REVISION NOTES- Manuscript "A hybrid framework for the spin-up and initialization of distributed coupled ecohydrological-biogeochemical models" (GMD, egusphere-2025-4796) by Lian et al.

Reviewer #2

The manuscript "A hybrid framework for the spin-up and initialization of distributed coupled ecohydrological-biogeochemical models" introduces a practical framework for initializing spatially distributed ecohydrological–biogeochemical models without the computational burden of long-term spin-up simulations. The authors combine a plot-scale spin-up that accounts for lateral water and solute fluxes with a machine-learning approach using random forests (RF) to extrapolate steady-state biogeochemical pools from a sample of representative grid cells to the entire domain. They implement and test this framework in the T&C-BG-2D model for the Erlenbach catchment, evaluating its performance across several synthetic scenarios that vary vegetation, soil, and topography. The results show that tracking fluxes in only a portion of cells (about 20–40%, depending on landscape heterogeneity and chosen predictors) can capture most of the spatial variation in soil organic carbon and nitrogen, while reducing computational costs by up to 90%.

Overall, the paper tackles a significant and practical challenge in spatially distributed ecohydrological and biogeochemical modeling, offering a creative hybrid solution that blends process-based and data-driven methods. However, several methodological gaps and clarity issues prevent a full evaluation of the robustness and generalizability of the proposed framework. Most notably, the machine-learning component is under-documented, the sampling strategy is not clearly reproducible, and the computational advantages are not quantified in sufficient detail. Addressing these issues will require substantial revision to strengthen transparency, reproducibility, and confidence in the proposed approach. Below, I present my comments to improve the manuscript's quality for publication.

> **Reply:** We thank the Reviewer for the positive assessment of our work and the numerous useful suggestions which allowed us to improve the clarity, accuracy, and reproducibility of our work. In the revised version, we have addressed all the major and minor comments provided, as detailed in our replies below. Specifically, we have (1) described in more detail the random forest (RF) model component and evaluated its robustness based on cross-validation and interpretability analyses, (2) clarified the pixel selection strategy and extended the discussion of the implementation of the proposed spin-up approach across different case studies, and (3) reported the computational saving by documenting the wall-clock times for the different components of the spin-up procedure.
>
> For revisions in the manuscript, **deleted text** is shown in gray as *Deleted:XX*, and **added** *text* is in blue. Line numbers refer to the **original submitted manuscript.**

Overall assessment

This is a strong, original contribution with clear practical relevance. The hybrid process-based and machine-learning framework is innovative and well motivated. The main weaknesses are the clarity and reproducibility of the RF component and the training data selection strategy. Addressing these areas would make the manuscript even more useful to the wider modeling community.

> **Reply:** We sincerely thank the Reviewer for the positive assessment of our work and for recognizing the novelty of the proposed hybrid framework. We agree that some portions of our work required more in depth description and discussion (similar points were also raised by Reviewer #1). We have thus substantially revised the manuscript to improve the description of the random forest implementation, clarify the predictor selection and sampling strategy, and explicitly discuss how the proposed framework can be adapted to different case studies. We believe these revisions have enhanced the transferability and accessibility of the proposed framework, making it easier to pick up by the wider modeling community.

Major comments:

1. Random forest model needs a clearer description and validation: The RF component is central to the framework, yet its implementation is described only at a high level. Important details are missing, including RF hyperparameters (e.g., number of trees, tree depth, minimum samples per leaf, feature selection strategy), whether a single RF model is trained for each target variable (SOC, SON, DOC) or separate models are used, the software/library used, and whether any tuning was performed and how training and prediction were conducted in practice (e.g., were all tracked cells used for training and then predictions made for all cells, or was any form of cross-validation used?). Moreover, performance is evaluated mainly by comparing spatial distributions against the benchmark simulation. While this is useful for assessing pattern similarity, it does not directly assess predictive accuracy at the grid-cell level. Some form of cross-validation or hold-out evaluation (e.g., training on a subset of tracked cells and testing on the remaining tracked cells) would substantially strengthen confidence in the RF component. These missing details limit reproducibility plus make it difficult for readers to assess how robust the RF results are.

> **Reply:** Thank you for these constructive suggestions. We agree that implementation details of the random forest (RF) model require a more in depth description and that a more robust validation is also needed (this point was also raised by Reviewer #1).
>
> In our study, the RF component was implemented using the *TreeBagger* function in MATLAB (Statistics and Machine Learning Toolbox), which implements the standard random forest algorithm of Breiman (2001) based on bootstrap-aggregated CART regression trees. For each carbon and nitrogen pool, a separate RF regression model was trained. Each model consisted of 100 regression trees and default settings were used for tree depth and minimum leaf size. At each split, a random subset of predictors (one-third of the total predictors, following the default MATLAB regression setting) was considered.
>
> Regarding the evaluation, in addition to comparing spatial patterns with the benchmark simulation, we performed k-fold cross-validation (k=10) to explicitly assess predictive robustness. For each pool, models were trained on 90% of tracked cells and evaluated on the remaining 10% subset. The predictive performance was quantified using the mean absolute percentage error (MAPE) and the normalized root mean squared error (NRMSE), calculated by normalizing RMSE with the range (maximum minus minimum)

of the observed values. These metrics are reported in **Table R1** below (same as Table S2 in the revised manuscript). This procedure directly evaluates grid-cell level predictive accuracy. We further conducted a SHAP-based interpretability analysis (Shapley values are reported in the same Table) for the RF trained using 40% of the tracked cells in the original scenario (as suggested by Reviewer #1). This allowed us to quantify the contribution of each predictor to the RF predictions.

