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. We have carefully considered all comments and have substantially revised the manuscript to address the issues raised.

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 have reported the number of parameter sets that satisfy each individual constraint as well as combinations of constraints in Table 2 (also shown below):

Table 2. Percentage of simulations satisfying each individual constraint and combinations of constraints among the 30 × 100,000 simulations performed for each catchment using randomly generated parameter sets. Constraints 1, 2, and 3 correspond to groundwater DOC concentration ($0.25 < C_{GW} < 2.66$ mg/L), the relationship between upland and riparian zone DOC concentrations ($C_{UL} < 0.5C_{RZ}$), and the relationship between C transformation rate in the upland and riparian zone ($k_{x\rightarrow y, UL} < k_{x\rightarrow y, RZ}$), respectively (Section 2.4).

Model	Lumped	Landscape-explicit						
Constraint number	1	1	2	3	1 & 2	1 & 3	2 & 3	1 & 2 & 3
Hassel	5.9	5.6	2.8	68.1	3.7	3.7	1.8	0.8
Kalte Bode	13.7	7.8	2.0	59.0	4.4	4.4	1.1	0.4
Warme Bode	14.5	9.3	2.5	55.6	5.0	5.0	1.3	0.6
Rappbode	13.4	9.2	2.5	60.9	5.4	5.4	1.4	0.7

We have modified the main text to reflect this aspect in the revised manuscript (L401-407)

> "It is seen that the effect of each constraint depends on the model structure, catchment, and the type of constraint applied (Table 2). For the same constraint (e.g., constraint 1 for groundwater DOC concentration), a smaller percentage of satisfied simulations with the landscape-explicit model than with the lumped model indicates greater uncertainty in the simulated internal DOC dynamics of the more complex model structure. Among the three constraints applied on the landscape-explicit model structure, constraint 2, which represents the relationship between upland and riparian-zone DOC concentrations, is the most restrictive. This highlights the importance of realistically representing spatial DOC gradients between upland and riparian zones."

Below we demonstrated how the constraints affect parameter distribution Figures S10 and S11 (as shown below). We report the case with all combined constraints as there are many parameters and combinations of different constraints. It is seen that the distribution of most parameters does not change much (Figs. S10 and S11). However, some parameters, especially the rate conversion from SOC to DOC in the upland or hillslope, show quite different distribution compared to its initial distribution, mentioned in L407-409:

> "However, a few parameters show substantial shifts from their initial distributions, particularly those controlling the rate of SOC to DOC conversion in the upland or hillslope."

We note, however, that these boxplots are univariate representations of the distribution of each parameter. They show how the distribution of each individual parameter changes after applying the constraints, but they do not describe the full joint distribution among multiple parameters. We recognize this aspect as well and interpret these plots as diagnostic summaries of marginal distributional shifts, rather than as a complete characterization of constrained parameter space. We mentioned it in the figure caption.

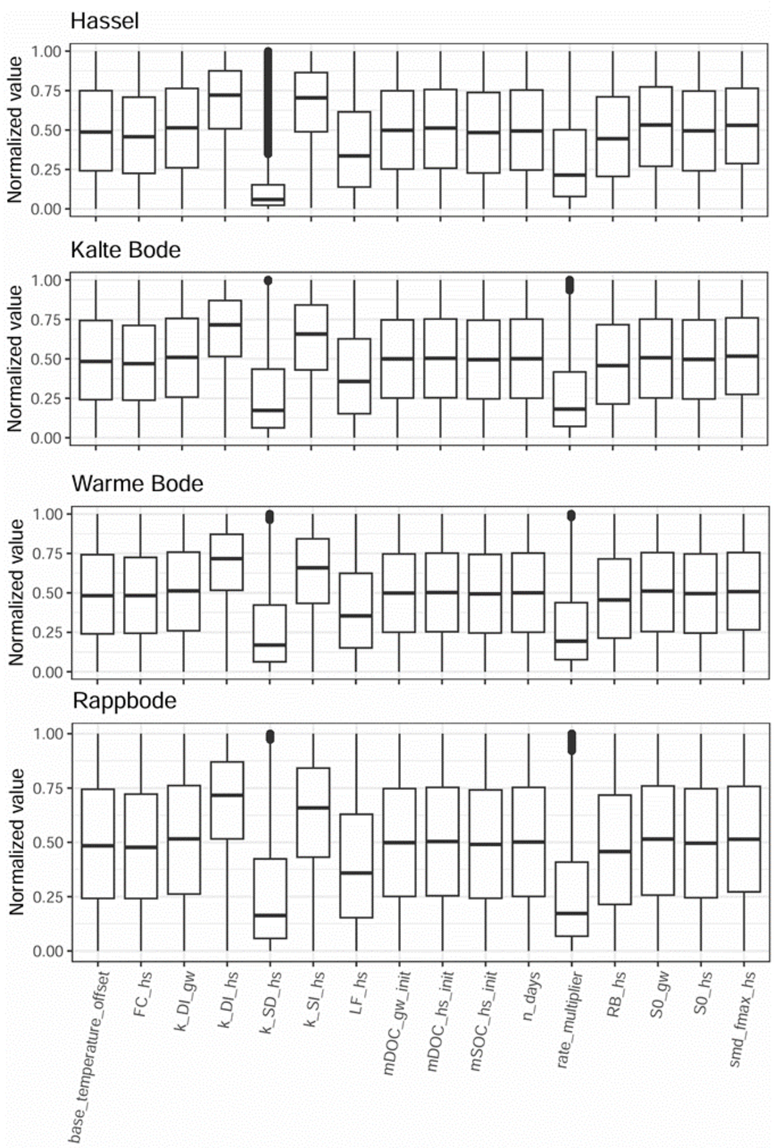

Figure S10: Boxplots of the parameters of the lumped model structure satisfying constraints 1. For visualization, each parameter range was normalized such that the minimum and maximum values correspond to 0 and 1, respectively. Therefore, the initial parameter distribution is centered at 0.5, with the first and third quartiles at 0.25 and 0.75, respectively. The parameter description, including the actual parameter ranges, is provided in the supporting document (doc_parameter.txt). These boxplots are univariate presentation the distribution of each parameter rather than a joint distribution among multiple parameters.

