Reviewer 1

General comments

This paper assesses how well hydrological models calibrated using the Kling-Gupta Efficiency (KGE) metric can reproduce local and regional flood characteristics. Stream-flow 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.

Reply: *Thank you very much for acknowledging the value of our work and for highlighting the need to revise the title. We did indeed not evaluate different calibration strategies. As the title may suggest otherwise, we revised it.*

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.

Reply: *Thank you for indicating that the title did not well reflect the content of the manuscript. We replaced the title by: 'Evaluating the suitability of hydrological models for flood impact assessments'.*

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.

Reply: *We highlight some key results in the abstract such as 'Our results show that both the modeling of local and spatial flood characteristics is challenging as models underestimate flood magnitude and flood timing is not necessarily well captured.'*

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.

Reply: *Thank you for highlighting that other factors influencing model performance merit more attention. We extended the discussion on input uncertainty (l.195-199) by streamflow observation uncertainty: 'In addition, model performance may depend on the uncertainty of streamflow observations* [*McMillan et al.*, 2010] *used for calibrating and evaluating the model or on input uncertainty, i.e. the precipitation product used to drive the models* [*Te Linde et al.*,

2007]*). Precipitation products may underestimate extreme rainfall or the spatial dependence of extreme precipitation at different locations because spatial smoothing or averaging during the gridding process reduces variability* [*Risser et al.*, 2019]*.*

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).
Reply: *We think that it is justified to talk about an underestimation of flow variability as an underestimation of high flows and an overestimation of low flows implies an underestimation of flow variability. We mentioned that NSE is widely used for high flow studies and added that using square errors enables focusing on high flows: 'For example, one widely-used metric that is considered integrative compared to others (e.g., bias, correlation) is the Nash-Sutcliffe efficiency (NSE), where the sum-of-sqaures error metric focuses attention on high flow.'*

Line 33: I do not completely follow this sentence – why non-flood-related signature?
Reply: *Thank you for pointing out the need for rephrasing. We rephrased the sentence to: 'The use of multiple objectives may, however, lead to a decrease in performance with respect to any individual flow signature not considered as an objective.*

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.
Reply: *We repeat the references provided in the introduction in the methods section.*

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.
Reply: *Thank you for this suggestion. We have created model structure diagrams to aid in the interpretation of between-model differences, which are provided in the appendix of the manuscript. Reproducing the model equations is infeasible, mainly because of the length of the VIC and mHM code. Instead of reproducing the equations here, we have updated the text with references to where the source code of each model may be found.*

The model diagrams are based on:

- *SAC-SMA: analysis of the model's description[National Weather Service NOAA, 2002]: https://www.nws.noaa.gov/oh/hrl/general/chps/Models/Sacramento_Soil_Moisture_Accounting.pdf.*
- *TUW HBV: analysis of the model's source code [Viglione and Parajka, 2020].*
- *VIC: descriptions of VIC in [Melsen et al., 2018; Melsen and Guse, 2019].*
- *mHM: analysis of the model's source code (https://git.ufz.de/mhm/mhm/-/tree/5.7) and a diagram provided in [Kumar et al., 2010].*

We hypothesize 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 introduces non-linearity to the system [*Viglione and Parajka*, 2020]*.'*

Line 70: "model parameters were calibrated on streamflow observations by minimizing 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.
Reply: *We specified that Melsen et al. (2018) used Sobol-based Latin hypercube sampling [Bratley and Fox, 1988] to calibrate VIC, HBV, and SAC. We also specified that mHM was calibrated by Mizukami et al.*, (2019) *for each basin individually using multi-scale parameter regionalization where the transfer function parameters were identified using the dynamically dimensioned search algorithm* [*Tolson and Shoemaker*, 2007].

