 | 48.2  |
|      | 30–50 | 4                        | 1.63                          | 48.4 | 43.3 | 8.4             | 1.32                         | 10.0| 50.0  |
| 5    | 0–20  | 4                        | 3.64                          | 28.9 | 68.3 | 2.8             | 1.37                         | 8.6 | 46.8  |
|      | 20–60 | 8                        | 1.44                          | 33.6 | 64.3 | 2.1             | 1.40                         | 8.9 | 46.7  |
|      | 60–165| 21                       | 0.71                          | 34.2 | 64.0 | 1.8             | 1.51                         | 8.7 | 44.0  |
| 6    | 0–30  | 6                        | 5.50                          | 45.6 | 51.2 | 3.2             | 1.04                         | 10.7| 59.0  |
|      | 30–45 | 3                        | 3.88                          | 47.6 | 48.3 | 4.1             | 1.29                         | 10.1| 51.3  |
Model description

The relevant processes for crop development and growth, unsaturated water flow, nitrogen and carbon turnover in the soil, evapotranspiration, and drainage water quantity and quality were simulated with the multi-model library Expert-N (Priesack 2006). Expert-N is a model system that facilitates a high degree of flexibility in selecting competing model formulations for the relevant processes in the soil-vegetation-atmosphere continuum. An example of a 48 member multi-model ensemble using Expert-N is the study about climate change impact on wheat and maize yield development in Ethiopia by Rettie et al. (2022). That study uses a model ensemble consisting of 48 unique model members set up in Expert-N. For the presented study we selected three different plant growth models within Expert-N: CERES (Ritchie et al. 1988), SUCROS (van Laar et al. 1997), and SPASS (Wang and Engel 2000; Gayler et al. 2002), which are coupled to the soil carbon and nitrogen turnover and transport models SOILN (Johnsson et al. 1987) and LEACHN (Hutson and Wagener 1995), and the Richard-son–Richards equation for variably saturated water flow as implemented in HYDRUS-1D (Smunek et al. 1998). Fluxes of heat and dissolved nitrogen in the soil were described by LEACHN and potential evapotranspiration as calculated by the Penman–Monteith equation (Allen 1998) modulated by crop coefficients. These models are described in “Pheno-
logical development”– “Soil nitrogen”. The model initial and boundary conditions, including a description of the uniformly distributed and bounded model parameter priors for sampling the prior predictive distribution, are given in “Process models”. In the following, SUCROS, CERES, and SPASS are described and refer to the versions implemented in Expert-N v.3.1. Model parameters and priors are listed in Tables 5, 6, and 7.

Phenological development

In Expert-N, the phenological development (BBCH) is modeled as a parametric function of thermal time, vernalization, photoperiod effect, and temperature sensitivity. While CERES is differentiated into nine development phases, SPASS and SUCROS are differentiated into three. All three adopted models distinguish the vegetative growth phase (BBCH ≤ 60) from the generative phase (BBCH > 60) and one for emergence. CERES (Ritchie and Godwin 1989; Jones 1986) and SUCROS (Spitters et al. 1989; Van Laar et al. 1992) are widely established models and SPASS is a combination and development out of the other two (Wang 1997; Gayler et al. 2002). Internally, phenological development is first simulated as a development stage on a scale from 0.5 to 2.0, and is subsequently converted to an externally reported BBCH variable using fixed lookup tables with 10 support points (11 in the case for SPASS). The support points in the lookup tables were considered as fixed. The simulated phenology acts as a boundary condition for the remaining part of the dynamic plant growth model, by setting the precondition after which certain other parts of the model are active (i.e. triggering submodules for, e.g. leaf area index or grain filling after anthesis at BBCH = 60). An important difference between the models is that winter wheat requires vernalization, which is the induction of flowering after a cold period. In contrast to CERES and SPASS, which contain routines for vernalization. Since SUCROS does not include vernalization, it is strictly speaking, a spring wheat model.

Root growth, root water uptake, and transpiration

Dynamic root growth in all plant models is simulated by roots growing downward up to a maximum root extension depth. The maximum growth rate is reached under optimum conditions. This is modulated by impacts of unfavorable environmental factors (temperature, soil moisture) in the layer of the currently greatest root extension. Specifically, these impedances are functions of temperature and low soil moisture (SPASS, SUCROS), or of low soil moisture and low mineral nitrogen contents (CERES). In each simulation layer, the active roots are the balance between root growth and senescence at each time step (Gayler et al. 2013). To calculate root water uptake, a root length density is required. CERES and SPASS use an identical approach: the root length growth rate is linearly related to the root biomass growth rate, and the vertical distribution is related to water and nitrogen availability in the respective soil layers. In SUCROS, the root length growth rate is derived for each simulation layer based on a crop-specific root depth distribution function, the root biomass, and a specific root length. The upper limit of root water uptake is limited to the potential transpiration as calculated by the Penman–Monteith equation (Allen 1998). In the case of SPASS and CERES, a maximum root water uptake rate per root length is additionally defined. All three models use a macroscopic approach in which root water extraction is distributed to the individual simulation layers proportionally to the relative root length in the layers, as long as the water supply is optimum. Impedance factors such as oxygen deficiency in (near-)saturated soils, soil compaction and structure, disease and pests, adverse chemical conditions (e.g. salts) are not considered in the models. To account for crop effects on potential evapotranspiration, all three models use crop coefficients, which we modeled as a piece-wise phenology-state dependent function with three parameters $k_{c\text{ini}}(-)$, $k_{c\text{mid}}(-)$, and $k_{c\text{end}}(-)$, which are the crop coefficient for the initial, mid and end of the vegetation period, respectively. These
parameters were considered uncertain with the uniform distribution in the ranges given in the Appendix, Table 4.

**CO₂ assimilation, biomass growth and leaf area development**

CERES adopts a robust “big leaf” approach to calculate carbon assimilation, using empirical adjustments to account for the depth-dependence of photosynthetic capacity and light response, with the interception of photosynthetically active radiation dependent on leaf area index. Biomass development depends on the partitioning of assimilates to different plant organs (roots, leaves, stem, and fruit). It is achieved through potential growth rates and a priority scheme for the allocation of assimilates to each organ (differentiated into five developmental stadia). In the juvenile phase, i.e. in the stadium between the emergence and, the development of the first apical spikelet, the leaf area index develops exponentially, and after the juvenile phase, leaf area develops proportionally to the leaf biomass development, depending on temperature, and water and nitrogen stress. In SUCROS and SPASS, the calculation of carbon assimilation is based on a multi-layer approach, which is more comprehensive compared to CERES. The aim of this approach is to differentiate between sunlit and shaded leaves and to account for the attenuation of direct and diffuse radiation. The two models differ in vertical resolution as SUCROS uses a three-layer approach, and SPASS uses five layers, but the calculation of leaf internal CO₂ concentration and net photosynthesis is similar, with small differences in the calculation of water stress and nitrogen response functions. In contrast to CERES, in SPASS and SUCROS carbohydrate allocation and hence organ growth follows an assimilate-partitioning scheme, which is fixed at optimum water supply and is determined solely by the development stage of the plant. However, in the case of water deficiency, root growth is favored in both models to counteract the cause of stress. In SUCROS, the leaf area growth is directly coupled to leaf biomass growth rate, whereas in SPASS, leaf area growth rate does account for water and nitrogen deficiency. More detailed presentations can be found in Priesack (2006), Priesack and Gayler (2009), Biernath et al. (2011), Wöhling et al. (2013) and references therein.