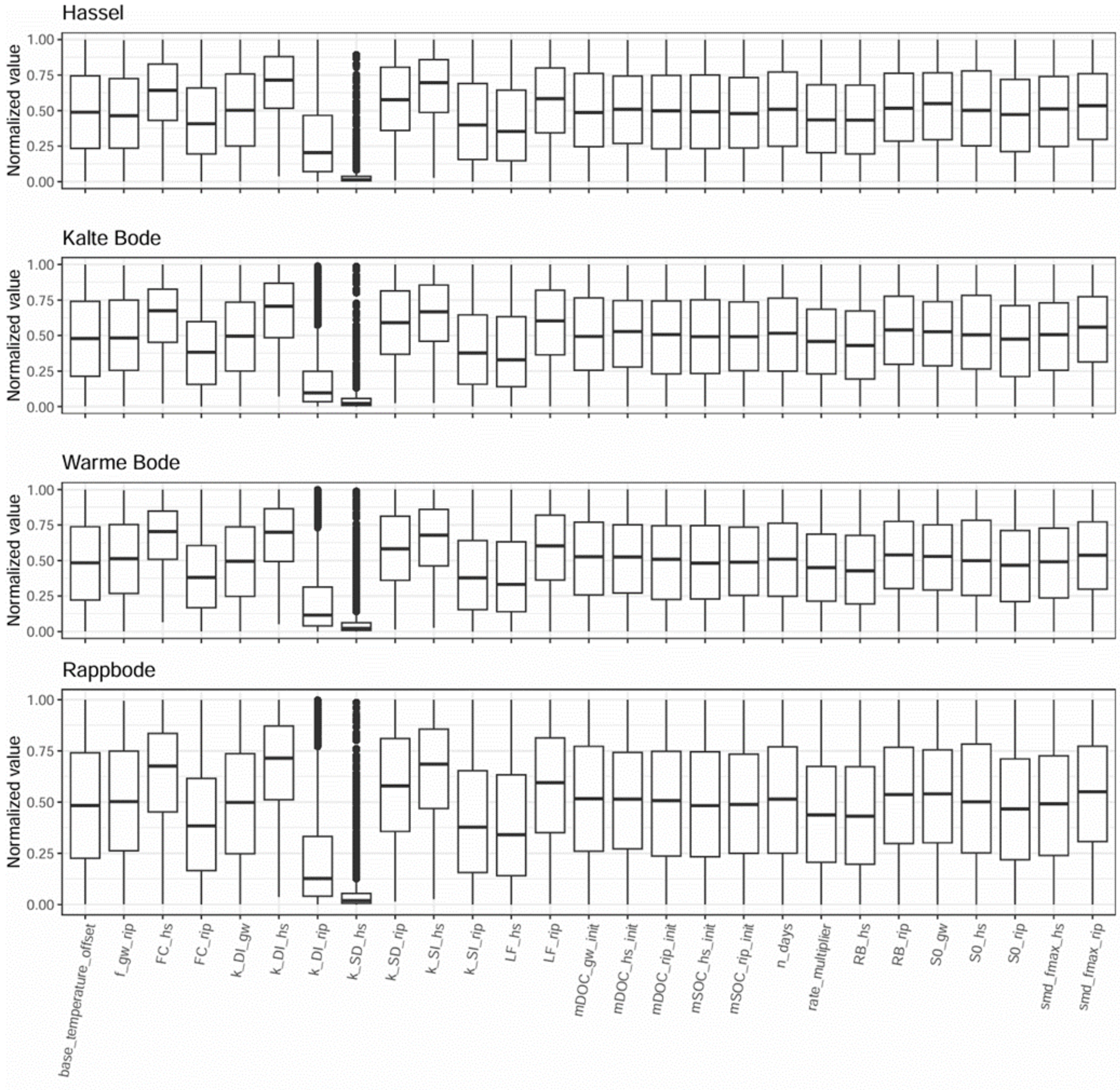

Figure S11: Boxplots of the parameters of the landscape-explicit model structure satisfying constraints 1, 2, and 3. For visualization, each parameter range was normalized such that the minimum and maximum values correspond to 0 and 1, respectively. Therefore, the initial parameter distribution is centered at 0.5, with the first and third quartiles at 0.25 and 0.75. The parameter description, including the actual parameter ranges, is provided in the supporting document (doc_parameter.txt). These boxplots are univariate presentation the distribution of each parameter rather than a joint distribution among multiple parameters.

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.

We agree that the delineation of riparian areas can be a source of uncertainty, particularly because the landscape-explicit model structure relies on the separation of upland and riparian units. In the revised manuscript, we clarified that the TWI threshold was selected to match the mapped extent of groundwater-dominated gley soils in the Rappbode catchment and that this delineation should be interpreted as a hydrologically meaningful near-stream landscape unit rather than an exact boundary of the riparian zone. To evaluate whether our main conclusions depend on the selected threshold, we added a sensitivity test in the Rappbode catchment in which the riparian area was reduced and increased by 25% relative to the reference delineation by adjusting the TWI threshold. The landscape-explicit model was then recalibrated using the same calibration procedure. We also rerun the constrained calibration and scenario analysis. In general, a wider range of model performance for stream DOC simulation was found, but the main findings (Section 5) remained unchanged.

We have revised the main text to reflect these aforementioned points and showed detailed results in the supporting information, a copied of those changes were shown below:

L202-209 (study area and data):

> "The area fraction of the riparian zone in the four catchments varies from 8.7 to 19.1% (Table 1). These riparian zone proportions were calculated using a threshold value of the topographic wetness index (TWI) derived from a 25 m European Digital Elevation Model (EU-DEM). The riparian zone was defined as the combined area with TWI above a specific threshold. A TWI threshold was previously shown to be a robust measure of the riparian zones within catchments in Germany (Musolff et al., 2018). Here, the reference threshold (TWI ≈ 6.8) was selected to match the areal extent of groundwater-dominated gley soils mapped in the Rappbode catchment (Werner et al., 2019) and then applied consistently across the four catchments. Because TWI-based delineation is a proxy for riparian extent rather than an exact hydropedological boundary, we evaluated the sensitivity of the main conclusions to this choice using an additional test in the Rappbode catchment (Section 4.4). "

L554-557 (limitation section):

> "In addition, we found that the extent of the riparian zone is sensitive to the defined TWI threshold, meaning that more precise information about the extent of the riparian zone is needed to constrain results based on the TWI. However, we tested for the Rappbode catchment that changing the riparian extent by ±25% led to wider range of model performance for stream DOC simulation, but the main findings (Section 5) remained unchanged."

Text S3. Effect of uncertainty in riparian zone delineation on the main findings

In the main manuscript, we acknowledge that the delineated riparian area is sensitive to the defined topographic wetness index (TWI). Here, we tested whether this uncertainty affects the main conclusions of the study. We selected the Rappbode, where the TWI threshold was defined and applied to the others. We adjusted the TWI threshold to generate riparian extents that were 25% smaller and 25% larger than the reference riparian extent. This changed the riparian area fraction of the Rappbode from 8.7% in the reference case to 6.5% and 10.9%, respectively. We then reran the landscape-explicit model structure only for the Rappbode with these new riparian zone fractions (6.5% and 10.9%) as examples, using the same baseline calibration, constrained calibration, and scenario simulation settings to test whether the main findings reported in Section 5 were maintained.

