Dear authors,

Three referees have now reviewed your revised manuscript: one of them had already seen the earlier version of your manuscript, while the two others are new reviewers. While all three reviewers find the objective of your paper laudable, their recommendations are quite varied (minor revisions, major revisions, reject). I should note that two of the reviewers have extensive teaching experience, while the third one has a bit less teaching experience. The simple/simplistic way of describing model structural uncertainties is a concern noted by at least one reviewer, as well as the small number of participants involved in the testing/evaluation of the teaching module. One reviewer, in particular, suggests that more participants be involved in the testing and evaluation of the teaching module, and that learning goals be better assessed. Another reviewer is concerned about the lack of consideration of (input/output) data uncertainty and parameter uncertainty (among other elements) in the teaching module. Would there be a way for you to better articulate the role of other sources off uncertainty in the teaching module, either through additional exercises or a case study? These issues are quite critical in the context of an educational paper focused on modelling and should be addressed, which is why I am returning your manuscript for major revisions. Do not hesitate to reach out to me if you need more time to address the concerns raised by the reviewers.

With best regards,

Genevieve Ali

Dear editor, dear reviewers,

Thank you for your (continued) effort on this manuscript. Based on your comments, we have summarized the main points of improvement of our paper as follows:

1. Evaluation of the course can be improved by increasing the number of participants and better assessment of learning goals;
2. Model structure uncertainty is discussed only in isolation and the module does not discuss data uncertainty and parameter uncertainty.

In response to the reviewers' comments, we updated our evaluation procedure to closely match current practice at the University of Saskatchewan and earlier procedures as reported in the HESS Special Issue "Hydrology education in a changing world", as well as including a direct assessment of students' understanding of our intended learning goals. We ran the course again at the TU Dresden earlier this year and collected additional student responses. We updated the main manuscript and its Supporting Materials to describe this updated procedure and its outcomes. Further details are provided in the response to Reviewer 1.

Based on the reviewers' suggestions, we quantified the impact of four different sources of uncertainty on our proposed four-way model comparison that uses single calibrated parameter sets only. Our analysis covers data uncertainty, parameter uncertainty, sampling uncertainty in objective function calculation and subjectivity in the choice of objective function. The analyses indicate that, while there is some uncertainty related to all these aspects, the relative model performance that underpins the core

of our manuscript (i.e. that both models obtain similar KGE scores in one catchment, and very different scores in another) remains visible. We added a new discussion section 4.1 (pages 12-16) that describes these analyses and their outcomes, and added a suggestion to Section 2.4 “Integration in current curriculum” that teachers discuss these other sources of uncertainty in a concluding lecture after the students have performed the exercises, to ensure that students understand that differences in model performance as measured by efficiency scores can originate from multiple different causes.

We also made various minor changes to improve the manuscript’s flow and clarity.

Responses to individual reviewer comments are provided on the following pages, in blue.

Kind regards,

Wouter Knoben & Diana Spieler

1 Reviewer comments

1.1 Reviewer 1

I appreciate the efforts of the authors in revising their manuscript. Many of the previous issues have been addressed, …

Thank you for your continued effort in reviewing our paper. We appreciate the time you spent on this and your efforts in outlining where we can improve.

… but I am afraid, two rather fundamental issues became apparent now:

1) Number of participants: this information was missing before. Frankly, I was surprised to read that there were only 11 participants, of which only 4 were 'normal' students, 2 were PhD students and almost half were faculty members (just as a minor comment, later the authors write about postdocs, it seems the term faculty member was used in an unusual way here – this section has been rewritten, see below). This number is clearly very low and it would be important to evaluate the course and the different design issues on a broader basis.

As indicated in our earlier response "the typical number of people in the MSc Hydrology course at TU Dresden [is] usually between 8 and 15 students per cohort". We have re-run the course at TU Dresden in early 2022 and gained another 10 evaluation responses. For what it's worth, this is the number of responses the University of Saskatchewan considers a minimum before student evaluation results can be considered stable (see below). Please note that the course was an extracurricular activity students voluntarily added to their schedule at the end of their semester. Responses show that the group of attendees feels more prepared to deal with model structure uncertainty issues after attending the course than they did before.

The diverging understanding of what a faculty member is relates to the German academic system and was rewritten to avoid confusion. We now use the term "academic or scientific staff".

