One Earth

CellPress
OPEN ACCESS

Perspective

Transdisciplinary coordination is essential for advancing agricultural modeling with machine learning

Lily-belle Sweet,[1,2,*] Ioannis N. Athanasiadis,[3] Ron van Bree,[3] Andres Castellano,[4] Pierre Martre,[5] Dilli Paudel,[3] Alex C. Ruane,[4] and Jakob Zscheischler[1,2,6]
[1]Department of Compound Environmental Risks, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
[2]Department of Hydro Sciences, TUD Dresden University of Technology, Dresden, Germany
[3]Wageningen University and Research, Wageningen, the Netherlands
[4]NASA Goddard Institute for Space Studies (GISS), New York, NY, USA
[5]LEPSE, Université Montpellier, INRAE, Institut Agro Montpellier, Montpellier, France
[6]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden-Leipzig, Germany
*Correspondence: lily-belle.sweet@ufz.de
https://doi.org/10.1016/j.oneear.2025.101233

SUMMARY

Crop models play a key role in the study of climate change impacts on food production as well as improving food systems resilience and analyzing the effect of potential adaptation interventions. Here, we illustrate opportunities that machine learning offers for tackling key challenges of agricultural modeling. However, to unlock the full potential of machine learning, and thereby accelerate progress toward a more secure and sustainable global food system, serious pitfalls must first be addressed. We argue that transdisciplinary coordination is needed to identify impactful research gaps, curate and maintain benchmark datasets, and establish domain-specific best practices.

INTRODUCTION

Despite large increases in the worldwide harvested area of staple crops, the United Nation's Sustainable Development Goal of eradicating hunger by 2030[1] is in jeopardy. Yields in some breadbasket regions are stagnating, and the share of crops being used directly for food is decreasing.[2–4] At the same time, the effects of climate change are already being felt by the agricultural sector, and these impacts are expected to increase over the coming decades.[5–11] Higher global temperatures are associated with reduced crop yields,[12,13] largely offsetting the recent gains associated with rising CO_2.[14] Food insecurity risk is expected to increase due to more widespread and pronounced extreme climate events such as droughts, heatwaves, and floods.[15–18] The challenge of improving the resilience of our global food system overlaps with an urgent need to mitigate its climate impact.[9]

The complexity of the food system necessitates a transdisciplinary systems thinking approach[19] that takes into account interactions and feedback mechanisms between economic, environmental, and biophysical processes at field, regional, national, or global scales.[20–22] This requires the use of empirical and process-based modeling approaches, with couplings between climatic, biophysical, and socioeconomic models, in order to accurately capture the behavior of food system decision processes[23] and disentangle underlying causal mechanisms.[24]

The biophysical impacts of climate on agricultural yields are usually simulated using process-based crop models, which track day-by-day plant development according to the interactions of genetic, environmental, and management factors (GxExM).[25] This process-level representation allows crop models to be used for applications outside of observed conditions, such as historical counterfactuals, future climate scenarios, novel crop traits, or adapted management systems. Process-based crop models can also simulate a large suite of information beyond yields, including water and nutrient use and loss, soil carbon sequestration, and greenhouse gas emissions—all of which are necessary for supply-side approaches to agricultural development and food security analysis. However, crop models differ in the mechanistic and functional components used, and some important plant processes are not captured.[26,27] For example, models often use "stress factors" as a practical way of representing the impact of heat or drought, but their performance in cases of combined stresses has rarely been evaluated, and most do not consider the effects of compounding biotic and abiotic stresses.[28] Furthermore, most models do not consider the impact of pests and diseases, or the spatial variability of soils or management, despite being increasingly applied at spatial and temporal scales at which these effects become relevant. Any resulting lack of explanatory skill may not be identified given that evaluation strategies are often insufficient in spatial contexts[29] and high-quality reference data are scarce. Consequently, although broad signals are robust, current crop models exhibit substantial uncertainties in future projections of global yields under climate change.[11,30]

Over the last few years, the use of data-driven machine-learning (ML) methods has risen rapidly in agricultural science[31] as well as many other scientific fields.[32] While statistical models have a long history in agricultural modeling, the flexibility of ML models and their ability to learn complex, potentially unknown interactions makes them promising tools for capturing crop

Box 1. A brief history of agricultural crop models

Agricultural modeling has been driven by a transdisciplinary network of scientists since the 1960s, and the field has repeatedly embraced emerging technologies such as dynamic programming, high-performance computing, and the open-source software movement.[25,41,42] Crop models have been used to improve our understanding of crop growth processes and their response to the environment and to support a variety of stakeholder decisions. These include the evaluation of GxExM interventions for climate adaptation and mitigation,[43,44] the identification of optimal growing periods,[45] and estimation of future warming attributable to worldwide food consumption.[46] Furthermore, simulations can provide valuable food-security information in regions where the availability, frequency, and/or quality of data are lacking.[47] Today, there exists a large diversity of crop models,[48] with over 40 developed for wheat alone.[49]

The release of global datasets of cropping areas, sowing dates, and yields in the early 2000s led to the establishment of data standards and harmonized inputs.[50] This then enabled the use of agricultural models to create projections of crop productivity, trade, food prices, and other impacts under future climate-change scenarios.[11,51,52] However, crop simulation experiments were primarily run with diverse protocols and data and without consistent accuracy assessment or uncertainty quantification,[53] meaning that results from different studies could not be synthesized or compared. In response to this, the Agricultural Model Intercomparison and Improvement Project (AgMIP) launched in 2010, fostering a global community of climate, crop, livestock, economics, and nutrition modelers to conduct multi-model ensemble intercomparison experiments.[54,55] AgMIP intercomparison activities and large-scale ensemble simulations have been used to evaluate and improve crop models[56,57] and to robustly assess current and future challenges to food systems as well as historic variability[58] and agro-climatic sensitivity.[59]

growth and development. Furthermore, ML models are skilled in dealing with missing and noisy data, which are common problems in agricultural settings. ML models have been used to forecast yields at multiple scales, to emulate existing process-based crop models,[33,34] to generate datasets of planting dates[35] and soil characteristics,[36] and to downscale simulations to the fine spatial resolutions needed to inform regional decision making[37] (which can be anywhere from the field to district level depending on the application). Their ability to extract information from image and text data has allowed researchers to exploit data sources such as satellite imagery[38,39] and news articles.[40]

However, some pitfalls in the use of ML for typical agricultural modeling applications must be recognized. Data-driven predictive models may not necessarily capture true biophysical processes, even if they exhibit excellent predictive performance. This means that ML model predictions outside of the training data distribution could be poor or even physically implausible, which impedes their use for model emulation or yield projections in future climate scenarios, alternative systems, or data-scarce regions. Additionally, crop models are used to encapsulate and improve scientific understanding of plant growth processes, but most ML models are "black boxes." While there has been considerable research effort into explainable or interpretable AI,[41] current methods may not be robust to dependencies among features or sampling variability.

While the existence of these theoretical limitations is well known, their relevance is dataset and task specific. Agricultural models are used for a wide range of applications, including farmer decision support, national yield forecasting, and the exploration of climate-change adaptation, each of which may have specific usability requirements (such as computational and data efficiency, auditability, ease of use without technical expertise, precision, or accuracy). To drive research progress at the intersection of ML and agricultural modeling, and to capitalize on the strengths of both ML and process-based modeling paradigms, transdisciplinary coordination and sustained engagement with stakeholders is vital. In this article, we aim to foster such collaboration by providing an introduction to the history, importance and pitfalls of both agricultural crop modeling (Box 1) and ML research (Box 2). We outline several ways researchers are leveraging ML to enhance existing crop models and identify critical priorities to advance research progress such as the creation and maintenance of comprehensive benchmark datasets, documentation of the diverse range of stakeholder requirements, establishment of model intercomparison protocols, and support for the development of innovative hybrid modeling approaches.

CURRENT LIMITATIONS OF CROP MODELS

Despite a relatively strong research focus leading to robust emerging information on the effects of climate change on agricultural yields,[55,71] crop models have been found to underestimate the impacts of heavy rainfall, extreme heat, and drought.[72–75] A number of non-linear responses of fertility and phenology to extreme climate events are not addressed adequately or at all in most models.[76,77] Additionally, the compounding effects of multiple non-extreme weather events can lead to severe yield impacts,[78] and scientific understanding of these interactions is lacking.[28] This was exemplified by the extreme 2016 wheat failure in France, which was not anticipated by forecasters until shortly before harvest.[79,80] In general, crop models do not explicitly consider interactions between multiple stresses,[28] but such combinations are critical for food security risk assessments (for example, heat and drought stress lead to intensified impacts when they co-occur[81]). Beyond climate factors, the multiple interactions between ozone,[82] CO_2,[83] salinity, nitrogen, phosphorus, soils, and genetic factors remain poorly understood.

