A HYBRID SURROGATE APPROACH FOR REDUCING PHOTOSYNTHESIS RUNTIME IN A MECHANISTIC PLANT MODEL

Aleksandar Joksimović*,
[0009-0008-5711-7636]

Dušan Kostić,
[0009-0000-3432-6074]

Tamara Naumović,
[0000-0001-9849-7665]

Petar Lukovac,
[0000-0003-4561-8886]

Zorica Bogdanović
[0000-0003-4799-1588]

University of Belgrade –
Faculty of Organizational Sciences,
Belgrade, Serbia

Abstract:

This paper examines the use of surrogate models to accelerate a mechanistic plant simulator in the context of plant digital twins. The research aims to reduce the computational cost of the simulator while preserving the biologically meaningful growth behaviour of the underlying model. To this end, the native C implementation was first profiled, which identified the iterative Farquhar–Jarvis photosynthesis routine as the dominant runtime bottleneck. Based on this finding, a simulation-generated dataset was constructed, a compact multilayer perceptron surrogate was trained for the high-irradiance regime, and the resulting model was integrated directly into the existing C simulation pipeline. The results showed that the surrogate achieved near-perfect agreement with the original net assimilation target in the selected regime, while substantially reducing the computational dominance of the original photosynthesis routine. At the same time, the surrogate-enabled simulator preserved relative growth rate and final leaf biomass with only minor deviations across the evaluated photoperiod scenarios. The main contribution of the paper lies in demonstrating that a narrow, regime-aware surrogate can significantly improve runtime efficiency without disrupting the mechanistic structure and system-level behaviour of the original plant model.

Keywords:

Mechanistic Plant Models, Surrogate Modelling, Plant Digital Twins, Hybrid Modelling.

INTRODUCTION

Mechanistic plant models remain important because they describe growth and development through explicit physiological processes such as photosynthesis, carbon allocation, nutrient dynamics, and biomass accumulation. Their value is not limited to predictive performance alone: by retaining process-level structure, such models support interpretation, causal reasoning, and scenario analysis, which are all particularly relevant in plant digital twin settings [1], [2]. At the same time, the very features that make these models scientifically attractive often make them computationally demanding, especially when biologically detailed submodules must be evaluated repeatedly during simulation.

Correspondence:
Aleksandar Joksimović

e-mail:
aleksandar.joksimovic@fon.bg.ac.rs

DOI: 10.15308/Sinteza-2026-260-266

This tension between physiological fidelity and computational tractability is well recognized in process-based plant and crop modelling [3]. In parallel, recent work on surrogate modelling and hybrid digital twin architectures suggests that data-driven approximations can play a useful supporting role when they are applied selectively to computationally expensive model components rather than used as wholesale replacements [4], [5]. However, there is still limited work on compact, natively deployable surrogates that replace a single dominant hotspot inside a mechanistic plant simulator while preserving downstream growth behaviour.

The aim of this paper is therefore to investigate whether the photosynthesis module of a mechanistic plant simulator can be approximated by a lightweight surrogate in a way that yields a meaningful computational benefit without materially altering the biologically relevant outputs of the simulator. More specifically, the study focuses on the iterative Farquhar–Jarvis photosynthesis routine, which was identified through profiling as the principal runtime bottleneck. The expected outcome is a hybrid simulator in which the surrogate is activated only in the operating regime where approximation quality is sufficiently high, while the original mechanistic routine is retained elsewhere.

The remainder of the paper is organized as follows. Section 2 explains the role of the photosynthesis module within the mechanistic simulator and presents the profiling-based identification of the bottleneck. Section 3 describes the surrogate design, including dataset generation, regime restriction, and the selected MLP architecture. Section 4 summarizes the integration of the trained surrogate into the native C simulation pipeline.

Section 5 presents the experimental evaluation in terms of predictive accuracy, runtime profiling, and preservation of plant growth indicators. Finally, the conclusion summarizes the main findings and outlines directions for future work.

2. PHOTOSYNTHESIS AS A COMPUTATIONAL BOTTLENECK IN PLANT SIMULATION

To examine the role of photosynthesis as a computational bottleneck, it is first necessary to consider its position within the overall simulator architecture. The mechanistic plant model employed in this study is formulated as a coupled dynamic system in which carbon assimilation, storage, transport, nutrient-related processes, and organ-level growth are tightly interconnected. In such a structure, the computational behaviour of an individual module cannot be interpreted in isolation, since its outputs directly affect the evolution of multiple downstream state variables.