In the methods section, we have added a paragraph after line 170 to report these additional analyses and information:

"Random forest models consisting of 100 regression trees were trained to predict each carbon and nitrogen pool. This was implemented using the TreeBagger function in MATLAB R2024b (Statistics and Machine Learning Toolbox). Default settings were used for tree growth and, at each split, a random subset of predictors (one-third of the total predictors) was considered. The robustness of the RF model was validated using k-fold (k=10) cross-validation (Stone, 1976), and by evaluating the mean absolute percentage error (MAPE) and the normalized root mean squared error (NRMSE). NRMSE was calculated by normalizing RMSE by the range (maximum minus minimum) of the observed values, and both metrics were averaged across the 10 folds (see Table S2)."

Table R1 (same as Table S2 in the supplementary information). *Mean absolute percentage error (MAPE) and normalized root mean squared error (NRMSE) of the random forest models trained using 40% of the tracked cells for predicting carbon and nitrogen pools in the original scenario. Both metrics are averaged over 10-fold cross-validation. Normalized mean absolute Shapley values are reported to indicate the relative importance of the six predictors.*

Soil organic carbon pools	Robustness Metrics		Normalized mean absolute Shapley values					
	MAPE (%)	NRMSE	Elevation	Slope	Flow accumulation area	Curvature	Vegetation type	Sand content
Below-ground Litter Metabolic	3.09	0.05	22.4%	1.9%	6.7%	1.2%	65.4%	2.5%
Below-ground Litter Structural - Cellulose/Hemicellulose	5.99	0.04	10.4%	2.3%	2.8%	1.1%	81.8%	1.6%
Below-ground Litter Structura - Lignin	12.06	0.05	7.3%	2.2%	2.4%	1.2%	85.1%	1.7%
SOM-POC- lignin	11.49	0.05	6.2%	3.0%	3.4%	0.9%	85.1%	1.5%
SOM-POC -Cellulose/Hemicellulose	2.59	0.07	30.1%	10.2%	12.1%	1.7%	44.4%	1.4%
SOM-MOC	2.73	0.06	24.7%	7.2%	7.2%	1.6%	58.0%	1.3%
DOC - for bacteria	3.19	0.13	58.0%	14.3%	9.7%	3.6%	3.9%	10.5%
DOC - for fungi	2.66	0.13	58.9%	11.7%	9.5%	2.8%	5.2%	11.9%
Enzyme for decomposition of POC-Bact	2.93	0.07	18.2%	13.3%	16.4%	2.2%	43.4%	6.6%
Enzyme for decomposition of POC-Fung	2.49	0.07	16.1%	7.1%	8.2%	2.0%	61.3%	5.3%
Enzyme for decomposition of MOC-Bact	2.94	0.08	20.1%	12.5%	15.1%	2.7%	43.8%	5.8%
Enzyme for decomposition of MOC-Fung	2.48	0.07	16.7%	6.5%	7.7%	1.8%	62.0%	5.3%
Bacteria pool	5.28	0.08	12.4%	14.7%	16.7%	2.3%	52.2%	1.7%
Fungi saprotrophic	4.27	0.07	10.1%	8.5%	10.3%	1.5%	68.2%	1.5%
AM-Mycorrhizal - C	2.28	0.04	7.5%	1.3%	1.7%	0.7%	87.1%	1.7%
Soil organic nitrogen pools								
Nitrogen Above-ground Litter	4.39	0.04	12.6%	0.9%	1.5%	0.7%	82.9%	1.4%
Nitrogen Above-ground Woody	0.90	0.05	87.1%	4.7%	3.4%	2.2%	0.0%	2.6%
Nitrogen Below-ground Litter	3.02	0.06	20.3%	3.6%	10.2%	1.8%	60.3%	3.8%
Nitrogen SOM	1.52	0.07	35.9%	5.0%	6.1%	1.1%	50.4%	1.5%
Nitrogen Bacteria	5.20	0.08	11.2%	16.2%	17.2%	1.9%	51.7%	1.9%
Nitrogen Fungi	4.22	0.06	9.2%	8.9%	9.4%	1.6%	69.8%	1.2%
AM Mycorrhizal - N	2.21	0.04	9.6%	1.4%	1.3%	0.9%	85.0%	1.9%
Nitrogen Ione Ammonium NH4+	2.25	0.06	16.5%	3.7%	3.5%	1.4%	72.4%	2.5%
Nitrogen Nitrate NO3-	6.03	0.15	15.5%	17.2%	17.6%	2.8%	30.3%	16.5%
DON	6.98	0.12	32.5%	5.9%	14.1%	4.0%	29.1%	14.4%

2. The procedure for selecting tracked cells is unclear: The authors state that the distributions of predictors in the training subsets preserve those of the full domain, which is good practice. However, the manuscript does not clearly describe how the tracked cells are selected. Is the sampling random, stratified, clustered, or manually curated? Are any restrictions imposed to ensure coverage of extremes (e.g., high/low elevation, wet/dry areas)? Is the sampling procedure deterministic or random? Since RF performance is sensitive to how well the training set covers the predictor space, this step is important. A reproducible sampling strategy (e.g., hierarchical sampling across bins of elevation, slope, and vegetation type) should be explicitly described and, ideally, formalized within the proposed framework.