Line 80: How do these meteorological forcing data differ? Are they both the same timestep?
Reply: *Both forcing datasets are at a daily resolution and both gridded datasets were derived from observed precipitation and temperature. However, the Maurer dataset with 12km has a coarser resolution than the Daymet dataset with 1km. We added these specifications to the text.*

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 hydro-logical years or calendar years? It would be helpful to give months here.
Reply: *The final analysis was performed on model simulations for the period 1981-2008 for all models. As the model simulations were generated in two different, prior studies, their calibration and evaluation periods do not match as indicated by the different year ranges provided in the text. However, we here used the period 1980-1981 as a spin-up period for all models and performed the analysis on the period 1981-2008. We specified that: 'All four models were finally run for the period 1980-2008 (calendar years), where the period 1980-1981 was used for spin-up and therefore discarded from the analysis.' We also specified that the period 1981-2008 refers to calendar years.*

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.
Reply: *The KGE values were not recalculated over the period 1981-2008, we used the original values provided by Mizukami et al. (2019) and Melsen et al. (2018). We indicate that mHM was evaluated over the period 1989-1999 while the other models were evaluated over the period 1985-2008.*

Line 85: I agree that performance is generally lowest for catchments with intermittent regimes, but there is a lot of overlap in performance.
Reply: *We add a note on this stating: 'However, there is a high within-class variability in model 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?
Reply: *Thank you for pointing out the need for clarification. We specified that: ‘We computed actual errors in flood connectedness by subtracting observed from simulated connectedness over all seasons and per season.’*

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.
Reply: *We followed the procedure introduced by Brunner et al.* (2020) *to define flood connectedness. We rephrased the sentence to make this clear: ‘Following the definition used by Brunner et al.* (2020)*, a catchment is connected to another catchment if they share a certain number of events. We here used an event threshold of 1% of the total or seasonal number of events to define connectedness (all months: 12 events, seasons: 3 events).*

Line 122: “Time of concentration is typically less than one day for small headwater basins.” This needs a reference.
Reply: *Besides catchment area, time of concentration also depends on other factors such as rainfall intensity or geology. We therefore made the sentence slightly less specific and provide a reference to a the book chapter by USDA-NRCS* (2010)*.*

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.
Reply: *We specify in the Methods section that ‘This procedure results in a first quartile of 36, a median of 40, and a third quartile of 47 events identified per basin.’ This indicates a relatively small variability in the number of events chosen per basin and justifies the use of actual errors. We provide a model schematic for each of the models considered in this study to aid the interpretation of model differences. We reason 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 introduces non-linearity to the system.’*

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 uncertain-ties for hydrology: rainfall, river discharge and water quality. Hydrological Processes,26(26), 4078-4111.
Reply: *We totally agree that underestimation may also result from data errors. We add the reference to McMillan et al. (2012) to our discussion about the influence of other uncertainty sources than model and parameter choice on flood characteristics (l. 195-199). We specify that: ‘Precipitation products may show observation uncertainties* [*Mcmillan et al.*, 2012] *and underestimate extreme rainfall or the spatial dependence of extreme precipitation at different*

locations because spatial smoothing or averaging during the gridding process reduces variability [*Risser et al.*, 2019]*.*

Line 152: "the use of lumped forcings may also artificially synchronize hydrologic response, which would lead to overestimation." – could this be further explained?
Reply: *We removed this sentence as it referred to underestimation of spatial dependence rather than magnitude and therefore did not fit into this section.*

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.
Reply: *We specify that: 'The overestimation of spatial dependence in winter for all regimes except the melt regime is likely related to…'*

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 over-estimated any more than other regime types. In winter, it seems to be in line with and below all regimes for SAC, VIC and mHM.
Reply: *We agree that this sentence was not correct. We rephrased it to: 'Connectedness overestimation by HBV is most pronounced for catchments with an intermittent regime. Otherwise, connectedness over-/underestimation seems to be independent of the regime.'*

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.
Reply: *We weakened the sentence by deleting the word 'clear'.*

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?
Reply: *Yes, we specified that we were referring to peak values.*

Line 183: I think this is also the case for SAC.
Reply: *Yes, SAC lies somewhere in between. We add the following sentence: 'Such low precipitation inputs can also lead to high peak discharge for SAC but to a lesser degree than HBV and mHM.'*

Line 186: "for SAC and VIC." I would add that to some extent this is also the case for HBV.
Reply: *We added: 'and to a lesser degree for HBV'.*

Line 199: Why is this the case?
Reply: *Future precipitation estimates are particularly uncertain because of climate model and scenario uncertainty* [*Lopez-Cantu et al.*, 2020]*, which was specified in the text.*

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.

