Reviewer #2: Toni Viskari

General Comments:

Comment #1

This is the review for the manuscript "Constraining boreal carbon allocation and turnover by assimilating forest growth dynamics in a differentiable framework" by Wu et al where they have applied biomass time series data to constrain forest growth.

It is probably easiest if I start with my recommendation here which is reject with encouragement to resubmit. While the foundation of the work is solid, at least from what I understood as I will soon further explain, I had fundamental issues with the work here. To the degree that I feel the experimental setup needs to be completely reconsidered, which is, for me, beyond major revisions.

There were multiple parts of the Methodology section where there were either major parts missing or that I didn't quite understand what was actually being done, for example the calibration of the initial state or how the growth time series was in effect used as calibration datapoints. Furthermore, there were also parts where I thoroughly disagreed with the validation work such as using the actual growth time series also for validation.

I've done my best to explain these issues in my line-by-line comments, hopefully to an understandable degree.

Additionally, the novelty of the work needs to be better conveyed as there have been previous studies about using biomass time series measurements for calibrating forest models. It did feel like the argument was partially build on DifferLand, but even that has been done before, as explained in the references of this work, and the application itself isn't unique in itself.

Apologies on my review coming across as harsh. I do still genuinely recommend on resubmission after clarification and reworking the experimental setups.

Response #1

Thank you for explaining these concerns in detail. We recognize that the original manuscript might not describe the experimental setup clearly enough. We have revised both the methodological description and the interpretation of the results.

First, the revised Methods describe how each experiment jointly optimized its own model parameters and initial pools. We clarified that the growth-curve-constrained experiment assimilated annual biomass constraints for stand ages 1–100 and aligned GPP, ET, RECO, and LAI observations from 2001–2012 with simulation years 89–100. We also distinguished these calibration simulations from the later comparison using a common prescribed initial state. We removed the claim that this comparison isolates the effects of physical parameters.

Second, we agree that agreement with the assimilated biomass–age curves cannot serve as independent validation. We now identify the age-100 biomass statistics as calibration fit and distinguish them from the 2013–2019 temporal evaluation and the independent field-allocation comparison. Neither comparison independently validates biomass accumulation over 100 years, and the revised manuscript explicitly acknowledges this limitation.

Finally, we expanded the discussion of previous biomass-assimilation studies and clarified the specific contribution. As explained in our response to **Reviewer #1 Comment #2**, our contribution is the regional-scale assimilation of grid-specific, century-scale biomass–age curves across the boreal domain in a differentiable DALEC framework. We compare the baseline experiment without biomass constraints, the single-year

biomass-constrained experiment, and the growth-curve-constrained experiment, and conduct a spatial analysis of the resulting effective allocation and turnover parameters. DifferLand provides the established computational framework for this analysis.

Abstract and Introduction:

Comment #2

Line 18: "Terrestrial biosphere models (TBMs) exhibit substantial…"

For ease-of-reading, I would suggest splitting this sentence into two.

Line 20: "Typical model-data fusion approaches…"

I'd change the word approaches to implementations as the used fusion method itself does not really care that much if higher- or lower-frequency data is used.

Response #2

We split the sentence and changed "approaches" to "implementations" as suggested.

Comment #3

Line 38: This is a very minor comment, but I would use another verb than regulate when discussing ecosystems as a part of the global carbon cycle. To regulate something refers to active participation to direct events towards a desired outcome, which does not apply here.

Line 40: "…terrestrial carbon and water fluxes and assess how…" -> "terrestrial carbon and water fluxes as well as assess how..."

Response #3

We removed "regulating" from the sentence and changed "and assess how" to "as well as assess how" as suggested.

Comment #4

Line 43: "These uncertainties primarily stem from differences in how models represent key biogeochemical and biophysical processes, such as photosynthesis and respiration, carbon allocation, vegetation dynamics, disturbance, and soil carbon turnover, and in how these processes respond to climate and CO2 (Canadell et al., 2021)-"

While technically correct, this sentence is essentially stating that ecosystem model uncertainties stem from how the different dynamics are presented in the model. Which is true, but kind of a moot point to make as obviously they are. Thus, this sentence should either be sharpened to better reflect the work in this manuscript or just removed.

Response #4

Rather than deleting the sentence beginning "These uncertainties primarily stem...," we revised it to provide a more balanced discussion of structural and parametric uncertainty, also following the suggestion from **Reviewer #1.** Please see our response to **Reviewer #1, Comment #6.**

Comment #5

Line 48: "…and together these uncertainties represent a major barrier to reliable future climate projections (Kaufhold et al., 2025)."

Does the uncertainties here refer to just the ecosystem related uncertainties or the combined uncertainties of the ecosystem and physical climate sensitivity? As it is a bit unclear in the current framing.

Response #5

We clarified that the barrier to reliable future climate projections arises from the combined uncertainties in terrestrial carbon-cycle responses and physical climate sensitivity. The revised sentence (**Line 52–54**):

"Uncertainty in terrestrial carbon-cycle responses is comparable in magnitude to the uncertainty in physical climate sensitivity (e.g., equilibrium climate sensitivity, ECS), and their combined uncertainties represent a major barrier to reliable future climate projections (Kaufhold et al., 2025)."

Reference

Kaufhold, C., Willeit, M., Talento, S., Ganopolski, A., and Rockström, J.: Interplay between climate and carbon cycle feedbacks could substantially enhance future warming, Environmental Research Letters, 20, 044027, 10.1088/1748-9326/adb6be, 2025.

Comment #6

Line 49: "A primary uncertainty in TBMs lies in the post-photosynthetic processes, especially the parameterization of carbon 50 allocation and biomass turnover (i.e., carbon residence time) (Bloom et al., 2016)."

The explicit implication of this sentence is that the actual structure is the same across all TBMs and the only things varying are the parameterizations of the actual equations. Which is naturally not the case. This sentence could actually be removed all-together as it is in a certain way repeated a little bit later in the paragraph.

Response #6

We deleted the sentence beginning "A primary uncertainty in TBMs lies...".

Comment #7

Line 54: "TBMs typically rely on Plant Functional Types (PFTs) to prescribe these parameters, but this PFT-based classification exhibits a limited ability to account for the observed spatial heterogeneity in carbon allocation and residence times (Bloom et al., 2016)."

Again, this statement is technically correct, although I'd argue this is simplifying the PFT application discussion. The work in this manuscript, however, does very little to address the core challenge with the PFT approach, so I am not certain of the benefit of highlighting the PFT issue here in this manner.

Line 57: "Using model-data fusion methods to constrain parameters faces the challenge of equifinality, where different parameter combinations can fit the observational data equally well (Famiglietti et al., 2021)."

True, but once more not something addressed in the manuscript.

If you are parameterizing with the same set of data and can find multiple parameter sets that produce the similar

fits, that is equifinality. But once you start changing the data used to constrain the models, then it is not an equifinality question anymore.

Response #7

We have removed both sentences.

Comment #8

Line 64: “Indeed, signals originating from large, slow-turnover carbon pools (such as woody 65 biomass) are difficult to detect in high-frequency flux data (Braswell et al., 2005).”

It would be helpful if the explanation here touched on why high-frequency data is relied on so much.

Additionally, in the start of the paragraph, the text mentions the use of both satellite and flux data, so I would remove the word flux from this sentence.

Response #8

We added a brief explanation of why high-frequency observations are widely used and removed “flux” from the sentence referring collectively to satellite- and tower-derived observations. The revised text states (**Line 62–63**):

"This reliance reflects their frequent temporal sampling and, for satellite products, broad spatial coverage."