Some important factors are not included in simulations because of the challenges that come with coupling external process models, such as diverging input data requirements. For example, the impacts of weeds, pests, and diseases are usually not considered in crop models, most of which were originally

Box 2. The rise of ML

While ML dates from the 1950s or earlier, more recent advances in computation speed, data availability, and the success of deep learning architectures have precipitated a new wave of interest in the field. The term "ML" encompasses a range of models, including decision trees, random forests, gradient boosted machines, and neural networks (deep learning). Algorithms can exploit labeled training examples (supervised learning), interact with a dynamic environment to maximize some reward (reinforcement learning), or find patterns in unlabeled data (unsupervised learning). Neural network architectures are able to exploit data properties such as temporal dependence, spatial patterns, or graph structures and have proved highly successful for complex tasks such as protein folding[60] and weather forecasting.[61,62] These methods can achieve impressive performance on high-dimensional input such as image or text data, but tree-based models may be preferable on tabular (low dimensional) or small datasets and in some domains such as time series forecasting.[63,64]

The culture of ML research prioritizes predictive skill over understanding the underlying data-generating process. For example, unlike traditional statistical models, where input variable interactions are clearly designed and the number of model parameters kept small, neural network models involve several layers of non-linear interactions and can contain billions of parameters. In general, ML methods are able to learn complex interactions that can improve model performance, but it is difficult to explain their predictions,[65] and models can overfit to the training data.[66] Motivated by the need to build trust in "black box" ML models, research in "explainable" or "interpretable" methods has grown rapidly in recent years. Usually, *post hoc* methods such as feature importances, partial dependence plots, and Shapley values are used to give human-understandable explanations of the reasoning behind individual predictions or a holistic overview of the underlying mechanisms.[67–70]

designed to capture biophysical processes relating to water, energy, and carbon. Separate models exist for simulating these effects, but they require spatially explicit and hourly data for variables not normally utilized by crop models.[84]

Crop models rely on knowledge of biophysical processes gained from field-scale experimental studies[72] and are usually run independently for each location or grid cell studied. At larger scales, the climate, topography, soil, and management data used to drive crop models are often spatially aggregated and may have substantial gap filling,[85,86] introducing an additional source of uncertainty.[87] Factors not considered in process-based models may have a larger effect at the regional, national, or global spatial scales relevant for some decision makers (e.g., farmer behavior, government policies, and the availability of workers). Some studies have found that scale-dependent parameterization can ameliorate these errors,[87,88] but validation of model performance requires high-quality data at similar scales. Furthermore, crop model evaluation frequently fails to adequately consider spatial context.[29]

Finally, in part to make up for incomplete representation of relevant processes happening at very fine scales, crop models contain a large number of parameters. These are derived from experimental data or empirically estimated by minimizing errors of simulated output against observations (often by manual trial and error).[89] Outside of experimental settings, often only bulk end-of-season yield or spatial production data are available for calibration, and a precise tuning is not feasible given that the number of parameters is often larger than the number of observations. Different calibration strategies can result in substantial variation in model skill, even when the same validation data are used.[87,90] Simulated yields and yield variability at the global scale have been found to be highly sensitive to parameterization.[91]

RECENT USE OF ML IN AGRICULTURAL MODELING

The successful use of ML methods could help researchers tackle several challenges in agricultural modeling. Their ability to extract information from remote-sensing data products, social media, or news articles,[39,40,92–94] for example, could improve yield predictions in regions where typical crop model input data are inaccessible or low quality. ML models can be trained on datasets comprising multiple crop types and cultivars and then fine-tuned for relatively under-studied crops, like millet, cassava, and groundnut. Although currently underrepresented in research studies, there are many opportunities for ML approaches to complement existing biophysical crop models (Figure 1). In this section, we highlight some of the ways in which ML methods have already been used in agricultural modeling, as well as some recent advances in related fields that suggest avenues for further exploration.

One of the simplest ways to make use of ML is to generate and gap-fill data that are required as input for current crop models. For instance, ML models have been used to map soils in high spatio-temporal resolutions from sparse observations[36] and to identify annual field-scale planting dates from satellite imagery,[35] both of which are key sources of uncertainty in global crop yield simulations.[95,96] ML could also help to provide estimates that can be used for calibrating process-based crop models, such as the days from sowing to anthesis and maturity, plant canopy characteristics such as chlorophyll content or leaf area index, or above-ground biomass at different stages of the growing season.

ML models have also been trained on simulated datasets to create emulators of existing process-based crop models (sometimes referred to as "meta-models" or "surrogate models"). These can be used to generate simulations for a wide range of scenarios comparatively cheaply in terms of computation time and/or using fewer input data.[33,34,97–99] Emulators can also be used for post-processing model output, for instance to downscale gridded simulations to the higher resolutions that are often required for local impact assessments.[37]

The increasing popularity and accessibility of explainable ML tools (section "the rise of ML") has led to a wave of research that makes use of these techniques to analyze the relationships between different factors, such as management practices, soil

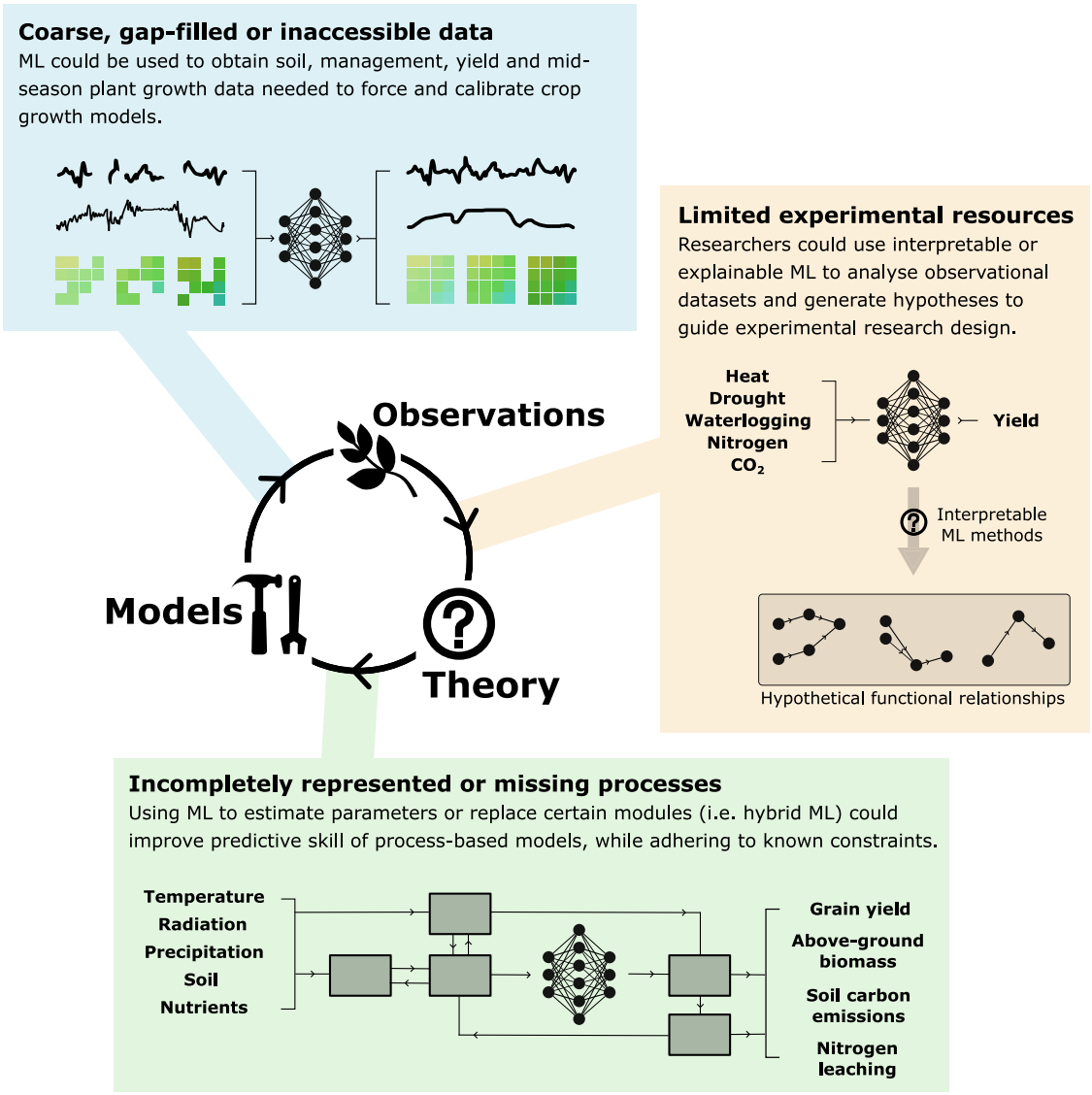

Figure 1. Examples of ways in which ML can be used to address current challenges in agricultural modeling

characteristics and health, climate, plant phenotype, crop growth, production, and food prices.[31,79,100–106] Given the infeasibility of conducting experimental trials that take into account all potential factors that influence crop growth, researchers have expressed hope that this avenue of research could improve scientific understanding of the compounding effects of multiple stressors.[28] Similarly, these methods can be used on emulators to study the relationships embodied in process-based models and diagnose incorrectly captured mechanisms.[37] In this type of research, ML models are usually trained to predict a single target variable (commonly, yield) from a number of predictive features. *Post hoc* interpretation methods are then used to identify important predictors and/or estimate the functional relationships learned.[69]

The hybridization of ML and process-based crop models is another promising avenue of research. Hybrid modeling aims to retain known biophysical relationships encoded in existing process-based models while using ML to better fit to the data by including missing processes or to select optimal parameter combinations.[99] This could make crop models relevant for new stakeholders and scientific communities by enabling the study of the biophysical and socioeconomic elements of transformative adaptation options (such as agroforestry, crop diversification, intercropping, pest control, or conservation agriculture).[107] Hybrid (or knowledge-guided) ML models lie on a spectrum between data-driven and process-based paradigms.[108] For example, a commonly used approach is to use simulated crop model output as input features for ML models[109,110] or as synthetic data to augment the training set. Some studies have used ML to replace components of process-based models where the underlying mechanisms may not be accurately represented rather than using model parameters.[111] This approach has been facilitated by the development of shared modeling frameworks and more modular and transparent crop models.[112,113] Similarly, ML

