Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-192-RC1, 2020

Hydrology and Earth System Sciences
Discussions

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

Anonymous Referee #1

Received and published: 26 May 2020

General Comments

This paper assesses how well hydrological models calibrated using the Kling-Gupta Efficiency (KGE) metric can reproduce local and regional flood characteristics. Streamflow simulations from four hydrological models are evaluated across a large sample of hydrologically varied catchments, for flood timing, magnitude and spatial variability. In addition, the authors explore the model sensitivities of high flows to precipitation and temperature. This is an interesting analysis and helps to explain model deficiencies for hazard and change impact assessments.

I enjoyed reading this paper, which is well-written, concise and easy to follow. The

figures are all relevant and well-presented. My main concern is that the title, and focus on deficiencies of integrated calibration metrics, does not accurately reflect the study. I think the title suggests that different model calibration strategies are going to be implemented and evaluated, or that there is going to be some assessment of performance for different calibration strategies. The study only looks at models calibrated using KGE, and it is therefore hard to distinguish if any failure of the models in representing flood characteristics is due to the calibration strategy or other factors such as quality of input and observed streamflow data, or model structural errors.

Overall, I think this paper would make an interesting contribution for HESS, following changes and clarifications to the manuscript. I have several specific comments which I outline below.

Specific comments

Title: as discussed in general comments, I am not sure the title best reflects the content of the paper. I think this title suggests evaluation of different calibration strategies, whereas KGE has been used throughout. A title focusing on key results/ what has been done (e.g. Evaluating hydrological model suitability for flood impact assessments across a large sample of catchments) may be more suitable.

Line 10: "Our results show that both the modelling of local and spatial flood characteristics is challenging." It could be helpful to highlight some the key results in the abstract to justify this statement, i.e. all models under predict the magnitude of events.

Line 12: "We conclude. . ." The manuscript focuses on models calibrated on KGE alone, and infers that deficiencies in model performance is due to the model calibration. I think this is quite a big leap – as there are other factors which could result in poor model performance (e.g. errors in observed precipitation and river flow data, particularly for peak flow events). It would be good to discuss these within the manuscript.

Introduction:

Line 25: There is a tendency for high values to be underestimated and low values to be overestimated (Gupta et al. 2009), but I am not sure it is correct to say that the optimal value actually underestimates flow variability. It could be worth mentioning that NSE is often used for high flow studies, based on the idea that by using squared errors it mostly constrains peaks and high flows (Mizukami et al. 2019).

Line 33: I do not completely follow this sentence – why non-flood-related signature?

Data and Methods: Whilst the methods section is clear overall, I felt that a few sections needed clarifying.

Line 68: It would be useful to add references for the models.

Line 68: It would be helpful to know some more about the differences/similarities between the models. In particular, any differences in modelling decisions that may contribute to the performance differences (it would be good to explain why HBV does so poorly compared to the other models). Perhaps a table of key differences or a figure giving model structure diagrams would be helpful.

Line 70: "model parameters were calibrated on streamflow observations by minimising the EKG" – How was the optimisation performed (e.g. which algorithm was used) and is this the same in both studies? Was mHM calibrated using multiscale parameter regionalisation, and if so was EKG evaluated across the region rather than for each catchment? It would be useful to know how the calibration differed, despite all being based on KGE.

Line 80: How do these meteorological forcing data differ? Are they both the same timestep?

Line 67-83: The dates used for the simulations are unclear. In the method a few different date ranges are given: Line 67: "we use daily streamflow simulations for the period 1981-2008", Line 82: "SAC, HBV and VIC were evaluated on the period 1985-2008", "mHM was calibrated on the period 1999-2008 and evaluated on the period

1989-1999." It seems that 1981-1985 were not used in the previous studies. It would be useful to know which period the model simulations were actually run for, whether a warm-up period has been given, and how long the warmup was. Also, over which period were SAC, HBV and VIC calibrated? Does the period 1981-2008 refer to hydrological years or calendar years? It would be helpful to give months here.

Line 85: Have you used the KGE values given by Mizukami et al. (2019) and Melsen et al. (2018), or were these re-calculated these over the period 1981-2008? I assumed all model performance was calculated over the same period, against the same observed discharge data, but this is not clear. Line 85: I agree that performance is generally lowest for catchments with intermittent regimes, but there is a lot of overlap in performance.

Line 114: "we then use the data sets resulting from Step 2 to evaluate how models reproduce overall and seasonal spatial flood dependence." It would be useful to have a bit more detail in this section. How was the error statistic calculated?

Line 117: It is not clear if 1% was the value used. This should be made clear, and it would help to have a reference/justification for why this value was chosen.

Line 122: "Time of concentration is typically less than one day for small headwater basins." This needs a reference.

