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.

Below, we summarize the key changes made to the manuscript. Line numbers refer to the **revised manuscript with tracked changes.**

- The **Abstract** and **Introduction** have been revised to improve clarity and provide more context for our work. The Abstract now more precisely quantifies the spatial patterns captured and clarifies the spatially distributed nature of the models to which the proposed framework can be applied (Lines 7-8 and 12). Additional references have been included in the Introduction (Lines 28, 55-56). The hypotheses tested in this study are now more clearly specified, and the novelty of this work is better explained (Lines 77-78, 85-88). Several sentences have also been revised throughout the Introduction to improve clarity and conciseness.
- The **presentation and structure of the methodology** have been improved. Table 1 has been added to summarize the different models referenced to in this study (Line 115), particularly those included in the workflow shown in Fig. 2, and Fig. 2 has been substantially revised to improve accessibility and clarity for readers. We also clarified assumptions related to the random soil scenario (Lines 215 and 316-319).
- The **Methods** section has been substantially revised to improve clarity, reproducibility, as well as to provide additional details on the Random Forest (RF) model component (e.g., Lines 168–178, 186–192, 199–203). Specifically, we clarified the sampling strategy, the configuration and parameters of the RF model, and added both a robustness evaluation and a SHAP-based interpretability analysis for the RF model. Additional details on the criteria used to determine steady-state conditions have been included (Lines 248-252).
- The **Results** section has been updated to include results related to the RF robustness evaluation and SHAP value analysis (Lines 294–296). Section 3.3 and Fig.6 have been revised to emphasize that the bias between plot-scale initialization and spatially informed initialization is strongly linked to grid-cell interactions. The computational cost and efficiency gains are now explicitly reported in a supplementary table and presented in the Results section (Lines 403–407).
- The **Discussion** section has been expanded to strengthen the interpretation and broader applicability of the approach. Specifically, a discussion on the RF robustness evaluation and SHAP value analysis has been added (Lines 446-450, 454-456). The Discussion has also been extended (Lines 465-480) to address site-specific sampling strategies when applying the approach to other sites, providing clear guidance on adapting predictor sets to different hydro-climatic regions, and outline recommended steps for implementing the proposed framework at new sites.

Reviewer #1

The study by Taiqi Lian et al. proposes a novel framework to reduce the computational requirements of the initialisation of a gridded eco-hydrological model with lateral transport using a mix of coupled/ uncoupled simulation and valorizing machine learning. The use of machine learning in process model spinup is a timely and important objective. Comparable approaches to my knowledge were not yet applied to process models with interdependent pixels.

However, critical gaps in the methodology, lack of demonstration of the robustness of the RF predictions, and a study design which fails to isolate the impact of the respective spinup components/assumptions on the results (failing to address hypotheses) are major shortcomings. In addition, the authors did not demonstrate that the new approach is accurate enough for typical model applications (in contrast Fig 7 seems to indicate that the underlying strategy of mixed uncoupled / coupled simulation is quite inaccurate already). As a consequence it is not possible to assess if the new approach (of combining a mix of coupled/ uncoupled simulation and valorizing machine learning) actually works sufficiently well.

> **Reply:** We thank the Reviewer for the positive feedback and for appreciating our work. We are also grateful for the detailed and constructive suggestions, which greatly helped us to improve the manuscript.
>
> In the revised version, we carefully addressed all the Reviewer's comments, as detailed in our replies below. Specifically, we have (1) clarified the criteria used to define steady-state conditions and showed that they are reached in the simulations here. (2) Besides, we performed a few additional simulations to confirm that the bias arising from the decoupled spin-up is acceptable in typical model and measurement contexts, but also better clarified the limitations of such an approach. (3) We also explicitly evaluated the robustness of random forest (RF) predictions using cross-validation and interpretability analyses, and (4) we quantified the computational savings of the proposed framework by reporting the wall-clock times for the different components of the spin-up procedure. Lastly, (5) we revised the introduction section to better state the novelty of our work.
>
> For revisions in the manuscript, we use *Deleted:XX* **in gray for deleted text**, and *text* **in blue for added text**. Line numbers refer to the **revised manuscript with tracked changes.**

Major comments:

1. The main aim of the new approach is to reach a stable state at a reduced computational time compared to conventional spinup procedures. The authors did not demonstrate (1) that steady states are reached and (2) that the new approach for spin-up leads to biases in fluxes which are acceptable for the typical application field of the model. This can be achieved by providing steady state criteria and conducting a test for a typical model application (e.g. transient response of C and fluxes under climate change). The authors should also give more information on the computational demand saving.

> **Reply:** Thank you for your comments, which we address in three parts.
>
> (1) Steady state conditions. As illustrated in Fig. 2h, two steady states are considered in the manuscript: the steady state obtained from the uncoupled spin-up (*biogeochemistry module*) and the one obtained from the fully coupled spin-up (*T&C-BG*). The uncoupled spin-up leads to a bias in SOC relative to the coupled steady state,

which is discussed in detail in point (2) below. Here we clarify that singular spectrum analysis was used to extract the long-term trend component of the time series. Steady state conditions were then defined as being reached when at least one of the following three criteria was satisfied for all soil carbon and nutrient pools: (i) the absolute value of the fitted linear trend slope (from the singular spectrum analysis) was smaller than 0.1 [gC/m^2/day or gN/m^2/day], (ii) the absolute difference between the first and last values of the fitted linear trend over the evaluation period was smaller than 0.1 [gC/m^2 or gN/m^2], or (iii) the relative change between the first and last values of the fitted linear trend was smaller than 1%. Under this condition, no systematic long-term trends are observed in any of the pools between the beginning and the end of the simulation period. This definition is consistent with the Reviewer's suggestion (minor comment 8 below) of computing linear trends over a given period and applying a threshold (e.g., 1% per year) to detect steady state conditions. In our case, a 9-year period was used in order to minimize the influence of interannual variability on soil carbon and nutrient pools.

To better illustrate this, we have now included time series of multiple soil biogeochemical pools over the final 9 years of the coupled simulation for a representative grid cell (**Fig. R1**; we added this in the supplementary information as Fig. S3). These results show that the pools exhibit only seasonal variability in response to climate forcings, without discernible long-term trends, indicating that steady state conditions are reached.

Accordingly, the manuscript has been revised as follows (from line 246):

"For every 9-year simulation, the long-term trend of all simulated soil carbon and nutrient pools was evaluated, using singular spectrum analysis to exclude interannual seasonalities. *The steady state was considered reached when* **the** *long-term* **trends** *in* Deleted: of **all soil carbon and nutrient pools** *satisfied at least one of three predefined criteria over the final 9-year simulation period: (i) an absolute linear trend (slope) smaller than 0.1 [gC/m^2/day or gN/m^2/day], (ii) an absolute difference between the first and last values of the fitted linear trend smaller than 0.1 [gC/m^2 or gN/m^2], or (iii) a relative change between the first and last values of the fitted linear trend smaller* Deleted: changed by less **than 1%** Deleted: simultaneously during a 9-year simulation *(see examples in Fig. S3)*."

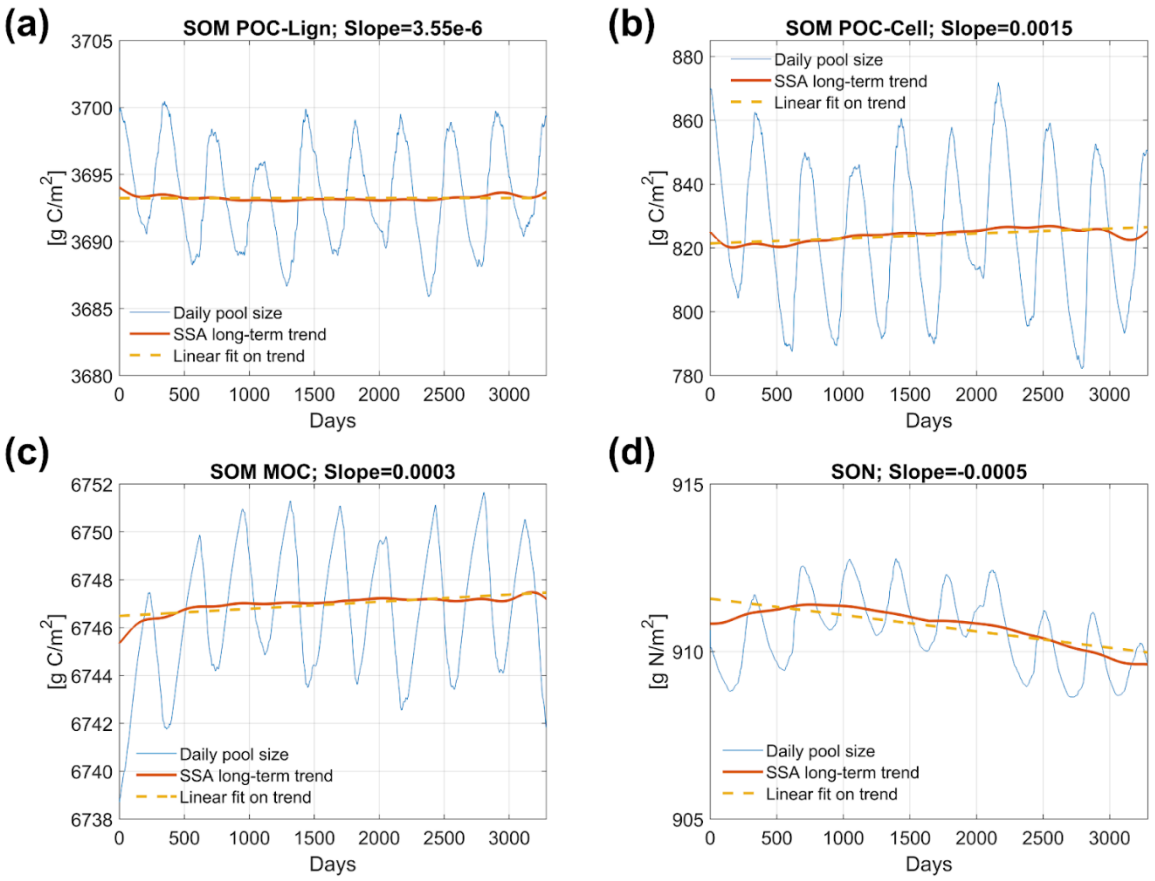

Fig. R1 (same as Fig. S3 in the supplementary information). *Time series of soil carbon and nitrogen pools during the 9-year coupled simulations for a representative grid cell. Long-term trends extracted using singular spectrum analysis (SSA), the corresponding linear fits, and the slopes of the fitted trends are shown. Carbon pools include soil organic matter particulate organic carbon (SOM–POC) associated with lignin (SOM–POC–Lign; panel a) and cellulose/hemicellulose (SOM–POC–Cell; panel b), and mineral-associated organic carbon (SOM–MOC; panel c). Nitrogen pools include nitrogen in soil organic matter (SON; panel d). All pools are considered to have reached steady-state conditions.*

(2) Potential bias introduced by the uncoupled spin-up method. As stated in the manuscript, the new spin-up approach leads to differences of 19.2% and 10.6% in soil carbon stocks for grassland and forested sites, respectively (line 389).

