European Journal of Agronomy 168 (2025) 127610

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Integrating machine learning with agroecosystem modelling: Current state and future challenges

Meshach Ojo Aderele [a], Amit Kumar Srivastava [b,c], Klaus Butterbach-Bahl [a,d], Jaber Rahimi [a,d,*]

[a] Pioneer Center Land-CRAFT, Department of Agroecology, Aarhus University, Aarhus, Denmark
[b] Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany
[c] Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany
[d] Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research (IMK-IFU), Garmisch-Partenkirchen, Germany

ARTICLE INFO

Keywords:
Process-based models (PBM)
Data-driven models(DDM)
Machine learning (ML)
Agroecosystem

ABSTRACT

Machine learning (ML), especially deep learning (DL), is gaining popularity in the agroecosystem modelling community due to its ability to improve the efficiency of computationally intensive tasks. By reviewing previous modelling studies using the PRISMA technique, we present several examples of ML applications in this domain. The potential of using such models is highligthed. The different types of integration and model-building methods are categorized into process-based modelling (PBMs) and data-driven modelling (DDMs), which simulate different aspects of agroecosystem dynamics. While PBMs excel at capturing complex biophysical and biogeochemical processes, they are computationally intensive and may not always be solvable using analytical methods. To address these challenges, machine learning (ML) techniques, including deep learning (DL), are increasingly being integrated into agroecosystem modelling. This integration involves replacing PBMs with data-driven models, using hybrid models that combine PBMs and ML, or constructing simplified versions of PBMs through meta-modelling. ML-based meta-models offer computational efficiency and can capture intricate patterns and non-linear relationships in complex agricultural systems. However, challenges such as interpretability and data requirements remain. This review highlights the importance of addressing gaps and challenges to fully realize the potential of ML to identify the most promising ways of field management in promoting sustainable agricultural systems. It also highlights specific considerations such as data requirements, interpretability, model validation, and scalability for the successful integration of ML with PBMs in agriculture and the transformative potential of combining ML with PBMs, particularly in extending simulations from field to global scales and streamlining data collection processes through advanced sensor technologies based on their applications.

1. Introduction

Agroecosystem models play a critical role in agriculture, offering valuable insights into crop yield, soil N transport and transformation, soil-heat transfer, soil-water movement, soil nitrogen (N) transport and transformation, soil organic matter (SOM) turnover, and greenhouse gas (GHG) emissions that can be used to optimize crop management practices and conduct environmental impact assessments (Li et al., 2000; Parton et al., 1998). These models are typically categorized into two main types: the process-based modelling (PBMs) approach and the data-driven modelling (DDMs) approach.

PBMs (such as DSSAT (Jones et al., 2003), APSIM (Holzworth et al., 2014), and LandscapeDNDC (Haas et al., 2013)), simulate various aspects of plant growth and ecosystem dynamics by incorporating mathematical equations that represent the complex interactions between weather, topography, soil properties, management practices, genetics, pests and diseases, and a range of other factors (Yin and Van Laar, 2005; Holzworth et al., 2014; Nowatzke et al., 2022). These models are capable of simulating complex biophysical and biogeochemical processes, including carbon and nutrient cycling, organic matter decomposition, and soil-plant interactions (Brilli et al., 2017). PBMs are considered as powerful tools that allow researchers to improve their understanding of the agroecosystem functioning and provide accurate predictions under different environmental conditions, even if some

* Corresponding author at: Pioneer Center Land-CRAFT, Department of Agroecology, Aarhus University, Aarhus, Denmark.
E-mail addresses: Jaber.rahimi@kit.edu, Jaber.rahimi@agro.au.dk (J. Rahimi).

https://doi.org/10.1016/j.eja.2025.127610
Received 27 September 2024; Received in revised form 29 January 2025; Accepted 9 March 2025
Available online 17 March 2025

factors change in the future (Larocque et al., 2016).

However, due to the many processes and interactions involved, the relationships can be highly complex and may not always amenable to analytical mathematical methods. That's why most PBMs still rely on the basic principles of systems analysis. Furthermore, PBMs have limitations due to their computationally intensive nature, especially for rapid assessments or large-scale simulations (Levin et al., 1997, Nichols et al., 2011, Dawson and Gerritsen, 2016). For instance, identifying climate adaptation strategies through crop or species migration, particularly at a large scale, requires capturing a comprehensive representation of all potential rotation possibilities. This task can be computationally intensive when relying on process-based models (PBMs). In contrast, machine learning (ML) approaches provide a more computationally efficient alternative for tackling such challenges (Rising and Devineni, 2020).

To address these challenges, scientists have turned to machine learning (ML) techniques, including deep learning (DL), to accelerate crop modelling and to capture intricate patterns and nonlinear relationships in complex agricultural systems without making a priori assumptions about the relationship between drivers and responses (Van Klompenburg et al., 2020). Furthermore, ML offers the potential for improved predictive accuracy and computational efficiency (Liu et al., 2022).

To our knowledge, existing efforts to integrate ML models into agroecosystem modelling can be classified into three main groups: (a) those that completely replace the PBM with a data-driven model, (b) those that use both a process-based and a data-driven model simultaneously, and (c) those that aim to mimic the behavior of the PBM using ML techniques and use it as a replacement for the PBM. Fig. 1 illustrates different ways of integrating ML models with PBMs.

ML models can be used stand-alone to simulate the desired variables (e.g., N_2O emissions, ecosystem service indicators such as crop yield and biomass, crop water use, etc.) by integrating in-season satellite imagery and combining it with meteorological features and ground survey data (e.g., Filippi et al., 2019; Jiang et al., 2020; Zhang et al., 2021; Jeong et al., 2022), or by relying solely on soil and meteorological features (Chang et al., 2023). This makes them applicable to future projections as well (Li et al., 2023) (mode 2 in Fig. 1). Despite the advantages of such models, they have drawbacks, such as being black-box models that make interpretation difficult and require large amounts of training data. To address the limitations inherent in both PBMs and ML, the integration of the two into hybrid models has gained popularity in agroecosystem modelling (Garg et al., 2021). Typically, these hybrid models that integrate ML with PBMs, could be classified into two different classes: either by incorporating ML as a component within the process for the PBM or by using the PBM to provide a set of inputs for ML (Mode 3 in Fig. 1).

In addition, another mode of integration involves constructing simplified versions of complex PBMs, called meta-modelling, which involves both surrogate modelling and emulation, to mimic the behavior of a model while being computationally cheaper to evaluate (Mode 4 in Fig. 1). Meta-models act as approximations of complex simulations, allowing researchers to accelerate evaluations and explore numerous scenarios within feasible computational timeframes. In this mode of integration, ML is trained using the output of existing agroecosystem PBM simulations, taking advantage of ML's ability to learn and encapsulate intricate relationships within datasets. ML models operate by learning patterns and correlations from input data, allowing them to create meta-models that encapsulate the essential features of the underlying system. Through this process, ML excels at distilling complex relationships, patterns, and dependencies that are difficult to articulate through traditional modelling approaches.

While hybrid and ML-based meta-models hold great promise, numerous challenges remain. Before proceeding, it is essential to systematically review the current state of research and applications in agroecosystem modelling. Identifying gaps and challenges in existing approaches will provide a more nuanced understanding of the potential benefits and limitations of meta-models in agroecosystem modelling. This review aims to synthesize the available literature, shed light on emerging trends, challenges, and opportunities, and guide future research efforts towards realizing the full potential of incorporating ML to advance field management practices in support of sustainable agriculture.