Finding 1: "Under the baseline calibration, both lumped and landscape-explicit model structures can reasonably represent stream DOC concentration dynamics, with none consistently outperforming the other across catchments and evaluation metrics. However, the simulated DOC concentrations in groundwater for both model structures were significantly higher than those from observations, indicating unrealistic internal DOC dynamics."

> Compared to the baseline simulation with a fixed value of the riparian zone (Fig. 4d), the KGE ranges are wider (lower minimum and higher maximum values; Fig. S13a). Some behavioral simulations even have higher KGE values, indicating that our conclusion that "landscape-explicit model structures can reasonably represent stream DOC concentration dynamics" is still valid. In line with previous findings, we found that while the lumped model (Fig. 4c) is worse than the landscape-explicit model structure during the calibration period, it is better during the validation period, showing that no model structure consistently outperforms the others. We also found that the performance varies across evaluation metrics. For example, the

KGE shows only a slight difference between the calibration and validation periods, while the NSE (Fig. S13b) shows clearly higher performance in the calibration period than in the validation period.

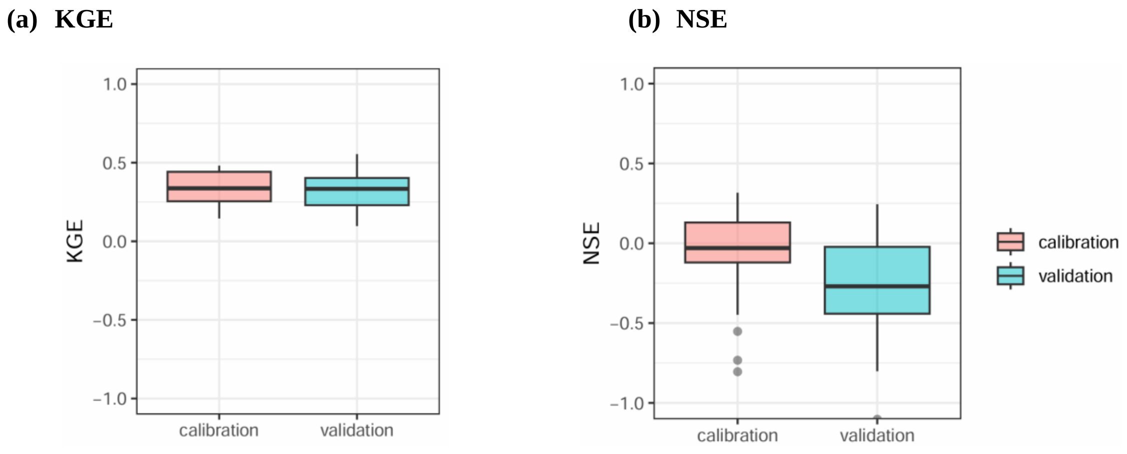

Figure S13. Boxplots of the (a) KGE and (b) NSE for stream DOC concentrations simulation for the calibration and validation periods of the Rappbode when decreasing and increasing the riparian zone area by 25%. The resulting riparian zone fractions are 6.5% and 10.9%.

Finding 2: "The constrained calibration (accounting for physically realistic DOC concentrations in the upland and groundwater compartments) slightly reduces model performance for stream DOC concentration simulation for the landscape-explicit model structure, while increasing the physical realism of the model in terms of internal DOC dynamics. In contrast, the performance of the lumped model structure is significantly reduced under the constrained calibration."

Here, we also found that, with the landscape-explicit model structure, the model performance for stream DOC concentration is slightly reduced under the constrained calibration compared to the baseline calibration. Furthermore, the model performance spreads over a wider range (wider interquartile range) when increasing and decreasing the riparian zone.

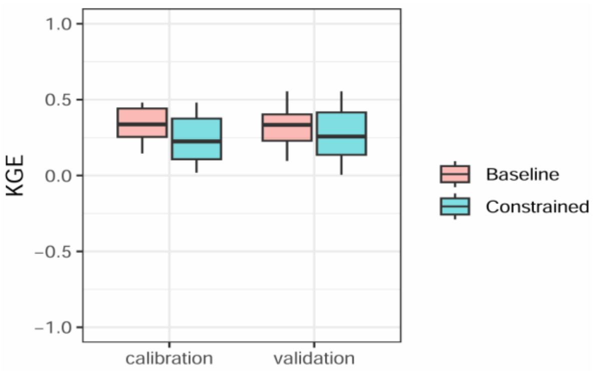

Figure S14. Boxplots of the KGE for stream DOC concentrations simulation for the baseline and constrained calibration of the Rappbode when decreasing and increasing the riparian zone area by 25%. The resulting riparian zone fractions are 6.5% and 10.9%.

Finding 3: "For evaluating the effectiveness of spatial management on stream DOC (e.g., changes that do not occur homogeneously over the entire catchment, such as increasing C input in the upland in this study), a landscape-explicit model structure provides credible results, while a lumped structure does not."

Consistent with the previous findings, after rerunning the scenario simulation with the new riparian zone fractions (6.5% and 10.9%), we found that increasing C input in the upland has no or only a very minor effect on stream DOC concentration (Fig. S15).

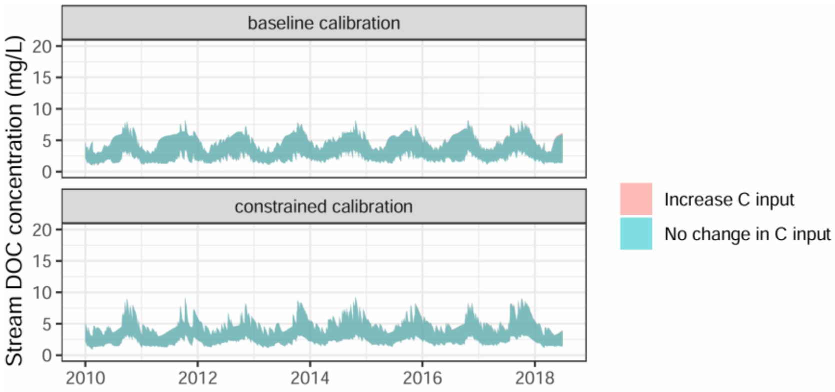

Figure S15. Effect of increasing C input (during the period 2015-2018) in the upland of the lumped model structure on simulated stream DOC concentrations in the Rappbode catchment. The results with the riparian zone fractions are 6.5% and 10.9%.