Figure 1 presents the overall structure of the mechanistic plant simulator used in this study. The model combines carbon storage and transport variables, nutrient-related processes, and organ-level biomass dynamics into a coupled dynamic system. Within this structure, the photosynthesis block occupies an upstream position: its output directly influences starch accumulation, sucrose dynamics, nutrient-related costs, and ultimately leaf and root growth. Because of this central role, changes in the photosynthesis routine propagate throughout the full simulation pipeline rather than remaining confined to a single local module.

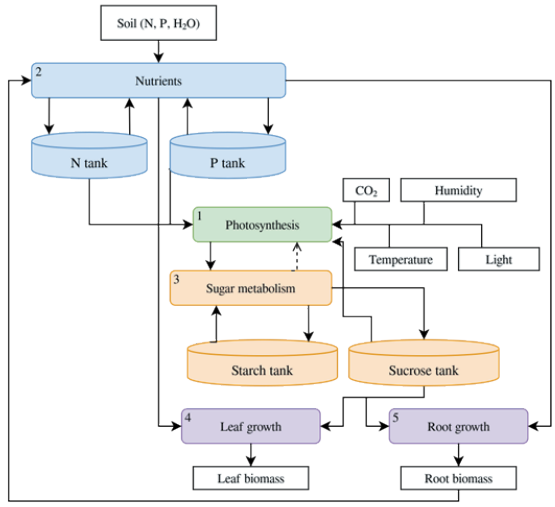

Figure 1. High-level structure of the mechanistic plant simulator and the position of the photosynthesis module within the overall dynamic workflow

2.1. ROLE OF THE PHOTOSYNTHESIS MODULE IN THE MECHANISTIC MODEL

The simulator considered in this work is based on a mechanistic whole-plant formulation derived from the Tedone framework, in which plant development is represented through explicit state variables and process-level dynamic interactions [6]. In this formulation, the plant is modelled through coupled variables such as starch, sucrose, nutrient pools, affinity coefficients, and organ-level biomass, thereby preserving a physiologically interpretable source–sink structure instead of relying on purely empirical growth laws.

To improve physiological realism, the photosynthesis module was extended using a Farquhar-type biochemical model of C3 photosynthesis [7] coupled with a Jarvis-type stomatal conductance formulation [8]. In this implementation the net assimilation rate is computed as

$$An = min(Ac, Aj) - Rd$$

Equation 1. Net assimilation rate

where Ac denotes the Rubisco-limited assimilation rate, Aj the light-limited assimilation rate, and Rd the dark respiration term [7].

Within the full model, the resulting photosynthesis output serves as a direct upstream input to starch production, sucrose production, nutrient costs, and downstream biomass accumulation. This makes the photosynthesis block both physiologically central and structurally influential in the dynamic behaviour of the simulator. The broader behaviour of the underlying mechanistic formulation, including growth responses under different photoperiod conditions, was examined previously in our earlier work [9].

2.2. PROFILING-BASED IDENTIFICATION OF THE BOTTLENECK

Since the objective of this study is not to replace the mechanistic simulator globally, but rather to accelerate its most computationally demanding component, the first step consisted of profiling the native C implementation under representative simulation workloads. The central question was whether the Farquhar–Jarvis photosynthesis module was merely biologically important or also computationally dominant.

This focus is strongly motivated by the structure of the model itself. In contrast to many other algebraic terms in the simulator, which are evaluated directly once per state update, the photosynthesis routine contains an iterative inner loop for resolving the coupled gs - Ci - An relationship. As a consequence, its cost is not limited to a single evaluation of the Farquhar equations but includes repeated recomputation of assimilation and intercellular CO_2 concentration until the diffusion-consistency condition is satisfied. In practical terms, this makes the photosynthesis module qualitatively different from simpler non-iterative expressions elsewhere in the model.

Profiling confirmed that this intuition was correct. In the baseline simulator, the farquhar_photosynthesis routine dominated sampled CPU overhead, accounting for the clear majority of execution cost in the evaluated workload. This established the photosynthesis component as the principal runtime bottleneck of the simulator. Such a result is consistent with the broader observation that increasing physiological realism in mechanistic plant models often leads to growing computational burden, especially when detailed process formulations are embedded in iterative simulation workflows [3], [10].

