Dear Dr. Peleg,

Thank you very much for the thorough assessment of our manuscript and for your invitation to resubmit our manuscript to HESS. We are glad that you and the three reviewers acknowledge the value of our approach. We also appreciate the three constructive reviews which helped to reframe the storyline and to clarify methodological details.

Please find our detailed answers to the reviewers' comments in our point-by-point response below. We hope that you find the revised version of our manuscript suitable for publication in HESS.

On behalf of all co-authors,

Manuela Brunner

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.*
Modification: Title

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'.*
Modification: Title

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.’*
Modification: l.9-10

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 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]*.’*
Modification: l.222-226

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-squares error metric focuses attention on high flow.’*
Modification: l.24-27

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, however, can lead to a decrease in performance with respect to any individual flow signature not considered as an objective.*
Modification: l.33-35

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.*
Modification: l.71-74

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]*.'*
Modification: Appendix A: Model illustrations; l.169-171

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].
Modification: l.74-77

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.*
Modification: l.86-90

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.*
Modification: l.90-92

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.*
Modification: l.89-90

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.'*
Modification: l.95

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.'*
Modification: l.132-133

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).'*
Modification: l.130-133

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).
Modification: l.137

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.'*
Modification: l.111-112; 169-171

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. 224). 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].
Modification: l.224-226

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…'*
Modification: l.188-189

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.'*
Modification: l.194-195

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'.*
Modification: l.203-304

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.*
Modification: l.207

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.'*
Modification: l.210

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'.*
Modification: l.213

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.*
Modification: l.227

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).'*
Modification: l.240-241

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.*
Modification: Figure 6

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.'*
Modification: Figure 6 caption

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 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]*.'*
Modification: l.251-254

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.'*
Modification: Figure 7 caption

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.'*
Modification: l.257

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.'*
Modification: l.273-274

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.*

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).
Modification: Title

[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]*.'*
Modification: l.74-77

[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.*
Modification: l.99-101

[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.*
Modification: l.169-171

[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.*
Modification: Title

[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'*
Modification: l.255-157

[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].
Modification: l.251-254

[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.*

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.
Modification: l.140-144

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.'*
Modification: l.86-88

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.*
Modification: Figure 3

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.*
Modification: l.125-134

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'*
Modification: l.145-146

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.*

Modification: Figure 5

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.*

Modification: Figure 3; l.165-168; Figure 2

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.'*

Modification: l.183-185

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.'*
Modification: l.180-181

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.*
Modification: Title

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.'*
Modification: l.268-271

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.'*
Modification: l.288-289

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.'*
Modification: Figure 1 caption

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.*
Modification: Figure 2 caption

l. 112: circular statistics???
Reply: *We specify that: 'circular statistics are suitable for defining central tendencies of variables with a cycle* [*Burn*, 1997].
Modification: l.122

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.'*
Modification: Figure 3 caption

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

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., 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.

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.

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.

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.

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.

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*.

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.

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.

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.

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)

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.

Evaluating the suitability of hydrological models for flood impact assessments

Manuela I. Brunner[1], Lieke A. Melsen[2], Andrew W. Wood[1,3], Oldrich Rakovec[4,5], Naoki Mizukami[1], Wouter J. M. Knoben[6], and Martyn P. Clark[6]

[1]Research Applications Laboratory, National Center for Atmospheric Research, Boulder CO, USA
[2]Hydrology and Quantitative Water Management, Wageningen University, Wageningen, Netherlands
[3]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder CO, USA
[4]Department Computational Hydrosystems, Helmholtz Centre for Environmental Research, Leipzig, Germany
[5]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha – Suchdol, Czech Republic
[6]University of Saskatchewan Coldwater Laboratory, Canmore, Canada

Correspondence: Manuela I. Brunner (manuelab@ucar.edu)

Abstract. Floods cause large damages, especially if they affect large regions. Assessments of current, local and regional flood hazards and their future changes often involve the use of hydrologic models. However, uncertainties in simulated floods can be considerable and yield unreliable hazard and climate change impact assessments. A reliable hydrologic model ideally reproduces both local flood characteristics and spatial aspects of flooding, which is, however, not guaranteed especially when using standard model calibration metrics. In this paper we investigate how flood timing, magnitude and spatial variability are represented by an ensemble of hydrological models when calibrated on streamflow using the Kling–Gupta efficiency metric, an increasingly common metric of hydrologic model performance. We compare how four well-known models (SAC, HBV, VIC, and mHM) represent (1) flood characteristics and their spatial patterns; and (2) how they translate changes in meteorologic variables that trigger floods into changes in flood magnitudes. 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. They further show that changes in precipitation and temperature are not necessarily well translated to changes in flood flow, which makes local and regional flood hazard assessments even more difficult for future conditions. We conclude that models calibrated on integrated metrics such as the Kling–Gupta efficiency alone have limited reliability in flood hazard assessments, in particular in regional and future assessments, and suggest the development of alternative process-based and spatial evaluation metrics.

1 Introduction

Many studies use a hydrological model driven by present or future meteorological forcing data to derive flood estimates for current and future conditions. However, data, model structure, and parameter uncertainties can be considerable (Clark et al., 2016) especially when considering extreme events such as floods (Brunner et al., 2019b; Das and Umamahesh, 2018) and when considering hydrological change. It is therefore challenging to produce statistically reliable estimates of future changes in flood hazard.

A suitable model ideally reproduces different aspects of flooding, including local characteristics such as event magnitude and timing. It has been shown, however, that capturing magnitude and timing is challenging when standard calibration metrics are used individually for parameter estimation (Lane et al., 2019; Brunner and Sikorska, 2018; Mizukami et al., 2019). For example, one widely-used metric that is considered integrative compared to others (e.g., bias, correlation) is the Nash–Sutcliffe efficiency (E_{NS}; Nash and Sutcliffe 1970), where the sum-of-squares error metric focuses attention on high flows. However, E_{NS} is formulated so that its optimal value actually underestimates flow variability (Gupta et al., 2009). Using a related metric, the Kling–Gupta efficiency (E_{KG}; Gupta et al. 2009) can partially overcome this deficiency and improve simulations of peak flows (Mizukami et al., 2019). Yet to achieve further improvement, a broader range of application-specific evaluation metrics is typically required, including objectives that directly characterize hydrologic phenomena (or 'signatures') such as peak flows, flood volumes and timing, recession rates, and seasonal hydrograph shape. Considering multiple objectives in a step-wise calibration sequence, either manual or automated, is common in agencies that implement models for applications such as flood forecasting (Hogue et al., 2000), and strengthens their ability to provide reliable flood predictions. The use of multiple objectives, however, can lead to a decrease in performance with respect to any individual flow signature not considered as an objective (Mizukami et al., 2019). Despite their deficiencies with respect to extremes, individual 'integrative' standard calibration metrics such as E_{NS} or E_{KG} are often used in research modeling studies, even when the focus is on floods and their future changes (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. Harrigan et al. 2020; Hirpa et al. 2018; Huang et al. 2018; Thober et al. 2018; Brunner and Sikorska 2018).

In addition to simulating the timing and magnitude of flow at individual catchments, it is also important to realistically reproduce spatial dependencies, i.e. the relationship of flood occurrence across gauging stations (Keef et al., 2013; De Luca et al., 2017; Berghuijs et al., 2019). An over- or underestimation of spatial dependencies across a network of gauging stations in regional flood hazard and risk assessments has been shown to under- or overestimate regional damage, respectively (Lamb et al., 2010; Metin et al., 2020). Prudhomme et al. (2011) have shown for a set of large-scale hydrological models that simulated high flow episodes are less spatially coherent than observed events. Despite their high relevance for impact, the spatial aspects of flooding have often been overlooked in past simulation studies.

In this paper we explore the suitability of hydrological models for local and regional flood hazard assessments under current and future conditions if calibrated with the commonly used E_{KG}, which has been shown to result in more accurate flood peak representations than E_{NS} in a recent study by Mizukami et al. (2019). We evaluate the extent to which hydrological models calibrated against this common individual calibration metric reproduce (1) local flood characteristics (e.g. flood magnitude and timing at any given gauging station), (2) spatial dependencies in flooding, and (3) relationships between changes in flood triggering variables and changes in flood magnitude. We assess which aspects of hydrological models may need to be improved if we want to bring hazard and change impact assessments to a point where we can make more reliable assessments of regional flood hazard and future changes in local and spatial flood characteristics.