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. Following the reviewer suggestion, we have revised part of the text in the Introduction and especially the scenario description to carefully describe the scenarios of forest dieback as a structural test rather than the comprehensive prediction of (forest-dieback) impact:

Lines 96-98:

"Furthermore, we conducted a structural stress test in which the two model structures (lumped and landscape-explicit) were subjected to changing boundary conditions to assess the robustness and plausibility of their simulated stream DOC responses."

Lines 298-300:

"We tested whether the lumped and landscape-explicit model structures could simulate stream DOC concentrations within a reasonable range under a structural stress test where boundary conditions were changed."

Lines 304-306:

"For example, Musolff et al. (2024) suggested that an increase in C availability in the upland (e.g., from forest dieback) will not lead to a significant increase in stream DOC concentrations due to the lack of direct hydrological connectivity."

Lines 311-313:

"We note that changes in DOC inputs are typically associated with changes in hydrological conditions. However, no changes were made to the hydrological simulations in this stress tess, as the objective was to isolate the effect of C inputs on stream DOC responses and enable a clearer assessment of their influence."

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 have revised the discussion to explicitly acknowledge that the landscape explicit model structure, while showing a significant advancement, is not yet a complete solution as the internal DOC dynamics are complex and may additionally modulate DOC fluxes and export.

Lines 540-542:

> "This is consistent with our understanding and experience, in which we found that although a landscape-explicit model structure represents a significant advancement in model structure, it does not necessarily yield physically plausible internal DOC dynamics. "

Yes, in the discussion we already indicated that riparian DOC concentrations may still be underestimated relative to observations, underscoring the need for further process refinement:

Lines 500-504:

> "Indeed, there is evidence from the Rappbode catchment that riparian zone DOC concentrations can be higher (observed up to 16 mg/L, Ledesma et al. 2025) than modelled here (range 2.5 to 9 mg/L, Fig. 7). Here, future model refinements may allow for a higher concentration range in the riparian zone, for example, by accounting for vertical heterogeneity in water flows and the associated depth-dependent DOC mobilization from riparian zones as described by Seibert et al. (2009) or Ledesma et al. (2018b). "

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 acknowledge that these issues likely arose during the manual transfer and reformatting of the original manuscript into the HESS template.

We have revised the section heading to **"3. Results"** and also conducted a thorough review of the entire manuscript to address formatting and technical inconsistencies. For example, we removed duplicated DOI prefixes (e.g., repeated "https://doi.org/"), corrected reference details (e.g., the publication year for Salmon-Monviola et al., 2025), and standardized unit formatting (e.g., km^2).

In addition, we carefully reviewed all equations and model descriptions to ensure consistency throughout the manuscript. We corrected Equations 6, 9, 11, and 12

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, other measures like NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018).

In the revised version, we also presented individual components of KGE (correlation, bias, and variability; Figs. S6-S9) 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 could be an appropriate metric for evaluating the overall model performance in this case.

We have modified the main text to reflect the above points (L240-250):

> “The objective function used for calibrating the hydrological model (streamflow simulation) and DOC model (stream DOC simulation) is the Kling–Gupta efficiency (KGE; Gupta et al., 2009) expressed as follows:
>
> $$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \quad (14)$$
>
> where r is the linear correlation coefficient between simulated and observed values (streamflow or stream DOC concentration), α is the ratio between the standard deviation of the simulated and the standard deviation of the observed values (hereafter referred to as variability), and β is the ratio between the mean simulated and mean observed values (hereafter referred to as bias). In addition, we also showed different components of the KGE (correlation r, variability α, and bias β) and the commonly used Nash–Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970; Text S2) to assess the performance of the calibrated models. We selected KGE as our primary evaluation criterion as it balances between correlation, variability, and bias between observed and simulated (Gupta et al., 2009). In contrast, NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018).”

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. As described above, we have now reported the number of behavioral models for each individual constraint as well as for combinations of constraints (see our response to the main comments).

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 have revised 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, splitting Figure 6a, b to Figures 6 and 7).

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 submitted the source code together with this revised manuscript to HESS. We are currently reviewing the data policy to determine whether we are permitted to publish data from these catchments. If permission is granted, we will also publish the dataset from these catchments in an open repository.

Reviewer #2 (RC2)

The present manuscript (MS) compares the performance of a lumped and a landscape-explicit version of a conceptual DOC catchment model. While both versions could satisfactorily simulate streamflow DOC, low groundwater DOC and increased C-input could only be simulated by the more complex, landscape-explicit model version. The paper is nicely written and gets to the point that a more complex model structure and additional information, i.e. low groundwater DOC, adds to the realism of model simulations.

We thank the reviewer for positive and encouraging assessment of our manuscript.

However, these results seem expectable at first sight and so far the contributions to general knowledge which go beyond this particular model and beyond this particular application to four specific German basins seem rather limited. Eventually, a more detailed representation of DOC formation, but for sure a more detailed data analysis are warranted to justify the model structure: is the presented model really a valid compromise between required complexity and computational efficiency and therefore might be regarded as a straight-forward tool for practitioners? As follows three major concerns are listed.

Thank you. We agree that some results may appear expected from a process-based perspective. However, this consistency is important because it shows that the model behaviour is in line with our understanding of system functioning, which is supported by field evidence.

We started with our specific models and catchments as a demonstrative example. While we acknowledge a certain uniqueness of place, we would like to emphasize that these are representative catchments for Central European landscape settings and also part of initiatives (see TERENO; https://www.tereno.net/) in which we compare their riparian wetland state and functioning to other established experimental catchments. More specifically we refer to Ledesma et al. (2025) and to the data newly provided in Trojahn et al. (2026). We embed these references into the manuscript to make more clear how the catchments are positioned in the international research on riparian zone functioning, L91-94:

> "We then applied the proposed models in four intensively studied catchments in the Harz Mountains, Germany as a demonstrative example. While the selected four catchments are located in a specific region, they are part of the initiative in which we compare the riparian state and functioning to other established experimental catchments (Ledesma, et al., 2025; Trojahn et al., 2026)."

Regarding "a more detailed representation of DOC formation, but for sure a more detailed data analysis is warranted to justify the model structure". This is an important point and we fully acknowledge; and have put our thoughts in the discussion (L533-536):

> "We argue for using only model structures that can be justified (Beven and Freer, 2001). In line with the arguments presented by Beven and Freer (2001), this justification should be based on observations or qualitative understanding of the catchment processes as demonstrated in this study, even when the overall performance of the justifiable models might be lower."