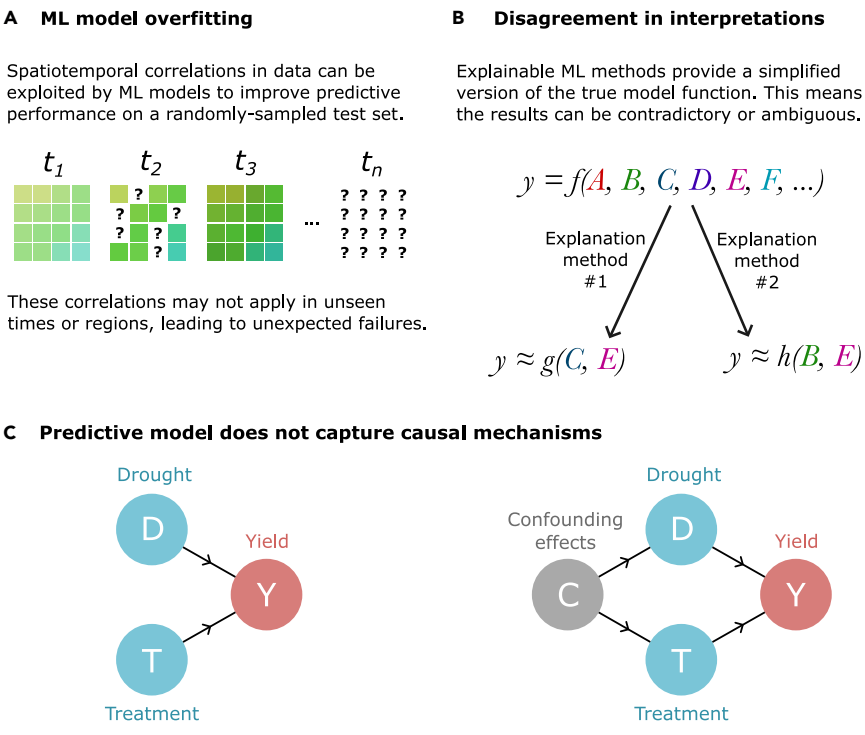

Figure 2. Some pitfalls in the use of ML for agricultural modeling applications

(A–C) ML model overfitting to spatiotemporal (or other) correlations in the training data (A), disagreement between interpretations from different explanation methods (B), and a predictive model not correctly capturing underlying causal mechanisms (C).

models can also be forced to adhere to known system dynamics by using knowledge-guided constraints,[114–116] which could reduce the data required for training and improve generalizability.

PITFALLS OF ML

The flexibility that makes ML methods so powerful can also, ironically, lead to serious impediments to their use for some crop-modeling applications (Figure 2). Researchers have called attention to the challenge of robustly evaluating ML models in spatial or spatiotemporal contexts, where data are not independent and identically distributed.[117–119] A recent review article identified data leakage issues leading to severely overoptimistic conclusions in studies from 17 scientific fields,[120] and there have been findings of previously undetected ML model failures in multiple scientific domains.[66,121] Replicating studies using ML for agricultural modeling is difficult, as research data are usually not openly accessible due to the financial interests and privacy concerns of stakeholders. This issue is, of course, not specific to ML, but it may have serious consequences for its use in agricultural modeling. Additionally, existing studies have made use of varying datasets, methods, and model evaluation strategies, which hinders comparison of results. This lack of reproducibility and comparability may contribute to a lack of trust and underutilization of ML methods in operational agricultural systems.[122]

Building trust requires transparency. Given the complex interactions influencing crop growth, and the potential ability of ML models to capture them, opening the black box is an attractive concept. However, researchers using explanation tools on ML models of crop yields have found ambiguous or contradictory results.[98,123] This may be partly due to the sensitivity of commonly used explanation methods to small data perturbations or correlations between features,[124–126] a lack of robustness to methodological choices such as the sampling method used for evaluation,[118,127] or the small size and inherent biases of observational agricultural datasets. The unreliability of these methods is particularly worrying because supplying inaccurate explanations for ML model predictions could still increase their perceived trustworthiness to stakeholders.[128]

Further study and more careful use of these methods could ameliorate these issues, but a fundamental problem remains: the difference between a predictive and causal model. Supervised ML methods exploit correlations to maximize predictive skill, and explanation methods are designed to diagnose bias and expose model mechanisms, not to identify the underlying causal structure of the data-generation process—a fundamentally different, and much larger, challenge.[129] Correlations learned by a predictive model may not hold outside the training distribution, leading to poor or physically implausible predictions.[66,130] This implies that, while ML models may be excellent tools for prediction, they should not be expected to give trustworthy answers to counterfactual questions or to capture the effect of interventions, particularly under new conditions (out of distribution). This is also an issue with traditional statistical models, unless they are carefully developed, as well as process-based models that are often calibrated under specific growth conditions, but the flexibility and high-dimensionality of ML models make them more susceptible while also making it more difficult to detect overfitting behavior.

These caveats are particularly concerning for some agricultural modeling applications. For example, a common crop model use case is the assessment of yield impacts under potential future climate scenarios, which entails extrapolation from the observed data distribution. Simulations using process-based models, which embody biophysical relationships identified through scientific experiments and theoretical understanding, can be assumed to adhere to physical constraints under these conditions; the same is not true for ML models. Furthermore, ML models that cannot be shown to reliably answer counterfactual "what-if" questions

may be inappropriate for applications such as climate-change adaptation planning or attribution analysis of the causal chains leading to yield failures. Suitable datasets and evaluation strategies to benchmark the performance of ML models, as well as crop models, under such scenarios are needed.

Hybrid modeling approaches aim to avoid these limitations, but this is not guaranteed. For example, the common approach of training ML models using simulated data does not necessarily mean that the relationships embodied by the process model are effectively learned. This has been shown in studies building crop model emulators, which then had difficulty generalizing to unseen regions or years.[97,98] Very rigorous evaluation is needed to identify model overfitting, and commonly used strategies might not be reliable. Random or temporal cross-validation, for example, can lead to severe overestimation of model skill.[118] Bayesian neural networks have been argued to help mitigate overfitting but come with an increase in computational complexity[131] and, in reality, are still vulnerable to overfitting due to the incorrect assumption of independent and identically distributed data.

Overall, these pitfalls suggest that, although ML methods can be a powerful tool, they must be used with caution. Ideally, predictions should be produced only where input data lie within the training-data distribution.[117] If not, the model used should be evaluated in a procedure corresponding to its intended application (for example, using a test set split by space or time that reflects operational usage). Where this is not possible, or where other pitfalls apply to the intended application, it may be that the use of ML is not appropriate.

ACCELERATING PROGRESS WITH TRANSDISCIPLINARY COORDINATION

To conduct research at the intersection of process-based crop modeling and ML without falling victim to these pitfalls, a deep understanding of the challenges and assumptions underlying both agricultural science and ML methods is needed. Although the limitations we describe can have serious implications, their impact will depend on the specific task, dataset, and modeling approach used. Many of these pitfalls also apply to process-based modeling to a certain degree, and good modeling practice can be applied to address some of them.[132,133] However, we argue this is not always enough. More research is needed to understand the extent to which these limitations impact agricultural modeling tasks and how this is influenced by the use of different methodologies, model architectures, or other factors.

We argue that a transdisciplinary community is required, consisting of ML experts and agricultural modelers as well as stakeholders and experts from other fields who make use of crop model simulations. We identify three key areas in which collaborative activities conducted by such a community could advance research progress (Table 1).

Exchanging knowledge, data, and domain-specific best practices

ML educational material is increasingly easy to access, but popular textbooks, courses, and software libraries often neglect issues that are relevant for agricultural modeling (such as spatial autocorrelation, sparse datasets, or susceptibility to covariates). Researchers applying ML have found that models often fail to generalize outside of the dataset on which it was trained and have suggested that this may be due to underspecification of the intended task in the modeling pipeline.[134] In other words, the training dataset and evaluation strategy used are not rigorous enough and/or are not capturing stakeholder requirements in real-world use. Commonly used strategies to address pitfalls may not be helpful for all models or tasks. For example, a popular method of addressing the challenges of imbalanced binary classification is to balance the data using duplicate or synthetic samples. However, some recent research suggests that using this method does not lead to improved model performance for strong classifiers.[135] Better communication of the assumptions underlying both ML methods and agricultural datasets, together with collaborative research to identify domain-specific best practices, is needed.

The transfer of knowledge from crop modelers, agronomists, and scientists from related fields to ML researchers is also important. We have highlighted some of the ways in which ML methods can be used in complement to existing process-based crop models, such as the development of hybrid models. Here, targeting opportunities with the most potential impact is key. For example, modules or parameters of existing crop models that are known to less accurately reflect the intended processes may be more likely to lead to improved predictive skill (or other desired behavior) if replaced with ML. Effective model development that builds on the advantages of both paradigms requires a deep understanding of the strengths and weaknesses of both existing crop models and state-of-the-art ML methods.

