We appreciate the comments from the reviewer. We still want to keep our paper as a research paper. We plan to add a more in-depth discussion to show the importance of the implicit scheme in large-scale hydrological simulation with big data and how our results align with the findings of Clark and Kavetski (2010) and Kavetski and Clark (2010), but we do not want to repeat their work that has been done thoroughly.

Since daily forcing and streamflow data are more commonly available and accessible, while hourly input data is often harder to obtain, in this work, our focus is solely on daily hydrological modeling using conceptual models. The numerical schemes employed within the framework of the differentiable model need to maintain stability for simultaneous large-scale simulations in minibatch while also allowing for gradient tracking. We conducted tests on the differentiable HBV model with various numerical schemes and time steps. In theory, with smaller time steps, we were supposed to configure forcing functions to provide hourly inputs that match the progression of time within a day, i.e., the inputs should reflect diurnal changes in forcing. However, as the datasets do not contain such hourly information and we are only doing a numerical comparison, we chose not to do that here.

We plan to include the following analyses in our discussion section (potentially as an appendix):
1. The fixed-step explicit Euler scheme with one day time step caused divergence in the large-scale simulation due to its instability. However, with a 4-hour time step, it exhibited better performance than the 4th-order Runge-Kutta (RK) explicit scheme but still lagged behind the sequential scheme that employed ad-hoc operation splitting (attached Table 1). The ad-hoc operation splitting updates the state variables after each operation has higher accuracy and stability. The implicit scheme yielded superior performance in terms of KGE and high and low flow metrics. Reducing time steps and using implicit schemes both improved model accuracy, aligning with the findings of Clark et al. in 2010.

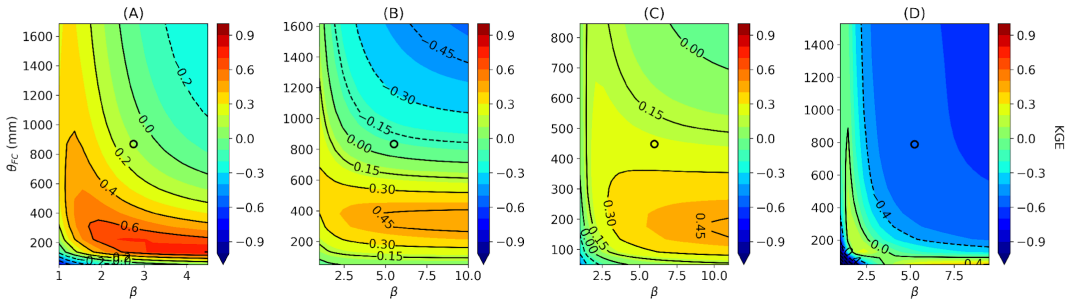
2. An important finding in Kavetski and Clark (2010) is that numerical approximation errors can be compensated for by distorted parameter values during calibration, resulting in a 'right result for the wrong reasons.' This phenomenon can affect parameter uncertainty analysis, hinder meaningful parameter interpretation and regionalization, and lead to erroneous internal model dynamics. To explore this further, we examine the metric (KGE) surface within the 2D slice defined by field capacity (θ_{FC}) and the parameter β, derived from various numerical solutions (attached Figure 1). Thanks to the neural network's ability for parameter regionalization, we did not observe a notable macroscale roughness as shown in Figure 1 of Kavetski and Clark (2010) when using explicit schemes. However, we still observed distortions and roughness in the KGE surface in models employing the RK scheme (site A and D). As we reduced the time steps and transitioned to implicit schemes, these distortions seem to have alleviated and converged toward the metric surface consistent with the correct numerical solution. That is, the 4-hourly patterns are more similar to the implicit results than the daily, and we expect further convergence when we are able to run the one-hourly model (this takes more time to prepare). The convergence toward the implicit scheme suggests that the implicit scheme results were more reliable.

3. Even with small time steps, we still observe that explicit schemes exhibit parameter distortion and potentially problematic internal model dynamics, as demonstrated in Table 2 and Figure 6 in the manuscript. Such issues have implications on our ability to learn a better model structure. Hence, we argue that implicit scheme has value.

4. Regarding the reviewer's recommendation to drop the analysis about the structural change, we have some reservations. Because we are providing comprehensive comparisons both with the change and without the change, it does not seem like the readers would lose anything by seeing the best configurations. In fact, it is to the benefit of the community to see what changes matter and by how much, and how to obtain the state of the art. We promise to present it in a clear and insightful way. While we fully respect the reviewer's opinion and could see where he is coming from, we respectfully insist on our paper's design here.

Table 1: Summary of streamflow metrics for models using different numerical schemes. As 1-hour time-step models take much more time, the comparison may be provided later in the revision, potentially as an appendix. Upon revision, a comparison of timing and memory use will also be provided.