Regarding the "compromise between required complexity and computational efficiency", we noted that the computational efficiency is not an issue as our model is a conceptual model with only few model compartments (not a gridded model), even increasing more DOC processes. The main thing to consider is the data to evaluate the validity of the model complexity (as indicated in the above statement). Here we show that even a lumped model is too complex if we only calibrate the model with stream DOC concentration (still there is significant uncertainty in groundwater DOC - as shown in Figure 6).

Please find below our responses to the three major concerns

1. Simplified representation of DOC origin

The conceptualization of carbon transformation follows the well-known INCA-C model by (Futter et al., 2007) but is even simplified by single net transformation between the soil organic carbon pool and DOC. While the authors state that this reduces equifinality, this very simple assumption combines various individual multicomponent reactions. An existing example in HESS shows that more physical basis is possible with reasonable computational effort (Wen et al. 2020, https://doi.org/10.5194/hess-24-945-2020). Those authors incorporated coupled elemental cycling, stoichiometry, thermodynamics and kinetics in their DOC catchment model already 6 years ago. This means that a more sophisticated representation of DOC formation is possible and might be warranted also in the present model. If the authors decide to maintain their simplified approach, this needs strong justification and more detailed checks.

We agree with the reviewer that the current models used a simplified representation of C transformation, e.g., single net transformation between soil organic carbon pool and DOC. We have checked the paper by Wen et al. 2020 and found that they also model the net transformation of SOC to DOC, considering the impact of soil moisture and temperature similar to the presented approach. However, they have an additional process "DOC produced from SOC can also sorb on soils". We did not consider the sorption of DOC on soil as implementing such processes requires data to constrain internal transformation pathways and associated parameters, which are not available for the present study and even harder to find in other, less monitored, catchments. The net transformation between soil organic carbon and DOC can be taken as choice from the practical viewpoint as well as from optimal model complexity considerations; i.e., it avoids introducing poorly constrained processes that could increase equifinality without improving model robustness. We have mentioned those points in the model description (L141-144) and discussion (L537-544) as a judgment for our simplified representation of DOC.

L141-144:

> "We represent the net transformation between SOC and DOC using an effective transformation rate ($k_{SOC \rightarrow DOC,z}(t)$) instead of representing the transformation of DOC to SOC and SOC to DOC explicitly, like the INCA-C model, or sorption of DOC on soil (e.g., Wen et al., 2020). This simplification assumes a net transfer of SOC to DOC and aims to reduce equifinality as data to evaluate the internal transformation pathways and its associate parameter is typically not available, at least in our study area (Section 2.2)."

L537-544:

> "With more complex models, e.g., SWAT-Carbon (Zhang et al., 2013), INCA-C (Futter et al., 2007), HYPE (Pers et al., 2016); BioRT-Flux-PIHM (Wen et al., 2020), the available data may be insufficient to constrain all model parameters. Still, the models can exhibit substantial equifinality (i.e., multiple parameter sets may produce equally acceptable simulations), resulting in a wide range of potential solutions, some of which may be unrealistic. This is consistent with our own experience, in which we found that although a landscape-explicit model structure represents an advancement in model structure, it may not always yield physically plausible internal DOC dynamics. Therefore, we recommend that modelers critically evaluate which model outputs can be used, prioritizing those that have been evaluated by direct observations or corroborated by expert knowledge."

2. Missing carbon mass balance

In the introduction, the authors criticize the model by Birkel et al. 2014 for not closing carbon mass balances. But in their own model application, they only study runoff and DOC concentrations separately, but do not calculate DOC loads or carbon mass fluxes. This would also be valuable for their "scenario" of forest dieback: Is their landscape-explicit model version really superior when it comes to a realistic carbon mass balance? By the way, also for drinking water reservoirs, realistic DOC loads are more important than reproducing DOC concentration patterns.

We would like to highlight that the carbon mass balance (carbon mass fluxes) is indeed close in our case for each model compartment and C pool (Equations (5) - (12)). The concentration was not shown in these equations, rather they are derived from their mass fluxes divided by the hydrological fluxes, for example in L179-184 show how to convert mass flux to concentration (as shown below):

> "Stream DOC concentration (mg L^{-1}) from both model structures can be derived from the C mass and hydrologic fluxes, as follows:
>
> $$C_{DOC,STREAM}(t) = 100 \cdot \frac{m_{DOC,STREAM}(t)}{Q_{STREAM}(t)}, \tag{13}$$
>
> where Q_{STREAM} (mm day^{-1}) is streamflow at time catchment outlet, 100 is the conversion unit from kg ha^{-1} mm^{-1} to mg L^{-1}."

In our viewpoint a good model performance for both DOC concentration and load is important. Having a good model performance for DOC concentration is more challenging than DOC load. This is because if the hydrological model for streamflow performs well (as is the case here, with median KGE values for streamflow ranging from 0.7 to over 0.9, Figure 3), the simulated DOC load is also likely to be good, even with nearly-constant DOC concentration. Therefore, the performance of the DOC model should be evaluated for their simulated concentration rather than load. This also aligns with the principle of obtaining the "right answers for the right reasons". Since all models are ultimately mathematical abstractions of reality, model evaluation should target the faithful representation of the controlling processes, rather than relying on potentially compensating effects that may still produce acceptable load estimates. We have revised L250-253 to reflect this:

> "We noted that calibration to stream DOC concentration is preferable to calibration to DOC load because DOC load is the product of streamflow and concentration. As a result, good performance for streamflow simulations can compensate for errors in simulated DOC concentrations, still producing acceptable load estimates. In contrast, calibrating to stream DOC concentration provides a more direct test of whether the model correctly represents the processes controlling DOC dynamics."

3. DOC-export during different flow conditions

The authors only test their models by overall performance during a multi-year period. They do not study DOC simulations during different flow conditions. Again the study of Wen et al., 2020 might be relevant here. Their study showed that DOC-stocks are formed during periods of high temperature, while the main DOC export occurs during following wet conditions or during high magnitude runoff events. They also showed that streamflow DOC equals groundwater DOC during dry conditions, riparian zone DOC during intermediate conditions, and hillslope DOC during real wet conditions. These results could easily be checked or challenged by the present application. This might also prove the physical soundness and robustness of the present approach. Potentially, a shortcut of riparian buffer zones during high magnitude events might also play a role which is not included in the present model so far. Especially in the "Warme Bode", the landscape-explicit model fails to predict single DOC peaks. Those are particularly evident following the 2015 European drought which was characterized by high temperatures and below average rainfall. The different behavior between of the four basins and how an improved model adequately represents this could be another test here.