Reply: In this work, tracked cells were selected using a stratified random sampling approach based on vegetation types. Within each vegetation class, cells were randomly sampled in proportion to their occurrence in the full domain, ensuring that the relative distribution of vegetation types was preserved across different sampling fractions (10%, 20%, 40%, 60%, and 80%). The sampling procedure was random but reproducible, as a fixed random seed was applied.

We only did the stratified sampling on the vegetation types, as it is the most important predictor in the case study (**Table R1**) and our primary objective was to assess RF performance in reproducing overall spatial distribution across sampling fractions rather than capturing extreme values (tails of the distribution). As shown in **Fig. R1** (same as Fig. S2), stratification based only on vegetation types also maintains broadly similar distributions of elevation, curvature, and soil texture characteristics across sampling fractions. We acknowledge that hierarchical sampling across bins of predictors could be applied in cases where predictors are highly spatially heterogeneous or patched.

We have revised the sentences in the Methods section (line 159) and Discussion section (line 403) to clarify the applied stratified sampling approach and acknowledge the alternative potential uses of a formalized sampling strategy.

Line 159: "*The covariates used as predictors include.....and vegetation type (forest or grassland). A stratified random sampling approach based on vegetation classes was used for pixels selection, meaning that grid cells were first grouped according to vegetation type (in this case grassland and forest) and, within each group, pixels were randomly sampled without replacement according to a prescribed sampling fraction (10%, 20%, 40%, 60%, and 80%). The steady-state pools predicted……. Because the distributions of topographic, soil, and vegetation characteristics across the full domain are known in advance and do not exhibit pronounced extreme tails, Deleted: the selected training cells were chosen to, stratified random sampling based on vegetation types leads to training subsets that largely preserve these distributions and thus ensure representativeness (Fig. S2). In cases where predictors exhibit strong spatial heterogeneity or patchiness, more structured hierarchical sampling across multiple predictor bins could be implemented to ensure a balanced representation of environmental gradients."*

Line 403: *"*.....*However, this is likely due to the fact that vegetation in the Erlenbach does not experience water limitation. Conversely, in drier regions, soil texture could become a major constraint and separate spin-ups for each soil type may be required instead of using an average soil texture. Apart from soil texture, we did not include elevation-based clusters in step (a) of the initialization procedure (Fig. 2), which is potentially necessary if the catchment has significant elevation changes or it spans climatic regimes where processes such as permafrost occurrence or soil freezing are relevant……."*

