The reviewers are generally positive about the revision and note that most of their comments have been addressed. However, several points remain that would benefit from further attention. After reviewing these comments, I find them to be reasonable and constructive, and addressing them would further improve the robustness and overall quality of the paper. I therefore recommend that the authors carefully consider these comments and revise the manuscript accordingly in a new version.

Reply: Thanks for giving us the opportunity to revise this manuscript again. We have carefully addressed the reviewers' comments and made targeted revisions throughout the manuscript. The point-to-point responses to all comments are presented below.

Reviewer 1

Most of my previous comments have been resolved satisfactorily, with the exception of the following points:

Line 254: Technically, the hydraulic head is not solely equal to the pressure head here. By definition (if we neglect the velocity head), the hydraulic head H should equal the summation of the pressure head ($P/\rho g$) and the elevation head (Z). While I agree that the pressure change ΔP depends only on the pressure head, it is not formally correct to state $H = P/(\rho g)$ directly. Please refer to Strack (1989), Groundwater Mechanics, p.8.

Reply: Yes, you are right. We are sorry for the confusion. Actually, the "head" calculated from equation (10) is the "pressure head (h)", while P denotes pressure, which includes contributions from both the pressure head and elevation. We have revised the relevant statements in the manuscript, please see line 248.

Figure 3: This figure remains a little unclear to me. For example, do the series of red points indicate that the mean flow velocity is increasing at different depths, or do they simply represent different simulation scenarios? Clarifying this would be helpful.

Reply: The series of points in Figure 3 represent different simulation scenarios. Different colors

indicate different median grain sizes of the streambed sediments. For each sediment grain, various streamflow conditions and bedform migration rates were simulated, which correspond to the velocity values shown on the y-axis. We have revised the figure legend and added a brief explanation in the manuscript to make it clearer. Please see lines 281-283.

Figure 5: The unit for RMSE appears to be missing. I noticed the unit might be too long to fit comfortably within the plot itself; perhaps you could state it in the figure caption instead?

Reply: The unit for RMSE has been added to the figure caption in the revised manuscript, please see lines 520-521.

Reviewer 2

I thank the authors for making efforts to address my comments in previous revision. I believe the authors have addressed most of my comments. After carefully reviewing the responses and revised manuscript, there are still two critical issues that need to be addressed to enhance model reliability and avoid misleading.

(1) The mass conservation accuracy for concentration is too low. I raised the potential mass conservation issue in major comment 3 in previous review comment. The authors provided quantitative evaluation of mass conservation for flow and concentration in Table S3. The results show that flow mass conservation error is 0.28%, which confirms flow mass conservation. However, the mass conservation error is 4.62% for groundwater-borne NO3 and 1.61% for stream-borne NO3. Considering that this work used mature commercial CFD software (COMSOL) as its computational engine and the computational domain (2D rectangular domain) is quite simple for CFD applications, such mass conservation error is generally treated as low accuracy which could undermine the reliability of all results. The authors should identify the causes of such high mass conservation error and improve the simulations results to a widely acceptable level for CFD-based reactive transport modeling. As a reference, 1% is recommended by USGS as a good mass balance error

https://wwwbrr.cr.usgs.gov/projects/GW_Unsat/vs2di/hlp/solute/massBalanceDiscussion.html.

Reply: Thank you very much for the careful and detailed comments. We fully agree with the reviewer's perspective as well as the relevant discussions on the USGS website. Based on our long-term experience with numerical modeling, it is generally relatively easy to achieve a mass conservation error of around 1% for flow models. However, for solute transport models, especially reactive solute transport models, the error is usually somewhat larger than that.

We have recalculated the mass conservation error for each time step of our transient model by strictly following the solute mass balance, which accounts for inflow, outflow, reaction, and the mass accumulated within the computational domain. The time series of the mass (c_sno_3, blue line and c_gno_3, green line) conservation errors was presented in the following Figure R1, where x-axis denotes iteration number and y-axis denotes the magnitude of the mass conservation error. The errors exhibit a periodic pattern, which arises from the spatiotemporal variations in pressure head imposed on the top boundary of the simulated domain.

In the previous manuscript, we calculated the conservation errors for stream-borne nitrate and groundwater-borne nitrate only at the final time step of the transient simulation. Given the periodic nature of the conservation error, we have now recalculated the average conservation errors over the entire simulation period (over 500 ripple passages). The resulting conservation errors for c_sno_3 and c_gno_3 are 0.09% and 2.98%, respectively (Table R1).