Thank you for the suggestion regarding further evaluation of the simulated stream DOC dynamics under different conditions. In this study, we provided a general evaluation (KGE) rather than focusing on a specific period (e.g., low or high flows). This aligns with our main quest to determine whether the model has reasonable performance for the overall stream DOC simulation (represented by the KGE) and has reasonable internal DOC dynamics. We noted that while other objectives and research questions are also valid, it is not within the scope of this study and could be done in future studies. Nevertheless, we put these detail in the discussion so that other studies could evaluate the model performance for different flow conditions, and the relative magnitude of DOC concentration from different model compartments compared to stream DOC concentration during different periods (high versus low flows) if these variables are of interest and observed data are available.

L546-550:

> "The calibration constraints implemented in this study were restricted to those that could be supported by available observational data and by our perceptual model of relative DOC concentrations among landscape units. Additional constraints, e.g., the relative magnitude of DOC concentrations in the upland, groundwater, and riparian zone compared to stream DOC concentration during different flow conditions could be used if such information is available (e.g., Wen et al., 2020). We did not conduct a separate evaluation of the simulated stream DOC under different flow conditions. Our analysis focused on whether the alternative model structures reproduced overall stream DOC dynamics and whether their internal DOC dynamics were consistent with available observations and process understanding. An extensive evaluation of model performance under different flow conditions is required if the focus is on event DOC dynamics and understanding the model's strengths and limitations under such conditions."

In the Wame Bode and other catchments, the observed peak DOC concentrations events were underestimated. This might require a more vertically refined representation of the riparian zone, e.g., through multiple soil layers, could account for depth-dependent DOC concentrations and for changes in the riparian source layer contributing to stream DOC under variable flow conditions. This could potentially improve the simulation of high DOC concentrations, especially for the Kalte Bode and Warme Bode catchments (Fig. 5b), as well as the representation of groundwater DOC. We added these points in the limitation section.

L558-561:

> "Furthermore, a more vertically refined representation of the riparian zone, e.g., through multiple soil layers, could account for depth-dependent DOC concentrations and for changes in the riparian source layer contributing to stream DOC under variable flow conditions. This could potentially improve the simulation of high DOC concentrations, especially for the Kalte Bode and Warme Bode catchments (Fig. 5b), as well as the representation of groundwater DOC concentrations. "

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 provided a short, high-level summary of the DOC model in the first paragraph of section 2.1, as shown below:

L102-107:

> "The DOC model developed in this study is a catchment-scale model operating at a daily timestep (Fig. 1). The DOC model uses hydrological states and fluxes from the mesoscale Hydrologic Model (mHM; Kumar et al., 2013; Samaniego et al., 2010). As the DOC and mHM models operate at different spatial scales, hydrological fluxes and states from mHM were first aggregated from the grid scale to the catchment scale using area-weighted averaging and then disaggregated, also using area-weighted allocation, to the landscape units of the DOC model (Text S1). The DOC model simulates carbon (C) transformation and balance for each model compartment and C pools. A more detailed description of the DOC model is provided below."

In the subsequent paragraph (L108-184) and in the supporting text S1, we detailed the model description, including how grid-based data from mHM were aggregated to the spatial units of the DOC model.

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, our model was based on the mass balance C for each C pool and model compartment, as shown in Equation 4 to 12 in our main manuscript. In other words, all C influxes and outfluxes associated with each C pool (Figure 1c) were explicitly represented in the model. However, only the dissolved organic carbon (DOC) fluxes were constrained through calibration against observed instream DOC concentrations. Other fluxes, such as the conversion of SOC to DIC and DOC to DIC, as well as the SOC and DIC storages, were not directly constrained and were only indirectly constrained through DOC calibration. Consequently, these fluxes and storage estimates remained subject to considerable uncertainty. Future studies could evaluate the validity of these simulated fluxes and C storage estimates as additional observational data become available. We added a limitation section with this aspect.

L568-573:

> "Our model was based on the mass balance of C for each C pool and model compartment. In other words, all C influxes and outfluxes associated with each C pool (Fig. 1c) were explicitly represented in the model. However, only the DOC fluxes were constrained through calibration against observed stream DOC concentrations. Other simulated C fluxes, such as the conversion of SOC to DIC and DOC to DIC, as well as the SOC and DIC storage estimates, were not directly constrained by observations. These fluxes and storages should therefore be interpreted with caution. Future studies could evaluate them more directly as additional observational data on C pools and transformation fluxes become available."

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).

To evaluate the model performance under different flow conditions, we show scatter plots of observed and simulated high flows (Q > Q10: the streamflow that is equaled or exceeded 10% of the time) and low flows (Q < Q90: the streamflow that is equaled or exceeded 90% of the time). Q10 and Q90 are defined from observed data. Our results show that there is no systematic under- or overestimation of high flows (Fig. S3). Similar findings were observed for low flows in the Kalte Bode and Warme Bode. However, in the Rappbode and Hassel, the model systematically overestimates low flows. We have modified the main text regarding these points.

L320-322:

> "There was no systematic under- or over-estimation of high flows across all catchments or low flows in the Kalte Bode and Warme Bode catchments (Figs. 3a and S3). However, low flows were systematically overestimated in the Rappbode and Hassel catchment (Fig. S3b)."