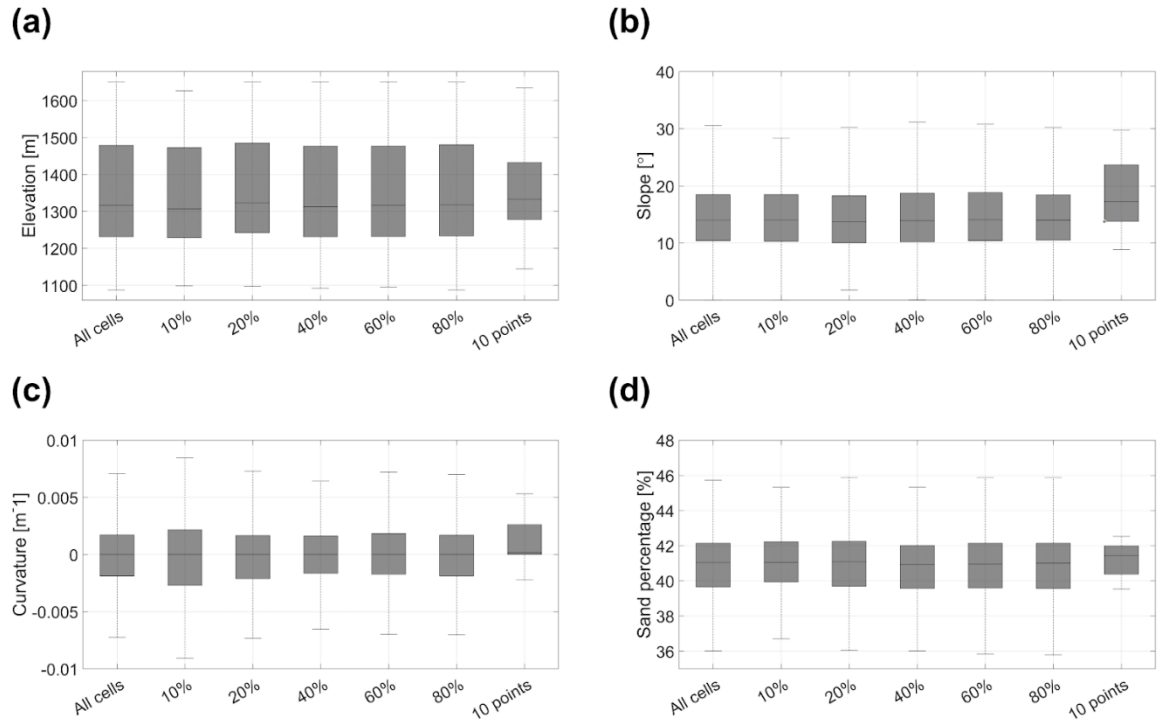

Fig. R1 (same as Fig. S2 in the supplementary information). *Distribution of topographic and soil attributes for all grid cells and selected subsets used for random forest model training. Panels show the distributions of (a) elevation, (b) slope, (c) curvature, and (d) sand percentage for all cells in the domain ("All cells"), different percentages of tracked training cells (10%–80%), and the 10 selected reference points for additional plot-scale simulation. The distributions indicate that the selected subsets preserve the key topographic and soil characteristics of the full domain.*

3. Applicability beyond the Erlenbach catchment: The authors emphasize that the framework is general and applicable to other catchments and models. While the conceptual approach is general, the numerical demonstration is limited to a single catchment with only two vegetation types, relatively homogeneous soils, and a specific alpine climate setting. The synthetic scenarios (randomized vegetation, soil, and modified topography) are useful stress tests, but they do not replace testing on a truly independent site. The manuscript would benefit from a more explicit discussion of how predictor sets should be adapted for different hydro-climatic regions, how many tracked cells might be needed in more heterogeneous landscapes, and whether the RF approach is expected to extrapolate reliably beyond the range of conditions represented in the tracked cells. Even without an additional case study, clearer direction or a conceptual workflow for transferring the method to new regions (e.g., how to choose predictors and sampling density in a new basin) would strengthen the paper.

Reply: We thank the Reviewer for this thoughtful and constructive comment. We agree that providing a clear workflow to transfer the spin-up method to a new case study will be useful for potential users. Combined with related comments from Reviewer #1, we have extended the discussion section to better discuss the limitation of the site-specific implementation in Erlenbach and to include clear working steps for implementing the proposed spin-up procedure in a different catchment.

Line 394: *".....However, simulations from the Random Soil scenario (Section 3.2) highlight the essential role of including soil information among the RF predictors, particularly in areas characterized by high soil spatial variability. Shapley values in*

Table S5 show that clay content is the most important predictor (after vegetation type) for soil carbon prediction, suggesting that clay content should be explicitly considered in generalized RF modelling frameworks."

Line 403: *".....However, this is likely due to the fact that vegetation in the Erlenbach does not experience water limitation. Conversely, in drier regions, soil texture could become a major constraint and separate spin-ups for each soil type may be required instead of using an average soil texture. Apart from soil texture, we did not include elevation-based clusters in step (a) of the initialization procedure (Fig. 2), which is potentially necessary if the catchment has significant elevation changes or it spans climatic regimes where processes such as permafrost occurrence or soil freezing are relevant. Furthermore, soil–vegetation coupling is inherently site-specific, and the decoupled spin-up here relies on a 9-year average of plant and soil dynamics. While this assumption is reasonable for the Erlenbach site, it may introduce biases in more extreme ecosystems (e.g., nutrient limited environments, more variable climatic conditions), where long-term average plant-soil biogeochemical dynamics cannot be adequately captured within a 9-year period. We therefore suggest evaluating this assumption by running longer-term simulations for a small subset of grid cells and verifying that the plant–soil dynamics averaged over the chosen period do not introduce systematic biases relative to longer-term simulations. In addition to the sensitivity analysis discussed in the previous paragraphs, we recommend that, when implementing the proposed spin-up procedure in a new case study, the first step should be to assess the distribution of key environmental predictors in the study area (particularly vegetation type and soil texture) and to apply a proper sampling strategy to represent the predictor space. It is also essential to evaluate whether the ecosystem is nutrient-limited and whether the meteorological forcing is representative of the long-term climatic conditions. Based on these considerations, a tracked-cell proportion of n = 10–40% may serve as an initial guideline, with the final choice determined by the trade-off between target accuracy and available computational resources."*

In addition, we note that the manuscript already clarifies that the proposed framework is currently designed for sites whose vegetation is assumed to be at a mature stage not subject to major disturbances (i.e., for which the steady-state assumption is reasonable). The final paragraph of the discussion section also outlines possible modifications required for applications under non–steady-state conditions, thereby completing the discussion on the generality of the proposed framework.

4. Report computational savings more concretely: The manuscript states that the approach lowers computational cost by up to ~90%, which is compelling. However, this is mostly framed in terms of reduced spin-up years and relative overhead. The argument would be stronger if the authors provided actual runtimes (e.g., wall-clock time or CPU hours) for the benchmark spin-up versus the proposed approach, the computational cost of training and applying the RF relative to the model simulations, and/or an explicit cost–accuracy tradeoff curve (e.g., runtime vs. PDF overlap as n increases). This would help readers assess whether the method is beneficial in their own computing systems.