For instance, based on my personal experience I would still argue that using local catchments helps the ('normal') students to see connections between model and real world and is good for their motivation.

As stated in our previous response: "[to teach] 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. Our trial application for a German audience suggests that using catchments in the United States was not detrimental to conveying these learning objectives and appetite among the audience for inclusion of more catchments was low." Our second edition of running this course for a German audience reinforces this point.

2) The authors now added an 'evaluation' of the course. Again, numbers are small (only 8), but more importantly: I don't think the value/effect of the course can be evaluated by this kind of questions. These questions evaluate more how happy the participants were with the course, but do not really evaluate what they learned.

These two points are rather fundamental. I strongly recommend that the authors test and evaluate their course with more participants and a better assessment of the learning goals before this work is published.

In response to the reviewers' comments, we updated our evaluation procedure to closely match current practice at the University of Saskatchewan and earlier procedures as reported in the HESS Special Issue "Hydrology education in a changing world", as well as including a direct assessment of students' understanding of our intended learning goals. We ran the course again at the TU Dresden earlier this year and collected additional student responses. We updated the main manuscript and its Supporting Materials to describe this updated procedure and its outcomes. Details of the analysis we used to inform this new procedure are given below.

Summary of approach and changes to evaluation procedure

To inform the new evaluation procedure, we collected various examples of existing evaluation procedures to better understand how learning is typically assessed. We have based this analysis on the HESS Special Issue "Hydrology education in a changing world" and on the University of Saskatchewan's internal course evaluation procedures, in the hopes of capturing both how evaluation is handled in other published manuscripts and in more practical settings.

Evaluation procedures (if present at all) in the articles in the HESS special issue are a mix of quantitative and qualitative evaluation approaches. Quantitative approaches appear exclusively based on students' self-assessment of what they learned. Such approaches are in line with the University of Saskatchewan's procedures, which also rely on self-reported assessments.

Major changes to our evaluation procedure are as follows:

- We now used a "before" and "after" survey, which lets us assess changes in student response and relate these to course effectiveness;
- We retained the 5-point answering system, because this is the same approach used at the University of Saskatchewan, and in Habib et al. (2012) and AghaKouchak et al. (2013) – the only examples in the HESS special issue on teaching that use a quantitative approach to evaluation of their proposed educational materials;
- Our new evaluation questions contain a combination of:
    - Questions related to general competence (e.g. "Through this course I gained knowledge and confidence in the general area of hydrological modelling");
    - Self-reported assessment of changes in competence (e.g. "Before/after the course, how familiar are you with model structural uncertainty");
    - Direct author-led assessment of the intended learning outcomes (e.g. "Would you agree that a model that works well in one catchment will also work well in another catchment?" – assessed before and after attending the course).

Evaluation procedures at the University of Saskatchewan

The University of Saskatchewan uses an internal course evaluation system that lets students self-report their experiences with the courses they attend. Evaluation outcomes are considered "stable" if $n > 10$. Students are asked to answer the following questions on a 5-point numeric scale:

- The course provided me with a deeper understanding of the subject matter

- I found the course intellectually stimulating
- The instructor created an environment that contributed to my learning
- Course projects, assignments, tests, and/or exams improved my understanding of the course material
- Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material
- Online-specific:
    - Online tools used to support course activities were easy for me to use. These activities could include: accessing content, submitting assignments, completing quizzes, accessing results/grades, etc.
    - The organization of online activities in the course was clear and easy to follow.
    - The online environment enriched or strengthened my learning of the course objectives/competencies.
    - The expectations for this online/remote course were made clear.
    - The instructor maintained a regular, engaged presence during online activities throughout the course.
- Overall, the quality of my learning experience in this course was: …
- Opportunity to add open comments

Evaluation procedures in the HESS special issue

Underlined text indicates paper titles, the contents of which are summarized as bullet point lists.