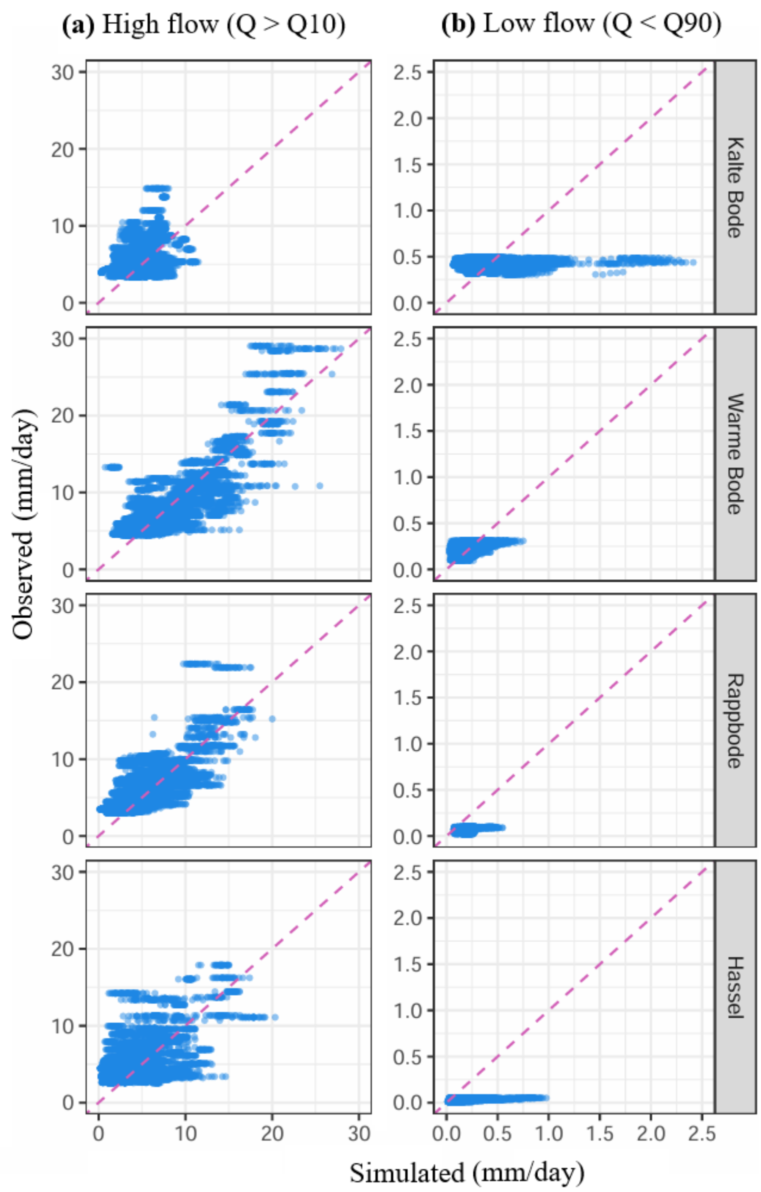

Figure S3. Simulated and observed high flows (Q > Q10: the streamflow that is equalled or exceeded 10% of the time) and low flows (Q < Q90: the streamflow that is equalled or exceeded 90% of the time). Q10 and Q90 are defined from observed data for each respective basin. Data are from both calibration and validation period.

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

In the revised manuscript, after conducting a more thorough check, we found that the model does not systematically under or over-estimate high flows (as discussed above).

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 reported the number of observations per period (calibration and validation) explicitly in the revised version of the manuscript.

Table S1. Number of DOC observations in calibration and validation periods.

Basins	Kalte_Bode	Warme_Bode	Rappbode	Hassel
Calibration period	97	94	91	88
Validation period	34	33	119	127

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 stream DOC observations, particularly during high DOC concentration events, would help balance the number of measurements under different DOC export conditions and could potentially improve model performance. This is likely because the observed DOC concentrations in the Warme Bode show a much wider range than other catchments, and more observations are needed to capture instream DOC dynamics when they vary over a wider range. We added this aspect in the revised manuscript.

L561-567:

> "Additional stream DOC observations, particularly during high DOC concentration events, would also help to better characterize DOC export across the full range of hydrological and biogeochemical conditions. This is especially relevant for Warme Bode, where observed stream DOC concentrations span a wider range than others (Fig. 4). Under such conditions, sparse weekly or biweekly sampling may miss short-lived DOC peaks and therefore provide limited information for evaluating event-scale model behaviour. High-frequency DOC measurements or suitable optical proxies, would help to constrain these dynamics more directly and could improve the representation of high-DOC events in the model."

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 showed each component (correlation, variability, and bias) in the supporting information (Figs. S6-S8).

For streamflow, we found that the role of these components to the overall KGE values (represented by its deviation from 1) differs. While these components are close to 1 for the Warme Bode and Rappbode, they deviate more substantially (within a range of 0.2) from 1 for the Kalte Bode and Hassel (Figs. S5–S8). We added this to the manuscript (L346-349):

> "It is also seen that, depending on the catchment, the contribution of each component of the KGE to its overall value (represented by its deviation from 1) differs. While these components are close to 1 for the Warme Bode and Rappbode, they deviate more substantially (within a range of 0.2) from 1 for the Kalte Bode and Hassel (Figs. S6–S8)"

For DOC, we found that the roles of these components are not the same to the overall KGE, and the model ranking varies depending of the performance metrics:

L371-375:

> "We note that the ranking of the model performance among catchments and periods is sensitive to the performance metrics. More specifically, the NSE, correlation bias, and variability (Figs. S5-S8) do not show the same trend across the four catchments, calibration, and validation periods as that of the KGE (Fig. 4). Results also show that the contribution of different KGE components to its overall value differs (Figs. S6-S8)."

Although going into a more detailed analysis of the KGE and its components as well as the roles of each component under different flow conditions and associated DOC dynamics is interesting, this is beyond the scope of this study. Therefore, we did not conduct such analyses.

Minor Comments:

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

In figures S4-S6 (which have become figures S5-S8), 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 now clearly indicated this aspect as suggested by the reviewer.

Reviewer #4 (RC4)

General comments

The following manuscript compares two modeling approaches – one that combines all landscape units within a single catchment (i.e., lumped) and one that treats landscape units independently (i.e., landscape-explicit) – for predicting dissolved organic carbon (DOC) transport from upland areas into streams. Previous such modeling efforts have performed model calibration using only in-stream DOC concentrations; here, the authors additionally perform model calibration using interior (i.e., groundwater) DOC concentrations. Overall, this manuscript does contribute to future modeling efforts, but in order to be most useful, the authors would do well to provide (1) additional rationale for their model validation approaches and (2) explicit recommendations for those interested in applying this approach in other catchments, since they note the challenges inherent in using this model in different systems.

Thank you for the summary and feedback. Please find our point-to-point responses below.

Specific comments

- The authors should provide more clarity in the methods to address how they reconciled using weekly to biweekly observations of DOC alongside daily estimated values. Was there any interpolation or other methods used to fill in the original dataset? How might this have affected the variability of their daily model estimates?

  In our study, observed stream DOC concentrations were measured at weekly and biweekly intervals from grab samples, as described in the data description section. Daily observations were not available, and no interpolation method was applied to generate daily data. Accordingly, the objective function (e.g., KGE) was calculated comparing the simulated concentrations on the respective day of the observations.

  Because the observations are not available at a daily resolution, short-term event-driven DOC may be underrepresented in the dataset. Consequently, when these data are used for model calibration, the model may have limited ability to capture high DOC concentrations during such events. We added additional text regarding these points and also in connection with comments from reviewer #2.

  L556-567:

  "Additional stream DOC observations, particularly during high DOC concentration events, would also help to better characterize DOC export across the full range of hydrological and biogeochemical conditions. This is especially relevant for Warme Bode, where observed stream DOC concentrations span a wider range than others (Fig. 4). Under such conditions, sparse weekly or biweekly sampling may miss short-lived DOC peaks and therefore provide limited information for evaluating event-scale model behaviour. High-frequency DOC measurements or suitable optical proxies, would help to constrain these dynamics more directly and could improve the representation of high-DOC events in the model."

