We thank the reviewer for their consideration of our manuscript. Below we discuss each comment in turn, outlining how we intend to change the manuscript in response or requesting clarifications where needed. Reviewer comments are shown in black, response in blue.

Kind regards,

Wouter Knoben
Diana Spieler

Major comments

Selection of the two catchments in this study: I find the selection of the two catchments a bit problematic. Mainly, the two catchments vary in several aspects besides the so-called aridity fraction (e.g. size). This makes comparison difficult.

As a general note, both models and catchments were selected from a much larger sample that consists of 36 conceptual bucket models of varying degrees of complexity calibrated for streamflow simulation in 559 catchments (Knoben et al., 2020). Readers may of course use or include their own catchments (or decide to choose different models). The two models and two catchments used and presented here are however specifically selected to convey the lessons described in section 2.1 (Learning Objectives) through a four-way comparison. The fact that the catchments vary in multiple aspects is in fact critical to two of the learning objectives we hope to convey, namely that (page 4, line 8):

"*Reinforcing the previous point, comparing the performance of model m03 across both catchments shows that the model achieves higher efficiency scores than model m02 in both places, while the catchments themselves are structurally very different (catchment descriptions are shown as part of the suggested exercises). This again shows that high efficiency scores are no guarantee of having used the "right" model.*"

And (page 4, line 12):

"*Choosing a model based on past performance should be done with care. Comparing the performance of model m02 across both catchments shows that the model performance is very different in both places and that having a "successful" model for one catchment is no guarantee that this model will perform equally well somewhere else.*"

We will add a summary of the intended learning objectives to the start of section 2.1, to clarify to the reader our intent of selecting these different catchments and models.

Furthermore, one of the catchments reports zero-flows. Here it is important to note that the used model variants are by design not able to simulate zero-flows.

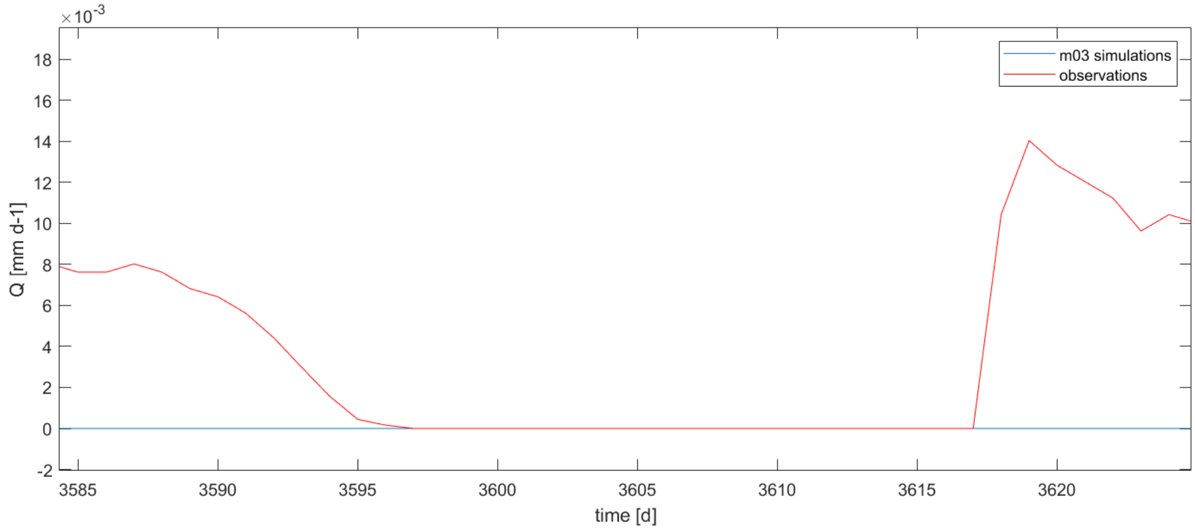
It is incorrect that our chosen models are not able to simulate zero flows. The threshold to flow generation in model m03 allows this model to produce zero flows (see also the figure below, obtained by running model m03 with data from catchment 08109700, showing zero flow simulations). That model m02 cannot produce zero flows is by design. Combined with using a catchment with occasional observed zero flows, this gives students a straightforward example of why choosing an appropriate model structure for a given catchment matters. This is key to the first learning objective (page 3, line 31):

“Model choice matters. Because all models are "hydrologic models" it is an easy assumption to make that the choice of model is largely one of taste or convenience, rather than one of suitability for the task at hand. Comparing the performance of both models in catchment c08109700 shows that this is not the case: the choice of model strongly affects the accuracy of obtained simulations.”

