## Reviewer 2

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.

**Reply:** *Thank you for pointing out the mismatch between the title and the analyses performed. This point was also raised by reviewer 1 and we changed the title to: 'Evaluating the suitability of hydrological models for flood impact assessments.' This rephrasing removes the emphasis from model calibration, whose effect on model simulations has been assessed by Mizukami et al.* (2019).

[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.

**Reply:** *We specified that: 'The model parameters were calibrated on streamflow observations by minimizing the 1-$E_{KG}$ by Melsen et al.* (2018) *using Sobol-based Latin hypercube sampling* [*Bratley and Fox*, 1988] *for SAC, HBV, and VIC and by Mizukami et al.* (2019) *for mHM using multi-scale parameter regionalization where the transfer function parameters were identified using the dynamically dimensioned search algorithm* [*Tolson and Shoemaker*, 2007]*.'*

[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?).

**Reply:** *We provide validation results for each of the four models in Figure 2 and specify that: 'Overall model performance decreases from mHM (median $E_{KG}$ 0.69), over SAC (median $E_{KG}$ 0.63) and VIC (median $E_{KG}$ 0.60) to HBV (median $E_{KG}$ 0.52).' So yes, the models are already different if we just look at $E_{KG}$ before considering any specific flood metric.*

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

**Reply:** *It is difficult to say why exactly HBV is performing worse than the other three models in reproducing flood characteristics. We think that: 'The overestimation of the number of events by HBV may be explained by its fast response to precipitation as expressed through its model parameter b, which introduced non-linearity to the system* [*Viglione and Parajka*, 2020]*.' and added this statement to the text.*

[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.

**Reply:** *We agree that the choice of the initial title could cause some confusion. Instead of comparing different calibration strategies as done in previous studies* [*Mizukami et al.*, 2019]*, we compare the representation of floods by different models calibrated with the same objective*

*function. As mentioned above, we have changed the title in order to eliminate focus on the calibration strategy itself. Our results show that even if models are calibrated using a calibration metric supposedly putting a lot of weight on high flows, they may not necessarily well represent local and regional features of floods.*

[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?
**Reply:** *As wrongly suggested by our initial title, the focus of this study is not supposed to be on the calibration strategy as the effect of the choice of an objective function on the quality of modeled flood flows has previously been assessed by Mizukami et al.* (2019). *They show that $E_{KG}$ leads to a better model performance with respect to flood flows than $E_{NS}$, which is very often recommended for calibrating a model aimed at simulating flood peaks/high flows. We show that even if one uses the metric found to lead to the best flood representation by Mizukami et al.* (2019)*, the reproduction of flood characteristics may still leave much to be desired. We rephrased this sentence to: 'We illustrate that reliance on an individual calibration metric ($E_{KG}$) can lead to simulation performance deficits for phenomena of interest, including an underestimation of streamflow variability and peak flood magnitudes and a misrepresentation of timing'*

[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?
**Reply:** *Our results show that models do not perform equally well in simulating flood characteristics when calibrated with the same objective function. We therefore think that the statement 'Our model comparison shows that all flood characteristics are not equally well represented by models calibrated with the widely used Kling–Gupta efficiency metric' is justified. We acknowledge that these limitations may not solely be related to model structure: 'These limitations may be related to input uncertainties* [*Te Linde et al.*, 2007]*, equifinality in process contributions for simulations with (very) similar efficiency scores, leading to an inability to unambiguously identify the appropriate relative process contributions* [*Khatami et al.*, 2019] *or insufficient model calibration* [*Fowler et al.*, 2016].

[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 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.
**Reply:** *As discussed above, the focus of this study was not supposed to be on a comparison of different model calibration strategies even though our initial title may have suggested otherwise. Rather, we wanted to show that using a calibration metric commonly recommended for model calibration in the case one is interested in floods may still lead to suboptimal model results. The development of an objective function targeted at optimizing local and spatial flood characteristics would be a study in itself. This is why we leave potential ways of improving calibration strategies to the discussion.*

**References used in this response to the reviewer**

Bratley, P., and B. L. Fox (1988), Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator, *ACM Trans. Math. Softw.*, *14*(1), 88–100, doi:10.1145/42288.214372.

Fowler, K. J. A., M. C. Peel, A. W. Western, L. Zhang, and T. J. Peterson (2016), Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, *Water Resour. Res.*, *52*, 1820–1846, doi:10.1111/j.1752-1688.1969.tb04897.x.

Khatami, S., M. C. Peel, T. J. Peterson, and A. W. Western (2019), Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty, *Water Resour. Res.*, *55*(11), 8922–8941, doi:10.1029/2018WR023750.

Te Linde, A. H., J. Aerts, H. Dolman, and R. Hurkmans (2007), Comparing model performance of the HBV and VIC models in the Rhine basin, in *Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management*, pp. 278–285.

Melsen, L., N. Addor, N. Mizukami, A. Newman, P. Torfs, M. Clark, R. Uijlenhoet, and R. Teuling (2018), Mapping (dis) agreement in hydrologic projections, *Hydrol. Earth Syst. Sci.*, *22*, 1775–1791, doi:10.5194/hess-22-1775-2018.

Mizukami, N., O. Rakovec, A. J. Newman, M. P. Clark, A. W. Wood, H. V. Gupta, and R. Kumar (2019), On the choice of calibration metrics for "high-flow" estimation using hydrologic models, *Hydrol. Earth Syst. Sci.*, *23*(6), 2601–2614, doi:10.5194/hess-23-2601-2019.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*(1), 1–16, doi:10.1029/2005WR004723.

Viglione, A., and J. Parajka (2020), TUWmodel: Lumped/Semi-Distributed Hydrological Model for Education Purposes, *TUWmodel Lumped/Semi-Distributed Hydrol. Model Educ. Purp.*, https://cran.r-project.org/web/packages/TUWmodel/i. Available from: https://cran.r-project.org/web/packages/TUWmodel/index.html (Accessed 25 June 2020)