**Soil hydrology**

In Expert-N, the standard process model for simulating variably saturated moisture fluxes in soils is the Richards–Richards equation (RRE) (Richardson 1922; Richards 1931). The solution of the RRE requires parameter functions to describe the soil hydraulic properties. Since we simulated root water uptake using a macroscopic approach (van Dam et al. 2008), it is sufficient to parameterize the RRE using the van Genuchten Mualem model (van Genuchten 1980). There would be physically more comprehensive soil hydraulic property models that account for water storage and conductivity in medium to dry soils (Weber et al. 2019; Streck and Weber 2020; Weber et al. 2020). These would influence the simulation of actual transpiration under water-stressed conditions only when using microscopic (not macroscopic) root water uptake models (van Dam et al. 2008), providing hydraulic uplift does not influence the simulation.

For each site, a top-soil/sub-soil differentiation was made, each with a different parameterization. The varied parameters per soil layer are $a$ (cm$^{-1}$) and $n$ (−), which are the van Genuchten shape parameters, the saturated hydraulic conductivity $K_s$ (cm d$^{-1}$), and the tortuosity parameter $τ$ (−), and we fixed $m = 1 – 1/n$. Instead of varying the saturated and residual water contents, $θ_s$ and $θ_r$, respectively, we vary the soil water content profile set as an initial condition. The soil model is discretized into simulation layers of 5 cm depth (see Appendix, Table 1). Here, in contrast to the observed soil profiles, we reduce the number of simulated soil horizons for the soil hydrological part to two by merging the second and third horizon at sites 1–5. The differentiation into more horizons is pedologically founded, but for the modeling purpose of this study not parsimonious, i.e. we group horizons with very similar hydraulic properties.

**Soil nitrogen**

Mineralization, nitrification, and denitrification are modeled following the SOILN approach, while urea hydrolysis, volatilization and dissolved nitrogen transport are modeled using the LEACHN approach.

The model concept of SOILN differentiates three pools of organic nitrogen representing the three different pools of organic carbon, available to the soil microbes. These pools are termed ‘litter’, ‘humus’, and ‘manure’. The ‘litter’ pool, a pool with fast turnover rates, represents fresh organic matter and microbial biomass. The ‘humus’ pool with a slow turnover of soil organic carbon, and the ‘manure’ pool, which represents the organic fertilizer. There are two essential assumptions of this model concept: (i) the N demand for the internal carbon cycle is governed by a constant C/N ratio of 10:1 in the microbial biomass and in the humus pool, and (ii) mineral nitrogen that is released or assimilated by the microbial biomass, follows this ratio. We varied the rate constants $k_{\text{miner.man}}$ (d$^{-1}$), $k_{\text{miner.lit}}$ (d$^{-1}$), $k_{\text{miner.hum}}$ (d$^{-1}$), which govern the rate of mineralization of the manure, litter, and humus pools. We also varied the rate coefficients for the nitrification $k_{\text{nitr}}$ (d$^{-1}$) and denitrification $k_{\text{nitr.lit}}$ (d$^{-1}$), where, analogously to the treatment of the soil hydraulic properties, we model the two horizons with different sets of kinetic rate constant parameters, except for $k_{\text{nitr.man}}$, which we set to 0 for the sub-soil. The rate constant for the urea
hydrolysis $k_{urea,hi}$ is a constant 0.36 (d$^{-1}$). The effectivity of decomposition $\alpha_e (-)$ describes the fraction of carbon that is re-immobilised after decomposition and set to a constant 0.45. The humus development constant $f_h$ (-) describes the fraction of decomposed litter that is added to the humus pool and was set to 0.2.

**Model setup and implementation**

**Process models**

As an upper boundary, we generally use in situ measured daily aggregated atmospheric temperatures (minimum, mean, and maximum), global radiation, wind speed, and precipitation. Potential evapotranspiration fluxes were calculated based on the Penman–Monteith approach. For the solutes, we use flux boundary conditions prescribing constant atmospheric NH4 deposition, and the timing, frequency, type, and amount of fertilizers. At the lower boundary, we used free drainage for the water flow module and a zero gradient for the solute and heat flux modules. Field management in terms of nitrogen input by fertilizers, and sowing and harvest dates were set to the farmer-reported data.

The selection of which parameters are considered as uncertain priors, and their respective ranges and distributional assumptions was guided by both model system and expert knowledge. Details are listed in Tables 4, 5, 6 and 7. We ran simulations based on $n_{MC} = 10,000$ parameter vector realizations generated by Latin Hypercube sampling per site-year and model. The gained forward simulation results resemble approximations to the prior predictive distributions for each of the 14 site-years and each of the plant models. This resulted in a total of 420,000 individual simulations performed on the High-performance Cluster bwFOR of the Federal State of Baden-Württemberg. The varied model parameters are listed in the Appendix in Tables 4, 5, 6 and 7.

**Similarity analysis**

We analyze the similarity of probabilistic predictions of CERES, SUCROS, and SPASS (Priesack 2006) via the energy distance between the predictive distributions of the models for different variables. The analyzed variables are yield, phenology, and leaf area index. For comparing model predictions and measurements, we use either the energy distance or the energy score as discussed in “Energy distance” and “Energy score”:

1. **Fitting a distribution to the observations and using the energy distance:** In the case of yield predictions, we can reasonably assume a Gaussian measurement error and, hence, define a distribution for the observations. Therefore, we can use the energy distance not only for the pairwise comparison of the models’ distributions among each other but also for the comparison with the distribution fitted to the observations.

2. **Using the median of the observations and the energy score:** In the case of the other two variables, leaf area index and phenology, making assumptions about measurement error and fitting a distribution to the observations is not as straightforward. Instead, we take the median of the measurements and use the energy score to compare the models and the observations. For the comparison of the models among each other, however, we still use the energy distance because the predictive distributions are available.

Please remember that similarities quantified by the energy score and the energy distance are on the same scale, and hence directly comparable.

We analyze data from six sites and up to three years per site to check how model performance and similarity vary under different conditions.

**Results and discussion**

First, we compare the models’ similarity (among each other) and their performance (i.e. similarity to observations) based on the end-of-season variable yield. Later, we analyze in-season variables to gain more insight into the processes that may have led to differences in the final yield predictions.