As suggested by the Reviewer, we conducted an additional test to evaluate whether the new spin-up strategy introduces acceptable biases in model transient fluxes. Specifically, we compared 9-year simulated gross primary productivity (GPP) and evapotranspiration from the T&C-BG model at one grassland and one forest site, using steady-state conditions obtained from the decoupled and coupled spin-up approaches as initial states. As shown in **Fig. R2**, the transient responses of carbon and water fluxes under the two initial conditions are highly similar, indicating that potential biases associated with the decoupled spin-up have a negligible impact on flux dynamics relevant for typical ecohydrological modeling applications.

We further note that the magnitude of the reported deviations falls well within commonly reported uncertainty ranges of SOC stock estimates arising from observational sources. For example, measurement-based SOC estimates can differ by 30–50% across different analytical laboratories and among samples collected at the same site (Even et al., 2025). Zhou et al. (2023) reported a normalized root mean square error of

approximately 48% for 0–30 cm SOC stocks when comparing two widely used SOC datasets (RaCA and gSSURGO). Furthermore, neglecting the vertical distribution of bulk density (as commonly done in most ecohydrological models) alone can introduce errors of about 17% in 0–30 cm SOC stock estimates (Fowler et al., 2023).

Accordingly, we revised the manuscript after line 387 as follows:

"A direct comparison of SOC between the uncoupled spin-up steady state (middle gray box in Fig. 7) and the reference steady state from the fully coupled plot-scale spin-up (red shading) reveals an average underestimation of 19.2% and 10.6% for grassland and forested sites, respectively (see also Table S3 Deleted: 2). These values are within typically reported uncertainty ranges of SOC stock estimates (e.g., Fowler et al., 2023; Zhou et al., 2023; Even et al., 2025), and the Deleted:. This underestimation is further reduced for the case of SON (see Fig. S8 Deleted: 7)."

Taken together, these comparisons indicate that the deviations introduced by the uncoupled spin-up approach are within the uncertainty envelope commonly encountered in SOC quantification, which supports the adequacy of the proposed spin-up strategy for typical model applications.

Considering this (as well as minor comment 1 and specific comment 20 below), we also extended the discussion about limitations related to site-specific applications and the bias from the decoupled spin-up process, after line 464:

"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."

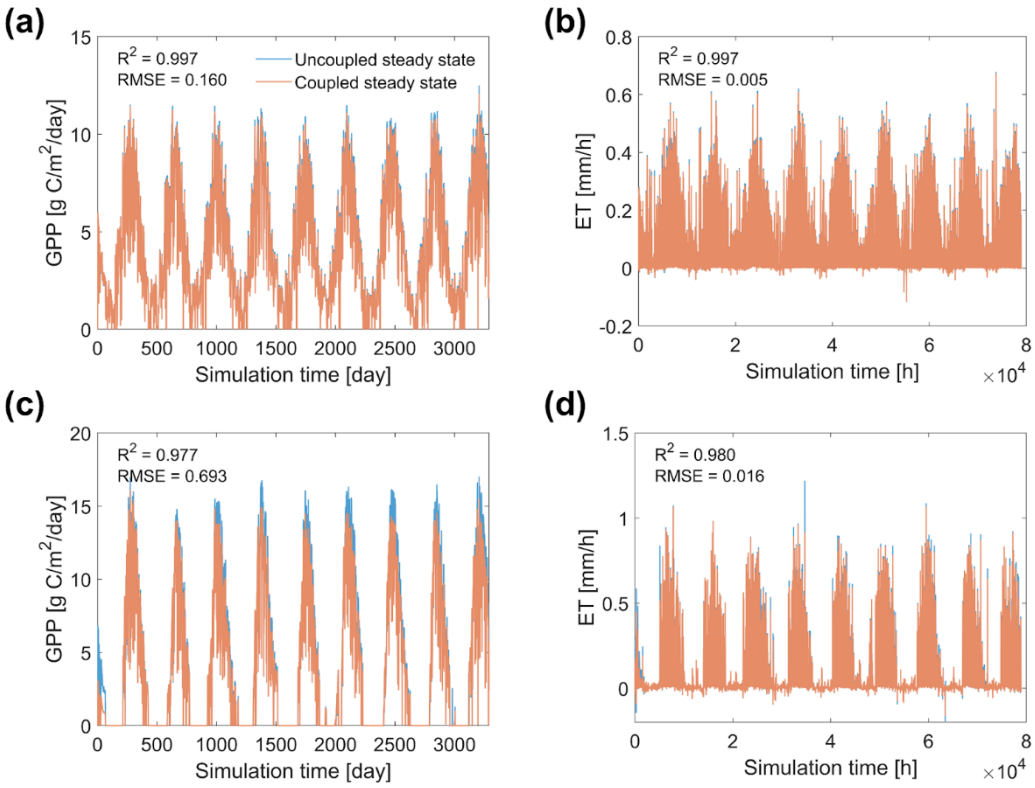

Fig. R2. *Simulated gross primary productivity (GPP) and evapotranspiration (ET) at one forest site (a, b) and one grassland site (c, d), using steady-state conditions derived from the decoupled and coupled spin-up approaches (see Fig. 7 in the main text) as initial conditions. The coefficient of determination (R^2) and root mean square error (RMSE) between the two simulations are shown in each panel. GPP is simulated at a daily time step and ET at an hourly time step. Negative ET values indicate dew formation. RMSE has the same units as the related variable, i.e. gC/m²/day for GPP and mm/h for ET.*

(3) Computational demand. Thank you for this suggestion. The computational cost of a traditional spin-up can be separated into two main components: (i) plot-scale initialization and (ii) long-term two-dimensional (2D) simulation required to reach steady state (300 years of simulations in this study). In contrast, the computational cost of the proposed hybrid spin-up approach consists of the following components: (i) plot-scale initialization, (ii) an initial 9-year 2D simulation with lateral fluxes tracked and saved, (iii) spin-up simulations for the tracked cells, (iv) RF model training, and (v) a second 9-year 2D simulation initialized using the RF-based estimates. Among these components, plot-scale initialization is identical for both approaches and represents only a small fraction of the total computational cost compared to the 2D simulations. The computational cost associated with RF training is even smaller (negligible). The dominant contribution to total runtime arises from executing the 2D model itself.

We reported the wall-clock time for different components of both the traditional and the hybrid spin-up approaches in **Table R1**. We find that the runtime of the 2D simulations increases with the fraction of tracked cells, as additional input/output (I/O) operations are required to store lateral fluxes. In contrast, the computational cost of the spin-up simulations for the tracked cells is also very small, and increases linearly with the number of tracked cells. Nevertheless, even in the hybrid approach, the overall computational cost remains dominated by the 2D simulations. Ideally, the hybrid approach reduces the required 2D simulation length from 300 years to two 9-year simulations (i.e., 18 years in total, corresponding to approximately 6% of the original simulation length). In practice, because of the heavy I/O operations especially for high

numbers of tracked cells, and the additional cost of other spin-up procedures, our recommended configuration that tracks 40% cells saves 86% computation time on our PC (Intel CPU, 40 cores, base speed 2.00 GHz, 384 GB memory). However, absolute runtimes depend on the computing platform, and we have observed higher computational saving when high-performance computing clusters are used for better performance in I/O operations. We further note that the relatively large I/O overhead observed here is likely attributable, at least in part, to the MATLAB-based implementation, as comparable models written in lower-level or more I/O-optimized languages (e.g., C/C++ or Julia) would be expected to incur lower data read/write costs.

In addition to the new table reporting wall-clock times (**Table R1**) we also revised the Results section at line 392 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 also revised lines 397-407 to better 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 R1 (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

2. The robustness of the random forest predictions is not demonstrated. The authors should provide results from the testing and validation of the RF, and deploy interpretable machine learning in order to provide evidence into their predictions by demonstrating the relationships between SOC (SON) and predictors aligned with existing evidence.

Reply: Thank you for your comment. We fully agree that demonstrating the robustness of the random forest (RF) models and the role of predictors is essential. Accordingly, we have added analyses to assess model robustness and predictor importance.

For the RF trained using 40% of the tracked cells in the original scenario, we evaluated model robustness using k-fold cross-validation (k=10), where, in each fold, 90% of the samples were used for training and the remaining 10% for validation. Model performance is 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, across the 10 folds (**Table R2**).

Besides, as suggested by the Reviewer, we now use SHAP-based interpretability analysis (Shapley values) to quantify the predictor contributions for the RF trained using 40% of the tracked cells in the original scenario and random soil scenario (**Tables R2-R3**, same as Table S2 in the manuscript).

We emphasize that, in this study, the RF model is not used to predict an unknown reference state. The steady state for 100% of grid cells is explicitly available (from simulations with 100% tracked cells, and no RF is needed) and serves as an independent benchmark. The role of the RF is solely to extrapolate steady-state information from a reduced set of tracked cells to the full domain, thereby minimizing the number of grid cells that require computationally expensive coupled spin-up simulations. Accordingly, the manuscript focuses on evaluating how well the RF-based extrapolation reproduces the spatial variability of SOC and SON relative to the known 100% steady-state reference (Fig. 3 and Table 1 in the manuscript).