Comment #9

Line 74: “Unlike high-frequency flux data and biomass snapshots, biomass growth curves capture the integrated outcomes of long-term carbon allocation and turnover processes over decades (Xu et al., 2026; Zhou et al., 2015)”

Is there any known reason why this data hasn’t been used before in model calibration?

Response #9

Repeated woody-biomass observations and chronosequence relationships have previously been assimilated at site and landscape scales, as now acknowledged in the literature review added in response to **Reviewer #1 Comment #2.** Their regional use has been more limited because spatially extensive datasets pairing biomass with time since disturbance have been scarce. We added the following clarification (**Line 81–83**):

"Regional use of biomass–age curves has been limited by the scarcity of spatially extensive datasets pairing biomass with time since disturbance, which require both long disturbance records and spatially consistent biomass observations."

Comment #10

Line 76: “A key challenge in TBM parameter calibration is equifinality…”

This line is repetition of the preceding paragraph.

Line 82: “…thereby reducing equifinality in post-photosynthetic parameter estimates.”

Now I am repeating myself from before, but once you add a new data source as a constraint, then it is not

reducing equifinality with regard to calibrations done without that dataset.

Response #10

We also deleted the repeated definition of equifinality beginning "A key challenge in TBM parameter calibration…".

Because the experiments assimilate different observational constraints, we no longer describe the addition of age-related biomass data as reducing equifinality. Instead, we revised the text to describe the specific information supplied by the different data types (**Line 91–95**):

"A single-year biomass stock does not directly measure the rate of biomass accumulation. Different combinations of carbon allocation and turnover can therefore reproduce the same stock, including high allocation with rapid turnover and low allocation with slow turnover. Age-related biomass data from forest inventories (Thum et al., 2017), chronosequences (Zhou et al., 2015), or remote sensing observations (Neigh et al., 2025) describe how biomass changes with stand age."

Reference

Neigh, C. S. R., Montesano, P. M., Sexton, J. O., Wooten, M., Wagner, W., Feng, M., Carvalhais, N., Calle, L., and Carroll, M. L.: Russian forests show strong potential for young forest growth, Commun Earth Environ, 6, 71, 10.1038/s43247-025-02006-9, 2025.

Thum, T., MacBean, N., Peylin, P., Bacour, C., Santaren, D., Longdoz, B., Loustau, D., and Ciais, P.: The potential benefit of using forest biomass data in addition to carbon and water flux measurements to constrain ecosystem model parameters: Case studies at two temperate forest sites, Agricultural and Forest Meteorology, 234-235, 48-65, 10.1016/j.agrformet.2016.12.004, 2017.

Zhou, T., Shi, P., Jia, G., Dai, Y., Zhao, X., Shangguan, W., Du, L., Wu, H., and Luo, Y.: Age-dependent forest carbon sink: Estimation via inverse modeling, Journal of Geophysical Research: Biogeosciences, 120, 2473-2492, 10.1002/2015JG002943, 2015.

Comment #11

Line 91: "To achieve this, we employed DifferLand, an advanced, JAX-based differentiable TBM (Fang & Gentine, 2024). Using the differentiable DALEC implementation provided within DifferLand, automatic differentiation computes gradients of the multi-source mismatch with respect to model parameters, enabling efficient gradient-based calibration against heterogeneous constraints that combine fast fluxes with slow carbon stocks/trajectories (Fang & Gentine, 2024; Lin, 2024; Shi et al., 2024)."

First, there is no explanation here what are JAX and DALEC. Second, stating that DifferLand is differentiable TBM that uses a differentiable DALEC implementation where the automatic differentiation computes gradients is a bit staggering to read. Third, it is not clear what differentiable means in this context.

Response #11

Thank you. We have revised this passage as follows (**Line 102–108**):

"We employed DifferLand, a data-assimilation framework implemented in JAX (Fang and Gentine, 2024), a numerical computing library that supports automatic differentiation (Bradbury et al., 2018). Within DifferLand, we used a differentiable implementation of the Data Assimilation Linked Ecosystem Carbon (DALEC) model. Here, "differentiable" means that gradients of the calibration objective with respect to the optimized model parameters and initial carbon pools can be computed through the model simulation. These gradients are then used to calibrate the model against observations spanning fast processes and slow biomass dynamics (Lin, 2024; Shi et al., 2024; Fang and Gentine, 2024)."

Reference

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and others: JAX: composable transformations of Python+ NumPy programs, 2018.

Fang, J. and Gentine, P.: Exploring Optimal Complexity for Water Stress Representation in Terrestrial Carbon Models: A Hybrid-Machine Learning Model Approach, Journal of Advances in Modeling Earth Systems, 16, e2024MS004308, 10.1029/2024MS004308, 2024.

Lin, M.: Automatic Functional Differentiation in JAX, in: International Conference on Representation Learning, 2024, 8139-8159,

Shi, Z., Hu, Z., Lin, M., and Kawaguchi, K.: Stochastic Taylor Derivative Estimator: Efficient amortization for arbitrary differential operators, in: Advances in Neural Information Processing Systems, 2024, 122316-122353, 10.52202/079017-3887,

---

Materials and Methods:

Comment #12

Line 105: "By leveraging automatic differentiation to compute…"

There should be some basic explanation on what automatic differentiation is since it is presented here as this central approach.

Line 107: "…along with data assimilation (using observation operators between model variables and observations)."

I don't quite understand what this part is attempting to state as joint calibration of parameters is data assimilation so what is the additional data assimilation here? Especially since the parentheses part about observation operators is not data assimilation, that is just a tool that allows the comparison of different data types while doing data assimilation.

Line 108: "Calibrating ecological processes across long timescales is particularly challenging, as error signals from slow-turnover pools are often difficult to attribute to the fast physiological drivers that govern them."

This is true in a sense, but there are multiples reasons for this. Even when ignoring the time-invariant assumption, we do not generally have the level of accurate driver data across the whole time series that would be necessary to more accurately constrain the model parameters in question. For example, we do not know how much radiation there was available for the growing forest at certain crucial times, further complicated by how models represent the internal distribution of radiation within the canopy.

DifferLand, as far as I can tell, does not address any of those issues nor does it solve the fact that the DALEC model structure itself is relatively simplistic which would further complicate representing with sufficient accuracy the division of vegetation carbon between different pools. There is nothing wrong with either of those challenges, by the way, as they are just issues ecosystem modelers have to deal with, but it does make the blanket statements about how DifferLand solves the issue represented here a bit confounding.

Response #12

We removed the phrase "along with data assimilation" and the statements that DifferLand "effectively bridges the timescale gap" and "resolves the trade-offs." The revised text distinguishes the computational role of gradient propagation from limitations imposed by the observations, objective formulation, forcing data, and

model structure (**Line 122–131**):

"In this context, automatic differentiation applies the chain rule through the sequence of model calculations to evaluate derivatives of the calibration objective with respect to model parameters and initial conditions. DifferLand uses these gradients for joint optimization. Observation operators convert simulated model variables into quantities directly comparable with the corresponding observations. One computational challenge in calibrating ecological processes across long timescales is tracing how mismatches in slow state variables (e.g., biomass) depend on model parameters through the full simulation. Automatic differentiation in DifferLand performs this calculation, including with respect to parameters associated with faster processes (e.g., photosynthesis) that influence biomass accumulation. The extent to which carbon allocation and turnover parameters are constrained nevertheless depends on the assimilated observations, objective formulation, forcing data, and DALEC model structure."