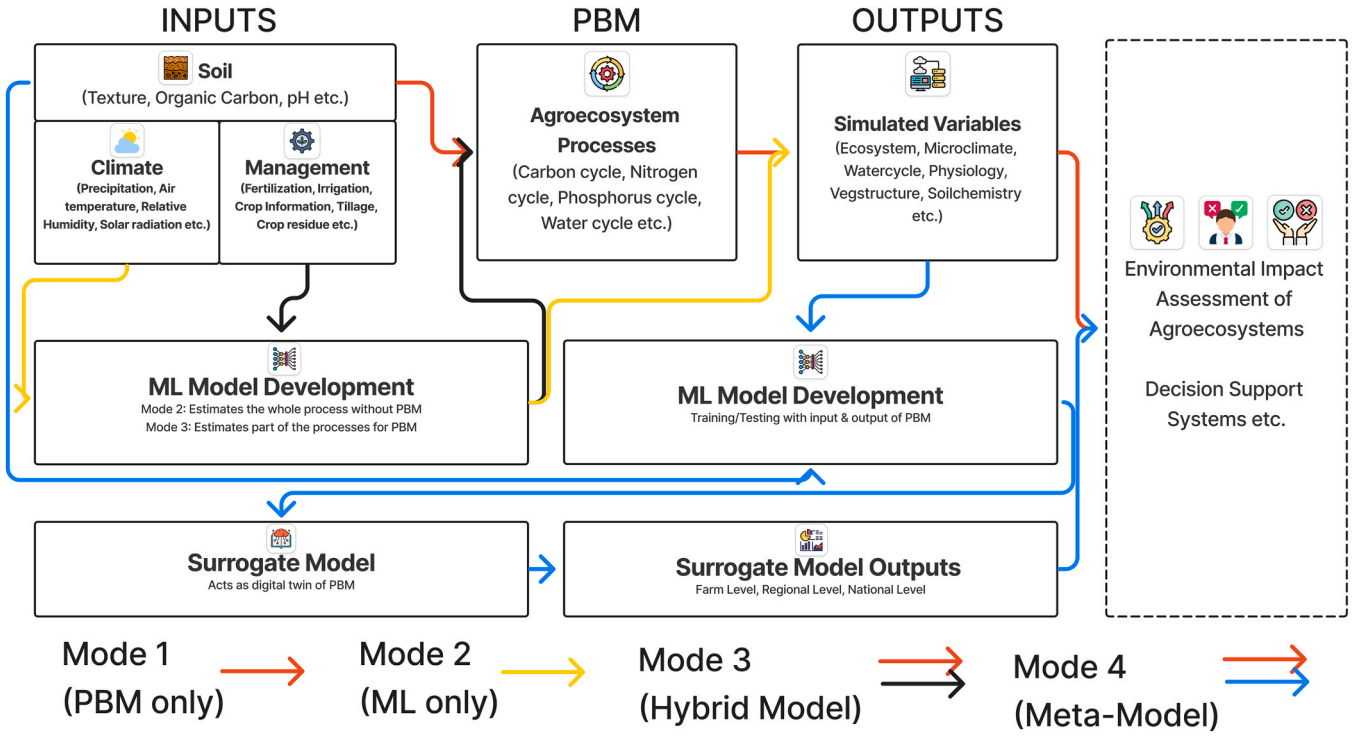

Fig. 1. Approaches for Integrating Machine Learning into Agroecosystem Modeling, with Color-Coded Arrows Representing Different Integration Modes, as Extracted from the Reviewed Articles.

2. Method

The study was organized according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Hutton et al., 2016). Data were collected from Google Scholar, Web of Science, and Scopus databases, which provide a comprehensive global dataset. The entire process of screening, eligibility assessment, and study inclusion is shown in Fig. 2.

The keywords used were "machine learning", "hybrid model", "surrogate model", "Meta-model", "metamodel", "agroecosystem modelling", "machine learning emulator", "crop modelling", "process based model" and "process-based model". Papers were selected based on the presence of either individual keywords or combinations of keywords. Specifically, studies were considered eligible if they used a PBM or ML model as the underlying framework for their ML-based hybrid or meta-model.

As shown in Fig. 2, a total of 1013 records were identified from databases. After removing 351 duplicate records, 183 records were excluded for other reasons such as misindexing and incorrect titles, which led to their initial inclusion with identified studies before screening, leaving 534 records for review. Of these, 283 records were excluded because their abstracts did not align with the scope of this review.

A further 251 reports were sought for retrieval, but 154 reports could not be retrieved due to reasons such as being conference abstracts or short articles without full texts, as well as duplicates with variations in indexing across different databases.

After retrieval, 97 reports were assessed for eligibility. Among these, 50 reports were excluded because they did not combine machine learning (ML) with process-based models (PBMs), 15 were review papers, and 7 were unrelated to agroecosystem modelling.

Finally, 25 studies were included in the review. Non-English articles, those published in non-peer-reviewed journals, and those published prior to year 2000 were not included in the analysis.

Key information extracted from the papers includes the agroecosystem applications of ML, the different PBMs used, study citations, the ML model used to build a hybrid or to create a meta-model, the best performing ML models used for these purposes, and the main conclusions and recommendations of each study.

The data was synthesized to identify the PBM for which most authors have attempted to build these integrated models, the different ML models that have been used to determine which one is the best, and to gather the opinions and recommendations of different authors on surrogate modelling. This synthesis aims to provide an overview of the current gaps and enable the recommendation of future research directions.

3. Results and discussion

In this section, we provide a summary of the literature review by outlining the characteristics of selected research studies and categorizing them based on the purpose of integration and model use, and discussing future recommendations for this integration.

The systematic literature review conducted highlights the

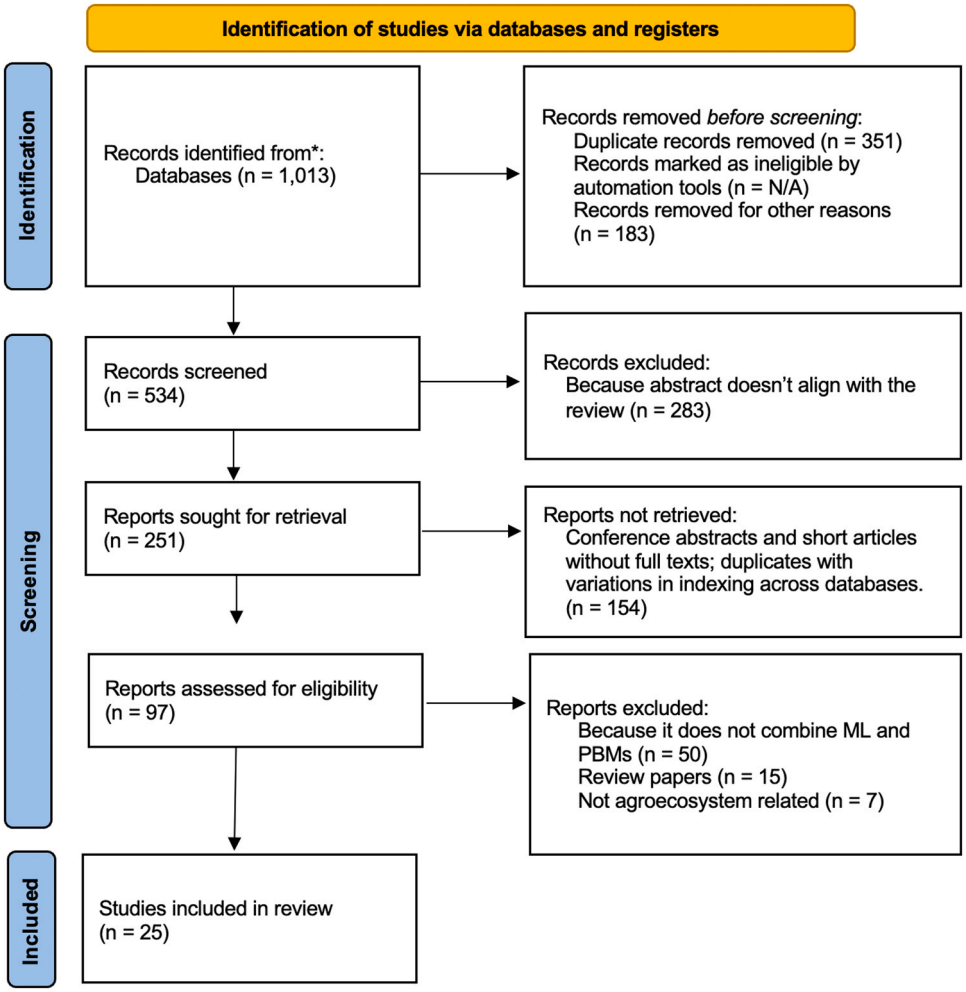

Fig. 2. The PRISMA screening process for selecting studies for inclusion in a systematic review.

widespread use of ML integration with agroecosystems modelling across different disciplines. As shown in Table 1, integrations have been carried out for different purposes such as yield prediction, estimation of gaseous N-fluxes (e.g., N_2O and NH_3), nitrate leaching calculations, carbon modelling, crop physiology, and phenology metrics, modelling water dynamics, grassland simulations, and agri-environmental policy. Furthermore, the creation of a meta-model is more common than the creatiion of a hybrid model. Meta-modelling is particularly common for large-scale simulations and optimization tasks (e.g., management practices) of a system or sensitivity analysis. The effectiveness of this approach depends on the accurate representation of agroecosystem dynamics within PBMs, which ensures the quality and relevance of the training data for ML algorithms. The important step in designing a meta-model is to generate numerous synthetic input samples and run them through the simulation process of the PBM. The outputs of these simulations together with the inputs, serve as the input for training the ML model, called the meta-model. There are various terms used for meta-modelling in the literature, the most common being surrogate modelling, emulators, proxy models, reduced order models, and lower fidelity models are the most common. However, we've noticed that these terms are sometimes used interchangeably, especially for "surrogate model" and "emulator". It's important to emphasize that the term "emulator," especially in the context of the Gaussian process approach, refers to approaches that provide a comprehensive probabilistic representation of simulation behavior, rather than just approximations of the outputs as the surrogate model does. To our knowledge, emulators aim to capture not only the mean behavior, but also the uncertainty associated with the predictions. They are often used when uncertainty quantification is essential. As an example of meta-modelling in agroecosystems, Johnston et al. (2023) used three ML algorithms (artificial neural networks, multivariate adaptive regression splines, and random forest) to replace the chickpea crop model of the APSIM-NextGen and showed that all ML models reasonably predicted the outputs for the training data set ($R^2 >$ 0.95). Furthermore, the study by Nguyen et al. (2019) used an artificial neural network was employed to develop a surrogate for the DayCent biogeochemical simulation model, aiming to optimize management practices. Their results indicate that the surrogate effectively captured over 99 % of the variation in DayCent's simulated outputs, resulting in a remarkable speed advantage by being 6.2 million times faster than the PBM.