When the original photosynthesis implementation was replaced with a lightweight approximation, the dominant hotspot disappeared and the runtime profile became substantially more balanced across the solver, right-hand-side evaluation, and the remaining algebraic computations. However, this naive simplification also degraded model fidelity. For that reason, the problem addressed in this paper is not the wholesale replacement of mechanistic photosynthesis, but the construction of a surrogate approximation that preserves the valuable behaviour of the original module while reducing its dominant runtime cost in the regime where approximation is sufficiently reliable.

3. SURROGATE MODEL DESIGN

The use of surrogate models in computational science is primarily motivated by the mismatch between the fidelity of mechanistic models and the runtime constraints of repeated evaluation. In digital twin settings, this tension becomes especially relevant because the model is not executed only once for offline analysis, but repeatedly for updating, scenario exploration, parameter adjustment, and potentially runtime decision support. For this reason, surrogate models are typically not introduced as full replacements for mechanistic formulations, but rather as targeted approximations of computationally expensive components whose repeated evaluation limits operational feasibility [1], [4].

The surrogate model targets the iterative Farquhar–Jarvis coupling by approximating the mappingf ^:u→An. Here, An is the net assimilation rate, while the input vectoru consists of variables available prior to the iterative solve: light_PAR, leaf_biomass, and stomatal_conductance. This submodule-level approach preserves the simulator's mechanistic structure while shifting the computational burden to offline training

This hybrid interpretation is consistent with recent developments in plant and ecosystem modelling, where learning-based approximations are increasingly used to complement rather than discard process-based formulations. For example, differentiable and physics-informed workflows have already been explored for photosynthesis-related simulations and parameter learning, showing that machine-learning-assisted approximations can be useful when embedded within scientifically meaningful model structures [11]. Likewise, recent reviews of agricultural digital twins also emphasize modular coupling between mechanistic, data-driven, and hybrid components rather than a strict opposition between them [5], [12].

Accordingly, the surrogate developed in this work is intentionally narrow in scope. It is trained only to reproduce the net assimilation output of the high-cost photosynthesis subroutine, while the remaining plant dynamics, state updates, and downstream growth processes remain mechanistic. This yields a targeted surrogate design that is both computationally useful and architecturally compatible with the existing simulator.

3.1. SIMULATION-GENERATED DATASET

The surrogate training data were generated directly from the mechanistic simulator. The initial exported dataset contained 14,126 rows and 18 logged variables. After cleanup, 11,361 valid samples remained. For each valid execution of the photosynthesis routine, the simulator logged the input quantities available before the iterative computation of intercellular CO2 and the resulting net assimilation output.

Although 18 variables were logged, several environmental and biochemical parameters remained constant under the nominal conditions considered here. For that reason, the final feature set was reduced to the variables that both varied across calls and were available prior to the iterative solve. The final surrogate inputs were light_PAR, leaf_biomass, and stomatal_conductance, while the target output was target_An.

This simulation-generated dataset reflects a typical surrogate workflow in which a higher-fidelity mechanistic model serves as the data generator for a reduced-cost approximation. Such an approach is particularly attractive when observational data are limited or when the immediate objective is computational acceleration of an already validated simulation core rather than direct empirical replacement [4].

3.2. REGIME RESTRICTION UNDER HIGH IRRADIANCE

Preliminary experiments with a single global MLP showed that the full photosynthesis mapping was not equally easy to approximate across the entire operating range. In particular, the low-light region near the compensation zone exhibited much less regular behaviour, whereas the sufficiently illuminated regime showed a smoother and more learnable relation between inputs and target assimilation rate. This observation is consistent with the fact that photosynthesis responses can change sharply across environmental regimes and that approximation quality depends strongly on the local structure of the mapping being learned [11].

For this reason, the current surrogate was restricted to the region where light_PAR ≥ 70, where approximation quality was substantially better. For lower irradiance values, the original mechanistic routine was retained. The resulting design is therefore regime-aware and hybrid: the surrogate is used only in the region where it achieves both high predictive fidelity and meaningful computational benefit.

3.3. MLP ARCHITECTURE AND TRAINING PROCEDURE

The final surrogate was implemented as a compact multilayer perceptron trained on the high-irradiance subset of the simulation-generated data. The network used three standardized inputs (light_PAR, leaf_biomass, stomatal_conductance), one hidden layer with 32 neurons, and tanh activation. The cleaned dataset was split into 7,952 training samples and 3,409 test samples. Model performance was assessed on the held-out test set, while 5-fold cross-validation was performed on the training subset. A linear regression model was used as a baseline for comparison.