Comment #13

Line 119: "Carbon enters the ecosystem through gross primary productivity (GPP)…"

This paragraph would benefit from more structured explanation of the model dynamics either by introducing equations or using a list approach to explain the different steps.

Also, I don't quite understand the use of NPP in this paragraph. To put it simply, NPP is a flux that can be determined (-ish) from flux tower measurements and is often used, quite understandably, to signal how much carbon remained within the plant itself. However, here NPP is just referred the amount of carbon distributed among the carbon pools after respiration which is not really the case.

This is not to be pedantic, but that distinction matters a lot from the mechanics point of view as the respired carbon comes from the pools within the plant and is affected by the traits of the plant itself. For example, the larger the plant, the larger the respiration. More importantly, that respiration still happens during nighttime when there is no GPP and during stress conditions the respiration can be higher than the GPP, resulting in an opposite sign for the NPP. Thus, referring to the NPP as it is here is a bit distracting as it doesn't reflect the dynamic of the system.

Response #13

Our use of NPP follows the definition NPP = GPP − autotrophic respiration. In the DALEC configuration used here, the calculation proceeds as follows: (i) the Aggregate Canopy Model estimates daily GPP, modified by the soil-water stress factor; (ii) autotrophic respiration is calculated as $R_a = f_{auto} \times GPP$, leaving $NPP = (1 - f_{auto}) \times GPP$; and (iii) this NPP is partitioned among the foliar, labile, fine-root, and wood pools.

Autotrophic respiration is represented as a fixed fraction of daily GPP, rather than being calculated separately from the sizes and maintenance requirements of individual plant carbon pools. The model does not resolve daytime and nighttime respiration separately, and, for non-negative GPP, this formulation cannot produce negative NPP when respiratory demand exceeds photosynthesis. We have clarified these limitations by adding the following sentence to the model description (**Line 144–148**):

"In this DALEC configuration, NPP is daily GPP minus autotrophic respiration. Autotrophic respiration is calculated as a fixed fraction of daily GPP, and the remaining NPP is allocated to the foliar, labile, fine-root, and wood pools. Respiration is not calculated from the sizes or maintenance requirements of individual plant pools. The model does not resolve daytime and nighttime respiration separately and cannot produce negative NPP when GPP is non-negative."

Comment #14

Line 129: “In total, the model comprises 26 tunable physical parameters (Table S1) governing these ecohydrological processes.”

This table should be a part of the main manuscript with all the prior assumptions included.

Response #14

Regarding the parameter table, we have retained Table S1 in the Supplement because this study evaluates biomass–age constraints using an established DALEC configuration, rather than introducing a new model structure. Keeping the complete parameter definitions, units, and bounds together in Table S1 allows the main text to focus on the model processes and calibration design relevant to this study. This organization also follows Fang and Gentine (2024), who provided the full parameter ranges in their supplementary Table S1. The revised Methods explicitly describe the empirical reference constraints: only ce, LCMA, and f_{auto} receive direct soft reference penalties, while the remaining parameters are constrained by their bounds, assimilated observations, and applicable EDCs. The complete objective is provided in Text S1, with further explanation in our responses to **Reviewer #2 Comment #19, #29.**

Reference

Fang, J. and Gentine, P.: Exploring Optimal Complexity for Water Stress Representation in Terrestrial Carbon Models: A Hybrid-Machine Learning Model Approach, Journal of Advances in Modeling Earth Systems, 16, e2024MS004308, 10.1029/2024MS004308, 2024.

Comment #15

Line 130: “For the subsequent analysis, the "allocation to leaf" is calculated as the sum of allocation fractions to the labile and foliar pools since they both support canopy development.”

Are the other labile pools root extracts or fine roots? Also why not just use the allocation to foliar pool as allocation to leaf?

Response #15

The labile pool in this configuration temporarily stores carbon for subsequent leaf development. Its release supplies the foliar pool through the CDEA phenology scheme. It does not supply fine roots, wood, or root exudates. Reporting only direct foliar allocation would therefore omit the NPP allocated to leaf development through this storage pathway. We define total leaf allocation as $f_{fol} + (1 - f_{fol})f_{lab}$, combining direct foliar allocation with allocation through the labile pool. This describes allocation to the leaf-development pathway, not the instantaneous carbon input to foliage. We added the following clarification (**Line 154–157**):

"In this configuration, the labile pool temporarily stores carbon for leaf development. Phenological release transfers this carbon only to the foliar pool, not to fine roots, wood, or root exudates. Total leaf allocation therefore combines direct NPP allocation to the foliar pool with NPP allocation to the labile pool. It describes the share of NPP assigned to leaf development, rather than the instantaneous carbon input to the foliar pool."

Comment #16

Line 145: “We assimilated multiple earth observation data…”

Just add a set in that it reads data -> data sets.

Also regarding this paragraph in general, I would recommend to changing it either to list or table format to make all the details easier to process. Especially the table format would be ideal as it gives you the opportunity to provide additional information about resolution and timesteps.

Response #16

We changed "data" to "data sets" and added a reference to Table S3 in the observational-data subsection. Table S3 now summarizes the products and their native spatial and temporal resolutions before aggregation to the 0.5° model grid cell.

Comment #17

Line 165: "2.2.3 Forest Biomass Constraints for Slow Processes"

Based on the section, the only disturbance used for setting the age-curve is fire? With no consideration of logging, large-scale drought mortality or pest damage?

This is more to check that all that data was filtered out and only locations that had forest fires were used here? Because if that is the case, it needs to be more explicitly explained and justified in the manuscript.

And then to the more important factor in that it is utterly unclear based on the explanations here how this data is used for calibration? Is it sampled at certain time-intervals? Is it just the end point? This part should not be as obfuscated as it is.

Response #17

The curves were fitted using forest pixels with detected fires. This selection does not imply that all effects of other disturbances or management were excluded. Xu et al. (2026a) also evaluated the curves against field biomass–age observations without restricting disturbance or management type. These comparisons support their use as general empirical biomass–age relationships. We clarified the sample selection and interpretation in the Methods (**Line 201–205**):

"Xu et al. (2026a) constructed the biomass–age curves using forest pixels with detected fires during 1985–2020. Stand age was estimated from the time since the last detected fire. Although the curves were fitted using post-fire observations, they provide general empirical biomass–age relationships that may offer a useful approximation of biomass accumulation following other types of disturbance, including harvesting, drought-induced mortality, insect outbreaks, and windthrow."

The curves constrained modeled biomass across stand ages 1–100, rather than only biomass at age 100. Modeled total live biomass was sampled at the end of each simulation year. These 100 annual values were compared with the curve values at stand ages 1–100 using NNSE. The revised Methods describe this implementation as follows (**Line 358–361**):

"For the growth-curve constraint, modeled total live biomass was calculated as the sum of the labile, foliar, root, and wood carbon pools at the end of each simulation year. The 100 annual values were compared with the biomass–age curve at stand ages 1–100 using NNSE in the observational objective."

Comment #18

Line 178: "This conversion was performed using spatially explicit Root-to-Shoot (R:S) ratios derived from a

global map of root biomass, which was generated by a random forest model trained on extensive field measurements (Huang et al., 2021)."

Does the conversion assume that this relationship stays same regardless of forest maturity? It is fine if it does, I would simply recommend adding a sentence here acknowledging that assumption.