We have now detailed the description of the RF model configuration and validation, and included information on its performance in the manuscript and supplementary information.

In the methods section, we have added a paragraph after line 186:

"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)."

We have clarified the purpose of our sensitivity analysis (remove one predictor at a time), and included the SHAP-base analysis in the Methods section. After line 199, the manuscript now reads as follows:

" ……how predictor choice may affect the sampling requirements (hypothesis H3). The primary purpose of this analysis was pragmatic: to evaluate which predictors provide essential information for reproducing spatial variability, and which variables could potentially be omitted when reliable spatial data are unavailable. Besides, for the original scenario and random soil scenario (see simulation scenarios in Section 2.3) with n=40% tracked cells, we further implemented a SHAP-based interpretability analysis (Shapley values) (Lundberg and Lee, 2017) to quantify the importance of each predictor in the RF model."

We included a sentence in line 294 to show the RF model robustness, now it reads:

Line 294 *".....n = 40% is sufficient to reproduce the full temporal dynamics of SOC spatial variability in the catchment here. Table S2 summarizes the robustness of the random forest models trained using 40% of the tracked cells in the Original scenario. For most carbon and nitrogen pools, MAPE values are below 10% and NRMSE values are below 0.1, indicating that the RF models exhibit satisfactory performance. These results provide direct support for……"*

We extended the paragraph about the sensitivity analysis in the discussion section (also related to minor comment 2 and specific comments 10 and 18), between lines 445 and 456:

*"A sensitivity analysis of predictor importance in the RF model indicates that the type of vegetation is the most influential variable (*Fig. S10 Deleted: 9*), followed by elevation. This statement is further confirmed by the Shapley values reported in Tables S2 and S3.* Deleted: This *These findings align* Deleted: s *with previous studies showing that vegetation strongly influences……, and that elevation is a dominant control, at least in pronounced topographies such as Erlenbach, due to the elevation-related lapse rate control on air temperature (Stähli et al., 2021; Lian et al., 2025b). 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."*

Table R2 (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%

Table R3 (same as Table S5 in the supplementary information). *Normalized mean absolute Shapley values of the seven predictors in random forest models trained using 40% of the tracked cells in the random soil scenario.*

Soil organic carbon pools	Normalized mean absolute Shapley values						
	Elevation	Slope	Flow accumulation area	Curvature	Vegetation type	Sand content	Clay content
Below-ground Litter Metabolic	8.0%	1.9%	5.5%	1.2%	54.0%	3.4%	26.1%
Below-ground Litter Structural - Cellulose/Hemicellulose	5.3%	2.7%	2.8%	0.7%	81.0%	1.4%	6.0%
Below-ground Litter Structura - Lignin	3.8%	2.4%	2.4%	0.6%	84.7%	1.0%	5.0%
SOM-POC- lignin	2.9%	2.5%	2.4%	0.6%	75.8%	0.6%	15.2%
SOM-POC -Cellulose/Hemicellulose	4.9%	4.1%	5.3%	0.7%	21.0%	1.7%	62.4%
SOM-MOC	5.2%	3.6%	4.0%	0.7%	34.4%	1.3%	50.9%
DOC - for bacteria	19.5%	1.9%	1.9%	0.9%	0.8%	28.9%	46.0%
DOC - for fungi	21.1%	2.3%	1.9%	1.0%	0.6%	28.2%	44.9%
Enzyme for decomposition of POC-Bact	6.9%	5.5%	9.5%	2.2%	25.5%	14.4%	35.9%
Enzyme for decomposition of POC-Fung	7.4%	2.9%	4.6%	1.7%	41.4%	11.3%	30.7%
Enzyme for decomposition of MOC-Bact	6.6%	6.0%	10.9%	2.1%	25.2%	14.2%	35.0%
Enzyme for decomposition of MOC-Fung	7.1%	3.3%	4.7%	1.5%	42.1%	11.4%	30.0%
Bacteria pool	3.1%	10.9%	21.6%	2.4%	45.9%	2.2%	14.0%
Fungi saprotrophic	2.9%	5.8%	11.3%	1.7%	68.1%	0.9%	9.3%
AM-Mycorrhizal - C	5.8%	1.0%	1.0%	0.6%	83.3%	1.0%	7.3%
Soil organic nitrogen pools							
Nitrogen Above-ground Litter	5.8%	1.3%	2.1%	0.5%	61.1%	1.8%	27.6%
Nitrogen Above-ground Woody	88.2%	3.4%	3.6%	0.8%	0.0%	0.9%	3.2%
Nitrogen Below-ground Litter	5.9%	2.8%	7.9%	1.4%	28.9%	4.8%	48.2%
Nitrogen SOM	9.5%	2.3%	2.1%	0.6%	32.8%	1.8%	51.0%
Nitrogen Bacteria	3.3%	10.2%	20.9%	2.7%	47.9%	2.3%	12.8%
Nitrogen Fungi	2.7%	5.9%	10.6%	1.5%	69.2%	0.9%	9.1%
AM Mycorrhizal - N	5.0%	1.1%	0.9%	0.7%	83.7%	1.3%	7.3%
Nitrogen Ione Ammonium NH4+	7.3%	1.9%	1.7%	1.5%	67.5%	2.6%	17.6%
Nitrogen Nitrate NO3-	14.0%	7.7%	7.7%	2.3%	8.2%	8.4%	51.8%
DON	10.5%	8.0%	12.1%	2.4%	31.6%	23.6%	11.8%

3. The experiment design does not allow to disentangle the effect of soil properties, climate, etc from the effect of vertical transport based on the experiments. As the deployment of RF from spinning up a model with vertical transport is the main novelty of this study I see that as a major shortcoming, and suggest an additional simulation is performed which differs from the benchmark case only from the omission of vertical transport.

Reply: Thank you for this point. We presume that the Reviewer refers to 'lateral transport' instead of 'vertical transport'. We understand the Reviewer's concern regarding the need to disentangle the effect of lateral transport from other sources of spatial variability such as variability in soil properties and climate. In the proposed framework, lateral transport is indeed an explicit process connecting grid cells, but it is not the only source of spatial heterogeneity considered during spin-up. In fact, in our simulations, heterogeneity in water and energy conditions is also introduced implicitly through topography-controlled processes, including spatially varying meteorological forcing (e.g., temperature lapse rate) and topography-induced radiation differences

(e.g., different sky views and shadowing effects). These factors jointly influence spatially heterogeneous soil moisture and energy states, plant activity, and biogeochemical dynamics, even in the absence of explicit lateral fluxes.

Some of the co-authors themselves have addressed disentangling these effects in a previous work - see Lian et al., WRR, 2025. From our previous work, we observed that land cover and soil type are generally the primary controls on soil biogeochemical patterns at the catchment scale. Topography-driven lateral transport introduces spatial heterogeneity and neglecting lateral fluxes may lead to systematic biases in simulated soil biogeochemical pools/fluxes, particularly along topographic gradients, underscoring the role of connectivity in shaping spatial patterns. The primary novelty of the study here is not to isolate individual contributions of these mechanisms, but rather to develop and evaluate a spin-up strategy that concurrently accounts for all these different sources of variability operating in a catchment (both explicit cell-to-cell connections through lateral transport and implicit spatial heterogeneity in water and energy conditions). Accordingly, our analysis focuses on comparing a spin-up approach that accounts for catchment-scale spatial variability with a conventional plot-scale spin-up that neglects such heterogeneity. While additional simulations omitting only lateral transport could further isolate individual process contributions, it would remove only part of the spatial variability considered in the model. In contrast, the scenario without tracked cells (i.e., the 0% tracking case) represents a fully plot-scale initialization and thus provides a clearer conceptual baseline for comparison. This scenario is already included and discussed in our analysis and shows a systematic underestimation in both SOC spatial variability and median values (Fig. 3, line 269), with a bias in SOC estimates up to 20–30% (line 342) and almost no spatial variability captured. Conversely, considering even only 10% of tracked cells, thus including minimal information on spatial heterogeneity and lateral transport in the model spin-up, can reproduce 75% of the SOC spatial variability (Table 1 in the original manuscript, Table 2 in the revised manuscript).

To better clarify this design choice and the scope of the proposed approach, we have revised the Introduction to more explicitly emphasize that the objective of this study is to account for catchment-scale spatial variability and cell-to-cell connectivity during the spin-up phase, rather than to decompose individual process contributions.

Line 76: *"(H1) Spatially explicit spin-up is required to capture spatial variability in soil carbon and nutrient pools* Deleted: across a landscape *at the catchment scale, where grid-cell hydrological connectivity and topography-driven heterogeneity in water and energy states concurrently act to shape soil biogeochemical spatial patterns;"*

Line 85: *"To our knowledge, this is the first study to address initialization in a catchment-scale, fully distributed coupled ecohydrological-soil biogeochemical model, which that mechanistically simulates coupled water, vegetation, and soil biogeochemical dynamics and explicitly accounts for water and solute lateral transport as well as topography-controlled variations in microclimatic conditions. The proposed approach provides……"*

4. The methodology lacks information on (1) the calculator of the computational time savings from the new approach, (2) more information on the random forest (training/ validation results, treatment of categorical variables, pixel selection, etc), (3) steady state criteria.

Reply: Thank you for this comment. In response to the Reviewer's concerns regarding the methodological description, the revised manuscript now provides detailed information on: (1) the calculation of computational time savings of the proposed

approach, (2) the training and validation of the random forest models, including cross-validation results and the treatment of predictor variables, and (3) the criteria used to determine steady-state conditions. These points are addressed in detail in our responses to the previous comments and corresponding revisions in the manuscript, i.e., (1) see reply to major comment 1, point 3, (2) see reply to major comment 2, and (3) see reply to major comment 1, point 1.

In addition to the previous points, we further noticed that the pixel selection strategy for RF model training was insufficiently described in the original manuscript. We have therefore added an explicit description of the pixel selection algorithm in the Methods section. In brief, pixels are selected using a stratified random sampling approach based on vegetation classes to ensure a reproducible and balanced representation of vegetation types across sampling fractions. This is now described in the Methods section (line 167), which reads as follows:

"*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."*

5. Novelty needs to be more clearly defined. The approach of using a RF for spinup of biogeochemical cycles in a land surface model has been proposed, applied and tested before in Sun, Yan, et al. "Machine learning for accelerating process-based computation of land biogeochemical cycles." Global Change Biology 29.11 (2023): 3221-3234.