Hybrid models, on the other hand, exploit the strengths of PBM in capturing complex agroecosystem processes, either by PBM providing inputs to ML or vice versa. As an example of such hybrid modelling, Dong et al. (2023) demonstrated how ML techniques could support PBM (SWAP) in addressing the root dynamics, one of the most challenging processes in crop growth simulation. In this approach, the PBM is used to simulate other variables, which are then fed back into the ML model for the next time step. Their study showed that this coupling can improve the performance of crop modelling. In another example, Droutsas et al. (2022) integrated ML into the PBM (GLAM-Parti) to generate inputs for stress factors under high temperatures. This integration simplified the physiological processes within the PBM by reducing the number of parameters to four, ultimately improving the model's performance in representing crop responses in different environments, including stressful conditions. In summary, the steps for building a hybrid model are in common and can be summarized as follows: Step 1, calibrate and validate the ML technique using experimental data; Step 2, set the initial conditions and boundaries of the simulation; Step 3, run the process-based model (PBM) to simulate crop/grass growth and related processes; Step 4, select outputs from the PBM (e.g., LAI, soil water content, soil salt content, and their corresponding days after sowing) to serve as inputs to run the ML model that predicts the variable of interest (e.g., root length density); Step 5, using the simulated variable from the ML model as inputs to run the PBM; and finally, Step 6, repeating steps 3 through 5 at each time step until the simulation is complete. In Table 1 provides a list of applications that involve the integration of ML models

Table 1
Current Applications of ML-PBM Integration.

Agroecosystem Application	Mode of Integration	Authors
Yield	Hybrid-modelling	Feng et al. (2019), Shahhosseini et al. (2021), Sun et al. (2022), Zhao et al. (2023)
	Meta-modelling	Shahhosseini et al. (2019), Nguyen et al. (2019), Xiao et al. (2022), Attia et al. (2022), Troost et al. (2022), Corrales et al. (2022), Kheir et al. (2023), Johnston et al. (2023), Kallenberg et al. (2023), Khan et al. (2023), Cunha, R. L et al., (2023)
Gases-N Flux	Hybrid-modelling	Saha et al. (2021)
	Meta-modelling	Villa-Vialaneix et al. (2012), Perlman et al. (2014), Shahhosseini et al. (2019), Nguyen et al. (2019), Ramanantenasoa et al. (2019), Liu et al. (2022), Genedy et al. (2023)
Nitrate Leaching	Meta-modelling	Garcet et al. (2006), Villa-Vialaneix et al. (2012), Nguyen et al. (2019)
Carbon	Hybrid-modelling	Hu et al. (2024)
	Meta-modelling	Luo et al. (2011), Nguyen et al. (2019), Xiao et al. (2022)
Crop Physiology & Phenology	Hybrid-modelling	Droutsas et al. (2022), Dong et al. (2023)
Water	Meta-modelling	Garcet et al. (2006), Nguyen et al. (2019), Attia et al. (2022)
Policy making	Meta-modelling	Shang et al. (2023)
Grassland simulation	Hybrid-modelling	Kenny et al. (2024)
	Meta-modelling	Pylianidis et al. (2022)

with PBMs, either through hybrid modelling or meta-modelling.

3.1. Key findings

Among the various PBMs used in agroecosystem modelling, APSIM (Agricultural Production Systems sIMulator) and DSSAT (Decision Support System for Agrotechnology Transfer) are prominent choices for surrogate modelling, as shown in Fig. 3. These models are selected by researchers for their comprehensive representation of agricultural processes, making them reliable options for predictive modelling in agroecosystems.

Furthermore, among the ML models for meta-modelling (Fig. 3), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) stand out as top choices. RF and XGBoost are favored for their robustness, scalability, and ability to find complex patterns. RF combines multiple decision trees to make accurate predictions and handle large datasets (Breiman 2001). It uses a bagging technique in which each tree is built independently on a random subset of the data, reducing overfitting and improving generalization, making it ideal for surrogate modelling with high-dimensional and noisy datasets.

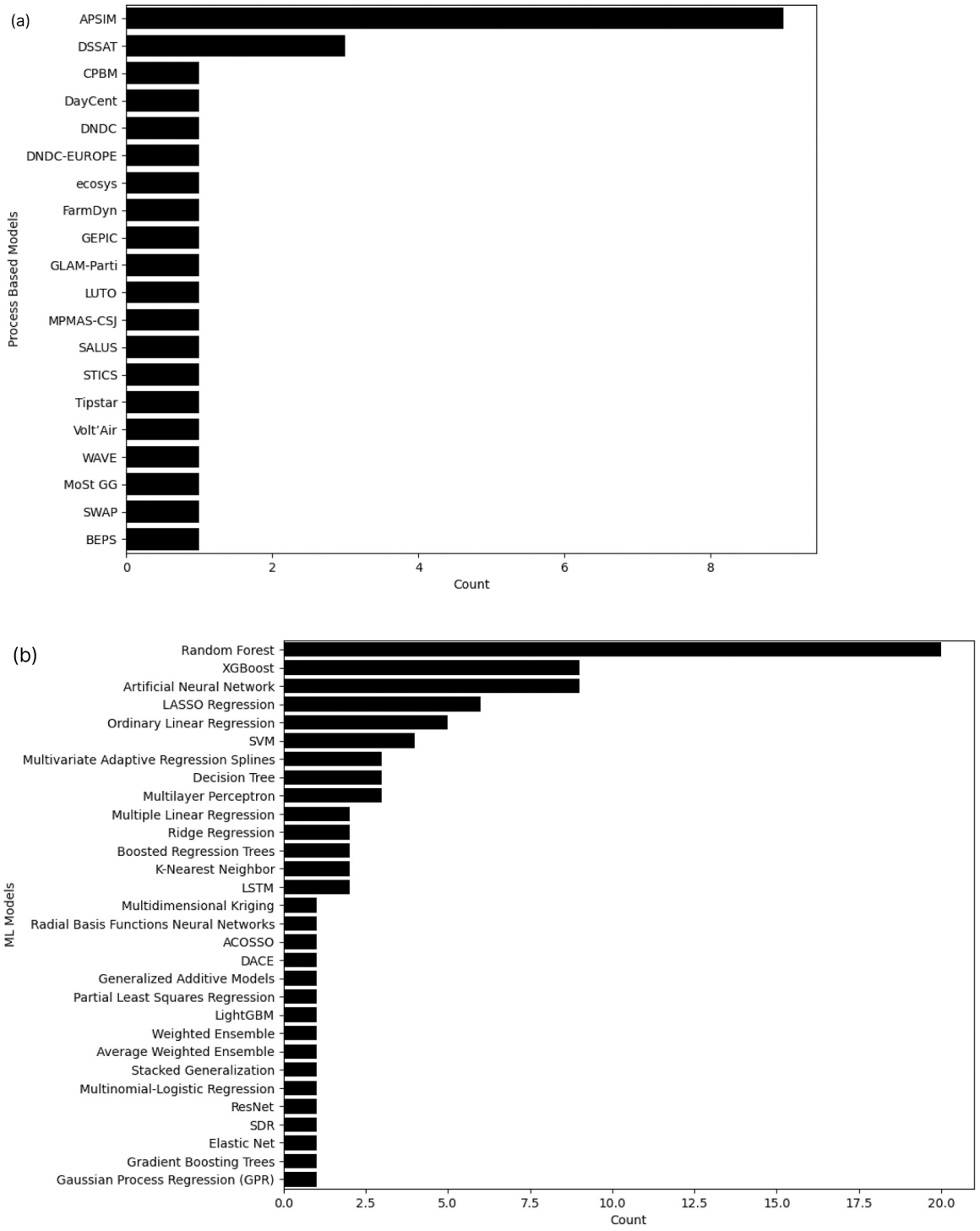

Fig. 3. Overview of Leading Process-Based Models (PBMs) (a) and Machine Learning (ML) Methods (b) in ML-PBM Integration, as Extracted from the Reviewed Articles.

