Reviewer 3

Though I agree with the title of the manuscript and with the main conclusions on l.254-259 (see below), there are several serious deficiencies in the approach and interpretation of results. The study looks like an initial stage only: let us calibrate four models for many relatively small catchments in USA using one metric, EKG, to see, how well the models will reproduce local flood characteristics and spatial aspects of flooding, and how well would they be prepared for climate impact assessment. The conclusion is that the models calibrated on the Kling–Gupta efficiency alone have limited reliability in flood hazard assessments. Such "negative" result could be expected, as there are several recent publications pointing on a necessity of comprehensive approaches for hydrological model calibration and evaluation (for mean flow and for extremes) and especially if they are intended for climate impact assessment (see e.g. Choi and Beven, 2007, Coron et al.,2012, Refsgaard et al., 2013, Thirel et al., 2015, Krysanova et al., 2018). Therefore, such an "initial stage" of the study should be supplemented by application of an extended approach: for example, including at least some of the further steps suggested in the papers listed above, like multi-site and multi-variable calibration (mentioned in the manuscript), DSS test checking for contrasting climate sub-periods, testing specifically for indicators of interest, i.e. for high flows and floods. Then the study would be much more valuable. There are also other deficiencies in the applied approach and in the interpretation of the obtained results. Therefore, the manuscript should be rejected in its present form.

Reply: *Thank you for your time reviewing our manuscript. We agree that the results highlight model deficiencies in the representation of local and regional flood characteristics particularly under climate conditions different from the current ones. We also agree that some previous studies have tried to highlight the necessity to evaluate model transferability to conditions different from the ones we live in. Still, we are surprised by how many studies (including some of our own studies) use E_{NS} or E_{KG} as a single calibration and evaluation metric for flood studies (for E_{NS} based calibration see e.g.* [*Hundecha and Merz*, 2012; *Köplin et al.*, 2014; *Vormoor et al.*, 2015; *Wobus et al.*, 2017] *and for E_{KG} based calibration see e.g.* [*Brunner and Sikorska*, 2018; *Hirpa et al.*, 2018; *Huang et al.*, 2018; *Thober et al.*, 2018; *Harrigan et al.*, 2020]*. The aim of this study is to clearly communicate to the hydrologic modeling community that such a focus on a single metric may not result in an accurate representation of flood characteristics, particularly not in a spatial and climate change context. Our study should contribute to expanding awareness of such issues within a field that we observe continues to rely (too) strongly on the E_{KG} and like metrics alone. We focused on E_{KG} because a previous study by* [*Mizukami et al.*, 2019] *has shown that E_{KG} results in a more reliable representation of peak discharge than E_{NS}.*

In all, we feel the presentation of multiple analyses of different aspects of model behavior for four models and hundreds of locations to shed further light on aspects of the simulations that are present but not directly indicated from a single E_{KG} score (which may be high) goes beyond an initial analysis. We do pay particular attention to the representation of spatial flood characteristics and the suitability of model setups in simulating floods under climate conditions different from the ones in the observations. To do so, we perform a resampling-based sensitivity analysis focusing on peak-over-threshold values, which has similar aims as the differential split sample test. We try to better explain this similarity by adding the following description to the introduction of this methodology: 'To do so, we look at how models translate changes in event temperature and precipitation into changes in POT discharge by performing a resampling-based sensitivity analysis. This sensitivity analysis aims at evaluating whether a model is still reliable under climate conditions different from the ones used in model calibration similar to split-sample or differential split-sample calibration/validation schemes [Coron et al., 2012; Refsgaard et al., 2014; Thirel et al., 2015].'

Besides highlighting potential modeling challenges related to the representation of floods, our discussion section points out potential avenues for further model improvement including the use of more tailored model calibration strategies, giving more weight to the variability component of an integrative metric, optimizing explicitly for key flood characteristics (e.g., peak flow, volume, timing) and/or metrics depicting the fidelity of the model representation of soil moisture and snowmelt, using a multi-objective model calibration process, or considering spatially distributed features of model response within a spatial calibration framework. This is thought as an encouragement to researchers who work on the development of innovative calibration techniques rather than an attempt to propose actual alternative calibration metrics ourselves. While we share the view that an expanded study which goes on to test the hypothesis that multi-objective, signature-aware and other types of calibration approaches would lead to more suitable models for flooding and change studies, adding that evidence to this study is not feasible given the substantial effort and time involved.