This model choice reflects a deliberate compromise between expressiveness and deployability. The network is sufficiently nonlinear to capture the target mapping in the selected regime, yet small enough to be exported and implemented directly in native C without requiring an external machine-learning runtime.

In this way, the surrogate is not only a statistical approximation, but also an implementation-oriented component designed for integration into the performance-critical simulation pipeline.

4. INTEGRATION INTO THE MECHANISTIC SIMULATOR

After training, the surrogate model was exported from Python into a parameterized form suitable for direct use in the native simulation environment. The exported data included the input and output standardization statistics together with the weight matrices and bias vectors of the trained multilayer perceptron. This allowed the surrogate to be integrated into the existing C-based simulator without introducing an external machine-learning runtime or modifying the surrounding numerical infrastructure.

The deployed surrogate was intentionally kept compact. It consists of a single-hidden-layer multilayer perceptron with 32 neurons and tanh activation, operating on three input variables: light_PAR, leaf_biomass, and stomatal_conductance. In the native C implementation, inference consists of constructing the input vector, standardizing it using the exported scaler parameters, evaluating the hidden and output layers, and transforming the predicted assimilation rate back to the original scale. Because the network is small, the resulting forward pass remains lightweight and compatible with the performance-oriented structure of the simulator.

The surrogate was not introduced as a global replacement for the original photosynthesis routine. Instead, the simulator follows a hybrid regime-aware logic: for sufficiently illuminated conditions (light_PAR >= 70), the iterative Farquhar-based computation of net assimilation is replaced by the surrogate-predicted value, while for lower irradiance levels the original mechanistic routine is retained. As a result, the simulator remains mechanistic in its overall structure, with the learned component serving only as a targeted approximation of the most expensive subroutine in the regime where approximation quality is high.

5. EXPERIMENTAL EVALUATION

The experimental evaluation was designed to assess the proposed surrogate along three complementary dimensions: local predictive accuracy, runtime impact after native integration, and preservation of system-level growth behaviour. First, the surrogate was evaluated as a regression model against the original mechanistic target An.

Second, the simulator was profiled before and after surrogate integration in order to determine whether the dominant computational hotspot had been effectively removed. Finally, the surrogate-enabled simulator was compared against the original implementation at the level of plant growth indicators under multiple photoperiod conditions.

The surrogate was trained and evaluated on simulation-generated data restricted to the high-irradiance regime (light_PAR ≥ 70). The final feature set consisted of light_PAR, leaf_biomass, and stomatal_conductance, while the target was the net assimilation rate An computed by the original iterative photosynthesis routine. On the held-out test set, the compact MLP achieved MAE = 0.014291, RMSE = 0.021259, and R^2 = 0.999742. Cross-validation results were similarly stable, with mean RMSE = 0.024546 and mean R^2 = 0.999663 across folds. A linear regression baseline performed substantially worse (MAE = 0.148727, RMSE = 0.175233, R^2 = 0.982503), indicating that the retained nonlinearity of the MLP was necessary even within the restricted operating regime.

As shown in Figure 2, the surrogate predictions closely follow the identity line, indicating near-perfect agreement with the original mechanistic target. This demonstrates that strong local fidelity can be obtained without a complex machine-learning component.

5.1. RUNTIME PROFILING BEFORE AND AFTER INTEGRATION

Profiling the baseline simulator confirmed the Farquhar–Jarvis module as the primary bottleneck, accounting for 81.46% of total CPU overhead. Following surrogate integration, the original routine's contribution dropped to 1.07%, while the MLP inference routine required only 2.02% of resources. As illustrated in Figure 3, replacing the iterative process created a balanced execution profile, effectively redistributing computational costs to the numerical solver and remaining simulator infrastructure.

5.2. EFFECT ON PLANT GROWTH INDICATORS

The surrogate impact was evaluated under four photoperiods (4h, 6h, 8h, and 12h). Building on the mechanistic baseline validated in our previous work [9], the evaluation focused on whether the hybrid simulator preserves established growth behaviour. Results in Table 1 show that RGR deviations remain close to 0.1%, while

final leaf biomass deviations remain below 1%. For instance, under a 12 h photoperiod, final leaf biomass was 77.87 compared to 77.33 in the original simulator, corresponding to a deviation of about 0.70%. This confirms that local submodule approximation does not disrupt long-term growth dynamics or the physiological realism of the simulator.