On the other hand, XGBoost improves predictions by repeatedly learning from mistakes (Chen and Guestrin, 2016). According to the literature, both RF and XGBoost employ ensemble learning, which uses the collective intelligence of multiple trees to improve prediction accuracy, making them suitable for optimization purposes.

Synthesizing findings from multiple studies reveals the multifaceted landscape of surrogate modelling in agricultural systems and provides nuanced insights into methodologies, applications, and implications. The summary of the pros and cons of surrogate models compared to processed-based models and ML models as found in the reviewed articles are highlighted in Table 2. Key findings on the integration of ML in process-based models are further discussed in this section as categorised in Table 1.

The integration of machine learning (ML) with process-based models (PBMs) has emerged as a promising approach to address the limitations of PBMs in crop yield prediction, particularly under extreme climate events (ECEs). PBMs, while valuable for simulating the effects of climate change on crops, often struggle with oversimplified processes, vague descriptions, and parameter uncertainties, making it challenging to model ECEs accurately (Feng et al., 2019). Hybrid models provide a pathway to enhance the accuracy and utility of these simulations.

Feng et al. (2019) pioneered the development of a hybrid model combining APSIM outputs with growth-stage-specific ECE indicators such as frost, drought, and heat stress, integrating them into a Random Forest (RF) model. This approach highlighted the potential of ML to address the gaps in PBM performance, particularly when benchmarked against traditional multiple linear regression models. Similarly, Sun et al. (2022) extended this concept by coupling the GIS-based Environmental Policy Integrated Climate (GEPIC) model with RF, reducing uncertainties in soybean yield predictions by up to 41.83 %. These studies demonstrate that hybrid approaches can substantially enhance the robustness of PBM outputs in the context of ECEs.

Moreover, Zhang et al. (2021) leveraged a hybrid model combining CEREC-Maize with ML to explore the impacts of climate change on maize hybrids. The results showcased the scalability of this approach and its ability to tackle the complex challenge of predicting the future performance of new maize hybrids. This indicates the broader applicability of hybrid models beyond current crop varieties, aiding in accelerating the development of resilient crop hybrids.

The importance of large datasets in ML-PBM integration was emphasized by Shahhosseini et al. (2019), who found that increasing training data reduced yield prediction errors by 10 %–40 %. This insight underscores the data-intensive nature of ML approaches but also their potential to refine predictions as more data becomes available. Xiao et al. (2022) supported this by demonstrating that ML-based emulators can replicate APSIM yield predictions efficiently across varying spatial

Table 2
Summary of pros and cons of PBM (Mode 1), ML (Mode 2), Hybrid models (Mode 3), and Meta-models (Mode 4) based on lessons learned in reviewed studies.

	Mode 1	Mode 2	Mode 3	Mode 4
Pros	Mechanistic understanding of agroecosystem processes (such as carbon, nitrogen, and phosphorus cycle)	Offers robust stability for predictive tasks once it has been trained	Blend different modelling approaches for better generalization, improving adaptability to new scenarios or data.	Despite being a black box at the moment, it effectively portrays relationships in PBMs
	Simulating the dynamic behavior of agroecosystems over time, allowing for the investigation of spatio-temporal patterns	Efficient handling of extensive historical data and experiences.	Merge process-based and data-driven components, offering a comprehensive view of complex systems.	Offers a faster alternative to running complex simulations. Once trained, they swiftly generate predictions, saving time and resources.
	Simulating multiple outputs simultaneously	Reduced susceptibility to biases	Integrates various techniques, capturing a wider range of factors for more accurate predictions compared to standalone models.	Distills complex systems into simpler representations, aiding analysis and decision-making.
	Uncertainties and errors are manageable and estimable	Low computational demand when assimilating long-term historical data	Adaptable to specific objectives by combining suitable techniques and tailoring the framework to application needs.	Combine outputs from different simulations for a comprehensive understanding and scenario exploration.
	Reduced susceptibility to biases	Highly scalable and can handle large volumes of data efficiently	Leverage the strengths of PBMs and ML, boosting performance and overcoming individual limitations	Predict behavior between data points and beyond, enhancing predictive abilities across conditions.
	Generalizable to problems with similar processes		Integrate domain expertise for informed interpretation of results and decision-making. Manage uncertainty by combining information sources and quantifying uncertainty propagation, making them useful for decision-making in uncertain conditions.	Quantify how input changes affect outputs, aiding in identifying critical system factors. Provide probabilistic assessments of predictions, enhancing decision robustness and risk management.
Cons	Require extensive input data	Black box nature	Combining different modelling approaches introduces additional sources of uncertainty, which can make it difficult to assess the reliability and accuracy of the predictions.	Potentially struggle to adapt to new or unseen scenarios, limiting applicability in real-world studies.
	Require calibration and validation against observed data to ensure model accuracy and reliability	Overfitting	Hybrid models may require ongoing maintenance and updates to ensure compatibility between the integrated components and to incorporate new developments in each modelling approach.	Heavily rely on the quality and representativeness of training data, potentially leading to inaccurate predictions if the data is biased or insufficient.
	Demands high computational resources for assimilating long-term data	Model predictions can mirror biases present in the data	The complexity of hybrid models may hinder their interpretability, making it challenging to understand the underlying mechanisms driving the model predictions.	Building and training meta-models involve expertise in simulation models and ML techniques, posing challenges in implementation and result interpretation.
	Prone to numerical instability	Poor generalization of unforeseen problems	Hybrid models often require more computational resources compared to individual models, leading to longer processing times and higher computational costs.	Potential lack interpretability, hindering insights into the relationships between input and output variables.
	Poor generalization to problems of different process		Integrating multiple models can increase the complexity of the overall system, making it challenging to understand and interpret the results.	Possibility to overfit, especially with limited or noisy data, resulting in poor performance on new data by capturing noise rather than underlying patterns. Keeping meta-models up to date with changes in simulation models or data requires retraining or recalibration, posing challenges.

resolutions, reducing computational demands while maintaining accuracy.

In regions particularly vulnerable to climate change, such as arid zones, ML-PBM hybrids have shown promise. Attia et al. (2022) found these models effective in predicting yield in challenging environments, highlighting their adaptability. Additionally, Corrales et al. (2022) noted the advantages of support vector regression and linear regression meta-models in addressing the persistent issue of poor grain yield predictions.

The spatial simulation capabilities of hybrid models were further explored by Kheir et al. (2023), who reported improved precision in APSIM wheat yield predictions when coupled with ML. This integration offers decision-makers faster and more accurate tools for spatially explicit yield forecasts, with precision improvements from 89 % to 95 %.

Hybrid modelling also facilitates significant performance improvements in ML models when PBM-generated features are incorporated. Shahhosseini et al. (2021) showed that integrating APSIM simulation variables into ML models boosted predictive performance by up to 29 %, underscoring the synergy between the two approaches.

Finally, Cunha, R. L et al., (2023) demonstrated the computational efficiency of ML meta-models, which perform crop yield predictions nearly 100 times faster than PBMs while maintaining similar accuracy levels. This efficiency highlights the practical benefits of hybrid models, particularly for large-scale applications.