For this assessment, we look at the model output of four widely used hydrological models (Addor and Melsen, 2019), namely, the Sacramento Soil Moisture Accounting model (SAC-SMA; Burnash et al., 1973) combined with SNOW–17 (Anderson,

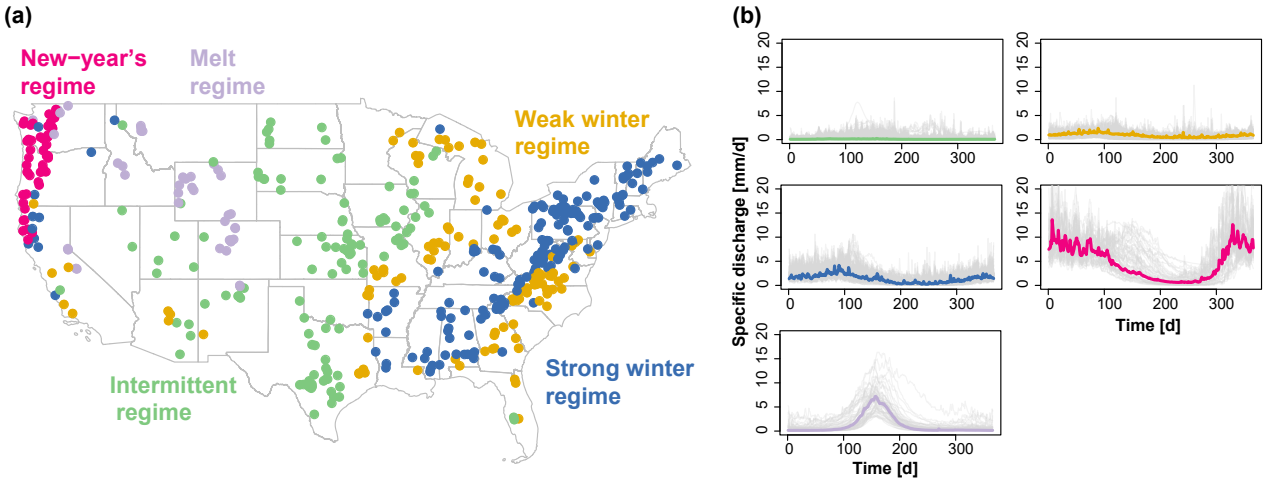

Figure 1. a) Map of the 488 catchments in the conterminous United States belonging to the five regime classes indicated by their gauge location: 1) Intermittent, 2) weak winter, 3) strong winter, 4) New Year's, and 5) melt. b) Median regime per regime class (colored lines) and variability of regimes within a class (on line per catchment, grey) (Brunner et al., 2020b).

1973), the Hydrologiska Byråns Vattenbalansavdelning model (HBV; Bergström, 1976), the Variable Infiltration Capacity model (VIC; Liang et al., 1994), and the mesoscale hydrologic model (mHM; Kumar et al., 2013; Samaniego et al., 2010). Identifying and documenting model weaknesses regarding regional and future flooding will highlight avenues for future model development and reveal potential deficiencies of a calibration strategy often applied for research studies on floods.

2 Data and Methods

To study how local and spatial flood characteristics are reproduced by hydrological models calibrated on streamflow using the individual calibration metric, E_{KG}, we compare observed to simulated flood event characteristics for a set of 488 catchments in the conterminous United States that have minimal human impact and catchment areas ranging from 4 to 2000 km^2 (Figure 1a) (Newman et al., 2015). The dataset comprises catchments with a wide range of climate and streamflow characteristics ranging from catchments with intermittent regimes and a very weak seasonality to catchments with a very strong seasonal cycle under the influence of snow (New Year's and melt regimes; Figure 1b; Brunner et al. 2020b). Observed streamflow time series are available from the U.S. Geological Survey (USGS, 2019).

2.1 Model simulations

We use daily streamflow simulations for the period 1981-2008 generated with four well-known hydrological models (Addor and Melsen, 2019) offering different model structures and complexity: the lumped SAC model (Figure A1; Burnash et al.,

1973), the lumped HBV model (Figure A2; Bergström, 1976), the lumped version of the VIC model (Figure A3; Liang et al., 1994), and the grid-based, distributed mesoscale hydrologic model mHM (Figure A4; Kumar et al., 2013; Samaniego et al., 2010). The model parameters were calibrated on streamflow observations by minimizing 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). E_{KG} is defined as:

$$E_{KG}(Q) = 1 - \sqrt{[s_\rho \cdot (\rho - 1)]^2 + [s_\alpha \cdot (\alpha - 1)]^2 + [s_\beta \cdot (\beta - 1)]^2}, \tag{1}$$

where ρ is the correlation between observed and simulated runoff, α is the standard deviation of the simulated runoff divided by the standard deviation of observed runoff, and β is the mean of the simulated runoff, divided by the mean of the observed runoff. s_ρ, s_α, and s_β are scaling parameters enabling a weighting of different components. When used individually, E_{KG} has been found to result in a better performance for annual peak flow simulation than the long-standing and related hydrologic model evaluation metric Nash–Sutcliffe efficiency (E_{NS}) (Mizukami et al., 2019).

For SAC, Melsen et al. (2018) calibrated and evaluated 18 out of the 35 parameters available in the coupled Snow-17 and SAC-SMA modeling system, for HBV 15 parameters, for VIC 17 parameters, and for mHM Rakovec et al. (2019) and Mizukami et al. (2019) calibrated and evaluated up to 48 parameters. 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; Thornton et al., 2012) and mHM with the forcing by Maurer et al. (2002) (12km resolution) both derived from observed precipitation and temperature. SAC, HBV, and VIC were calibrated and evaluated on the period 1985–2008 while mHM was calibrated on the period 1999–2008 and evaluated on the period 1989-1999. After calibration, all four models were run for the period 1980–2008 (calendar years), where the period 1980-1981 was here used for spin-up and therefore discarded from the analysis.

Model performance in terms of E_{KG} varies spatially and is related to the hydrological regime (Figure 2). It is overall lowest for catchments with intermittent regimes and a weak seasonality and highest for catchments with a strong seasonality such as a melt and New Year's regime. However, there is a high within-class variability in model performance. The finding that intermittent regimes are challenging to model successfully is well known in hydrology and reproduced in many studies, e.g., Unduche et al. (2018), who show that hydrological modeling on Prairie watersheds is very complex (Hay et al., 2018). Intermittent regimes may suffer in calibration if they rely solely on correlation-type measures because their day to day variation is more difficult to reproduce than a more pronounced and regular seasonality. 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). In addition to streamflow, we use areal precipitation and simulated soil moisture to explain potential differences in model performance.

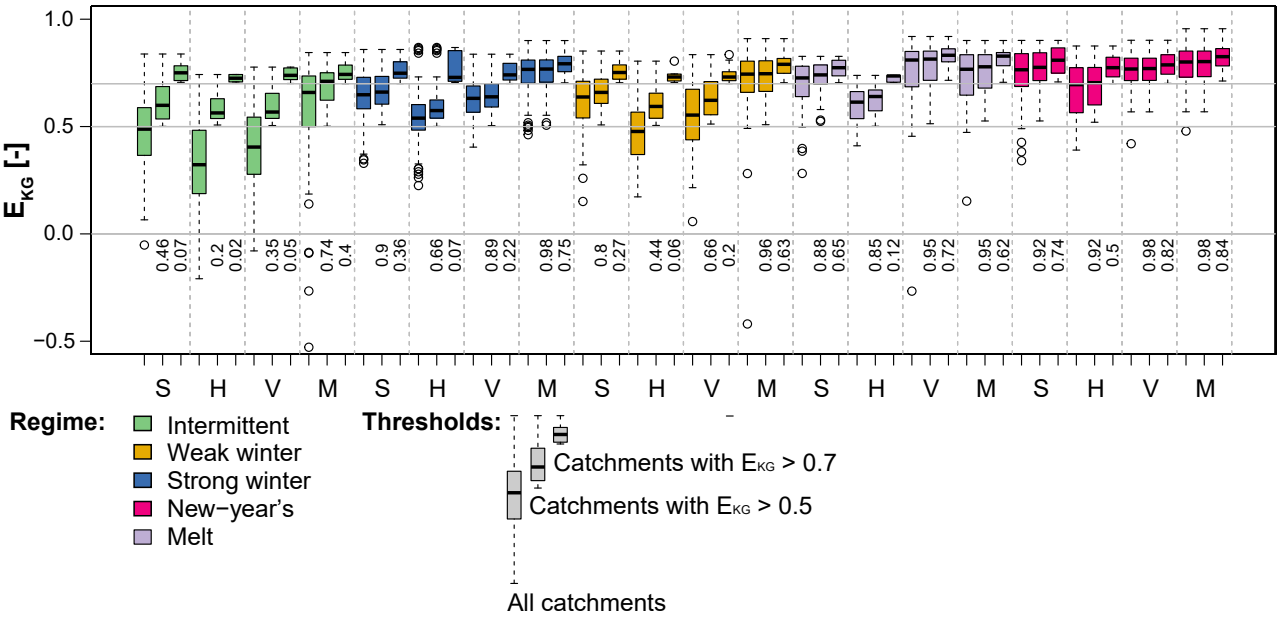

Figure 2. Model performance in terms of E_{KG} over the period 1981–2008 for the four models SAC (S), HBV (H), VIC (V), and mHM (M) per hydrological regime: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40). For each model and regime, three boxplots are shown: all catchments, catchments with $E_{KG} > 0.5$, and catchments with $E_{KG} > 0.7$. The percentage [-] of catchments of a regime class above the corresponding threshold is indicated below the 0 line.