Response #18

We have added the following sentence to acknowledge this assumption:

"This conversion assumes that the spatially explicit R:S ratio remains constant across stand ages within each grid cell."

Comment #19

Line 182: "2.2.4 Data for Parameter Priors"

In section 2.1, it is established that the work here calibrates 26 parameters. It is not clear which of those parameters the data discussed in this section is used as constraints or how the rest of the parameters are then constrained. Prior information, expert opinion, something else?

Response #19

Thank you for pointing out this ambiguity. We have renamed this subsection "Empirical Reference Constraints" and clarified which parameters were linked to the datasets described therein. We added the following text (**Line 233–238**):

"Of the 26 optimized physical parameters, only ce, LCMA, and f_{auto} received direct soft reference penalties. CUE was linked to f_{auto} through CUE = 1 − f_{auto}. The leaf lifespan–LCMA relationship was not an additional parameter-specific prior but constrained the realized annual foliar turnover generated by τ_{foliar}, c_{rfall}, and d_{fall} relative to the optimized LCMA. No direct empirical references were assigned to the other 23 physical parameters, which were optimized within the bounds listed in Table S1 and jointly constrained by the assimilated data streams and applicable EDCs."

Comment #20

Line 199: "We specifically extracted the 1985–2020 mean areal fractions of boreal needleleaf evergreen (ENF), boreal broadleaf summergreen (DBF), and boreal needleleaf deciduous (DNF)."

I raised this question before, but when calculating these fractions, is the assumption here that there are no other disturbances except fire affecting these systems? If so, what is it based on?

Response #20

No assumption that fire was the only disturbance was made when calculating or using the ENF, DBF, and DNF fractions. These fractions describe mean vegetation composition and were used as ancillary predictors after calibration. They did not enter the disturbance-history assignment, biomass-curve construction, or DALEC calibration. We have clarified their role in the Methods (**Line 248–251**):

"The 1985–2020 mean areal fractions of boreal needleleaf evergreen, broadleaf summergreen, and needleleaf deciduous forests were used as predictors in the post-calibration spatial analysis. They were not used to determine stand age, describe disturbance history, construct the biomass–age curves, or constrain DALEC

calibration."

The disturbance scope of the biomass–age curves is addressed separately in our response to **Reviewer #2 Comment #17.**

Comment #21

Line 202: "2.2.6 Independent Validation Data"

In the following section, the manuscript defines that it is using parts of the time series for independent validation and this particular data is not mentioned. This is especially confusing in the Results section where the validation dataset discussed is somewhat confusing in itself and this dataset is not referred to as the validation data.

Also I don't quite comprehend the implementation of this as why would the data from plot-level be aggregated to grid-level if it is being used for validation? Why not run the model just on a plot level? And what does the aggregation even exactly mean in this context?

Response #21

We renamed the subsection "Independent Allocation Comparison Data" to identify the field-derived allocation estimates from Xia et al. (2019), which were not assimilated during calibration. Separately, observations of GPP, RECO, ET, and LAI from 2013–2019 were withheld for temporal evaluation. The revised Methods and Results distinguish these two uses of evaluation data.

The field-allocation comparison was conducted at the grid scale because the model forcing, observational constraints, and retrieved effective parameters were defined at 0.5° resolution. We compared the field-derived allocation fractions with the model allocation fractions for the same grid cells. We did not conduct separate simulations calibrated to individual plot conditions. We clarified this reason and the aggregation procedure in Text S4:

"The field-allocation comparison was conducted at the grid scale because the model forcing, observational constraints, and retrieved effective parameters were defined at 0.5° resolution. We assigned the field observations to their corresponding 0.5° × 0.5° grid cells and calculated the unweighted arithmetic mean of the reported allocation fractions within each cell for comparison with the model allocation fractions for that cell."

This aggregation allows comparison at the same grid scale, but does not eliminate differences in spatial representativeness between sampled plots and grid-cell model results. Text S4 and the Discussion now explicitly acknowledge this limitation, together with the absence of observation-level uncertainty estimates.

Reference

Xia, J., Yuan, W., Lienert, S., Joos, F., Ciais, P., Viovy, N., Wang, Y.-p., Wang, X., Zhang, H., Chen, Y., and Tian, X.: Global Patterns in Net Primary Production Allocation Regulated by Environmental Conditions and Forest Stand Age: A Model-Data Comparison, Journal of Geophysical Research: Biogeosciences, 124, 2039-2059, doi.org/10.1029/2018JG004777, 2019.

Comment #22

Line 211: "In other words, within the differentiable framework, initial conditions can be treated as parameters and we can compute the gradients/sensitivity of the model to those initial conditions."

I didn't quite understand this statement as the initial conditions can be treated as parameters by every calibration method. Thus, this isn't something allowed only by the differentiable framework as implied here.

The challenge is, and why it is not that often done, is that it introduces a massive equifinality/uncertainty issue with the results and this possibility is not tested for at all in the resulting analysis.

Furthermore, it remains completely unclear to me that at what exactly is the initial state calibrated here? Like I understand that it is the model pools, but at what time? Since you are modeling the forest regrowth after a forest fire, does that not start from zero? If this is the initial state at the start of the validation period, which I am concerned of, this would essentially everything in that work somewhat pointless.

Response #22

We deleted the statement implying that optimizing initial conditions was specific to a differentiable framework and clarified the timing and interpretation of the initial pools in the Methods (**Line 269–270**):

"The optimized initial pools in the growth-curve-constrained experiment correspond to the beginning of the 100-year simulation rather than to a specific calendar year (Text S5)."

We added further details on the initial pools in Text S5:

"**Text S5 Initial states and calibration**

The growth-curve-constrained experiment simulates biomass accumulation over stand ages 1–100. The optimized initial pools correspond to the beginning of this simulation rather than to a specific calendar year. The initial live biomass was not fixed to zero because the current DALEC configuration does not explicitly represent seedling establishment or recruitment. The initial litter, soil organic carbon, and water pools were also optimized and need not be zero.

Across all grid cells in the growth-curve-constrained experiment, the mean initial live biomass was 7.17 Mg ha^{-1}, corresponding on average to 4.68% of the biomass at year 100. This indicates that the optimized initial live biomass was small but non-zero. In all three experiments, the model parameters and initial pools were estimated during calibration. The 2013–2019 temporal evaluation continued from each experiment's calibrated end state, without refitting the parameters or initial pools to the withheld observations. Evaluating alternative initial-state assumptions would require recalibration of both the model parameters and initial pools."

We acknowledge that the joint uncertainty of the model parameters and initial pools has not been quantified. The five optimization runs assess optimization stability. They do not establish uniqueness or characterize posterior uncertainty. We also added a paragraph to the Discussion acknowledging that the jointly optimized model parameters and initial pools may compensate for one another and that the retrieved initial pools are not uniquely identified carbon stocks (**Line 637–641**):

"Because the initial pools were optimized jointly with the model parameters, different parameter and initial-state combinations may provide similar fits to the assimilated observations. The retrieved initial pools should therefore be interpreted as effective model states conditional on the model structure and observational constraints, rather than uniquely identified carbon stocks. The forward comparison may also be affected by incompatibility between the prescribed initial pools and parameter sets calibrated in other experiments."

Comment #23

Line 214: "To capture the long-term legacy effects of carbon turnover and accumulation, we designed a 107-year continuous simulation experiment (growth-curve-constrained experiment) for each grid cell."