These findings confirm that local submodule approximation does not disrupt physiological realism or long-term growth dynamics. The proposed surrogate therefore achieves the desired efficiency trade-off by reducing computational cost while preserving system-level outputs.

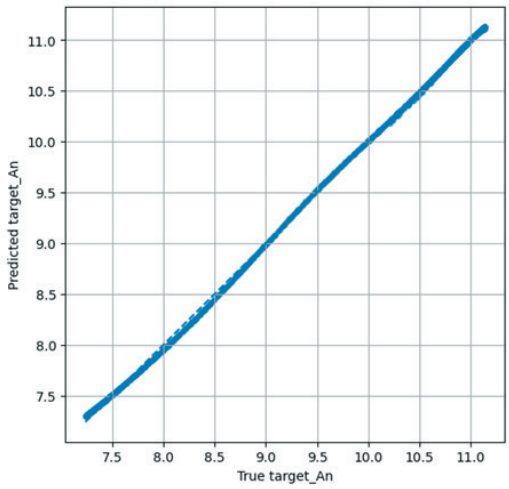

Figure 2. Comparison between mechanistic and surrogate-predicted net assimilation values in the high-irradiance regime

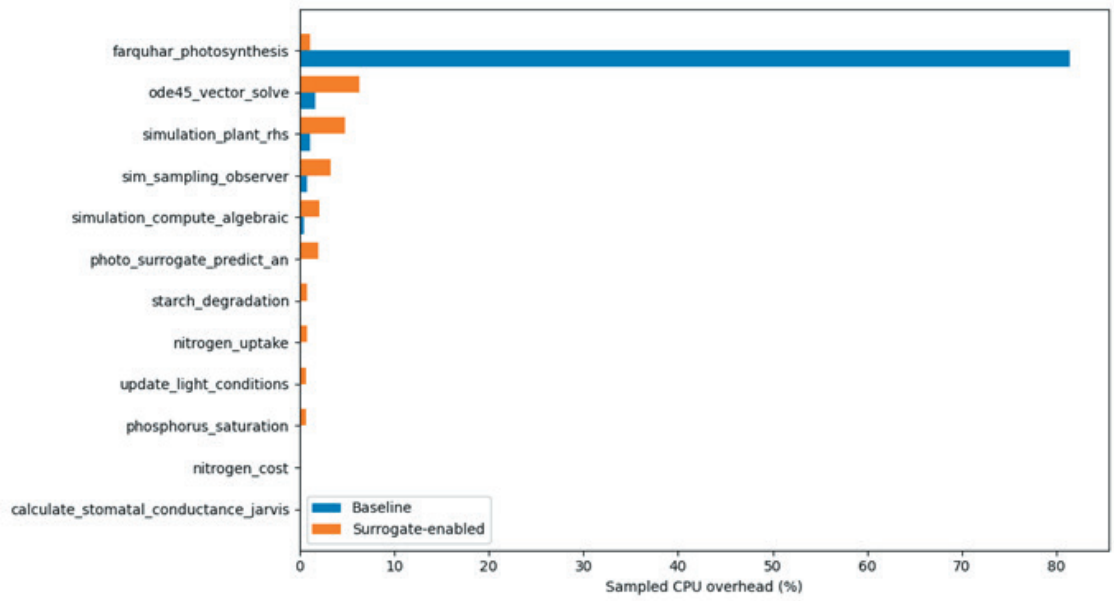

Figure 3. Comparison of sampled CPU overhead contributions before and after surrogate integration

Table 1. Comparison of growth indicators between the original and surrogate simulator across multiple photoperiods

Photoperiod	Original RGR	Surrogate RGR	Original leaf end	Surrogate leaf end
4 h	0.068784	0.068712	0.294014	0.293399
6 h	0.116960	0.116850	1.188824	1.185034
8 h	0.182653	0.182476	7.988965	7.948023
12 h	0.260930	0.261171	77.330114	77.873393

6. CONCLUSION

This paper addressed the problem of high computational cost in a mechanistic plant simulator by focusing on its dominant runtime hotspot, the iterative Farquhar–Jarvis photosynthesis routine. To address this issue, a compact surrogate model was designed for the high-irradiance regime, trained on simulation-generated data, and integrated directly into the native C simulation framework as a regime-aware hybrid component.

