Anonymous Referee #2

Received and published: 4 November 2020

It is very useful to have new models so that one can compare different approaches, especially when the models have open source code like this one. However, with very many existing catchment-scale biogeochemical water quality models, it should be more clearly stated what this model provides that others don't, and how it can shed light on catchment processes that were previously not well understood. Or alternatively how this new software makes the job of the model user easier. For instance, the process descriptions for Phosphorous, DOC and Nitrogen don't look too dissimilar to existing models. My impression is that this model is toward the complex end of the spectrum when it comes to parametric complexity. I miss an explicit analysis of this. For instance, how many parameters have to be calibrated and cannot be sufficiently constrained by (easily obtainable) measurement or literature values?

Response: Thanks for the comment. We have revised two introduction paragraphs to be more explicit about the existing model gap and how our integrated watershed model help understand complicated watershed processes. We have also added a paragraph for calibration process.

Line 74 – 80: "Reactive transport models (RTMs) have traditionally centered on transport and multi-component biogeochemical reactions, typically in groundwater systems, which often have limited interactions with climate and other surficial watershed processes (Steefel et al., 2015;Li et al., 2017b;Mayer et al., 2002). Biogeochemical reactions in shallow soils that are often driven by environmental factors such as soil temperature and moisture cannot be well simulated in these models."

Line 90 – 108: "While many of these models can simulate reaction processes such leaching of nutrients from agriculture lands (Lindström et al., 2005;Lindström et al., 2010;Bailey et al., 2017), most of them do not explicitly solve the multi-component reactive transport equations. In other words, they have relatively crude representations of solute leaching out of element bulk mass as part of the solute export but do not represent kinetics and thermodynamics of multi-component biogeochemical reactions typically included in reactive transport models (RTMs). They also do not simulate processes such as chemical weathering. As an example, nutrient leaching is often calculated based on empirical equations without explicitly solving reactive transport equations. Reaction rates are often represented using first-order decay (Gatel et al., 2019), assuming reaction rate constants do not change with time and environmental conditions. However, biogeochemical processes including carbon decomposition and nutrient cycling are highly variable in space and time, depending on local environments such as substrate availability, soil temperature, and soil moisture (Li et al., 2017a;Suseela

et al., 2012;HARTLEY et al., 2007). In filling in this model need, recently we augmented our watershed model RT-Flux-PIHM (Bao et al., 2017) with new developments of microbially mediated reactions, which allows us to model the interactions between biogeochemical reactions and environmental factors that are driven by land surface and hydrological processes."

Line 491 – 497: "A typical model application requires 20 to 30 hydrological parameters to be calibrated. These parameters include land surface parameters (e.g., canopy resistance, surface albedo), soil and geology parameters (e.g., hydraulic conductivity, porosity, Van Genuchten, macropore properties) (Shi et al., 2013). Reaction-related parameters (e.g., reaction rate constant, mineral surface area, Q_{10}, $S_{w,c}$, and n) are additionally needed for calibration, the number of which depends on the numbers of reactions involved in a particular system."

Knowledge of this is of vital importance to a user. For instance, if one wants to do a very thorough investigation of the processes in a single catchment one maybe has some time to spend to do a very detailed model setup. However, in some applications one needs to model all the inputs from land into a whole coastline or a large set of lakes. In such an application one often relies on autocalibration and upscaling, and in such applications high parametric complexity can be detrimental. On a similar note, data availability of data that can be used as model drivers can vary between locations. Does this model accommodate for locations with low data availability?

Response: Thanks for the interesting point. The short answer for low data availability is mostly YES as our model can be operated on the spatially lumped version, which requires much less data for spatial details and data points in different locations. And the model is flexible and can take inputs either from online data portals or user's own measurements (e.g., elevation, soil properties). The model uses a global coefficient approach to reduce parameter dimension and facilitate parameter calibration (*Line 490 – 491*).

Line 197 – 200: "Despite the model complexity, the model is flexible for taking inputs from online data portals or local measurements and it can accommodate low data availability (see the following section of 5 for data need and domain setup)."

Line 445 – 447: "the model domain can be set up using elevation, land cover, soil and geology maps supplied by the user or from the data portal of Geospatial Data Gateway (https://datagateway.nrcs.usda.gov)."

Line 455 – 456: "Local measurements from meteorological stations and field campaigns (e.g., land cover, soil, geology) can also be used in the model."

Line 481 – 483: "A simple domain can be set up with only two land grids representing two sides of a watershed connected by one river cell. This setup uses averaged properties without needs for larger spatial data."

Line 490 – 491: "Auto-calibration is not built into the model, but a global calibration coefficient approach is used to reduce parameter dimension and facilitate manual calibration."

