

General remarks

We are grateful for the evaluation of the manuscript and have now addressed the issues and questions that will certainly improve the comprehensiveness of the text in a point to point manner.

Specific answers to the remarks ([RC: Reviewer's Comment](#); AC: Authors' Response)

Reviewer 1

RC: I24, this is too strong given what is actually done. A tree-based surrogate trained on model outputs does not preserve mechanistic fidelity in any strict sense. At best, it approximates the response surface of the parent model within the sampled domain.

AC: Thank you for your valuable feedback. We have reviewed the manuscript and updated it as seen below.

I22-24

"Here we present mLDNDCv1.0, a tree-based machine-learning surrogate of LandscapeDNDC that allows for the rapid exploration of large decision spaces while maintaining high fidelity to the parent process-based model's input-output behaviour."

RC: I30, this is somewhat misleading. The so-called "real" conditions still rely on model-generated outputs, not independent observations (except partially for yield). This is not a true real-world validation, and the sentence overstates the level of external validation I would like to say.

AC: Thank you for this valuable observation. We acknowledge that our original phrasing overstated the nature of the validation. To clarify: the surrogate model was trained on synthetic management data with corresponding LandscapeDNDC-simulated outputs (N₂O, NO₃, yield, and SOC) as labels. For validation, real field activity data were used as inputs to both the process-based model and the surrogate, and the surrogate predictions were compared against the LandscapeDNDC outputs. We recognise that this constitutes a model-to-model comparison rather than independent real-world validation. However, for crop yield, we were additionally able to compare the surrogate predictions against observed yield data not derived from any model. We have revised the text to clearly distinguish between these two levels of validation.

I30-32

"When evaluated on real field activity data from Denmark, the surrogate model closely reproduced the process-based model outputs, with particularly strong agreement for crop yield, which was further corroborated by independent observational data."

RC: I32, this sounds more general than it is. It would be more precise to state that the optimization is conditional on the model and the defined decision space.

AC: Thank you for this observation. We agree that the original phrasing could be interpreted as implying a general, model-independent optimisation. We have revised the text to explicitly state that the optimisation operates within a predefined decision space and is conditional on the surrogate model. The identified optimal solutions are therefore subject to both the boundaries of the decision variables considered and the fidelity of the surrogate to the underlying process-based model. We believe the revised wording makes this conditionality clear.

I33-35

“Coupling mLDNDC with the multi-objective evolutionary algorithm NSGA-II allowed us to optimise millions of management combinations within a predefined decision boundary across all winter wheat fields in Denmark.”

RC: I36, 2200 ± 400 Mg CO₂-eq ha⁻¹ yr⁻¹ is highly questionable. The magnitude appears unrealistically large for cropland systems.

AC: Thank you for flagging this. Upon review, we identified a unit error. The reported value of 2200 ± 400 Mg CO₂-eq yr⁻¹ represents the aggregate annual soil-net GHG balance across all winter wheat fields considered, not a per-hectare estimate. The inclusion of ha⁻¹ was erroneous. We have corrected the unit in both the main text (line 36 now 38) and the Figure 7 caption, which now reads:

I774-775

“Figure 7. Soil-net GHG balance (Mg CO₂-eq yr⁻¹) under baseline and optimised management scenarios, averaged over the 10-year period 2011–2020; negative values indicate greater net climate-mitigation potential.”

RC: I39, the study demonstrates a model-based optimization workflow, not a validated decision-support system. "Real-time decision support" in particular is not demonstrated.

AC: Thank you for this important distinction. While development as a real-time and validated decision support tool remains a future goal, we recognise this is beyond the scope of the current work and have ensured the revised manuscript does not imply otherwise. We have removed all references to real-time decision support and revised the sentence to accurately reflect what this study contributes and have rephrased the paragraph as shown below.

I40-42

“Thus, mLDNDC provides a scalable, transparent framework for country-wide scenario comparison and strategic planning at an annual time scale in climate-smart agriculture”

RC: I133, it is not clear whether the design leads to unrealistic combinations, especially when multiple categorical and continuous variables are combined. This needs more justification.

AC: Thank you for this valuable feedback. We address unrealistic combinations using two complementary filters: field-level activity data and scientific reality.

First, field-level activity data served as a baseline to identify and remove implausible combinations. For instance, fertiliser replication (a variable representing how many times fertiliser was applied) could still carry a non-zero value even when both synthetic and organic fertiliser amounts are 0 kg, which is scientifically meaningless. Such rows were removed.

