Reviewer #3 (RC3)

This manuscript evaluates if accurate DOC simulations at catchment outlets reliably indicate realistic internal DOC dynamics. By comparing lumped and landscape-explicit models across diverse headwater catchments, the study assesses how model structures and calibration strategies (baseline and constrained) perform under varying conditions. The results demonstrate that good performance at the catchment outlet does not necessarily imply realistic internal process representation, as stream-level agreement can mask substantial internal discrepancies. The analysis further shows that landscape-explicit models incorporating physically motivated internal constraints provide enhanced interpretability and are better suited for scenario analyses. Overall, the study highlights that robust DOC modelling requires an adequate representation of internal dynamics and supports the use of landscape-explicit, internally constrained model structures for catchment-scale applications.

In summary, the manuscript shows clear potential; however, the outcome is somehow expectable. Several issues related to clarity, the linkage between hydrology and DOC dynamics, data uncertainty, and the interpretation of performance metrics need to be addressed:

Thank you for checking our manuscript in detail and the constructive comments. We confirm that the summary provided by the reviewer is accurate and reflects our work. Please find below our point to point response to the reviewer's comment.

1. Although the DOC model description is detailed, a short, high-level overview would benefit readers less familiar with the framework—particularly regarding how hydrological simulation outputs are incorporated into the DOC simulations (e.g. whether and how grid-based data are aggregated).

We will revise the model description (Section 2.1 in the main manuscript), providing a short, high-level, step-wise overview of the DOC and hydrological model and how they are coupled together.

2. The introduction refers to a carbon mass balance. Please clarify whether the model applies a closed carbon mass balance and discuss any implications this assumption may have for the results.

Yes, the model is based on a dissolved carbon mass balance framework, as described in Equations 4–12 of the main manuscript. Under this assumption, carbon cannot arbitrarily appear or disappear in the model to improve the fit to the observed instream DOC concentrations. All carbon influxes and outfluxes associated with each carbon pool (Figure 1c) are explicitly represented in the model. However, some fluxes are weakly constrained because of limited data availability (e.g., the conversion of SOC to DIC). Similarly, changes in carbon storage within different pools are not directly constrained by observations. As a result, the simulated carbon fluxes and pool dynamics remain uncertain and should be interpreted and evaluated with caution before further application. We will add this to the model limitation section.

3. Hydrology plays a key role in the modelling experiments.

The 30 selected simulations exhibit similarly strong hydrological performance, including comparable representations of individual flow components (Figure S3). Expanding the analysis to include simulations with contrasting hydrological behavior—such as improved high-flow performance—could provide additional insight and strengthen the discussion and conclusions, even if overall performance is slightly reduced but still acceptable.

Thank you. Yes, the 30 selected simulations exhibit similarly strong hydrological performance (KGE values for streamflow ranging from 0.7 to over 0.9; Figure 3). Representation of different flow components among these behavioral simulations are comparable (e.g., Hassel catchments with relatively low range of variabilities in groundwater flow, upland flow, and groundwater recharge)

while other catchments show a relatively wide range of variations (e.g., upland flow in the Kalte Bode varies from 55-70%, Figure S3). Furthermore, we will provide additional analyses for simulations exhibiting the model's skill in capturing seasonal dynamics during both high-flow and low-flow periods.

Regarding the suggestion of improving the model performance for high-flow, we provide some possible solutions in the discussion section, such as, using different objective functions that have more weight for high flow, e.g., NSE and RMSE. In this study, we focus more on the overall model performance rather on specific flow conditions.

Please discuss how the underestimation of high flows affects the simulated DOC dynamics.

We will check (e.g., C-Q plot) and add to the discussion if the model underestimates high flows and if this affects the simulated DOC dynamics.

4. Figure 4 indicates an uneven distribution of calibration and validation observations. In particular, Kalte Bode and Warme Bode have comparatively few validation observations at different times. The number of observations per period should be explicitly reported, as this imbalance may influence flow-condition representation and the interpretation of model performance.

Thank you. We will report the number of observations per period (calibration and validation) explicitly in the revised version of the manuscript. We will also discuss in the discussion section if this imbalance could potentially affect the model performance for the validation period.

5. Please address the uncertainty associated with the DOC measurements. Could higher observation frequency or event-based (e.g. daily) DOC observations potentially improve model performance.

In the Kalte Bode and Warme Bode catchments, additional observations, particularly during high-flow events, would help balance the number of measurements between high- and low-flow conditions and could potentially improve model performance. This is likely because the observed DOC concentrations in the Kalte Bode and Warme Bode show a much wider range (approximately 2–19 mg /L) than those in the Rappbode and Hassel catchments (approximately 2–13 mg/L; Figure 4) and more observation is needed to capture instream DOC dynamics if they vary in a wider range.

6. The Kling–Gupta Efficiency (KGE) is used as the primary performance metric, integrating correlation, variability, and bias between simulated and observed DOC concentrations. Given that these components can contribute differently to the overall KGE—as suggested by NSE, R^2 and bias in the supplementary material—further discussion of their respective roles, as well as the strengths and limitations of KGE in this context, would be valuable. In particular, it would be informative to assess whether one component dominates KGE and how this relates to different flow conditions and associated DOC dynamics.

Thank you for the detailed suggestion. Yes, the KGE considers correlation, variability, and bias between simulated and observed (DOC concentrations, also streamflow) and each component could contribute differently to the KGE. We will detail each component (correlation, variability, and bias) in the supporting information. We will discuss the role of each component and check if one component dominates KGE and if it relates to different flow and DOC dynamics among catchments. Based on that, we will evaluate the strength and limitations of the KGE.

Minor Comments:

- Figures S4–S6: Please clarify in the captions whether the results refer to the calibration period, the validation period, or both.

In figure S4-S6, we showed results for the calibration and validation separately, as indicated by the boxplot color and shown in the figure legend. In the caption, we will indicate it again.