Reply: Thank you for this suggestion. We agree that the novelty of the study should be more explicitly defined in relation to previous work applying machine learning to accelerate land biogeochemical model spin-up, including the work of Sun et al. (2023), which is now explicitly mentioned in our manuscript.

While Sun et al. (2023) addressed spin-up challenges in grid-cell–based land surface modeling frameworks, typically applied at regional to global scales (typical spatial resolution ranging from 0.5° to 2° in latitude-longitude grids) without accounting for lateral transfer fluxes, the novelty of our work lies in extending the RF-based spin-up concept to fully spatially distributed, catchment-scale ecohydrological-soil biogeochemical models. Here, spatial heterogeneity emerges not only from variability in soil and vegetation features, but also from topography-driven microclimate variability (e.g., lapse-rate-controlled temperature gradients and terrain-induced radiation shading) and explicit lateral water and solute transport between cells.

The revised Introduction now explicitly highlights this distinction at several points:

Line 28: *"In contrast, soil biogeochemical models often require much longer spin-up periods, sometimes up to thousands of years (Randerson et al., 2009; Qu et al., 2018; Sun et al., 2020),....."*.

Line 55-56: *"Recent advances in artificial intelligence and machine learning (ML) have provided new opportunities for environmental modeling, including applications in hydrology, ecology, and soil science (e.g., Peters et al., 2007; Mahdavi et al., 2018; Tyralis et al., 2019; Carranza et al., 2021; Gupta et al., 2021; Sun et al., 2023; Pawusch et al., 2025)"*

We further clarified in the text that that the novelty of this study lies in addressing model initialization in a spatially distributed ecohydrological–soil biogeochemical model with explicit lateral connectivity among grid cells:

Line 85: *"To our knowledge, this is the first study to address initialization in a catchment-scale, fully distributed coupled ecohydrological-soil biogeochemical model, which that mechanistically simulates coupled water, vegetation, and soil biogeochemical dynamics and explicitly accounts for water and solute lateral transport as well as topography-controlled variations in microclimatic conditions. The proposed approach provides……"*

Minor comments:

1. Two plot-scale simulations were performed for the vegetation-soil combinations. It is unclear how they can capture the variation in SOC from climate. Fig 2 b) suggests that only two combinations were performed for a single location. This neither variation in climate (e.g. temperature/elevation), nor soil texture are accounted for therefore additional sensitivity simulations are performed aiming to disentangle the respective effects. This is an approximation as drivers interact, the discussion falls short to fully reflect this,

Reply: Thank you for this comment. We fully agree on this point regarding the simplification of vegetation-soil combinations. In the current study, two plot-scale simulations (for the vegetation-soil combinations available at Erlenbach) were performed. This design does not explicitly span climatic variability (e.g., temperature/elevation gradients) or strong contrasts in soil texture. We emphasize that this simplification is motivated by the characteristics of the Erlenbach catchment, where vegetation cover is relatively simple and soil texture (based on currently available information) is comparatively uniform at the scale relevant for this application. More generally, for larger or more heterogeneous catchments, plot-scale initialization would ideally be performed at multiple representative locations that capture major gradients in climate and soil properties. These location-specific steady-state pools could then be used as initial conditions across the corresponding subregions of the domain. However, this is often not possible due to a lack of information in most applications at the small to medium catchment scales.

Although the plot-scale initialization (step a in Fig. 2) is simplified, the subsequent spatially distributed simulations (step c in Fig. 2) explicitly account for spatial variability in climatic drivers and soil texture combinations across the catchment. Therefore, the effects of climate and soil heterogeneity in soil biogeochemical pools is still considered during the spin-up (Fig. 2d) and is reflected in the resulting steady-state patterns (Fig. 2f).

Combined with major comment 1 and specific comment 20, we have extended the discussion (after line 464) about the site-specific limitations of our application, and provided clear working steps for implementing the proposed spin-up procedure in a new catchment.

"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."

Also, in the work here, we considered different scenarios as a first test to evaluate how, in this catchment configuration, different soil distributions, vegetation, and topography could affect the spin-up procedure. A comprehensive evaluation across a broad range of catchment types or biomes would be required to fully disentangle these effects. Such an analysis, however, is beyond the scope of the present study and is left for future work. We have therefore expanded the discussion to explicitly acknowledge this limitation and to clarify the site-specific nature of the presented results. In addition, SHAP-based analyses were introduced to improve the interpretability of the random forest models and to quantify predictor contributions within the RF extrapolation framework, rather than to replace process-based sensitivity analyses. We have revised the manuscript accordingly (see also our response to major comment 1 and 2).

Between line 445 and line 456:

*"A sensitivity analysis of predictor importance in the RF model indicates that the type of vegetation is the most influential variable (*Fig. S10 Deleted: 9*), followed by elevation. This statement is further confirmed by the Shapley values reported in Tables S2 and S3.* Deleted: This *These findings align* Deleted: s *with previous studies showing that vegetation strongly influences……, and that elevation is a dominant control, at least in pronounced topographies such as Erlenbach, due to the elevation-related lapse rate control on air temperature (Stähli et al., 2021; Lian et al., 2025b). 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."*

2. The conservation of mass is of critical importance in the field of hydrological and biogeochemical modelling. I didn't find any statement concerning mass conservation in the literature on T&C-BG. The authors should indicate the basic principles underlying the model.

Reply: Thank you for this comment - we agree that ensuring mass conservation is a fundamental requirement in hydrological and biogeochemical modelling applications. Tethys–Chloris (T&C) and its biogeochemical extension (T&C-BG) are process-based models formulated using mass-balance principles. For each grid cell, the temporal evolution of water, carbon, and nutrient pools is computed as the balance between incoming and outgoing fluxes. Vertical and lateral water fluxes, plant carbon assimilation and allocation, soil organic matter turnover, and biogeochemical transformations are all represented explicitly through conservation equations. For the spatially distributed version, lateral water and solute transport between grid cells is treated conservatively, such that fluxes leaving one cell enter neighboring cells or the channel network. Details for the mass balance can be found in cited early works on the model (e.g., Fatichi et al., 2012). As such, mass conservation is always checked in model simulations - an example is shown in **Fig. R3**, presenting the mass balance of water (left) and a carbon pool (right) for a 9-year 2D simulation without applying any spin-up.

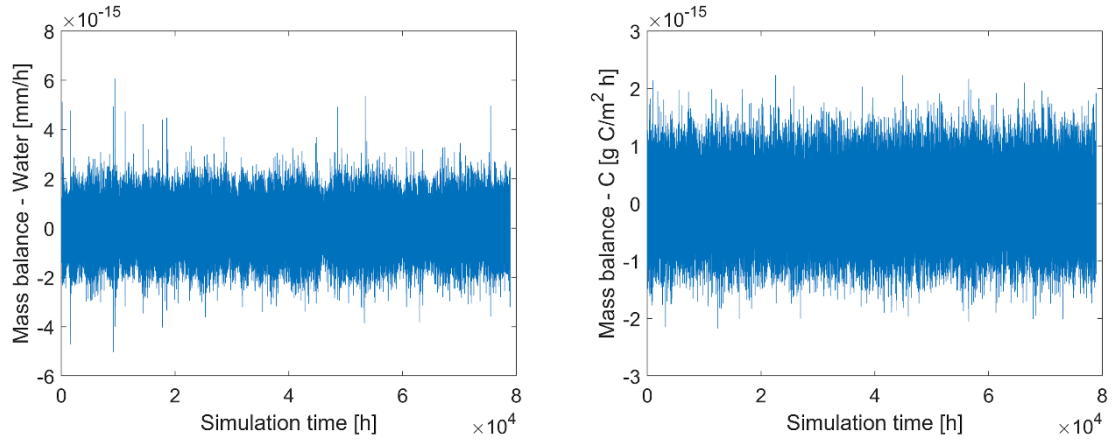

Fig. R3. Examples of water and carbon mass balance from a 9-years 2D simulation (without spin-up).

3. The steady state criterion of 'trends of [...] pools changed by less than 1%' (Line 218) makes no sense to me. Did you mean pools changed by less than 1%? Steady state is not reached when the trends are stabilised (changes less than 1%) but when they are negligible. Usually we compute the linear trend over the given period and use a threshold of eg 1% per year to detect a steady state.

Reply: Thank you for this comment. This point is addressed in detail in our response to the first major comment. Briefly, the steady-state criterion is based on the magnitude of the long-term linear trend of each pool, evaluated over the 9-year period, rather than on short-term dynamic changes.

4. The investigation of the number of pixels required for RF training (section 3.2) is not very informative. It is not clear how pixels were selected nor what motivates this analysis of the different scenarios. Approaches like k-means clustering are available in order to guide the sampling.

Reply: We agree that the motivation and criteria for the choice of pixels for RF training should be better clarified.

The objective of investigating different fractions of tracked pixels (10%, 20%, 40%, 60%, and 80%) is to evaluate the performance of the RF-based extrapolation and how the resulting steady-state biogeochemical pools reproduce the benchmark results across applications with different levels of complexity. This analysis therefore aims to determine a minimum fraction of tracked pixels that provides an acceptable approximation of the benchmark case (i.e., reproduces more than 90% of the SOC/SON spatial variability), which is central to the goal of reducing computational cost while maintaining accuracy.

The purpose of designing different scenarios was to test whether the required number of tracked cells and predictors changed considerably under different levels of spatial complexities. We clarified this point in the revised manuscript (line 205) which now reads:

"In addition to the original simulation scenario (hereafter Original), a set of modified simulation setups was designed to evaluate the generality of the proposed initialization method, Deleted:and to *assess its performance, and evaluate variations in the required number of tracked cells and predictors under different* Deleted: varying *levels of landscape complexity."*

Regarding pixel selection, as clarified in the reply to major comment 4, pixels are selected using a stratified random sampling approach based on vegetation classes. We acknowledge that other sampling strategies, such as clustering-based approaches (e.g., k-means), can be used to explicitly structure sampling in the predictor space. In this study, given that predictor distributions are known and that the focus is on robustness across different sampling fractions rather than on optimal sampling design, we apply the transparent and reproducible stratified random approach.