**Analysis of the end-of-season variable yield**

For yield, we assume a Gaussian measurement error, describing the distributions based on the replicates’ mean and variance. The resulting distributions are shown in Figs. 2 and 3. Based on Monte Carlo samples of these distributions, we calculate the energy distance between models and observations.

**Predictive distributions**

Figures 2 and 3 depict the distributions for yield predicted by the three models and the measurements for the sites 1–3 and 4–6, respectively. The probability density functions represent the prior model predictions, i.e. the models have not been calibrated and represent the full range of plausible parameters as defined in Tables 5, 6 and 7.

The mean observed values for yield range from 7.0 t/ha (site 2, 2013) to 9.2 t/ha (site 4, 2014), the corresponding standard deviations range from 0.39 t/ha $\approx$ 5% (site 2, year 2011) to 2.1 t/ha $\approx$ 23% (site 6, year 2014).

The probability density functions show that the predictions made by SPASS (yellow) have the highest variance for
Fig. 2 Probability density functions of the yield predicted by the three models and observed (gray), sites 1–3. For better visualization, y-axis scales are not the same across all sub-plots.

Fig. 3 Probability density functions of the yield predicted by the three models and observed (gray), sites 4–6. For better visualization, y-axis scales are not the same across all sub-plots.
all sites and years and it predicts ranges of very low yield with a higher probability than CERES (blue) and SUCROS (red) do. For most cases, SUCROS shows the smallest variance in the yield predictions and, from the visual impression, the best goodness-of-fit to the measurements (gray).

**Energy distance-based similarity analysis**

To get a more aggregated and objective comparison of the predictive distributions of all models and the data, we quantify their similarity according to their prior predictive distributions using the energy distance (Eq. 1). In the following, we discuss corresponding visualizations with radar charts, dendrograms, and heat maps.

**Visualizing model similarities using radar charts**

Figure 4 shows four radar diagrams that represent the similarity of the models and the observations, each centered on one of the models, or the observations, respectively. In each chart, each model is represented by points that are connected across all radial axes. Each axis represents one site-year. The closer a point to the center, the lower the energy distance between the respective distributions, i.e. the more similar the distributions.

The background color of each segment is color-coded according to the weather conditions of the respective site-year. For this color-coding, we calculated the ratio of mean precipitation and mean temperature from April to June $p/T$ and represent low values as red and high values as blue. This color-code is useful to investigate if there is any obvious relationship between weather conditions and model performance or similarity.

From Fig. 4a we can see that SUCROS is closest to the observed data for most site-years. In parts, this can be explained by SUCROS showing the lowest variance in the yield predictions, while being reasonably centered on the observations (see Figs. 2, 3).

Comparing model performances for sites 1 and 2 in 2013, we notice that CERES and SPASS are much closer to the data for site 1 than they are for site 2. Considering that the conditions at both sites are very similar, it may be surprising that the models perform so differently. The model predictions for both sites are indeed highly similar. However, the observations’ variances differ considerably between both sites (see Fig. 2). Please recall the property of the energy distance that decreasing the variance of a distribution while keeping the same mean increases the energy distance between two distributions (see Fig. 1). This effect is clearly visible in the case of model performances for site 1 and 2 in 2013.

Focusing on the weather conditions during the growing season, we see that all models perform relatively poorly for site 5, 2010, which was a rather wet year at this site. Here, SPASS is slightly closer to the measurements than the other two models. This is due to the high variance of its predictions (see Fig. 2), which leads to a larger overlap of the predictive distribution with the measurement distribution, even though the modes of all three predictive model distributions are relatively similar, all overestimating the yield. For the other wet year 2013, we cannot observe a similarly poor performance of all models. Only the CERES predictions of site 2, 2013 show a rather high distance to the observations again overestimating the measured yield (see Fig. 2). However, for the relatively dry conditions at site 3 in 2010, CERES shows a similar distance as it again overestimates yield (see Fig. 2). Hence, in the current data set, the poor performance of the models cannot be clearly explained by specific weather conditions.

Next, we check whether any of the models perform above its average during certain weather conditions: CERES performs well for site 6 in 2014, which was a rather dry year. SUCROS shows equally good performance under both dry and wet years. SPASS performs comparatively well for site 4 in 2011, which was the second driest year in our data set. However, this is not true for other dry years in this data set.

In summary, we cannot observe a specific pattern of good or poor model performance under certain weather conditions. This result indicates that the data do not provide evidence for identifying systematic mispredictions based on climate or site. An exception is site 5, which, however, only has one replicate year, so that no general statement can be made.

In Fig. 4b–d, we focus on the similarity between the models. When CERES is in the center of the radar chart (subfigure (b)), we can see the high discrepancy between CERES and SPASS for site 3, 2014, site 1 and 2, 2015, site 3, 2010 (rather dry years). The highest distance between CERES and SUCROS occurs for site 1 and 2, 2013 (wet years). When SUCROS is centered (subfigure (c)), it is apparent that SPASS is closer, i.e. more similar for most site-years. Subfigure (d), with SPASS being centered does not contain new information that has not yet been shown in figures (a)–(c) and is only shown for the sake of completeness.

**Visualizing model similarities using dendrograms**

In Fig. 5, an alternative visualization of model similarities based on dendrograms is presented. Here, the data of sites 1–3 are shown, the corresponding Figure for sites 4–6 can be found in the appendix (Fig. 11). Please note that the order of the models is not necessarily the same for all dendrograms. Rather, this so-called leaf order was optimized such that the sum of the similarities between adjacent leaves is maximized (Novoselova et al. 2015).

From the way models and observations are merged into clusters and from the height at which two objects are joined together in the dendrograms, we can intuitively see their similarity.
As can be seen from Figs. 5 and 11, different clusters are formed for different site-years. This shows that both, model similarity and goodness-of-fit, vary depending on the site-years.

Visualizing model similarities using heatmaps
In the heatmaps shown in Figs. 6 and 12, small values are represented by light colors and large values by dark colors. To dissect the individual components that the energy distance consists of, we plot its constituent parts $E[||X - Y||_2]$ and $E[||X - X'||_2]$ separately. The main diagonal entries represent the spread within the predictive distribution of a single model $E[||X - X'||_2]$, while the off-diagonal entries represent the similarity of two objects $E[||X - Y||_2]$.
From the color-coding, we can intuitively see that the highest dissimilarity among the models occurs for site 3 in 2010 and 2014, as well as for sites 1 and 2 in 2015, as these heatmaps are overall darker than the others. In the same manner, it is immediately obvious that SPASS clearly differs from the other models and the measurements. This is neither visible in Fig. 4 (radar charts), nor in Fig. 5 (dendrograms). The last columns/rows show the goodness-of-fit to the observations. Here, it is clearly visible that SUCROS performs best and SPASS performs worst for most site-years.