Table R1. Model-calculated average conservation results over the entire stage

Boundary	fluid in (m^2/s)	fluid out (m^2/s)	c_gno_3 in (mg/m h)	c_gno_3 out (mg/m h)	c_sno_3 in (mg/m h)	c_sno_3 out (mg/m h)	R_sno_3 (mg/m h)	R_gno_3 (mg/m h)
BC (top)	6.16×10^{-8}	1.12×10^{-7}	1.247	1.928	1.305	0.547	0.756	1.908
AB (right)	7.87×10^{-9}	2.25×10^{-8}	0.189	0.693	0.045	0.177		
AB (left)	2.25×10^{-8}	7.86×10^{-9}	0.693	0.189	0.177	0.045		
AD (bottom)	5.04×10^{-8}	0.00	2.733	0.000	0.000	0.000		
Sum	1.42×10^{-7}	1.42×10^{-7}	4.862	2.809	1.527	0.770		
Error (%)	0		2.98		0.09			

Exactly, as noted by the USGS, the initial mass balance error is larger due to the instantaneous changes in the initial conditions caused by the imposed boundary conditions. We calculated the conservation errors between 400 and 500 ripple passages, the resulting conservation errors for c_sno_3

and c_gno_3 are 0.02% and 0.01%, respectively (Table R2). We confirm that the numerical conceptualization and boundary conditions of the model are correct, and the observed errors arise from unavoidable systematic numerical errors.

Table R2. Model-calculated average conservation results during the steady stage

Boundary	fluid in (m^2/s)	fluid out (m^2/s)	c_gno_3 in (mg/m h)	c_gno_3 out (mg/m h)	c_sno_3 in (mg/m h)	c_sno_3 out (mg/m h)	R_sno_3 (mg/m h)	R_gno_3 (mg/m h)
BC (top)	6.15×10^{-8}	1.12×10^{-7}	1.350	2.083	1.304	0.547	0.757	1.989
AB (right)	8.12×10^{-9}	2.23×10^{-8}	0.207	0.725	0.048	0.177		
AB (left)	2.23×10^{-8}	8.11×10^{-9}	0.725	0.207	0.176	0.048		
AD (bottom)	5.04×10^{-8}	0.00	2.722	0.000	0.000	0.000		
Sum	1.42×10^{-7}	1.42×10^{-7}	5.005	3.015	1.528	0.771		
Error (%)	0		0.01		0.02			

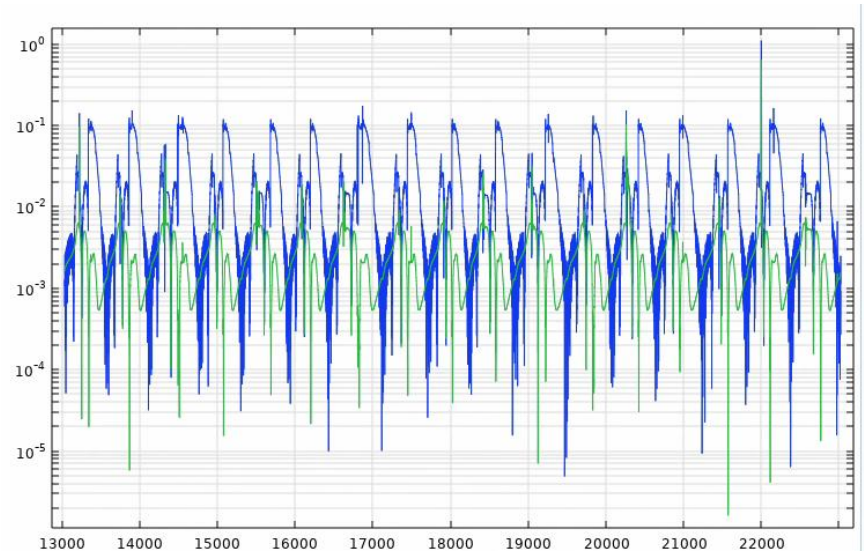

Figure R1. The variation of mass conservation error with the numerical iteration number.

(2) The texts that justify replacing triangle bed with flatbed should avoid overstatement. Figure 4 shows the comparison of flow and concentration distributions from ripple and flat beds. The line 443-444 describes “their overall patterns are most similar”. This description gives me the impression that the difference between the two geometries are not important. However, by carefully checking the regions beneath the flat surface in, I see the magnitude and locations of maximum/minimum values shift a lot (roughly 10-30% from visual observation). In hyporheic zone study, the magnitude and

locations of max/min values represent the hot spots and hot moments, which is the key study target. The Figure 4 actually suggests the model is not reliable for studying hot spots/moments if ignoring the triangle beds. Such information should be reflected in the Paragraph ending at line 453. Further, Line 459 says "the flat bed exhibit slightly larger flux…..". However, the following sentences says "10%" and "20%" errors. From a pure CFD model perspective, 10%-20% errors are never slight errors, but moderate to high errors, especially for simple geometry computational domains used in this work. Such descriptions will likely cause immediate concern for future readers whose work involves detailed CFD modeling. I suggest the authors revise such texts to be more objective to avoid potential overstatement. Also, I recommend that authors adding a figure to show the distributions of relative errors of all variables used in Figure 4 as a supplementary figure, which will show the most comprehensive difference between ripple and flatbed results.