Finally, accessing, handling, and processing agricultural data come with many domain-specific obstacles. Often, data used in individual studies cannot be shared due to privacy, ownership, and other governance issues. Publicly available data, where available, also tend to have issues related to size and quality. Coverage is often limited to major crops (such as wheat and maize). Other crops relevant for food security or nutritional diversity, such as millet and sorghum, may not have data available of the required size or quality. Most applications will require a range of data types (crop yields, weather, soil characteristics and moisture, planted and harvested areas, sowing and harvesting dates, irrigation and fertilization practices, to name a few) that vary in their spatial and temporal resolution, coverage, collection, and preprocessing needed. Combining these datasets for modeling is a daunting task that requires both deep and wide expertise. Data harmonization and exchange standards, protocols, and shared tools are also lacking.[136] AgMIP has made important contributions in this domain, but these have not been disseminated and used as widely as the challenge would require.

Robustly quantifying performance of ML methods for agricultural modeling tasks

So far, studies have reported varying abilities of different ML methods for agricultural modeling tasks, and reconciling these results is difficult due to the range of performance metrics, modeling methodologies, and datasets used. Community efforts to define benchmark datasets, evaluation strategies, and protocols that well represent the intended task are vitally needed. These benchmarks should adhere to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles, and have defined procedures for reproducing resulting model performance

Table 1. Transdisciplinary community activities to tackle ML pitfalls in agricultural modeling

Knowledge exchange	Quantifying ML performance	ML method development
Pitfall: ML model overfitting and poor generalization		
For specific applications, identify where and how overfitting or lack of generalization ability might be problematic. Find thresholds of required model generalization performance in discussion with stakeholders and model users.	Create and maintain benchmark simulated and observational datasets (and corresponding robust evaluation strategies) designed to detect unacceptable overfitting for specific applications.	Use simulated or large observational datasets to test and develop improved model evaluation and uncertainty quantification methods.
–	Intercompare existing ML methods on a wide variety of datasets and tasks.	Develop and intercompare hybrid models that enforce certain physical processes, model architectures that adhere to physical constraints or explicitly handle spatiotemporal autocorrelation.
Pitfall: Ambiguous model interpretations		
Assess and communicate the limitations of commonly used interpretable ML methods, identify state-of-the-art methods or inherently interpretable models that have promise for agricultural modeling applications.	Intercompare existing and state-of-the-art interpretable ML methods on simulated data where underlying processes are known.	Develop interpretable ML tools tailored to agricultural modeling tasks, research questions, and stakeholders.
Identify and make explicit the diverse needs of different users of model interpretations (e.g., model developers, scientists using interpretable ML for process understanding).	Investigate the relevance of providing model interpretations to different stakeholders and model users (e.g., increasing model trustworthiness).	Develop domain-specific interpretable ML evaluation and uncertainty quantification tools and platforms.
Pitfall: Predictive models not capturing causal mechanisms		
Establish guidelines for the use of predictive ML models for scientific research in agricultural modeling.	Measure and intercompare abilities of current ML methods for counterfactual and intervention analysis in different tasks using simulated or experimental datasets.	Develop and intercompare causal ML methods for interventional and counterfactual queries on agricultural modeling applications.
Identify agricultural modeling tasks that require causal models (e.g., estimating the effect of a proposed climate adaptation strategy) and assess if assumptions of causal methods may be violated (e.g., the presence of non-stationarity or feedback loops).	–	Develop and intercompare hybrid or knowledge-guided models, in close collaboration with domain experts, to enforce consistency with known biophysical processes.

scores. The curation and maintenance of these resources requires a transdisciplinary, coordinated effort with awareness of the specific needs of different model users and stakeholders.

For many crop-modeling applications, labeled ground-truth data are not available (for example, assessment of future climate-change impacts to crop yields). However, simulated data from process-based crop models could be used to create benchmark datasets with similar characteristics. Such set-ups enable in-depth study of the impact of evaluation strategy[118] or training data availability[98] on model behavior. Similar approaches have been used to provide evidence for improved phenotypic prediction across different environments[137] and to estimate bias in the modeled temperature sensitivity of global crop yields arising from uneven spatial representation of observation data.[138]

Facilitating the development of new ML methods for agricultural modeling applications

The quest for improvement in model performance on benchmark datasets drives and quantifies progress in ML research. Providing such datasets and evaluation criteria for specific agricultural modeling applications could enable ML experts with little domain knowledge to contribute, leading to the development of improved ML models.[139] For example, the recent release of a benchmark dataset for medium-range weather forecasting (WeatherBench[145]) has led to rapid technological advances, with several ML models now reporting performance comparable to, or better than, physics-based models.[61,62]

Experimental or simulated data from agricultural modeling studies could be repurposed for ML research in relevant subdomains, such as domain generalization, explainability, and robustness. For example, the use of computer vision for image-based plant phenotyping was accelerated by the compilation of benchmark datasets for a challenge hosted at a 2014 computer vision conference.[140] Since then, these datasets have been used in hundreds of published studies and are now a standard benchmark for multi-instance image segmentation.

Finally, ML researchers have called attention to the negative consequences of excessive focus on model scores on specific metrics and datasets. Benchmark dataset suites that represent

multiple locations, cultivars, and management strategies, including granular evaluation criteria aligned with the needs of specific end-users, could help to avoid ML models achieving the right answer for the wrong reasons and improve generalizability. However, we note that there is a tradeoff between the use of detailed evaluation criteria, which can help to democratize ML utilization by identifying performant approaches given the computation, data and expertise available, and the use of simple aggregate metrics that enable straightforward model intercomparison.

RECOMMENDATIONS AND WAYS FORWARD

ML methods are already driving scientific progress in the agricultural modeling domain, and, with effective transdisciplinary collaboration, we expect these tools to lead to significant research advances in the coming years. For researchers working at the intersection of ML and agricultural modeling, we have several recommendations.

First, consider any potential impact of the fundamental limitations of ML methods on the intended use case of the model. There are many ways to integrate the scientific knowledge of biophysical processes embodied by existing crop models with ML (via parameterization, data generation, calibration, data assimilation, post-processing, or the use of hybrid modeling approaches) that may lead to better outcomes than a purely data-driven approach.

Secondly, ML models should be evaluated more rigorously, with evaluation criteria that take into account the specific requirements and difficulties of stakeholders (such as the necessary lead time for forecasting, or time lag in data availability) as well as the presence of spatiotemporal autocorrelation or other dependencies between features. Ideally, multiple metrics should be used. For example, model performance under climate extremes, such as heatwaves or droughts, could be reported separately.[141] Predictions used for model evaluation could be made available in public repositories to enable follow-up analyses.[142]

It is important to note that defining and weighting diverse and context-specific model criteria is not strictly a scientific question but a matter of balancing the priorities of multiple stakeholders.[143] Optimizing agricultural yields, for example, can lead to biodiversity loss, degraded soils, and increased pollution.[144,145] Therefore, deciding on criteria by which models are evaluated should take place in an open conversation between ML researchers, crop modelers, stakeholders from multiple communities, and experts from other scientific disciplines such as soil scientists, climatologists, agronomists, and ecologists. This transdisciplinary dialogue is likely to be as important for advancing scientific understanding as the resulting models, so these discussions and corresponding lessons learned should be documented.[146] Furthermore, quantifying model requirements can be challenging, and stakeholder priorities will change over time. The emphasis should be on establishing an efficient process for iterative improvement rather than the creation of a single definitive product.

For this type of undertaking, global and transdisciplinary coordination is needed. In 2023, we established the AgMIP ML community (AgML) as a first step toward addressing the above-mentioned challenges. While we were met with considerable enthusiasm and engagement from both ML researchers and crop modelers, channeling collective ideas and discussion into useful output is more difficult. The following recommendations are based on our experience launching and working with this large, open, and diverse community.

Focus on application-driven challenges

It is important to decide on a specific problem or research question early on in the process, which should be tied to the needs of a real-world model user or researcher. For example, if we consider crop yield forecasting, the user could be an individual farmer trying to maximize profits or a governmental institution hoping to anticipate and avoid food insecurity. This decision will dictate the scale and type of data required, the forecasting lead time, suitable model architectures and corresponding learning algorithms, and the evaluation criteria. However, despite the narrow scope, receiving regular input from a wide range of participants is key. To facilitate this, we have found it useful to hold regular open meetings where the progress of multiple activities can be discussed, along with other relevant topics. This provides enough value to merit continual high attendance and has allowed small issues to be noticed by community members with specific expertise. We note that motivating engagement and collaboration between researchers is easier than with real-world users of crop models and other non-academic stakeholders.

Develop a common language

Although transdisciplinary community discussion is important, in order to produce legible output (such as the publication of benchmark datasets), the problems tackled must be explained and motivated in field-specific language. To attract interest from the wider ML community, publications in ML conferences would be prioritized. These are highly competitive and have a fast-paced review system. Reviewers are unlikely to be familiar with crop modeling or adjacent topics, so this context and the research problem should be explained clearly using domain-specific vocabulary. A more general problem may be that interdisciplinary research careers may not fulfill field-specific criteria needed to obtain professorship positions.

Reward interdisciplinary work and dataset curation

Producing high-quality research and datasets in coordination with a large, interdisciplinary community has long been disincentivized by academic structures. Research is driven by a need to publish quickly and preferentially focuses on novelty over iterative improvement. The actual costs involved in collecting agricultural data and the inherent time lag in generating them (i.e., a growing season may span several months) often do not receive proper attention. In ML, the hard work of benchmark dataset curation has been undervalued, although there has been a recent push for "data-centric ML," which aims to improve this. Similarly, data journals in the environmental sciences offer prominent venues and credit for the often-laborious creation of relevant auxiliary data.