5. The labelling and the design of the different catchment scenarios can be improved, e.g. the random soil texture case is not only random but spans a much wider predictor space.

Reply: Thank you for this comment. The Reviewer is correct. We have revised the scenario labels and descriptions accordingly to better reflect the wider range of predictor space compared to the reference case. Specifically, the full name of the scenario is now "randomized and extended soil texture distribution". However, for conciseness throughout the text, we still refer to it as "Random Soil".

Line 203: *"Soil texture (Fig. 1c): the original soil texture map is replaced either with a randomized and extended spatial distribution (Random Soil, gray points in Fig. 1c) or with a uniform soil texture across the domain (Homog. Soil, red point in Fig. 1c). Note that the Random Soil scenario introduces a much wider soil texture predictor space."*

Specific comments

1. Line 7-8: provide quantification of the degree to which variability and pattern are captured

Reply: Thank you for your suggestion. We revised the sentence which now reads: "*Applied to T&C-BG-2D, a fully coupled distributed ecohydrological-soil biogeochemical model, the scheme reconstructs over 90% of the spatial variability in* Deleted: of *soil carbon and nutrient patterns in terms of probability distribution similarity while reducing……*".

2. Line 11: 'easily applied' this is too vague. Better would be to list which conditions need to be met (e.g model characteristics) in order for the spinup approach to be applicable.

Reply: We revised the sentence to replace the term "easily applied" with an explicit mention of the fact that the proposed approach is intended for spatially distributed ecohydrological-soil biogeochemical models. The sentence now reads:

"*The framework developed here can be* Deleted: easily *applied to other spatially distributed models that explicitly account for lateral transfer fluxes* Deleted: and across diverse catchments*, enabling large-scale distributed ecohydrological-biogeochemical model initializations under constrained computational budgets*."

In the discussion (lines 427-431), we also listed a few models that could potentially employ our spin-up technique, such as RHESSys (Tague and Band, 2004) and Flux-PIHM-BGC (Shi et al., 2018).

3. Line 52-54: there is at least one study which actually deployed RF for model spinup in this context which could be added: Sun, Yan, et al. "Machine learning for accelerating process-based computation of land biogeochemical cycles." Global Change Biology 29.11 (2023): 3221-3234.

Reply: Thank you for pointing us to this relevant work, which is now cited in the revised manuscript along with other relevant papers (line 28, lines 55-56).

4. Line 74: The H1 is trivial, not sure it is needed.

Reply: Thank you for this comment. We agree that, conceptually, H1 may appear straightforward. However, while spatially explicit processes are generally expected to influence soil carbon and nutrient patterns, to our knowledge no previous study has explicitly examined this requirement in the context of model spin-up at the catchment scale, where spatial variability arises not only from heterogeneous inputs but also from topography-controlled microclimate and explicit lateral transport between grid cells. Therefore, we believe it remains important to explicitly state and test this hypothesis in the context of catchment-scale, spatially distributed ecohydrological-soil biogeochemical model spinup.

To clarify the focus of this research, we revised hypothesis 1 as:

Line 76 *"H1: Spatially explicit spin-up is required to capture spatial variability in soil carbon and nutrient pools* Deleted: across a landscape *at the catchment scale, where grid-cell hydrological connectivity and topography-driven heterogeneity in water and energy states concurrently act to shape soil biogeochemical spatial patterns;"*

5. Line 81: 'this is the first study to address initialization in a fully distributed coupled ecohydrological-soil biogeochemical model that mechanistically simulates coupled water, vegetation, and soil biogeochemical dynamics' that sentence does not make it clear whether the model's pixels are interdependent. Thus the claim is invalid (see Sun et al 2023).

Reply: Thank you for pointing this out. In this study, the term "fully distributed" refers to a modelling framework in which grid cells are explicitly interdependent through lateral transport, and where spatial heterogeneity emerges dynamically from topography-controlled fluxes and microclimate variability. This goes beyond grid-based land

surface models that neglect lateral transfers. We revised this sentence to clarify the novelty of this study line 85:

"To our knowledge, this is the first study to address initialization in a catchment-scale, fully distributed coupled ecohydrological-soil biogeochemical model, which that mechanistically simulates coupled water, vegetation, and soil biogeochemical dynamics and explicitly accounts for water and solute lateral transport as well as topography-controlled variations in microclimatic conditions."

6. Line 100: what about soil P and K: How do they affect / are affected by vegetation? E.g. Does soil fertility control plant growth ?

Reply: Thank you for this comment. In T&C-BG, vegetation growth is indeed influenced by soil fertility, including nitrogen (N), phosphorus (P), and potassium (K). Changes in plant nutrient content depend on the dynamics of carbon pools (e.g., leaves, living sapwood, fine roots, carbohydrate reserves, flower and fruits, and heartwood). For each plant carbon pool, a target stoichiometry is defined to describe the quantity of nutrients required for a given amount of carbon, and this stoichiometry is flexible and responds to nutrient availability. Nutrients can also be temporarily stored in plant reserves. This flexibility is represented through a two-step scheme: nutrient reserves can first buffer imbalances between uptake and demand without altering tissue concentrations and, if reserve limits are exceeded, nutrient concentrations in nonstructural tissues adjust accordingly. Under nutrient limitation, insufficient nutrient availability may prevent the construction of plant tissues, thereby constraining plant growth. More details of the plant nutrient dynamics can be found in Fatichi et al. (2019).

7. Line 131 continued: Soil texture variation was not accounted for as you state later (L 137/138). I would suggest revising this section to be frank about this from the start.

Reply: Thank you for your helpful suggestion. We moved the sentence in lines 144-146 to line 139, to clearly state that the soil texture variation is not included in the soil–vegetation combination plot-scale simulation. Now the paragraph reads:

"The scheme follows several steps, as shown in Fig. 2. First, a plot-scale simulation using T&C model is run for each soil–vegetation combination to obtain climatologically driven, nutrient-unstressed water and vegetation fluxes (dashed box in Fig. 2a). Ideally, one should perform this step for each major vegetation and soil type combination. For the application here, given the limited variability in soil texture across the catchment, an averaged soil texture was considered. Then, we run the T&C-BG biogeochemistry module……".

8. Line 144: 'may not fully equilibrate' this is misleading. Given the slow turnover of soil organic matter it's certain they won't fully equilibrated within 9 years.

Reply: Thank you for your suggestion. Indeed it is highly unlikely that the soil biogeochemistry pool can reach a steady state in 9 years. We revised line 152 as:

"……,although we should note that soil carbon and nutrient pools respond slowly to spatial gradients and are unlikely to Deleted: may not fully equilibrate within a short simulation period"

9. Line 145 continued : 'In the 2D simulation in Fig. 2c, ideally, the entire available forcing period should be used' which is what you did as you wrote earlier. There is no need to demonstrate that one can do worse than one did.

Reply: Thank you for raising this point. We agree and have revised the sentence by removing the wording "ideally" to reflect the fact that the full available forcing period should be used. Now it reads:

"In the 2D simulation in Fig. 2c, Deleted: ideally, the entire available forcing period should be used to capture representative vegetation dynamics and to track biogeochemical fluxes as accurately as possible"

10. Lines 160: how were categorical variables treated in the RF ? Isn't clay content besides sand accounted for?

Reply: Thank you for this comment. In all the scenarios excluding the randomized and extended soil texture range scenario (Random soil scenario), only sand content was used as a predictor. This choice reflects the relatively homogeneous soil texture in the Erlenbach catchment, where preliminary analyses indicated a limited explanatory power of soil texture variability compared to topographic and hydrological predictors. Consequently, clay content was not included in the initial RF configuration.

In contrast, for the random soil scenario, both sand and clay contents were included as predictors. In this scenario, we found that excluding clay content led to a noticeable decline of the RF performance, whereas including both sand and clay substantially improved the prediction accuracy (Table 1, values in parenthesis). This result highlights the increased importance of soil texture information when it is more spatially heterogeneous. In the manuscript, we also highlighted the importance of including clay content in the RF model in the Discussion section, lines 445-456:

*"A sensitivity analysis of predictor importance in the RF model indicates that the type of vegetation is the most influential variable (*Fig. S10 Deleted: 9*), followed by elevation. This statement is further confirmed by the Shapley values reported in Tables S2 and S3. Deleted: This These findings align Deleted: s with previous studies showing that vegetation strongly influences……, and that elevation is a dominant control, at least in pronounced topographies such as Erlenbach, due to the elevation-related lapse rate control on air temperature (Stähli et al., 2021; Lian et al., 2025b). 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."*

We have revised the Methods section (lines 167-169) to explicitly mention that sand was used in the original scenario and clay was added in the random soil scenario. Now the paragraph reads:

"The covariates used as predictors include topographic features (elevation, slope, drainage area, and profile curvature), soil properties (Deleted: percentage of sand content; clay content is additionally included in the random soil texture scenario described in the following section), and vegetation type (forest or grassland)."

11. Lines 161: be more clear on the approach of selecting representative pixels. Did you use a formalized approach like k-mean clustering?

Reply: Thank you for your constructive comment. The pixel selection strategy is clarified in detail in our response to major comment 4 and minor comment 9, where we elaborate on the stratified random sampling approach used for selecting representative pixels.

12. Line 171: The sequence of removing predictors matters. There are formalized approaches to account for this like recursive feature elimination and predictor importance ranking. I would suggest using partial dependence plots or Shapley values in order to investigate the relationship between pools and predictors. RF is prone to overparameterization and assessment of these relationships can help to provide trust into the black box RF.

Reply: Thank you for this constructive comment. In the original manuscript, our sensitivity analysis consisted of removing one predictor at a time. The primary purpose of this analysis was pragmatic: to evaluate which predictors provide essential information for reproducing spatial variability, and which variables could potentially be omitted when reliable spatial data are unavailable. Because predictors were removed independently and not sequentially, the analysis does not rely on or assume any specific order of predictor removal.

Following the Reviewer's suggestion, we have extended this analysis by incorporating Shapley values to quantify predictor contributions in a formal way. This additional analysis increases the interpretability and transparency of the RF-based extrapolation. This comment is closely related to major comment 2 and minor comment 6, where we present the Shapley analysis and the corresponding manuscript revisions, and where we also address the robustness of the RF through k-fold cross-validation. To avoid repetition, we refer the reader to those responses for further details.