Students are asked to follow this line of reasoning in the document that describes exercise 2:

“Compare calibration and evaluation scores of both models for catchment 08109700 (Middle Yegua Creek). Based on what you know of this catchment's streamflow regime and the two model structures, which difference between the model structures do you think causes the difference in performance?”

We will update the description of the first learning objective to include that the occurrence of zero flows and a model’s (in)ability to produce these can be important.



The occurrence of zero-flows also makes the use of log-transformations for the computation of performance measures challenging.

Zero flows are indeed problematic for log-transformations. That said, our suggested exercises do not ask students to compute log flows and this is thus not an issue for the exercises as suggested. If one wishes to deviate from our suggested exercises and teach students how to log-transform flows, discussing the case of catchments with zero flows seems extremely relevant and having data from a catchment with zero flows readily available could be helpful with this. Therefore, we do not agree with the reviewer that the occurrence of zero-flows in one of these catchments is a major drawback to our setup and we do not agree with the implied suggestion that this catchment should be removed/replaced.

I am also a bit confused by the selection of the two model variants, why just these two?

As indicated in section 2.1, *"Both models and catchments have been specifically selected out of a sample of 40+ models and 500+ catchments for the lessons that can be conveyed by each comparative exercise."* These two models are sufficient to convey several important learning objectives to students. We will clarify at the start of section 2.1 that our purpose with this paper and these catchments & models is to convey general lessons about model structure and model evaluation, and that the two models and catchments should be seen as tools to achieve this. Detailed understanding of either models or catchments is not a goal in itself, beyond the understanding needed to grasp the learning objectives in section 2.1.

I am missing an evaluation of how successful the suggested module is. As it is now, basically the same claims that the authors make to motivate their module are also used to describe its success, which is not convincing. What would be needed is some form of evaluation by surveying students who took the class.

This is a very helpful suggestion. We circulated a survey among the students that included multiple questions to be answered on a 1-5 scale and three open questions. We currently summarize the response as a general statement that students found that the course *"was easy to follow and complete, and that the main messages were clear. Various attendees specifically noted that the exercises were helpful for better understanding the material covered during the seminar, […]"* (section 4.3, page 11, line 11). We will include an additional figure that shows responses to specific questions on the survey in an anonymized way and expand on the current discussion of the module's application at TU Dresden.

I am also missing information on how many students and with which background participated in the course in Dresden.

We will add a brief description of the hydrology curriculum at the TU Dresden and thus the expected background of the students who completed our survey to section 4.3.

The authors claim that their module could be added into 'any hydrology course with minimal effort' (P10L10). I'm afraid I have to disagree for several reasons:

1. If at all, then it can be added to courses in hydrological modelling, but not all hydrology courses.
2. If Matlab is not used in a particular class, including this module is by no means trivial
3. Teaching materials are not provided; this would be important as a service to a potential teacher who wants to adapt this module in their course.
4. The fixed selection of catchments and models might limit the utility of the module.

Regarding point 1, we make the implicit assumption that modelling is part of most hydrology courses. We realize that this is not necessarily correct and will rephrase the manuscript accordingly.

Regarding point 2, the prerequisite that "the module requires either Matlab or Octave, […]" already appears:

- In the abstract (page 1, line 10)
- At the end of the introduction (page 3, line 3)

- In the description of the MARRMoT framework (page 5, line 24)
- In the section “Software requirements” (page 7, line 10)

We will add this statement to the conclusion section as well to ensure this message is present in all locations a reader is likely to look for it. We will rephrase any sentences that talk about inclusion into existing curriculums as well to mention the need to have access to Matlab/Octave.