In conclusion, we advocate for open, accessible communities to facilitate important dialogues, help scientists to reap the potential benefits of ML for agricultural modeling while avoiding pitfalls, and conduct activities such as benchmark dataset creation

and model intercomparison experiments. These actions will help quantify the domain-specific utility of current ML methods and build trust in the use of these tools where warranted. We believe that such efforts will enable the development of more robust, usable, and trustworthy crop models that can be used to address the multiple challenges that face our global food system.

ACKNOWLEDGMENTS

We acknowledge funding for a Short-Term Scientific Mission from the European COST Action DAMOCLES (CA17109) provided to L.-b.S., which helped initiate and develop this paper. L.-b.S. and J.Z. acknowledge the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, grant agreement VH-NG-1537). P.M. acknowledges support from the metaprogram Agriculture and Forestry in the Face of Climate Change: Adaptation and Mitigation (CLIMAE) of INRAE. A.C.R. and A.C. receive NASA GISS Climate Impacts Group support from the Earth System Directorate. I.N.A. acknowledges support from the European Union's Digital Europe Programme under grant agreement N° 101100622 (AgrifoodTEF). D.P. acknowledges support from Wageningen University and Research investment theme Data Driven Discoveries in a Changing Climate. This manuscript benefited from the input and feedback from the agricultural and ML modeling community at the ninth AgMIP Global Workshop, the first AgML workshop in January 2024 in Wageningen, and from regular AgML community meetings.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. United Nations (2015). Transforming Our World: The 2030 Agenda for Sustainable Development. Resolution Adopted by the General Assembly on 25 September 2015, 42809, 1-13. https://doi.org/10.1007/s13398-014-0173-7.2
2. Ray, D.K., Sloat, L.L., Garcia, A.S., Davis, K.F., Ali, T., and Xie, W. (2022). Crop harvests for direct food use insufficient to meet the UN's food security goal. Nat. Food *3*, 367–374. https://doi.org/10.1038/s43016-022-00504-z.
3. Ray, D.K., Mueller, N.D., West, P.C., and Foley, J.A. (2013). Yield Trends Are Insufficient to Double Global Crop Production by 2050. PLoS One *8*, e66428. https://doi.org/10.1371/journal.pone.0066428.
4. Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F.-X., and Huard, F. (2010). Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. Field Crops Res. *119*, 201–212. https://doi.org/10.1016/j.fcr.2010.07.012.
5. Ray, D.K., West, P.C., Clark, M., Gerber, J.S., Prishchepov, A.V., and Chatterjee, S. (2019). Climate change has likely already affected global food production. PLoS One *14*, e0217148. https://doi.org/10.1371/journal.pone.0217148.
6. Ortiz-Bobea, A., Ault, T.R., Carrillo, C.M., Chambers, R.G., and Lobell, D.B. (2021). Anthropogenic climate change has slowed global agricultural productivity growth. Nat. Clim. Change *11*, 306–312. https://doi.org/10.1038/s41558-021-01000-1.
7. Kummu, M., Heino, M., Taka, M., Varis, O., and Viviroli, D. (2021). Climate change risks pushing one-third of global food production outside the safe climatic space. One Earth *4*, 720–729. https://doi.org/10.1016/j.oneear.2021.04.017.
8. Rezaei, E.E., Webber, H., Asseng, S., Boote, K., Durand, J.L., Ewert, F., Martre, P., and MacCarthy, D.S. (2023). Climate change impacts on crop yields. Nat. Rev. Earth Environ. *4*, 831–846. https://doi.org/10.1038/s43017-023-00491-0.
9. Mbow, C., Rosenzweig, C., Barioni, L.G., Benton, T.G., Herrero, M., Krishnapillai, M., Liwenga, E., Pradhan, P., Rivera-Ferre, M.G., Sapkota, T., et al. (2019). Food security. In Climate Change and Land: An IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems, P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.-O. Pörtner, D.C. Roberts, P. Zhai, R. Slade, S. Connors, and R. van Diemen, et al., eds. (Intergovernmental Panel on Climate Change).
10. Bezner Kerr, R., Hasegawa, T., and Lasco, R. (2022). Food, fibre, and other ecosystem products Climate Change 2022: Impacts, Adaptation, and Vulnerability. In Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change ed HO Pörtner et al.
11. Jägermeyr, J., Müller, C., Ruane, A.C., Elliott, J., Balkovic, J., Castillo, O., Faye, B., Foster, I., Folberth, C., Franke, J.A., et al. (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. Nat. Food *2*, 873–885. https://doi.org/10.1038/s43016-021-00400-y.
12. Zhu, P., Burney, J., Chang, J., Jin, Z., Mueller, N.D., Xin, Q., Xu, J., Yu, L., Makowski, D., and Ciais, P. (2022). Warming reduces global agricultural production by decreasing cropping frequency and yields. Nat. Clim. Change *12*, 1016–1023. https://doi.org/10.1038/s41558-022-01492-5.
13. Asseng, S., Ewert, F., Martre, P., Rötter, R.P., Lobell, D.B., Cammarano, D., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., et al. (2015). Rising temperatures reduce global wheat production. Nat. Clim. Change *5*, 143–147. https://doi.org/10.1038/nclimate2470.
14. Helman, D., and Bonfil, D.J. (2022). Six decades of warming and drought in the world's top wheat-producing countries offset the benefits of rising CO2 to yield. Sci. Rep. *12*, 7921. https://doi.org/10.1038/s41598-022-11423-1.
15. Seneviratne, S.I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., et al. (2021). Weather and Climate Extreme Events in a Changing Climate. In Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Cambridge University Press), pp. 1513–1766. https://doi.org/10.1017/9781009157896.013.
16. Hasegawa, T., Sakurai, G., Fujimori, S., Takahashi, K., Hijioka, Y., and Masui, T. (2021). Extreme climate events increase risk of global food insecurity and adaptation needs. Nat. Food *2*, 587–595. https://doi.org/10.1038/s43016-021-00335-4.
17. Brás, T.A., Seixas, J., Carvalhais, N., and Jägermeyr, J. (2021). Severity of drought and heatwave crop losses tripled over the last five decades in Europe. Environ. Res. Lett. *16*, 065012. https://doi.org/10.1088/1748-9326/abf004.
18. Ranasinghe, R., Ruane, A.C., Vautard, R., Arnell, N., Coppola, E., Cruz, F.A., Dessai, S., Saiful Islam, A., Rahimi, M., Carrascal, D.R., et al. (2021). Climate change information for regional impact and for risk assessment. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Cambridge University Press)), pp. 1767–1926. https://doi.org/10.1017/9781009157896.014.
19. (2020). Systems thinking, systems doing. Nat. Food *1*, 659. https://doi.org/10.1038/s43016-020-00190-9.
20. Ewert, F., Baatz, R., and Finger, R. (2023). Agroecology for a Sustainable Agriculture and Food System: From Local Solutions to Large-Scale Adoption. Annu. Rev. Resour. Econ. *15*, 351–381. https://doi.org/10.1146/annurev-resource-102422-090105.
21. Agathokleous, E., Frei, M., Knopf, O.M., Muller, O., Xu, Y., Nguyen, T.H., Gaiser, T., Liu, X., Liu, B., Saitanis, C.J., et al. (2023). Adapting crop production to climate change and air pollution at different scales. Nat. Food *4*, 854–865. https://doi.org/10.1038/s43016-023-00858-y.
22. Liu, J., Hull, V., Godfray, H.C.J., Tilman, D., Gleick, P., Hoff, H., Pahl-Wostl, C., Xu, Z., Chung, M.G., Sun, J., and Li, S. (2018). Nexus approaches to global sustainable development. Nat. Sustain. *1*, 466–476. https://doi.org/10.1038/s41893-018-0135-8.
23. Ruane, A.C., and Rosenzweig, C. (2019). Climate change impacts on agriculture. In Agriculture & Food Systems to 2050: Global Trends, Challenges and Opportunities (World Scientific), pp. 161–191.
24. Schlüter, M., Brelsford, C., Ferraro, P.J., Orach, K., Qiu, M., and Smith, M.D. (2023). Unraveling complex causal processes that affect sustainability requires more integration between empirical and modeling approaches. Proc. Natl. Acad. Sci. USA *120*, e2215676120. https://doi.org/10.1073/pnas.2215676120.
25. Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., et al. (2017). Brief history of agricultural systems modeling. Agric. Syst. *155*, 240–254. https://doi.org/10.1016/j.agsy.2016.05.014.
26. Antle, J.M., Basso, B., Conant, R.T., Godfray, H.C.J., Jones, J.W., Herrero, M., Howitt, R.E., Keating, B.A., Munoz-Carpena, R., Rosenzweig, C., et al. (2017). Towards a new generation of agricultural system data, models and knowledge products: Design and improvement. Agric. Syst. *155*, 255–268. https://doi.org/10.1016/j.agsy.2016.10.002.
27. Antle, J.M., Jones, J.W., and Rosenzweig, C. (2017). Next generation agricultural system models and knowledge products: Synthesis and

strategy. Agric. Syst. *155*, 179–185. https://doi.org/10.1016/j.agsy.2017.05.006.

