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.

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, 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 using the available weekly and biweekly observations. Because the observations are not available at a daily resolution, short-term event-driven DOC peaks 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 will add this to the discussion/limitation section.

- 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). In contrast, NSE does not adequately account for variability (Gupta et al., 2009; Santos et al., 2018). We also note that the selection of performance metrics ultimately depends on the specific objectives of the modelling study.

  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 will add to the discussion regarding groundwater DOC concentration. Specifically, simulated groundwater DOC in some of the behavioral simulations under the baseline calibration of is much higher than expected because the model is calibrated using stream DOC rather than DOC concentration in individual compartments (hillslope, riparian, and groundwater). Under such calibration, there is interplay between groundwater DOC and hillslope DOC (or riparian DOC). In other words, high DOC in the groundwater can be compensated by low DOC from the hillslope (or riparian) during the lowflow period and vice versa to match with stream DOC during low flow condition. In most cases, especially with lumped model structure, stream DOC under based line calibration is higher versus constrained as in the constrained calibration groundwater DOC is constrained to lower.

  Simulated groundwater DOC is higher than expected, showing that the model is not able to account for the vertical transport of DOC. Specifically, DOC is transported vertically from the organic-rich upper soil layers into the deeper mineral subsoil and quickly gets adsorbed to the mineral phase, resulting in low DOC concentrations in the percolating water that recharges the groundwater.

  So in the baseline calibration, the role of groundwater DOC might be masked by hillslope or riparian DOC

  We will discuss the aforementioned points in the discussion.

- 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 will search for similar studies/models/areas and discuss our results in the context of these studies.

- 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 will add a rationale for using KGE over other performance metrics (e.g., NSE, BIAS, and R^2), as described in our response above. Furthermore, we will present the individual components of KGE (correlation, bias, and variability) separately so that readers can better understand the strengths and weaknesses of our model. We believe that reporting the different components of KGE provides clearer insight into why model performance may decrease "slightly" or "significantly," even when some metrics (e.g., BIAS or R^2) show little or no change, as noted by the reviewers.

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 will clarify this point in the manuscript

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 will revise the subscript of equations 11 (change RZ to HS) and 12 (change UL to RZ).

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

Thank you. Yes, that is right, we will correct this