13. Figure 2: is not very clear.fd (g and f) indicates RF was used but there is a pathway from (d) to (f) so there is no RF involved. Do not use 'model' without specifying which model is meant.

Reply: Thank you for these useful suggestions. Reviewer #2 also provided some suggestions to improve the clarity of Fig. 2. We have now revised Fig. 2 (same as **Fig. R4** below) to distinguish the pathway with and without RF and specify the term 'model' in panel (e). We also included **Table R4** (same as Table 1 in the revised manuscript) to clarify the acronyms of the models used in the spin-up framework.

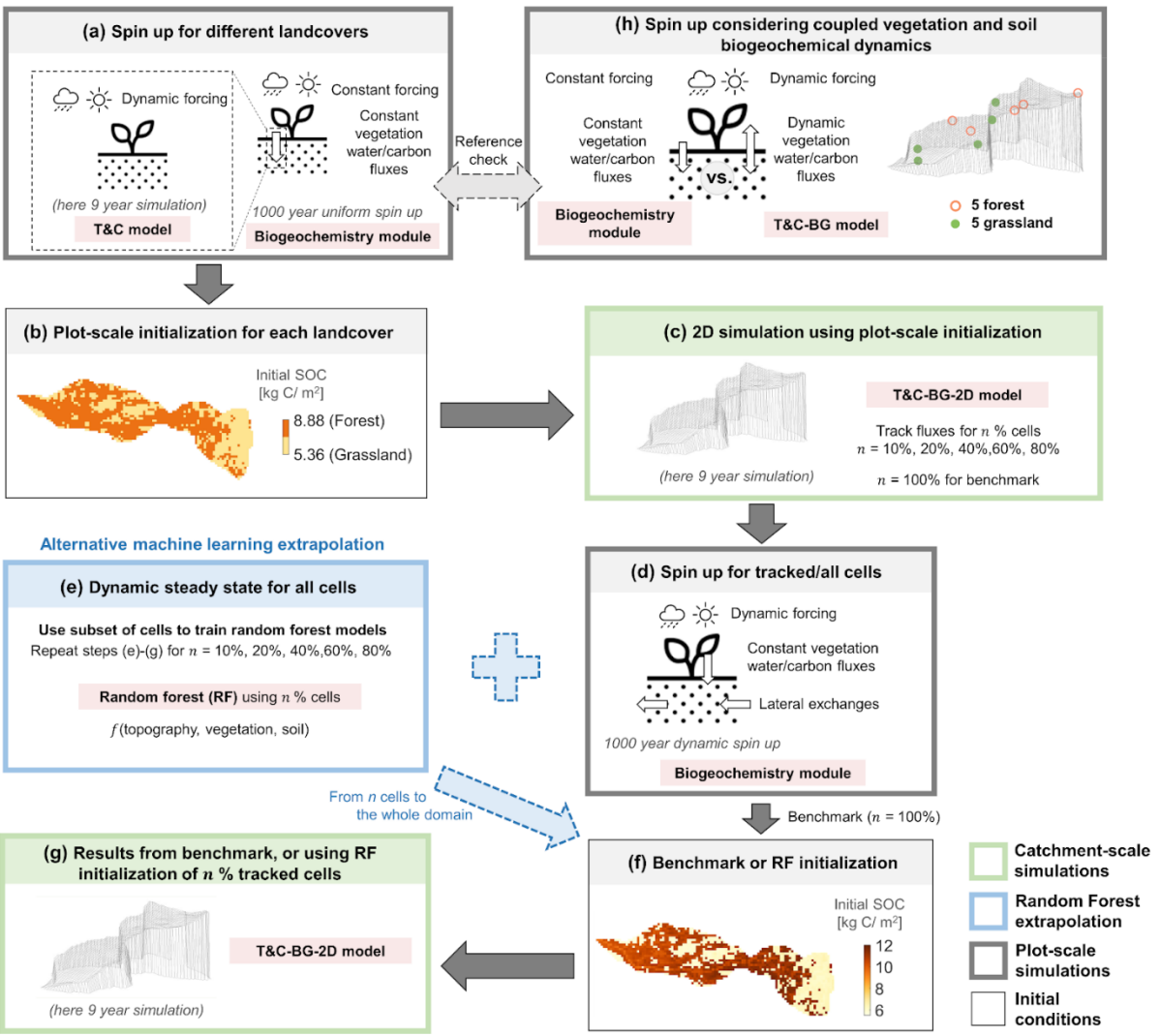

Fig. R4 (same as revised Fig. 2 in the manuscript). *Workflow of the model initialization procedure. (a) For each soil–vegetation combination, the Deleted: perform T&C model is used to obtain Deleted: to get averaged water and plant fluxes, which are used in a 1D spin-up via the T&C-BG biogeochemistry 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.*

Table R4 (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)

14. Line 200: specify the criteria for 'satisfactory performance' (e.g. % variance explained)

Reply: We have now specified the meaning of satisfactory performance and revised the sentence as follows:

"For the soil texture and topography scenarios, results for n equal to 60% and 80% are omitted, as a satisfactory performance (i.e., more than 90% of SOC and SON spatial variability captured) was already achieved with n = 40% of tracked cells ."

15. Line 211: the decoupling approach of Krinner et al 2025 was developed for model components which are coupled one-way, i.e. there is no feedback of the state of component 2 (soil) on the state component 1 (vegetation). Is this the case in T&C-BG which has nutrient cycles (and thus soil fertility affects vegetation processes which in return affect soil fertility)?

Reply: Thank you for your comment. In T&C-BG the vegetation and soil components are two-way coupled. Soil biogeochemical processes influence vegetation growth through nutrient availability (as detailed in our previous reply to specific comment 6), while vegetation dynamics affect soil carbon and nutrient pools through litter inputs, root turnover, and plant exports to belowground compartments. As described in Section 2.4, the decoupled spin-up is intentionally an approximation to save computational time by avoiding simulating long-term coupled plant-soil dynamics. We evaluated this approximation by comparing the resulting decoupled steady state with the fully coupled steady state (Section 3.4), and showed that the resulting bias remains within an acceptable range for typical applications (details are provided in our reply to major comment 1).

16. Figure 4: It is not clear why time series (of coefficient of variances) are analyzed . Does one expect large variations in SOC over the course of 9 years? Why not aggregate over time?

Reply: Thank you for this comment. We emphasize that the coefficient of variation (CV) analysed in Fig. 4 refers to the spatial coefficient of variation of SOC across the catchment at each time step, rather than to temporal variations of SOC. The aim of this analysis is to examine how spatial variability develops and stabilizes during the spin-up process. In simulations without RF initialization, the spatial SOC distribution starts from a homogeneous state (CV=0) and increases over time as spatial patterns gradually emerge. In contrast, simulations initialized using the RF-based approach exhibit a relatively stable spatial CV from the beginning, with only short-term fluctuations reflecting climate-driven responses rather than a systematic temporal trend (similar to the benchmark). The time series of the spatial CV is used as a diagnostic of spatial pattern stabilization rather than of temporal SOC variability. To avoid misinterpretations, we revised the caption of Fig. 4, to clarify this as follows:

"Figure 4. Temporal evolution of the coefficient of variation (CV) of spatial soil organic carbon (SOC) distribution over a 9-year simulation period……"

17. Figure 4: The discussion of the underlying reason for the much higher CV of random texture (but also much wider parameter space) compared to other experiments (and the implications) is not very clear.

Reply: The Reviewer is correct that the wider parameter space leads to a higher magnitude of CV in the random soil scenario in Fig.4. Specifically, this effect is largely driven by the wider range of clay and silt contents across the study area, as carbon associated with these fine particles contributes to a large proportion of SOC.

We expanded our discussion of the random soil scenario in the sentences between lines 313-321, which now read:

"However, in the Random Soil scenario, where soil texture is highly heterogeneous, adding supplementary predictors such as the percentage of clay significantly improves the model's ability to reproduce benchmark SOC spatial patterns (Table 1). This highlights the importance of including sufficient soil information when simulating spatial variability under heterogeneous soil conditions. Notably, the spatial CV is also higher in this scenario relative to the others, reflecting higher spatial variability of SOC. This is due to the expanded soil texture parameter space, especially variations in silt and clay contents, as this fine-particle-associated carbon contributes substantially to SOC (Six, et al., 2002; Feng et al., 2013; Matus 2021) (see also Table S5). These findings are consistent with previous studies showing that soil properties are among the most influential predictors in digital Deleted: soil mapping of SOC (e.g., Meersmans et al., 2008; Zeraatpisheh et al., 2022; Martín-López et al., 2023)."

18. Line 278: This is speculation. To provide evidence that your RF captures the relationship between SOC and texture you should show their relationship using interpretable machine learning techniques like partial dependence plots.

Reply: We agree that an improvement in RF performance alone does not fully demonstrate that the learned relationship between SOC and soil texture is physically meaningful. Motivated by the Reviewer's comment, we therefore complemented the analysis with Shapley values for the random soil scenario, where clay content is included as one of the predictors (see **Table R3**). This analysis confirms that soil texture

variables play a substantial role (the most important after vegetation type) in explaining SOC spatial variability, supporting the interpretation of the RF results.

We revised the sentences between lines 313 and 321 (as shown above in the reply to specific comment 17), and we extended the paragraph between lines 445 and 456 as follows:

*"A sensitivity analysis of predictor importance in the RF model indicates that the type of vegetation is the most influential variable (*Fig. S10 Deleted: 9*), followed by elevation. This statement is further confirmed by the Shapley values reported in Tables S2 and S3.* Deleted: This *These findings align* Deleted: s *with previous studies showing that vegetation strongly influences……, and that elevation is a dominant control, at least in pronounced topographies such as Erlenbach, due to the elevation-related lapse rate control on air temperature (Stähli et al., 2021; Lian et al., 2025b). 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."*

19. Line 300-320, Figure 5F: Is it relevant to analyse the bias in the 0% simulations? I would assume it is common sense in the spatial modelling community (which is the prime audience for this article). I would prefer to see more discussion about the effect of the number of training data on the relationships in figure 5&6.