Integration ML into PBM has also been used to address the challenges of nitrogen gas flux modelling, particularly for nitrous oxide (N_2O) and ammonia (NH_3), which has long been hindered by limitations in process-based models (PBMs). These challenges include high variability in fluxes, limited geographic coverage, and computational inefficiencies. Recent studies demonstrate how integrating machine learning (ML) with PBMs can address these issues, enhancing accuracy and scalability. Saha et al. (2021) highlighted the challenges in predicting N_2O fluxes in intensively managed cropping systems, where PBMs often fail to achieve more than 20 % accuracy in daily to monthly emissions, even with site-specific calibration. By coupling ML with a cropping systems model to simulate unmeasured soil parameters, they demonstrated improved predictions of N_2O emissions. For example, their coupled model explained 51 % of the variation in weekly to biweekly flux measurements from corn, compared to 38 % with ML alone. This coupled approach underscores the potential of hybrid models to enhance both field-level predictions and global agricultural mitigation strategies. At a continental scale, Villa-Vialaneix et al. (2012) addressed the computational limitations of PBMs like DNDC-EUROPE by integrating ML. They found that spline approaches work effectively for small datasets, while support vector machines (SVMs) and Random Forest (RF) algorithms provided faster and more accurate solutions for larger datasets. This adaptability of hybrid models offers significant potential for scaling predictions across large and diverse geographic areas. Similarly, Perlman et al. (2014) tackled global analysis challenges by developing metamodels of the DNDC mechanistic model to estimate N_2O emissions from maize and wheat fields. Their metamodels achieved 97 % accuracy for maize and 91 % for wheat. The study highlighted the sensitivity of metamodels to soil properties, particularly soil organic carbon content, and emphasized the importance of interpretability in hybrid approaches. They also noted that global emission estimates were more affected by soil data aggregation than climate data, revealing critical factors for improving model precision. Ramanantenasoa et al. (2019) explored the potential of ML metamodels for ammonia (NH_3) emissions, addressing the long computation times of PBMs like Volt'Air. Their metamodel, developed from Volt'Air simulations in France, offered a computationally efficient alternative for regional and national NH_3 emission inventories, making it more accessible for public environmental agencies.

Nitrate leaching modelling, essential for understanding nutrient loss and its environmental impacts, has traditionally relied on deterministic nutrient balance models. While these models, such as WAVE and DayCent, offer detailed insights into nutrient dynamics under varying soil, crop, and climatic conditions, their application at larger scales is limited by computational inefficiencies and the need for extensive parameterization. Recent advancements integrating machine learning (ML) with these models have addressed these challenges, enabling faster and more scalable solutions.

Garcet et al. (2006) emphasized the constraints of deterministic models for large-scale or optimization studies due to their computational demands and the extensive parameters required. As an alternative, they developed a metamodel tailored for specific applications using a limited number of deterministic simulations from the WAVE nitrogen leaching model. Their metamodelling analysis significantly reduced computational time while maintaining high model efficiency (close to 1) and low root mean squared error. This success demonstrates the potential of metamodels to retain accuracy while overcoming the limitations of traditional deterministic approaches.

Similarly, Villa-Vialaneix et al. (2012) combined DNDC-EUROPE with ML techniques to predict nitrate leaching. This hybrid approach provided accurate results and significantly faster solutions, underscoring the effectiveness of ML in enhancing the scalability and efficiency of PBMs for nutrient leaching studies.

Nguyen et al. (2019) further pushed the boundaries of computational efficiency by employing an artificial neural network (ANN) as a surrogate for the DayCent biogeochemical simulation model. Their surrogate captured over 99 % of the variation in DayCent's outputs and achieved a staggering improvement in computational speed—6.2 million times faster than the original model. This finding highlights the transformative potential of ML in creating surrogate models that retain the precision of PBMs while enabling real-time or large-scale applications.

For carbon dynamics simulation, Luo et al. (2011) emphasized the difficulty of upscaling PBM results for regional soil carbon change assessments due to the lack of detailed spatial information and the high computational requirements of traditional models. Using Agricultural Production Systems sIMulator (APSIM), they developed meta-models that explained 85 % and 87 % of the variation in soil carbon content and changes, respectively, as simulated by APSIM. These meta-models, requiring only basic soil and climate data, demonstrated the capability to effectively replicate PBM outputs with significantly reduced input requirements. This advancement enables broader applications, such as regional and continental soil carbon assessments, with substantial reductions in computational complexity.

Similarly, Xiao et al. (2022) constructed a meta-model using APSIM for the Huang-Huai-Hai Plain in China. Their findings demonstrated that ML-based emulators could accurately and efficiently replicate APSIM predictions for soil carbon dynamics under varying nitrogen and water management scenarios. Moreover, these meta-models facilitated the rapid identification of optimal nitrogen management strategies for enhancing soil carbon sequestration. By reducing input requirements and computational overhead, ML-based approaches present a transformative solution for large-scale soil carbon modelling.

In the case of crop physiology and phenology, ML integration with PBMs offers promising advancements in simulating plant growth and development. Droutsas et al. (2022) embedded ML algorithms into the GLAM-Parti crop model for daily predictions of key physiological metrics, such as radiation use efficiency and days to maturity. This hybrid approach achieved high accuracy, with errors of less than 20 % for metrics like above-ground biomass, grain yield, and phenological measures. Additionally, ML eliminated the empirical stress factors traditionally used in PBMs to simulate crop responses to high-temperature environments, significantly simplifying physiological parameters. This framework underscores the potential of ML-driven PBMs to deliver more precise and scalable crop modelling solutions.

Dong et al. (2023) extended this integration by addressing the simulation of crop growth under varying water and salinity management practices. Their hybrid model, which coupled ML algorithms with the

SWAP crop model, accurately predicted soil water content, soil salt content, leaf area index, and dry matter production. Notably, Random Forest (RF) algorithms delivered more stable and precise predictions with lower root mean square errors (RMSE). This highlights the versatility of hybrid models in tackling complex interactions between soil and crop systems, particularly under stress conditions.

Accurately predicting water use and closing yield gaps are critical in addressing agricultural challenges, especially in regions vulnerable to climate change and water scarcity. While process-based crop models (PBMs) have been effective tools for such predictions, their limitations and uncertainties necessitate the integration of machine learning (ML) to improve accuracy and interpretability. Attia et al. (2022) highlighted the potential of combining PBMs with ML algorithms to enhance predictions of crop yield and water use. They utilized the DSSAT-CERES-Maize model to simulate long-term yield and evapotranspiration data under diverse management and environmental conditions. These simulations served as training inputs for ML models, effectively bridging the predictive gaps of PBMs. The resulting hybrid model demonstrated strong performance, achieving R^2 values greater than 0.82 and relative root mean square errors (RRMSE) below 9 % when using XGBoost as the ML algorithm. This high level of accuracy underscores the robustness of ML-enhanced PBMs in capturing the complexities of crop-water interactions across varying conditions. Beyond prediction accuracy, the integration of ML facilitated enhanced interpretability of the hybrid model. Attia et al. employed global and local SHAP (SHapley Additive exPlanations) values to identify key factors influencing yield and evapotranspiration predictions. These factors included:

- Maximum and minimum temperatures
- Available water content
- Soil organic carbon
- Irrigation practices
- Cultivar selection
- Soil texture
- Solar radiation
- Planting dates

This interpretability adds significant value, enabling stakeholders to understand the drivers of yield gaps and water use, and make data-informed decisions to optimize agricultural practices. The success of the ML-PBM hybrid model positions it as a promising tool for arid regions, where water resources are scarce and climate change exacerbates agricultural vulnerabilities.

Part of the studies reviewed also touched on the integration of ML with PBM with respect to Agri-environmental policy assessment. Agri-environmental policies play a critical role in shaping sustainable agricultural practices and structural transformations in farming systems. To assess the impacts of these policies effectively, models must integrate both technological advancements and environmental performance indicators. Shang et al. (2023) highlighted the significance of using farm-level models enriched with technological and environmental data, coupled with agent-based models, to simulate dynamic farm interactions. Despite the advantages of integrating farm-level and agent-based models, Shang et al. emphasized the considerable challenges involved in this approach. These include:

- **Model Development Complexity**: Integrating models with rich technology and environmental details requires advanced development efforts.
- **Debugging Difficulties**: Identifying and resolving issues in such integrated systems can be resource-intensive.
- **Computational Demands**: Simulations covering broad regions often require substantial computational power, limiting scalability.

To address these challenges, Shang et al. explored the use of surrogate models developed using deep learning techniques. In their study, the authors built surrogates of the FarmDyn farm model using various neural network architectures, achieving significant improvements in model usability. Among the surrogate models, Multilayer Perceptron (MLP) emerged as a particularly effective choice. The surrogate model built with MLP achieved near-top performance across all evaluation criteria, including accuracy and fit to the original FarmDyn model. It also dramatically reduced inference time, making simulations feasible even for broad regional applications. This demonstrates the potential of MLP-based surrogate models to streamline simulations, enabling faster and more scalable analysis of agri-environmental policies.