2.2 Model evaluation for floods

We compare local and spatial flood characteristics extracted from the observed time series to those of the series simulated with the four models for the period 1981–2008.

2.2.1 Flood event identification

Flood events are identified for each of the five time series (one observed, four simulated) using a peak-over-threshold (POT) approach similar to the one used in Brunner et al. (2019a, 2020b). This approach consists of two main steps and results in two data sets each, which are used for the local and spatial analysis, respectively: (1) POT events in individual catchments and (2) event occurrences across all catchments. In Step 1, independent POT events are identified in the daily discharge time series of the individual catchments using the 25th percentile of the corresponding time series of annual maxima as a threshold (Schlef et al., 2019) and by prescribing a minimum time lag of 10 days between events (Diederen et al., 2019). This procedure results in a first quartile of 36, a median of 40, and a third quartile of 47 events identified per basin. In Step 2, a data set consisting of the dates of flood occurrences across all catchments is compiled. This set is converted into a binary matrix which specifies for each catchment (columns) whether or not it is affected by a specific event (rows). We consider a catchment to be affected by a certain event if it experiences an event within a window of $\pm$ 2 days of that event to take into account travel times. In addition

to a binary matrix over all events, we set up seasonal binary matrices (winter: Dec–Feb, spring: Mar-May, summer: June–Aug, fall: Sept–Nov).

2.2.2 Flood characteristics at individual sites

We use the data sets resulting from Step 1, the POT events at individual catchments, to evaluate how well the models reproduce flood statistics at individual sites. We focus on the total number of events n (actual error: $n_s - n_o$, where s represents simulations and o observations), magnitude in terms of mean peak discharge x (relative error: $(x_s - x_o)/x_o$), and mean timing (absolute error: circular statistics suitable for defining central tendencies of variables with a cycle (Burn, 1997)).

2.2.3 Spatial flood dependence

We then use the data sets resulting from Step 2 to evaluate how models reproduce overall and seasonal spatial flood dependence. To do so, we use the connectedness measure introduced by Brunner et al. (2020a), 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. (2020a), 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). We computed actual errors in flood connectedness by subtracting observed from simulated connectedness over all seasons and per season.

2.2.4 Flood triggers

To explain potential differences in model performance, we look at the relationship of simulated peak discharge with the two flood triggers: precipitation and soil moisture on the day of flood occurrence. We focus on the day of occurrence because time of concentration is typically small for small headwater basins (USDA-NRCS, 2010).

2.2.5 Floods under change

In addition to assessing model performance under current climate conditions, we would like to understand potential, additional challenges arising when interested in future conditions. 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). To perform this sensitivity analysis, we generate surrogate time series of temperature, precipitation, and streamflow for each catchment (Wood et al., 2004; Brunner et al., 2020b). 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. For each of the surrogate series, we again extract POT flood events using the same procedure as described under Step 1. For each of the extracted events we then determine temperature and precipitation. We use the sets of peak discharge, event temperature and event precipitation to compute mean event discharge, temperature, and precipitation, which enables the derivation of a relationship between mean POT discharge and the two meteorological variables during events. We repeat the resampling $n = 500$ times to derive a relationship between changes in mean event temperature and precipitation and changes in mean POT streamflow. This resampling experiment results in a response surface of POT discharge spanned by mean event temperature and mean event precipitation for each catchment. We summarize the results obtained at individual locations by computing horizontal and vertical sensitivity gradients on these reaction surfaces using a linear regression model. The horizontal gradient describes the strength of POT discharge changes in response to event temperature changes while the vertical gradient describes the strength of change in response to changes in event precipitation. Conducting this experiment for both observed and simulated time series allows for the determination of whether the models react to changes in mean event temperature and precipitation in the same way as the real world system and are therefore suitable for the use in climate change impact assessments on floods. If models produce different climate sensitivities than the ones seen in the observations, the use of models to simulate sets of flood events for future conditions may preclude reliable change assessments.

3 Results

3.1 Flood characteristics at individual sites

Model performance at individual sites with respect to the number of events, event magnitude, and timing varies by model and hydrological regime type (Figure 3). 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 (Brunner et al., 2020b). The overestimation of the number of events by HBV may be explained by its fast response to precipitation as expressed through its model parameter β, which introduces non-linearity to the system (Viglione and Parajka, 2020). Event magnitude in terms of peak discharge is generally underestimated for all regime types independent of the model. Underestimation is in line with previous studies showing that using E_{KG} individually results in an underestimation of peak flow (Mizukami et al., 2019) due to an underestimation of variability, which will result in an under-representation of extremes (Katz and Brown, 1992). Another factor potentially contributing to this underestimation is that the models were forced with spatially lumped instead of distributed data, which may smooth the simulated discharge response.

Absolute flood timing errors are present in all models. They are the highest in catchments with intermittent regimes with a high variability in flood timing and low in catchments with a New Year's and melt regime where the flood season is limited to

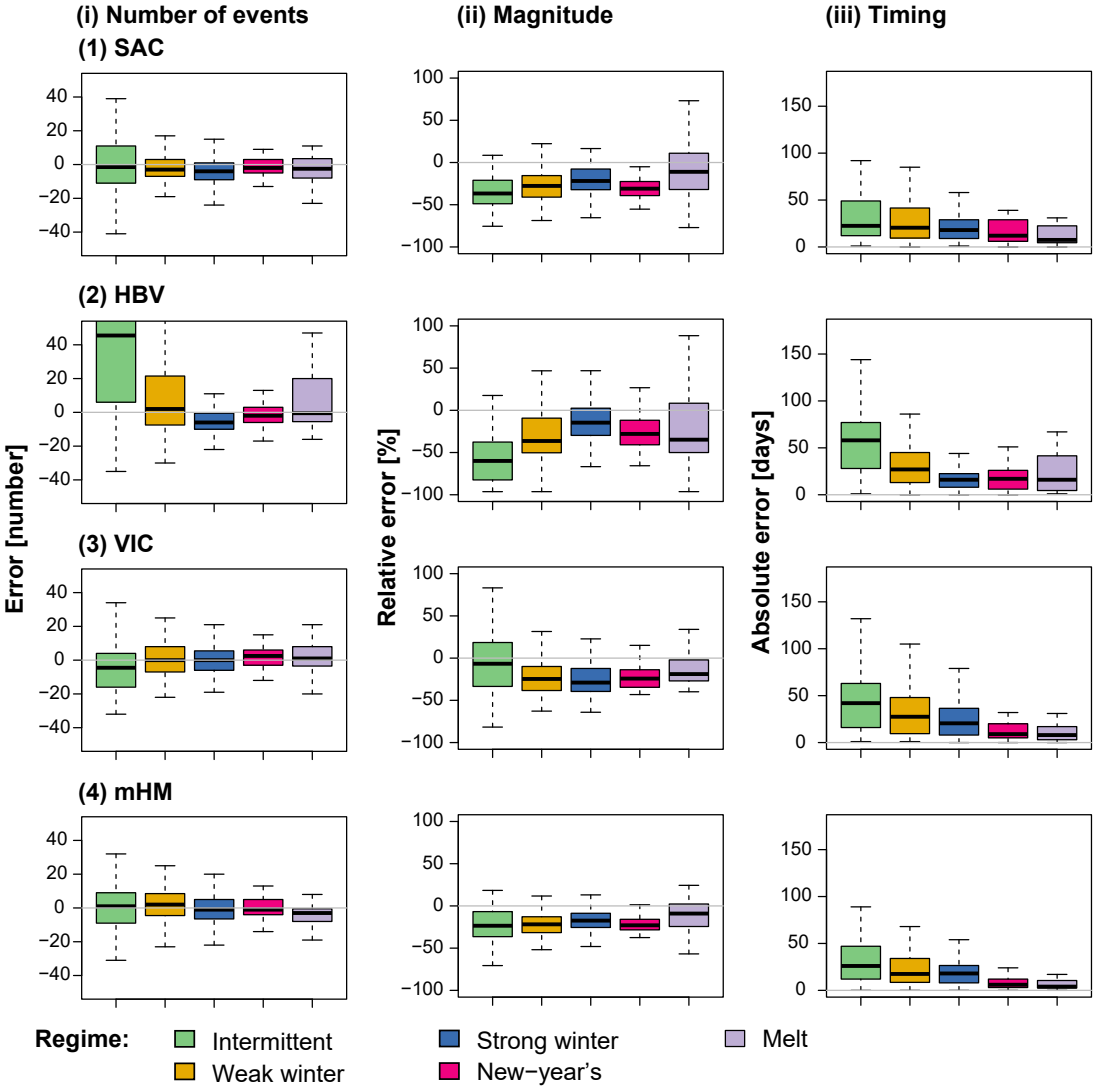

Figure 3. Model errors per regime type computed over the period 1981–2008: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40) (Figure 1). Errors are shown for (i) number of events (error in number of events), (ii) magnitude (mean relative error), and (iii) timing (mean absolute error in days) for the four models (1) SAC, (2) HBV, (3) VIC, and (4) mHM. The boxplots are composed of one value per catchment belonging to the respective regime class.

a few months (Brunner et al., 2020a). Over all, there is no clear tendency of one model to perform better than the other ones. 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.