Reply: Thank you for this helpful suggestion. We agree that, even if the computational savings depends on the computation platform used, it is helpful to report the actual computation time in our devices (this point was also raised by Reviewer #1). As such, in the revised manuscript, we have now detailed the computation time for different components of the spin-up procedure.

The computational cost of a traditional spin-up consists primarily of two components: (i) plot-scale initialization and (ii) long-term two-dimensional (2D) simulations required

to reach steady state (300 simulation years in this study). In contrast, the proposed hybrid spin-up approach includes: (i) plot-scale initialization, (ii) an initial 9-year 2D simulation with lateral fluxes tracked and stored, (iii) spin-up simulations for the tracked cells, (iv) RF model training, and (v) a second 9-year 2D simulation initialized using RF-based estimates.

In general, plot-scale initialization is identical in both approaches and represents only a minor fraction of total cost. The dominant contribution to wall-clock time is from executing the 2D simulations. The computational cost associated with RF training is negligible (even in relation to the plot-scale simulations), and the cost of spin-up simulations for tracked cells increases approximately linearly with their number but remains relatively small.

We now report wall-clock times for each component of both the traditional and hybrid approaches in **Table R2**. In the hybrid configuration, the 2D simulation length is reduced from 300 years to two 9-year simulations (18 years total, approximately 6% of the original simulation duration). Runtime increases with the fraction of tracked cells due to additional input/output (I/O) operations required to store lateral fluxes. For the recommended configuration (tracking 40% of cells) the total runtime is reduced by approximately 86% on our workstation (Intel CPU, 40 cores, base frequency 2.00 GHz, 384 GB RAM). Notice that absolute runtimes depend on hardware and I/O performance. We observe that computational savings are generally larger on high-performance computing systems where I/O operations are more efficient. Also, the relatively large I/O load in our implementation is partly attributable to the MATLAB-based framework. Implementations in lower-level or more I/O-optimized languages (e.g., C/C++ or Julia) would be expected to further reduce runtime.

In addition to **Table R2** that reports the wall-clock time, we also revised the result section at line 342 as follows:

"This uncoupled spin-up steady state can be reached by first running the fully coupled model for only a short period (here 9 years) to obtain average vegetation fluxes, which are then used to drive the biogeochemistry-only spin-up. This provides a substantial gain in computational efficiency (the computation time for decoupled spin-up is negligible, see Table S6) despite the slight disagreement in steady state, thus……"

We have revised lines 347-352 to describe the computation efficiency gain:

"In summary, the domain used here contains 1859 cells in total. Tracking lateral fluxes in all cells increases computational cost significantly (e.g., by approximately 50% compared to a simulation without tracking any fluxes when n=20%), whereas tracking only n = 10% of the cells has less than 10% impact on its overall spin up procedure (Table S6) Delated: simulation time*. In Erlenbach, SOC requires over 300 years of simulation to reach steady state (Fig. S9* Deleted: 8*) even without coupled vegetation-soil biogeochemical dynamics. The proposed initialization framework reduces this demand by collapsing the full 300-year 2D spin-up into two 9-year simulations (corresponding to Figs. 2b and 2f), combined with a 1D plot-scale spin-up. Ideally, the hybrid spin-up procedure requires two 9-year 2D simulations instead of 300 years (i.e., 18 years in total, corresponding to approximately 6% of the original simulation length). Wall-clock times of different components of the spin-up procedure are reported in Table S6. While these times are expected to change based on the specific computational platform used, for the case here the hybrid spin-up procedure* Deleted: This *resulted* Deleted: s *in a computational saving of approximately 90% using the recommended n=40% configuration."*

Table R2 (same as Table S6 in the supplementary information). *Wall-clock times for different components of the original and hybrid spin-up procedures. Results are shown for different fractions of tracked cells (n = 10%, 20%, 40%, and 100%). In the 2D simulations, T_w denotes the wall-clock time required to simulate one year and N_y is the number of simulated years. For the tracked cell spin up, T_w represents the wall-clock time per tracked cell and N_{cell} denotes the number of tracked cells.*

Computation Demand	Plot-initialization (9 years) [h]	2D simulation (T_w * N_y) [h]				Sum [h]
Original spin-up	0.11	2031 = (6.77*300)				2031.11
Hybrid spin-up with n=X%	Plot-initialization (9 years) [h]	2D simulation with tracking (T_w * N_y) [h]	Tracked cells spin up (T_w * N_{cell}) [h]	RF traning	Second 2D simulation [h]	Sum [h]
n=10%	0.11	75.33 = (8.37*9)	2.12= (0.0114*186)	Negligible	60.93	138.49
n=20%	0.11	90.90 = (10.1*9)	4.24 = (0.0114*372)	Negligible	60.93	156.18
n=40%	0.11	216.00 = (24*9)	8.48 = (0.0114*744)	Negligible	60.93	285.52
n=100%	0.11	433.53 = (48.17*9)	21.19 = (0.0114*1859)	Negligible	60.93	515.76

5. Clarifying the workflow and Figure 2: Figure 2 is central to understanding the proposed framework, but it is quite dense and may confuse readers unfamiliar with T&C-BG-2D. While the caption is detailed, the figure could benefit from a clearer labeling of which steps are 1D vs. 2D simulations, an explicit indication of where the RF is trained and where it is applied, and possibly splitting the figure into two panels (mechanistic spin-up vs. ML extrapolation) for clarity. I believe that improving the clarity of this figure would significantly enhance the manuscript's accessibility.