Finally, Machine learning (ML) integration into process-based models (PBMs) has also shown promise for enhancing grassland modelling. Kenny et al. (2024) investigated this approach to predict grass growth, beginning with a comparison of the predictive capabilities of PBMs and ML models when used independently. The study revealed complementary strengths: PBMs excelled at handling out-of-distribution events, while ML models demonstrated superior adaptability to temporary fluctuations in event variables, such as shifts in climate factors. By coupling the two approaches, the final model provided stable and accurate predictions across a wide range of conditions, including unprecedented temperature fluctuations, highlighting its practical utility. Pylianidis et al. (2022) expanded on this by addressing the limitations in input data typically faced by both PBMs and ML models when used individually. They employed a process-based model to generate simulation data, which they aggregated to lower resolutions to mimic real-world scenarios. Using a fraction of these inputs, they developed ML models trained at varying scales and tested them on both sampled and unsampled locations. The combined approach yielded accurate and generalizable predictions, underscoring its effectiveness for tactical decision-making in grassland management. This methodology showcases the potential of integrating PBMs and ML to overcome traditional limitations, enabling precise and reliable forecasting in grassland ecosystems.

The consensus among authors who have used integrated machine learning into process-based modelling for agroecosystem applications reflects a collective recognition of their utility, while also acknowledging opportunities for refinement and improvement. This consensus emerges from a nuanced understanding of the methodological intricacies and practical implications inherent in surrogate/hybrid modelling in understanding nutrient fluxes and crop performance in agroecosystems.

3.2. *Future perspective*

The integration of ML with PBMs has shown promise in several studies, although several gaps remain to be addressed. A critical gap is the need for a meta-model capable of representing interactions between different components as e.g. between plant and soil pools for nutrients, as evidenced by the need for ML-based surrogate models to fully capture phenomena such as the dynamics of agroecosystem carbon and nitrogen budgets. Such meta-models can reduce the computational intensity and time constraints associated with PBMs, thereby facilitating more comprehensive assessments of mitigation strategeies. However, while the meta-modelling process is valuable for enhancing model adaptability and predictive accuracy, it also introduces significant challenges as highlited from the studies reviewed. These challenges include issues related to data quality and quantity, where insufficient or noisy datasets can hinder model training and optimization. The interpretability of meta-models is another critical concern, as the complexity introduced by multiple layers of abstraction can obscure how predictions are made. Similarly, model validation becomes increasingly difficult due to the need to rigorously test and benchmark hybrid systems to ensure reliability and consistency. Additionally, the demand for computational resources escalates as optimization techniques in meta-modelling often

require significant processing power and memory, particularly when handling large datasets or complex scenarios. Another key challenge is uncertainty quantification, where identifying, propagating, and mitigating uncertainties in predictions is essential to maintain confidence in model outputs. Finally, achieving scalability and generalization remains a persistent difficulty, as models must adapt to varying datasets and conditions without compromising performance. To address these challenges, we have proposed several strategies.

3.2.1. Data quality and quantity

Since ML models are data dependent and require significant amounts of data for effective training, collecting such large data sets can often be a significant challenge, especially in the agricultural sector where data collection can be labor-intensive and time-consuming. Thus, the need for large data sets can be a major barrier to the successful implementation of ML models in agriculture. However, this can be addressed by encouraging cooperation between farmers, food system companies, research institutions, and agricultural organizations to share data that can enrich the variety and size of the datasets. Using methods such as data augmentation, transfer learning from existing models, and incorporating crowdsourcing efforts can effectively aggregate and enrich datasets, reducing the need for extensive labeled data. In addition, the creation of synthetic data and the use of active learning strategies can address data scarcity issues. For example, Liu et al. (2024) demonstrated the use of synthetic data from ecosys to KGML-ag-Carbon model, a model used for accurate and cost-effective quantification of the carbon cycle for agroecosystems at decision-relevant scales. Their findings highlight that leveraging synthetic data generated by process-based (PB) models is several orders of magnitude more cost-effective compared to collecting real-world observational data. Furthermore, the use of sensor networks and IoT (Internet of Things) devices can generate real-time data streams, enriching datasets essential for training ML models. This collaborative approach not only improves data quality but also promotes innovation and efficiency in agricultural practices through advanced analytics.

3.2.2. Interpretability

The integration of a "black-box" model such as ML into PBM introduces additional uncertainties that cannot be avoided. This integration has the potential to complicate the interpretability of the original PBM, raising concerns about the reliability and transparency of the overall system. Therefore, the decision to replace components of the PBM with ML models should be made with careful consideration. It is advisable to replace only those modules within the PBM that are based on simple empirical relationships and assumptions, while ignoring the underlying mechanisms of the simulated process. Lian et al. (2023) improved the accuracy and interpretability of streamflow estimation by replacing the classical evapotranspiration (ET) module of process-based hydrological models with a data-driven submodel. This cautious approach ensures that the benefits of using ML, such as improved predictive accuracy and adaptability to complex patterns, are balanced with the need for transparency and traceability in the modelling process. By selectively replacing specific modules, the integrity and interpretability of the PBM can be maintained while leveraging the strengths of ML to improve modelling capabilities. Additionally, techniques such as feature importance analysis, partial dependence plots, and SHapley Additive exPlanations (SHAP) can be used to make ML models more interpretable (Molnar, 2020). These methods help identify how input variables influence model predictions, providing transparency without compromising accuracy (Villa-Vialaneix et al., 2012). Through thoughtful decision making and rigorous evaluation, the synergistic integration of ML with PBMs can lead to more robust and insightful modelling results.

3.2.3. Model validation

Validating the accuracy of hybrid models that integrate machine learning (ML) with process-based models (PBMs) is critical to ensuring their reliability, consistency, and practical applicability. This process involves testing the model on unseen data and assessing its performance using robust metrics to evaluate accuracy, generalizability, and robustness. Proper validation is essential to prevent overfitting or underfitting, both of which can compromise the model's ability to make reliable predictions.

To enhance validation efforts, specific techniques and criteria should be employed. For instance, cross-validation methods, such as k-fold or leave-one-out cross-validation, provide insights into model performance across different subsets of data. Performance metrics such as root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) should be used to quantify predictive accuracy and identify potential biases. Furthermore, coupling meta-modelling with interpretability measures can yield simplified insights, as emphasized by Villa-Vialaneix et al. (2012). This approach helps streamline the hybrid modelling process while preserving essential details. Rigorous feature selection, as suggested by Khan et al. (2023), can reduce data redundancy and improve model efficiency by focusing only on the most relevant predictors. Evaluating multiple ML approaches and ensembles tailored to specific datasets, as advocated by Shahhosseini et al. (2019), can further improve reliability of the model. Ensembles often mitigate individual model biases and enhance predictive robustness by combining the strengths of diverse algorithms.

3.2.4. Uncertainty quantification

Effectively managing and quantifying and managing uncertainty is crucial for ensuring the reliability and credibility of hybrid models that integrate machine learning (ML) with process-based models (PBMs).

Several techniques can be employed to address uncertainty in hybrid models. Quantification approaches, such as Bayesian inference and Monte Carlo simulations, allow for the characterization of parameter uncertainty and its impact on model outputs. These methods provide probabilistic estimates that help model users better understand the range and likelihood of potential outcomes. Propagation techniques, such as interval analysis and polynomial chaos expansion, can further assess how uncertainties in input variables affect final model predictions (Abdar et al., 2021). Sensitivity analysis is also critical to identify key parameters and inputs that contribute most to uncertainty, enabling targeted refinement of model components (Ankenbrand et al., 2021). Future research should also focus on integrating remote sensing and Geographic Information System (GIS) data, as suggested by Shahhosseini et al. (2021), to enhance prediction accuracy and data representativeness. Remote sensing data, with its high spatial and temporal resolution, can help reduce uncertainties by providing comprehensive environmental information. Similarly, GIS tools enable spatial analysis and integration of multi-source data, offering an avenue to better contextualize predictions within real-world settings.

3.2.5. Scalability and generalization

It is important to note that the validation process should be an ongoing one. As new data becomes available, the model should be retrained and revalidated to ensure that it continues to perform well. This is particularly important in areas such as agriculture, where conditions can change rapidly and models need to be able to adapt to these changes. Moreover, future research should prioritize the generalizability of surrogate models across different scenarios and applications to improve their utility and scalability, as proposed by Saha et al. (2021). To make a a model scalebale and generalisable to new dataset, a robust ML operational framework should include an automated pipeline for training, retraining, and deployment. Such a pipeline ensures that models remain relevant by incorporating new data streams and adapting to shifts in underlying data distributions. Model retraining can be triggered based on predefined criteria, such as a drop in model performance metrics (e.g., accuracy, F1-score) on a validation set or when significant data drift detected (Mallick et al., 2022). For example, systems like Google's TFX (TensorFlow Extended) enable automated model

retraining pipelines by integrating data ingestion, feature engineering, model validation, and deployment into a cohesive workflow (Baylor et al., 2017). These pipelines can monitor incoming data streams, detect anomalies or shifts, and initiate retraining when necessary, ensuring that the model remains generalizable to new scenarios. Model registries also help with scalbility and generalisability, they play a pivotal role in managing the lifecycle of models as they are trained and retrained. Platforms such as MLflow or Amazon SageMaker provide tools to version models, store their metadata (e.g., training data, hyperparameters, and evaluation metrics), and track lineage. Registries enable seamless transitions between model versions, ensuring that updates do not disrupt production systems. For instance, MLflow's Model Registry supports lifecycle stages (e.g., "Staging," "Production," "Archived") and allows automated workflows for promoting models based on validation results (Chen et al., 2020). These workflows help in managing multiple models efficiently and ensure the best-performing model is always in production.