3.2 Spatial flood dependencies

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 (Figure 4). Seasonally, most models over- or underestimate spatial dependence in certain regions. In winter, connectedness is overestimated by most models except for VIC and the strength of overestimation is strongest for HBV. In spring, most models tend to underestimate spatial dependence except for HBV that results in an overestimation of spatial dependence for catchments with an intermittent regime. The overestimation of spatial dependence in winter for all regimes except the melt regime is likely related to higher simulated than observed snowmelt as high soil moisture and snow availability have been shown to increase spatial flood connectedness (Brunner et al., 2020a). Related to this, the underestimation of spatial connectedness in spring may be related to the subsequent missing snowmelt contributions. Spatial connectedness in summer has been shown to be generally weak due to the occurrence of localized, convective events (Brunner et al., 2020a), which is reflected by most models except for HBV in the case of intermittent and melt regimes. Spatial flood connectedness has also been shown to be weak in fall (Brunner et al., 2020a) but is overestimated by most models. Connectedness overestimation by HBV is most pronounced for catchments with an intermittent regime. Otherwise, connectedness over-/underestimation seems to be independent of the regime. The finding that there is room for improvement regarding the representation of spatial flood dependencies is in line with previous studies showing that large-scale hydrological models have a weakness in reproducing regional aspects of floods (Prudhomme et al., 2011).

3.3 Flood triggers

The differences in model performance regarding local and spatial flood characteristics may be partially explained by differences in their structure and how they transform precipitation into runoff. Figure 5 shows how simulated peak discharge is related to event precipitation, event precipitation plus snowmelt, and simulated soil moisture over all catchments for the four hydrologic models. The SAC and VIC models show similar simulated relationships for all three variable pairs. There is a positive relationship between peak discharge and precipitation and peak discharge and rainfall plus snowmelt, i.e. the higher the precipitation input or rainfall and snowmelt combined, respectively, the higher the resulting peak discharge. This relationship is slightly more expressed for VIC than for SAC. In both models, soil moisture and event magnitude are also positively related with lower peak values potentially associated with lower soil moisture states than more severe events. The peak discharge–precipitation relationship of HBV and mHM is less straightforward than the one of SAC and VIC. HBV and mHM also show high discharge when precipitation input is high, but may in some cases still produce high discharge values even for low precipitation inputs. Such low precipitation inputs can also lead to high peak discharge for SAC but to a lesser degree than HBV and mHM. However, peak discharge and rainfall plus snowmelt show a strong linear relationship, i.e. the higher the combined rainfall and

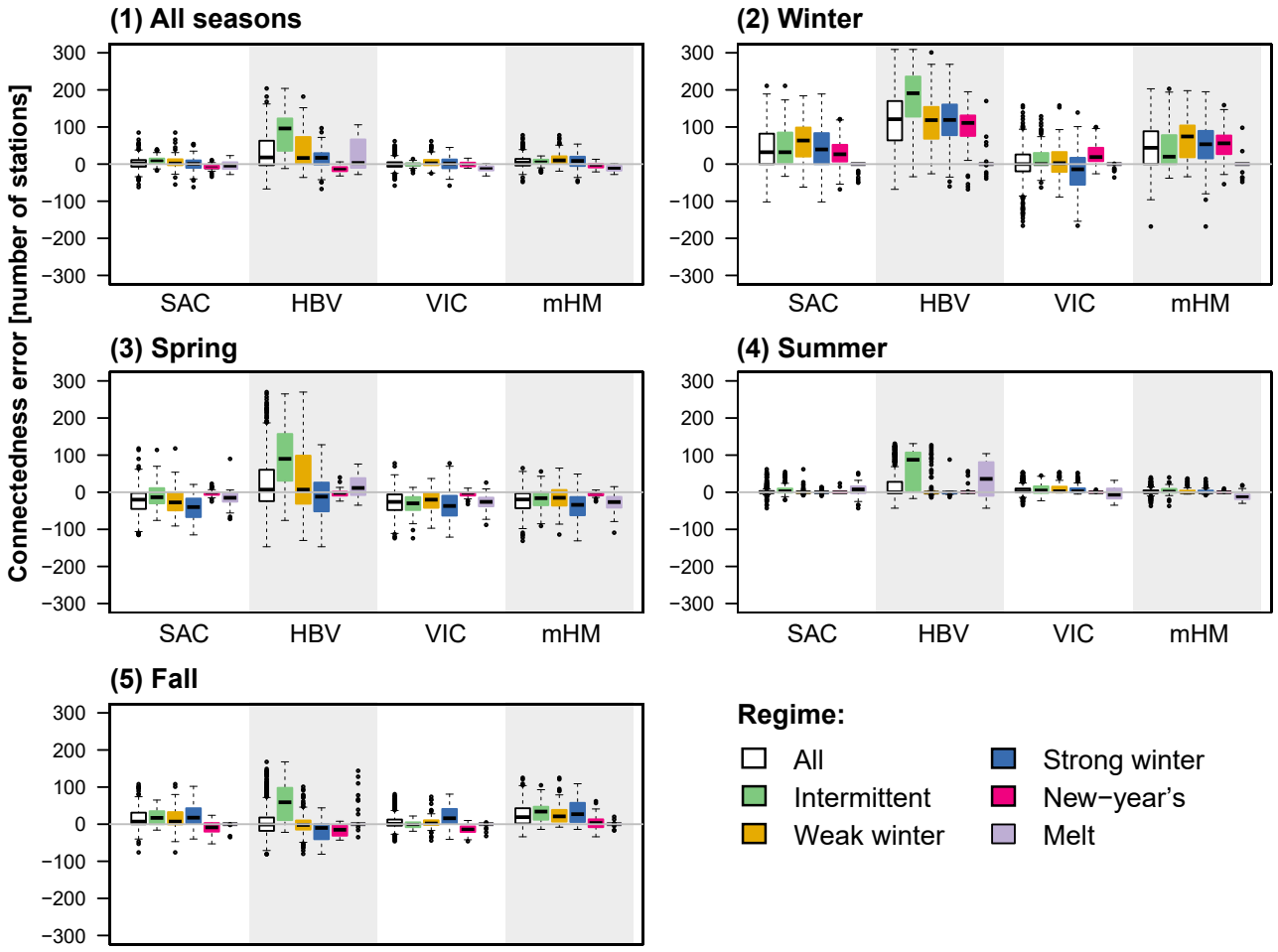

Figure 4. Overall (1) and seasonal (2–5) errors in flood connectedness (simulated minus observed connectedness), i.e. number of catchments a catchment is sharing at least 1% of the total number of flood events with, for the four models SAC, HBV, VIC, and mHM over all regimes and per regime: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40).