Comparison of the visualization methods

Table 2 summarizes the properties of the three visualization methods and their applicability for different use cases.

With heatmaps and dendrograms, all objects (here: models) can be compared at one glance, however, only for one condition (here: site-year). Therefore, for comparing many models under specific conditions (e.g. per site-year), heat maps and dendrograms are suitable. In contrast, radar charts are useful for comparing one object (e.g. measurement data) to a small number of other objects (e.g. models) under many different conditions (e.g. site-years).

Dendrograms make it easy to identify clusters. Such clusters can, for example, indicate settings in which all models are similar, but are far from measurements. Such a case may point to the fact that an important process was not considered in any of the models. Examples for such a setting are the yield predictions for site 5, year 2010 and site 4, year 2014 (see Appendix B1). Of course, we can also see from the density functions in Fig. 3 or the radar chart centered on the observations (Fig. 4) that none of the models fits the measurements for this site-year well. However, the analysis based on density functions is only trivial in 1D cases such as the exemplary yield predictions used here. Imagine if we wanted to compare higher dimensional predictions. In such cases, this task would be much easier using dendrograms based on probability metrics. Also the analysis based on radar charts would be less convenient for identifying clusters: we can tell from the radar chart centered on the observations that all models are far off. However, we cannot tell based on the same radar chart whether the models are close to each other. To get this insight, we would have to analyze several radar charts with different models being centered.

Summary regarding the Expert-N model set

Summarizing the comparison of the Expert-N model set, SUCROS performs best in predicting yield based on the similarity of its prior predictive distribution and the distribution of the observations. The highest similarity among the models is between SPASS and SUCROS, while the biggest differences appear for SPASS and CERES. Our analysis confirms that yield predictions vary significantly between different models as Asseng et al. (2013) and Palosuo et al. (2011) found in earlier studies.
Using the energy distance to aggregate the information and visualizing it with radar charts, dendrograms, and heat maps have been found insightful when inspecting influencing factors such as different sites or different weather conditions.

**Analysis of the in-season variable phenology**

After analyzing the end-of-season variable yield, we now focus on time series of the in-season variables phenological development stage (BBCH) and leaf area index (LAI, “Analysis of the in-season variable leaf area index”). Exemplary plots of the BBCH time series for site 1, year 2011 are shown in Fig. 7. The corresponding plots for all site-years are provided in the Appendix (Fig. 25).

We start with a qualitative analysis of the time series in “Predictive distributions”. Next, we quantify how similar or different the models behave at each daily time step by calculating the energy distance between the predictive distributions in “Energy distance-based similarity analysis”. We compare this between-model distance to the spread within
each model. This spread is calculated as the square root of
the mean Euclidean distance between all samples of the pre-
dictive distribution (see Eq. 1).

In contrast to the analysis of the yield predictions, we do not
assume a distribution for the measurement errors for BBCH or
LAI. Therefore, we use the median of the replicates and calcu-
late the energy score, i.e. the counterpart of the energy distance
for comparing distributions to a single observation.

Predictive distributions

From Fig. 7 we can see that the prior predictive distribu-
tions for the development stage generated by SPASS show
a very small variance until January, whereas the predictions
of CERES and SUCROS initially have little variance, but the
spread increases already in November. This can be observed
for all site-years (see Fig. 25).

Comparing the predictions to the measurements shows
that early BBCH stages are usually overestimated by all
three models. Starting approximately in March, when the
development stage reaches values of 20, the predictions
become more accurate.

The mean predictions of SPASS are closest to the meas-
urements for most site-years, while particularly CERES,
and to a lesser extent SUCROS, overestimate earlier phen-
ological development. The discernible steps in the SPASS
simulations are a direct result of and consistent with the
model structure: in contrast to the two other models, which
represent the secondary growth stages as fractions of the
temperature sums required for each principal growth stage
(BBCH = 10–20, 20–30, 30–40, …), SPASS simulates the
secondary growth stages during the early development
(BBCH = 10–40) based on the number of emerged main
stem leaves (BBCH = 11, 12, …), tillers (BBCH = 21, 22,
…), and main stem nodes (BBCH = 31, 32, …) Wang and
Engel (1998). Therefore, in this model simulated BBCH may
not be a continuous function of time. For example, if only
five main stem leaves have unfolded by the day on which
the principal growth stage “tillering” (BBCH = 20) has
been reached, a discontinuity from BBCH = 15 to BBCH =
20 would be simulated. The model behaves similarly with
respect to the number of tillers on day of principal growth
stage “stem elongation” (BBCH = 30) and the number of
nodes at principal growth stage “booting” (BBCH = 40).

We note that the better predictions of BBCH by SPASS are
in contrast to the worst performance in yield. We can
put these results into perspective with the yield predic-
tions. While our belief about plausible parameter ranges
of SUCROS and CERES led to an early onset of BBCH
development, on average, the grain filling period from anthe-
thesis (BBCH = 60) until maturity (BBCH = 90) is longest.
The price of this is a worse match in BBCH, in contrast
to SPASS. While SPASS reaches maturity approximately
similarly as SUCROS, it is the shorter grain filling duration
in SPASS that results in a tendency for lower yields, shown
by the heavy tails on the left of the predictive distribution
functions (see Figs. 2, 3).

For the stated reasons, this delayed BBCH development
(i.e., not-achieved maturity, leading to lower grain yields)
does not occur in CERES. From this insight, we can update
our formulation of plausible parameter ranges by ensuring
that the grain filling rate in the SPASS model parameters is
increased. Similar updates are possible for SUCROS and
CERES, where a delayed BBCH would have been matched
with an increased grain filling rate. Since, in practice, fertili-
dation dates are co-informed by BBCH, this result is of great
significance to enhance the prior predictive capabilities of
the models for these types of environments.
Energy distance-based similarity analysis

Figure 8 (a) shows time series of the spread of each model, (b) the energy distance between pairs of models, and (c) the energy score between models and observations. The corresponding figures for all other site-years are provided in the Appendix (Figs. 19, 20, 21, 22, 23 and 24).

From Fig. 8 we can see phenomena that can be observed in most site-years: Within-model spreads of phenology predictions (Fig. 8a) show an increasing trend until June and decrease steadily thereafter. This is to be expected since all models aim at reaching fill plant maturity by the time of harvest. While the curves for CERES (blue) and SUCROS (red) are similar, the one representing SPASS (yellow) is often shifted towards a later maximum in July and therefore shows the highest spread in its predictions at harvest date. From this, we can identify that the largest predictive uncertainty occurs around BBCH = 60, although it is one of the most important predicted stages. Within all models, the anthesis date at BBCH=60 is very important, as it marks the point at which grain filling starts. In principle, a very late start and short grain filling period (visible as steep slopes in the curves after BBCH=60 in Fig. 7) can be compensated with implausibly high grain filling rate parameters, such that reasonable yields can nevertheless be simulated. This can be achieved in all of the models.