Reply: We very much appreciate this suggestion regarding Figure 2. We agree that the original Figure 2 was visually dense and may have been difficult to follow. We have now revised the figure (see **Fig. R2** showing the revised Fig. 2) to improve readability. Specifically, we clearly separated the workflows into the benchmark without machine learning technique applied, and the alternative extrapolations using the random forest. We also explicitly labeled all 1D simulations, 2D simulations, and random forest models with color-blind friendly colors, to clarify the model used at each step.

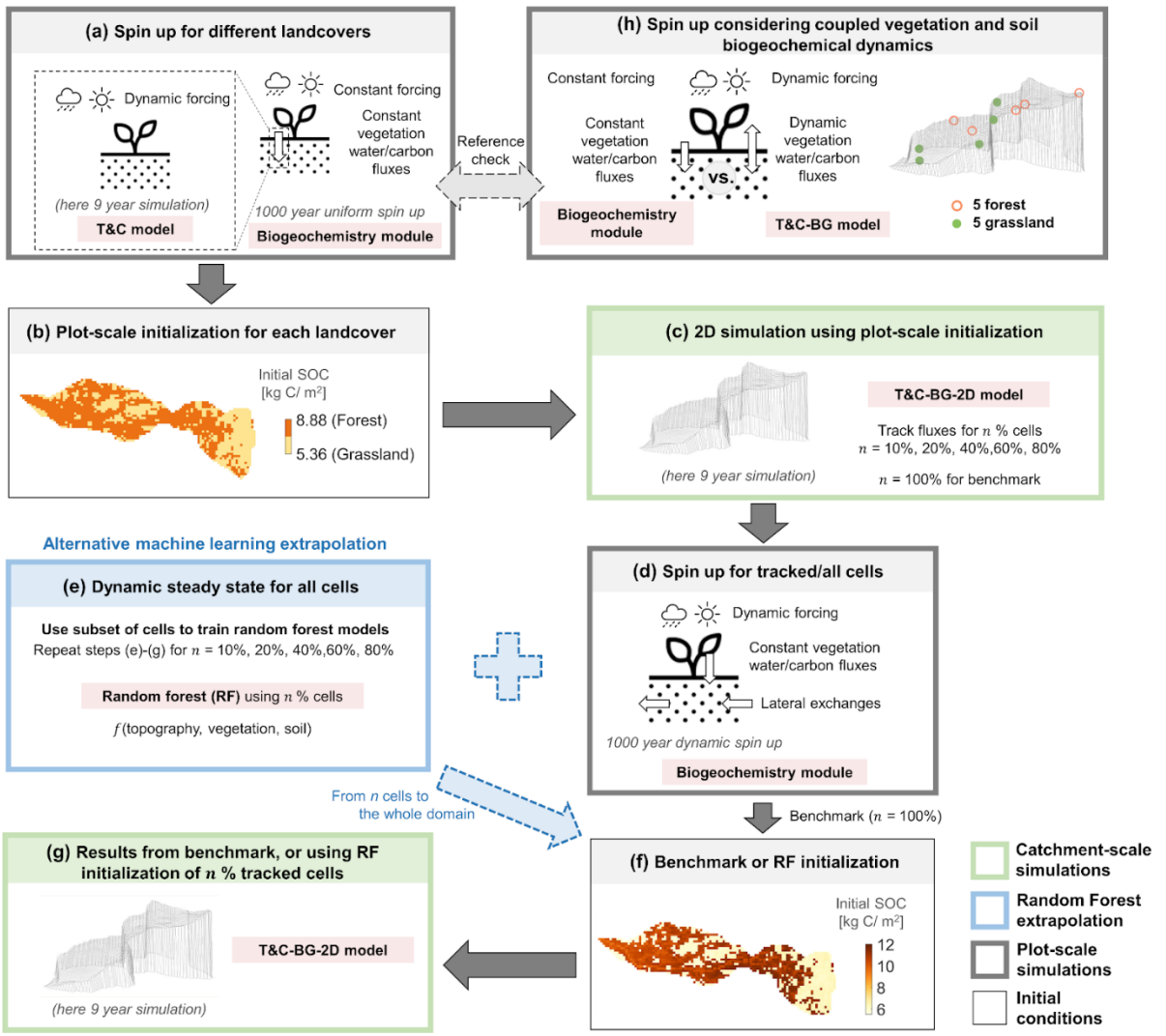

Fig. R2 (same as revised Fig. 2 in the manuscript). *Workflow of the model initialization procedure. (a) For each soil–vegetation combination, the Deleted: perform T&C model is used to obtain Deleted: to get averaged water and plant fluxes, which are used in a 1D spin-up via the T&C-BG biogeochemistry module using long-term average meteorological inputs. The resulting steady-state pools are then assigned to all cells with the same soil–vegetation combination (b). (c) A 2D simulation (T&C-BG-2D) is run with these initial conditions, tracking fluxes for all cells (benchmark) or for a given percentage n of all grid cells. (d) For these cells, the T&C-BG biogeochemistry module 1D spin-up with dynamic meteorology and imposed lateral fluxes is run to achieve a dynamic steady state in all tracked cells. (e) A random forest (RF) model is trained on results from different percentages (n) of the subset of these cells Deleted: (n) using topographic, soil, and vegetation attributes as predictors to reconstruct the initial conditions for the entire catchment domain (f), which are then used to initialize the final simulations (g). For each simulation scenario, steps (e) to (g) are repeated for different percentages of tracked cells n to investigate the optimal number of cells for the RF algorithm. (h) Additional plot-scale simulations are performed for 10 representative cells (5 forest, 5 grassland) to obtain a reference steady state using the fully coupled T&C-BG model (i.e., also considering dynamic vegetation fluxes), which is compared to the initial states generated in step (a). The solid arrows represent steps needed for the benchmark simulation, the dashed arrows represent the alternative machine learning extrapolation (blue) and the reference check (gray). See Table 1 for details on models used in the spin-up procedure and their acronyms.*

Minor comments:

1. The manuscript is generally well written, though some sections, especially the introduction and methods, are a bit dense. Editing for conciseness would improve readability.

Reply: Thank you! We revised some sentences to be more concise in the Introduction and Methods sections, some of which are reported below.

Line 32: *"Thornton and Rosenbloom (2005) introduced several acceleration techniques in the Biome-BGC model, including accelerated decomposition, nitrogen addition, and multivariate minimization Deleted:,. These approaches Deleted: which reduced spin-up time by up to 75% Deleted: in the Biome-BGC model."*

Line 45: *"In the most extreme case, Christensen et al. (2008) applied a spatially distributed spin-up Deleted: period of more than 2500 years to reach steady state Deleted: conditions for the of soil carbon and nitrogen pools. However, Deleted: the prolonged spin-up periods Deleted: required for stabilizing the biogeochemical pools can become computationally prohibitive when applied to thousands of grid cells."*

