Reviewer #1 (RC1)

This manuscript addresses an important question in catchment-scale dissolved organic carbon (DOC) modelling: whether acceptable outlet DOC simulations are achieved for the right reasons—namely, through physically plausible internal DOC dynamics. By comparing lumped and landscape-explicit model structures across four headwater catchments, and by contrasting baseline and constrained calibration approaches, the study makes a valuable contribution to the field. The central conclusion is compelling: satisfactory stream DOC performance alone does not guarantee realistic internal behavior, and a landscape-explicit structure paired with internal constraints offers a more robust framework for interpretation and scenario analysis.

Overall, the manuscript is promising; however, several substantial issues must be addressed before it is suitable for publication.

We thank the reviewer for their positive assessment to our work.

Major Comments:

1. The constrained calibration framework requires greater transparency. While the manuscript explains that 100,000 parameter sets were generated, screened against internal constraints, and then filtered for the highest-KGE solution, it remains unclear how restrictive these constraints are and the extent to which they reduce the feasible solution space. I encourage the authors to explicitly report the number of parameter sets that satisfy each individual constraint, as well as all constraints jointly. Furthermore, it would be highly beneficial to better illustrate how the constrained calibration alters the distributions of both parameters and internal states.

Thank you. We will report (in a table) the number of parameter sets that satisfy each individual constraint as well as combinations of constraints. In addition, we will illustrate how the distributions of parameters and internal states change under each individual constraint and under the full set of constraints. We believe this analysis will provide a clearer basis for evaluating the constraints.

2. The delineation of riparian areas necessitates stronger justification. Because the core argument of the paper is build on the explicit separation of upland and riparian units, the conclusions are highly likely to be sensitive to the definition of the riparian extent. Incorporating a brief sensitivity analysis, or at a minimum, a more comprehensive discussion regarding the uncertainties associated with the chosen delineation approach, would significantly strengthen the manuscript.

With the current approach, we found the delineated riparian zones are distributed close to the stream network, matching the location of the actual riparian zone (i.e., groundwater influenced soils along the streams). Nevertheless, in the revised version, we will check how sensitive is the extension of the delineated riparian zone with different topographic wetness indices. We noted that in the current model we do not aim to explicitly represent the exact location of the riparian zone. Rather we conceptualized it as a compartment at the bottom of the hillslope (Fig. 1). Furthermore, our main conclusions (Section 5) are valid for all four basins with different riparian zone areal fractions (Table 1). We will include the aforementioned points in the discussion.

3. The scenario analysis should be interpreted with greater caution. Although the manuscript notes that the experiment focuses solely on increased upland carbon inputs without representing the hydrological alterations associated with forest dieback, this limitation must be stated more prominently. The exercise should be explicitly framed as a structural stress test rather than a comprehensive prediction of forest-dieback impacts.

We thank the reviewer for this important clarification. We agree with the reviewer that the current framing may overstate the scope of the scenario analysis. In the revised manuscript, we will clearly frame this analysis as a structural stress test of the system rather than a comprehensive prediction of forest-dieback impacts, and adjust the wording throughout the manuscript accordingly.

4. The authors should more explicitly acknowledge that while the landscape-explicit structure represents a significant advancement, it is not yet a complete solution. The study demonstrates that baseline outlet calibration can yield unrealistic internal DOC dynamics, and that the landscape-explicit model performs better under constrained calibration and scenario testing. Nevertheless, the discussion also indicates that riparian DOC concentrations may still be underestimated relative to observations, underscoring the need for further process refinement.

Thank you for this insightful comment. We will revise the Discussion and/or Conclusion sections to explicitly acknowledge the reviewer's point that, while the landscape-explicit structure represents a significant advancement, it is not yet a complete solution. In the discussion, we will include in the discussion that the landscape-explicit model (although more complex already than the lumped model) is still a simplified model, but one that can still be used and constrained, while further process refinement is an option, the challenge would be to develop constraints for those again.

Minor Comments:

1. Several formatting and technical errors must be addressed. Notably, the heading "3 Methodology" mistakenly appears at the beginning of the Results section. Additionally, the export equations should be carefully verified for consistency with the conceptual model description.

Thank you for pointing this out. We will carefully check the entire manuscript for these issues. We note that they likely arose during the manual transfer and reformatting of the original manuscript into the HESS format.

2. The reliance on Kling-Gupta Efficiency (KGE) as the primary evaluation criterion should be explicitly justified, particularly given the authors' observation that model ranking is sensitive to the choice of performance metric.

Thank you. We agree that the choice of KGE over other metrics (e.g., NSE, R^2, and BIAS) should be justified. In this study, we used KGE because it simultaneously accounts for correlation, variability, and bias between observed and simulated values (Gupta et al., 2009). In contrast, NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018).

Furthermore, we will present the individual components of KGE (correlation, bias, and variability) separately. This will allow readers to better understand that strong performance in one component does not necessarily imply strong performance in the others, and why KGE is an appropriate metric for evaluating model performance in this case.

3. The manuscript would benefit from a clearer presentation of how many behavioral models satisfy the imposed internal constraints. For instance, the finding that only 53% of behavioral landscape-explicit models meet all constraints under baseline calibration is highly significant and warrants greater emphasis.

Thank you. We will report the number of behavioral models for each individual constraint as well as for combinations of constraints.

4. Several figures are visually dense and should be streamlined. Shifting some of the highly detailed material to the Supplementary Information would enhance overall readability.

We agree that some figures are dense. We will revise them, by adopting alternative plot types or moving selected panels to make the figures look better (e.g., combine boxplot of Fig. 3b together and rearrange them, naming the KGE plot in figures 4 and 5 as subfigures c, d, moving part of figure 6a,b to the supporting information).

5. Given that the study's novelty stems largely from its model implementation and calibration design, making the code and computational workflow openly accessible would substantially elevate the paper's impact and reproducibility.

We will make the underlying data and source codes publicly available by providing an accessible link upon resubmission of the manuscript.