An educational model for ensemble streamflow simulation and uncertainty analysis

Reference: (Aghakouchak, Nakhjiri and Habib, 2013)

Ensemble modelling with HBV in Matlab. Contains student feedback & evaluation. Evaluation approach:

- Very brief description of student background (2 sentences)
- Anonymous survey (n = 56), with 5 options per question (1-5). Questions:
    - As a result of your work with this education toolbox in the class, what gains did you make in each of the following?
        - Hydrologic modeling in general
        - Water budget analysis
        - Rainfall-runoff processes, their mathematical formulations and the required calculations to estimate the flood resulting from a given precipitation event
        - The effect of evapotranspiration on rainfall-runoff processes, its mathematical formulation and the required calculations
        - Model calibration and ensemble simulation
        - Sensitivity analysis
        - Differences between empirical and physically-based parameters
        - Enthusiasm for the subject of hydrologic modeling and analysis
        - Confidence in performing hydrologic modeling
    - How each of the following aspects and attributes of the developed teaching tool contributed to your learning gains?
        - The use of a practical case study with actual data

- The use of hands-on calculations in the lecture
- The fact that you could change the model parameters and their effects
- The requirement of a hydrologic modeling project using this hands-on toolbox

HydroViz: design and evaluation of a Web-based tool for improving hydrology education

Reference: (Habib et al., 2012)

Online educational tool. Contains evaluation (n_student = 182, n_teacher = 6). Evaluation questions:

- How effective is the conceptual design and software features of HydroViz in facilitating students' learning and delivering the embedded educational contents on hydrologic concepts and related skills?
- What are students' perceptions of various features and characteristics of HydroViz?
- What are students' perceptions of HydroViz as a part of the curriculum?
- How effective is HydroViz in developing freshmen engineering students' interest in hydrology as a subject area?
- Do students in different classes and universities differ in their learning of the hydrologic concepts and perceptions of HydroViz?
- What can be done to improve HydroViz?

Answer to questions obtained from:

- Tasks given to students as part of exercises and correctness of answers assessed;
- Online survey where students were asked to quantify their own knowledge gains on a 5-point scale, using 17 statements;
- Informal interviews

Computer-supported games and role plays in teaching water Management

Reference: (Hoekstra, 2012)

Board game about stakeholder interaction. No evaluation to speak of.

Web 2.0 collaboration tool to support student research in hydrology – an opinion

Reference (Pathirana, Gersonius and Radhakrishnan, 2012)

Wiki for use during thesis projects. Evaluation through qualitative "Students' / educators' impressions" (n_student = 5, n_teacher = 1?), based on an interview of 5 open questions. Not anonymous. Full responses in supplement.

Water management simulation games and the construction of knowledge

Reference: (Rusca, Heun and Schwartz, 2012)

Ravilla simulation game to introduce people to Integrated Water Resource Management problems. Qualitative evaluation only.

Teaching hydrological modeling with a user-friendly catchment-runoff-model software package

Reference: (Seibert and M. J.P. Vis, 2012)

Stand-alone HBV model with suggested exercises. No formal evaluation performed. Evaluation limited to a few qualitative statements in conclusions.

Irrigania – a web-based game about sharing water resources

Reference: (Seibert and M. J. P. Vis, 2012)

Browser game about stakeholder interaction. No formal evaluation of learning goals, apart from qualitative statements about observed behavior shown by students.

1.2 Reviewer 3

I have been pondering about this manuscript for quite a while and I am still not sure what really to make of it. Quite clearly, the authors are right in underlining the importance of model structure uncertainty or, more pointedly, our past (and ongoing) failure to formulate general, catchment-scale theories (and thus models) from available data (cf. Nearing et al., 2021). Similarly, I agree with the authors that there are multiple (interlaced) facets to uncertainty, many of which are far from straightforward to grasp – in particular for many students at Master level.

Thank you for these positive thoughts about the usefulness of our work.

However, I am nevertheless surprised by the suggestions made by the authors and I am concerned that they could end up doing disservice to our students. The reason for this is the overly simplistic and informal – almost leisurely – way to define, explain and demonstrate uncertainties in the suggested experiment.

As I understand it, the experiment consists of a simple application of a modular modelling framework, in which 2 models are selected (why these ones?) …

As described in Section 2.1, *"[b]oth 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."* The chosen models and catchments, when calibrated against streamflow observations from both catchments, achieve the Kling-Gupta Efficiency scores shown in Figure 1. Four-way comparison of these scores leads students to our chosen learning objectives.

… and their runs with optimum(?) parameter sets are compared.

This is correct and you are right to point out that this was not particularly clear in the earlier version of this manuscript. We have clarified the caption of Figure 1 and the corresponding text in section 2.1 to specifically mention that this concerns a single calibrated parameter set, and that calibration is performed by the students using adaptations of existing scripts in the MARRMoT repository.