Reply: We have revised the descriptions in Section 3.1.1 to avoid overstatement. Instead of comparing exchange fluxes, mixing fluxes, and reaction rates at the final time step of the transient model (as in the original version), we now compare their average values, taking into account the periodic patterns observed in these parameters (differences of 5%–10% in the revised version). We have also modified the concluding paragraph (ending at Line 476) to emphasize the differences in hotspots and nitrate removal efficiency between the triangular ripple and flat bed models.

In fact, both the triangular ripple and flat bed model are conceptual approaches used to describe water flow and solute transport under moving bedforms. The results from the ripple model should not be interpreted as representing the true or absolute values of the physical processes. Therefore, the observed differences of approximately 5–10% between the two models do not necessarily indicate absolute errors in the flat-bed approximation. To provide a more comprehensive comparison, we have added supplementary rows showing the relative difference distributions of all variables presented in Figure 4. This figure clearly illustrates the spatial pattern of differences between the triangular ripple and flat bed simulations.

There are also a few minor issues, listed below for consideration:

Line 189: Equation 2 is incorrect. Gradient v ci is a tensor, but the rest terms are scalars. The correct

way to write it is gradient dot (vc).

Reply: We have revised the equation 2 (in the revised version): $\frac{\partial c_i}{\partial t} - \nabla \cdot \left(D_{ij} \nabla c_i \right) + \nabla \cdot \left(\boldsymbol{v} c_i \right) = R_i$

Figure 2: as mentioned in previous comments, the governing equations do not include P, but only h. Please considering replacing P in Figure 2 (boundary AB, CD) by h.

Reply: We have revised Figure 2 by replacing P with h in the boundary condition labels (AB and CD).

Line 478: changed "method" to 'empirical equation'. From Fox et al. 2018, they proposed an equation to back-calculate the HEF based on experimentally-validated equation using concentration profile. This is an empirical equation-based approach, which is different from the "integration of volumetric flux at the boundary BC" used in numerical experiments. Replacing "method" by "empirical equation" will help potential readers understand how experiment HFEs were derived.

Reply: We have replaced “method” to “empirical equation” in the revised manuscript, please see line 492.

Line 487: please clarify “with their parameters adjusted to fit …”. Do you modify parameters to match the observational HEF/Fin/Fout/Ao from experiment? Usually, if you say model validation. This means you should use the same parameters from experiments, without parameter fitting. If you fit parameters to match, this is called model calibration and cannot be used a model accuracy evaluation. Please clarify this sentence.

Reply: In our study, all physical and hydraulic parameters (e.g., ripple geometry, streamflow, bedform migration velocity, porosity, and permeability) were used exactly as reported in Wolke’s experiments without any adjustment. Only the biogeochemical reaction parameters (e.g., dispersivities, maximum oxygen consumption rate, and half-saturation constant for oxygen) were

calibrated to match the observed oxygen plume, as these parameters are difficult to measure directly in dynamic streambed environments and are commonly calibrated in reactive transport models. We have revised this sentence, please see lines 496-505.

Figure 5: Figure 5e, f, h experiment data have more than 1 points at different velocity. But numerical model has one point at each velocity. How do you deal with the multiple data in experiment? For the RMSE, is the difference computed as the difference between experiment and flat mode numerical results?

Reply: For Figures 5e, 5f, and 5h, where multiple experimental data points exist at the same velocity, we first averaged the experimental values at each velocity. The RMSE was then calculated between these experimental averages and the corresponding numerical results from the flat bed model.

Reviewer 3

I appreciate the substantial revisions and improved clarity, particularly the distinction from Ping et al. (2022). However, several key issues remain that require further revision.

1.Please provide a complete specification of the conservative tracer's initial and boundary conditions. Clarify whether the groundwater tracer is a continuous source or a pulse, as the current description of "mixing start/end" implies a pulse, which contradicts a continuous source assumption.