3.2.6. Computational resources

The integration of machine learning (ML) with process-based models (PBMs) can be computationally demanding, primarily due to the high resource requirements associated with model training and retraining. These processes typically involve processing large datasets, which can result in increased processing times and higher operational costs, particularly when using complex ML algorithms such as deep learning (Bengio, 2009). However, it is important to note that the high computational consumption is not constant. The resource-intensive tasks mainly occur during the training phase, which is done infrequently (e.g., a few times per year, depending on the model and data availability). Once the model has been trained, making predictions requires significantly fewer resources and is less computationally expensive.

To optimize computational efficiency during training, strategies such as batch processing, parallel processing, and cloud computing can be employed. Batch processing divides the dataset into smaller, manageable subsets (batches), reducing memory requirements and accelerating convergence by updating model weights incrementally (Ruder, 2016). Parallel processing leverages multiple processors to handle computations simultaneously, while cloud computing provides scalable resources to adjust computational capacity as needed. Additionally, optimized algorithms or model compression techniques can further mitigate computational demands without sacrificing predictive accuracy (Dantas et al., 2024).

Together, these recommendations collectively pave the way for a more effective and robust use of ML in conjunction with PBMs for sustainable resource management and decision making.

4. Conclusion

The integration of machine learning (ML) with process-based modelling (PBM) offers a transformative solution to overcoming the computational and time constraints often faced in large-scale simulations. ML enhances PBMs by providing efficient tools for spatial interpolation, global data downscaling, and the processing of large datasets, such as those from remote sensing. This integration reduces the computational burden by enabling faster model calibration, parameterization, and prediction across diverse spatial scales. ML algorithms speed up simulation workflows that would otherwise be computationally prohibitive. In addition, ML's ability to adapt to new data allows for real-time model updates and dynamic recalibration, addressing time constraints associated with long-duration simulations. These capabilities make it possible to scale up from field-level models to global simulations, enhancing the representation of environmental processes while maintaining accuracy and efficiency.

Ultimately, the synergy between ML and PBM not only resolves critical challenges in computational resource demands but also accelerates decision-making by enabling the timely and accurate assessment of complex issues in agroecosystem. This integrated framework is poised to significantly advance research and decision-making in fields such as agriculture, climate science, and ecosystem management, providing actionable insights for policy making at both local and global scales.

CRediT authorship contribution statement

Rahimi Jaber: Writing – review & editing, Writing – original draft, Supervision, Data curation, Conceptualization. **Butterbach-Bahl Klaus:** Writing – review & editing, Supervision. **Srivastava Amit Kumar:** Writing – review & editing. **Aderele Meshach Ojo:** Writing – original draft, Data curation.

Author Contribution Statement

M.A., J.R., and K.B.B. conceived and designed the study. M.A. performed the analysis. M.A. and J.R. wrote the first draft, while other co-authors (K.B.B and A.S) have contributed to improving the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare that they have no conflict of interest.

Acknowledgement

This study was supported by the Pioneer Center for Research in Sustainable Agricultural Futures (Land-CRAFT), DNRF grant number P2, Aarhus University, Denmark.

Data availability

No data was used for the research described in the article.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297.

Ankenbrand, M.J., Shainberg, L., Hock, M., Lohr, D., Schreiber, L.M., 2021. Sensitivity analysis for interpretation of machine learning based segmentation models in cardiac MRI. BMC Med. Imaging 21, 1–8.

Attia, A., Govind, A., Qureshi, A.S., Feike, T., Rizk, M.S., Shabana, M.M.A., Kheir, A.M.S., 2022. Coupling Process-Based Models and Machine Learning Algorithms for Predicting Yield and Evapotranspiration of Maize in Arid Environments. Water (Switz.) 14 (22). https://doi.org/10.3390/w14223647.

Baylor, D., Breck, E., Cheng, H.T., Fiedel, N., Foo, C.Y., Haque, Z., Zinkevich, M., 2017. Tfx: A tensorflow-based production-scale machine learning platform (August). *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* 1387–1395.

Bengio, Y., 2009. Learn. Deep Archit. AI.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Bellocchi, G., 2017. Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. Sci. Total Environ. 598, 445–470.

Chang, Y., Latham, J., Licht, M., Wang, L., 2023. A data-driven crop model for maize yield prediction. Commun. Biol. 6 (1), 439.

Chen, A., Chow, A., Davidson, A., DCunha, A., Ghodsi, A., Hong, S.A., Zumar, C., 2020. Developments in mlflow: A system to accelerate the machine learning lifecycle. Proceedings of the fourth international workshop on data management for end-to-end machine learning, pp. 1–4.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Corrales, D.C., Schoving, C., Raynal, H., Debaeke, P., Journet, E.P., Constantin, J., 2022. A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France. Comput. Electron. Agric. 192. https://doi.org/10.1016/j.compag.2021.106578.

Cunha, R. L. de F.., Silva, B., & Avegliano, P.B. (2023). A Comprehensive Modelling Approach for Crop Yield Forecasts using AI-based Methods and Crop Simulation Models. ⟨http://arxiv.org/abs/2306.10121⟩.