As write this, I have read this section now multiple times and still have really no grasp in how the work here is using the growth curve as a constraint, especially since a few paragraphs later it is stated that for all experiments use the observational data from 2001 to 2012.

Are you repeating the initial growth period with the different parameter sets? As in that case your calibration period wouldn't be the one specified later as it would cover a longer time window. What is the actual measurement used as a constraint here and how is it implemented in the calibration?

Response #23

We have revised the Methods to clarify the constraint data, their temporal alignment, and their implementation in the calibration. **Line 263–266**:

"To capture the long-term legacy effects of carbon turnover and accumulation, we designed a 100-year simulation (growth-curve-constrained experiment) for each grid cell. The meteorological and atmospheric CO_2 drivers from 1985 to 2020 (36 years) were cycled twice, followed by the first 28 years of a third cycle (1985-2012), to provide forcing for the 100-year simulation."

As explained in our response to **Reviewer #2, Comment #17,** annual modeled biomass was compared with the biomass–age curve at stand ages 1–100 using NNSE. We also clarified how the biomass–age curve was combined with GPP, ET, RECO, and LAI observations from 2001–2012 (**Line 280–286**):

"GPP, ET, RECO, and LAI observations during 2001–2012 were used for calibration, and observations during 2013–2019 were withheld for temporal evaluation. In the growth-curve-constrained experiment, the biomass–age curve constrained all 100 simulation years. The 2001–2012 GPP, ET, RECO, and LAI observations were aligned with simulation years 89–100. This alignment was based on ecosystem state rather than calendar year: simulation years 89–100 represent the mature stage of the fitted biomass–age curve, while the GPP, ET, RECO, and LAI products were specifically aggregated over mature-forest pixels. It therefore does not imply that all stands experienced disturbance in the same calendar year."

The first 88 simulation years were therefore constrained by the biomass–age curve but not by the GPP, ET, RECO, and LAI observations from 2001–2012. The additional 2013–2019 simulation was a post-calibration temporal evaluation and was not part of the 100-year calibration. The different parameter sets were compared in separate 100-year forward simulations only after calibration, as clarified in the revised Methods and Fig. 1.

Comment #24

Line 215: "The meteorological drivers from 1985 to 2020 (36 years) were cycled three times, with the final year (2020) excluded from the last cycle to align the simulation end date (2019) with the extent of the observational record."

From the manner which I have understood the logic behind the cycling meteorological driver, and how I have implemented myself, is that one can choose a period to repeat because of an underlying assumption that the general climate averages remain the same and this introduces noise to the system. When doing this, the periods chosen have usually been from the latter half of the 20th century.

Here, though, the chosen period includes the trends due to climate change. So when cycling like this, the work is actually restarting a trend three times while also projecting the growth of the forest in a much warmer environment than it actually occurred in. Especially since the northern latitudes where the forests chosen here are located have been much more strongly affected by climate change.

If I remember correctly, the CRU-JRA dataset does also have the earlier periods. So why implement the cycling in this manner? Why not run cycle with a different climate than which you then use for the actual simulation runs?

Response #24

The selected 1985–2020 period contains trends in both meteorology and atmospheric CO_2, and cycling this period repeats those trends. The exact 100-year forcing sequence is 1985–2020, 1985–2020, and 1985–2012. We selected this period because the Xu curves are biomass–age relationships estimated from recent space-for-time observations (1985-2020) rather than long-term observations of the same stands. Using an earlier forcing sequence would assume a history of climate conditions for those stands that is not recorded in the source curves.

We have revised the Methods to state (**Line 266–269**):

"The biomass–age curves combine biomass observations with stand ages derived from the 1985–2020 fire record using a space-for-time method. We thus cycled the 1985–2020 meteorological and atmospheric CO_2 forcing to simulate biomass accumulation, rather than to reconstruct the historical development of individual stands."

We have also clarified the limitation of this design in the Discussion (**Line 632–636**):

"The biomass‐age curves derived from the space-for-time method may not fully capture future growth under climate change and increasing atmospheric CO_2. Cycling the 1985–2020 drivers repeats their climate and CO_2 trends. Each simulated stand age is exposed to a particular year of that repeated record. This design does not separately test the effects of stand age and climate. The retrieved parameters therefore depend on the chosen forcing design."

Comment #25

Line 219: “To quantify the specific impact of assimilating forest biomass growth dynamics, we conducted two additional experiments, a baseline experiment and a single-year biomass-constrained experiment”

How do you initialize the forest state for these two experiments? As that would have a drastic impact on the results.

Response #25

We clarified how the baseline and single-year biomass-constrained experiments were initialized and corrected the description of the calibration duration. **Line 271–279**:

"We compared the growth-curve-constrained experiment with two additional experiments: a baseline experiment and a single-year biomass-constrained experiment. The baseline experiment follows the existing DifferLand configuration and provides a reference for comparison with the two biomass-constrained experiments. During calibration, these two experiments independently optimized their own initial pools together with the model parameters. For both experiments, the optimized initial pools defined the model state at the beginning of the 2001–2012 calibration simulation and were not taken from the growth-curve-constrained experiment. Both calibration simulations spanned 12 years (2001–2012), driven directly by the corresponding meteorological forcing without cycling. The baseline experiment assimilated only GPP, ET, RECO, and LAI, whereas the single-year biomass-constrained experiment additionally assimilated a mature-forest biomass stock derived from the 2020 biomass map."

Comment #26

Line 227: “For all experimental setups, the calibration process utilized observational data from the period 2001–2012 (corresponding to the simulation years 89–100 for the 107-year run, and years 1–12 for the 19-year run), while data from 2013–2019 were reserved for independent validation.”

Literally at the end of the previous paragraph it is stated that the biomass data used for the comparison experiment is from the year 2020. While I can imagine that there is just that gap in calibration data where the model is ran for the total biomass, this just feels like a really clumsy manner in which to do this.

Response #26

The alignment of the 2001–2012 GPP, ET, RECO, and LAI observations with simulation years 89–100 in the growth-curve-constrained experiment is explained in our response to **Reviewer #2 Comment #23.** For the single-year biomass-constrained experiment, we clarified how the 2020 biomass estimate was used and how the simulations continued after calibration (**Line 286–290**):

"In the single-year biomass-constrained experiment, the 2020 biomass map provided a single mature-forest stock constraint rather than a biomass time series. This stock estimate was compared with the modeled total biomass at the end of the 12-year calibration simulation and was not interpreted as a biomass observation for calendar year 2012. Following calibration, the model state was propagated with meteorological forcing from 2013–2019 to evaluate GPP, ET, RECO, and LAI against the temporally withheld observations."

Comment #27

Why wasn’t the validation done by putting aside individual sites to use those purely as validation? As that would have also given you information how well the results from the calibration here can be generally applied for the PFTs in question.

Response #27

The calibration in this study retrieved a separate parameter set for each grid cell. It did not estimate a shared parameter set for each PFT. Testing transfer to withheld sites would require a rule for assigning parameters calibrated elsewhere to those sites. Such parameter transfer was not evaluated in this study.

Our temporal evaluation instead assessed performance in withheld years at locations used for calibration, retaining the calibrated parameters and continuing from the calibrated model states without refitting. We acknowledge that this does not establish spatial transferability and added the following clarification to the Discussion (**Line 653–656**):

"Model parameters were calibrated separately for each grid cell rather than as shared PFT-specific parameter sets. The temporal evaluation assesses performance in withheld years at calibrated locations. It does not establish whether the retrieved parameters can be transferred to locations that were not calibrated."