Line 57: *"These successes highlight Deleted: its strong capability their ability to extrapolate Deleted: the spatial distributions of soil carbon and nutrient pools* using Deleted: based on a suite of *environmental covariates,* based on Deleted: starting from *a limited number of locations with known steady-state conditions (Martin et al., 2014; Parvizi and Fatehi, 2025; Zeraatpisheh et al., 2022)."*

Line 59: *"In this context, if steady-state conditions are known for a subset of grid cells, RF can be used to infer the spatial distribution of biogeochemical Deleted carbon and nutrient pools across the entire catchment."*

Line 151: *"In this case, the dynamic steady-state conditions from all grid cells are Deleted: directly used to initialize the Deleted: as the initial state for a subsequent 9-year simulation (from Fig. 2d to Fig. 2f)."*

Line 165: *"The performance of different initialization schemes (i.e., without RF (n = 0%) and with n < 100%) is assessed by comparing the spatial distributions of Deleted: soil carbon and nutrient pools — specifically SOC, DOC, and SON Deleted: — with Deleted: those from the benchmark case (n = 100%)."*

2. There are quite a few acronyms and model components (like T&C, T&C-BG, T&C-BG-2D), which could confuse readers. A short table summarizing these components would be helpful.

Reply: Thank you for this comment. We have included **Table R3** as Table 1 in the manuscript and referred to it at the end of the Introduction.

Line 110: *".....shaping spatio-temporal carbon and nutrient patterns across a landscape. For clarity, Table 1 summarizes the different models referred to in this study. The proposed hybrid initialization technique is implemented in the coupled ecohydrological–biogeochemical T&C-BG-2D model."*

Table R3 (same as new Table 1 in the manuscript). *Acronyms, full names, spatial scale, and brief description with key references of the models mentioned in this study. Note that the proposed hybrid initialization technique is for the fully distributed coupled ecohydrological-soil biogeochemical T&C-BG-2D model.*

Acronyms	Model full name	Scale	Description	Key references
T&C	Tethys-Chloris	Plot and catchment scale	Mechanistic ecohydrological model that simulates coupled dynamics of energy, water and vegetation at the land surface	Fatichi et al. (2012a, b)
T&C-BG	Tethys-Chloris-Biogeochemistry	Plot scale	Extension of T&C to include modules simulating soil biogeochemistry and plant nutrient dynamics	Fatichi et al. (2019)
T&C-BG-2D	Tethys-Chloris-Biogeochemistry-2 Dimensional	Catchment scale	Extension of T&C-BG that considers lateral transport of carbon and nutrients	Lian et al. (2025)
RF	Random Forest	-	Machine learning algorithm used here to extrapolate spatially heterogeneous initial conditions from flux-tracking simulations	Breiman (2001)

3. The RF feature importance analysis is informative. A brief discussion of how predictor importance could guide targeted sampling, for example, focusing tracked cells in areas where important predictors vary most, would be helpful.

Reply: Thank you for this helpful suggestion. This comment is related to major comments 2 and 3 above, regarding the (site-specific) sampling strategy. We agree that the RF feature importance analysis can provide useful guidance for designing targeted sampling strategies. We have clarified that a proper sampling strategy is necessary when the predictor exhibits strong spatial heterogeneity. In particular, vegetation type and soil texture should be explicitly considered in the selection of tracked cells to ensure adequate coverage of the predictor space. We also included the suggested workflow for transferring the proposed technique to a new case study.