The results demonstrate that this approach provides both strong predictive fidelity and substantial computational benefit. The surrogate closely reproduced the original net assimilation target in the selected operating regime, significantly reduced the computational dominance of the original photosynthesis module, and preserved system-level growth indicators such as relative growth rate and final leaf biomass with only minor deviations.

The main contribution of the paper is therefore not the replacement of mechanistic plant simulation by a black-box predictor, but the demonstration that a narrow, implementation-oriented surrogate can remove a major computational bottleneck while preserving the broader structure and behaviour of the model. Future work should extend this approach to additional environmental regimes, particularly the low-light region, and investigate more general strategies for hybrid runtime-efficient plant digital twins.

REFERENCES

[1] C. Pylianidis, S. Osinga, and I. N. Athanasiadis, "Introducing digital twins to agriculture," *Comput. Electron. Agric.*, vol. 184, p. 105942, May 2021, doi: 10.1016/J.COMPAG.2020.105942.

[2] M. Escribà-Gelonch, S. Liang, P. van Schalkwyk, I. Fisk, N. V. D. Long, and V. Hessel, "Digital Twins in Agriculture: Orchestration and Applications," *J. Agric. Food Chem.*, vol. 72, no. 19, pp. 10737–10752, May 2024, doi: 10.1021/ACS.JAFC.4C01934.

[3] D. Wallach, D. Makowski, J. W. Jones, and F. Brun, "Working with dynamic crop models: Methods, tools and examples for agriculture and environment," Working with Dynamic Crop Models: Methods, *Tools and Examples for Agriculture and Environment*, pp. 1–613, Jan. 2018, doi: 10.1016/C2016-0-01552-8.

[4] Á. Bárkányi, T. Chován, S. Németh, and J. Abonyi, "Modelling for Digital Twins—Potential Role of Surrogate Models," *Processes 2021*, Vol. 9, Page 476, vol. 9, no. 3, p. 476, Mar. 2021, doi: 10.3390/PR9030476.

[5] S. Rao et al., "Integrated Modeling Approaches for Agricultural Digital twins: the role of Process Based Models, Agent Based Models, Machine Learning, and Model Coupling," *In Silico Plants*, vol. 8, no. 1, 2026, doi: 10.1093/INSILICOPLANTS/DIAG002.

[6] F. Tedone, "The mathematical modeling of plant growth and applications to robotics," 2020. Accessed: Aug. 19, 2025. [Online]. Available: https://hdl.handle.net/20.500.12571/14997

[7] G. D. Farquhar, S. von Caemmerer, and J. A. Berry, "A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species," *Planta*, vol. 149, no. 1, pp. 78–90, Jun. 1980, doi: 10.1007/BF00386231.

[8] P. G. Jarvis, "The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 273, no. 927, pp. 593–610, Feb. 1976, doi: 10.1098/RSTB.1976.0035.

[9] A. Joksimović, A. Mujezinović, D. Kostić, M. Jolović, and P. Lukovac, "Physiological Simulation of Arabidopsis thaliana for Plant Digital Twin," *E-business technologies conference proceedings*, vol. 4, no. 1, Dec. 2025, Accessed: Apr. 18, 2026. [Online]. Available: https://ebt.rs/journals/index.php/conf-proc/article/view/260

[10] J. Buckley Paules, S. Fatichi, B. Warring, and A. Paschalis, "T&C-CROP: Representing mechanistic crop growth with a terrestrial biosphere model (T&C, v1.5) - Model formulation and validation," *Geosci. Model Dev.*, vol. 18, no. 4, pp. 1287–1305, Mar. 2025, doi: 10.5194/GMD-18-1287-2025.

[11] D. Aboelyazeed et al., "A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: demonstration with photosynthesis simulations," *Biogeosciences*, vol. 20, no. 13, pp. 2671–2692, Jul. 2023, doi: 10.5194/BG-20-2671-2023.

[12] A. C. Tagarakis, L. Benos, G. Kyriakarakos, S. Pearson, C. G. Sørensen, and D. Bochtis, "Digital Twins in Agriculture and Forestry: A Review," *Sensors 2024*, Vol. 24, Page 3117, vol. 24, no. 10, p. 3117, May 2024, doi: 10.3390/S24103117.