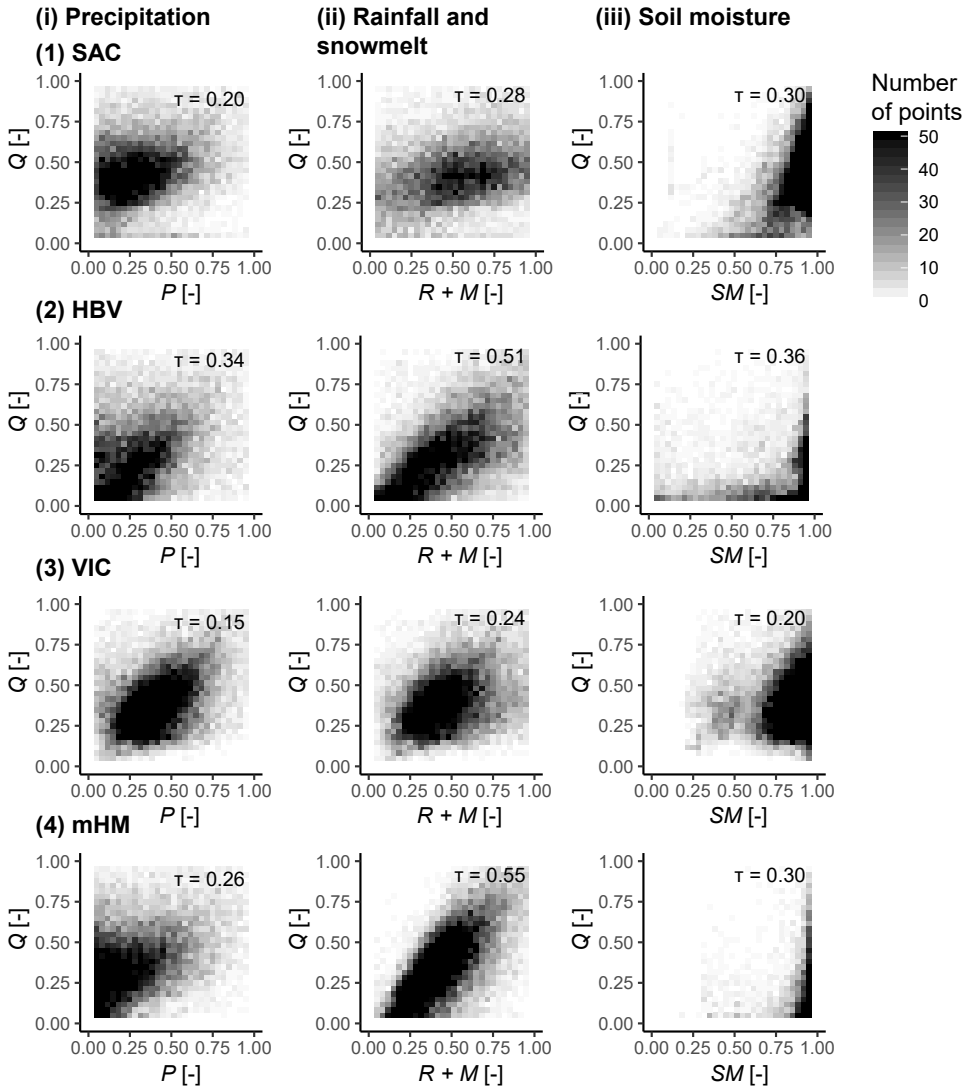

Figure 5. Simulated relationships between normalized flood discharge (Q) and normalized precipitation (i, P), rainfall and snowmelt (ii, $R+M$), and soil moisture (iii, SM, upper two soil layers for mHM) over all catchments represented by a binned scatter plot for the four hydrologic models (1) SAC, (2) HBV, (3) VIC, and (4) mHM. The darker the color, the higher the number of points within a bin (one point per catchment and event). Kendall's correlation coefficients are provided in the upper right corners of the subplots.

snowmelt input to the system, the higher is peak discharge. High flows are in most cases related to nearly full storage states but can occasionally also be triggered when soil moisture is low for SAC and VIC and to a lesser degree for HBV. These two types of responses may be related to differences in model behavior. VIC and SAC show more linearity in their event precipitation and peak discharge relationship than HBV and mHM, possibly because VIC and SAC have the capability to generate surface runoff when precipitation intensity exceeds infiltration capacity (Burnash et al., 1973; Liang et al., 1994). In this case, incoming precipitation is directly translated into flood discharge. In contrast, HBV and mHM, which is based on the HBV model structure (Kumar et al., 2013), do not include a surface runoff component and all discharge originates in the model stores (Bergström, 1976). This introduces a non-linearity in model response and may explain why a smaller precipitation input may still generate high peak flows in these models.

We here show that model performance to some degree depends on model choice and that many different combinations of forcing and model states can simulate floods. 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 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). The importance of input uncertainty is particularly pronounced if we are interested in future changes because of climate model and scenario uncertainty (Chen et al., 2014; Lopez-Cantu et al., 2020).

3.4 Floods under change

In addition to looking at how well local and spatial flood characteristics are represented by models, we look at how changes in temperature and event precipitation are translated into changes in flood flows to assess each models' suitability for climate impact assessments on floods. Our sensitivity analysis shows that the models have difficulty translating changes in event temperature and precipitation into sensitivities of flood flows (Figure 6), which can be problematic if we would like to use such models in climate change assessments. Generally, flood flows show a relatively low sensitivity to changes in mean event precipitation and temperature. This is in contrast to the behavior for mean flow, which is strongly influenced by changes in mean precipitation as demonstrated in a similar experiment by Brunner et al. (2020b). The much stronger relationship between mean precipitation and flow than between event precipitation and flow might arise because mean flow is a climate signal (Knoben et al., 2018), whereas floods are more an event (higher frequency, short-term) signal. However, some catchments, e.g. the Tucca Creek (New Year's regime) show a clear relationship between peak magnitude and both event temperature and precipitation. 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). In the case of melt regimes, the misrepresentation of flood sensitivities by models suggests that they may have difficulty simulating snow-influenced flooding.

This relatively poor model performance in capturing observed flood sensitivities can be generalized to the larger set of catchments studied here (Figure 7). Temperature sensitivities are found to be positive or negative, i.e. an increase in temperature

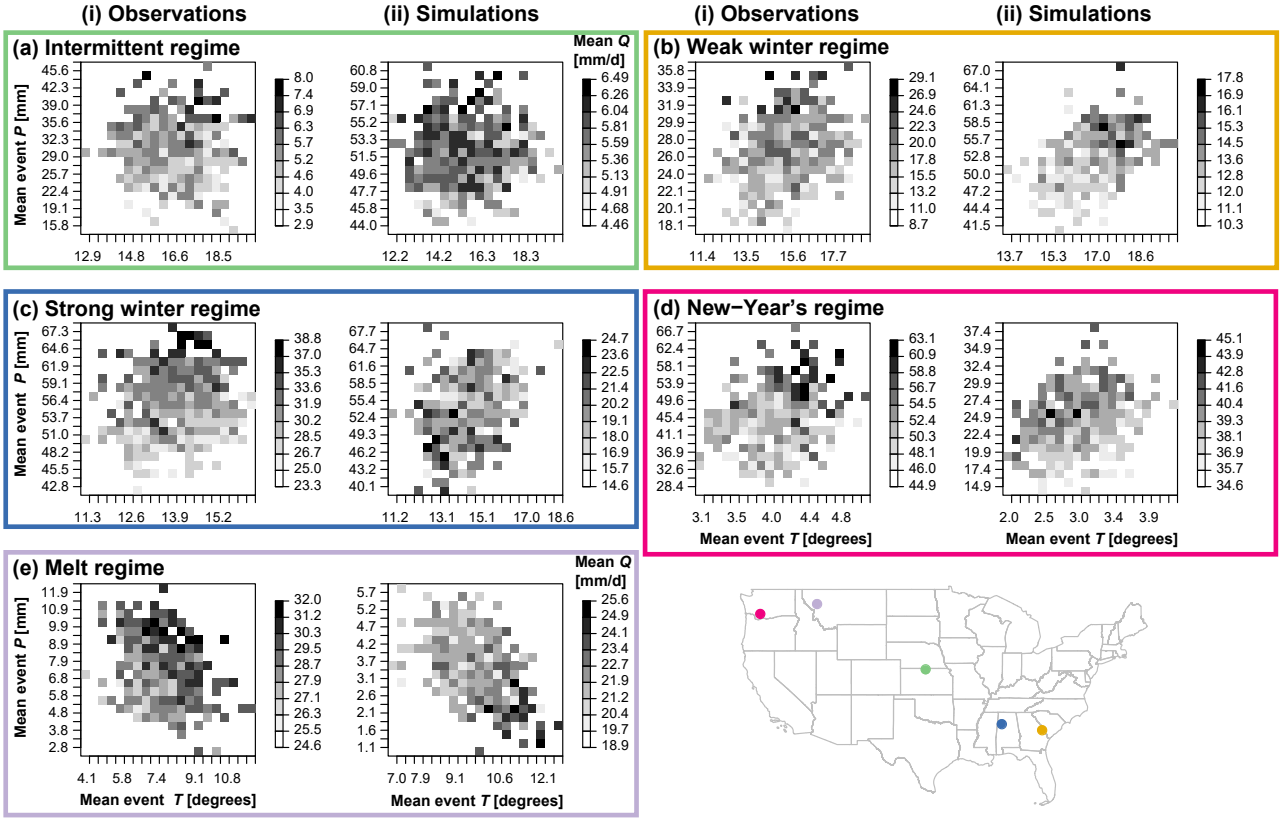

Figure 6. Climate sensitivity analysis for the VIC model: Dependence of mean POT magnitude (Q) on mean flood event precipitation (1-day; P) and mean flood temperature (T) for five example catchments, those with the best E_{KG} per regime type: intermittent regime (green; USGS ID 09210500 Fontanelle Creek near Fontanelle, WY; $E_{\mathrm{KG}} = 0.78$), weak winter regime (yellow; USGS ID 02369800 Blackwater River near Bradley, AL; $E_{\mathrm{KG}} = 0.83$), strong winter regime (blue; USGS ID 11522500 Salmon River above Somes, CA; $E_{\mathrm{KG}} = 0.84$), New Year's regime (pink; USGS ID 14303200 Tucca Creek near Blaine, OR; $E_{\mathrm{KG}} = 0.9$), and melt regime (purple; USGS ID 13011500 Pacific Creek at Moran, WY; $E_{\mathrm{KG}} = 0.92$). Grid axes and grey scales differ between plots where darker colors indicate higher flood magnitudes.