The related revisions are:

Line 161: "*Because the distributions of topographic, soil, and vegetation characteristics across the full domain are known in advance and do not exhibit pronounced extreme tails, Deleted: the selected training cells were chosen to, stratified random sampling based on vegetation types leads to training subsets that approximately preserve these distributions and thus ensure representativeness (Fig. S2). In cases where predictors exhibit strong spatial heterogeneity or patchiness, more structured hierarchical sampling across multiple predictor bins could be implemented to ensure a balanced representation of environmental gradients."*

Line 394: *".....However, simulations from the Random Soil scenario (Section 3.2) highlight the essential role of including soil information among the RF predictors, particularly in areas characterized by high soil spatial variability. Shapley values in Table S5 show that clay content is the most important predictor (after vegetation type)*

for soil carbon prediction, suggesting that clay content should be explicitly considered in generalized RF modelling frameworks."

Line 402: *".....However, this is likely due to the fact that vegetation in the Erlenbach does not experience water limitation. Conversely, in drier regions, soil texture could become a major constraint and separate spin-ups for each soil type may be required instead of using an average soil texture. Apart from soil texture, we did not include elevation-based clusters in step (a) of the initialization procedure (Fig. 2), which is potentially necessary if the catchment has significant elevation changes or spans climatic regimes where processes such as permafrost occurrence or soil freezing are relevant. Furthermore, soil–vegetation coupling is inherently site-specific, and the decoupled spin-up here relies on a 9-year average of plant and soil dynamics. While this assumption is reasonable for the Erlenbach site, it may introduce biases in more extreme ecosystems (e.g., nutrient limited environments, more variable climatic conditions), where long-term average plant-soil biogeochemical dynamics cannot be adequately captured within a 9-year period. We therefore suggest evaluating this assumption by running longer-term simulations for a small subset of grid cells and verifying that the plant–soil dynamics averaged over the chosen period do not introduce systematic biases relative to longer-term simulations. In addition to the sensitivity analysis discussed in the previous paragraphs, we recommend that, when implementing the proposed spin-up procedure in a new case study, the first step should be to assess the distribution of key environmental predictors in the study area (particularly vegetation type and soil texture) and to apply a proper sampling strategy to represent the predictor space. It is also essential to evaluate whether the ecosystem is nutrient-limited and whether the meteorological forcing is representative of the long-term climatic conditions. Based on these considerations, a tracked-cell proportion of n = 10–40% may serve as an initial guideline, with the final choice determined by the trade-off between target accuracy and available computational resources. "*

4. The discussion of uncoupled vs. fully coupled spin-up is thoughtful, but a clearer statement of when the uncoupled approach might fail would be valuable.

Reply: Thank you for this comment. In the revised manuscript, we have expanded the discussion to explicitly outline several situations where the decoupled approach may fail or introduce bias. Specifically, we now emphasize that the uncoupled spin-up may be inadequate in hydroclimatic regimes characterized by strong water or nutrient limitation, where soil–vegetation interactions are tightly coupled, and situations where the meteorological forcing period is not sufficiently representative of long-term climatic conditions. This, combined with the suggested workflow and the discussion on non-steady state conditions (last paragraph in the discussion), should now provide the reader with sufficient information to assess the applicability of the proposed framework to their case study. The related paragraph now reads:

Line 402: *".....However, this is likely due to the fact that vegetation in the Erlenbach does not experience water limitation. Conversely, in drier regions, soil texture could become a major constraint and separate spin-ups for each soil type may be required instead of using an average soil texture. Apart from soil texture, we did not include elevation-based clusters in step (a) of the initialization procedure (Fig. 2), which is potentially necessary if the catchment has significant elevation changes or it spans climatic regimes where processes such as permafrost occurrence or soil freezing are relevant. Furthermore, soil–vegetation coupling is inherently site-specific, and the decoupled spin-up here relies on a 9-year average of plant and soil dynamics. While this assumption is reasonable for the Erlenbach site, it may introduce biases in more extreme ecosystems (e.g., nutrient limited environments, more variable climatic conditions), where long-term average plant-soil biogeochemical dynamics cannot be*

adequately captured within a 9-year period. We therefore suggest evaluating this assumption by running longer-term simulations for a small subset of grid cells and verifying that the plant–soil dynamics averaged over the chosen period do not introduce systematic biases relative to longer-term simulations. In addition to the sensitivity analysis discussed in the previous paragraphs, we recommend that, when implementing the proposed spin-up procedure in a new case study, the first step should be to assess the distribution of key environmental predictors in the study area (particularly vegetation type and soil texture) and to apply a proper sampling strategy to represent the predictor space. It is also essential to evaluate whether the ecosystem is nutrient-limited and whether the meteorological forcing is representative of the long-term climatic conditions. Based on these considerations, a tracked-cell proportion of n = 10–40% may serve as an initial guideline, with the final choice determined by the trade-off between target accuracy and available computational resources. ”

References

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.