Second, scientific reality was applied to constrain continuous variable ranges. Based on observed field activity data, there are realistic thresholds below which a second or third fertiliser application would not occur in practice. For example, a 30 kg fertiliser amount would not realistically be applied two or three times in Denmark. A comprehensive set of conditions derived from both field data and domain knowledge was therefore used to filter such combinations.

We have further clarified this in the manuscript and updated the relevant section accordingly.

I135-142

“Second, we varied these variables within their agronomic limits using a full factorial design to generate a large set of unique management combinations representing the full range of plausible practices within the Danish agricultural context. While the factorial design is comprehensive by nature, it inevitably produces unrealistic combinations. These were systematically filtered out based on agronomic constraints and consistency with observed field practice. For instance, fertiliser replication (the number of times fertiliser is applied) was set to zero wherever both synthetic and organic fertiliser amounts were 0 kg, and application frequency was constrained by realistic quantity thresholds. For example, a total fertiliser amount of 30 kg would not feasibly support two or three separate applications.”

RC: I148, the optimization results will be entirely conditioned on these boundaries, yet it is not clear how they were defined, how strict they are, or how sensitive results are to these choices.

AC: Thank you for this valuable feedback. You are correct that the optimisation results are conditioned on the defined boundaries. These boundaries were derived directly from field-level activity data, capturing the observed range of current agricultural practices and regulations in Denmark. The rationale is intentional: the goal is to discover management options that fields have not yet explored but that remain within the bounds of current agricultural reality, ensuring that optimised strategies are practically feasible rather than theoretically ideal.

Regarding strictness, the boundaries are data-driven and therefore reflect what has actually been practised in Denmark, making them neither arbitrarily conservative nor overly permissive.

We acknowledge, however, that results are inherently conditioned on these choices, and we have made this more transparent in the updated manuscript as follows.

l155-161

“Management boundaries were identified from Danish agricultural registry data and agronomic guidelines to represent the observed range of current agricultural practices in Denmark. These boundaries served as reference points for defining the limits of each management variable. A management library was then constructed, encapsulating all possible combinations of these management parameters for further use in the factorial design stage. The optimisation algorithm operates within these boundaries, ensuring that identified management strategies are grounded in current agricultural reality while still revealing unexplored but feasible options.”

RC: l214, This is a key step but not described in sufficient detail. How many combinations were removed? How sensitive are results to these rules? This directly affects the training distribution.

AC: Thank you for raising this point. We agree that this filtering step is critical and warranted more detailed description. We have expanded the relevant section accordingly.

In total, approximately 9.1 million rows were removed from the roughly 13.6 million generated combinations, retaining around 4.5 million realistic management scenarios for further use. Regarding sensitivity, the filtering rules are grounded in field-level activity data and scientific knowledge rather than arbitrary thresholds, which limits subjectivity in their application. Rather than distorting the training distribution, this filtering step improves it by removing implausible combinations that would otherwise introduce noise and reduce model reliability. The retained dataset therefore reflects a more focused and realistic representation of the management space, which we argue is preferable for training a model intended to support practical decision-making. The number of combinations removed are stated in section 2.2.

The manuscript has been updated as follows.

l224-228

“Therefore, the synthetic dataset was systematically screened and cleaned using rule-based plausibility checks to eliminate such combinations and ensure that the remaining scenarios reflect realistic and interpretable management configurations. Rather than undermining the training distribution, this step refines it by concentrating the training data on agronomically plausible scenarios, improving both model reliability and the interpretability of optimisation results.”

RC: l245, despite this, the manuscript assumes that prior validation is sufficient for all variables and conditions considered here, which may need explicitly stating here.

AC: We appreciate the reviewer’s comment and agree that our original wording could be read as implying a too comprehensive validation across all variables and conditions. We have revised the

manuscript and now explicitly state that LDNDC has been evaluated against multiple independent outputs (SOC, N₂O, yield) under diverse conditions, which supports the consistency of its process representation and justifies its use as the parent model for our surrogate. However, at the same time, we emphasise that neither LDNDC nor the surrogate can be assumed to be correct in all applications, and that use of the surrogate requires careful evaluation, as is would be the case for the parent model. The paragraph has now been updated as shown below.

l257-263

"Given the extensive international and Danish-level validation of LDNDC cited above, and its demonstrated agreement with observations for multiple quantities (e.g., SOC dynamics, N₂O fluxes, crop yields) under diverse conditions, the model's process representation is well-supported across a range of conditions. LDNDC therefore provides a defensible foundation for the surrogate-modelling work carried out here. Nevertheless, neither LDNDC nor its surrogate can be assumed valid for every variable, management regime, or environmental condition considered. Applications of the surrogate therefore require careful evaluation, as would be expected for the parent process-based model itself."