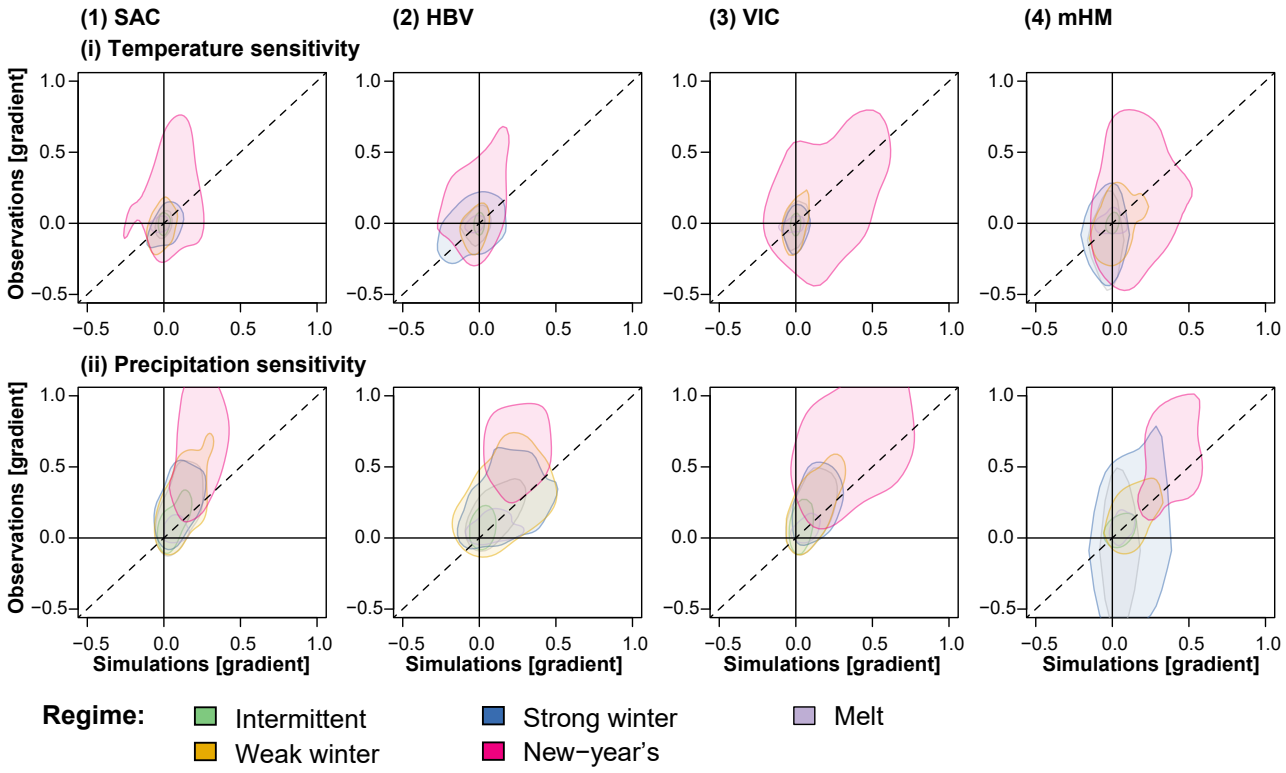

Figure 7. Observed vs. simulated (i) horizontal (temperature) and (ii) vertical (precipitation) climate sensitivities for floods represented by two-dimensional kernel density estimates for the four models (1) SAC, (2) HBV, (3) VIC, and (4) mHM for the five regime types: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40) (Figure 1). 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.

could lead to an increase or decrease of peak flow depending on the catchment. In general, these temperature sensitivities are relatively weak (i.e. gradients are close to zero), which may be the reason why they are difficult to capture. In contrast, precipitation sensitivities are mostly positive, i.e. an increase in event precipitation leads to an increase in peak flow. However, the strength of these sensitivities is underestimated by all models, i.e. a change in precipitation leads to a too small change in peak flow. This underestimation of sensitivity can be understood by the underestimation of flood magnitude in general.

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). 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 (Mizukami et al., 2019) and peak flood

magnitudes and a misrepresentation of timing. As is evident in some existing practice-oriented applications of hydrological models (Hogue et al., 2000; Unduche et al., 2018; World Meteorological Organization, 2011), the simulation of floods and other hydrologic phenomena is likely to be improved by using more tailored model calibration strategies. The former would include either giving more weight to the variability component of an integrative metric such as the E_{KG} (Pool et al., 2017; Mizukami et al., 2019); whereas the latter might include 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, within a multi-objective model calibration process (Moussa and Chahinian, 2009; Sikorska et al., 2018; Sikorska-Senoner et al., 2020). The spatial representation of extremes may also be improved by considering spatially distributed features of model response within a spatial calibration framework (Dembélé et al., 2020; Koch et al., 2018).

4 Conclusions

Our model comparison shows that all flood characteristics are not equally well represented by models calibrated with the widely used Kling–Gupta efficiency metric. 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. These model deficiencies in reproducing local flood characteristics, especially timing, can lead to a misrepresentation of spatial flood dependencies, particularly in winter, because the temporal and spatial dimension of flooding are closely linked. We therefore conclude that the representation of magnitude, timing and spatial connectedness can be improved. The limited capability of the models in reproducing local and spatial flood characteristics is partly attributed to a reliance of the calibration on an individual variable (streamflow) and calibration metric (E_{KG}). While E_{KG} is integrative of certain properties (bias, variance, correlation), it does nonetheless not explicitly focus on high flow values, their spatial dependencies, or processes generating high flow values. Such focus could be improved by giving more weight to the variability component of E_{KG}, if a single metric is used, or by including indicators of extremes in in a multi-objective framework when calibrating and validating the model. The spatial concern could be addressed by applying spatial calibration procedures. Such steps are recommended if we would like to improve the reliability of local and regional flood hazard assessments.

Our sensitivity analysis also shows that climate sensitivities of floods, especially to changes in precipitation, are not well represented in models even if the model can be deemed 'well-calibrated' via the individual E_{KG} metric. These sensitivities are generally underestimated by models independent of the geographical areas considered, i.e. an increase in event precipitation may not be translated into a strong enough increase in flood peak. The mis-estimation of these sensitivities may undermine the reliability of future flood hazard assessments relying on such models.

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. Despite its shortcomings, this practice has become increasingly more common and accepted in

the research literature. Yet, our analysis illustrates that the development and adoption of more comprehensive multi-objective and multi-variable calibration strategies are needed to significantly improve model performance regarding floods under both current and future climate conditions.

Data availability. Observed streamflow measurements were made accessible by the USGS and can be downloaded via the website https://waterdata.usgs.gov/nwis. Simulated streamflow, precipitation, and storage time series can be requested from Lieke Melsen (lieke.melsen@wur.nl) for the SAC, HBV, and VIC models and for the mHM model from Oldrich Rakovec (oldrich.rakovec@ufz.de).

Appendix A: Model illustrations

This section provides illustrations of the model structures used in this work. Model schematics summarize the model states and fluxes. Schematics and equations use model-specific names as they are used in the model code. For clarity, these descriptions do enforce that fluxes are shown in lower case and states in upper case. The model diagrams are based on:

- Snow17/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 and source code.
- 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) and on analysis of the v4.1.2h source code (https://github.com/UW-Hydro/VIC/releases/tag/VIC.4.1.2.h).
- 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).

A1 Snow17/SAC-SMA

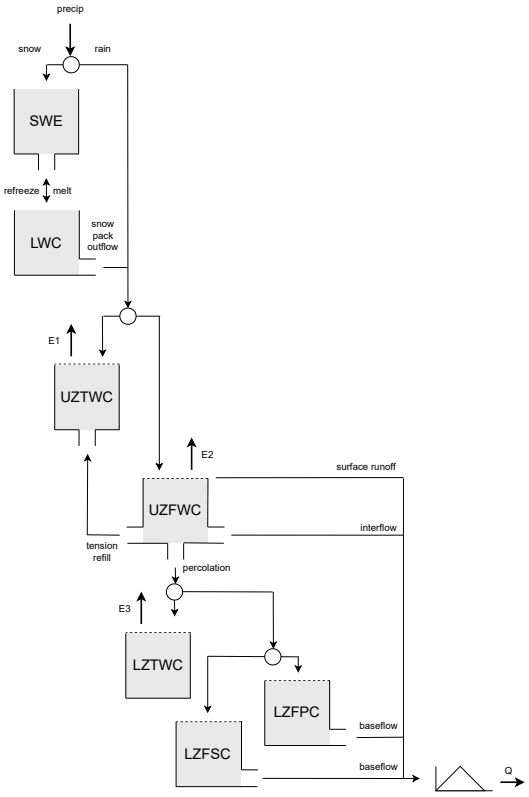

Figure A1. Structure of the Snow17/SAC-SMA model. Fluxes: precipitation (precip), snow, rain, snowmelt (melt), refreeze, snowpack outflow, evapotranspiration (E1, E2, and E3), tension refill, surface runoff, interflow percolation, baseflow, simulated discharge (Q). States: snow-water-equivalent (SWE), liquid water content (LWC), (UZTWC), upper zone free water contents (UZFWC), lower zone tension water contents (LZTWC), lower zone free primary contents (LZFPC), lower zone free supplemental contents (LZFSC).