How, in the absence of (1) any quantification of uncertainties related to data (or the use thereof), (2) any meaningful quantification of the uncertainties in the parameters or (3) any considerations of the impact of the choice of performance metric, can the authors suggest that the differences in the models' skill to reproduce stream flow is indeed linked to model structural errors? I believe that this conveys a way too simplistic view of uncertainty to the students. What will they take from such an example? Without any representation of the other sources of uncertainty, the risk is that many students may learn from that example that all the differences between the models are due to uncertainty in structure of the deterministic model. That is of course wrong.

Although I welcome the authors very laudable intention to sensibilize students for different sources of uncertainty, I believe that this requires a much more in-depth analysis.

It is good to know you see merit in this particular piece of work. You are right that central to our manuscript is a four-way comparison of the performance of calibrated parameter sets for two models in two catchments. In one catchment, models perform similarly in terms of KGE scores while in the other catchment model performance is very different. Based on your comments, we investigated:

- The impact of data uncertainty through modifying the data used for model calibration and evaluation;
- The impact of parameter uncertainty through:
    - First confirming through Latin Hypercube sampling that our calibration algorithm returns solutions in those regions where highest model performance is found through sampling;
    - Then assessing whether using any of the 100 best parameter sets identified through sampling would substantially alter relative model performance in either catchment.
- The impact of sampling uncertainty in calculation of the objective function;
- The impact of using a different objective function.

The analyses indicate that, while there is some uncertainty related to all these aspects, the relative model performance that underpins the core of our manuscript (i.e. that both models obtain similar KGE scores in one catchment, and very different scores in another) remains visible.

The analyses summarized above are discussed in the new section 4.1 (pages 12-16).

We added a suggestion to Section 2.4 “Integration in current curriculum” that teachers discuss these other sources of uncertainty in a concluding lecture after the students have performed the exercises, to ensure that students understand that differences in model performance as measured by efficiency scores can originate from multiple different causes.

Please also note that although I am teaching hydrology at Master level (including classes and hands-on examples on different uncertainties), I do not consider myself as educational expert. I therefore cannot provide a valid assessment of the educational value of the suggested experiment that goes beyond my hydrology-related concerns about the experiment and my 20 years in-class experience.

1.3 Reviewer 4

This is a useful contribution towards teaching structural uncertainty based on model comparative analysis. Please follow my detailed comments in the attachment.

This work presents a set of (open source - Octave and licensed-Matlab) computational exercises that help teach hydrological model structural uncertainty, particularly model choice as an example of structural uncertainty. As the analysis and coverage of structural uncertainty are limited, this work is a useful contribution. Below are my comments and suggestions.

Thank you for your time and the comments on our work. It is encouraging to read that you see merit in this work.

1. Model adequacy is closely related but different from structural uncertainty (Gupta et al., 2011). Although the authors have indicated the limitation of statistical metrics such as KGE in diagnosing model adequacy (page 4 line 3-18), they used the term adequacy in their core objective plot (Figure 1 – lessons in the three boxes). The work (and Figure 1) is based on the relative performance of two models in two catchments. As such, 'adequate model performance' is not the right phrase to use. I suggest the use of relative terms such as 'better' and/or 'a relatively high' performance.
   Similarly, it is a stretch to use strong words such as 'appropriate' and 'accurate' (on page 4 line 26 and line 31) based on comparative analysis.
   Thank you for this comment. We agree with this sentiment and have made multiple changes throughout the manuscript (including Figure 1) to be more precise in our use of language. We decided to retain the use of the word "accurate" as this describes the mathematical (mis)match between simulations and observations. We added a new sentences to Section 2.1 to explicitly define these terms:

   "In the remainder of this work, we refer to the calculation of efficiency scores as the accuracy *of a model's simulation, in the sense that simulations with higher efficiency scores more accurately resemble observations than the simulations from models with lower efficiency scores. This is contrasted by the term* adequacy *which is more commonly used to refer to a model's degree of realism (see e.g. Gupta et al., 2011)."*

2. The manuscript needs to explain why model 'm03' performs better in the two distinctly different catchments while 'm02' performs poorly in one of the catchments. Although the manuscript mentions simulating zero flows and the basis of the models' development, it is important to briefly discuss these points directly. This may support both educators and students to articulate the causes.
   We agree that such a discussion would clarify the manuscript. We have added the necessary explanations to Section 3.2 where the relevant exercise is discussed.