Other major concerns:

APPROACH

l. 81-82: were driven with Daymet meteorological forcing (Thornton et al., 2012) and mHM with the forcing by Maurer et al. (2002):→how are they comparable with the observed climate? Was the comparison done or not? If not, it would be reasonable to do.

Reply: *Thank you for pointing out the need for clarification. Both the Daymet and Maurer datasets represent current climate conditions, were derived from observed precipitation and temperature, and have been shown to result in similar mean daily precipitation fields* [*Newman et al.*, 2015]*. We specified that 'All the models were driven with daily, spatially lumped meteorological forcing data representing current climate conditions: SAC, HBV, and VIC were driven with Daymet meteorological forcing (1km resolution) and mHM with the forcing by Maurer et al. (2002) (12km resolution) both derived from observed precipitation and temperature. So yes, they represent observed climate.*

l. 82: SAC, HBV, and VIC were evaluated on the period 1985–2008:→and calibrated for which period?

Reply: *Melsen et al.* (2018) *ran the three models using a large number of parameter sets for the period 1985-2008. These parameter sets were generated by first sampling 100 base runs based on the average parameter values. Subsequently, they sampled each parameter 100 times by applying perturbations to the base runs. This implies that for each of the 605 basins, SAC was run 1900 times, VIC 1800 times, and HBV 1600 times. From these runs, we here chose the best parameter set in terms of* E_{KG}*, which represents the calibration step in a wider sense as the definition of calibration is identifying parameters. This procedure does not correspond to a classical calibration-validation scheme where the model is evaluated over a validation period independent of the calibration period but rather to a sampling procedure.*

l. 110-112:→would be good to express the relative error in %, and define thresholds for acceptable performance (e.g. based on literature) for all 3 indicators. For example, is a relative error of 25% acceptable or not? The thresholds could be shown in Fig. 3 by horizontal lines to enable distinguishing the good/acceptable and poor performances. Sec. 3.1:→to discuss performance based on the pre-defined thresholds

Reply: *We expressed the relative errors in Figure 3 in %. Defining a threshold for acceptable performance is a great idea. However, such an acceptability threshold is likely to depend on the problem at hand and a general threshold is therefore difficult to define.*

l. 116-118: a catchment is connected to another catchment if they share a certain number of events, i.e. at least 1% of the total or seasonal number of events:→is 1%of shared events really sufficient to define their connectivity??? Due to that, the whole section 3.2 is questionable.
Reply: *Thank you for expressing your concern and highlighting the need for clarification. We provide additional information on how many flood events were in the data set and how this translates into thresholds used: 'To do so, we use the connectedness measure introduced by Brunner et al.* (2020)*, which quantifies the number of catchments with which a specific catchment co-experiences floods. The number of concurrent flood events for a pair of stations is determined based on a data set consisting of the dates of flood occurrences across all catchments. This set is converted into a binary matrix which specifies for each catchment whether or not it is affected by a certain event. The matrix compiled using observed streamflow time series contained 1164 events among which 258 occur in winter, 291 in spring, 324 in summer, and 291 in fall. 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).' These values are similar to the absolute values used by Brunner et al.* (2020) *who used thresholds of 10 events for the annual and 5 events for the seasonal analysis. We consider these thresholds high enough to avoid defining a pair of stations as connected coincidentally.*

l. 127: we generate surrogate time series of temperature, precipitation, and streamflow for each catchment by resampling the available hydrological years with replacement:→the procedure is not quite clear, and should be better explained!
Reply: *Thank you for pointing out the need to provide more specifics on the sampling strategy. We specify that: 'To generate these series, we randomly sample a series of years with replacement in the period 1981-2008 which we use to compose time series consisting of the daily values corresponding to these years for each of the three variables'*