A2 TUW-HBV

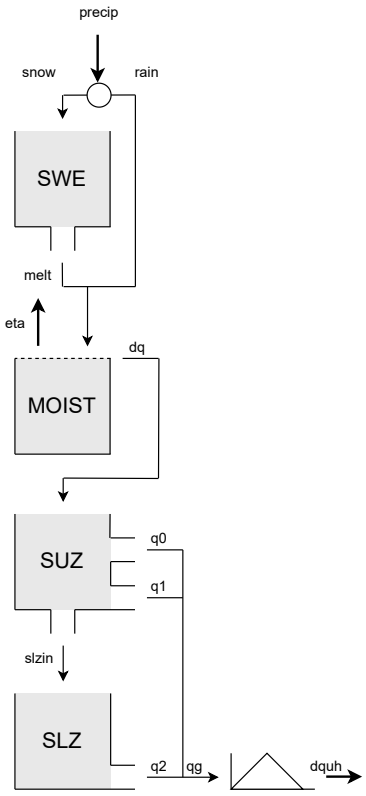

Figure A2. Structure of the TUW HBV model. Fluxes: precipitation (precip), snow, rain, snowmelt (melt), actual evapotranspiration (eta), runoff (dq), surface runoff (q0), subsurface runoff (q1), baseflow (q2), simulated runoff (qg), simulated discharge (dquh), input from upper to lower storage (slzin). States: snow-water-equivalent (SWE), soil moisture (MOIST), upper storage zone (SUZ), lower storage zone (SLZ).

A3 VIC

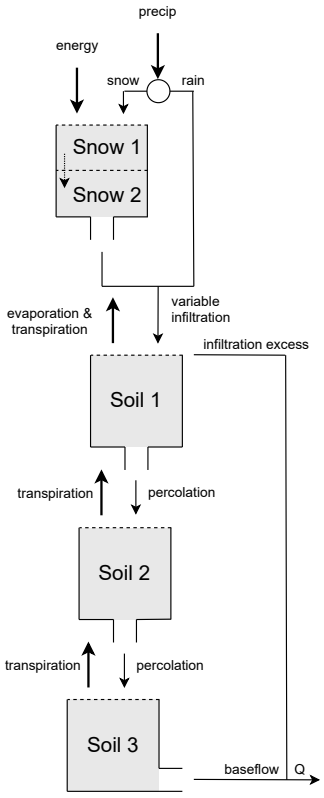

Figure A3. Fluxes: precipitation (precip), energy, snow, rain, variable infiltration, evaporation and transpiration, infiltration excess, baseflow, percolation, transpiration, simulated runoff (Q). Storage: snow layer 1 (Snow 1), snow layer 2 (Snow 2), soil layer 1 (Soil 1), soil layer 2 (Soil 2), soil layer 3 (Soil 3).

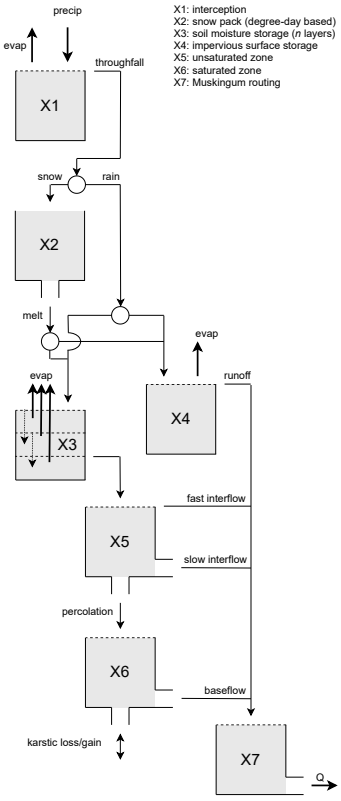

Figure A4. Fluxes: precipitation (precip), evapotranspiration (evap), throughfall, snow, rain, snowmelt (melt), runoff, fast interflow, slow interflow, percolation, baseflow, karstic loss/gain, simulated discharge (Q). Storage: Interception storage (X1), snow pack (X2), soil moisture storage (X3), impervious surface storage (X4), unsaturated zone (X5), saturated zone (X6), routing (X7).

Author contributions. MIB and MPC developed the study design. NM, OR, and LAM provided the model simulations and together with MIB, MPC and WK interpreted the model output. AW assisted with the paper's background and messaging and proposed the climate sensitivity strategy. WK produced the model illustrations. MIB wrote the first draft of the manuscript and all co-authors revised and edited the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Swiss National Science Foundation via a PostDoc.Mobility grant (P400P2_183844, granted to MIB). We acknowledge co-author support by the Bureau of Reclamation (CA R16AC00039), the US Army Corps of Engineers (CSA 1254557), and the NASA Advanced Information Systems Technology program (award ID 80NSSC17K0541). We also acknowledge support from the Global Water Futures research programme. We thank the three reviewers for their constructive feedback, which helped to reframe and clarify the storyline.

References

Addor, N. and Melsen, L. A.: Legacy, rather than adequacy, drives the selection of hydrological models, Water Resources Research, 55, 378–390, https://doi.org/10.1029/2018WR022958, 2019.

Anderson, E. A.: NOAA technical memorandum NWS-HYDRO-17: National Weather Service river forecast system-snow accumulation and ablation model, Tech. rep., U.S. Depertment of Commerce. National Oceanic and Atmospheric Administration. National Weather Service, Washington, DC, 1973.

Berghuijs, W. R., Allen, S. T., Harrigan, S., and Kirchner, J. W.: Growing spatial scales of synchronous river flooding in Europe, Geophysical Research Letters, 46, 1423–1428, https://doi.org/10.1029/2018GL081883, 2019.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments. Swedish Meteorological and Hydrological Institute (SMHI) RHO 7, Tech. Rep. January 1976, Sveriges Meteorologiska och Hydrologiska Institut, Norrköping, 1976.

Bratley, P. and Fox, B. L.: Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator, ACM Transactions on Mathematical Software (TOMS), 14, 88–100, https://doi.org/10.1145/42288.214372, 1988.

Brunner, M. I. and Sikorska, A. E.: Dependence of flood peaks and volumes in modeled runoff time series: effect of data disaggregation and distribution, Journal of Hydrology, 572, 620–629, https://doi.org/10.1016/j.jhydrol.2019.03.024, 2018.

Brunner, M. I., Furrer, R., and Favre, A.-C.: Modeling the spatial dependence of floods using the Fisher copula, Hydrology and Earth System Sciences, 23, 107–124, https://doi.org/10.5194/hess-23-107-2019, 2019a.

Brunner, M. I., Hingray, B., Zappa, M., and Favre, A. C.: Future trends in the interdependence between flood peaks and volumes: Hydroclimatological drivers and uncertainty, Water Resources Research, 55, 1–15, https://doi.org/10.1029/2019WR024701, 2019b.

Brunner, M. I., Gilleland, E., Wood, A., Swain, D. L., and Clark, M.: Spatial dependence of floods shaped by spatiotemporal variations in meteorological and land-surface processes, Geophysical Research Letters, 47, e2020GL088 000, https://doi.org/10.1029/2020GL088000, 2020a.

Brunner, M. I., Newman, A., Melsen, L. A., and Wood, A.: Future streamflow regime changes in the United States: assessment using functional classification, Hydrology and Earth System Sciences Discussions, p. in press, https://doi.org/10.5194/hess-2020-54, 2020b.

Burn, D. H.: Catchment similarity for regional flood frequency analysis using seasonality measures, Journal of Hydrology, 202, 212–230, 1997.

Burnash, R. J., Ferral, R. L., and McGuire, R. A.: A generalized streamflow simulation system. Conceptual modeling for digital computers, Tech. rep., Joint Federal-State River Forecast Center, Sacramento, 1973.

Chen, H., Sun, J., and Chen, X.: Projection and uncertainty analysis of global precipitation-related extremes using CMIP5 models, International Journal of Climatology, 34, 2730–2748, https://doi.org/10.1002/joc.3871, 2014.

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., Fowler, H. J., Prudhomme, C., Arnold, J. R., and Brekke, L. D.: Characterizing uncertainty of the hydrologic impacts of climate change, Current Climate Change Reports, 2, 55–64, https://doi.org/10.1007/s40641-016-0034-x, 2016.

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

Das, J. and Umamahesh, N. V.: Assessment of uncertainty in estimating future flood return levels under climate change, Natural Hazards, 93, 109–124, https://doi.org/10.1007/s11069-018-3291-2, 2018.

De Luca, P., Hillier, J. K., Wilby, R. L., Quinn, N. W., and Harrigan, S.: Extreme multi-basin flooding linked with extra-tropical cyclones, Environmental Research Letters, 12, 1–12, https://doi.org/10.1088/1748-9326/aa868e, 2017.

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., and Mariéthoz, G.: Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite datasets, Water Resources Research, p. e2019WR026085, https://doi.org/10.1029/2019WR026085, 2020.

Diederen, D., Liu, Y., Gouldby, B., Diermanse, F., and Vorogushyn, S.: Stochastic generation of spatially coherent river discharge peaks for continental event-based flood risk assessment, Natural Hazards and Earth System Sciences, 19, 1041–1053, https://doi.org/10.5194/nhess-19-1041-2019, 2019.