RC: l305, I would say that a simple random split may not be sufficient to test generalization across management regimes.

AC: Thank you for this valuable feedback. We appreciate this important point. It is worth clarifying that the random split at this stage served a specific and limited purpose: to benchmark and select the best-performing model for full training, rather than to assess generalisation across management regimes. XGBoost was identified as the best-performing model through this comparison and was subsequently taken forward for hyperparameter tuning. The final model was then trained on the full dataset using cross-validation with the tuned hyperparameters, which provides a more robust assessment of generalisation performance. The manuscript has been updated as follows.

l321-325

"Before training on the full data, it is vital to select which model will be used for this final training. The dataset was therefore split into two parts: 80% for model training and the remaining 20% withheld as an independent test set for evaluating generalisation performance. This ratio is widely adopted in machine learning practice as it balances the data available for learning with the need for reliable evaluation (Kuhn & Johnson, 2013)."

RC: l351, but validation against yield alone is not sufficient to claim general transferability, especially for the other target variables.

AC: Thank you for this valuable feedback. We fully acknowledge this limitation. External validation data were only available for winter wheat yield, which constrained our ability to validate transferability across all target variables. We have therefore been careful not to claim general transferability beyond yield in the manuscript. The limitation of lacking external validation data for the remaining target variables has been explicitly stated in Section 4 as part of the study's limitations. The relevant manuscript text has been updated as follows.

l368-379

"This dataset represents real-world observations across diverse management and environmental conditions, providing a meaningful basis for evaluating model transferability with respect to winter wheat yield beyond the simulation domain. While site-scale observations for N₂O, NO₃, and SOC exist in Denmark and have been previously evaluated ((Aderale et al., 2025; Grados et al., 2024; Kollmer, 2023; Rahimi et al., 2024)), national-scale validation data for these variables were unavailable, precluding transferability assessment at the scale considered in this study. It should be noted, however, that LDNDC has been evaluated at site scale for winter wheat systems in other European contexts, including Germany (Haas et al., 2021; Kasper et al., 2018; Molina-Herrera et al., 2016), as well, supporting confidence in its broader process representation. Nonetheless, transferability claims in this study remain limited to yield, and this is acknowledged as a key limitation. Future work should seek to validate surrogate performance across the full set of agroecosystem indicators as suitable observational datasets become available."

RC: l404-406, I think it effectively removes all trade-offs and selects only "win-win" solutions. As a result, the reported improvements are no longer representative of the Pareto space but of a heavily filtered subset. The authors should clarify this point.

AC: We appreciate the reviewer for the valuable feedback. The filtering step was intentionally designed to retain only management solutions that simultaneously outperform the baseline across all target indicators, effectively selecting "win-win" solutions. We acknowledge that this approach does not represent the full Pareto space but rather a filtered subset of it. This choice was deliberate, as the practical goal of the study is to recommend management strategies that offer clear and unambiguous improvements for farmers and policymakers, rather than solutions that involve trade-offs between indicators. Nevertheless, we recognise that this should be stated transparently, and the manuscript has been updated accordingly.

l431-438

"Specifically, an optimised solution was retained only if it produced lower nitrous oxide emissions and lower nitrate leaching while achieving higher crop yield and greater soil organic carbon content than the baseline management for that field. This requirement ensured that

recommended solutions deliver clear improvements across all key sustainability indicators rather than merely redistributing environmental impact. It is important to note that the reported improvements therefore reflect a filtered subset of the full Pareto space, specifically those solutions meeting all four criteria simultaneously, and should be interpreted as such rather than as a representation of all available trade-off solutions."

RC: l483, this applies only to yield, but other variables are not independently validated. The sentence generalizes beyond what is actually shown.