Results: A key advantage of this study is the application of multiple model structures to a large sample of catchments. Throughout the methods/results it would be useful to have more discussion of the differences between the models. In particular, it would be useful to know why HBV performs so poorly compared to the other models for flood magnitudes.

Line 145: "For most catchments, the number of flood events is relatively well simulated by most models. . ." It would be useful to know the number of observed events, to put these errors into context. I am assuming that the number of events is similar between

all regime types due to the selection of the threshold. Otherwise a percentage error may be easier to interpret.

Line 150: Underestimation of peak flow is attributed to the KGE metric underestimating variability, and spatially lumped model inputs. This could also be due to data errors – for example, McMillan et al. (2012) show that there can be large uncertainties associated with precipitation products. It would be useful to add this to the discussion. McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrological Processes, 26(26), 4078-4111.

Line 152: "the use of lumped forcings may also artificially synchronize hydrologic response, which would lead to overestimation." – could this be further explained?

Line 163: "the overestimation of spatial dependence in winter is likely related to higher simulated than observed snowmelt." I was not sure which regime(s) this comment was referring to. The melt regime is the only one that doesn't show an overestimation of spatial dependence in winter for any model.

Line 170: "Connectedness overestimation is most pronounced. . ." I don't agree with this sentence. For the other 3 models intermittent regime does not seem to be overestimated any more than other regime types. In winter, it seems to be in line with and below all regimes for SAC, VIC and mHM.

Line 177: "There is a clear positive relationship . . ." I would not say there is a clear positive relationship for SAC. Perhaps a slight positive relationship.

Line 180: "soil moisture and event magnitude are also positively related. . ." I would interpret this a little differently for VIC - at full saturation we see events of all magnitudes. It is just the upper level of flows that is increasing with soil moisture. 'lower values' - does this mean lower values of peak Q?

Line 183: I think this is also the case for SAC.

Line 186: "for SAC and VIC." I would add that to some extent this is also the case for HBV.

Line 199: Why is this the case?

Line 211: "These relationships are, however, not necessarily captured by the models. . ." It may be worth highlighting that in some areas these relationships are generally captured by the models: e.g. weak winter regime broadly captures the precipitation relationship, and New-Year's regime captures precipitation and temperature relationship.

Figure 6: This figure has a lot of text, which can be distracting from the plots. I think it would be help to simplify the y axes and colorbar scales to 2 significant figures (i.e. no decimal places).

Figure 6: It would be clearer if the colour scales matched between the observations and simulations for a specific catchment, and also the x and y axis ranges. Otherwise, it would be useful to point out that the scales differ, and explain why this has been done, i.e. colours ranging from the largest to smallest flood.

Line 221: I do not follow this link -could this be explained better? It feels like there is a jump from models inadequately representing the sensitivity of peak flows to precipitation to errors in precipitation data being the cause.

Line 223: ".. may be related to insufficient model calibration. . ." This feels like quite a big leap. Having only looked at models calibrated using KGE it doesn't feel like there is enough information to attribute poor performance to calibration metrics. Could it be the model structures more generally, or the input data errors, that are causing these model deficiencies rather than the calibration metric?

Figure 7: It would be helpful to have a more thorough explanation of this figure. Perhaps a sentence explaining that positive values mean an increase in the variable leads to an increase in peak flows, and values falling on the dotted line indicate simulations match

observations.

Line 226: This sentence implies an underestimation in timing. Only absolute errors in day of flood timing are given, not the direction of change within the year. Maybe rephrase this sentence.

Conclusions:

Line 235: In the introduction a key aim is 'assess which aspects of hydrological models may need to be improved ....' and 'identifying and documenting model weaknesses regarding regional and future flooding will highlight advances for future model development.' .. These aims/questions could be more directly addressed in the conclusions section.

Technical corrections

Line 86: "successfully" should be "success"

Line 160: "underestimates" should be "underestimate"

HESS Review Checklist

In the full review and interactive discussion, the referees and other interested members of the scientific community are asked to take into account all of the following aspects:

1) Does the paper address relevant scientific questions within the scope of HESS? YES

2) Does the paper present novel concepts, ideas, tools, or data? YES

3) Are substantial conclusions reached? YES

4) Are the scientific methods and assumptions valid and clearly outlined? YES

5) Are the results sufficient to support the interpretations and conclusions? MOSTLY

6) Is the description of experiments and calculations sufficiently complete and precise

to allow their reproduction by fellow scientists (traceability of results)? YES

7) Do the authors give proper credit to related work and clearly indicate their own new/original contribution? YES

8) Does the title clearly reflect the contents of the paper? NO

9) Does the abstract provide a concise and complete summary? YES

10) Is the overall presentation well structured and clear? YES

11) Is the language fluent and precise? YES

12) Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? YES

13) Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? MINOR CLARIFICATIONS TO METHODS

14) Are the number and quality of references appropriate? YES

15) Is the amount and quality of supplementary material appropriate? YES

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