28. Webber, H., Rezaei, E.E., Ryo, M., and Ewert, F. (2022). Framework to guide modeling single and multiple abiotic stresses in arable crops. Agric. Ecosyst. Environ. *340*, 108179. https://doi.org/10.1016/j.agee.2022.108179.
29. Pasquel, D., Roux, S., Richetti, J., Cammarano, D., Tisseyre, B., and Taylor, J.A. (2022). A review of methods to evaluate crop model performance at multiple and changing spatial scales. Precis. Agric. *23*, 1489–1513. https://doi.org/10.1007/s11119-022-09885-4.
30. Müller, C., Franke, J., Jägermeyr, J., Ruane, A.C., Elliott, J., Moyer, E., Heinke, J., Falloon, P.D., Folberth, C., Francois, L., et al. (2021). Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. Environ. Res. Lett. *16*, 034040. https://doi.org/10.1088/1748-9326/abd8fc.
31. Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. Artif. Intell. Agric *6*, 257–265. https://doi.org/10.1016/j.aiia.2022.11.003.
32. Van Noorden, R., and Perkel, J.M. (2023). AI and science: what 1,600 researchers think. Nature *621*, 672–675. https://doi.org/10.1038/d41586-023-02980-0.
33. Liu, W., Ye, T., Müller, C., Jägermeyr, J., Franke, J.A., Stephens, H., and Chen, S. (2023). he statistical emulators of GGCMI phase 2: responses of year-to-year variation of crop yield to CO 2, temperature, water, and nitrogen perturbations. Geosci. Model Dev. (GMD) *16*, 7203–7221. https://doi.org/10.5194/gmd-16-7203-2023.
34. Pylianidis, C., Snow, V., Overweg, H., Osinga, S., Kean, J., and Athanasiadis, I.N. (2022). Simulation-assisted machine learning for operational digital twins. Environ. Model. Software *148*, 105274. https://doi.org/10.1016/j.envsoft.2021.105274.
35. Deines, J.M., Swatantran, A., Ye, D., Myers, B., Archontoulis, S., and Lobell, D.B. (2023). Field-scale dynamics of planting dates in the US Corn Belt from 2000 to 2020. Remote Sens. Environ. *291*, 113551. https://doi.org/10.1016/j.rse.2023.113551.
36. Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., and Sanderman, J. (2021). Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil Sci. *72*, 1607–1623. https://doi.org/10.1111/ejss.12998.
37. Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M. (2019). Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. Agric. For. Meteorol. *264*, 1–15. https://doi.org/10.1016/j.agrformet.2018.09.021.
38. Nevavuori, P., Narra, N., and Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. Comput. Electron. Agric. *163*, 104859. https://doi.org/10.1016/j.compag.2019.104859.
39. Cartuyvels, R., Fierens, T., Coppieters, E., Moens, M.-F., and Sileo, D. (2023). Spatiotemporal self-supervised pre-training on satellite imagery improves food insecurity prediction. Br. J. Hist. Sci. *2*, e48. https://doi.org/10.1017/eds.2023.42.
40. Balashankar, A., Subramanian, L., and Fraiberger, S.P. (2023). Predicting food crises using news streams. Sci. Adv. *9*, eabm3449. https://doi.org/10.1126/sciadv.abm3449.
41. Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., et al. (2014). APSIM – Evolution towards a new generation of agricultural systems simulation. Environ. Model. Software *62*, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009.
42. Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., and Foster, I. (2014). The parallel system for integrating impact models and sectors (pSIMS). Environ. Model. Software *62*, 509–516. https://doi.org/10.1016/j.envsoft.2014.04.008.
43. Chenu, K., Porter, J.R., Martre, P., Basso, B., Chapman, S.C., Ewert, F., Bindi, M., and Asseng, S. (2017). Contribution of Crop Models to Adaptation in Wheat. Trends Plant Sci. *22*, 472–490. https://doi.org/10.1016/j.tplants.2017.02.003.
44. Smerald, A., Kraus, D., Rahimi, J., Fuchs, K., Kiese, R., Butterbach-Bahl, K., and Scheer, C. (2023). A redistribution of nitrogen fertiliser across global croplands can help achieve food security within environmental boundaries. Commun. Earth Environ. *4*, 315. https://doi.org/10.1038/s43247-023-00970-8.
45. Minoli, S., Jägermeyr, J., Asseng, S., Urfels, A., and Müller, C. (2022). Global crop yields can be lifted by timely adaptation of growing periods to climate change. Nat. Commun. *13*, 7079. https://doi.org/10.1038/s41467-022-34411-5.
46. Ivanovich, C.C., Sun, T., Gordon, D.R., and Ocko, I.B. (2023). Future warming from global food consumption. Nat. Clim. Change *13*, 297–302. https://doi.org/10.1038/s41558-023-01605-8.
47. Mehrabi, Z., Delzeit, R., Ignaciuk, A., Levers, C., Braich, G., Bajaj, K., Amo-Aidoo, A., Anderson, W., Balgah, R.A., Benton, T.G., et al. (2022). Research priorities for global food security under extreme events. One Earth *5*, 756–766. https://doi.org/10.1016/j.oneear.2022.06.008.
48. Muller, B., and Martre, P. (2019). Plant and crop simulation models: powerful tools to link physiology, genetics, and phenomics. J. Exp. Bot. *70*, 2339–2344. https://doi.org/10.1093/jxb/erz175.
49. White, J.W., Hoogenboom, G., Kimball, B.A., and Wall, G.W. (2011). Methodologies for simulating impacts of climate change on crop production. Field Crops Res. *124*, 357–368. https://doi.org/10.1016/j.fcr.2011.07.001.
50. White, J.W., Hunt, L.A., Boote, K.J., Jones, J.W., Koo, J., Kim, S., Porter, C.H., Wilkens, P.W., and Hoogenboom, G. (2013). Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. Comput. Electron. Agric. *96*, 1–12. https://doi.org/10.1016/j.compag.2013.04.003.
51. Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., et al. (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proc. Natl. Acad. Sci. USA *111*, 3268–3273. https://doi.org/10.1073/pnas.1222463110.
52. Rosenzweig, C., and Parry, M.L. (1994). Potential impact of climate change on world food supply. Nature *367*, 133–138. https://doi.org/10.1038/367133a0.
53. Rötter, R.P., Carter, T.R., Olesen, J.E., and Porter, J.R. (2011). Crop–climate models need an overhaul. Nat. Clim. Change *1*, 175–177. https://doi.org/10.1038/nclimate1152.
54. Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., et al. (2013). The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agric. For. Meteorol. *170*, 166–182. https://doi.org/10.1016/j.agrformet.2012.09.011.
55. Ruane, A.C., Rosenzweig, C., Asseng, S., Boote, K.J., Elliott, J., Ewert, F., Jones, J.W., Martre, P., McDermid, S.P., Müller, C., et al. (2017). An AgMIP framework for improved agricultural representation in integrated assessment models. Environ. Res. Lett. *12*, 125003. https://doi.org/10.1088/1748-9326/aa8da6.
56. Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R.P., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., et al. (2017). The uncertainty of crop yield projections is reduced by improved temperature response functions. Nat. Plants *3*, 1–13. https://doi.org/10.1038/nplants.2017.102.
57. Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R.P., Ruane, A.C., Semenov, M.A., Wallach, D., Wang, E., et al. (2017). Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. Field Crops Res. *202*, 5–20. https://doi.org/10.1016/j.fcr.2016.05.001.
58. Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S., et al. (2017). Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. Geosci. Model Dev. (GMD) *10*, 1403–1422. https://doi.org/10.5194/gmd-10-1403-2017.
59. Franke, J.A., Müller, C., Elliott, J., Ruane, A.C., Jägermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P.D., Folberth, C., et al. (2020). The GGCMI Phase 2 experiment: global gridded crop model simulations under uniform changes in CO 2, temperature, water, and nitrogen levels (protocol version 1.0). Geosci. Model Dev. (GMD) *13*, 2315–2336. https://doi.org/10.5194/gmd-13-2315-2020.
60. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
61. Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2023). Learning skillful medium-range global weather forecasting. Science *382*, 1416–1421. https://doi.org/10.1126/science.adi2336.
62. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. Nature *619*, 533–538. https://doi.org/10.1038/s41586-023-06185-3.
63. Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2207.08815.

64. Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. Int. J. Forecast. *38*, 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013.

65. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

66. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F.A. (2020). Shortcut learning in deep neural networks. Nat. Mach. Intell. *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z.

67. Roscher, R., Bohn, B., Duarte, M.F., and Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.

68. Flora, M.L., Potvin, C.K., McGovern, A., and Handler, S. (2024). A Machine Learning Explainability Tutorial for Atmospheric Sciences. Artif. Intell. Earth Syst. *3*, e230018. https://doi.org/10.1175/AIES-D-23-0018.1.

69. Jiang, S., Sweet, L., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., and Zscheischler, J. (2024). How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences. Earths Future *12*, e2024EF004540. https://doi.org/10.1029/2024EF004540.

70. Paudel, D., de Wit, A., Boogaard, H., Marcos, D., Osinga, S., and Athanasiadis, I.N. (2023). Interpretability of deep learning models for crop yield forecasting. Comput. Electron. Agric. *206*, 107663. https://doi.org/10.1016/j.compag.2023.107663.

71. Silva, J.V., and Giller, K.E. (2020). Grand challenges for the 21st century: what crop models can and can't (yet) do. J. Agric. Sci. *158*, 794–805. https://doi.org/10.1017/S0021859621000150.

72. Fu, J., Jian, Y., Wang, X., Li, L., Ciais, P., Zscheischler, J., Wang, Y., Tang, Y., Müller, C., Webber, H., et al. (2023). Extreme rainfall reduces one-twelfth of China's rice yield over the last two decades. Nat. Food *4*, 416–426. https://doi.org/10.1038/s43016-023-00753-6.