AC: Thank you for this observation. We agree that the original sentence generalised beyond what was directly demonstrated. The independent validation using Statistics Denmark yield data applies solely to winter wheat yield, and we have revised the manuscript to state this explicitly. While external observational datasets for the remaining target variables were unavailable, we note that the surrogate model shows strong agreement with the parent process-based model (LandscapeDNDC) across all output variables, as demonstrated in the cross-validation results, which provides confidence in its predictive capacity more broadly. The absence of independent external validation for these variables is acknowledged as a limitation in Section 4. The relevant manuscript text has been updated as follows.

l512-519

"When model predictions on national data were compared to Denmark national yield data for winter wheat from Statistics Denmark (<https://www.dst.dk>), the fully trained XGBoost model achieved an R^2 of 0.77, indicating strong agreement between predicted and observed winter wheat yields (Figure S.1). This independent validation was limited to yield, as national-scale estimates for the remaining target variables are unavailable at the crop level. However, the surrogate model's close reproduction of LandscapeDNDC outputs across all target variables in cross-validation suggests reliable predictive performance beyond yield, though independent confirmation remains a priority for future work."

RC: l485, not fully convinced; agreement with yield alone does not demonstrate that the model captures underlying processes, especially for nitrogen and carbon dynamics.

AC: Thank you for this comment. We respectfully note that the surrogate model is designed to approximate the outputs of LandscapeDNDC, a process-based model whose representation of nitrogen and carbon dynamics has been extensively validated in prior studies (as detailed in Section 2.2.1). The surrogate's role is not to independently capture biophysical processes, but to faithfully reproduce the behaviour of the parent model. Our cross-validation results demonstrate that the surrogate achieves this across all target variables, including nitrogen and carbon indicators. The agreement with independent yield observations provides an additional layer of confidence by confirming that both the PBM and the surrogate produce outputs consistent with

real-world data for at least one key variable. We have revised the text to clarify this distinction between process representation, which is inherited from the PBM, and predictive skill, which is evaluated through both cross-validation and independent yield data.

1519-527

"This level of performance demonstrates the surrogate model's predictive skill for winter wheat yield at a national scale and, together with the strong cross-validation agreement across all target variables, supports the model's capacity to faithfully approximate LandscapeDNDC's process-based outputs. Since the surrogate inherits its representation of nitrogen and carbon dynamics from the parent model, whose process fidelity has been established through extensive prior validation, agreement with independent yield observations provides additional confidence in the overall modelling framework. Comprehensive validation of nitrogen and carbon indicators against independent estimates at the crop level remains a priority for future work as suitable datasets become available."

RC: 1548, that's true, yet the interpretation in the section occasionally goes beyond this (consistency check that the surrogate has not learned spurious relationships). Given that the surrogate is trained on model-generated outputs, the SHAP analysis primarily reflects the behavior of the parent model instead of independent evidence of processes. This distinction could be made more explicit in the section.

AC: Thank you for this observation. We agree that this distinction warrants explicit statement and have updated the manuscript accordingly. As the surrogate is trained on LandscapeDNDC outputs, the SHAP analysis naturally reflects the behaviour of the parent model, and we do not claim otherwise. However, we would note that this is precisely the purpose of the analysis in our context: to verify that the surrogate has faithfully learned the input-output relationships encoded in the process-based model rather than fitting to spurious patterns in the training data. The fact that the resulting SHAP patterns align with well-established agronomic relationships documented in the literature (as referenced throughout the section) provides two-fold assurance: first, that the surrogate reliably reproduces the parent model's process logic, and second, that the parent model's own representations are consistent with domain knowledge. We have revised the text to make the interpretive scope of the SHAP analysis explicit.

1589-596

"Since the surrogate is trained on LandscapeDNDC-generated outputs, the SHAP analysis presented here reflects the learned behaviour of the parent model rather than providing independent empirical evidence of biophysical processes. This is by design: the analysis serves to verify that the surrogate faithfully captures the process logic of the parent model rather than fitting to spurious patterns. The consistency of the resulting SHAP attributions with agronomically

established relationships reported in the referenced literature provides confidence in both the surrogate's fidelity and the coherence of the parent model's process representations."

General remarks

We are grateful for the evaluation of the manuscript and have now addressed the issues and questions that will certainly improve the comprehensiveness of the text in a point to point manner.

Specific answers to the remarks ([RC: Reviewer's Comment](#); AC: Authors' Response)

Reviewer 2

RC: First is about the mechanistic fidelity claim. The paper repeatedly claims about preserving mechanistic fidelity but actually the surrogate model only preserve the input-out mapping of the process based model, not its mechanistic processes and it cannot report the intermediate state variables. This distinction should be stated clearly.