In Fig. 8b, we can analyze the distance, i.e. dissimilarity between pairs of models. The distance between CERES and SPASS (green curve) is the highest during most phases for all site-years. After reaching the maximum distance in May, it decreases again. The distance between SUCROS and SPASS (orange curve) shows a similar development, however, the distance between these models is smaller most of the time. The distance between CERES and SUCROS (purple curve) does not show this characteristic maximum in May, rather it increases more or less steadily over the season and thus, the difference between CERES and SUCROS becomes the highest at the harvest date.

The curves in Fig. 8c can be interpreted as the models’ goodness-of-fit to measurements. A low energy score means a low distance to the observations and hence a good model performance. The time series of CERES and SUCROS show similar behavior with a maximum energy score in May to June, followed by a decline. For most site-years, CERES has the highest energy score. Although in some cases SPASS starts with the highest energy score, for most site-years, it is closest to the observations during most phases. Hence, from a model selection perspective, SPASS would be considered best overall. Nonetheless, the analysis with the energy score clearly shows that, especially in certain time windows, employing additional alternative models increases reliable predictive coverage.
Analysis of the in-season variable leaf area index

Predictive distributions

After the analysis of BBCH, we now study the in-season variable LAI. Exemplary plots of the time series for site 1, year 2011 are shown in Fig. 9. The corresponding plots for all site-years are provided in the Appendix (Fig. 26).

The adopted measurement technique (Weber et al. 2021) for field observations of LAI does not differentiate between green leaf (i.e. photosynthetically active) and dead leaves. After the onset of leaf senescence, which dominates the LAI evolution after maximum LAI, we can consider that the measured LAI values contain an unknown amount of green leaf as well as dead leaves. However, the modeled LAI is green leaf LAI. For LAI, we can see clear differences between the predictions of the three models: CERES and SUCROS overestimate the LAI in the initial phase, however, the predictions of CERES do not drop as significantly as the ones of SUCROS. As for the phenological development, SPASS can describe the measured data most accurately. In all cases, the peak of the median simulated LAI is much earlier than the peak in the measurements, sometimes by far. Since green leaf LAI of winter wheat is 0 at harvest, we see that, out of the three models, only a few individual simulation runs of SUCROS achieve this. Recall that the SUCROS wheat model does not simulate the vernalization of winter wheat. In other words, it is a summer wheat model, providing a feasible explanation of the, comparatively, very early development of LAI. This overestimation leads to a premature maximum and underestimates the measurements in the decreasing phase during senescence.

Figure 9 indicates that the prior predictive of SUCROS at harvest is close to zero, however with both an early and large peak in simulated LAI. Both CERES and SPASS are closer to the data but do not reach a green leaf of 0 at harvest. For SPASS this could relate to the fact that the simulated phenology had not fully reached maturity by the time of the harvest. In other words, if in the model the harvest date had been set to full maturity, and not to the farmer-reported harvest date, we would surely observe a further decrease in LAI. While the development stage of SPASS is “too slow”, we can learn that the senescence of CERES is not fast enough.

Energy distance-based similarity analysis

Similar to BBCH, the within-model spread of the LAI predictions in Fig. 10a increases slowly until June, followed by a decrease until harvest. The curves representing CERES and SPASS are relatively similar. SUCROS shows the largest spread for most of the simulation time until it declines starting in June. At the harvest date, it is mostly SPASS that has the highest spread in its predictions.

In Fig. 10b, the energy distance between the models shows that, during most of the seasons, the predictions of CERES and SPASS are the most similar ones. The time series of the energy distance between SUCROS and CERES and between SUCROS and SPASS are similar, with the ones of SUCROS and CERES being usually lower (i.e. the models are more similar) and showing a minimum in June–July, followed by a rising phase until harvest.

Figure 10c shows that SPASS has the lowest distance to the observations for most months, which means it performs best. The curve representing CERES’ energy score ranges in the middle, and SUCROS performs worst, having the highest distance. The case of SUCROS highlights one of the major benefits of analyzing model predictive distributions using energy distance and energy score: While SUCROS shows the worst predictions according to Fig. 10c, it provides these with the highest confidence in the time from July to August. Energy statistics support this insight in a straightforward way on an easily-interpretable scale.

Compared to the energy-statistics-based analysis of BBCH (Fig. 8), the within-model spreads of the LAI predictions (Fig. 10) are smaller and increase less during the growing season. From the energy distance between the models, we can observe that CERES and SPASS are the most dissimilar models with regard to BBCH predictions, while they are the most similar ones considering LAI predictions. Here, the biggest differences are between the LAI predictions of SUCROS and SPASS. The smallest energy score and, hence, the best goodness-of-fit during most phases was calculated, for both quantities of interest, by SPASS. The worst performing model regarding LAI predictions is SUCROS, while considering BBCH predictions, it is CERES.
Our analysis of the prior predictive distributions revealed that a model that performs well during season might still end with an imprecise yield prediction: During the season, SPASS is best in predicting LAI and BBCH, whereas CERES and SUCROS clearly deviate from the observations, especially in the case of LAI. In predicting yield, however, SUCROS performs best, while SPASS is worst due to its very broad prior predictive distribution that covers even very low values with a relatively high probability. This confirms what was observed in the study by Martre et al. (2015). They compared the goodness-of-fit for calibrated models: a model that cannot reproduce in-season measurements well might do a better job in predicting end-of-season variables. Given our analysis, one could add to this statement that also the reverse can be true, i.e. a model that performs well during the season might still fail to predict yield reasonably.

In addition, the similarity among the models was not consistent across different variables: while the LAI predictions of CERES and SPASS are the most similar ones and the ones of SUCROS and SPASS are the most dissimilar ones, the opposite is true considering the yield predictions.

Summary and conclusions

We analyzed the similarity of predictions by the three plant growth models CERES, SPASS and SUCROS and their goodness-of-fit to observed data in a probabilistic framework. The goal of this study was to find methods for gaining deeper insights into the model set. An intuitive understanding of similarities between the models and the measurements can help model developers to improve both the individual models and the multi-model methods. The presented method can be used to identify different model settings, e.g. situations in which all models form a cluster while being distant from the measured data. This may indicate that all models are highly similar and that a relevant process is not considered in any of the models. An intuitive visualization of model similarities can guide the multi-model process, e.g. when it comes to assigning model weights for averaging. Therefore, we propose to combine specific visualization methods that make modelers aware of the (dis)similarities in the predictions of the considered model set. Each method highlights another piece of information and adds to a comprehensive overview of the considered model set.

The analysis is based on so-called energy statistics introduced by Rizzo and Székely (2016). The energy distance between the probabilistic predictions is used to quantify model similarities. With the same method, we can also assess model performance by calculating the energy distance between model predictions and noisy measurements. For comparing probabilistic model predictions to deterministic observations, the so-called energy score is used. It acts on the same scale as the energy distance making both intuitively comparable. Therefore, energy statistics proved to be widely applicable, as energy distance and energy score can be used.

