Response to Reviewer #1

We would like to thank the anonymous referee for his/her interest and the comments on our manuscript. Below, reviewer comments are in italic font and our replies are in plain blue font.

The authors improved the manuscript from the previous version. However, I still struggle with a lack of clarity regarding what the toolbox and the models it can build are like. The choice of comparable models and applications seems – to me – rather different from the details of the model described by the authors. Could the authors try to still be clearer in this regard? I am sorry if this is annoying the authors, but I believe more clarity will help the readers.

We thank the reviewer for his/her helpful comments. However, we are a bit surprised, that the reviewer came up with completely new points, he/she did not mention in the first round.

(1) The authors still do a poor job in describing what the toolbox is for. If the toolbox is meant to run with forcing data in the order of minutes and spatial resolutions up to 25m2, then clearly the objective is not to build continental scale models. However, one must read significantly into section 2 before this becomes apparent. The example models used in the introduction section vary widely in model style (e.g. HYPE and HYDRUS). I think that the proposed model is a valuable addition to the current set of models by adding a very high resolution and small-scale model to the mix. Why not say so? Why do the authors not define what the model is for (and what it is not for)?

We clarify the description of the toolbox and elaborate the scale of application. We added the missing information about the model objectives to the introduction. We further stress the value added by RoGeR in the introduction. In lines 169f., we provide guidance on the spatial resolution. This should avoid potential misuse. Moreover, capabilities of the models are listed in section 2.2 and section 2.3. With the information about the considered processes and temporal resolution, any hydrologist should be able to judge if the model is suitable for a certain application he/she has in mind.

(2) The current motivation (in the abstract) suggests that the model is meant to improve water management. However, the model needs very high-resolution data on forcing (<10 minutes) and system properties (<25m2). This is not what a typical hydrological model for water management uses, even if it does include water quality. The model is much more like DHSVM [a] than to HYPE or other models used in the text. Instead, the model select the models they include by the language they have been written in, which is less helpful to understand what this specific model actually does.
[a] https://www.pnnl.gov/projects/distributed-hydrology-soil-vegetation-model

We modified the motivation in the abstract and refined the scale of the water management. We already provide a brief description of the capabilities of HYDRUS, EcH2O-iso and mHM in the beginning of the paragraph (see lines 35ff.). We include the programming language since knowledge in certain programming plays an important role for the application of the considered model.

(3) In lines 31ff. („The main reasons for this lack of reproducibility…") you cite Reinecke et al. (2022) several times in a single sentence. It is perfectly fine to just cite it once at the end of the sentence.

We merged the citation and moved the citation to the end of the sentence.

(4) The authors state in their response that "Modular means that certain processes (e.g. lateral subsurface runoff, capillary rise from groundwater, etc.) can be activated/deactivated." That is a unique definition used by the authors and rather different from the typical use of the term modular in the context of model building in hydrology, which is to define the option to represent the same process with multiple algorithms (e.g. Knoben et al. 2019, GMD). I would at a minimum expect that the authors clarify that their use of the term modular is different from the more common use. Though preferably the authors use a different and less confusing terminology.

Knoben et al. (2019) use the following definition of modularity:
"MARRMoT takes inspiration from earlier modular frameworks (e.g. FUSE, Clark et al., 2008; FLEX, Fenicia et al., 2011) and uses modular code with individual flux equations as the basic building blocks. Multi-model frameworks benefit from modular implementation because this simplifies the programming of the framework and makes it easier to (i) reuse components of a model in a different context, including cases in which the same basic equation is used by multiple models, and (ii) add new options to the framework."
Similarly, we provide the flux equations in process-specific modules that are used as basic building blocks.
Therefore, we argue that we do not deviate from definition.

(5) I am afraid that I am unclear about the organization proposed in table 1. Software complexity from low to high is used to organize guidelines for open software. I do like the guidelines, but I do not see how these points organize along a complexity axis. Online documentation or unit tests are only useful to high complexity software? I think unit tests should be done for any software of the type proposed here. If the authors keep the complexity organization, then they need to explain why these guidelines follow this logic.

We further explain the rationale behind the complexity organization in Section 2.1.

(6) The authors conclude by linking their model to others like VIC and the potential for integration with Earth Science Models. However, models like VIC and certainly Earth Science Models run at spatial scales of kilometers, where 1km2 resolution is considered hyper resolution. Your model runs at meters. Do the authors really suggest that VIC is the right model to compare their model against? I mentioned DHSVM earlier. DHSVM and VIC come from the same group but have been built to fulfill very different purposes. The model the authors describe is very much like DHSVM, so why do the authors compare it to VIC?

We removed the potential integration into earth science models. Instead, we link towards potential model coupling opportunities. We also removed the comparison to VIC und use DHSVM instead.

Response to Reviewer #2

We would like to thank the anonymous referee for his/her interest and the comments on our manuscript. Below, reviewer comments are in italic font and our replies are in plain blue font.

Authors have addressed all of my comments now. Thank you for including a real-world example in the paper, even though, it would have been better to use a watershed where real-world O18 were available. Also, I was referring to the assumption made about the bedrock conductivity in the paper. It would be nice to see a rationale for using a high value. Finally, some information about the study catchments such as climatological and hydrological conditions, soil properties and geology would be nice to include.

These are just some suggestions that the author may address. Otherwise, I recommend the publication of the paper.

We thank the reviewer for his/her helpful comments.

The rationale for using a high bedrock conductivity is the assumption of free drainage. We are assuming free drainage since knowledge about the lower boundary condition (i.e. the depth of the groundwater table) is not available. We clarify the assumption in Section 4. Regarding information about climatology, soil properties and land cover, we would like to refer to Section S3 in the supplement.