AC: We thank the reviewer for this clarification and agree that the phrasing "mechanistic fidelity" could be misinterpreted. Our intended meaning is that the surrogate preserves the input-output behaviour of a mechanistic model i.e., it reproduces LandscapeDNDC's responses with high accuracy, rather than that it replicates the underlying process representations or intermediate state variables. We have revised the manuscript to make this distinction explicit, replacing "mechanistic fidelity" with language that more precisely describes what is preserved (e.g., "fidelity to the process-based model's predictions").

l22-24

"Here we present mLDNDCv1.0, a tree-based machine-learning surrogate of LandscapeDNDC that allows for the rapid exploration of large decision spaces while maintaining high fidelity to the parent process-based model's input-output behaviour."

RC: Second is the path dependency and legacy effects in agricultural systems. This approach ignores the trajectory-dependent nature of agroecosystem state variables. Soil organic carbon accrual, microbial community composition and other soil health related variables are slow variables whose current states constrain future management responses and have feedback effects. The discussion should acknowledge that two-year rotation categorical variable is an insufficient proxy for these cumulative legacy dynamics.

AC: We thank the reviewer for raising this important point. We agree that agroecosystem state variables such as soil organic carbon and microbial community composition are trajectory-dependent, and that a two-year categorical rotation variable cannot fully capture these cumulative legacy dynamics. We note that LandscapeDNDC does internally simulate these slow variables over the full modelling period, and so the surrogate's training data implicitly reflects their influence on outputs. However, the surrogate itself has no explicit representation of soil state trajectories, and cannot generalise to management histories outside those captured in the

training simulations. We have added a paragraph to Section 4 discussing this design boundary and outlining how future versions could incorporate temporally explicit management sequences or slow state variables as additional inputs.

1863-873

"The current surrogate framework represents management history through a categorical two-year crop rotation variable, which indexes distinct management trajectories as simulated by LandscapeDNDC. Because the process-based model tracks trajectory-dependent state variables, including soil organic carbon accrual and microbial community development, internally over the full modelling period, the surrogate's training data implicitly reflects their influence on modelled outputs for the rotation sequences considered. However, the surrogate does not explicitly represent the evolution of these slow state variables, and its applicability is therefore bounded by the management histories included in the training simulations. Future development could extend the framework by incorporating temporally explicit management sequences or slow state variables as additional inputs, further enhancing its capacity to represent path-dependent system behaviour across a broader range of management histories."

RC: The third issue is temporal specificity and the real-time decision support claim. The surrogate predicts annual totals from static or seasonally aggregated features. Claiming suitability for real-time decision support overstates the tool's capability. In operational terms in farming conditions, real time decision making is much difficult because of within season interactions effects of uncertain weather, crop phenology, soil moisture dynamics among others. It would be better to state it as scenario-comparison and strategic planning tool.

AC: Thank you for this point. We agree that the original phrasing did not accurately reflect the temporal resolution of the surrogate, which operates on annual totals derived from static or seasonally aggregated inputs. The current framework is designed for evaluating and comparing management strategies at a planning level rather than for guiding within-season operational decisions, which would require resolving sub-annual dynamics that are beyond the scope of mLDNDC in its present form. We have revised the relevant text in the conclusions to describe the tool in terms consistent with its actual capabilities.

140-42

"Thus, mLDNDC provides a scalable, transparent framework for country-wide scenario comparison and strategic planning at an annual time scale in climate-smart agriculture".

RC: Overall, the manuscript reads well. As reported by the first reviewer, the author needs to address the circular validation architecture used in the study as a limitation. Also, it would be

better to be more concise in SHAP discussion sections as the post-hoc interpretability of a statistical surrogate does not substitute for mechanistic diagnosis.

AC: Thank you for this feedback. We have been more explicit with the SHAP discussion sections and revised the text to clearly situate the interpretability analysis within its appropriate scope. Regarding the circular validation architecture, this has been addressed as a limitation in Section 4, as also requested by the first reviewer.

1589-596

"Since the surrogate is trained on LandscapeDNDC-generated outputs, the SHAP analysis presented here reflects the learned behaviour of the parent model rather than providing independent empirical evidence of biophysical processes. This is by design: the analysis serves to verify that the surrogate faithfully captures the process logic of the parent model rather than fitting to spurious patterns. The consistency of the resulting SHAP attributions with agronomically established relationships reported in the referenced literature provides confidence in both the surrogate's fidelity and the coherence of the parent model's process representations."