Model	Numerical scheme	Time step	Median NSE	Median KGE	Median low flow RMSE (mm/day)	Median peak flow RMSE (mm/day)
δHBV	Fixed-step explicit	1 day	-	-	-	-
δHBV	The fourth-order Runge-Kutta explicit	1 day	0.69	0.70	0.061	3.25
δHBV	Fixed-step explicit	4 hours	0.72	0.71	0.09	2.50
δHBV	Sequential	1 day	0.73	0.73	0.074	2.56
δHBV.adj	Implicit adjoint	1 day	0.72	0.75	0.048	2.47



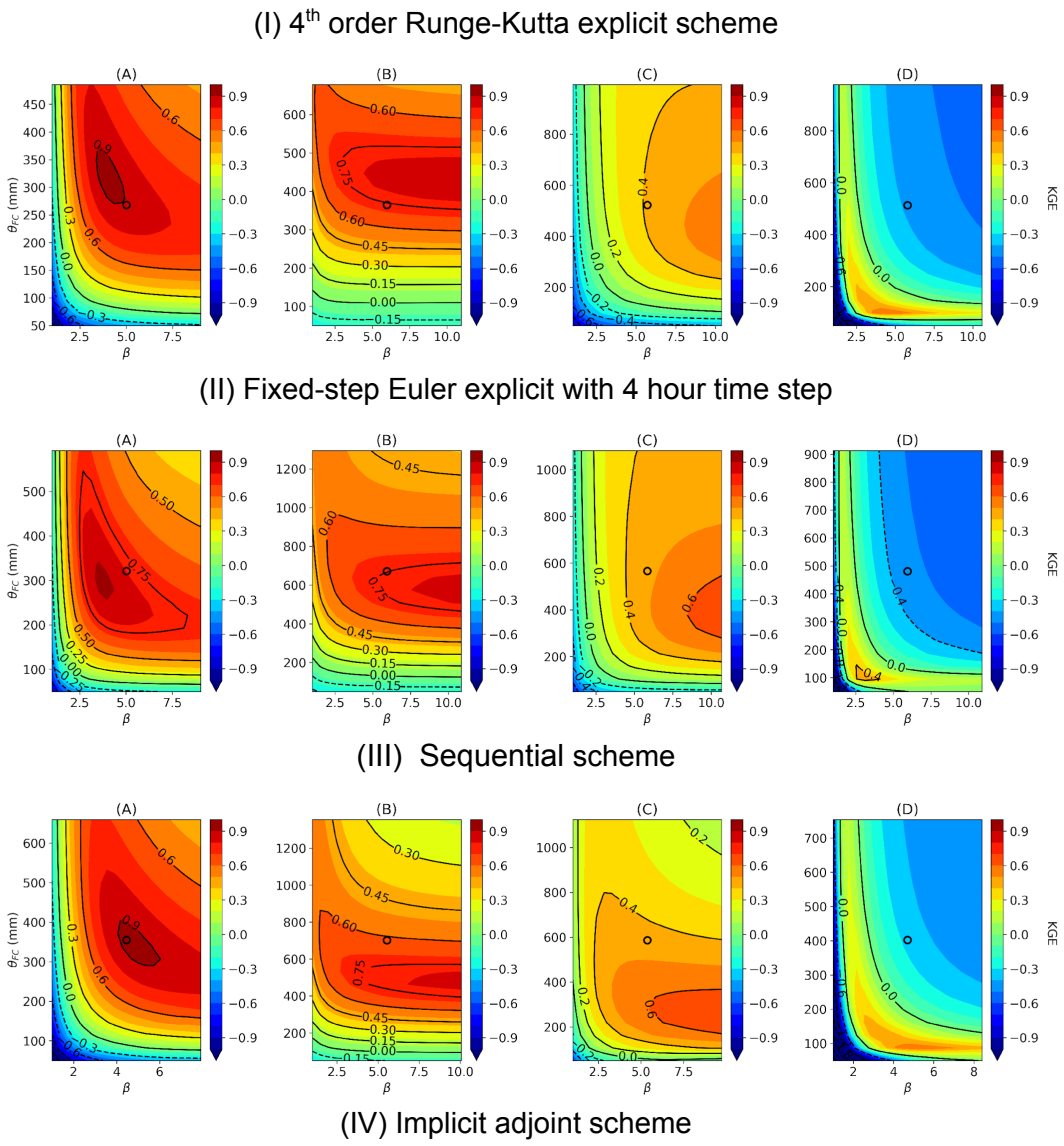


Figure 1: Impact of numerical schemes on the KGE surface of HBV model: The contour of KGE calculated from the (I) 4th order Runge-Kutta explicit scheme, (II) Fixed-step Euler explicit with 4 hour time step with 4 hour time step, (III) sequential scheme, and (IV) implicit adjoint scheme on the 2D slice of field capacity (θ_{FC}) and parameter β. The predicted parameter values are positioned at the central point of the contours delineated by circles. The locations of selected sites are annotated in Figure 6