Reply: Thank you for this helpful comment. We agree that a systematic bias in the 0% simulation may appear intuitive to the spatial modelling community. We thought that the analysis of the no cell track case (n=0%) could be informative for clarifying the physical origin of these biases in a fully distributed ecohydrological model.

In particular, the analysis of the 0% scenario in Fig. 5 is not intended to demonstrate that the approach fails when no spatial information is used, but rather to diagnose how biases emerge in the absence of spatially explicit spin-up. We show that SOC is systematically underestimated in the n=0% simulation, and that this underestimation decreases with increasing topographic wetness index (TWI). This behaviour reflects the role of topography-controlled lateral water and carbon redistribution in shaping SOC patterns. Examining the bias along a topographic gradient therefore provides mechanistic insight that may be relevant for other distributed modelling studies.

Moreover, a key characteristic of our spin-up framework is the explicit, topography-driven connectivity between grid cells. As shown in Fig. 6 c and d (same as **Fig. R5** below), even the n=0% simulation reproduces the qualitative relationships between fast fluxes and topography (e.g., soil respiration). In contrast, spatial patterns in slow changing pools (such as SOC) cannot be captured without a spatially explicit initialization (Fig. 6 a and b). This contrast highlights the different sensitivities of different fluxes and pools to the spin-up strategy.

We fully agree that a more in-depth discussion on the choice of the number of training data is relevant here. Following the Reviewer's suggestion, we included results for the 10% and 40% RF-based initialization simulations in Fig. 6. These additional results demonstrate that tracking a limited fraction of grid cells is sufficient to recover the benchmark SOC and soil respiration variations with TWI, together with the increasingly better performance for higher track numbers. We have revised the discussion as follows to better elaborate on these points (line 354):

"However, the benchmark simulation shows a decreasing SOC trend with increasing TWI for TWI < 11, reflecting a long-term steady-state pattern shaped by soil respiration processes during spin-up. Simulations with n = 10% and n = 40% successfully capture the SOC–TWI relationship, with results from n = 40% being closer to the benchmark simulation. This relationship is consistent with previous findings…….However, under excessively wet conditions (such as, TWI > 11), this relationship reverses due to oxygen limitation (Moyano et al., 2013; Ruiz et al., 2015). Fig. 6c,d further shows that while the n=0% case reproduces the correct qualitative relationship between fast carbon fluxes (e.g., soil respiration) and TWI, correct flux magnitudes are only captured when n = 10% of the cells are tracked, and the performance is further improved for n = 40%. Because the SOC distribution…….Using only n= 10% of the cells for RF training already eliminates most of the systematic bias (Fig. 5b), while n= 20 (Fig. 5c) and n= 40 (Fig. 5d) eliminates it almost entirely. The analysis here clearly highlights that the bias between plot-scale initialization and spatially-informed ones is strongly linked to grid-cell interactions via topography-controlled lateral transport."

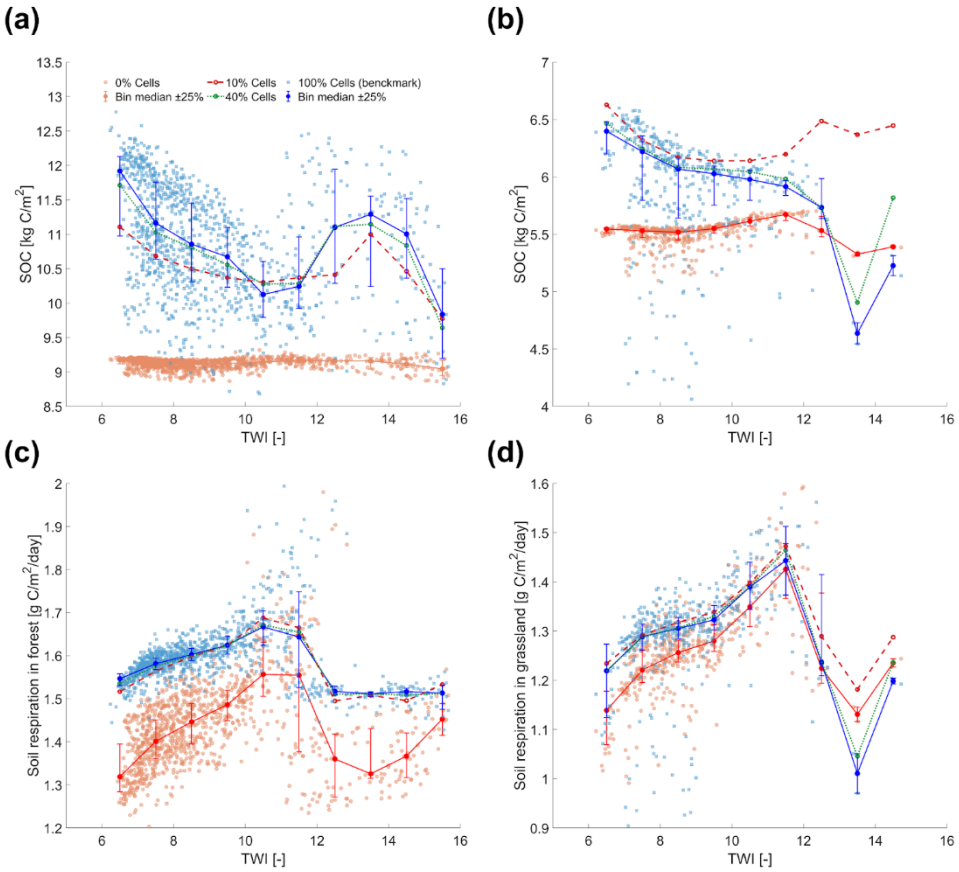

Fig. R5 (same as Fig. 6 in the revised manuscript). *Relationship between the topographic wetness index (TWI) and (a, b) soil organic carbon (SOC) and (c, d) soil respiration in the Original simulation scenario, shown separately for forest (left) and grassland (right). Red points and lines represent results from n=0% simulations; blue squares and lines show results from n= 100% simulations. Filled circles indicate the median values within TWI bins of width 1, with vertical bars showing the ±25% range within each bin. Dashed red lines and dotted green lines present results from n=10% and n=40%, respectively.*

20. Section 3.4: The direction of the soil - vegetation coupling is expected to be strongly site dependent. It is not very informative as it is likely very case specific.

Reply: Thank you for this comment. We agree that soil-vegetation coupling is strongly site-dependent. The purpose of this section is to illustrate, for the present case study,

how vegetation dynamics influence the initialization of soil biogeochemical pools in a one-dimensional spin-up framework, and whether the bias from the uncoupled simplification is acceptable. Importantly, site dependence is already accounted for in our methodology. During the spin-up procedure (Fig. 2d), the soil biogeochemical module is driven by vegetation dynamics obtained from step c in Fig. 2. These vegetation states and fluxes (e.g., litter inputs, nutrient constraints) reflect site-specific controls such as climate, soil properties, and topography. In turn, soil constraints on vegetation are consistent with the vegetation states derived in step c. As a result, the vegetation inputs used to initialize the soil biogeochemistry differ across scenarios and sites.

We note that, in our framework, we only consider the soil-vegetation coupling for 9 years, and use the average state and fluxes over this period to approximate the long-term steady-state conditions. We have expanded the discussion to acknowledge the site-specific nature of this assumption, and the comparison between coupled and uncoupled approaches. We have clarified this after line 464 (same as reply to major comment 1):

"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."

21. Figure 7: what is the purpose of 3 x 9 year long simulations? One cannot expect that 27 years are sufficient to equilibrate a coupled soil - forest model. Can one rule out that the steady-state is independent of the initial state? If not, it is speculated that if one would continue the 3 x 9 year long simulation reaches the same state as the coupled-spin-up (T&C BG).

Reply: Thank you for your comment. Indeed 27 years are not sufficient to fully equilibrate a coupled soil–vegetation system. The purpose of the 3×9-year simulations is to illustrate the convergence behavior of SOC under successive coupled simulations. After enough simulation time the system will reach the coupled steady state. We have clarified this intention in the revised text.

In line 373, now it reads:

"The next three columns show SOC values after 9, 18, and 27 years of simulation using the full T&C-BG model (i.e., with coupled soil biogeochemical and vegetation dynamics). These three columns are used to exemplify the gradual convergence of SOC toward steady state conditions. The gray box in the middle shows the SOC state obtained using the biogeochemistry-only spin-up with 9-year average vegetation fluxes after 1000 years……"

22. Line 344: I cannot follow the logic here. A 10-20 % bias in average 'steady-state' SOC stocks when using the mixed uncoupled - coupled spinup strategy can - depending on the composition of SOC among pools - lead to C fluxes which are potentially larger than the impact on C fluxes from changing environmental conditions. The authors should demonstrate that such biases lead to C fluxes which are negligible compared to the typical C cycle response investigated with this type of model.

Reply: Thank you for this important comment. We conducted an additional test comparing the transient ecosystem carbon and water flux responses using initial conditions from uncoupled and coupled steady states. The results show that the resulting transient responses of key carbon and water fluxes are very similar between the two initialization strategies (**Fig. R2**). For details, please refer to our reply to major comment 1.

23. Line 361: Where is the computation time reduction shown, how was it estimated (I would assume it depends on the machine (e.g # or processors used, etc).

Reply: Thank you for this instructive comment. We have revised the paragraph and reported the wall time of different components of the spin-up procedure in Table R1. You can find our detailed reply to the major comment 1. The revised paragraph at lines 396-407 now reads:

"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."

References

Even, R. J., Machmuller, M. B., Lavallee, J. M., Zelikova, T. J., & Cotrufo, M. F. (2025). Large errors in soil carbon measurements attributed to inconsistent sample processing. Soil, 11(1), 17-34.

Fatichi, S., Ivanov, V. Y., & Caporali, E. (2012). A mechanistic ecohydrological model to investigate complex interactions in cold and warm water-controlled environments: 1. Theoretical framework and plot-scale analysis. Journal of Advances in Modeling Earth Systems, 4(2).

Fatichi, S., Manzoni, S., Or, D., & Paschalis, A. (2019). A mechanistic model of microbially mediated soil biogeochemical processes: A reality check. Global Biogeochemical Cycles, 33(6), 620-648.