Comment #28

Line 229: “By aligning the calibration period with simulation years 89–100, we assume that after approximately 90 years of recovery following the stand-replacing disturbance…”

This would be true only if you assumed that every stand burned down in the year 1911. Is the implication here that your calibration data doesn’t have any stands in varying states of regrowth, but all are already recovered

and mature?

Line 254: “By applying the physical parameters retrieved from the baseline and the single-year biomass-constrained experiments to this same initial state from the growth-curve-constrained experiment and comparing the result against the growth curve-constrained experiment, we isolated the divergence in 100-year biomass accumulations caused solely by the differences in physical parameters.”

Again, the estimated initial state at what year?

Furthermore, this setup is fundamentally flawed. The initial state and the parameters that produced it are inherently linked. If you are running the model with an initial state that is not at balance with the model parameters, that will affect how the system actually develops over the given time frame. Hence, the results here don’t really give an isolated perspective on the impact of the physical parameters alone.

Response #28

No common fire year was assumed. Fire history was used as a proxy for stand age to construct empirical biomass–age curves from stands of different ages. These curves provide the age-related biomass constraints. In contrast, the GPP, ET, RECO, and LAI products for 2001–2012 were aggregated over mature-forest pixels within each 0.5° grid cell. They were aligned with simulation years 89–100 because these years represent the mature stage of the fitted biomass–age curve, as explained in our response to **Reviewer #2, Comment #23.**

We clarified the purpose and limitations of using common initial pools in the Methods (**Line 307–314**):

"For each grid cell and optimization run, we performed 100-year forward simulations using all three calibrated parameter sets. Each simulation used the initial pools from the corresponding growth-curve-constrained experiment. These initial pools define the starting state of the biomass–age simulation, rather than a state in a specific calendar year. This common initialization standardized the starting pools across the forward simulations. With the same forcing, this comparison shows how the three calibrated parameter sets produce different biomass accumulation from the same starting state. However, those initial pools were jointly optimized with the growth-curve parameter set, not with the baseline or single-year parameter sets. The comparison therefore remains dependent on the chosen initial pools and their compatibility with each parameter set."

Fang and Gentine (2024) also optimized initial carbon pools together with model parameters. Bloom and Williams (2015) allowed DALEC carbon pools to differ from equilibrium and discussed adapting the constraints for growing forests. These studies provide precedent for jointly calibrating initial pools and parameters without requiring an equilibrium initial state. The possible effects of using the growth-curve initial pools with the other parameter sets are discussed in our response to **Reviewer #2, Comment #22.**

Reference

Bloom, A. A. and Williams, M.: Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model–data fusion framework, Biogeosciences, 12, 1299-1315, 10.5194/bg-12-1299-2015, 2015.

Fang, J. and Gentine, P.: Exploring Optimal Complexity for Water Stress Representation in Terrestrial Carbon Models: A Hybrid-Machine Learning Model Approach, Journal of Advances in Modeling Earth Systems, 16, e2024MS004308, 10.1029/2024MS004308, 2024.

Comment #29

Line 286: “The observational loss (Jobs) quantifies the mismatch between model simulations and data using

the negative Normalized Nash-Sutcliffe Efficiency (NNSE). The NNSE transforms the standard error metric onto a bounded range of (0, 1], preventing variables with large absolute magnitudes from dominating the optimization gradient."

Generally, when calculating the cost functions, the uncertainties will scale the different types of measurements to be comparable with each other. While this approach here is not inherently wrong at all, but it remained unclear to me why it was necessary in this particular situation?

Line 293: "These priors specifically target the canopy efficiency (ce), LCMA and the fraction of GPP allocated to autotrophic respiration (fauto)."

Why only use these few parameter priors instead for all the calibrated parameters?

Also, are these being scaled in the same manner as with the observations.

Line 298: "Finally, we imposed ecological and dynamical constraints (EDCs) to ensure biologically realistic process rates."

I understand the constraints themselves, but now how they are implemented in the cost function?

Response #29

We thank the reviewer for these questions. We agree that observational uncertainties are commonly used to scale heterogeneous data streams when comparable and defensible error estimates are available. In our calibration, however, uncertainty estimates were not available for all calibration inputs. We therefore used NNSE to place data streams with different units, magnitudes, and variability on a bounded numerical scale. We have clarified that this is numerical normalization rather than uncertainty weighting and that it does not constitute a likelihood-based treatment of observational uncertainty.

We also tested the effect of halving and doubling the weight of the biomass term. The results are reported in our response to **Reviewer #1, Comment #3**.

Direct reference penalties were used for canopy efficiency, LCMA, and CUE because spatial reference information could be linked directly to these model quantities. The leaf economics spectrum provided a relationship constraint, rather than a reference for another individual parameter. The other parameters and initial pools were constrained by their bounds, the assimilated observations, and the applicable EDCs. We did not assign empirical priors where defensible reference information was unavailable.

These reference penalties were not scaled in the same manner as the observational terms. The objective included NNSE terms for GPP, ET, RECO, and LAI with negative signs and a coefficient of 1. The growth-curve-constrained experiment used the same coefficient for the NNSE between modeled biomass and the biomass–age curve. The single-year biomass constraint instead used a term based on the squared relative difference between modeled and observed mature biomass. The CE, LCMA, and CUE penalties used squared differences between the logarithms of model quantities and reference values. The LES penalty used the same form to compare realized annual foliar turnover with the turnover predicted from LCMA.

To clarify these points, we revised the Methods accordingly:

"All three experiments included NNSE terms for GPP, ET, RECO, and LAI. The baseline experiment included no biomass term. The single-year biomass-constrained experiment used a term based on the squared relative difference between modeled and observed mature biomass. The growth-curve-constrained experiment used NNSE to compare modeled biomass with the biomass–age curve. NNSE scales the fit terms for different datasets but does not weight observations by their uncertainties."

Line 367–373:

"Direct reference penalties were used only where spatial reference information could be linked to a model quantity. The other parameters and initial pools were constrained by their bounds, the assimilated data, and applicable EDCs. Static parameter references (Canopy efficiency, LCMA, and CUE) entered the objective J_{prior} as separate penalties based on squared differences between the logarithms of ce, LCMA, and CUE and the logarithms of their respective reference values. The parameters could depart from those values during calibration. The penalty coefficients were fixed regularization strengths rather than inverse observational variances. Their complete formulation is provided in Text S1."

Line 389–390:

"For the leaf economics spectrum constraint, we used the squared difference between the logarithms of the realized annual foliar turnover and the turnover predicted from LCMA."

Supplementary Text S1 now gives the complete objective:

"During optimization, we minimized an objective function that included the NNSE terms with negative signs. Valid-data masks selected available observations only and were not inverse-variance weights. Defining R_{fast} as the sum of the NNSE values for GPP, ET, RECO, and LAI, the complete objective was:

$$J = -R_{fast} - R_B + P_{CE} + P_{LCMA} + P_{CUE} + P_{LES} + P_{preEDC} + P_{postEDC}$$

R_B was zero in the baseline experiment, $1 - \left[\left(\widehat{B} - B_{obs}\right)/B_{obs}\right]^2$ in the single-year biomass-constrained experiment, and the NNSE between the 100 modeled and observed annual biomass values in the growth-curve-constrained experiment. The static-reference penalties were $P_{CE} = 0.1\left[\ln(ce) - \ln\left(ce_{opt}\right)\right]^2$, $P_{LCMA} = [\ln(LCMA) - \ln(LCMA_{ref})]^2$, and $P_{CUE} = [\ln(CUE) - \ln(CUE_{ref})]^2$. These penalties were applied only where valid reference values were available. P_{LES} was the squared log difference between realized annual foliar turnover and the annual turnover predicted from LCMA. The pre-integration term, P_{preEDC}, summed smooth negative-log-sigmoid penalties for violations of the parameter-based turnover and allocation inequalities. The post-integration term, $P_{postEDC}$, was evaluated from the model simulations and penalized the applicable carbon-pool input-output imbalances and pool-boundary violations. Both were non-negative soft penalties added to J rather than hard rejection criteria. The growth-curve-constrained experiment allowed vegetation and litter pools to accumulate over the 100-year simulation, whereas the baseline and single-year biomass-constrained experiments additionally applied the mature-state pool-balance penalties."