Reply: The groundwater-borne conservative tracer is treated as a continuous source with a constant concentration boundary condition applied at the bottom inlet of the domain. The initial concentration of the tracer within the streambed is set to zero. The terms "mixing start/end" in the manuscript do not imply a pulse input. Instead, they are used to define the period over which we analyze the mixing dynamics between upwelling groundwater and downwelling surface water. "Mixing start" corresponds to the time when groundwater first encounters surface water within the streambed. "Mixing end" is defined as the time when the tracer, introduced continuously from the bottom

boundary, has spread sufficiently throughout the domain. We have revised the manuscript to clarify that the groundwater tracer is a continuous source and that "mixing start/end" refers to the analysis period rather than the duration of tracer input. Please see lines 395-409.

2.Since mixing metrics (e.g., Amix, MF) are sensitive to physical dispersion, please include a sensitivity analysis for dispersivities aL, aT) to demonstrate that the reported trends are not artifacts of the parameterization.

Reply: We have added the sensitivity analysis for dispersivities a_L and a_T in the revised manuscript and supporting information, please see lines 365-369 and Text S4.

3.The comparison to Wolke et al. validates hyporheic exchange but does not cover the gaining conditions central to this study. Consequently, it does not directly validate the specific mixing structure or the derived metrics. To strengthen credibility, please seek an independent benchmark specifically for mixing (e.g., tracer data of experiments with mixing process).

Reply: The comparison with Wolke et al. serves to validate the model's ability to represent bedform migration. To our knowledge, no published experiments explicitly measure mixing structures or mixing metrics under gaining conditions. Most previous studies investigating mixing dynamics and mixing-triggered reactions have relied on numerical simulations (Hester et al., 2013, 2014, 2019; Nogueira et al., 2022, 2024). Santizo et al. (2020) conducted a flume experiment and simulated a mixing-triggered chemical reaction (sulfite oxidation), using stream-borne DO to estimate mixing thickness. However, DO is not a conservative tracer and thus cannot reliably represent mixing dynamics. We apologize for the lack of further validation, which is primarily due to the scarcity of direct experimental data for mixing under gaining conditions.

4.Partitioning nitrate with nonlinear Monod kinetics can be inconsistent because f(c_tot) not equals to f(c_s)+f(c_g). Please explicitly state whether the denitrification rate is controlled by total nitrate or calculated for each pool separately. If based on total nitrate, specify how consumption is allocated

between sources. If calculated separately, justify the physical validity.

Reply: In our model, denitrification is calculated separately for stream- and groundwater-borne nitrate. This definition aims to distinguish between non-mixing-dependent and mixing-dependent denitrification reactions in the model. Each nitrate type, differentiated by its origin, has its own set of boundary conditions. We separately solve for the transport of these two nitrate fractions in the porous medium using the advection-dispersion-reaction equation. We quantified the inputs and outputs of both stream-borne and groundwater-borne nitrate, as well as their internal consumption via denitrification. Notably, the relative error remains below the acceptable threshold, confirming the conservation of solute mass.

5.The pressure condition P(0) = P(\lambda) + \Delta P is not strictly periodic but quasi-periodic (periodic with an imposed mean gradient).

Reply: We agree that the pressure condition $P(0) = P(\lambda)+\Delta P$ represents a quasi-periodic (periodic with a mean gradient) rather than strictly periodic boundary condition. We have revised the relevant description in the manuscript, please see line 240.

Reference

Hester, E. T., Young, K. I., and Widdowson, M. A., (2013). Mixing of surface and groundwater induced by riverbed dunes: Implications for hyporheic zone definitions and pollutant reactions. Water Resources Research, 49, 5221-5237.

Hester, E. T., Young, K. I., Widdowson, M. A., (2014). Controls mixing-dependent denitrification in hyporheic zones induced by riverbed dunes: A steady state modeling study. Water Resources Research, 50, 9048-9066.

Hester, E. T., Eastes, L. A., Widdowson, M. A., (2019). Effect of surface water stage fluctuation on mixing-dependent hyporheic denitrification in riverbed dune. Water Resources Research, 55, 4668-4687.

Nogueira, G. E. H., Schmidt, C., Partington, D., Brunner, P., and Fleckenstein, J. H., (2022). Spatiotemporal variations in water sources and mixing spots in a riparian zone. Hydrology Earth System Science, 26 (7), 1883-1905.

Nogueira, G. E. H., Partington, D., Heidbuchel, I., and Fleckenstein, J. H., (2024). Combined effects of geological heterogeneity and discharge events on groundwater and surface water mixing. Journal of Hydrology, 638, 131467.

Santizo, K. Y., Widdowson, M. A., Hester, E. T., (2020). Abiotic mixing-dependent reaction in a laboratory simulated hyporheic zone. Water Resources Research, 56, e2020WR027090.