l. 202-203: "to assess each model' suitability for climate impact assessments on floods":→how the resampling could help to assess suitability? It would be better to test for contrasting climate sub periods, or to compare trends in discharge, high flows and POT series.
Reply: *This resampling procedure allows us to look at whether the models react to changes in mean event temperature and precipitation in the same way as the real world system. E.g. if higher observed event precipitation results in higher observed peak discharge, this should ideally be reflected in the modeling system which should show higher peak discharge for events with higher precipitation (i.e. the gradients derived from the observed and simulated response surfaces should be similar). If the model does not reproduce this behavior, its process representation in terms of floods is probably not ideal. We agree that differential split sample testing would be another way of looking at how transferable a model is to climate conditions, which differ from the ones used for calibrating and validating the model* [*Seibert*, 2003]*. However, we think that our resampling procedure goes beyond a split sample test because it enables analyzing gradients in the P-T-Q space instead of just comparing two periods that might differ with respect to certain characteristics. Within the observation period (1981-2008) less than 5% of the 488 catchments show statistically significant trends in POT values according to the non-parametric Mann-Kendall test. A comparison of observed vs. simulated trends is therefore not going to be a very useful evaluation metric with respect to the transferability of the model to changed climate conditions.*

Sec. 3.3 and Fig.5:→Maybe to add correlation coefficients to better characterize the relationships?

Reply: *Thank you for this suggestion. We added Kendall's rank correlation coefficients to each of the subplots to characterize the relationships between the pair of variables.*

INTERPRETATION

l. 145-146: "For most catchments, the number of flood events is relatively well simulated by most models":→this is not evident, if a threshold is not defined. It is only visible that medians are close to zero for three models, and there is no under- or over-estimation for the whole set of 40 – 176 catchments, but nothing more! After defining the threshold, the interpretation could be different! Besides, it would make sense to normalize over the number of catchments in every regime? And it would be reasonable to cut Y scale for (i) at -50 and +50, even if one box for HBV will not be fully visible.

Reply: *Thank you for these suggestions. We scaled the axis of panel (i) to -50 and +50. We indicate the number of catchments per regime in the figure caption to highlight that not all regimes have the same sample size. However, we do not understand your suggestion to normalize as each catchment represents one data point forming the boxplot. We agree that a threshold of model acceptability would be desirable and think that such a threshold would depend on the problem at hand. It is therefore difficult to define a generally valid threshold separating bad from good model performance. To not make any specific judgement, we rephrase the sentence and specify the actual error ranges for each of the models rather than talking about good and bad model performance in the updated version of the manuscript: 'For most catchments, the median deviation between the simulated and observed number of flood events lies close to zero (SAC: -3 events, HBV: -1, VIC: -1, mHM: 0). However, the simulations result in over- and underestimations of the number of events depending on the catchment (1st and 3rd quartiles for SAC: -9, 4; HBV: -8, 15; VIC: -7, 6; mHM: -6, 6). The overestimation is strongest for HBV, which overestimates the number of events for catchments with intermittent, weak winter, and melt regimes.' To still provide some guidance for the reader, we included different thresholds in Figure 2. For each model and regime, we broke up the results into three categories (and boxplots): all catchments, catchments with $E_{KG} > 0.5$, and catchments with $E_{KG} > 0.7$. For the last two categories, we provide the percentage of catchments in the regime under consideration exceeding the respective threshold.*

l. 158: Over all seasons, most models show an acceptable performance (i.e. median error close to zero):→if the median error is close to zero, it does not mean that most models show an acceptable performance!!! It only means that there is no tendency to over- or underestimation for catchments in five regimes, nothing more!

Reply: *Good point. We agree that a median error of zero does not necessarily imply acceptable model performance as positive and negative errors can cancel out. We rephrase this sentence in a neutral tone: 'Over all seasons, most models show a median error close to zero for flood connectedness. Flood connectedness can be over- and underestimated dependent on the catchment by most of the models while HBV overestimates spatial dependence in most catchments particularly in the Western part of the US.'*

l. 156: Over all, there is no clear tendency of one model to perform better than the other ones. →Based on thresholds, this could be better visible.