Comment #30

Line 312: “We developed separate machine learning models to predict the calibrated values of key process parameters, specifically leaf lifespan, carbon turnover rates for wood and fine roots, and NPP allocation fractions to leaf, wood, and fine roots.”

I don’t understand the purpose of this as aren’t these the parameters you are calibrating in this work? So why develop a separate machine learning models to then predict them? Apparently this is saying that it allows you to project them to different environments, but you are already using those for the calibration?

Line 313: “To attribute the spatial heterogeneity of the calibrated parameters to environmental drivers…”

While the sentence is grammatically correct, I would suggest writing this out a bit more as in the current phrasing it feels initially a bit confusing. Additionally, you are using the word attribute twice in this sentence

Response #30

We used the machine-learning analysis to understand how climate, soil properties, and vegetation composition are associated with spatial differences in the calibrated parameters. The calibrated parameters were the response variables, and the environmental covariates were the predictors. This was an analysis of the calibration results, rather than a new calibration or a projection to other environments. The revised text states (**Line 395–397**):

"Following grid-cell calibration, we used gradient boosting regression to examine how spatial variation in the calibrated parameters was associated with climate, soil properties, and vegetation composition (Friedman, 2001; Pedregosa et al., 2011)."

Reference

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, 29, 1189-1232, 2001.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and others: Scikit-learn: Machine learning in Python, the Journal of machine Learning research, 12, 2825-2830, 2011.

Results:

Comment #31

Line 330: "The baseline experiment, which constrained the DALEC model solely using high-frequency flux and LAI observations…"

Because there has been no explanation how the model state has been initialized for the different experiments, it is quite impossible to evaluate the performance here. Especially since the validation window is apparently 10 years in a slow-developing ecosystem where there are no disturbances, at least if I understood your implementation correctly, which essentially means that your results are going to be heavily impacted by the initial state.

Line 335: "Consequently, the baseline simulation yielded a poor fit with the 100-year biomass derived from the forest growth curves (R2 = 0.02), with a large RMSE of 138.7 Mg ha-1."

Alright, first of all this is not neither of the validation datasets introduced in the Methodology section.

Second, and even more crucially, isn't the whole point of the work introduced here to calibrate with the forest growth curves? Because if that is the case, is the qualifier for improvement that the two comparison experiments didn't perform as well compared to this dataset as the experiment where that very same data is used for calibration?

If that is not what was done here, then this part here is very misleading. But if that is instead what was done, then this comparison is so fundamentally flawed.

Line 347: "This constraint largely improved the spatial performance of the model simulation, increasing the R2 to 0.99 and reducing the RMSE to 11.3 Mg ha-1."

I am going to repeat myself here somewhat, but this is simply, for me, really crucial to get across.

Based on what was explained above, the implication is that their performance is with the same data that was used for calibration. At which point getting a R2 value that high makes sense. The other explanation would be that the initial states mentioned above are for the start of the validation period, in which case since you are

apparently running mature forests, there is almost no change expected in such a short period of time.

With that stated, and I want to stress this again here, if you are getting a R2 value of 0.99 with the model and datasets used here, that should have been an instant sign that there was something severely wrong with your test setup.

Response #31

Yes, the biomass–age curves were used to calibrate the growth-curve-constrained experiment. The reported R^2 of 0.99 compares simulated age-100 biomass with those same curve values across grid cells. It measures calibration fit. It is not an independent validation score. A better fit to the assimilated curves alone does not demonstrate better predictive skill.

Each experiment jointly optimized its own model parameters and initial pools, as clarified in our responses to **Reviewer #2, Comments #22 and #25.** The temporal evaluation (2013–2019) continued from each calibrated end state without refitting parameters or initial pools to the withheld observations. These simulations remain influenced by the calibrated states. The short evaluation period limits what it can show about slow carbon pools. The separate 100-year comparison used common growth-curve initial pools. It compares the parameter sets under those initial pools and does not isolate physical-parameter effects, as explained in our response to **Reviewer #2, Comment #28.**

The 2013–2019 evaluation and the field-allocation comparison address different questions. The three experiments performed similarly against withheld GPP, RECO, ET, and LAI observations (Fig. S6). The field-derived allocation estimates provide an external comparison of the retrieved allocation fractions (Fig. S8). However, their observation-level uncertainties were unavailable, and sampled plots may not represent the full grid cell. Neither comparison independently validates biomass accumulation over 100 years. We revised the manuscript to distinguish these comparisons from calibration fit, as detailed below.

In the Abstract, we now state (**Line 27–30**):

"As expected, the growth-curve-constrained calibration closely reproduced the age-100 biomass values from the assimilated biomass–age curves (RMSE = 11.3 Mg ha^{-1}), while maintaining comparable performance against the temporally withheld GPP, ET, RECO, and LAI observations."

In the Results, we revised the description of all three experiments (**Line 415–417**):

"The baseline experiment, which constrained the DALEC model solely using high-frequency flux and LAI observations, showed large discrepancies from the reference biomass–age curves in the 100-year comparison with the same initial pools (Fig. 2)."

Line 424–425:

"Its agreement with the reference age-100 biomass was greater than that of the baseline experiment (R^2=0.35 and RMSE = 87.5 Mg ha^{-1}), although substantial systematic differences remained."

Line 430–433:

"As expected, the growth-curve-constrained simulation closely reproduced the age-100 biomass values from the biomass–age curves used in calibration (Fig. 2c, R^2 = 0.99, RMSE = 11.3 Mg ha^{-1}). These statistics measure calibration fit, rather than independent validation. Performance against the withheld 2013–2019 observations of GPP, RECO, and ET remained similar across the three experiments (Fig. S6)."

We added the clarification to the Fig. 2 caption (**Line 454–456**):

"For the growth-curve-constrained experiment, the reference biomass–age curves were also used during

calibration, and therefore the displayed correspondence represents calibration fit rather than independent validation."

Finally, the corresponding Discussion passage now reads (**Line 613–615**):

"Assimilating the biomass–age curves substantially improved the fit to these curves, with the RMSE decreasing from 138.7 to 11.3 Mg ha^{-1}. The latter value represents calibration fit rather than independent validation."

Comment #32

Line 385: "The gradient boosting models captured ecologically consistent signals from environmental drivers…"

Ecologically consistent based on what? Also what are the sequential R2 values in reference to?