Feng, W., Plante, A. F., & Six, J. (2013). Improving estimates of maximal organic carbon stabilization by fine soil particles. Biogeochemistry, 112(1), 81-93.

Fowler, A. F., Basso, B., Millar, N., & Brinton, W. F. (2023). A simple soil mass correction for a more accurate determination of soil carbon stock changes. Scientific Reports, 13(1), 2242.

Lian, T., Fatichi, S., Stähli, M., & Bonetti, S. (2025). Assessing spatial patterns of carbon and nutrient dynamics in catchments of complex topography. Water Resources Research, 61(10), e2025WR040260.

Matus, F. J. (2021). Fine silt and clay content is the main factor defining maximal C and N accumulations in soils: a meta-analysis. Scientific Reports, 11(1), 6438.

Pawusch, L., Scheurer, S., Nowak, W., & Maxwell, R. (2025). HydroStartML: A combined machine learning and physics-based approach to reduce hydrological model spin-up time. arXiv preprint arXiv:2504.17420.

Shi, Y., Eissenstat, D. M., He, Y., & Davis, K. J. (2018). Using a spatially-distributed hydrologic biogeochemistry model with a nitrogen transport module to study the spatial variation of carbon processes in a Critical Zone Observatory. Ecological Modelling, 380, 8-21.

Six, J., Conant, R. T., Paul, E. A., & Paustian, K. (2002). Stabilization mechanisms of soil organic matter: implications for C-saturation of soils. Plant and soil, 241(2), 155-176.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological), 36(2), 111-133.

Sun, Y., Goll, D. S., Chang, J., Ciais, P., Guenet, B., Helfenstein, J., ... & Zhang, H. (2020). Global evaluation of nutrient enabled version land surface model ORCHIDEE-CNP v1. 2 (r5986). Geoscientific Model Development Discussions, 2020, 1-65.

Sun, Y., Goll, D. S., Huang, Y., Ciais, P., Wang, Y. P., Bastrikov, V., & Wang, Y. (2023). Machine learning for accelerating process-based computation of land biogeochemical cycles. Global Change Biology, 29(11), 3221-3234.

Tague, C. L., & Band, L. E. (2004). RHESSys: Regional Hydro-Ecologic Simulation System—An object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling. Earth interactions, 8(19), 1-42.

Zhou, W., Guan, K., Peng, B., Margenot, A., Lee, D., Tang, J., ... & Wang, S. (2023). How does uncertainty of soil organic carbon stock affect the calculation of carbon budgets and soil carbon credits for croplands in the US Midwest?. Geoderma, 429, 116254.

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 **revised manuscript with tracked changes.**

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 186 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 167) and Discussion section (line 464) to clarify the applied stratified sampling approach and acknowledge the alternative potential uses of a formalized sampling strategy.

Line 167: "*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 464: *"*.....*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……."*

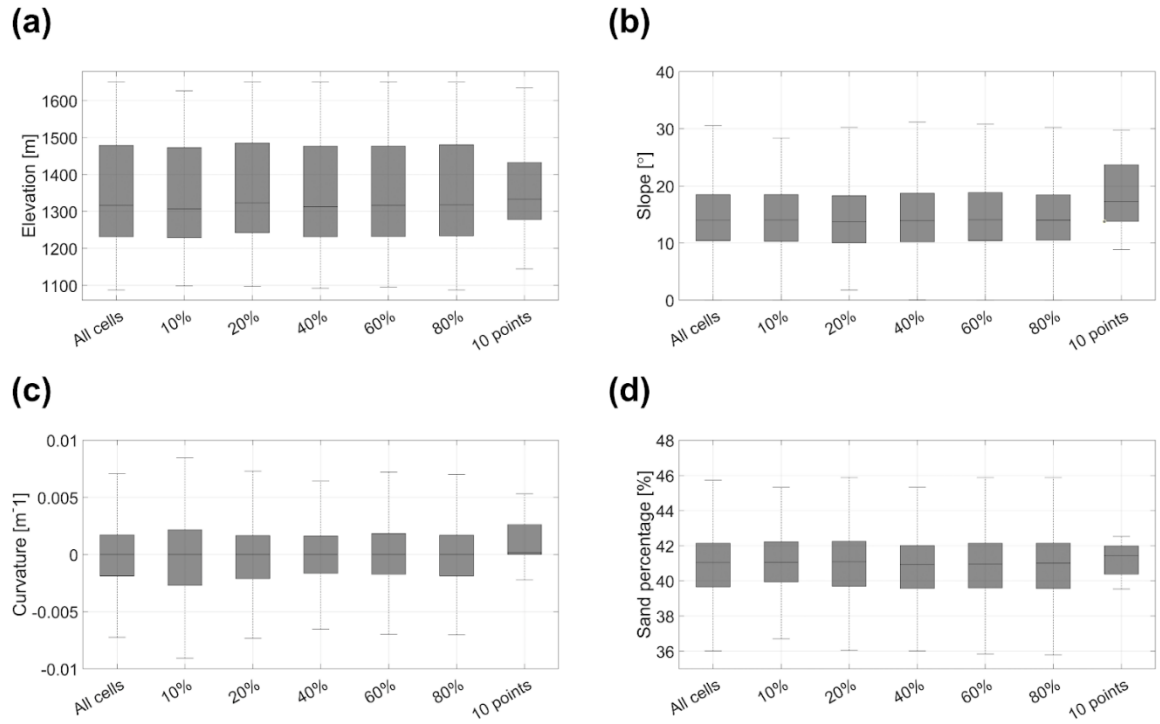

Fig. R1 (same as Fig. S2 in the supplementary information). *Distribution of topographic and soil attributes for all grid cells and selected subsets used for random forest model training. Panels show the distributions of (a) elevation, (b) slope, (c) curvature, and (d) sand percentage for all cells in 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 452: *".....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 464: *".....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 392 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 397-407 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.

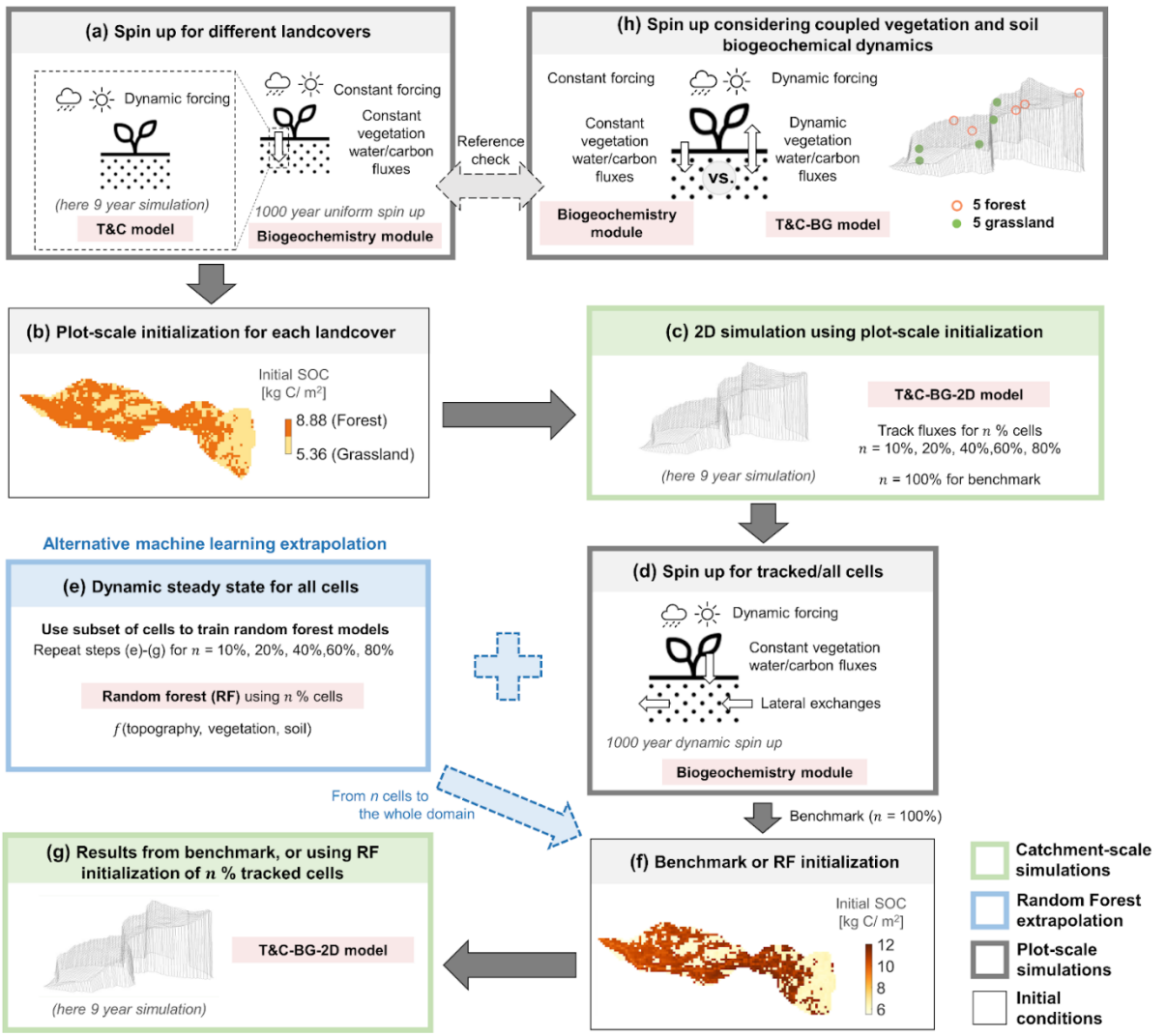

Fig. R2 (same as revised Fig. 2 in the manuscript). *Workflow of the model initialization procedure. (a) For each soil–vegetation combination, the Deleted: perform T&C model is used to obtain Deleted: to get averaged water and plant fluxes, which are used in a 1D spin-up via the T&C-BG biogeochemistry 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 33: *"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 47: *"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 59: *"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 62: *"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 159: *"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 179: *"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 114: *".....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 174: "*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 452: *".....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 464: *".....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 464: *".....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.