Dantas, P.V., Sabino da, Silva Jr, W., Cordeiro, L.C., Carvalho, C.B., 2024. A comprehensive review of model compression techniques in machine learning. Appl. Intell. 54 (22), 11804–11844.
Dawson, C., Gerritsen, M.G., 2016. Computational Challenges in the Geosciences. Springer, Berlin.
Dong, L., Lei, G., Huang, J., Zeng, W., 2023. Improving crop modelling in saline soils by predicting root length density dynamics with machine learning algorithms. Agric. Water Manag. 287, 108425.
Droutsas, I., Challinor, A.J., Deva, C.R., Wang, E., 2022. Integration of machine learning into process-based modelling to improve simulation of complex crop responses. Silico Plants 4 (2). https://doi.org/10.1093/insilicoplants/diac017.
Feng, P., Wang, B., Liu, D.L., Waters, C., 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.
Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D., Pozza, L.E., Ugbaje, S.U., Bishop, T.F., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. 20, 1015–1029.
Garcet, J.P., Ordonez, A., Roosen, J., Vanclooster, M., 2006. Meta-modelling: Theory, concepts and application to nitrate leaching modelling. Ecol. Model. 193 (3-4), 629–644.
Garg, P., Chakravarthy, A.S., Mandal, M., Narang, P., Chamola, V., Guizani, M., 2021. Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities. ACM Trans. Internet Technol. (TOIT) 21 (3), 1–18.
Genedy, R.A., Ogejo, J.A., Chung, M., Senger, R.S., Shortridge, J.E., 2023. Integr. Mach. Learn. Into Process-Based Model. Predict Ammon. Losses Stored Liq. Dairy Manure.
Haas, E., Klatt, S., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., Grote, R., Breuer, L., Butterbach-Bahl, K., 2013. LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. Landsc. Ecol. 28, 615–636.
Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Keating, B.A., 2014. APSIM–evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350.
Hu, X., Shi, L., Lin, L., Li, S., Deng, X., Li, L., Lian, X., 2024. A novel hybrid modelling framework for GPP estimation: Integrating a multispectral surface reflectance based Vcmax25 simulator into the process-based model. Sci. Total Environ., 171182
Hutton, B., Catala-Lopez, F., Moher, D., 2016. The PRISMA statement extension for systematic reviews incorporating network meta-analysis: PRISMA-NMA. Med. Cl. ínica (Engl. Ed.) 147 (6), 262–266.
Jeong, S., Ko, J., Shin, T., 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 (1), 9030.
Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Lin, T., 2020. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. Glob. Change Biol. 26 (3), 1754–1766.
Johnston, D.B., Pembleton, K.G., Huth, N.I., Deo, R.C., 2023. Comparison of machine learning methods emulating process driven crop models. Environ. Model. Softw. 162. https://doi.org/10.1016/j.envsoft.2023.105634.
Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18 (3-4), 235–265.
Kallenberg, M.G.J., Maestrini, B., van Bree, R., Ravensbergen, P., Pylianidis, C., van Evert, F., Athanasiadis, I.N., 2023. Integr. Process. -Based Models Mach. Learn. Crop yield Predict. http://arxiv.org/abs/2307.13466.
Kenny, E.M., Ruelle, E., Keane, M.T., Shalloo, L., 2024. A Hybrid Model that Combines Machine Learning and Mechanistic Models for Useful Grass Growth Prediction. Comput. Electron. Agric. 219, 108805.
Khan, M.S., Moallemi, E.A., Nazari, A., Thiruvady, D., Bryan, B.A., 2023. Quantifying the Safe Operating Space for Land-System SDG Achievement via Machine Learning and Scenario Discovery. Earth'S. Future 11 (3). https://doi.org/10.1029/2022EF003083.
Kheir, A.M.S., Mkuhlani, S., Mugo, J.W., Elnashar, A., Nangia, V., Deware, M., Govind, A., 2023. Integrating APSIM model with machine learning to predict wheat yield spatial distribution. Agron. J. https://doi.org/10.1002/agj2.21470.
Larocque, G.R., Komarov, A., Chertov, O., Shanin, V., Liu, J., Bhatti, J.S., Wang, W., Peng, C., Shugart, H.H., Xi, W., Holm, J.A., 2016. Process-Based Model.: a Synth. Model. Appl. Address Environ. Manag. Issues.
Levin, S.A., Grenfell, B., Hastings, A., Perelson, A.S., 1997. Mathematical and computational challenges in population biology and ecosystems science. Science 275 (5298), 334–343.
Li, C., Aber, J., Stange, F., Butterbach-Bahl, K., Papen, H., 2000. A process-oriented model of N2O and NO emissions from forest soils: 1. Model development. J. Geophys. Res.: Atmospheres 105 (D4), 4369–4384.
Li, L., Zhang, Y., Wang, B., Feng, P., He, Q., Shi, Y., Yu, Q., 2023. Integrating machine learning and environmental variables to constrain uncertainty in crop yield change projections under climate change. Eur. J. Agron. 149, 126917.
Lian, X., Hu, X., Bian, J., Shi, L., Lin, L., Cui, Y., 2023. Enhancing streamflow estimation by integrating a data-driven evapotranspiration submodel into process-based hydrological models. J. Hydrol. 621, 129603.
Liu, L., Xu, S., Tang, J., Guan, K., Griffis, T.J., Erickson, M.D., Frie, A.L., Jia, X., Kim, T., Miller, L.T., Peng, B., Wu, S., Yang, Y., Zhou, W., Kumar, V., Jin, Z., 2022. KGML-ag: a modelling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating N2O emission using data from mesocosm experiments. Geosci. Model Dev. 15 (7), 2839–2858. https://doi.org/10.5194/gmd-15-2839-2022.
Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., Wang, G., 2022. Machine learning crop yield models based on meteorological features and comparison with a process-based model. Artif. Intell. Earth Syst. 1 (4), e220002.
Liu, L., Zhou, W., Guan, K., 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.
Luo, Z., Wang, E., Sun, O.J., Smith, C.J., Probert, M.E., 2011. Modeling long-term soil carbon dynamics and sequestration potential in semi-arid agro-ecosystems. Agric. For. Meteorol. 151 (12), 1529–1544.
Mallick, A., Hsieh, K., Arzani, B., Joshi, G., 2022. Matchmaker: Data drift mitigation in machine learning for large-scale systems. Proc. Mach. Learn. Syst. 4, 77–94.
Molnar, C., 2020. Interpret. Mach. Learn.: A Guide Mak. Black Box. Models Explain. (Available at: https://christophm.github.io/interpretable-ML-book/).
Nguyen, T.H., Nong, D., Paustian, K., 2019. Surrogate-based multi-objective optimization of management options for agricultural landscapes using artificial neural networks. Ecol. Model. 400, 1–13. https://doi.org/10.1016/j.ecolmodel.2019.02.018.
Nichols, J., Kang, S., Post, W., Wang, D., Bandaru, V., Manowitz, D., Izaurralde, R., 2011. HPC-EPIC for high resolution simulations of environmental and sustainability assessment. Comput. Electron. Agric. 79 (2), 112–115.
Nowatzke, M., Damiano, L., Miguez, F.E., McNunn, G.S., Niemi, J., Schulte, L.A., Heaton, E.A., VanLoocke, A., 2022. Augmenting agroecosystem models with remote sensing data and machine learning increases overall estimates of nitrate-nitrogen leaching. Environ. Res. Lett. 17 (11), 114010.
Parton, W.J., Hartman, M., Ojima, D., Schimel, D., 1998. DAYCENT and its land surface submodel: description and testing. Glob. Planet. Change 19 (1-4), 35–48.
Perlman, J., Hijmans, R.J., Horwath, W.R., 2014. A Meta-modelling approach to estimate global N2O emissions from agricultural soils. Glob. Ecol. Biogeogr. 23 (8), 912–924. https://doi.org/10.1111/geb.12166.
Pylianidis, C., Snow, V., Overweg, H., Osinga, S., Kean, J., Athanasiadis, I.N., 2022. Simulation-assisted machine learning for operational digital twins. Environ. Model. Softw. 148, 105274.
Ramanantenasoa, M.M.J., Génermont, S., Gilliot, J.M., Bedos, C., Makowski, D., 2019. Meta-modelling methods for estimating ammonia volatilization from nitrogen fertilizer and manure applications. J. Environ. Manag. 236, 195–205. https://doi.org/10.1016/j.jenvman.2019.01.066.
Rising, J., Devineni, N., 2020. Crop switching reduces agricultural losses from climate change in the United States by half under RCP 8.5. Nat. Commun. 11 (1), 4991.
Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.
Saha, D., Basso, B., Robertson, G.P., 2021. Machine learning improves predictions of agricultural nitrous oxide (N2O) emissions from intensively managed cropping systems. Environ. Res. Lett. 16 (2). https://doi.org/10.1088/1748-9326/abd2f3.
Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V., 2021. Coupling machine learning and crop modelling improves crop yield prediction in the US Corn Belt. Sci. Rep. 11 (1). https://doi.org/10.1038/s41598-020-80820-1.
Shahhosseini, M., Martinez-Feria, R.A., Hu, G., Archontoulis, S.V., 2019. Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. 14 (12). https://doi.org/10.1088/1748-9326/ab5268.
Shang, L., Wang, J., Schäfer, D., Heckelei, T., Gall, J., Appel, F., Storm, H., 2023. Surrogate modelling of a detailed farm-level model using deep learning. J. Agric. Econ. https://doi.org/10.1111/1477-9552.12543.
Sun, Q., Zhang, Y., Che, X., Chen, S., Ying, Q., Zheng, X., Feng, A., 2022. Coupling Process-Based Crop Model and Extreme Climate Indicators with Machine Learning Can Improve the Predictions and Reduce Uncertainties of Global Soybean Yields. Agric. (Switz.) 12 (11). https://doi.org/10.3390/agriculture12111791.
Troost, C., Parussis-Krech, J., Mejaíl, M., Berger, T., 2022. Boosting the Scalability of Farm-Level Models: Efficient Surrogate Modelling of Compositional Simulation Output. Comput. Econ. https://doi.org/10.1007/s10614-022-10276-0.
Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709.
Villa-Vialaneix, N., Follador, M., Ratto, M., Leip, A., 2012. A comparison of eight Meta-modelling techniques for the simulation of N2O fluxes and N leaching from corn crops. Environ. Model. Softw. 34, 51–66. https://doi.org/10.1016/j.envsoft.2011.05.003i.
Xiao, L., Wang, G., Zhou, H., Jin, X., Luo, Z., 2022. Coupling agricultural system models with machine learning to facilitate regional predictions of management practices and crop production. Environ. Res. Lett. 17 (11). https://doi.org/10.1088/1748-9326/ac9c71.
Yin, X., Van Laar, H.H., 2005. Crop systems dynamics: an ecophysiological simulation model for genotype-by-environment interactions. Wageningen Academic Publishers.
Zhang, L., Zhang, Z., Tao, F., Luo, Y., Cao, J., Li, Z., Xie, R., Li, S., 2021. Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning. Environ. Res. Lett. 16 (12). https://doi.org/10.1088/1748-9326/ac32fd.
Zhao, Y., Xiao, D., Bai, H., Tang, J., Liu, D.L., Qi, Y., Shen, Y., 2023. The prediction of wheat yield in the North China plain by coupling crop model with machine learning algorithms. Agriculture 13 (1), 99.