Reply: *This might be true with respect to a specific application, where one is interested in a specific regime or flood characteristic. We here intended to make a statement valid independent of a problem and therefore refrained from setting a 'somewhat arbitrary' threshold for model performance. As shown in Figures 3 and 4, SAC, VIC, and mHM perform similarly well regarding most of the flood characteristics assessed here and we therefore think that this statement is valid. We add that: 'However, there are slight differences in model performance which suggests*

that a 'most suitable model' could be identified for a specific application at hand, where a certain region or variable is of interest.'

l. 224-225: reliance on an individual calibration metric (EKG) rather than a broader suite of performance metrics can lead to simulation performance deficits for phenomena of interest, including an underestimation of streamflow variability:→Not only the metric, but the calibration approach is general!!!
Reply: *Yes, we agree with the reviewer, that the calibration approach/strategy, which includes the selection of time periods for training and validating model skill, screening for sensitive model parameters, and selecting a number of model evaluation metrics, includes a large number of subjective choices and might influence model results. We therefore based this manuscript on previously published work and we believe best possible calibration settings given past computer and resources availability. By changing the title and removing the work "calibration" from it, we believe to have removed the focus on any new calibration exercise/strategy.*

l. 238: the number of flood events in a simulation time series, which tend to verify well:→disagree, see above!
Reply: *We agree that a more nuanced statement is required here and rephrased the sentence to: 'The number of floods, flood magnitude, and timing are not always well captured by hydrological models in many catchments. The number of flood events were over- or underestimated depending on the catchment, flood magnitudes were underestimated by all models in most catchments, and the ability of the model to accurately reproduce event timing was proportional to the hydroclimatic seasonality.'*

l. 245-246: Such focus could be improved by giving more weight to the variability component of EKG →or including indicators of extremes in the calibration/validation!!!
Reply: *Yes, we tried to express this by writing: or by using a suite of appropriate and targeted metrics in a multi-objective framework. We rephrased this to: 'or by including indicators of extremes in in a multi-objective framework when calibrating and validating the model.'*

Minor corrections needed:
Fig. 1: catchments are indicated by the gauge location?
Reply: *Correct. We clarify this in the figure caption: 'Map of the 488 catchments in the conterminous United States belonging to the five regime classes indicated by their gauge location.'*

Fig. 2: for which period(s) is this statistics?
Reply: *These values refer to the period 1981-2008, which was specified in the figure caption.*

l. 112: circular statistics???
Reply: *We specify that: 'circular statistics are suitable for defining central tendencies of variables with a cycle* [*Burn*, 1997]*.*

Fig. 3: to explain what is represented by each box with whiskers: comparison for all catchments in a regime over which period: 1981-2008? To add this to the caption.
Reply: *We clarify in the figure caption that: 'The errors were computed over the period 1981-2008', that we looked at 'mean' errors for magnitude and timing and that 'The boxplots are composed of one value per catchment belonging to the respective regime class.'*

l. 159-160: particularly in the Western part of the US:→not, in the middle part (intermittent regime)
Reply: *The Western part is actually correct. Because we do not explicitly show this, we removed this sub-sentence though.*

I agree with the authors on the following:
l. 222-223: The results of this study indicate that the limited capability of hydrological models used in this study to reproduce observed hydrologic sensitivities during flooding may be related to insufficient model calibration: FULLY AGREE!
l. 247: The spatial concern could be addressed by applying spatial calibration procedures:→Agree!
l. 254-256: We conclude that calibration using only an individual model performance metric or variable can result in model implementations that have limited value for specific model applications, such as local and in particular spatial flood hazard analyses and change impact assessments: AGREE!
l. 258: more comprehensive multi-objective and multi-variable calibration strategies are needed: AGREE!
Reply: *We are glad that we have some common ground here.*

References
Choi and Beven, 2007, doi:10.1016/j.jhydrol.2006.07.012
Coron et al.,2012, doi:10.1029/2011WR011721
Refsgaard et al., 2013, doi:10.1007/s10584-013-0990-2
Thirel et al., 2015, doi:10.1080/02626667.2015.1050027
Krysanova et al.,2018, DOI: 10.1080/02626667.2018.1446214

References used in this response to the reviewers

Brunner, M. I., and A. E. Sikorska (2018), Dependence of flood peaks and volumes in modeled runoff time series: effect of data disaggregation and distribution, *J. Hydrol.*, *572*, 620–629, doi:10.1016/j.jhydrol.2019.03.024.