Fig. 10  a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on LAI, site 1, year 2011.
jointly to compare two probability distributions as well as a probability distribution and a deterministic reference.

Our results confirmed that “there is no single best model” (Hagedorn et al. 2005; Palosuo et al. 2011; Martre et al. 2015): none of the investigated models performed consistently better or worse than the others when considering different variables. While SPASS showed the best goodness-of-fit regarding in-season variables LAI and BBCH, its overly wide yield predictions lead to poor performance for this end-of-season variable. Therefore, combining the models in an ensemble might give more robust predictions as a broader range of possible predictions is covered.

Generally, we suggest analyzing model similarities when using multi-model ensembles, as redundancies in the model set lead to an overly high weight of certain predictions and therefore, model weights should be diluted (George 2010; Garthwaite and Mubwandarikwa 2010). Similar to the results regarding goodness-of-fit, we also found that model similarities vary for different variables: two models that gave similar predictions for one variable showed clear differences in predicting another one. Therefore, no general dilution priors can be defined for this model set. Rather, they need to be chosen depending on model similarities for each quantity of interest.

We also investigated whether model similarities or performance are dependent on the weather conditions during the growing season. To this end, we used radar charts to visualize the similarities and color-coded them according to the wetness or dryness of the respective site-year. Although there was no apparent effect of the weather conditions on the model predictions visible, we suggest this approach of visualization to be studied further. We assume that, for other scenarios and model sets, this might be a straightforward tool to display the influence of different boundary conditions on the prediction accuracy and similarity of models.

By assessing the within-model spread, the distance between the models, and the goodness-of-fit on the same scale, we can gain a better understanding of the model set.

Our study was based on prior predictions, i.e. the models have not been calibrated. There are two main reasons for this: (1) (not only) in the crop modeling community, different groups use different calibration approaches (Wallach et al. 2020) and hence, there is a lack of consistency. (2) Model structural errors are often compensated by choosing non-physical parameters (e.g. Wallach 2011). This leads to good model performance for the variable the model has been calibrated on, but poor performance for others. Therefore, we support the suggestion of Vogel and Sankarasubramanian (2003) to validate the model structure prior to calibration. As in any Bayesian framework, a subjective choice of prior distributions based on expert knowledge is needed. Future research should assess the sensitivity of the analysis regarding the priors. Another promising way to go is the assessment of structural model similarity, e.g. based on information-theoretic methods as done by Bennett et al. (2019). As our analysis suggests that a combination of the individual models into an ensemble prediction might yield more robust results, our introduced model evaluation workflow might also inform different model combination methods and weighting schemes of future applications.

### Appendix 1: Model description

See Tables 3, 4, 6, 5, 7.

| Characteristic                        | CERES          | SPASS          | SUCROS         |
|---------------------------------------|----------------|----------------|----------------|
| Leaf area/light interception          | Simple         | Detailed       | Detailed       |
| Light utilization                     | Radiation use efficiency approach | Gross photosynthesis-respiration | Gross photosynthesis-respiration |
| Yield formation                       | Tot. (above-ground) biomass, number of grains | Number of grains, partitioning during reproductive stages | Partitioning during reproductive stages |
| Phenology                             | Temperature, photoperiod (day length), vernalization | Temperature, photoperiod (day length), vernalization | Temperature |
| Root distribution over depth          | Exponential    | Exponential    | Exponential    |
| Environmental constraints             | Water limitation, N limitation | Water limitation, N limitation | Water limitation, N limitation |
| Type of water stress                  | Actual to potential evapotranspiration ratio, soil available water in root zone | Actual to potential evapotranspiration ratio, soil available water in root zone | Actual to potential evapotranspiration ratio, soil available water in root zone |
| Water dynamics                        | Richards approach | Richards approach | Richards approach |
| Evapotranspiration                    | Penman–Monteith | Penman–Monteith | Penman–Monteith |
| Soil CN-model                         | CN model, 3 organic matter pools, microbial biomass pool | CN model, 3 organic matter pools, microbial biomass pool | CN model, 3 organic matter pools, microbial biomass pool |
| No. of cultivar parameters            | 7              | 5              | 2              |
Table 4 Description of model parameters varied and parameter bounds of the uniform prior where $a = \text{lower bound and } b = \text{upper bound}, \text{subscript 1 denotes the top-soil and 2 the sub-soil as defined in Table 1}$

| Module       | Parameter | Unit | $a$  | $b$  | Description                                                                 |
|--------------|-----------|------|------|------|----------------------------------------------------------------------------|
| Soil water   | $a_1$, $a_2$ | cm$^{-1}$ | 0.002 | 0.03 | van Genuchten $a$                                                           |
|              | $n_1$, $n_2$ | –     | 1.2  | 2.2  | van Genuchten $n$                                                           |
|              | $K_{s,1}$, $K_{s,2}$ | cm d$^{-1}$ | 10  | 500  | saturated hydraulic conductivity                                            |
|              | $r_1$, $r_i$ | –     | $-1$ | 8    | Tortuosity parameter                                                        |
| $ET_a$       | $k_{\text{ini}}$ | –     | 0.2  | 1    | Crop coefficient for the initial vegetation period (dev stage 0)             |
|              | $k_{\text{mid}}$ | –     | 0.5  | 1.8  | Crop coefficient for the mid vegetation period (dev stage 0.75–1.5)         |
|              | $k_{\text{end}}$ | –     | 0.2  | 1    | Crop coefficient for the end of the vegetation period (dev stage 2.0)        |
| Nitrogen     | $\text{Nit}_1$ | d$^{-1}$ | 0.1  | 1    | Nitrification rate                                                          |
|              | $\text{Nit}_2$ | d$^{-1}$ | 0.05 | 0.6  |                                                                           |
|              | $\text{DeNit}_1$ | d$^{-1}$ | 0.1  | 1    | Denitrification rate                                                        |
|              | $\text{DeNit}_2$ | d$^{-1}$ | 0.0  | 0.01 |                                                                           |
|              | $\text{MiLit}_1$ | d$^{-1}$ | 0.01 | 0.1  | Mineralization rate constant of the litter pool                             |
|              | $\text{MiLit}_2$ | d$^{-1}$ | 0.01 | 0.1  |                                                                           |
|              | $\text{MiHum}_1$ | d$^{-1}$ | $1e^{-5}$ | $1e^{-4}$ | Mineralization rate constant of the humus pool                             |
|              | $\text{MiHum}_2$ | d$^{-1}$ | $1e^{-6}$ | $1e^{-5}$ |                                                                           |
|              | $\text{MiMa}$ | d$^{-1}$ | 0.01 | 0.1  | Mineralization rate constant of the manure pool                             |
| Initial condition | $\theta_{\text{ini},1}$ | vol% | 10    | 40  | Initial soil profile water content                                          |
|              | $\theta_{\text{ini},2}$ | vol% | 15    | 45  |                                                                           |
### Table 5 Description of SPASS model parameters and parameter bounds of the prior