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

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA4 operational global river discharge reanalysis 1979-present, Earth System Science Data, pp. 1–23, 2020.

Hay, L., Norton, P., Viger, R., Markstrom, S., Steven Regan, R., and Vanderhoof, M.: Modelling surface-water depression storage in a Prairie Pothole Region, Hydrological Processes, 32, 462–479, https://doi.org/10.1002/hyp.11416, 2018.

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

Hogue, T. S., Sorooshian, S., Gupta, H., Holz, A., and Braatz, D.: A multistep automatic calibration scheme for river forecasting models, Journal of Hydrometeorology, 1, 524–542, 2000.

Huang, S., Kumar, R., Rakovec, O., Aich, V., Wang, X., Samaniego, L., Liersch, S., and Krysanova, V.: 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, Environmental Research Letters, 13, 124 005, https://doi.org/10.1088/1748-9326/aae94b, 2018.

Hundecha, Y. and Merz, B.: Exploring the relationship between changes in climate and floods using a model-based analysis, Water Resources Research, 48, https://doi.org/10.1029/2011WR010527, 2012.

Katz, R. W. and Brown, B. G.: Extreme events in a changing climate: variability is more important than averages, Climatic Change, 21, 289–302, 1992.

Keef, C., Tawn, J. A., and Lamb, R.: Estimating the probability of widespread flood events, Environmetrics, 24, 13–21, https://doi.org/10.1002/env.2190, 2013.

Khatami, S., Peel, M. C., Peterson, T. J., and Western, A. W.: Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty, Water Resources Research, 55, 8922–8941, https://doi.org/10.1029/2018WR023750, 2019.

Knoben, W. J., Woods, R. A., and Freer, J. E.: A quantitative hydrological climate classification evaluated with independent streamflow data, Water Resources Research, 54, 5088–5109, https://doi.org/10.1029/2018WR022913, 2018.

Koch, J., Demirel, M. C., and Stisen, S.: The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models, Geoscientific Model Development, 11, 1873–1886, https://doi.org/10.5194/gmd-11-1873-2018, 2018.

Köplin, N., Schädler, B., Viviroli, D., and Weingartner, R.: Seasonality and magnitude of floods in Switzerland under future climate change, Hydrological Processes, 28, 2567–2578, https://doi.org/10.1002/hyp.9757, 2014.

Kumar, R., Samaniego, L., and Attinger, S.: The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics, Journal of Hydrology, 392, 54–69, https://doi.org/10.1016/j.jhydrol.2010.07.047, 2010.

Kumar, R., Samaniego, L., and Attinger, S.: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resources Research, 49, 360–379, https://doi.org/10.1029/2012WR012195, 2013.

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., Dunning, P., and Batstone, C.: A new method to assess the risk of local and widespread flooding on rivers and coasts, Journal of Flood Risk Management, 3, 323–336, https://doi.org/10.1111/j.1753-318X.2010.01081.x, 2010.

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A., and Reaney, S. M.: Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain, Hydrology and Earth System Sciences, 23, 4011–4032, https://doi.org/10.5194/hess-23-4011-2019, 2019.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research, 99, 14 415, https://doi.org/10.1029/94JD00483, 1994.

Lopez-Cantu, T., Prein, A. F., and Samaras, C.: Uncertainties in future U.S. extreme precipitation from downscaled climate projections, Geophysical Research Letters, 47, 1–11, https://doi.org/10.1029/2019GL086797, https://doi.org/10.1029/2019GL086797, 2020.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, Journal of Climate, 15, 3237–3251, 2002.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrological Processes, 24, 1270–1284, https://doi.org/10.1002/hyp.7587, 2010.

Mcmillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, Hydrological Processes, 26, 4078–4111, https://doi.org/10.1002/hyp.9384, 2012.

Melsen, L. and Guse, B.: Hydrological drought simulations: How climate and model structure control parameter sensitivity, Water Resources Research, pp. 1–21, https://doi.org/10.1029/2019wr025230, 2019.

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

Metin, A. D., Dung, N. V., Schröter, K., Vorogushyn, S., Guse, B., Kreibich, H., and Merz, B.: The role of spatial dependence for large-scale flood risk estimation, Natural Hazards and Earth System Sciences, 20, 967–979, https://doi.org/10.5194/nhess-2019-393, 2020.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrology and Earth System Sciences, 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

Moussa, R. and Chahinian, N.: Comparison of different multi-objective calibration criteria using a conceptual rainfall-runoff model of flood events, Hydrology and Earth System Sciences, 13, 519–535, https://doi.org/10.5194/hess-13-519-2009, 2009.

Nash, J. E. and Sutcliffe, I. V.: River flow forecasting through conceptual models Part I - A discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

National Weather Service NOAA: Conceptualization of the Sacramento Soil Moisture accounting model, Tech. rep., NOAA, https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/23sacsma.pdf, 2002.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set char-

acteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Pool, S., Vis, M. J., Knight, R. R., and Seibert, J.: Streamflow characteristics from modeled runoff time series - Importance of calibration criteria selection, Hydrology and Earth System Sciences, 21, 5443–5457, https://doi.org/10.5194/hess-21-5443-2017, 2017.

Prudhomme, C., Parry, S., Hannaford, J., Clark, D. B., Hagemann, S., and Voss, F.: How well do large-scale models reproduce regional hydrological extremes: In Europe?, Journal of Hydrometeorology, 12, 1181–1204, https://doi.org/10.1175/2011JHM1387.1, 2011.

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., and Samaniego, L.: Diagnostic evaluation of large-domain hydrologic models calibrated across the Contiguous United States, Journal of Geophysical Research: Atmospheres, 124, 13 991–14 007, https://doi.org/10.1029/2019JD030767, 2019.

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., Hamilton, D. P., Jeppesen, E., Kjellström, E., Olesen, J. E., Sonnenborg, T. O., Trolle, D., Willems, P., and Christensen, J. H.: A framework for testing the ability of models to project climate change and its impacts, Climatic Change, 122, 271–282, https://doi.org/10.1007/s10584-013-0990-2, 2014.

Risser, M. D., Paciorek, C. J., Wehner, M. F., O'Brien, T. A., and Collins, W. D.: A probabilistic gridded product for daily precipitation extremes over the United States, Climate Dynamics, 53, 2517–2538, https://doi.org/10.1007/s00382-019-04636-0, 2019.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 1–25, https://doi.org/10.1029/2008WR007327, 2010.

Schlef, K. E., Moradkhani, H., and Lall, U.: Atmospheric circulation patterns associated with extreme United States floods identified via machine learning, Scientific Reports, 9, 1–12, https://doi.org/10.1038/s41598-019-43496-w, 2019.

Sikorska, A. E., Viviroli, D., and Seibert, J.: Effective precipitation duration for runoff peaks based on catchment modelling, Journal of Hydrology, 556, 510–522, https://doi.org/10.1016/j.jhydrol.2017.11.028, 2018.

Sikorska-Senoner, A. E., Schaefli, B., and Seibert, J.: Downsizing parameter ensembles for simulations of extreme floods, Natural Hazards and Earth System Sciences Discussions, p. under review, https://doi.org/10.5194/nhess-2020-79, 2020.

Te Linde, A. H., Aerts, J., Dolman, H., and Hurkmans, R.: 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, 313, pp. 278–285, 2007.

Thirel, G., Andréassian, V., and Perrin, C.: De la nécessité de tester les modèles hydrologiques sous des conditions changeantes, Hydrological Sciences Journal, 60, 1165–1173, https://doi.org/10.1080/02626667.2015.1050027, http://dx.doi.org/10.1080/02626667.2015.1050027, 2015.

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

Thornton, P., Thornton, M., Mayer, B., Wilhelmi, N., Wei, Y., and Cook, R.: Daymet: daily surface weather on a 1 km grid for North America, 1980-2012, 2012.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resources Research, 43, 1–16, https://doi.org/10.1029/2005WR004723, 2007.

Unduche, F., Tolossa, H., Senbeta, D., and Zhu, E.: Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed, Hydrological Sciences Journal, 63, 1133–1149, https://doi.org/10.1080/02626667.2018.1474219, 2018.

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

USGS: USGS Water Data for the Nation, https://waterdata.usgs.gov/nwis, 2019.

Viglione, A. and Parajka, J.: TUWmodel: Lumped/Semi-Distributed Hydrological Model for Education Purposes, https://cran.r-project.org/web/packages/TUWmodel/index.html, 2020.

Vormoor, K., Lawrence, D., Heistermann, M., and Bronstert, A.: 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, https://doi.org/10.5194/hess-19-913-2015, 2015.

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

Wood, A. W., Leung, L. R., Sridhar, V., and Lettenmaier, D. P.: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs, Climatic Change, 62, 189–216, https://doi.org/10.1023/B:CLIM.0000013685.99609.9e, 2004.

World Meteorological Organization: Manual on flood forecasting and warning, Tech. Rep. 1072, WMO, Geneva, 2011.