Reply: *We rephrased this section to: 'While these relationships are captured for some catchments (e.g. Blackwater River, weak winter regime or Tucca Creek, New Year's regime), they aren't in other catchments. The simulated sensitivities may even point in another direction than the observed ones (e.g. Pacific Creek, melt regime).'*

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).
Reply: *We agree that Figure 6 would profit from 'decluttering'. We reduced the number of digits displayed to 1 wherever possible and added colored boxes to improve the reading flow.*

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.
Reply: *We chose different color scales and axes for the observed and simulated floods to compare gradients rather than differences in magnitudes. Plotting the grids on the same grid and using the same colors would lead to non-centered grid clouds and to very weak colors in the case floods are underestimated. We adjusted the figure caption and point out that the different grids are shown on different scales: 'Grid axes and grey scales differ between plots where darker colors indicate higher flood magnitudes.'*

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.
Reply: *We agree that the link between the two sub-sentences was not evident. As we discuss precipitation errors as a potential source elsewhere in the manuscript (l.195-199), we removed its mention from here.*

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?
Reply: *Yes, model structure and input data uncertainty are definitely also part of the story. We acknowledge this in the newly phrased paragraph: 'The results of this study indicate that the hydrological models used in this study have limited capability in reproducing observed hydrologic sensitivities during flooding. 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]*.'*

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.
Reply: *Thank you for pointing out the need for clarification. We added that: 'Positive and negative values indicate positive and negative associations of precipitation and temperature with peak flow, respectively. Values on the dashed line indicate correspondence between observed and modeled sensitivity gradients.'*

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.
Reply: *We rephrased this to: 'including an underestimation of streamflow variability and peak flood magnitudes and a misrepresentation of timing.'*

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.
Reply: *We try to more explicitly address these aims by adding: 'We therefore conclude that the representation of magnitude, timing and spatial connectedness can be improved.'*

Technical corrections
Line 86: "successfully" should be "success"
Reply: *We think that the phrasing is correct and retained successfully.*

Line 160: "underestimates" should be "underestimate"
Reply: *We removed the 's'.*

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?
YES12) Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?
YES13) 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
Reply: *We changed the title and added a few clarifications to the methods section as discussed in detail in the individual comments above.*

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.

Brunner, M. I., E. Gilleland, A. Wood, D. L. Swain, and M. Clark (2020), Spatial dependence of floods shaped by spatiotemporal variations in meteorological and land-surface processes, *Geophys. Res. Lett.*, *47*, e2020GL088000, doi:10.1029/2020GL088000.

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.

Kumar, R., L. Samaniego, and S. Attinger (2010), The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics, *J. Hydrol.*, *392*(1–2), 54–69, doi:10.1016/j.jhydrol.2010.07.047.

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.

Lopez-Cantu, T., A. F. Prein, and C. Samaras (2020), Uncertainties in future U.S. extreme precipitation from downscaled climate projections, *Geophys. Res. Lett.*, *47*(9), 1–11, doi:10.1029/2019GL086797.

Mcmillan, H., T. Krueger, and J. Freer (2012), Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, *Hydrol. Process.*, *26*(26), 4078–4111, doi:10.1002/hyp.9384.

McMillan, H., J. Freer, F. Pappenberger, T. Krueger, and M. Clark (2010), Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, *Hydrol. Process.*, *24*(10), 1270–1284, doi:10.1002/hyp.7587.

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.

Melsen, L. A., and B. Guse (2019), Hydrological drought simulations: How climate and model structure control parameter sensitivity, *Water Resour. Res.*, 1–21, doi:10.1029/2019wr025230.

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.

National Weather Service NOAA (2002), *Conceptualization of the Sacramento Soil Moisture accounting model*.

Risser, M. D., C. J. Paciorek, M. F. Wehner, T. A. O’Brien, and W. D. Collins (2019), A probabilistic gridded product for daily precipitation extremes over the United States, *Clim. Dyn.*, *53*(5–6), 2517–2538, doi:10.1007/s00382-019-04636-0.

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.

USDA-NRCS (2010), Time of concentration, in *National Engineering Handbook: Part 630 Hydrology*, pp. 1–15, U.S. Department of Atriculture (USDA), Fort Worth.

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)