- As the authors found that results (under the baseline calibration) varied by evaluation metric, there should be additional justification added to the methods to provide rationale for their focus on using Kling-Gupta efficiency (KGE) values in the main text and primarily including NSE, BIAS, and R2 results only in the supplement.

  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), as indicated by the Reviewer #3 as well. In contrast, NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018). We added this to the main text (L245-247). We also note that the selection of performance metrics ultimately depends on the specific objectives of the modelling study.

L247-248:

> "We selected KGE as our primary evaluation criterion as it balances between correlation, variability, and bias between observed and simulated (Gupta et al., 2009). In contrast, NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018)."

Reference

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, *377*(1-2), 80-91.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrol. Earth Syst. Sci., 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

- Additional text should be added to the discussion to further examine why simulated DOC concentrations were unrealistic in the groundwater and stream pools under the different calibrations. What might be some additional mechanisms/processes that the model is unable to account for or capture the variability of? And which might be masked or biased under baseline and constrained calibrations?

We thank the reviewer for this clarification. From our analysis viewpoint, simulated groundwater DOC in some of the behavioral simulations under the baseline calibration is much higher than expected from observations because the model is calibrated using stream DOC rather than DOC concentrations in individual compartments (hillslope, riparian, and groundwater). Under the baseline calibration, there is an interplay between groundwater DOC and hillslope DOC (or riparian DOC). In other words, high groundwater DOC can be compensated by low hillslope (or riparian) DOC to match observed stream DOC. So in the baseline calibration, the role of groundwater DOC might be masked by hillslope or riparian DOC. We have modified the main text to reflect these points:

L464-468:

> "In fact, with the lumped model structure, most of the behavioural models under the baseline calibration do not meet our understanding regarding low groundwater DOC concentrations. More specifically, the lumped model needs implausibly high groundwater DOC concentrations to compensate for the low DOC concentration from the hillslope during the high-flow period in order to match observed stream DOC. Also, with the landscape-explicit model structure, only 53% of behavioural models (across all four catchments) under the baseline calibrations meet all of our constraints (section 2.4)."

L477-487:

> "The interplay between different parameters affecting DOC production, together with the added complexity of having those parameters acting in different model compartments (i.e., upland, groundwater, and riparian zone) result in a myriad of model simulations that are able to reproduce stream DOC dynamics, but not necessarily able to represent landscape-internal DOC dynamics. With more complex model structures, there is a higher degree of freedom and a higher uncertainty in

> model parameterization, leading to higher uncertainty in the internal DOC dynamics. For example, wider simulated groundwater DOC concentration ranges were found from the landscape-explicit model structure than from the lumped model structure in the baseline calibration (Figs. 6, 7). This is because the amount of groundwater that flows directly to the stream in the landscape-explicit model structure is less than that in the lumped model structure (Fig. 1 and section 3.1), making the information content of observed stream DOC concentration ineffective for constraining groundwater DOC concentration. Similarly, there is also a high uncertainty in the upland DOC concentration from the baseline calibration, as DOC in the upland does not flow directly into the stream."

An additional process that the model might not accurately account for is the vertical representation of the riparian zone, e.g., through multiple soil layers, could account for depth-dependent DOC concentrations and for changes in the riparian source layer contributing to stream DOC under variable flow conditions. This could potentially improve the simulation of high DOC concentrations, especially for the Kalte Bode and Warme Bode catchments (Fig. 5b), as well as the representation of groundwater DOC concentrations. We have added this to the discussion (L558-561):

> "Furthermore, a more vertically refined representation of the riparian zone, e.g., through multiple soil layers, could account for depth-dependent DOC concentrations and for changes in the riparian source layer contributing to stream DOC under variable flow conditions. This could potentially improve the simulation of high DOC concentrations, especially for the Kalte Bode and Warme Bode catchments (Fig. 5b), as well as the representation of groundwater DOC concentrations. "

- Broadly, the discussion would also benefit from more analysis of results in the context of other models/results from other systems. Currently, it reads very narrowly focused on this particular study and location.

  Thank you. We added some related references to the discussion section, regarding other models and other study areas, e.g., SWAT-Carbon (Zhang et al., 2013), INCA-C (Futter et al., 2007), HYPE (Pers et al., 2016); BioRT-Flux-PIHM (Wen et al., 2020) (L537-538).

  We noted that in the Discussion, we discussed the implications of our results in a broader context, while the results from our study are cited as scientific evidence for such implications.

- Per my earlier comment, additional rationale for the authors' primary reliance on KGE values would help to provide added confidence in their findings comparing baseline and constrained calibration values. Currently, it is hard to reconcile the "slight" decrease in landscape-explicit model performance with no change in NSE, BIAS, and R2 values between calibrations as well as the "significant" decrease in the lumped model performance with no change in R2 values. If the authors are using all four evaluation metrics jointly to describe model performance, this should be clearly stated in the methods; if instead, they weigh KGE values most in determining model performance, this instead should be clarified.

  Thank you. We have added the rationale for using KGE over other performance metrics (e.g., NSE, BIAS, and R^2), as described in our response above.

  L247-250:

  > "We selected KGE as our primary evaluation criterion as it balances between correlation, variability, and bias between observed and simulated (Gupta et al., 2009). In contrast, NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018)."

> Furthermore, we also presented the individual components of KGE (correlation, bias, and variability; shown in Figs. S5-S9 in the supporting information) separately so that readers can better understand the strengths and weaknesses of our model. We noted that the KGE is used as primary evaluation criteria (as stated above).

Technical Corrections

Line 40 – It is unclear why the authors focus specifically on temperate/boreal systems in this sentence. If the manuscript plans to do the same, this should be introduced prior.

The temperate and boreal systems cover a large portion of the Earth's land surface, and the underlying DOC dynamics are sufficiently general to be applicable across other systems as well. In addition, this is just an example intended to illustrate the importance of considering different landscape types within a catchment. We have clarified this point in the manuscript.

L40-43:

> "Riparian zones in temperate and boreal systems, which cover a large portion of the Earth's land surface and exhibit DOC dynamics that are general enough to be applicable across other systems, are typically wet with shallow groundwater levels, rich in organic matter, enhancing lateral hydrological connectivity between the DOC source zone and the stream network (Ledesma et al., 2018a)."

Line 170 – Please review the subscripts in Equations 11 and 12; they do not appear to match the compartments displayed in Figure 1.

Thanks for pointing this out, we have changed the subscripts in Eqs. 11-12 to match with the model concept (Fig. 1). Before this is the typo error.

Line 293 – I believe this heading should read as "3 Results".

Thank you. Yes, that is right, we have corrected this