

Interactive comment on “Flood hazard and change impact assessments may profit from rethinking model calibration strategies” by Manuela I. Brunner et al.

Anonymous Referee #2

Received and published: 13 June 2020

This is a well-written journal with appropriate content for HESS. I think this is a nice study, though I do have some suggestions related to the framing of the work and its discussion. This could be a very nice paper if the focus was actually on the calibration strategy.

My comments are:

[1] The title of the study suggests a wide-ranging assessment of different calibration strategies in the context of flood modelling. However, the study is essentially an assessment of the value of using KGE for flood modelling. The actual focus is fine, but I think it should be reflected in the title of the manuscript to avoid confusion.

C1

[2] Given that the focus of the manuscript is on the calibration strategy, I was surprised to not find any details on what strategy was used to find the best KGE values? What algorithm was used etc would be helpful information for the reader to understand what has been done. While this might be covered in previous papers in detail, it would be good to see at least some basic description here as well.

[3] It would also be helpful to have some calibration/validation results for each model to distinguish them at this point already (if they differ?).

[4] Section 3.1: Why is HBV so poor? Especially given its focus on snow/cold regions?

[5] Section 3.1: I am a bit confused by this assessment. Are you assessing the model or the metric used for calibration? The paper title suggests that the focus is on the calibration strategy, so my question is why using the same calibration strategy results in different model performance? Significant differences between very similar models is surprising if the models have been calibrated in the same manner.

[6] Lines 224-225: But how do you know that if you only assessed one metric? The authors do a very nice job of including multiple models, but if the focus is on the calibration strategy, then why do you not include variability in how they calibrate the models? How can you make conclusions about the calibration strategy if you did not vary it. Would putting more weight on fitting the variability have produced a better fit to variability (using a weighted KGE)? You have this as a discussion point, but why is this not part of your actual study?

[7] Line 236: But how do you know that? Maybe all the models have the same problem regardless of calibration metric used? Maybe you did not look hard enough for an optimum parameter set?

[8] Line 245: As stated above, I find it dissatisfying to make such a conclusion. Testing this suggestion is a very minor effort given the work already presented in this paper. Why can the authors not try this? This – to me – would be part of the main tests the

C2

authors should have done in this paper. You cannot test the implications of choices about the calibration strategy if you do not test different choices. Using multiple models does not compensate for this omission.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., <https://doi.org/10.5194/hess-2020-192>, 2020.