Response #32

We replaced "ecologically consistent signals" with a description of predictive performance and explicitly identified the quantities corresponding to the reported R^2 values. The revised Results state (**Line 485–491**):

"The gradient boosting models predicted spatial variation in the calibrated parameters, explaining a moderate proportion of the spatial variance with R^2 values ranging from 0.27 to 0.76 (Fig. S12). Predictive performance was highest for canopy-related traits, specifically the fraction of NPP allocated to leaves (R^2 = 0.60) and leaf lifespan (R^2 = 0.76). In contrast, parameters associated with slower or belowground processes, including NPP allocation fractions to wood and fine roots and their respective mean carbon residence times, showed lower explainability (R^2 = 0.32 and 0.42 for wood and fine-root allocation, and 0.32 and 0.27 for their respective residence times)."

Comment #33

Line 418: "For the fraction of NPP allocated to wood, VPD emerged as the most important predictor, followed by annual mean temperature and solar radiation (Fig. 4b)"

This perhaps would benefit in analysing together and with more depth with the previous paragraphs as in your current result wood turnover, in essence mortality, is most explained by solar radiation while NPP allocation, thus new growth, is most driven by VPD.

This isn't to question the results, but rather that their combination is a bit surprising as this is the growth/diminishing of the forest.

Response #33

We added a joint interpretation of the solar-radiation and VPD associations with wood allocation and residence time (**Line 524–528**):

"In the SHAP analysis, solar radiation was associated with a greater fraction of NPP allocated to wood and shorter wood residence time (Fig. 3c and Fig. 4g), whereas high VPD was associated with a lower fraction of NPP allocated to wood. A greater fraction of NPP allocated to wood does not necessarily mean faster net biomass accumulation. Carbon input to wood also depends on total NPP, and carbon loss depends on the existing wood stock and its turnover rate."

Comment #34

Line 462: “Although both frameworks perform grid-level parameter retrieval, they differ critically in the observational constraints used for boreal ecosystems.”

Is CARDAMOM also calibrating the initial state? Because if not, then it is important to highlight the further differences in the process.

Response #34

Yes. The publicly documented DALEC4 implementation used by CARDAMOM v13 jointly optimizes initial carbon pools and process parameters, as clarified in Text S8. We therefore do not treat initial-state optimization as a distinction between the two studies. The revised comparison is detailed in our response to **Reviewer #1 Comment #5.**

Discussion:

Comment #35

Line 490: “4. Discussion”

This was a challenging section for me to comment on as, for me, there was so much vagueness and questionable comparisons in the Methodology/Results sections. Because of that, it is difficult to truly to evaluate the claims/analysis.

I do not write to be dismissive this section, but rather that this section should be gone through again after the revisions to the previous sections to see if they are still inline. I have also made some comments about the general structure that I feel would improve the manuscript.

Line 491: “A major uncertainty in terrestrial biosphere modeling arises from equifinality, where different parameter combinations yield similar fits to observational data (Bloom et al., 2016; Famiglietti et al., 2021).”

Apologies for the repetition, but this work hasn’t really dealt with equifinality at all. Different data is used for calibration which then perform differently with the validation dataset. Which is fine, but it makes it distracting that the Discussion section is started with the statement.

Line 493: “In our baseline experiment, 493 constrained solely by high-frequency observations (GPP, RECO, ET, and LAI), the model reproduced seasonal flux 494 dynamics well but showed large biases in long-term biomass stocks.”

I would recommend reworking the whole first paragraph of the discussion as for some reason, the focus here is on the comparison experiments instead of the actual new research presented in this work.

Line 498: “In our baseline experiment…”

This is the second time in this paragraph that a sentence is started with these words.

Line 508: “To constrain these slower processes…

Both this and previous paragraphs feel more like repetition of Methods and Results instead of Discussion. I would recommend starting this by instantly connecting the central parts of this research with the larger context.

Response #35

We have revised the Discussion to match the clarified Methods and Results. Its opening now focuses on what biomass–age information adds to calibration, rather than repeating the experimental setup or leading with the baseline comparison. We also removed claims about reducing equifinality and about attributing the benefit of the additional data specifically to DifferLand. The revised opening states (**Line 588–590**):

"At the boreal-domain scale, our comparison of the three experiments indicates that grid-specific biomass–age curves add information on the rate and shape of long-term live-biomass accumulation. This information goes beyond that provided by high-frequency observations or a single contemporary biomass stock."

We revised the second occurrence of "In our baseline experiment" and deleted the methods-style sentence beginning "To constrain these slower processes, …"

We further replaced statements in these opening paragraphs about exploiting, resolving, or reducing equifinality with interpretations supported directly by the comparison of the three experiments.

Line 597–599:

"Our results further indicate that incorporating a snapshot of mature biomass information (single-year biomass-constrained experiment) provides an additional constraint on wood residence times but less information on carbon allocation than the biomass–age curve."

Line 617–618:

"Adding the biomass–age curve to the calibration objective penalizes parameter sets that reproduce the high-frequency observations but are inconsistent with that curve."

Comment #36

Line 532: "However, by optimizing parameters against the full accumulation trajectory, DifferLand retrieves effective mean parameters that robustly represent the long-term carbon balance of the ecosystem."

But this isn't a question about using DifferLand as you used that will all the experiments? You are just using more data, so all calibration methods would also produce similar results?

Response #36

Other calibration methods could also use the same biomass–age information. We used DifferLand in all three experiments and did not compare it with other calibration methods**.** We revised the attribution accordingly (**Line 660–663**):

"However, the growth-curve-constrained calibration retrieved effective mean parameters by optimizing against the biomass–age curve. DifferLand provided the differentiable DALEC implementation used to compute gradients through the full simulation and jointly optimize model parameters and initial conditions."

Comment #37

Line 545: "Beyond forest biomass growth curves, additional long-term constraints could further reduce equifinality, particularly for slow soil carbon pools which were not directly constrained in this study."

I would argue one cannot reduce equifinality, either it exists for the system or it doesn't.

Response #37

The revised sentence states (**Line 675–676**):

"Beyond forest biomass growth curves, additional long-term observations could provide complementary information on slow soil carbon pools which were not directly constrained in this study."

Comment #38

Line 569: “The increasing fire frequency (Magney & Pierrat, 2025) and drought-induced mortality (Peng et al., 2011) are already reducing carbon residence times.”

And harvests?

Response #38

We added harvest to the sentence on processes that can shorten carbon residence times (**Line 698–700**):

"Increasing fire frequency (Magney and Pierrat, 2025), harvest, and drought-induced mortality (Peng et al., 2011) can all shorten carbon residence times."

Reference

Magney, T. S. and Pierrat, Z. A.: Shifting equilibria in a warming boreal forest, Proceedings of the National Academy of Sciences, 122, e2424669122, 10.1073/pnas.2424669122, 2025.

Peng, C., Ma, Z., Lei, X., Zhu, Q., Chen, H., Wang, W., Liu, S., Li, W., Fang, X., and Zhou, X.: A drought-induced pervasive increase in tree mortality across Canada's boreal forests, Nature Clim Change, 1, 467-471, 10.1038/nclimate1293, 2011.

Comment #39

Line 576: “This implies that future ESM projections relying on fixed turnover rates may underestimate the risk of a rapid release of stored carbon as these bioclimatic barriers are removed.”

But since your approach produces fixed turnover rates based on the historical growth rate, how would it help with this issue?

Response #39

Our fixed turnover rates do not resolve the issue raised for future projections. We removed that implication from the Discussion (**Line 706–709**):

"The turnover rates retrieved here are fixed in time and describe turnover under the calibration conditions. They do not describe future changes in mortality or disturbance. Representing such changes in projections requires turnover formulations that respond to climate and disturbance, including fire, harvest, drought, and insect outbreaks."