If the stated goal of the model is to be a research model targeted at understanding catchment processes, rather than a model that can also be used as an input source for oceanic models or to be used by government officials to inform policy decisions on a large scale, then this is maybe not as big of a concern. But that could be made more explicit.
Response: We respectfully disagree. How to use the model should be determined by users, not model developers. For example, HBV is a simple hydrology model but it has been used by many in the world for policy making and guidelines for decisions. Users should decide how they would like to use the code.

Can you argue why the model complexity is justified? Some studies show that simple models can give as good predictions as complex ones while taking much less time and data to deploy. I can see that you have a plot of sensitivity to turning off various nitrate processes, but what about sensitivity to simpler or more complex descriptions of these processes? What is the sensitivity of model results to perturbations in the parameters?
Response: This is a generic question to the whole modeling community, not just about this manuscript. Nonetheless, we added some discussions on model simplicity and complexity. In a nutshell, we advocate for simple models and adding complexity only if necessary. We added the following:

Line 723 – 756: "The model presented here is complex and process-based. The computational cost of solving a spatially distributed, nonlinear, multi-component reactive transport model is high, posing challenges for the application of ensemble-based uncertainty analysis and model weighting/selection methods (Song et al., 2015). With additional reaction and transport processes, the model includes more functions (such as reaction kinetic rate laws) and parameters (e.g., reaction rate constants, surface area) than hydrological models, which have already been criticized for their complexity, equifinality, uncertainty, and data demands (Beven, 2001, 2006;Kirchner et al., 1996). These issues will persist even though reactive transport models will be constrained by additional chemical data. A major source of uncertainty in these models lies in epistemic uncertainties, i.e., the lack of specific knowledge in forcing data and details of reactivities (e.g., spatial distribution and abundance of reactive materials), on top of uncertainties related to hydrology (Beven, 2000;Beven and Freer, 2001). The model's conceptual foundations also represent a major source of uncertainty.

It is in this spirit of "balancing" the cost and gain that we present both spatial distributed and lumped modes for the BioRT model. Compared to the distributed version, the spatially implicit model requires less spatial data, is computationally inexpensive, and

is relatively easy to set up. It can assess the average dynamics of the water and solute dynamics and focus on the interactions among processes without resolving spatial details. The lumped approach can also accommodate basins with low data availability, and it can be easier for students to learn to use the model. In contrast, spatially explicit representations enable the exploration of the "hot spots" (e.g., swales and riparian zones with high soil water DOC concentrations in Figure 10e) and their contribution to stream chemistry at different times. Spatial heterogeneities in watershed properties (e.g., soil types and depth, lithology, vegetation, biomass, and mineralogy) are ubiquitous in natural systems. However, a general understanding of the linkage between local catchment features and catchment-scale dynamics (e.g., stream concentration dynamics and solute export pattern) is often lacking. We generally do not understand how spatial heterogeneity affects water flow paths, stream water chemistry, and biogeochemical reaction rates. The spatially distributed model provides a tool to further explore these questions. Ultimately, the choice of the model complexity level depends on the research questions that the model is set to answer. At the end, we all need to balance cost and gain when deciding to use a simple or complex model, striving to be "simple but not simplistic" (Beven and Lane, 2019)."

Similarly, what is the sensitivity to subdividing the land into many cells? Is having 100 cells warranted, or can you get just as good predictions just using a couple of cells describing the different land use types? (I understand the argument about identifying hot spots, but it could also be interesting to see if the subdivision has impact on the stream concentration predictions).

Response: This is an interesting upscaling-related question that deserves its own manuscript. With lumped model, one gets the simplicity and average dynamics and bound to lose spatial details. Our discussion above provides some guidelines. But the other two reviewers are also saying this paper is too long and has too much information so we are not adding more here. It would be interesting to have an independent paper asking just this upscaling question.

What is the calibration process like for the user? Are any autocalibration tools set up for the model?

Response: This is a good point. We have added a few sentences to describe the calibration part.

Line 490 – 497: "Auto-calibration is not built into the model, but a global calibration coefficient approach is used to reduce parameter dimension and facilitate manual calibration. A typical model application requires 20 to 30 hydrological parameters to be calibrated. These parameters include land surface parameters (e.g., canopy resistance, surface albedo), soil and geology parameters (e.g., hydraulic conductivity, porosity, Van

Genuchten, macropore properties) (Shi et al., 2013). Reaction-related parameters (e.g., reaction rate constant, mineral surface area, Q_{10}, $S_{w,c}$, and n) are additionally needed for calibration, the number of which depends on the numbers of reactions involved in a particular system."