Regarding point 3, teaching materials are provided on GitHub, as indicated in the introduction, section 2.3.1 and the section *Code and data availability.* We will clarify that these materials do not include lecture slides, but they do provide:

- Prepared data (meteorological time series and data describing both catchments);
- Pdf’s and LaTeX source files for the two exercises described in section 3;
- An example script showing how to complete exercise 2;
- Calibrated parameter sets for all combinations of models and catchments, resulting in the calibration and evaluation results shown in Table 1.

This is sufficient to run the suggested exercises with minimal effort and for adaptation with new models or data by a teacher wishing to do so. If the reviewer disagrees, we would welcome more detailed comments about what they think is missing.

Regarding point 4, summarizing our earlier responses, these models and catchments were selected for the general lessons they can convey. We believe this is a good introduction into several important aspects related to model structure uncertainty. Given that both CAMELS data (for multiple countries) and the MARRMoT toolbox are freely available, those wishing to give their students an expanded experience can easily do so.

I would recommend describing the module first in generic terms. Both catchments and model variants could be left open to be selected as appropriate for a particular course. Forst of all, there is great value in using catchments that the students are familiar with. Using US catchments might not be the most pedagogical choice in many cases.

Briefly, this is true if one is trying to teach locally relevant hydrologic understanding. As argued before, for general understanding of difficulties relating to modelling, specifically selecting a few catchments and models precisely for their ability to convey this general understanding seems logical to us. While finding local examples that show our intended learning objectives would be great, this takes effort on part of the teacher (not to mention that local data or models may not be available at all). Our goal is to reduce the initial effort needed to teach these concepts at all. Those who wish to go beyond our provided setup are of course welcome to do so.

Furthermore, depending on which programming language/modelling frameworks are used in a course, it might also be more useful to use an alternative to the option presented here.

In cases where Matlab or Octave are not available, a teacher is of course welcome to simply use the learning objectives presented in our paper if they think them relevant enough to include. For those that do teach Matlab (as is still common in universities), this paper presents a useful tool. Teachers/Students requiring an open source software may also use the Octave code provided in the MARRMoT framework, avoiding the need for (expensive) licenses.

In a second step, a concrete implementation of the module could be described (=as it is described now) and guidance could be given on alternatives.

We assume guidance on alternatives is meant in the context of the comments above, e.g. meaning how to use different catchments, models or programming languages. This seems so broad to us that any guidance will be either obvious (e.g. “one can look for data from local catchments instead”) or unhelpful (e.g. “if Matlab or Octave are not available, one can consider converting this exercise to their programming language of choice”). We will consider expanding section 4.2 “Possible follow-up teaching topics” to cover this but it is currently unclear to us how to do so in a meaningful manner.

Finally, it is crucial to evaluate the module in some way (e.g. student survey before-after)

Agreed, and we will add this to the extent possible.

Minor comments

P3L26 (mathematically) accurate – I think you just mean ‘better’. Note that a model can be mathematically accurate but still totally useless.

The choice of “(mathematically) accurate” is deliberate; precisely to convey that a high NSE or KGE score does not necessarily mean a hydrologically useful model. To us, “better” implies the latter more than it does the former. We will rephrase this sentence to make our intent clear.

P5L10 Aridity fraction: please explain this term and how it is computed

We will add this as requested.

P8L2: The statement that instructions are straightforward is followed by a ‘fork and clone’ statement that might be not at all straightforward to most readers.

We will clarify that more detailed instructions regarding the forking and cloning of Github repositories are provided as part of our suggested exercise 1.

P10L4: formalize? Do you mean formulate?

We mean this in the sense of “formalize your thoughts” by writing them down.

P10L30: Does this mean it was an one-day course in practice?

Yes. We will rephrase this.

P11L16: sorry, but the choice of one single student can’t be really used as a convincing argument

We will either rephrase or remove this sentence.

Figure 2 is hard to read and needs to be improved. I am olso a bit wondering about the shown precip data, for me it does not look as if “on average 294 days have < 1 mm precipitation” from this figure

This is an unfortunate consequence of squashing 20 years of data into a few centimeters of graphic. We will consider the usefulness of this figure and change or remove as appropriate.

References

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. Water Resources Research, 56, e2019WR025975. https://doi.org/10.1029/2019WR025975