| Parameter   | Unit          | a   | b   | Description                                                                 |
|-------------|---------------|-----|-----|------------------------------------------------------------------------------|
| PMAX        | kg CO$_2$ ha$^{-1}$ h$^{-1}$ | 38  | 45  | Gross photosynthesis rate at light saturation and CO$_2$ 340 ppm             |
| LUE         | g J$^{-1}$    | 0.55| 0.7 | Light use efficiency                                                         |
| TMINPS      | °C            | 0   | 5   | Minimum temperature for photosynthesis                                       |
| TOPTPS      | °C            | 20  | 25  | Optimum temperature for photosynthesis                                       |
| TMAXPS      | °C            | 32  | 40  | Maximum temperature for photosynthesis                                       |
| PDD1        | d             | 32  | 48  | Duration from emergence to anthesis                                           |
| PDD2        | d             | 20  | 36  | Duration from anthesis to emergence                                           |
| VERN        | d             | 24  | 46  | Minimum value of vernalization days                                           |
| PDL         | –             | 0.01| 0.25| Photoperiod sensitivity factor                                                |
| DLOPT       | h             | 18  | 20  | Optimal photoperiod length                                                    |
| TMININDEV   | °C            | 0   | 2   | Minimum temperature of vegetative development                                |
| TOPTDEV     | °C            | 22  | 26  | Optimum temperature of vegetative development                                |
| TMAXDEV     | °C            | 32  | 38  | Maximum temperature of vegetative development                                |
| TMINDEV2    | °C            | 2   | 6   | Minimum temperature of reprod. development                                    |
| TOPTDEV2    | °C            | 26  | 30  | Optimum temperature of reprod. development                                    |
| TMAXDEV2    | °C            | 38  | 45  | Maximum temperature of reprod. development                                    |
| G1          | # g$^{-1}$    | 24  | 35  | Number of grains per stem weight at anthesis                                  |
| SPCLW       | k g$^{-1}$ leaf | 350 | 500 | Specific leaf weight                                                         |
| RESR        | –             | 0.36| 0.45| Fraction of stem weight as reserves                                           |
| G2          | mg grain$^{-1}$ d$^{-1}$ | 2.4 | 3.6 | Maximum grain filling rate                                                    |
| NACCR       | mg N grain$^{-1}$ d$^{-1}$ | 0.02| 0.06| Nitrogen accumulation rate                                                    |
| REXT        | cm d$^{-1}$   | 1.5 | 3.0 | Maximum root extension rate                                                  |
| SPCRL       | cm g$^{-1}$   | 8000| 12000| Specific root length                                                         |
| RWUR        | cm$^3$ cm$^{-1}$ d$^{-1}$ | 0.024| 0.036| Maximum water uptake rate per root length                                     |
| RNUR        | kg N cm$^{-1}$ d$^{-1}$ | 0.006| 0.01 | Maximum nitrogen uptake rate per root length                                  |
| RDMAX       | cm            | 100 | 200 | Maximum rooting depth                                                       |
| DVSSEN      | –             | 1   | 1.3 | Development stage at which senescence begins                                 |
| RDRL        | –             | 0.015| 0.025| Relative death rate of leaves                                                 |
| RDRR        | –             | 0.015| 0.025| Relative death rate of roots                                                  |

### Table 6 Description of CERES model parameters and parameter bounds of the prior

| Parameter   | Unit            | a   | b   | Description                                                                 |
|-------------|-----------------|-----|-----|------------------------------------------------------------------------------|
| P1D         | –               | 0.001| 0.008| Daylength coefficient                                                       |
| P1V         | h               | 25  | 60  | Inverse of optimum vernalization rate                                        |
| PHINT       | –               | 70  | 150 | Phyllochroine interval                                                       |
| P1          | °C d            | 170 | 400 | Thermal Time from emergence-to terminal spikelet                            |
| P4          | °C d            | 120 | 200 | Thermal Time from end of pre-anthesis ear growth-begin of grain filling      |
| P5          | °C d            | 400 | 700 | Thermal Time for grain filling (phase 5)                                     |
| G1          | # g$^{-1}$      | 20  | 40  | Number of grains per stem weight at anthesis                                 |
| G2          | mg grain$^{-1}$ d$^{-1}$ | 1 | 4 | Maximum grain filling rate                                                  |
| RWUR        | cm$^3$ cm$^{-1}$ d$^{-1}$ | 0.01 | 0.1 | Maximum water uptake rate per root length                                   |
| RNUR        | kg ha$^{-1}$    | 0.003| 0.027| Maximum nitrogen uptake rate per root length                                 |
Appendix B: Figures

See Figs. 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and 26.

![Dendrograms based on the energy distance between models and observations (yield predictions)](image)

Table 7 Description of SUCROS model parameters and parameter bounds of the prior

| Parameter | Unit | a  | b  | Description                                                                 |
|-----------|------|----|----|-----------------------------------------------------------------------------|
| PMAX      | kg CO₂ ha⁻¹ leaf⁻¹ h⁻¹ | 38 | 45 | Gross photosynthesis rate at light saturation and CO₂ 340 ppm              |
| LUE       | g J⁻¹ | 0.55 | 0.7 | Light use efficiency                                                        |
| TBASE1    | °C   | 0 | 2  | Base temperature for phen. dev. vegetative phase                            |
| TSUM1     | °C d | 700 | 1500 | Temperature sum of vegetative growth phase                                  |
| TBASE2    | °C   | 2 | 5  | Base temperature for phen. dev. generative phase                            |
| TSUM2     | °C d | 600 | 1400 | Temperature sum of generative growth phase                                  |
| LA0       | m² plant⁻¹ * 10,000 | 0.45 | 0.7 | Initial leaf area                                                           |
| RGRL      | °C⁻¹ d⁻¹ | 0.005 | 0.01 | Relative growth rate of leaf area                                           |
| G1        | #g⁻¹ | 24 | 35 | Number of grains per stem weight at anthesis                                |
| SPCLW     | kg DW ha⁻¹ leaf⁻¹ | 350 | 500 | Specific leaf weight                                                        |
| REXT      | cm d⁻¹ | 1.5 | 3.0 | Maximum root extension rate                                                 |
| SPCRL     | m kg⁻¹ | 8000 | 12,000 | Specific root length                                                        |
| RDMAX     | cm   | 100 | 200 | Maximum rooting depth                                                       |
Fig. 12  Heatmap reflecting the similarities between models and observations based on yield predictions for site 4–6. The color-coding represents the values of the individual components of the energy distance: $E[|X - Y|_2$ (main diagonal entries) and $E[|X - X'|_2$ (off-diagonal entries).