73. Liu, W., Sun, W., Huang, J., Wen, H., and Huang, R. (2021). Excessive rainfall is the key meteorological limiting factor for winter wheat yield in the middle and lower reaches of the yangtze river. Agronomy *12*, 50. https://doi.org/10.3390/agronomy12010050.

74. Heinicke, S., Frieler, K., Jägermeyr, J., and Mengel, M. (2022). Global gridded crop models underestimate yield responses to droughts and heatwaves. Environ. Res. Lett. *17*, 044026. https://doi.org/10.1088/1748-9326/ac592e.

75. Jin, Z., Zhuang, Q., Tan, Z., Dukes, J.S., Zheng, B., and Melillo, J.M. (2016). Do maize models capture the impacts of heat and drought stresses on yield? Using algorithm ensembles to identify successful approaches. Glob. Change Biol. *22*, 3112–3126. https://doi.org/10.1111/gcb.13376.

76. Barlow, K.M., Christy, B.P., O'Leary, G., Riffkin, P.A., and Nuttall, J.G. (2015). Simulating the impact of extreme heat and frost events on wheat crop production: A review. Field Crops Res. *171*, 109–119. https://doi.org/10.1016/j.fcr.2014.11.010.

77. Eitzinger, J., Thaler, S., Schmid, E., Strauss, F., Ferrise, R., Moriondo, M., Bindi, M., Palosuo, T., Rötter, R., Kersebaum, K.C., et al. (2013). Sensitivities of crop models to extreme weather conditions during flowering period demonstrated for maize and winter wheat in Austria. J. Agric. Sci. *151*, 813–835. https://doi.org/10.1017/S0021859612000779.

78. van der Wiel, K., Selten, F.M., Bintanja, R., Blackport, R., and Screen, J.A. (2020). Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions. Environ. Res. Lett. *15*, 034050. https://doi.org/10.1088/1748-9326/ab7668.

79. Nóia Júnior, R. de S., Deswarte, J.-C., Cohan, J.-P., Martre, P., van der Velde, M., Lecerf, R., Webber, H., Ewert, F., Ruane, A.C., Slafer, G.A., et al. (2023). The extreme 2016 wheat yield failure in France. Glob. Change Biol. *29*, 3130–3146. https://doi.org/10.1111/gcb.16662.

80. Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van der Velde, M., and Makowski, D. (2018). Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. Nat. Commun. *9*, 1627. https://doi.org/10.1038/s41467-018-04087-x.

81. Matiu, M., Ankerst, D.P., and Menzel, A. (2017). Interactions between temperature and drought in global and regional crop yield variability during 1961-2014. PLoS One *12*, e0178339. https://doi.org/10.1371/journal.pone.0178339.

82. Mills, G., Sharps, K., Simpson, D., Pleijel, H., Frei, M., Burkey, K., Emberson, L., Uddling, J., Broberg, M., Feng, Z., et al. (2018). Closing the global ozone yield gap: Quantification and cobenefits for multistress tolerance. Glob. Change Biol. *24*, 4869–4893. https://doi.org/10.1111/gcb.14381.

83. Toreti, A., Deryng, D., Tubiello, F.N., Müller, C., Kimball, B.A., Moser, G., Boote, K., Asseng, S., Pugh, T.A.M., Vanuytrecht, E., et al. (2020). Narrowing uncertainties in the effects of elevated CO2 on crops. Nat. Food *1*, 775–782. https://doi.org/10.1038/s43016-020-00195-4.

84. Donatelli, M., Magarey, R.D., Bregaglio, S., Willocquet, L., Whish, J.P.M., and Savary, S. (2017). Modelling the impacts of pests and diseases on agricultural systems. Agric. Syst. *155*, 213–224. https://doi.org/10.1016/j.agsy.2017.01.019.

85. Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K.J., Büchner, M., Foster, I., Glotter, M., Heinke, J., Iizumi, T., et al. (2015). The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0). Geosci. Model Dev. (GMD) *8*, 261–277. https://doi.org/10.5194/gmd-8-261-2015.

86. Müller, C., Elliott, J., Kelly, D., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Hoek, S., Izaurralde, R.C., et al. (2019). The Global Gridded Crop Model Intercomparison phase 1 simulation dataset. Sci. Data *6*, 50. https://doi.org/10.1038/s41597-019-0023-8.

87. Xiong, W., Asseng, S., Hoogenboom, G., Hernandez-Ochoa, I., Robertson, R., Sonder, K., Pequeno, D., Reynolds, M., and Gerard, B. (2019). Different uncertainty distribution between high and low latitudes in modelling warming impacts on wheat. Nat. Food *1*, 63–69. https://doi.org/10.1038/s43016-019-0004-2.

88. Iizumi, T., Tanaka, Y., Sakurai, G., Ishigooka, Y., and Yokozawa, M. (2014). Dependency of parameter values of a crop model on the spatial scale of simulation. J. Adv. Model. Earth Syst. *6*, 527–540. https://doi.org/10.1002/2014MS000311.

89. Seidel, S.J., Palosuo, T., Thorburn, P., and Wallach, D. (2018). Towards improved calibration of crop models – Where are we now and where should we go? Eur. J. Agron. *94*, 25–35. https://doi.org/10.1016/j.eja.2018.01.006.

90. Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., et al. (2021). The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environ. Model. Software *145*, 105206. https://doi.org/10.1016/j.envsoft.2021.105206.

91. Folberth, C., Elliott, J., Müller, C., Balkovič, J., Chryssanthacopoulos, J., Izaurralde, R.C., Jones, C.D., Khabarov, N., Liu, W., Reddy, A., et al. (2019). Parameterization-induced uncertainties and impacts of crop management harmonization in a global gridded crop model ensemble. PLoS One *14*, e0221862. https://doi.org/10.1371/journal.pone.0221862.

92. Jeong, S., Ko, J., Shin, T., and Yeom, J.M. (2022). Incorporation of machine learning and deep neural network approaches into a remote sensing-integrated crop model for the simulation of rice growth. Sci. Rep. *12*, 9030. https://doi.org/10.1038/s41598-022-13232-y.

93. Kamir, E., Waldner, F., and Hochman, Z. (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J. Photogrammetry Remote Sens. *160*, 124–135. https://doi.org/10.1016/j.isprsjprs.2019.11.008.

94. You, J., Li, X., Low, M., Lobell, D., and Ermon, S. (2017). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. Proceedings of the AAAI Conference on Artificial Intelligence 31. https://doi.org/10.1609/aaai.v31i1.11172.

95. Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L.B., Obersteiner, M., and van der Velde, M. (2016). Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. Nat. Commun. *7*, 11872. https://doi.org/10.1038/ncomms11872.

96. Constantin, J., Raynal, H., Casellas, E., Hoffmann, H., Bindi, M., Doro, L., Eckersten, H., Gaiser, T., Grosz, B., Haas, E., et al. (2019). Management and spatial resolution effects on yield and water balance at regional scale in crop models. Agric. For. Meteorol. *275*, 184–195. https://doi.org/10.1016/j.agrformet.2019.05.013.

97. Xu, H., Zhang, X., Ye, Z., Jiang, L., Qiu, X., Tian, Y., Zhu, Y., and Cao, W. (2021). Machine learning approaches can reduce environmental data requirements for regional yield potential simulation. Eur. J. Agron. *129*, 126335. https://doi.org/10.1016/j.eja.2021.126335.

98. Shahhosseini, M., Martinez-Feria, R.A., Hu, G., and Archontoulis, S.V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. *14*, 124026. https://doi.org/10.1088/1748-9326/ab5268.

99. Maestrini, B., Mimić, G., van Oort, P.A., Jindo, K., Brdar, S., Athanasiadis, I.N., and van Evert, F.K. (2022). Mixing process-based and data-driven approaches in yield prediction. Eur. J. Agron. *139*, 126569. https://doi.org/10.1016/j.eja.2022.126569.

100. Elrys, A.S., Wang, J., Meng, L., Zhu, Q., El-Sawy, M.M., Chen, Z., Tu, X., El-Saadony, M.T., Zhang, Y., Zhang, J., et al. (2023). Integrative knowledge-based nitrogen management practices can provide positive effects on ecosystem nitrogen retention. Nat. Food *4*, 1075–1089. https://doi.org/10.1038/s43016-023-00888-6.

101. Kinnunen, P., Heino, M., Sandström, V., Taka, M., Ray, D.K., and Kummu, M. (2022). Crop Yield Loss Risk Is Modulated by Anthropogenic Factors. Earths Future *10*, e2021EF002420. https://doi.org/10.1029/2021EF002420.

102. Yin, X., Leng, G., and Yu, L. (2022). Disentangling the separate and confounding effects of temperature and precipitation on global maize yield using machine learning, statistical and process crop models. Environ. Res. Lett. *17*, 044036. https://doi.org/10.1088/1748-9326/ac5716.

103. Guilpart, N., Iizumi, T., and Makowski, D. (2022). Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change. Nat. Food *3*, 255–265. https://doi.org/10.1038/s43016-022-00481-3.

104. Vogel, E., Donat, M.G., Alexander, L.V., Meinshausen, M., Ray, D.K., Karoly, D., Meinshausen, N., and Frieler, K. (2019). The effects of climate extremes on global agricultural yields. Environ. Res. Lett. *14*, 054010. https://doi.org/10.1088/1748-9326/ab154b.

105. Zelingher, R., and Makowski, D. (2023). Investigating and forecasting the impact of crop production shocks on global commodity prices. Environ. Res. Lett. *19*, 014026. https://doi.org/10.1088/1748-9326/ad0dda.