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.

Burn, D. H. (1997), Catchment similarity for regional flood frequency analysis using seasonality measures, *J. Hydrol.*, *202*, 212–230.

Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resour. Res.*, *48*(5), 1–17, doi:10.1029/2011WR011721.

Harrigan, S., E. Zsoter, L. Alfieri, C. Prudhomme, P. Salamon, C. Barnard, H. Cloke, and F. Pappenberger (2020), GloFAS-ERA4 operational global river discharge reanalysis 1979-present, *Earth Syst. Sci. Data*, (January), 1–23.

Hirpa, F. A., P. Salamon, H. E. Beck, V. Lorini, L. Alfieri, E. Zsoter, and S. J. Dadson (2018), Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data, *J. Hydrol.*, *566*(September), 595–606, doi:10.1016/j.jhydrol.2018.09.052.

Huang, S., R. Kumar, O. Rakovec, V. Aich, X. Wang, L. Samaniego, S. Liersch, and V. Krysanova (2018), Multimodel assessment of flood characteristics in four large river basins at global warming of

1.5, 2.0 and 3.0 K above the pre-industrial level, *Environ. Res. Lett.*, *13*(12), 124005, doi:10.1088/1748-9326/aae94b.

Hundecha, Y., and B. Merz (2012), Exploring the relationship between changes in climate and floods using a model-based analysis, *Water Resour. Res.*, *48*(4), doi:10.1029/2011WR010527.

Köplin, N., B. Schädler, D. Viviroli, and R. Weingartner (2014), Seasonality and magnitude of floods in Switzerland under future climate change, *Hydrol. Process.*, *28*(4), 2567–2578, doi:10.1002/hyp.9757.

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.

Newman, A. J., M. P. Clark, J. Craig, B. Nijssen, A. Wood, E. Gutmann, N. Mizukami, L. Brekke, and J. R. Arnold (2015), Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States, *J. Hydrometeorol.*, *16*(6), 2481–2500, doi:10.1175/jhm-d-15-0026.1.

Refsgaard, J. C. et al. (2014), A framework for testing the ability of models to project climate change and its impacts, *Clim. Change*, *122*(1–2), 271–282, doi:10.1007/s10584-013-0990-2.

Seibert, J. (2003), Reliability of Model Predictions Outside Calibration Conditions, *Nord. Hydrol.*, *34*(5), 477–492.

Thirel, G., V. Andréassian, and C. Perrin (2015), De la nécessité de tester les modèles hydrologiques sous des conditions changeantes, *Hydrol. Sci. J.*, *60*(7–8), 1165–1173, doi:10.1080/02626667.2015.1050027.

Thober, S., R. Kumar, N. Wanders, A. Marx, M. Pan, O. Rakovec, L. Samaniego, J. Sheffield, E. F. Wood, and M. Zink (2018), Multi-model ensemble projections of European river floods and high flows at 1.5, 2, and 3 degrees global warming, *Environ. Res. Lett.*, *13*(1), doi:10.1088/1748-9326/aa9e35.

Vormoor, K., D. Lawrence, M. Heistermann, and A. Bronstert (2015), Climate change impacts on the seasonality and generation processes of floods – projections and uncertainties for catchments with mixed snowmelt/rainfall regimes, *Hydrol. Earth Syst. Sci*, *19*, 913–931, doi:10.5194/hess-19-913-2015.

Wobus, C., E. Gutmann, R. Jones, M. Rissing, N. Mizukami, M. Lorie, H. Mahoney, A. W. Wood, D. Mills, and J. Martinich (2017), Climate change impacts on flood risk and asset damages within mapped 100-year floodplains of the contiguous United States, *Nat. Hazards Earth Syst. Sci.*, *17*, 2199–2211, doi:10.5194/nhess-17-2199-2017.