3. In this work, 'calibrated' parameters are used to support the comparative analysis. But, it is important to indicate/discuss the non-uniqueness of these parameters and the interplay of parameter and structural uncertainties (Clark et al., 2011; Moges et al., 2021). As separating the two uncertainties is not always straightforward, a brief discussion with references for further

reading will be helpful.
We investigated the impact of various sources of uncertainty on our proposed exercise and added those findings to the new Section 4.1. We also added a suggestion in Section 2.3 that teachers discuss these concepts in a concluding lecture after the students completed the proposed exercise, so that the students may be in a better position to appreciate these (reasonably complex) concepts.

Technical comments:

1. It is good not to repetitively use the term "this section describes". If necessary, it is enough to use it once (e.g., the first case on page 3 line 15 – 20). Using this term in other places (e.g., page 3 lines, 24 - 26; page 4 line 2; page 5 line 11) is just a distraction.
   Thank you for this comment. We have found that descriptions such as these are useful to manage reader expectations – particularly at the start of section 2 - but agree that we may have overdone things a bit. We removed the mentions on page 4 line 2 and page 5 line 11 to streamline the text.

2. Page 2 line 21, avoid the use of the term 'For a variety of reasons'. State a few of the reasons or rewrite the sentences.
   We have rewritten these sentences as follows: *"Regrettably, suitability of a given model for the task at hand is not always the main driver in model selection. Prior experience with a given model combined with lacking insights into model strengths and weaknesses often lead to a certain attachment of hydrologists to their model of choice (Addor and Melsen, 2019). Hands-on experience with model structure uncertainty in a classroom setting, particularly through exercises that show that the choice of model can have a strong impact on the quality of simulations for a given catchment, will prepare students to think beyond their 'model of choice'. This will prepare students for when they will need to design modeling studies or interpret modeling results in their future careers."*

3. Page 4 lines 19 - 31 referred the catchments and models by their CAMELS and model ID. It is better to first introduce the catchment names, ID and the models' names earlier. Perhaps, on page 3 lines 14 – 15 where the objective of the paper and the experimental designs are indicated.
   We moved the sentences used to introduce "Section 2.2 Catchments and models" to the suggest location, and so ensure that the reader knows the catchment and model names and IDs before the reach the learning goals in Section 2.1.

References:

Clark, M.; Kavetski, D.; Fenicia, F. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 2011, 47.

Gupta, H.V.; Clark, M.P.; Vrugt, J.A.; Abramowitz, G.; Ye, M. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 2012, 48.

Moges E, Demissie Y, Larsen L, Yassin F. Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis. Water. 2021; 13(1):28

2 References

Aghakouchak, A., Nakhjiri, N. and Habib, E. (2013) 'An educational model for ensemble streamflow simulation and uncertainty analysis', *Hydrology and Earth System Sciences*, 17(2), pp. 445–452. doi:10.5194/hess-17-445-2013.

Habib, E. *et al.* (2012) 'HydroViz: design and evaluation of a Web-based tool for improving hydrology education', *Hydrology and Earth System Sciences*, 16(10), pp. 3767–3781. doi:10.5194/hess-16-3767-2012.

Hoekstra, A.Y. (2012) 'Computer-supported games and role plays in teaching water management', *Hydrology and Earth System Sciences*, 16(8), pp. 2985–2994. doi:10.5194/hess-16-2985-2012.

Pathirana, A., Gersonius, B. and Radhakrishnan, M. (2012) 'Web 2.0 collaboration tool to support student research in hydrology – an opinion', *Hydrology and Earth System Sciences*, 16(8), pp. 2499–2509. doi:10.5194/hess-16-2499-2012.

Rusca, M., Heun, J. and Schwartz, K. (2012) 'Water management simulation games and the construction of knowledge', *Hydrology and Earth System Sciences*, 16(8), pp. 2749–2757. doi:10.5194/hess-16-2749-2012.

Seibert, J. and Vis, M. J. P. (2012) 'Irrigania – a web-based game about sharing water resources', *Hydrology and Earth System Sciences*, 16(8), pp. 2523–2530. doi:10.5194/hess-16-2523-2012.

Seibert, J. and Vis, M. J.P. (2012) 'Teaching hydrological modeling with a user-friendly catchment-runoff-model software package', *Hydrology and Earth System Sciences*, 16(9), pp. 3315–3325. doi:10.5194/hess-16-3315-2012.