Fig. 13  a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_i}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on LAI for site 1.
Fig. 14  

(a) Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$,  
(b) energy distance $d$ between pairs of models, and  
(c) root energy score $\sqrt{ES}$ between models and observations based on LAI for site 2

Fig. 15  

(a) Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$,  
(b) energy distance $d$ between pairs of models, and  
(c) root energy score $\sqrt{ES}$ between models and observations based on LAI for site 3
**Fig. 16** a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on LAI for site 4.

**Fig. 17** a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on LAI for site 5.
Fig. 18  a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on LAI for site 6.

Fig. 19  a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on phenology for site 1. The dashed lines indicate the date when the mean predictions reach BBCH = 60.
Fig. 20  

(a) Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$,  
(b) energy distance $d$ between pairs of models, and  
(c) root energy score $\sqrt{ES}$ between models and observations based on phenology for site 2.  
The dashed lines indicate the date when the mean predictions reach BBCH = 60.

Fig. 21  

(a) Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$,  
(b) energy distance $d$ between pairs of models, and  
(c) root energy score $\sqrt{ES}$ between models and observations based on phenology for site 3.  
The dashed lines indicate the date when the mean predictions reach BBCH = 60.
Fig. 22  a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $\Delta$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on phenology for site 4. The dashed lines indicate the date when the mean predictions reach BBCH = 60.

Fig. 23  a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $\Delta$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on phenology for site 5. The dashed lines indicate the date when the mean predictions reach BBCH = 60.
Fig. 24. a Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, b energy distance $d$ between pairs of models, and c root energy score $\sqrt{ES}$ between models and observations based on phenomenology for site 6. The dashed lines indicate the date when the mean predictions reach BBCH = 60.

Fig. 25 Time series of phenology predictions for all site-years. The shaded intervals represent the 90% credible intervals (blue: CERES, red: SUCROS, yellow: SPASS), the dashed lines represent the model means and the black points represent the medians of the measurements.
Acknowledgements The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Collaborative Research Center 1253 CAMPOS (DFG, Grant Agreement SFB 1253/1 2017) and the Cluster of Excellence EXC 2075 “Data-integrated Simulation Science (SimTech)” at the University of Stuttgart under Germany’s Excellence Strategy-EXC 2075-390740016. We also thank Dr. Irene Witte for providing the Monte Carlo simulation results.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Data and code can be downloaded here: https://bwsyn.candshare.kit.edu/s/4gZf4qEAA4TFm.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abramowitz G (2010) Model independence in multi-model ensemble prediction. Aust Meteorol Oceanogr J 59(1SP):3–6. https://doi.org/10.22499/2.5901.002

Abramowitz G, Gupta H (2008) Toward a model space and model independence metric. Geophys Res Lett. https://doi.org/10.1029/2007GL032834
crop–environment–management interactions. Glob Change Biol. https://doi.org/10.1111/gcb.14411

Wallach D, Palosuo T, Thorburn P, Hochman Z, Gourdain E, Andrianasolo F, Asseng S, Basso B, Buis S, Crott N, Dibari C, Dumont B, Ferrise R, Gaiser T, Garcia C, Gayler S, Gahramani A, Hiremath S, Hoek S, Horan H, Hoogenboom G, Huang M, Jabloun M, Janssen PE, Jing Q, Justes E, Kersebaum KC, Klosterhalfen A, Launay M, Lewan E, Luo Q, Maestri B, Mielzen H, Moriondo M, Zadeh HN, Padovan G, Olesen JE, Poyda A, Priesack E, Pullens JWM, Qian B, Schütze N, Shelia V, Souissi A, Specia X, Srivastava AK, Stella T, Streck T, Trombi G, Wallor E, Wang J, Weber TKD, Weihermüller L, de Wit A, Wöhling T, Xiao L, Zhao C, Zhu Y, Seidel SJ (2020) The chaos in calibrating crop models. Plant Biol. https://doi.org/10.1111/pb.12947

Wang E (1997) Development of a generic process-oriented model for simulation of crop growth. Ökologie (Munich, Germany). Utz, Wissenschaft, Munich

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

Wang E, Engel T (2000) SPASS: a generic process-oriented crop model with versatile windows interfaces. Environ Model Softw 15(2):179–188. https://doi.org/10.1016/S1364-8152(99)00033-X

Weber TKD, Durner W, Streck T, Diamantopoulos E (2019) A modular framework for modelling unsaturated soil hydraulic properties over the full moisture range. Water Resour Res. https://doi.org/10.1029/2018WR024584

Weber TKD, Finkel M, Conceição Gonçalves M, Vereecken H. (2020) Pedotransfer function for the Brunswick soil hydraulic property model and comparison to the van Genuchten–Mualem model. Water Resour Res. https://doi.org/10.1029/2019WR026820

Weber TKD, Ingwersen J, Högy P, Poyda A, Wizemann H-D, Demyan MS, Bohm K, Eshonkulov R, Gayler S, Kremer P, Laub M, Nkwain YF, Troost C, Witte I, Cadisch G, Müller T, Fangmeier A, Wulfmeyer V, Streck T (2021) Multi-site, multi-crop measurements in the soil-vegetation-atmosphere continuum: a comprehensive dataset from two climatically contrasting regions in South West Germany for the period 2009–2018. Earth Syst Sci Data Discuss 2021:1–32. https://doi.org/10.5194/essd-2020-396

Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134(630):241–260. https://doi.org/10.1002/qj.210 (accessed 2018-07-24)

Winter CL, Nychka D (2010) Forecasting skill of model averages. Stoch Environ Res Risk Assess 24(5):633–638. https://doi.org/10.1007/s00477-009-0350-y

Wöhling T, Geiges A, Nowak G, Gayler S, Högy P, Wizemann HD (2013) Towards optimizing experiments for maximum-confidence model selection between different soil–plant models. Procedia Environ Sci 19:514–523. https://doi.org/10.1016/j.proenv.2013.06.058

Wöhling T, Schöniger A, Gayler S, Nowak W (2015) Bayesian model averaging to explore the worth of data for soil–plant model selection and prediction. Water Resour Res 51(4):2825–2846. https://doi.org/10.1002/2014WR016292

Xu R, Wunsch D (2008) Clustering. IEEE Press Series on Computational Intelligence. Wiley, New Jersey

Yao Y, Vehtari A, Simpson D, Gelman A (2018) Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal 13(3):917–1007. https://doi.org/10.1214/17-BA1091

Yun K, Hsiao J, Jung M-P, Choi I-T, Glenn DM, Shim K-M, Kim S-H (2017) Can a multi-model ensemble improve phenology predictions for climate change studies? Ecol Model 362:54–64. https://doi.org/10.1016/j.ecolmodel.2017.08.003

Ziel F, Berk K (2019) Multivariate forecasting evaluation: on sensitive and strictly proper scoring rules. arXiv:1910.07325 [econ, stat]

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.