106. Newman, S.J., and Furbank, R.T. (2021). Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. Nat. Plants *7*, 1354–1363. https://doi.org/10.1038/s41477-021-01001-0.

107. Farrell, A.D., Deryng, D., and Neufeldt, H. (2023). Modelling adaptation and transformative adaptation in cropping systems: recent advances and future directions. Curr. Opin. Environ. Sustain. *61*, 101265. https://doi.org/10.1016/j.cosust.2023.101265.

108. Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Trans. Knowl. Data Eng. *29*, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168.

109. Feng, P., Wang, B., Liu, D.L., Waters, C., and Yu, Q. (2019). Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. Agric. For. Meteorol. *275*, 100–113. https://doi.org/10.1016/j.agrformet.2019.05.018.

110. Shahhosseini, M., Hu, G., Huber, I., and Archontoulis, S.V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci. Rep. *11*, 1606. https://doi.org/10.1038/s41598-020-80820-1.

111. Droutsas, I., Challinor, A.J., Deva, C.R., and Wang, E. (2022). Integration of machine learning into process-based modelling to improve simulation of complex crop responses. Silico Plants *4*, diac017. https://doi.org/10.1093/insilicoplants/diac017.

112. Midingoyi, C.A., Pradal, C., Enders, A., Fumagalli, D., Lecharpentier, P., Raynal, H., Donatelli, M., Fanchini, D., Athanasiadis, I.N., Porter, C., et al. (2023). Crop modeling frameworks interoperability through bidirectional source code transformation. Environ. Model. Software *168*, 105790. https://doi.org/10.1016/j.envsoft.2023.105790.

113. Midingoyi, C.A., Pradal, C., Enders, A., Fumagalli, D., Raynal, H., Donatelli, M., Athanasiadis, I.N., Porter, C., Hoogenboom, G., Holzworth, D., et al. (2021). Crop2ML: An open-source multi-language modeling framework for the exchange and reuse of crop model components. Environ. Model. Software *142*, 105055. https://doi.org/10.1016/j.envsoft.2021.105055.

114. Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., Zhu, Q., Till, J., Jia, X., Jiang, C., et al. (2024). Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. Nat. Commun. *15*, 357. https://doi.org/10.1038/s41467-023-43860-5.

115. Moon, T., Kim, D., Kwon, S., and Son, J.E. (2023). Process-Based Crop Modeling for High Applicability with Attention Mechanism and Multitask Decoders. Plant Phenomics *5*, 0035. https://doi.org/10.34133/plantphenomics.0035.

116. Han, J., Shi, L., Pylianidis, C., Yang, Q., and Athanasiadis, I.N. (2023). DeepOryza: A Knowledge guided machine learning model for rice growth simulation. In 2nd AAAI Workshop on AI for Agriculture and Food Systems. https://openreview.net/forum?id=L9ankU4Ge-v

117. Meyer, H., and Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. *13*, 2208. https://doi.org/10.1038/s41467-022-29838-9.

118. Sweet, L.B., Müller, C., Anand, M., and Zscheischler, J. (2023). Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. Artif. Intell. Earth Syst. *2*. https://doi.org/10.1175/AIES-D-23-0026.1.

119. Maier, H.R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., Loritz, R., Wu, W., Guo, D., Bennett, A., et al. (2023). On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization. Environ. Model. Software *167*, 105779. https://doi.org/10.1016/j.envsoft.2023.105779.

120. Kapoor, S., and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.

121. Chekroud, A.M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P.R., Koutsouleris, N., Krumholz, H.M., et al. (2024). Illusory generalizability of clinical prediction models. Science *383*, 164–167. https://doi.org/10.1126/science.adg8538.

122. Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., and Athanasiadis, I.N. (2021). Machine learning for large-scale crop yield forecasting. Agric. Syst. *187*, 103016. https://doi.org/10.1016/j.agsy.2020.103016.

123. Lischeid, G., Webber, H., Sommer, M., Nendel, C., and Ewert, F. (2022). Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. Agric. For. Meteorol. *312*, 108698. https://doi.org/10.1016/j.agrformet.2021.108698.

124. Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence AAAI'19/IAAI'19/EAAI'19 (AAAI Press), pp. 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681.

125. Hooker, G., Mentch, L., and Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat. Comput. *31*, 82. https://doi.org/10.1007/s11222-021-10057-z.

126. Silva, S.J., and Keller, C.A. (2024). Limitations of XAI Methods for Process-Level Understanding in the Atmospheric Sciences. Artif. Intell. Earth Syst. *3*. https://doi.org/10.1175/AIES-D-23-0045.1.

127. Mamalakis, A., Barnes, E.A., and Ebert-Uphoff, I. (2023). Carefully Choose the Baseline: Lessons Learned from Applying XAI Attribution Methods for Regression Tasks in Geoscience. Artif. Intell. Earth Syst. *2*, 1–18. https://doi.org/10.1175/AIES-D-22-0058.1.

128. Eiband, M., Buschek, D., Kremer, A., and Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems CHI EA '19 (Association for Computing Machinery)), pp. 1–6. https://doi.org/10.1145/3290607.3312787.

129. Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Marí, J., et al. (2019). Inferring causation from time series in Earth system sciences. Nat. Commun. *10*, 2553. https://doi.org/10.1038/s41467-019-10105-3.

130. McGovern, A., Ebert-Uphoff, I., Gagne, D.J., and Bostrom, A. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. J. Inflamm. Res. *1*, e6. https://doi.org/10.1017/eds.2022.5.

131. Izmailov, P., Vikram, S., Hoffman, M.D., and Wilson, A.G.G. (2021). What Are Bayesian Neural Network Posteriors Really Like? In Proceedings of the 38th International Conference on Machine Learning (PMLR)), pp. 4629–4640.

132. Crout, N., Kokkonen, T., Jakeman, A.J., Norton, J.P., Newham, L.T.H., Anderson, R., Assaf, H., Croke, B.F.W., Gaber, N., Gibbons, J., et al. (2008). Chapter two good modelling practice. Environ. Model. Softw. Decis. Support *3*, 15–31. https://doi.org/10.1016/s1574-101x(08)00602-9.

133. Jakeman, A.J., Elsawah, S., Wang, H.-H., Hamilton, S.H., Melsen, L., and Grimm, V. (2024). Towards normalizing good practice across the whole modeling cycle: its instrumentation and future research topics. Socio-Environmental Systems Modelling *6*, 18755. https://doi.org/10.18174/sesmo.18755.

134. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., et al. (2020). Underspecification Presents Challenges for Credibility in Modern Machine Learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2011.03395.

135. Elor, Y., and Averbuch-Elor, H. (2022). To SMOTE, or not to SMOTE?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2201.08528.

136. Porter, C.H., Villalobos, C., Holzworth, D., Nelson, R., White, J.W., Athanasiadis, I.N., Janssen, S., Ripoche, D., Cufi, J., Raes, D., et al. (2014). Harmonization and translation of crop modeling data to ensure interoperability. Environ. Model. Software *62*, 495–508. https://doi.org/10.1016/j.envsoft.2014.09.004.

137. Messina, C.D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). Eur. J. Agron. *100*, 151–162. https://doi.org/10.1016/j.eja.2018.01.007.

138. Wang, X., Zhao, C., Müller, C., Wang, C., Ciais, P., Janssens, I., Peñuelas, J., Asseng, S., Li, T., Elliott, J., et al. (2020). Emergent constraint on crop yield response to warmer temperature from field experiments. Nat. Sustain. *3*, 908–916. https://doi.org/10.1038/s41893-020-0569-7.

139. Dueben, P.D., Schultz, M.G., Chantry, M., Gagne, D.J., Hall, D.M., and McGovern, A. (2022). Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook. Artif. Intell. Earth Syst. *1*. https://doi.org/10.1175/AIES-D-21-0002.1.

140. Tsaftaris, S.A., and Scharr, H. (2019). Sharing the Right Data Right: A Symbiosis with Machine Learning. Trends Plant Sci. *24*, 99–102. https://doi.org/10.1016/j.tplants.2018.10.016.

141. Watson, P.A.G. (2022). Machine learning applications for weather and climate need greater focus on extremes. Environ. Res. Lett. *17*, 111004. https://doi.org/10.1088/1748-9326/ac9d4e.

142. Burnell, R., Schellaert, W., Burden, J., Ullman, T.D., Martinez-Plumed, F., Tenenbaum, J.B., Rutar, D., Cheke, L.G., Sohl-Dickstein, J., Mitchell, M., et al. (2023). Rethink reporting of evaluation results in AI. Science *380*, 136–138. https://doi.org/10.1126/science.adf6369.

143. Thomas, R.L., and Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. Patterns *3*, 100476. https://doi.org/10.1016/j.patter.2022.100476.

144. Niroula, S., Cai, X., and McIsaac, G. (2023). Sustaining crop yield and water quality under climate change in intensively managed agricultural watersheds—the need for both adaptive and conservation measures. Environ. Res. Lett. *18*, 124029. https://doi.org/10.1088/1748-9326/ad085f.

145. Tzachor, A., Devare, M., King, B., Avin, S., and Ó hÉigeartaigh, S. (2022). Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities. Nat. Mach. Intell. *4*, 104–109. https://doi.org/10.1038/s42256-022-00440-4.

146. Hammer, G., Messina, C., Wu, A., and Cooper, M. (2019). Biological reality and parsimony in crop models—why we need both in crop improvement. in silico Plants *1*, diz010. https://doi.org/10.1093/insilicoplants/diz010.